MONASH University

A Bioinformatics Study of Protein Folding, Aggregation and Disease

Chen Li
M.Eng. (*Computer Science*)

A thesis submitted for the degree of Doctor of Philosophy at
Monash University in 2015
Department of Biochemistry and Molecular Biology

Copyright notice

Table of contents

Summary

The objective of this PhD thesis is to provide insightful analyses and biological data resources of protein structural and sequence features that are strongly associated with protein function and disease. In terms of protein sequence features, my research focuses on data collection, analysis and knowledgebase construction in order to allow the generation of new hypothesis on protein functions and follow-up studies of human diseases. Protein structure and disorder are introduced in Chapter 1. Chapter 2 focuses on analysing protein sequence features of human polyglutamine (polyQ) proteins that contain consecutive glutamine repeats in their sequences. A number of studies have demonstrated that expanded polyQ repeats are responsible for human neurodegenerative disease including Huntington disease and spinocerebellar ataxia. Building upon the previously published PolyQ database, an updated database, named PolyQ 2.0, has been constructed by incorporating functional and structural annotations for human disease- and non-disease associated polyQ proteins. Chapter 3 describes a novel knowledge-base, 'KinetochoreDB', a relational database that describes kinetochore and its related proteins. Kinetochore plays a crucial role during cell mitosis and meiosis by pulling sister chromatids apart. A number of disease-associated mutations have been verified and located with the kinetochore and its related proteins. It is envisaged that this new database will be useful for studies of kinetochore proteins and related diseases.

From a protein structure perspective, two important structural features have been investigated: protein coiled-coil domains (CCDs) and disordered regions. CCD is a type of protein tertiary structure consisting of an ensemble of helices binding together. It has been estimated that approximately 10% of eukaryotic proteins contain CCDs. Previous reports have revealed that some mutations occurring within the CCDs are responsible for human diseases due to the instability of CCDs caused by these mutations. On the other

hand, CCDs have been widely used as drug delivery systems due to their capability for molecular binding and recognition. From a bioinformatics perspective, this thesis mainly focused on analysing computational approaches for accurate identification of CCDs and their oligomeric states. Chapter 4 presents a comprehensive and critical performance evaluation of 12 currently available computational approaches for protein CCD and oligomeric state prediction, using carefully curated independent test datasets. A study of nine human polyQ disease-associated proteins was also performed to illustrate the prediction inconsistency amongst different CCD and oligomeric state predictors and highlighted the useful directions for development of improved predictors.

Intrinsically disordered proteins (IDPs) lack stable and well-defined three-dimensional structures. Proteins with disordered regions are often biologically important and play crucial roles in molecular binding and recognition, protein post-translational modification, protein regulation and other important biological processes. Using well-annotated datasets of human disease-associated mutations and state-of-art computational algorithms for protein disorder prediction, I investigated four different types of structural transitions due to single point mutations, all underlying human pathogenic mutations and polymorphisms; Disorder-to-Order (D→O), Disorder-to-Disorder (D→D), Order-to-Disorder (O→D) and Order-to-Order (O→O). Chapter 5 presents a bioinformatics analysis of these four structural transitions and proposes a mechanism, named 'structural capacitance' that may lead to *de novo* generation of microstructure in previously disordered regions (for D→O structural transition).

A summary, discussion and future directions of all the topics covered in this thesis are provided in Chapter 6.

Declaration of Authenticity

In accordance with Doctorate Regulations 17.2 of Monash University, the following declarations are made:

I hereby declare that, except where otherwise stated, this thesis contains no material which has been accepted for the award of any other degree or diploma at any University or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

Chapter 2 presents a submitted manuscript co-authored with Jeremy Nagel, Steve Androulakis, Jiangning Song and Ashley M. Buckle. The proportional contribution of each co-author involved in this manuscript has been declared in the following declaration of Chapter 2. The sections of this submitted paper have been renumbered in order to generate a consistent presentation within the thesis. The original PDF copy of this submitted manuscript has been attached in the "Publications" section.

Chapter 3 presents a submitted manuscript that is currently under revision and co-authored with Steve Androulakis, Ashley M. Buckle and Jiangning Song. The proportional contribution of each co-author involved in this manuscript has been declared in the following declaration of Chapter 3. The sections of this submitted paper have been renumbered in order to generate a consistent presentation within the thesis. The original PDF copy of this submitted manuscript has been attached in the "Publications" section.

Chapter 4 presents an article (Li *et al.,* 2015) co-authored with Catherine Chin Han Chang, Jeremy Nagel, Benjamin T. Porebski, Morihiro Hayashida, Tatsuya Akutsu, Jiangning Song and Ashley M. Buckle. The proportional contribution of each co-author involved in this article has been declared in the following declaration of Chapter 4. The sections of this published paper have been renumbered in order to generate a consistent presentation within the thesis. The original PDF copy of this published article (Li *et al.,* 2015) has been attached in the "Publications" section. I contributed to the publication of one compared coiled-coil oligomeric state predictor, namely RFCoil (Li *et al.* 2014). Even though my contribution is less than 50%, this publication is significant to my thesis. Therefore, I have listed this paper as one of my publications and appended this paper in the Publications.

Chapter 5 contains a perspective paper, which is currently under preparation. This chapter has been re-written in a traditional chapter style. The manuscript will be submitted for peer review once the experiments in progress are completed.

Chen Li

Date: 31/8/2015

General Declaration

Declaration for thesis based or partially based on conjointly published or unpublished work

In accordance with Monash University Doctorate Regulation 17.2 Doctor of Philosophy and Research Master's regulations the following declarations are made:

I hereby declare that this thesis contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

This thesis includes 1 original papers published in peer reviewed journals and 2 unpublished publications. The core theme of the thesis is 'a bioinformatics study of protein folding, aggregation and disease'. The ideas, development and writing up of all the papers in the thesis were the principal responsibility of myself, the candidate, working within the Department of Biochemistry and Molecular Biology under the supervision of A/Prof Ashley Buckle.

The inclusion of co-authors reflects the fact that the work came from active collaboration between researchers and acknowledges input into team-based research.

In the case of Chapters 2, 3 and 4, my contribution to the work involved the following:

Thesis chapter	Publication title	Publication status*	Nature and extent of candidate's contribution
2	**PolyQ 2.0: an Updated Database of Human Polyglutamine Proteins**	**Under review**	**Data collection, database construction, Paper writing and revision (90%)**
3	**KineotchoreDB: a Comprehensive Online Resource for the Kinetochore ar Its Related Proteins**	**Under revision**	**Data collection, database construction, paper writing and revision (90%)**
4	**Critical Evaluation of *in silico* Methods for Prediction of Coiled-coil Domains in Proteins**	**Published**	**data analysis, model testing, paper writing and revision (85%)**

I have renumbered sections of submitted or published papers in order to generate a consistent presentation within the thesis.

Signed: **……**  **……**

Date: **…31/8/2015……**

Declaration of Thesis Chapter 2

Declaration by candidate

In the case of Chapter 2, the nature and extent of my contribution to the work was the following:

Nature of contribution	Extent of contribution (%)
Data collection and analysis, Database construction, Paper writing and revision	90

The following co-authors contributed to the work. If co-authors are students at Monash University, the extent of their contribution in percentage terms must be stated:

Name	Nature of contribution	Extent of contribution (%) for Student co-authors only
Jeremy Nagel	Data analysis	
Steve Androulakis	Paper writing for database techniques description	
Jiangning Song*	Supervised the study, Paper revision	
Ashley M. Buckle*	Designed and supervised the study, Paper revision	

The undersigned hereby certify that the above declaration correctly reflects the nature and extent of the candidate's and co-authors' contributions to this work*.

Candidate's Signature	[redacted]	**Date 31/8/2015**
Main Supervisor's Signature	[redacted]	**Date 31/8/2015**

*Note: Where the responsible author is not the candidate's main supervisor, the main supervisor should consult with the responsible author to agree on the respective contributions of the authors.

Declaration of Thesis Chapter 3

Declaration by candidate

In the case of Chapter 3, the nature and extent of my contribution to the work was the following:

Nature of contribution	Extent of contribution (%)
Data collection and analysis, Database construction, Paper writing and revision	90

The following co-authors contributed to the work. If co-authors are students at Monash University, the extent of their contribution in percentage terms must be stated:

Name	Nature of contribution	Extent of contribution (%) for student co-authors only
Steve Androulakis	Paper writing for database techniques description	
Ashley M. Buckle*	Supervised the study, Paper revision	
Jiangning Song*	Designed and supervised the study, Paper revision	

The undersigned hereby certify that the above declaration correctly reflects the nature and extent of the candidate's and co-authors' contributions to this work*.

Candidate's Signature	[redacted]	**Date 31/8/2015**

Main Supervisor's Signature	[redacted]	**Date 31/8/2015**

*Note: Where the responsible author is not the candidate's main supervisor, the main supervisor should consult with the responsible author to agree on the respective contributions of the authors.

Declaration of Thesis Chapter 4

Declaration by candidate

In the case of Chapter 4, the nature and extent of my contribution to the work was the following:

Nature of contribution	Extent of contribution (%)
Data analysis, model testing, paper writing and revision	85

The following co-authors contributed to the work. If co-authors are students at Monash University, the extent of their contribution in percentage terms must be stated:

Name	Nature of contribution	Extent of contribution (%) for student co-authors only
Catherine Ching Han Chang	Paper section writing	2
Jeremy Nagel	Case study	
Benjamin T. Porebski	Sequence blast	1
Morihito Hayashida	Sequence blast	
Tatsuya Akutsu	Sequence blast	
Jiangning Song*	Designed and supervised the study, paper revision	
Ashley M. Buckle*	Designed and supervised the study, paper revision	

The undersigned hereby certify that the above declaration correctly reflects the nature and extent of the candidate's and co-authors' contributions to this work*.

Candidate's Signature	[redacted]	**Date 31/8/2015**

Main Supervisor's Signature	[redacted]	**Date 31/8/2015**

*Note: Where the responsible author is not the candidate's main supervisor, the main supervisor should consult with the responsible author to agree on the respective contributions of the authors.

List of Publications

Publications resulting from this thesis

Chen Li, Catherine Chin Han Chang, Jeremy Nagel, Benjamin T. Porebski, Morihiro Hayashida, Tatsuya Akutsu, Jiangning Song and Ashley M. Buckle. **Critical Evaluation of *in silico* Methods for Prediction of Coiled-coil Domains in Proteins**. *Briefings in Bioinformatics*, 2015, DOI: 10.1093/bib/bbv047. (IF: 9.617 – Q1 in Biochemical Research Methods)

Chen Li, Xiao-Feng Wang, Zhen Chen, Ziding Zhang and Jiangning Song. **Computational Characterisation of Parallel Dimeric and Trimeric Coiled-coils Using Effective Amino Acid Indices**. *Molecular BioSystems*, 2015 Feb; 11(2): 354-60, DOI: 10.1039/C4MB00569D (IF: 3.183 - Q2 in Biochemistry and Molecular Biology)

Chen Li, Steve Androulakis, Ashley M. Buckle and Jiangning Song. **KinetochoreDB: a Comprehensive Online Resource for the Kinetochore and Its Related Proteins**. Submitted to *Database* (under revision)

Chen Li, Jeremy Nagel, Steve Androulakis, Jiangning Song and Ashley M. Buckle. **PolyQ 2.0: an Updated Database of Human Polyglutamine Proteins**. Submitted to *Database* (under review)

Other publications

Chen Li and Xue-Liang Hua. **Towards Positive Unlabeled Learning for Parallel Data Mining: a Random Forest Framework**. **(Best paper award)** The 10th International Conference on Advanced Data Mining and Applications (ADMA'14), 21st December, Guilin, China, DOI: 10.1007/978-3-319-14717-8_45.

Arash Arjomand, Mark A. Baker, **Chen Li**, Ashley M. Buckle, David A. Jans, Kate L.

Loveland and Yoichi Miyamoto. **The α-importome of Mammalian Germ Cell Maturation Provides Novel Insights for Importin Biology**. *FASEB J*, 28(8): 3480-3493, DOI: 10.1096/fj.13-244913, 2014 (IF: 5.704 - Q1 in Biology and Biochemistry)

Oral presentations

Presenting author underlined

Chen Li and Xue-Liang Hua. (2014) **Towards Positive Unlabeled Learning for Parallel Data Mining: a Random Forest Framework**. The 10th International Conference on Advanced Data Mining and Applications (ADMA'14), 21st December, Guilin, China.

Chen Li, Adrian Woolfson, Jiangning Song and Ashley M. Buckle. (2012) **Structural Capacitance in Protein Evolution and Human Diseases**. 3rd ISCB-RSG Australia 2012 Student Symposium, 13th December, Melbourne.

Poster presentations

Presenting author outlined

Chen Li, Adrian Woolfson, Jiangning Song and Ashley M. Buckle. (2013) **Structural Capacitance in Protein Evolution and Human Diseases**. The 21st Annual International Conference on Intelligent Systems for Molecular Biology/12th European Conference on Computational Biology (ISMB/ECCB), Berlin, Germany, 2013

Chen Li, Adrian Woolfson, Jiangning Song and Ashley M. Buckle. 2012. **A Bioinformatics Study of Casual Relationships Between Disorder-order Transitions Induced by Point Mutation and Human Disease**. 2012 Biochemistry and Molecular Biology Postgraduate Research Conference, Monash University, Victoria, Australia, 22nd – 23rd, November, 2012.

Chen Li, Jeremy Nagel, Amy L. Robertson, M. A. Bate, Steve G. Androulakis,

Stephen P. Bottomley, Jiangning Song and Ashley M. Buckle. 2012. **A Bioinformatics Investigation of the Role of Polyglutamine Repeats in Disease**. 37th Lorne Conference on Protein Structure and Function, Lorne, Victoria, Australia, February, 2012.

Acknowledgements

Foremost, I would like to express my great appreciation and thanks to my supervisors A/Prof. Ashley Buckle and Dr. Jiangning Song for their continual and generous help during my PhD study. As a student from computer science and without any biological background, I have experienced lots of difficulties regarding my PhD research project. However, it is their tremendous mentorship that has helped me through every milestone. They have taught me not only the knowledge, but also, more importantly, how to be an independent researcher. What I have learned from them will undoubtedly be my invaluable treasure for my career and my life. Specially, I would like to thank Dr. Jiangning Song for the inspiring discussions on development of my scientific career. His wisdom and experiences have brightened the future of my career.

I would like to thank the past and present members of the Buckle Lab: Itamar Kass, Bindu Jayakrishnan, Olga Ilyichova, David Hoke, Benjamin Porebski, Blake Riley, Emilia Marijanovic, Melissa Honeydew, Sarah Le and Sebastian Brøndum for sharing their knowledge and for offering their help with my English and presentations. Special thanks go to my co-workers, Itamar Kass, Bindu Jayakrishnan, David Hoke and Ben Porebski for showing me around the lab and offering me a warm welcome when I arrived at Monash University four years ago.

I also want to express my thanks to my PhD committee for the help and guidance on my PhD research project. The members of my PhD committee are Prof. Mibel Aguilar, Prof. Matthew Wilce and A/Prof. Martin Stone. I would like to express my appreciation for their great support and suggestions during my first-year confirmation, mid-candidature review and pre-submission milestone seminars. I also want to thank Mrs. Pip Carmen and Mrs. Liz Kemp for their hard work and time on organizing my milestone seminars. I

greatly appreciate the helpful and warm discussions with Prof. Mibel Aguilar on my PhD study and life in Melbourne.

I would like to thank my collaborators working with my PhD project; it is my great honour to work with them. I would like to thank Adrian Woolfson and Ben Porebski for the inspiring discussions on the structure capacitance paper, Steve Androulakis for the two database papers, Jeremy Nagel, Catharine Chang, Morihiro Hayashida and Tatsuya Akutsu for the evaluation of coiled-coil prediction algorithms. Their efficient effort and work facilitated the fast publications during my PhD study.

"A good friend is my nearest relation". Life is never easy for international students. I would like to thank my good friends, Wei Yang, Steve Heaton, Guojie Wu, Jackie Mao and Sue Ekkel for their help during my scientifically dark days. Their words made me feel warm and helped me continue with my work on the 'rocky' science road.

I would like to thank my scholarship provider, Chinese Scholarship Council (CSC) and Monash University, for providing me such a great opportunity to pursue my PhD degree here in Monash University.

Last but not least, I would like to express my great thanks to my parents for their support and encouragement during my PhD study. Calling my parents on my way back home after a whole day at work was the most enjoyable time. Talking with them made me feel relieved and cheerful. They made me understand the importance of being optimistic when I encountered hurdles. Their wisdom always motivated me to find the solutions to my difficulties both in research and in life.

"Now we are not afraid, although we know there's much to fear.

We were moving mountains long before we knew we could."

-'When you believe'

By Mariah Carey and Whitney Houston

List of Abbreviations

Å	Angstrom
AUC	Area Under the Curve
CCD	Coiled-coil Domain
CD	Circular Dichroism
D→D	Disorder-to-Disorder
D→O	Disorder-to-Order
IDP	Intrinsically Disordered Protein
MoRFs	Molecular Recognition Feature
LDR	Long Disordered Region
MSA	Multiple Sequence Alignment
NMR	Nuclear Magnetic Resonance
O→D	Order-to-Disorder
O→O	Order-to-Order
PolyQ	Polyglutamine
ProS	Protean Segment
SDR	Short Disordered Region
SVM	Support Vector Machine

List of Figures

List of Tables

Amino Acids Abbreviations

Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic Acid	Asp	B
Cysteine	Cys	C
Glutamic Acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

Chapter 1 Introduction

The aim of this thesis is to develop bioinformatics resources and perform insightful analyses of protein sequence and structural features, which are highly associated with protein functions and human diseases. Two particular types of protein structural features are investigated: disordered regions (Section 1.1) and coiled-coil regions (Section 1.2). For protein sequence features, this thesis focuses on building relational biological knowledgebases for polyglutamine (polyQ) proteins (Section 1.3), and kinetochore and related proteins (Section 1.4), to shed light on the sequence-structure-function relationship of these proteins and their association with human diseases. The aims of this thesis are described and discussed in detail in Section 1.5.

1.1 Protein structure and intrinsically disordered regions (IDRs)

1.1.1 Protein structure and folding

Protein structure is the functional, 3-dimensional folded form adopted by its amino acid sequence. The amino acid sequence enables proteins to fold into many different types of structures to perform a wide array of biological functions. As such, protein function is dictated by its structure, which in turn is dictated by its sequence: this can be referred to as the 'sequence-structure-function' paradigm [1]. Four kinds of structures, including primary, secondary, tertiary and quaternary structure, will be discussed.

The sequence of amino acids in the protein chain is described as the protein primary structure. There are a total of 20 common and natural amino acids and they vary in terms of size, charge and hydrophobicity. The most common protein secondary structures include alpha-helix and beta sheet, representing the local conformations of peptides. The alpha-helix was firstly described by Pauling *et al.* in 1951 [2], and is a spiral conformation. The average length of a typical alpha-helix is 10 amino acids and there are on average 3.6 amino acids in each turn of the helix [2]. Beta sheets consist of several beta strands held together by hydrogen bonds. There are three types of beta sheet in terms of the orientation: parallel beta sheet, antiparallel beta sheet and mixed based on their directions. Random coil, on the other hand, is not a true type of secondary structure, but a conformational that lacks regular secondary structure. The examples of alpha-helix, parallel/antiparallel/mix beta sheets and random coil are shown in Figure 1.1. Protein tertiary structure can be regarded as an 'ensemble' of two or more protein secondary structure units. For instance, the coiled-coil region is a type of tertiary structure, which has two or more alpha-helices [3]. Another example of protein tertiary structure is the beta-sandwich architecture, which

consists of two or more beta sheets. Figure 1.2 shows examples for coiled-coil and beta-sandwich architecture, respectively. Protein quaternary structure is a multi-unit complex with folded protein subunits from more than one polypeptide chains [4]. The quaternary state of proteins is important for protein function [5].

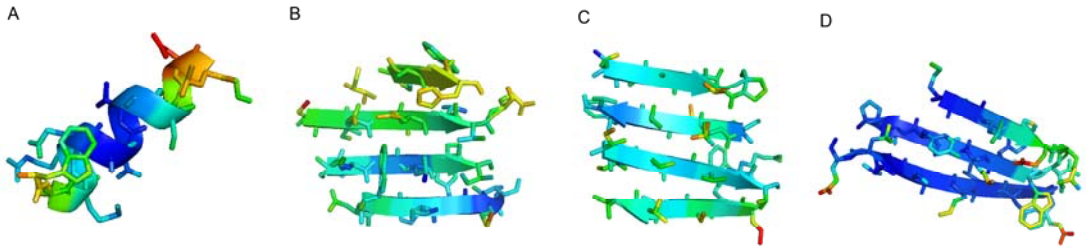


Figure 1.1 Examples of protein secondary structures including (A) alpha-helix (PDB: 2XRC – Human complement factor I [6]); (B) parallel beta sheet (PDB: 1JQD – Human histamine methyltransferase complexed with adoHcy and histamine [7]); (C) antiparallel beta sheet (PDB: 1DZO – Truncated PAK pilin from *Pseudomonas aeruginosa* [8]) and (D) mix of parallel and antiparallel beta sheet (PDB: 1O22 – Crystal structure of an orphan protein from *Thermotoga maritima* [9]).

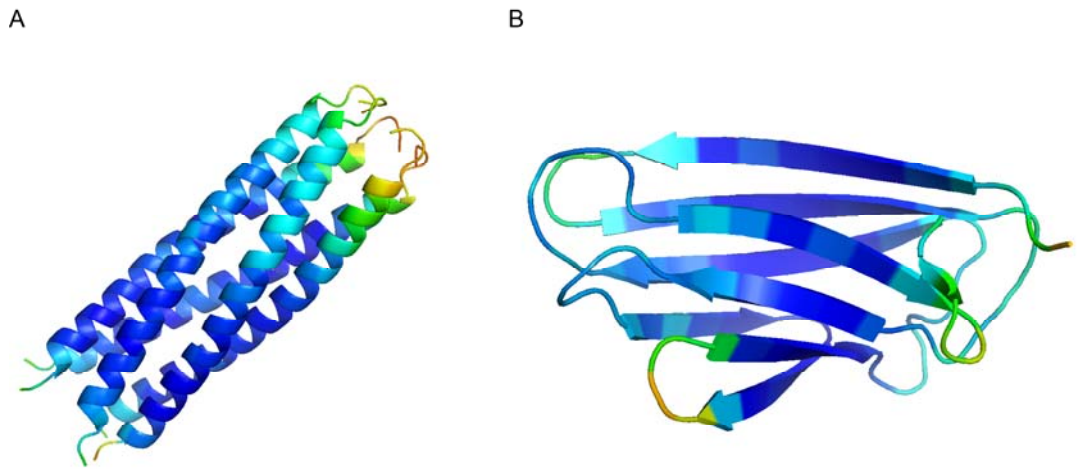


Figure 1.2 Examples of protein tertiary structure. (A) Coiled-coil (PDB: 1MZ9 – The crystal structure of the coiled-coil domain in complex with vitamin D3.) [10] and (B) Beta-sandwich architecture (PDB: 2CWR – Crystal structure of chitin binding domain of chitinase from *Pyrococcus furiosus*) [11].

Protein folding is a complex biological process where a linear chain of amino acids that adopts no defined structure folds into a functional and well-defined structure [12, 13]. The 'Levinthal Paradox', showed that due to the large configurational space available to each amino acid, for a protein of 100 amino acids it would take approximately 10^{27} years to exhaust all the possible conformational states if protein folding occurred by a purely random search [14, 15]. In reality, most proteins fold spontaneously on timescales of seconds or less. This paradox has been resolved by observing that folding is initiated and guided by local interactions, dictated by a 'funnel-like energy landscape' [16, 17]. When the free energy decreases to favor the folding, the number of attempts to search for the folding states will significantly decrease, forming a 'folding funnel' [14, 17]. It is now clear that protein folding in the cell is a highly complex physical process that is tightly controlled and regulated by many other factors, for example chaperones and post-translational modifications [18].

1.1.2 Intrinsically disordered proteins (IDPs) and disordered regions

Intrinsically disordered proteins (IDPs) lack stable well-defined 3D structures [19-23]. Protein disordered regions, on the other hand, refer to the consecutive regions of amino acids that are disordered. It is estimated that nearly one third of eukaryotic proteins contain disordered regions of longer than 50 consecutive amino acids [1, 20]. Normally, disordered regions longer than 30 amino acids are referred to as long disordered regions (LDRs), while disordered regions shorter than 30 amino acids are regarded as short disordered regions (SDRs) [20, 24-27]. The complexity of disordered region is relatively low and favours hydrophilic amino acids [19, 23]. This thesis focuses mainly on LDRs

rather than IDPs. The mutations located in LDRs and their structural impacts are investigated in Chapter 5.

1.1.3 Function of IDPs

The well-established sequence-structure-function paradigm has been used to illustrate that the protein structure is the obligatory prerequisite for protein function [1]. However, it is difficult to apply this paradigm to explain the function of IDPs (proteins with disordered regions) or disordered regions, given that they lack well-defined structures.

Regarding the function of protein disordered regions, Dunker *et al.* [28] has proposed a detailed classification scheme with 28 function features based on careful review on 150 proteins with LDRs. To summarise, I classify these function features into the following categories:

a. Molecular binding/recognition

   (1) protein-protein binding, (2) protein-DNA binding, (3) protein-rRNA binding, (4) protein-tRNA binding, (4) protein-mRNA binding, (5) protein-genomic RNA binding, (6) protein-lipid interaction, (7) polymerization, (8) cofactor/heme binding, (9) metal binding and (10) substrate/ligand binding;

b. Post-translational modification

   (1) phosphorylation, (2) acetylation, (3) glycosylation, (4) methylation, (5) fatty acylation (myristolation and palmitoylation) and (6) ADP-ribosylation;

c. Regulatory

   (1) autoregulatory and (2) regulation of proteolysis *in vivo*;

d. Entropy

   (1) entropic spring, (2) entropic bristle and (3) entropic clock

e. DNA re-shaping

(1) DNA unwinding and (2) DNA bending;

f. Others

(1) flexible linkers/spacers, (2) structural mortar, (3) transport thru channel, (4) protein detergent and (5) unknown.

Both experimental and computational studies have confirmed that protein disordered regions harbour a variety of post-translational modification sites (PTMs) [29-32], indicating the biological significance of disordered regions. On the other hand, protein disordered regions also serve as binding interfaces with different types of binding partners due to the highly dynamic nature of disordered regions, thereby mediating protein function.

1.1.4 IDPs, mutations and diseases

Given the biological significance of disordered regions, it is not surprising that mutations occurring within these disordered regions can be pathogenic. A number of studies have focused on investigating the relationship between protein disordered regions and human disease [21, 22, 33-36]. The common conclusion is that the functional disruptions within the disordered regions are responsible for diseases. Notably, in [21], the authors built a map of different disease classes and the protein disorder content of their associated proteins, highlighting the necessity of focusing on the different content of disordered regions in different disease categories.

1.1.5 Human disease mutation datasets

Multiple databases with human disease mutation annotations are currently available for bioinformatics studies and experimental investigation.

Gene level

Human Gene Mutation Database. HGMD (The Human Gene Mutation Database; http://www.hgmd.cf.ac.uk/ac/index.php) [37] is the most popular genetic mutation database, which contains 121,002 entries with 67,439 nonsense mutations for academic use as of 30-08-2015. In addition, HGMD also provides disease annotation for its data entries.

Protein Level

Human polymorphisms and disease mutations dataset [38]. This dataset (http://www.uniprot.org/docs/humsavar) contains 69,141 human mutations including 24,646 human disease mutations, 37,931 polymorphisms mutations and 6,564 unclassified mutations. Disease mutations have been annotated by the disease names and their OMIM (Online Mendelian Inheritance in Man; www.omim.org/) accession numbers. The annotated disease mutations are assigned according to literature reports on probable disease association, including those based on theoretical reasons. For each mutation, this dataset provides detailed information, including UniProt [39] (http://www.uniprot.org/) accession number of original protein, mutated position, wild-type and mutated amino acid and mutation type (i.e., disease, polymorphism and unclassified). This dataset is being updated every four weeks.

SNPeffect database. With a total number of 63,410 human mutations, SNPeffect [40] (http://snpeffect.switchlab.org/) focuses on both functional and structural effects the mutations bring to proteins. This database employed several computational approaches to predict changes in protein aggregation (using TANGO [41]), amylogenicity (using WALTZ [42]), chaperone binding (using LIMBO [43]), and structural profile (using FOLDX [44]).

MSV3d database [45] (currently unavailable). This database was released in the beginning of 2012. There are several highlights in this database comparing with other similar datasets and databases: (1) more mutations are involved in MSV3d. Mutations in

this database are from multiple data sources: dbSNP [46] and UniProt. In total, there are 445,574 mutations and 20,199 proteins are involved (according to the version of 09-Jan-2012) and (2) if possible, the structure of the original protein is also made available. Currently, there are 10,713 3D homology models available in the database (according to the version of 09-Jan-2012). For each mutation, some physical-chemical and structural changes are also provided.

1.1.6 Computational resources for disorder-to-order transition upon binding

A number of studies [47-50] have been performed to investigate the disorder-to-order transition upon binding to different types of partners including peptide [51, 52] and DNA [53]. There are also several databases and computational methods for disorder-to-order transition through binding and protein disordered binding region.

Databases for disorder-to-order transition through binding

ComSin. ComSin (http://antares.protres.ru/comsin/) [54] is a specialized database that focuses on the bound and unbound analysis in disordered regions, since disorder-to-order (D→O) and order-to-disorder (O→D) transitions can be triggered by binding to partners. In total, the database has 24,910 pairs of homologous proteins observed in unbound and bound states. This database is a good resource for analysis of the bound and unbound states in disordered regions.

Computational methods for protein disordered binding region prediction

ANCHOR. To the best of my knowledge, ANCHOR [55] is the first sequence-based framework to predict the disordered binding regions. ANCHOR aims at finding regions in disordered areas that can bind and interact with a globular protein partner. These regions are called disordered binding regions. Based on pairwise energy estimation approach (which is also the basis of IUPred [56]), ANCHOR can find long binding sites in

disordered regions. Many of these predicted binding sites have been validated by experiments. Moreover, the performance of ANCHOR is independent from amino acid composition and secondary structure.

MoRFpred. Located in disordered regions, MoRFs (Molecular Recognition Features) are short regions that bind with other binding partners leading to disorder-to-order transitions [57]. Since only a limited number of annotated and experimentally verified data for MoRFs is available, the computational method for predicting such regions in disordered regions is necessary and important. This predictor was designed to find MoRFs in the disordered regions especially in long disorder regions. Using Support Vector Machines (SVMs), the framework collects different features including amino acids indices, predicted disordered regions, relative solvent accessibility, B-factor and PSSM (Position Specific Scoring Matrix) with the help of different predictors. An empirical study showed that MoRFpred outperformed ANCHOR [55] in terms of accuracy and AUC (Area Under Curve).

1.1.7 Current bioinformatics studies of disorder prediction

The importance of protein disordered regions has motivated the development of a number of computational approaches to facilitate fast prediction of sequence-based protein disordered region. In this section, several popular and well-established predictors for disordered region are introduced.

IUPred (http://iupred.enzim.hu/). IUPred [56, 58] maintains two versions of IUPred including IUPred-S and IUPred-L. Here, 'S' and 'L' refer to the long LDRs and SDRs, respectively. For 'S' option, the model was trained using a dataset corresponding to missing residues in the protein structures. These residues are missing in the protein structures due to the missing electron density in the X-ray crystal structures. These

disordered regions are usually short. Conversely for the 'L' option, the dataset used to train models corresponds to long disordered regions from the Protein Database [59] that are validated by various experimental techniques such as X-ray crystallography, NMR etc. According to the instruction for IUPred, residues with scores above 0.5 can be regarded as disordered.

PONDR-VSL2B. VSL2B [60] is a widely used sequence-based predictor for intrinsically disordered regions. It has achieved outstanding prediction performance for disordered regions prediction [61] using SVM. Amino acid with score above 0.5 are considered as being disordered.

DynaMine. Trained with carefully curated nuclear magnetic resonance (NMR) dataset. DynaMine [62] aims to accurately predict protein disordered regions with only sequence information as the input. The empirical results demonstrate that the performance is comparable with other algorithms such as IUPred [56], ESpritz [63] and RONN [64].

1.1.8 Current databases/resources on IDPs

Given the important biological function of intrinsically disordered proteins and protein disordered regions, several computational and experimental resources have been published to facilitate in-depth investigation of protein disorder and computational tools for protein disordered region prediction.

DisProt (http://www.disprot.org/index.php**) [65].** To the best of our knowledge, DisProt is the first database harbouring experimentally verified intrinsically disorder proteins and disordered regions (verified at least once). In addition, for each entry in this database, DisProt provides detailed function classification, function description and experimental evidence. The advantage of this database is that the disordered regions harboured in DisProt are all experimentally confirmed results. Therefore, the results are

reliable so that the database could be a useful resource for development of new disordered region predictors. There are 1,539 disordered regions and 694 proteins involved (according to the version of 24-May-2013).

IDEAL (http://www.ideal.force.cs.is.nagoya-u.ac.jp/IDEAL/) [66, 67]. IDEAL (Intrinsically Disordered proteins with Extensive Annotation and Literature) also contains protein disordered regions annotated based on experimental results. Moreover, by defining ProS (Protean Segments) using available experimental evidence, this database also annotates small flexible regions that are likely to bind with their partners (i.e., the disordered regions supported by experimental evidence for both the isolated disordered state and the ordered partner-bound state).

D^2P^2 (http://d2p2.pro/) [68]. Unlike the two databases introduced above, D^2P^2 is an online knowledgebase for protein disordered regions prediction results using nine different tools for protein disorder prediction, including PONDR VLXT [19, 69], PONDR VSL2B [60], IUPred (short and long versions) [56, 58], PV2 [61], Espritz-D, Espritz-X and Espritz-N [63]. All the prediction results for a single entry of D^2P^2 have been integrated as ensemble display so that the users can easily find the predicted disorder agreement. In addition, in the updated version of D^2P^2 database, MoRF regions (predicted by ANCHOR) and PTM site annotations have been collected and visualized, which are helpful for further investigation of protein binding and function within the disordered regions.

1.2 Coiled-coils – an important type of protein tertiary structure

1.2.1 What is a coiled-coil domain (CCD)?

Coiled-coil domains (CCDs) consist of ***n*** (two or more) alpha-helices twisting around each other [70, 71]. CCD motif is ubiquitous and can be found in approximately 10% of eukaryotic proteins [72]. This supercoiled motif is represented by seven-residue pattern $[\boldsymbol{abcdefg}]_{\boldsymbol{n}}$, where ***a*** and ***d*** are predominantly hydrophobic residues, ***b*** and ***c*** are usually hydrophilic resiudes, ***e*** and ***g*** are charged residues according to the 'Peptide Velcro' hypothesis [73]. Depending on the number of helices binding the CCDs, CCDs can be further categorized into several groups, including antiparallel dimer, parallel dimer, trimer, and tetramer. Among these four kinds of coiled-coil oligermeric states, dimeric coiled-coil is the most prevalent type, while the number of tetramer coiled-coils is very small and limited number of proteins has been detected to contain tetrameric coiled-coils.

1.2.2 CCDs and diseases

CCDs have been revealed to play a fundamental role in many biological processes including subcellular infrastructure and controlling trafficking of eukaryotic cells [74, 75]. In addition, CCDs are highly versatile protein motifs that can function as molecular recognition system [76]. A recent computational study showed that CCDs also play an important role in the evolution of the centrosome and expand the function of centrosome, which is critical for cell division [77].

A number of experimental studies have revealed that the mutations occurring within the CCDs are pathogenic. Mutations occurring within coiled-coil domain can cause diseases including neurodegenerative disease, progeria, cancer and severe skin fragility [75, 78-85], which are probably caused by the damage of the stability of the coiled-coil domains in corresponding proteins. For example, a mutation in the CCD of the KIF5A gene is associated with the late-onset hereditary spastic paraplegia [86]. Another example

shows that the primary ciliary dyskinesia can be caused by a nonsense mutation harboured in the coiled-coil domain of 151 gene [87].

1.2.3 Brief summary of current bioinformatics approaches for CCD prediction and design

Given a protein structure, it is now easy to detect the CCDs located in the current protein by applying computational tools. SOCKET [88], for example, is an easy-to-use software to identify CCDs in the given protein structures. With the help of SOCKET and manual annotation, a database, namely CC+ [89], has been proposed to provide experimentally verified CCDs and enable the protein CCD prediction by offering high quality training samples. CC+ database covers a wide range of CCDs with different oligermeric states, including dimer, trimer, tetramer, pentamer and hexamer. The CC+ database is freely available at http://coiledcoils.chm.bris.ac.uk/ccplus/search/.

A variety of computational tools have been proposed to perform CCD prediction based on protein sequences. There are basically two tasks for CCD prediction. The first task is to predict CCD with the given protein sequences. The next task, which is more advanced, is to identify the oligomeric states for the given CCDs. When predicting the oligomeric states for CCDs, it is also important to identify the helix orientation (i.e., parallel or antiparallel). Sequence-based predictors, including COILS [90], Paircoil [91, 92], CCHMM_PROF [93], MARCOIL [94], Spiricoil [95] and Multicoil [71, 96], have been designed and implemented for the first task; while LOGICOIL [97], SCORER2.0 [98], RFCoil [99], and PrOCoil [100] have been proposed to address the second task. Structural modelling/prediction methods have also been employed for dimeric helix orientation [101]. However, it is generally believed that the structural modelling based method is slower than the sequence-based methods due to the high computational complexity.

Given the advanced structural studies and the importance of CCDs in human disease and drug delivery systems, the development of computational approaches and algorithms for CCD design has been accelerated. The sequence-to-structure relationships have made the CCDs and their oligomeric state design straightforward (e.g., parallel dimer, antiparallel dimer, trimer and tetramer) [102]. A set of *de novo* coiled-coil peptides has been proposed for rational protein design based on literature and current databases with experimentally verified CCDs [103]. With the help of the current advances of bioinformatics and computational biology, more CCDs can be designed for different proteins and purposes [104, 105]. A coiled-coil designing tool, namely, CC Builder, has recently been implemented to facilitate fast CCDs design and future experimental investigation [106]. Note that the traditional protein secondary structure predictors and CCD design tools serve different purposes. Protein secondary structure predictors aim to predict protein secondary structures with given amino acid sequences. Whether the predictive secondary structure contains CCD or not is subject to CCD predictors and experimental investigations. CCD design tools help users create CCD domains they want, with help of specific CCD physical, chemical and structural properties and parameters. Therefore, the CCDs from design tools are expected to be more reliable and accurate.

1.3 Polyglutamine (PolyQ) proteins

1.3.1 What is a polyQ protein?

Repetitive protein sequences are ubiquitous and over 3% of human proteins contain the single amino acid repeats [107, 108]. The polyglutamine (polyQ) stretch, which is a common repeat in eukaryotic proteins [109], is a peptide with a consecutive tract of glutamine (Q) residues. PolyQ is a normal sequence feature of many human proteins

[110], indicating that polyQ repeats are biologically important. Recent studies have revealed that polyQ repeats play crucial roles in stabilizing protein-protein interaction [110] and functional modulation [111].

1.3.2 Expanded polyQ repeats and protein aggregation

Abnormally longer polyQ tracts are caused by the expanded tri-nucleotide sequence CAG repeats in the corresponding genes [112, 113]. These expanded polyQ tracts tend to interact with their coiled-coil partners, leading to protein aggregation [114]. Although there are many human polyQ-containing proteins [115], only nine proteins have been experimentally verified to cause diseases [116-118]. Several neurodegenerative diseases, including Huntington Disease, spinobulbar muscular atrophy and spinocerebellar ataxias, are caused by the pathogenic proteins with abnormal expansions of polyQ tract [119].

1.3.3 Current resources for polyQ protein annotations

To the best of my knowledge, the PolyQ database [120] is the only available data resource for human polyQ proteins. This database contains 128 human poly proteins, among which nine proteins are disease-associated proteins due to the expansion of polyQ repeats. In the PolyQ database, all the entries have been further categorised according to the distribution of polyQ tracts and Pfam (protein family) domains [121] to illustrate the context of polyQ repeats and their flanking Pfam domains. However, PolyQ database contains only basic information for each data entry, and lacks essential annotations for proteins in terms of function and structure. Therefore, much work is urgently needed to expand and update the PolyQ database to enrich the annotation for all the data entries.

1.4 Kinetochore and its related proteins

1.4.1 Kinetochore and its function

The kinetochore is the macromolecular complex that plays an important role during cell mitosis and meiosis by attaching on the chromosomes and pulling sister chromatids apart [122]. The centromeric chromatin is the place where the kinetochore is constructed during mitosis and meiosis. Kinetochore attaches on the chromosome to enable chromosome segregation [123]. There are two main regions for the kinetochore: outer kinetochore (outer plate) and inner kinetochore (inner plate). The outer kinetochore is mainly responsible for the binding/interacting the microtubules during cell division.

1.4.2 Kinetochore and disease

Kinetochore-microtubule (kMT) dynamics plays important roles in the cell cycle and evidence has been revealed that the deregulation of the kMT attachments is strongly related to diseases [124, 125]. A good example is the Mosaic Variegated Aneuploidy (MVA) Syndrome. The clinical features of this syndrome include mental and growth retardation and severe microcephaly [126, 127]. It is believed that this syndrome can be caused by the mutations in BubR1 protein, which is a serine/threonine kinase [124], by inducing instability to the kMT dynamics. On the other hand, besides germline mutations, some somatic mutations can putatively increase the stability of kMT dynamic in the overexpressed cells, leading to cancers [124]. For example, the somatic mutations occurring on proteins including CenpH, Cyclin E and MCT-1, are believed to increase the stability of kMT and eventually trigger the cancer [124, 128-130].

1.4.3 Current databases/resources for kinetochore and its related proteins

Despite its biological significance and the increasing awareness of its important roles in human diseases, there is currently a paucity of publically available databases or resources that focus on providing comprehensive functional annotations of the kinetochore and its related proteins. MiCroKiTS [131], for example, is an integrated online resource for kinetochore, midbody, telomere, centrosome and spindle proteins. However, important annotations on entries in MiCroKiTS are not available in terms of protein 3D structure, protein interaction partners, metabolic/signaling pathways etc., all of which are crucial aspects for follow-up functional studies of kinetochore and its related proteins.

1.5 Thesis aims

The motivation to understand the mechanisms of human disease related to protein sequence and structural features has been increasingly bringing together experimental and computational biology. In light of this, this thesis focuses on two areas. The first area describes the construction of data resources for proteins with specific sequence features. One important sequence feature is polyQ. Nine human proteins with expansion of polyQ tracts are closely associated with neurodegenerative diseases due to their high aggregation propensity. Hence, in order to provide useful resources for the polyQ proteins, this thesis develops a user-friendly web-based database for efficient storage and rapid search of these proteins supplemented with comprehensive structural and functional annotations. To further explore the sequence-structure-function relationship, this thesis also implements a novel biological database, namely 'KinetochoreDB' for kinetochore and its related proteins. Kinetochore plays an important role in cell mitosis and meiosis. Experimental studies have revealed a number of mutations that occur on the kinetochore proteins are associated with human diseases.

The second area of focus is on protein structural features that are associated with human disease, namely protein disordered regions and coiled-coil regions. In this regard, many experimental and bioinformatics studies of structural and functional properties in disease-associated mutations in proteins have been conducted. For protein disordered regions, I investigate the structural changes upon mutation in these regions and provide data with experimentally-testable candidates that cause disorder-to-order structural change. This analysis is then used to propose a novel hypothesis called 'structural capacitance'. In the case of protein coiled-coil regions, a comprehensive survey on current computational approaches is performed for protein CCD prediction. A comparison of their performance using carefully curated independent test datasets is also provided. This thesis also uses a specific case study dataset with nine human polyQ disease associated proteins to provide useful insights into sequence-structure-function relationship of such proteins.

More specifically, this thesis seeks to:

(a) update a previously published database for polyglutamine (polyQ) proteins by integrating descriptions of human neurodegenerative disease-associated and non-disease-associated human polyQ proteins with complete structural and functional annotations (Chapter 2);

(b) construct a comprehensive knowledge base for kinetochore and its related proteins by providing detailed annotations in terms of protein function, 3D structure, disease-associated mutation, signalling/metabolic pathway and multiple sequence alignment (Chapter 3);

(c) better understand and compare the prediction performance of current computational approaches for protein coiled-coil prediction and coiled-coil oligomeric state identification (Chapter 4);

(d) perform bioinformatics analysis using disorder prediction algorithms to identify mutations predicted to generate regions of microstructures in disordered regions, and accordingly propose experimental candidates for newly proposed mechanism, termed ‘structural capacitance’, which results in *de novo* generation of microstructures in disordered regions upon mutation (Chapter 5).

In summary, this PhD thesis aims to comprehensively interrogate the sequence-structure-function relationships between protein aggregation, folding, function and human disease. To address this, novel bioinformatics approaches and databases are developed and deployed to improve our understanding of these important aspects of proteins and their implications in human diseases.

Chapter 2 PolyQ 2.0: an Updated Database of Human Polyglutamine Proteins

Proteins with expanded polyglutamine (PolyQ) repeats are involved in human neurodegenerative diseases, via a gain-of-function mechanism of neuronal toxicity involving protein conformational changes that result in the formation and deposition of β-sheet-rich aggregates. Aggregation is dependent on the context and the properties of the host protein, such as domain architecture and location of the repeat tract. In order to explore this relationship in greater detail, this chapter describes PolyQ 2.0, an updated database that provides a comprehensive knowledgebase for human polyQ proteins. This database details domain context information, protein structural and functional annotation, single point mutations, predicted disordered regions, protein-protein interaction partners, metabolic/signaling pathways, post-translational modification sites and evolutionary information. Several new database functionalities have also been added, including search with multiple keywords, and new data entry submission. Currently the database contains nine reviewed disease-associated polyQ proteins, 105 reviewed non-disease polyQ proteins and 146 un-reviewed polyQ proteins. It is envisaged that this updated database will be a useful resource for functional and structural investigation of human polyQ proteins.

Database URL: http://lightning.med.monash.edu/polyq2/

2.1 Introduction

The polyglutamine (PolyQ) repeat family of proteins contain a stretch of multiple consecutive glutamines [132]. Expansion of the polyQ tract due to the instability of the cognate CAG codon can lead to a toxic gain-of-function via a conformational change within the protein and the deposition of β-sheet-rich amyloid-like fibrils [133-135]. As such, polyQ repeats are implicated in several neurodegenerative diseases, including Huntington disease and spinocerebellar ataxia [136-142]. While the length of the polyQ repeat is critical to the pathogenesis, the polyQ domain context (i.e. the domains flanking the polyQ tract) is also important [143-146]. Since polyQ repeats are highly aggregation prone [144], it is difficult to experimentally determine the crystal structure of the expanded polyQ repeats [147]. Most studies to date have proposed that polyQ repeats have a beta sheet or intrinsically disordered structure [95]. Recent evidence has further suggested that the misfolding mechanism is context-dependent, and that properties of the host protein, including the domain architecture and location of the repeat tract, can modulate aggregation.

Given the importance of polyQ repeats and their domain context information, we recently performed a bioinformatics investigation of the protein context of polyglutamine repeats [148], and constructed a web-accessible database of all human proteins containing a polyQ repeat greater than seven glutamines in length [120]. Although the PolyQ database provides basic information for each entry, it lacks in both depth and breadth of annotation as well as functionality. Here, we present PolyQ 2.0, a substantially updated knowledgebase for human polyQ proteins. PolyQ 2.0 contains a variety of structural and functional annotations, broad protein information, and domain context of polyQ repeats. In addition, the usability of the web interface has been improved, which now offers database search with multiple keywords as well as user data submission. PolyQ updates

the MySQL relational database that stores entries, and enhances the web interface through the use of modern Javascript tools for visualization and interaction. Apache Tomcat mediates users access to the database through Java Servlets and JavaServer Pages (JSP).

2.2 Update of database entries

Whereas PolyQ contained two types of polyQ proteins, namely disease and non-disease-associated, in PolyQ 2.0 all entries are categorized into three groups according to the annotation of disease involvement and review completeness. Here disease-associated proteins refer to those proteins causing neurodegenerative diseases due to the abnormal expansion of polyQ repeats rather than other proteins with common disease-associated mutations. These groups are: reviewed disease-associated polyQ proteins, reviewed non-disease polyQ proteins and un-reviewed polyQ proteins. We first validated all the data entries in the previous PolyQ database with their UniProt annotation in order to ensure that only high quality data entries are included in PolyQ 2.0. Proteins were included as reviewed entries according to their annotation in the UniProt database. We incorporated polyQ proteins that have not been manually verified from UniProt as un-reviewed polyQ proteins for potential future reference. As a result, we obtained nine reviewed disease-associated polyQ proteins, 105 reviewed non-disease polyQ proteins and 146 un-reviewed polyQ proteins, respectively (Figure 2.1A).

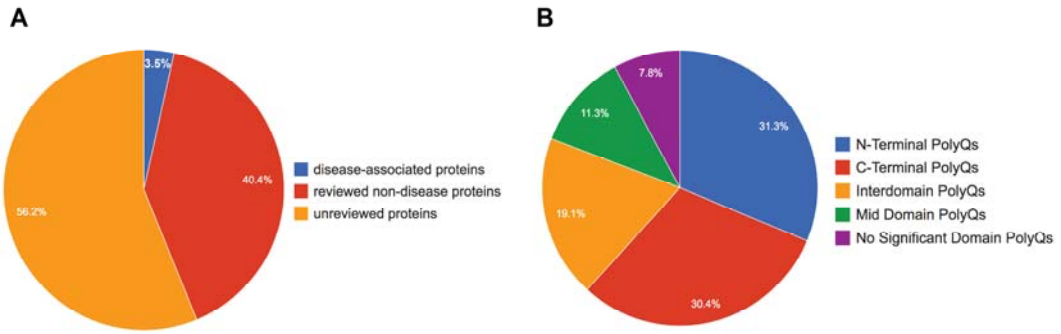


Figure 2.1 Statistics of data entries in PolyQ 2.0. (A) Distribution of disease-associated proteins, reviewed non-disease proteins and un-reviewed proteins; (B) Distribution of the sequence context of different types of polyQ domains for reviewed entries only.

Following the classification system set out previously in PolyQ, we further classified all reviewed 114 sequences into six categories based on the locations and context of polyQ repeats relative to Pfam domains [121]: (1) *N-Terminal PolyQs* – the first polyQ repeat appears before all Pfam domains; (2) *C-Terminal PolyQs* – the last polyQ repeat appears after all the Pfam domains; (3) *Interdomain PolyQs* – the polyQ tracts appear between the first Pfam and last Pfam domain; (4) *Mid Domain PolyQs* –the polyQ repeat appears in the middle of a Pfam domain or overlaps with a Pfam domain; (5) *No Significant Domain PolyQs* – sequences that do not contain any significant Pfam domains; (6) *Unclassified PolyQs* – sequences that do not fit into any of the above categories. The majority of polyQ domains are either *N*- or *C-Terminal PolyQs* while only 7.8% of the reviewed polyQ containing entries do not harbor any significant Pfam domains (Figure 2.1B).

2.3 Update of content and annotation

For PolyQ 2.0, the information content and annotations for entries have been significantly improved and expanded. The updated content includes basic protein information, protein structural information, predicted disordered regions, protein-protein

interaction partners, metabolic/signaling pathways, single point disease- and non-disease associated mutations, and protein post-translational modification sites. In addition, we also performed BLAST search and generated multiple sequence alignments (MSA) in order to provide evolutionary information for each protein entry. A comparison of protein annotations provided in PolyQ and PolyQ 2.0 is shown in Table 2.1.

Table 2.1 The comparison of protein annotation in PolyQ and PolyQ 2.0

Content	**PolyQ**	**PolyQ 2.0**
Protein information	Sequence and unstructured FASTA headers	Structured protein information (function, gene name, protein accession…)
Protein 3D structure	No	Yes
Pfam domain	Yes	Yes
Protein disordered regions	No	Yes
Protein interaction partner	No	Yes
Metabolic/signaling pathway	No	Yes
Single point mutation	No	Yes, incorporating both disease-associated and nonsense mutations
Post-translational modification sites	No	Yes
Multiple sequence alignment	No	Yes

Annotations were extracted and reviewed from a variety of different publicly available resources, including UniProt [149], Protein Data Bank [150], BioGrid [151], KEGG [152], SUPERFAMILY [153] and Pfam [121]. We employed VSL2B [60] to annotate predicted disordered regions. Homologous sequence search was conducted using PSI-BLAST [154] (with an *E*-value of 0.001) against the Swiss-Prot database (http://www.uniprot.org/downloads). Multiple sequence alignments were generated using

Clustal Omega [155]. A summary of the database contents and annotations is shown in Table 2.2.

Table 2.2 Summary of the database contents and annotations of PolyQ 2.0.

Number of protein structures	356
Number of protein interactions	4,081
Number of single point mutations	704
Number of KEGG pathways	41
Number of Pfam domains	498
Number of post-translational modification sites	569

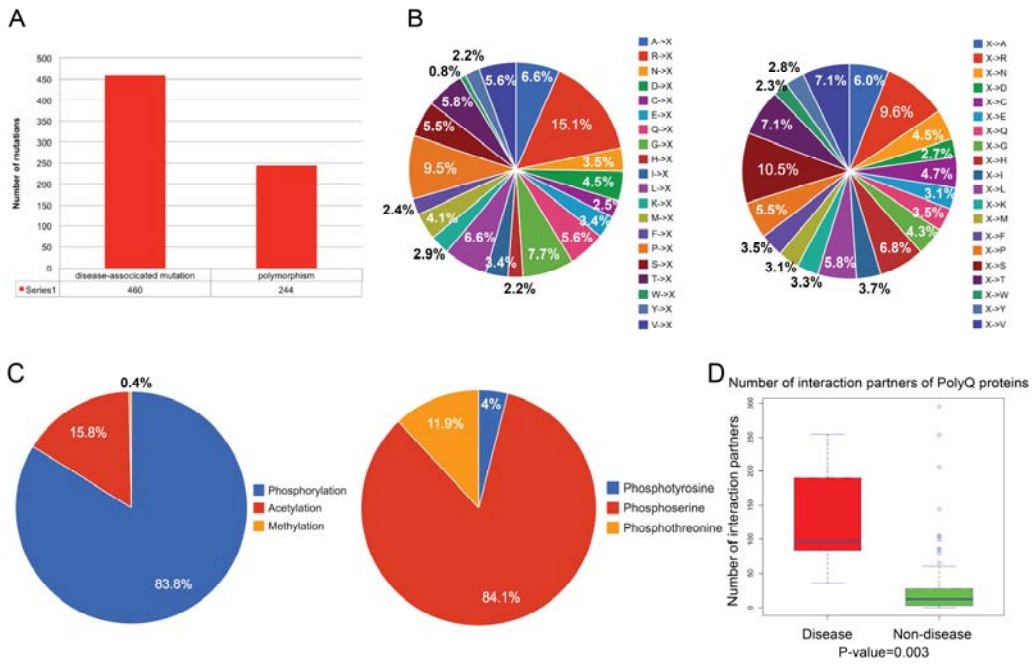


Figure 2.2 Statistical analysis of database content in terms of distributions of disease-associated mutations, post-translational modification site and number of protein-protein interaction partners. (A) Distribution of disease-associated mutation and polymorphism; (B) Distribution of the number of mutations with respect to two mutation patterns (where *X* means any amino acid); (C) Distribution of types of protein post-translational modification with detailed distribution of sub-types of phosphorylation; (D) Number of protein-protein interaction partners of reviewed polyQ disease-associated proteins and non-disease proteins.

We analyzed the database content in terms of distribution of disease-associated mutations, post-translational modification sites and number of protein-protein interaction partners. From a total of 704 single point mutations within the 260 data entries, 460 (65.3%) mutations are disease-associated, while 244 (34.7%) mutations are polymorphisms (Figure 2.2A). By analyzing the distribution of different types of mutations associated with polyQ proteins, we found that arginine is the most frequently mutated amino acid (approximately 15% of the mutated residues; Figure 2.2B). Phosphorylation is the most frequently observed post-translational modification (Figure 2.2C). Disease-associated polyQ proteins have significantly more protein interaction partners than non-disease polyQ proteins (p-value = 0.003; Figure 2.2D).

2.4 Database functionality and web interface improvements

PolyQ 2.0 features several important improvements of the user interface as well as new functionality, including database search with multiple types of keywords and new entry submission. A comparison of database functionality between PolyQ and PolyQ 2.0 is listed in Table 2.3.

Table 2.3 Database functionality comparison between PolyQ and PolyQ 2.0

Functionality		**PolyQ**	**PolyQ 2.0**
Database search	Database ID/UniProt ID	No	Yes
	Protein name	Yes	Yes
	Pfam domain	Yes	Yes
	Disease	No	Yes
	PTM	No	Yes
	PTM kinase	No	Yes
	Interaction partner	No	Yes
User submission		No	Yes

The search functionality in PolyQ 2.0 has been considerably improved, with search options available using multiple keywords, in addition to the options of protein name and Pfam domain offered by the previous version. The database can be searched by PolyQ/UniProt ID, protein name, Pfam domain, disease, type of protein post-translational modification sites/kinase and protein-protein interaction partner name. The PolyQ ID is composed of "PD" followed by five digits. As there are in total 260 entries in PolyQ 2.0, the PolyQ ID ranges from "PD00001" to "PD00260". An example of the result of database search with UniProt ID=P54252 (Ataxin-3) is shown in Figure 2.3, comprising nine main sections related to different annotations.

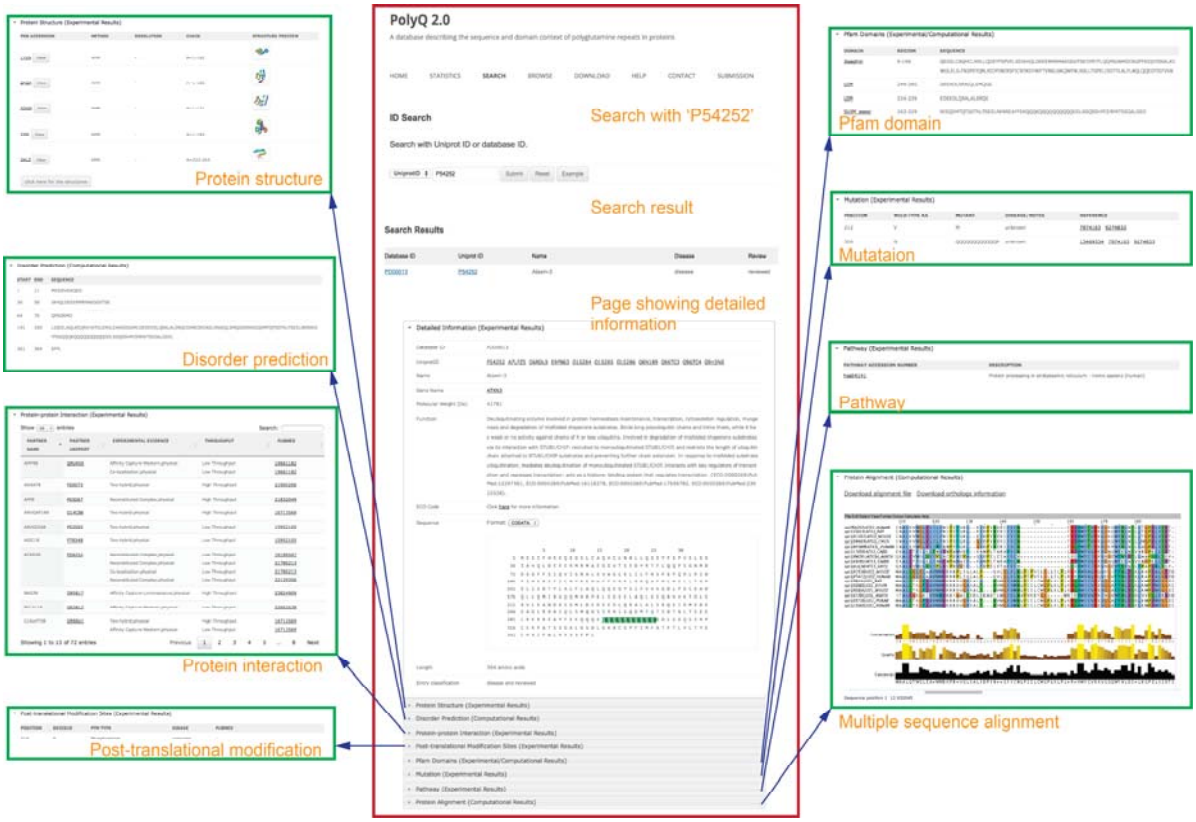


Figure 2.3 Typical search results in PolyQ 2.0 using the UniProt ID P54252 as an example. The results are summarized and displayed in nine main sections, including protein information, protein structure, metabolic/signalling pathway, protein interaction, post-translational modification site, Pfam domain, disorder region prediction, protein mutation and multiple sequence alignment.

Several plug-ins were employed to enhance visualization of database entries. In the protein basic information section, we embedded a protein feature view plug-in in order to show protein functional sites/domains and basic structural information (Figure 2.4A). PV (http://biasmv.github.io/pv/) and pViz [156] were also used to allow detailed examination of protein structures (Figure 2.4BC). Multiple sequence alignment is displayed using JalView [157] to visualize sequence conservation (Figure 2.4D).

Browsing of data entries has also been improved. The entries can now be categorized in terms of disease involvement and completeness of review and annotation. In addition, detailed context annotations, which show the distribution of polyQ domain, protein superfamily domain and protein post-translational modification sites are available. A webpage showing database statistics is available, giving users a one-page snapshot of database contents as well as convenient navigation around the database. Detailed user help and instructions are also provided. Finally, we have built a data submission page, enabling users to deposit data in the database, which are made publically available after checking, curation and approval by the site administrator.

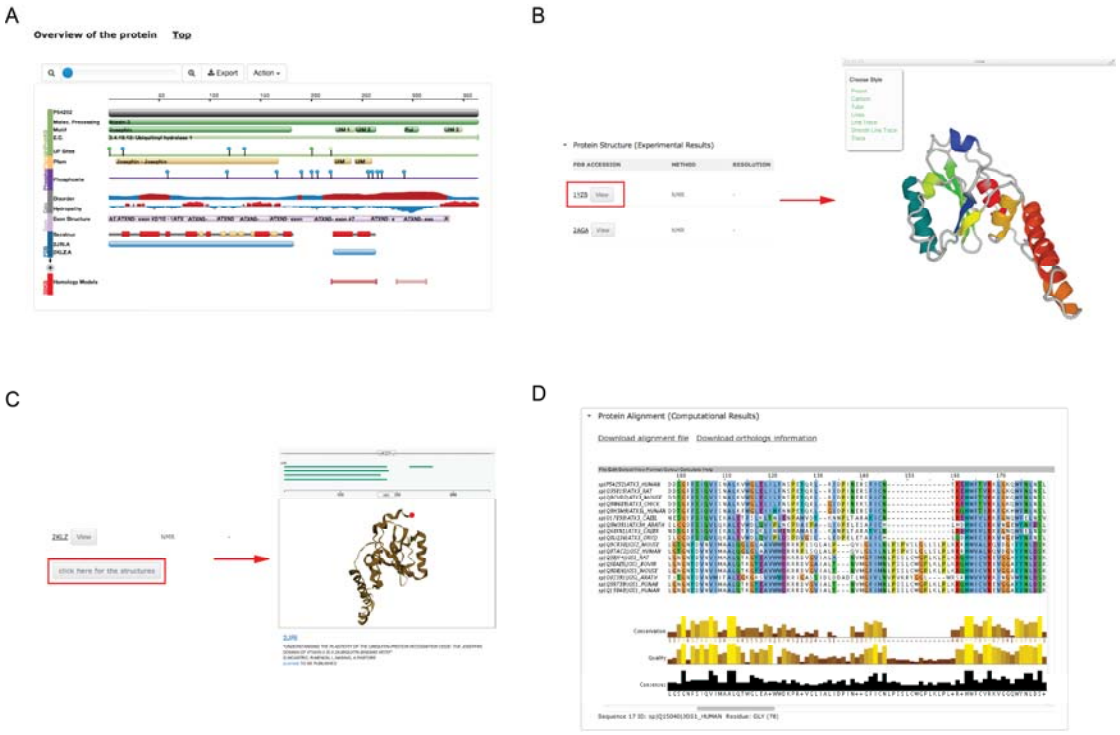


Figure 2.4 Plug-ins in PolyQ 2.0 to enhance database visualization. (A) Protein feature plug-in; (B) PV showing protein structure; (C) pViz for visualizing multiple structures; (D) Jalview displaying MSAs.

2.5 Conclusions

Based on our previous PolyQ database for human polyQ proteins, in the present study we have developed an updated database, PolyQ 2.0, to provide comprehensive protein functional, structural and evolutional annotations together with domain context information for human polyQ proteins. Integrating publicly available annotations and computational resources, PolyQ 2.0 offers a variety of annotations in terms of protein basic information, protein structure, predicted intrinsically disordered domain, protein-protein interaction, protein functional site/domain, single point mutation, metabolic/signaling pathway and multiple sequence alignment. We anticipate that this

updated knowledgebase will benefit functional and structural studies of human polyQ proteins and their role in neurodegenerative diseases.

Chapter 3 KinetochoreDB: a Comprehensive Online Resource for the Kinetochore and Its Related Proteins

In this chapter, a novel database, namely KinetochoreDB, for the kinetochore and its related proteins has been constructed to integrate sequence features, structural and functional annotations, and disease associations of kinetochore and its related proteins. It provides comprehensive annotations on 1,554 related protein entries in terms of their amino acid sequence, protein 3D structure, predicted intrinsically disordered region, protein-protein interaction, post-translational modification site, functional domain and key metabolic pathways, integrating several public databases, computational annotations and experimental results. KinetochoreDB provides interactive and customizable search and data display functions that allow users to interrogate the database in an efficient and user-friendly manner. It uses PSI-BLAST searches to retrieve the orthologs of all entries and generate multiple sequence alignments that contain important evolutionary information. This knowledge base also provides annotations of single point mutations for entries with respect to their pathogenicity, which may be useful for generation of new hypotheses on their functions and follow-up studies of human diseases.

Database URL: http://lightning.med.monash.edu/kinetochoreDB2/

3.1 Introduction

During cell mitosis and meiosis, the kinetochore plays a critical role of locating the attachments on chromosomes and pulling sister chromatids apart. It is assembled on centromeric chromatin through complex pathways and functions during the cell cycle [158-163]. During the last few decades, numerous studies of the kinetochore and its related proteins have characterized its function, architecture and the repertoire of its related proteins using biochemistry, structural biology and cell biology techniques [161, 164-168]. Both the stability of the kinetochore–microtubule interface and mutations occurring in the kinetochore and its related proteins, are associated with a number of human diseases [169-172]. Dynamics studies of the kinetochore have also shown that deregulation of the kinetochore-microtubule dynamics frequently results in chromosome instability, leading to the development of cancer [167, 168, 173]. Other experimental studies reveal that mutations of the kinetochore and its related proteins are closely linked to human diseases. For example, the adenomatous polyposis coli protein, found in both centrosome and kinetochore, contains approximately 30 disease mutations that cause Familial Adenomatous Polyposis (FAP) [171, 172] and Medulloblastoma (MDB) [169].

Despite its biological significance and our increasing awareness of its potential roles in human diseases, there is currently a paucity of publically available databases or resources that focus on comprehensive functional annotations of the kinetochore and its related proteins. MiCroKiTS [131], for example, is an integrated online data resource for kinetochore, midbody, telomere, centrosome and spindle proteins. However, important annotations on entries in MiCroKiTS are not available in terms of protein 3D structure, protein interaction partners, metabolic/signaling pathways etc., all of which are crucial aspects for follow-up studies of their functions (Table 3.1).

Table 3.1 Comparison between KinetochoreDB and MiCroKiTS.

Annotation/Function	KinetochoreDB	MiCroKiTS
Target	Kinetochore and its related proteins	Kinetochore, centrosome, midbody, telomere and spindle proteins
Protein 3D structure	Yes, detailed structural information available; customizable display	No
Protein intrinsic disorder	Yes, predicted by VSL2B	No
Protein interaction partner	Yes, detailed information available	No
Metabolic pathway	Yes	No
Disease-associated mutations	Yes, incorporating both OMIM disease-associated and nonsense mutations	No
Evolutionary conservation	Yes, multiple sequence alignments curated and displayed using Jalview	No
User enquiry and submission	Yes	No

In an effort to address this gap, we created KinetochoreDB, which integrates several public databases, computational annotations and experimental results for currently 1,554 related entries. KinetochoreDB is featured by the following aspects:

(1) It provides annotations of protein 3D structure when structural information is available. For protein entries with available structural information, the PDB ID and their related information are provided. In addition predicted intrinsic disorder information is provided. This is particularly important for obtaining structural insights into those entries in KinetochoreDB whose 3D structures have not been solved.

(2) It provides comprehensive annotations of single point mutations and possible pathogenic effects. These mutations are classified as pathogenic and nonsense mutations in KinetochoreDB. For disease-associated pathogenic mutations,

KinetochoreDB provides a link to the OMIM (Online Mendelian Inherited Mutations in Man) database (http://www.omim.org/) (22). Moreover, it allows users to search the entire database with the disease name of interest, provides user-friendly options to browse the related kinetochore proteins that harbor such disease-associated mutations.

(3) It provides metabolic pathway information for each entry by cross-referencing the KEGG database, which is important for understanding the functions of kinetochore proteins from a metabolic/signaling network perspective. In particular, the pathway information and the link to KEGG will be provided if an entry has available pathway information in KEGG. It is worth noting that such important information is not available in MiCroKiTS.

(4) It provides multiple sequence alignments (MSAs) for all included entries, allowing users to easily identify evolutionarily conserved regions within the family of the kinetochore protein. In addition the visualization of MSAs implemented by Jalview is user-friendly and customizable.

(5) It provides convenient user enquiry and new entry submission options by allowing users to automatically upload their newly discovered sequences into the online database.

3.2 Database construction and features

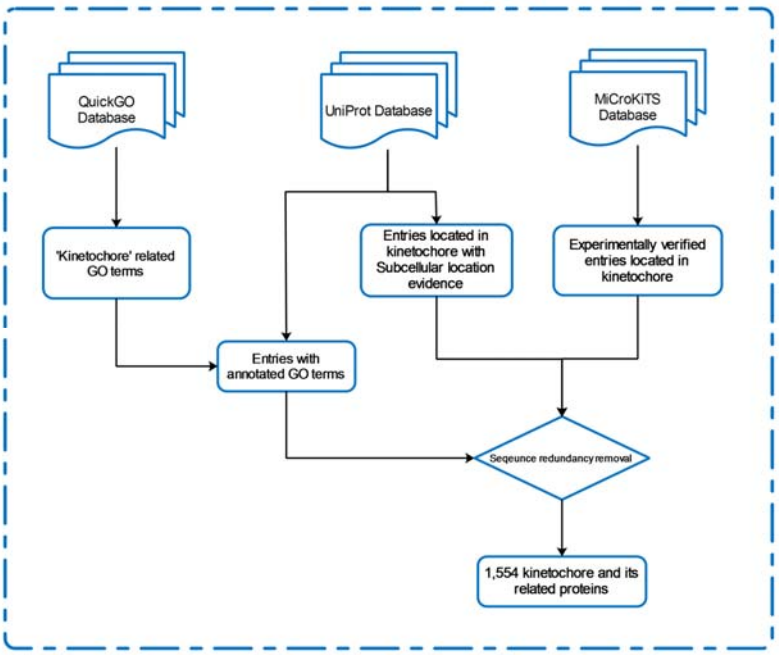


Figure 3.1 The schema of database construction and data collection processes

We define 'kinetochore and its related proteins' in terms of protein subcellular location and Gene Ontology terms. The entries of KinetochoreDB originate from three resources including QuickGo database [174], UniProt database [175] and MiCroKiTS. From MiCroKiTS, we obtained data entries that have been experimentally verified to be located in kinetochore. By searching GO terms from QuickGo database with keyword 'kinetochore', we obtained 64 GO terms related to kinetochore. For each GO term, we searched and filtered the reviewed entries from the UniProt database to ensure that all the downloaded entries contain the GO annotation. This process resulted in 53 GO terms remaining including 25 cellular component terms, 2 molecular function terms and 26 biological process terms (Table S3.1). In addition, we queried 'subcellular localization' with the keyword 'kinetochore' from UniProt and downloaded entries with published experimental evidence from search results. After the removal of redundant entries, the resulting dataset contains 1,554 carefully reviewed entries in total. The detailed processes

of database construction and data collection are illustrated in Figure 3.1 and the statistical summary can be found in Figure 3.2 and Table 3.2.

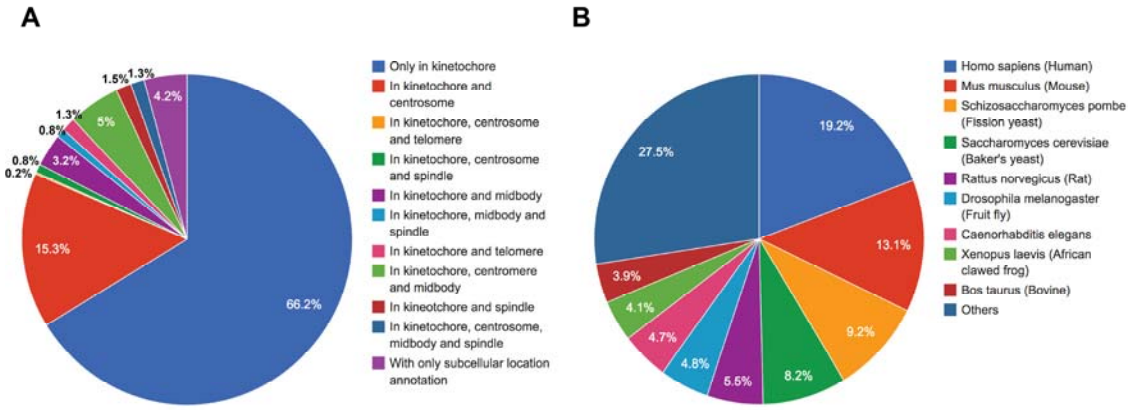


Figure 3.2. **Statistics summary of location and species of KinetochoreDB entries. (A) Distribution of protein locations according to MiCroKiTS and protein subcellular location annotation from UniProt database. (B) Distribution of species of all KinetochoreDB entries.**

Table 3.2 Statistical summary of the information contained in KinetochoreDB.

Number of entries	1,554
Number of protein structures	1,163
Number of protein interactions	47,675
Number of mutations	2,429
Number of KEGG pathways	430
Number of Pfam domains	4,000
Number of post-translational modification sites	2,165

For each entry, KinetochoreDB integrates several public resources including the UniProt database, RCSB Protein Data Bank (PDB) [59], OMIM [176], BioGRID 3.2 [177], Pfam database [121] and KEGG PATHWAY [152] to provide a comprehensive description in terms of basic protein information, protein structure, function, mutation and conservation. An important feature of KinetochoreDB is the provision of 3D structure. To achieve this, we manually searched all the entries against the PDB database using their corresponding UniProt identifiers and protein names. For protein complex structures, we identified the PDB chain for each entry and annotated the entry with that chain. In

addition, we generated multiple sequence alignments (MSAs) using all the homologous sequences for each protein entry. The homologous sequences were retrieved by PSI-BLAST [154] searches against the Swiss-Prot dataset from UniProt. The alignment was generated using Clustal Omega [155]. We also predicted the natively disordered regions for all protein entries using one of the most widely used disorder predictors, VSL2B [60], which predicted a residue to be in disordered region if its prediction score was greater than 0.5.

We used Jmol (http://jmol.sourceforge.net/) and pViz [156] for visualization of protein structures, and Jalview [157] for customizable editing and display of MSAs for each protein entry. The information stored within KinetochoreDB resides in a MySQL relational database. The highly interactive web front-end to the data was constructed using the Javascript framework, JQuery. Apache Tomcat handles serving of data to users on the web, utilizing a set of Java Servlets and JavaServer Pages (JSP) for searching and viewing of data.

3.3 Database utility

The 'Search' page (http://lightning.med.monash.edu/kinetochoreDB2/Search.jsp) (Figure 3.3) allows users to search the database in different ways. These search options can be generally classified into two groups: search with ID or search with keywords. Examples are provided below to help users to understand how to perform the search. When searching the database with IDs, we provide two different kinds of IDs to facilitate the search: UniProt ID and KinetochoreDB ID. The latter is composed of 'KD' and 5 digits, e.g. KD00095. As there are a total of 1,554 entries in the database, the database ID ranges from KD00001 to KD01554. In addition we offer alternative search options with keywords. These include protein name, kinase name, post-translational modification type, name of protein interaction partner and name of diseases caused by single point mutations.

After selecting ‘Submit’ button, the corresponding searching results will be shown on the webpage. For each entry, there are generally nine sections of structural and functional aspects, including general information, protein structure, disordered regions prediction, interaction partner, post-translational modification site, Pfam domain, protein mutation, metabolic/signaling pathway and protein alignment with homologs. To illustrate the annotations for each entry in KinetochoreDB, we use ‘UniProt ID = O14965’ (KinetochoreDB ID = ‘KD01531’) as an example query. The resulting page and nine sections are shown in Figure 3.4.

For protein overview, we used pViz [156] to facilitate the general description of protein entries including functional sites and domains (Figure S3.1A), allowing a more detailed inspection of its domains. For protein structure overview, we provide two different ways to review the 3D structures. A single structure for the current protein entry can be examined by clicking the ‘View’ button to launch Jmol, a Java applet for displaying protein 3D structures (Figure S3.1C). Multiple structures can also be viewed together as an ensemble using pViz (Figure S3.1B).

For protein-protein interaction, we highlighted the interaction partner if this protein is also an entry of KineotchoreDB. We noticed that due to our search strategy (see section 2.1 for details), some proteins in MSAs are not included in current KinetochoreDB. To facilitate the comparison between entries in KinetochoreDB and their homologs, we also archived these homologs by extracting protein UniProt ID. These files are available in the ‘Protein alignment’ section of the webpages for protein detailed information.

KinetochoreDB will be updated on a regular basis with newly available entries from various databases, to allow an up-to-date archive of recent results of the kinetochore and its related proteins. To allow an up-to-date archive of recent results of the kinetochore

and its related proteins, we allow users to submit new sequences and their structural and functional annotations to KinetochoreDB (Figure S3.2). After careful review and verification, new data will be included in KinetochoreDB and made publically available.

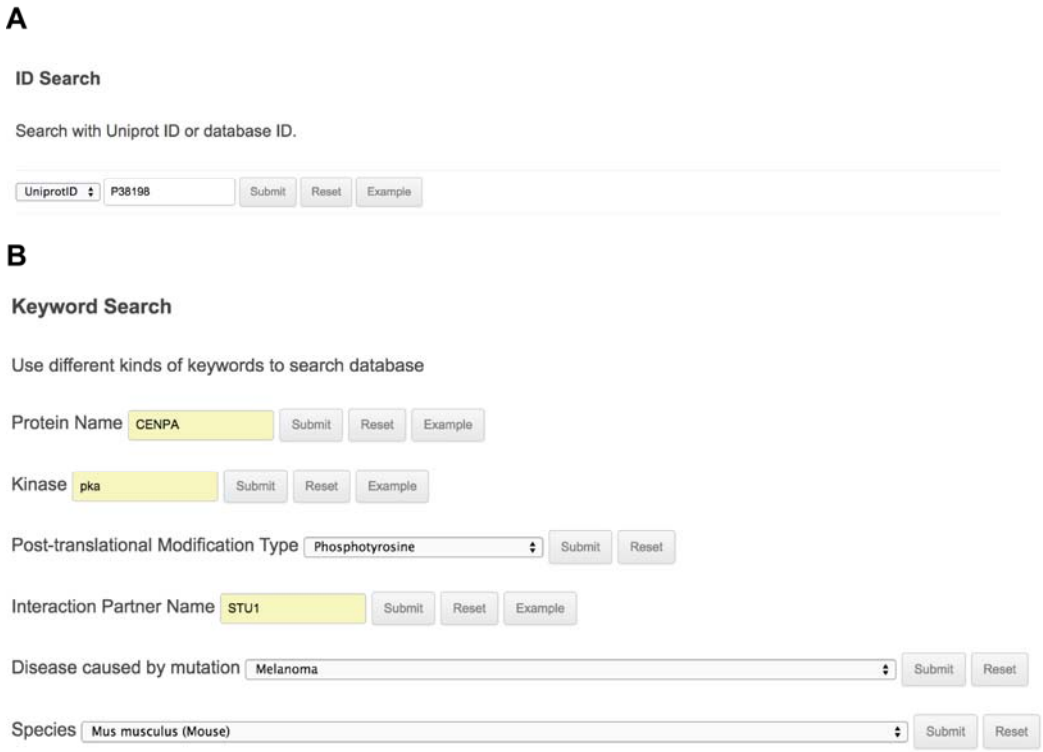


Figure 3.3 The search options provided by KineotchoreDB. (A) Protein ID search with UniProt ID or KineotchoreDB ID. (B) Keyword search with protein name, kinase, post-translational modification type, interaction partner name, disease or species.

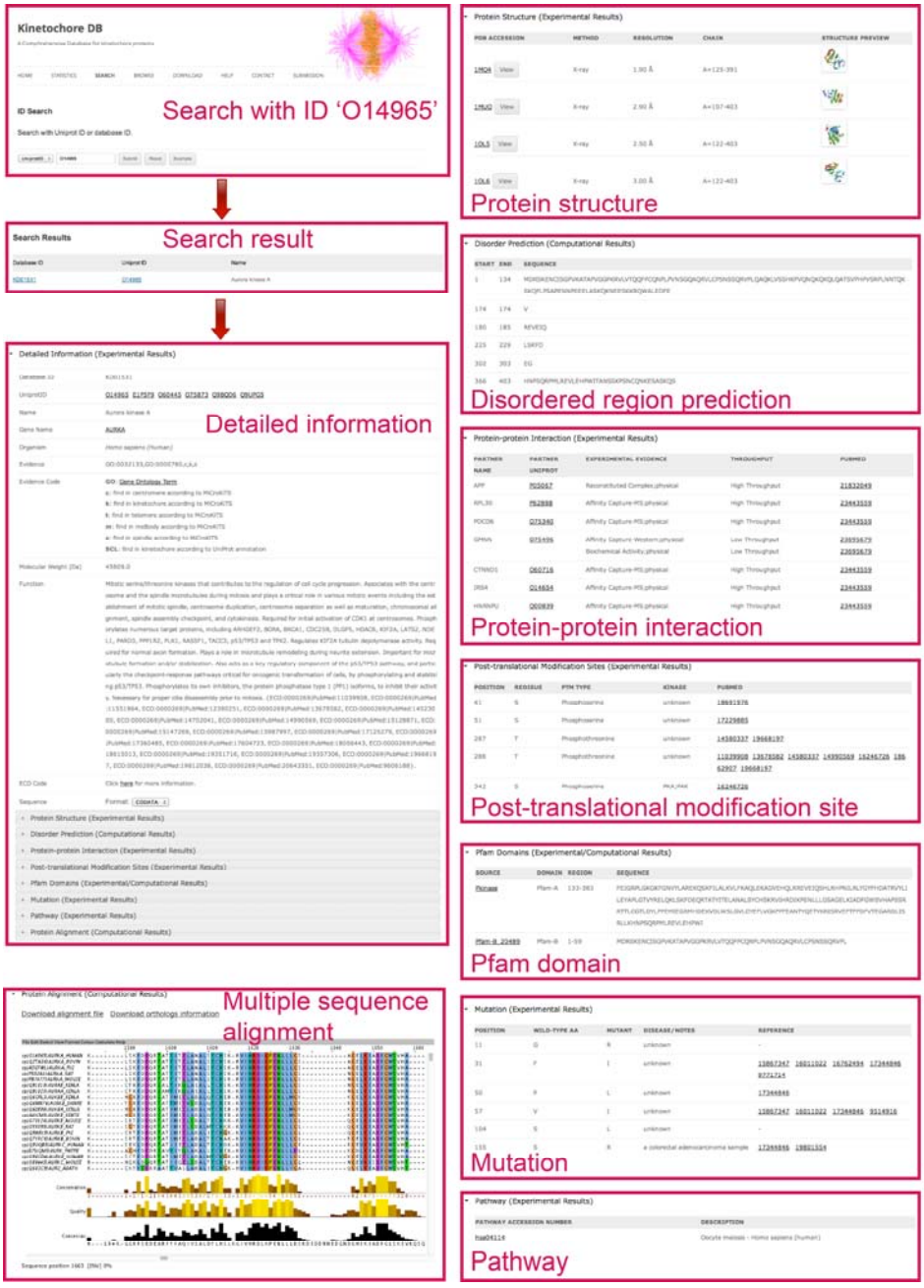


Figure 3.4 Typical search results in KinetochoreDB using the UniProtID O14965 as an example. The results are summarized and displayed in nine sections including protein information, protein structure, metabolic/signaling pathway, protein interaction, post-translational modification site, Pfam domain, disorder region prediction, protein mutation and multiple sequence alignment.

3.4 Discussion

Many of the entries in our KinetochoreDB harbour mutations. We hence provide a preliminary statistical analysis. There are 2,429 mutations occurring in the 1,554 entries in KinetochoreDB. Among these, 1,146 (47.2%) mutations cause diseases while 1,283 (52.8%) mutations are nonsense mutations (Figure 3.5A). We analyzed the distributions of different types of mutations in Figure 3.5B, where *X* means any type of amino acid. It can be noticed that there is no significant difference between the two types of mutation patterns (Figure 3.5B). Post-tranlational modifications (PTM), on the other hand, attached new chemical groups and small molecules, thereby extending the chemical repertoire of amino acids. With the availability of PTM annotations from our KinetochoreDB, we further analysed the distribution of different types of PTM for all the entries in KinetochoreDB. We noticed that kinetochore and its related proteins possess many PTM sites, among which, the top three types are phosphorylation, acetylation and methylation, as shown in Figure 3.5C. We also list different sub-types of acetylation and phosphorylation in Figure 3.5C, respectively.

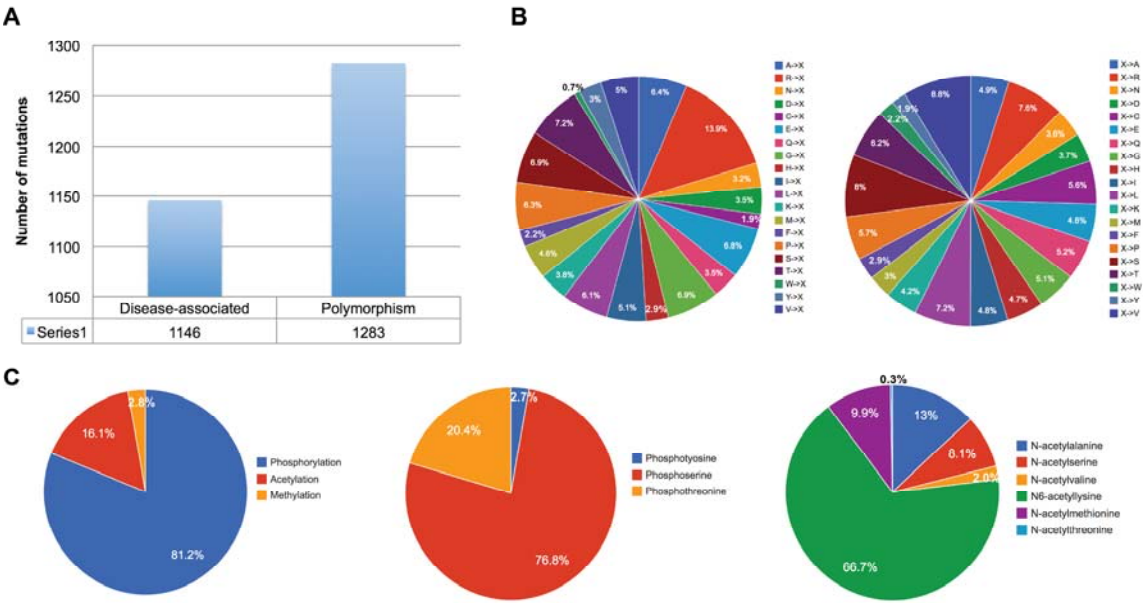


Figure 3.5 Statistical analysis regarding single point mutation and protein post-translational modification. (A) Distribution of disease-associated mutation and polymorphism. (B) Distribution of the number of mutations with respect to two mutation patterns. (C) Distribution of types of protein post-translational modification with detailed distribution of sub-types of phosphorylation and acetylation.

In addition, with the comprehensive dataset from KinetochoreDB, we conducted a statistical analysis regarding the number of proteins invovled in different GO terms including cellular component, molecular function and biological process, respectively. As shown in Figure 3.6, we ranked all the GO terms according to the number of proteins with annotation of current GO term and selected 10 top GO terms, which are listed in Figure 6. In total, 414, 285 and 72 proteins contain the annotation of condensed chromosome kinetochore (GO:0000777), microtubule motor activity (GO:0003777) and protein localization to kinetochore (GO:0034501).

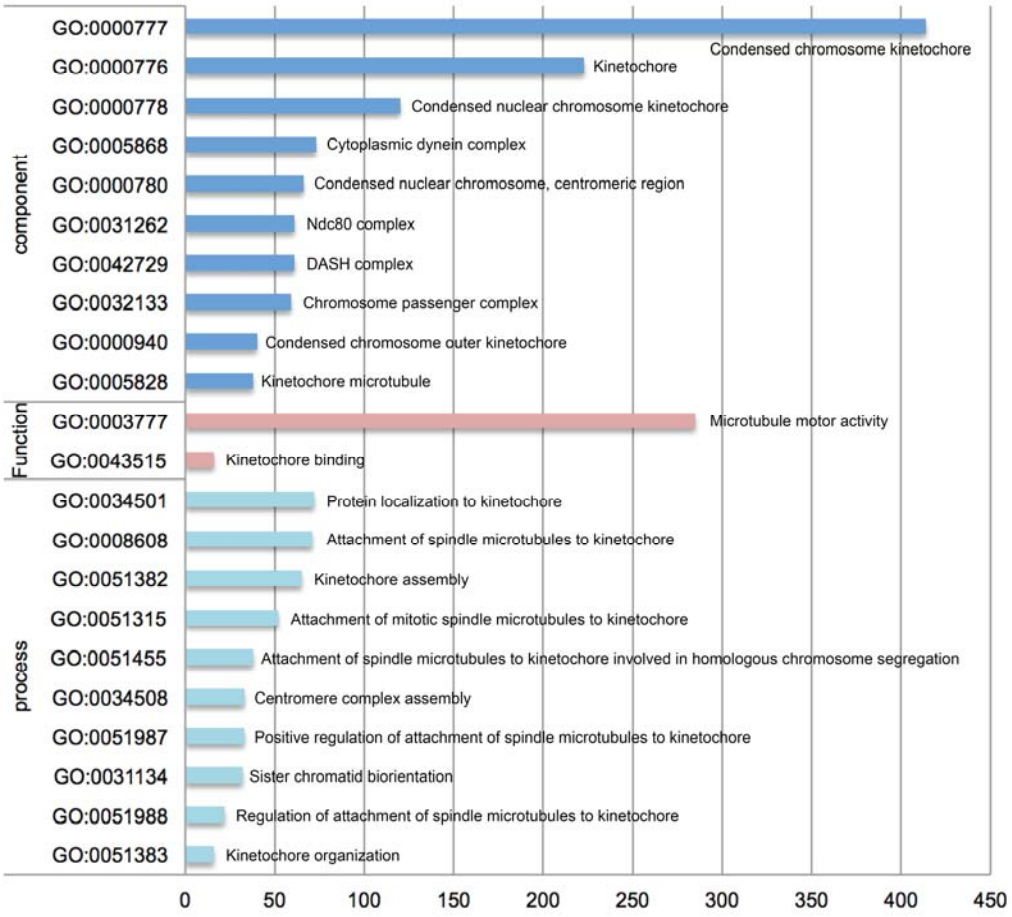


Figure 3.6 Statistical analysis regarding the number of entries from KinetochoreDB involved in different GO terms.

The kinetochore and its related proteins play extremely important roles during cell division and mitosis. In the past few decades, research on this topic has attracted a great deal of interests not only because they are important in cell cycle, mitosis and meiosis, but also because they are closely associated with human diseases upon mutation. In this context, databases such as KinetochoreDB that provide comprehensive annotations on the repertoire of kinetochore-related proteins will greatly facilitate in-depth functional investigation of these proteins and their pathological relationships with human diseases. Through effective data integration from multiple public resources, KinetochoreDB has collected large amounts of information for relevant protein entries with respect to their

amino acid sequence, protein 3D structure, biological function and evolutionary conservation. By providing comprehensive functional annotations of all available kinetochore-related proteins, we believe that this online resource will be used as a powerful tool to bridge the functional characterization and disease-associated mutation studies of this important class of proteins.

Chapter 4 Critical Evaluation of in silico Methods for Prediction of Coiled-coil Domains in Proteins

One of the two protein structural features focused in this thesis is the protein coiled-coil domain (CCD). Coiled-coils refer to a bundle of helices coiled together like strands of a rope. It has been estimated that nearly 3% of protein-encoding regions of genes harbour coiled-coil domains. Experimental studies have confirmed that CCDs play a fundamental role in subcellular infrastructure and controlling trafficking of eukaryotic cells. Given the importance of coiled-coils, multiple bioinformatics tools have been developed to facilitate the systematic and high-throughput prediction of CCDs in proteins. This chapter has reviewed and compared twelve sequence-based bioinformatics approaches and tools for coiled-coil prediction. These approaches can be categorised into two classes: coiled-coil detection and coiled-coil oligomeric state prediction. These methods in terms of their input/output, algorithm, prediction performance, validation methods and software utility have been reviewed and compared. All the independent testing datasets are available at http://lightning.med.monash.edu/coiledcoil/. In addition, this chapter describes a case study of nine human polyglutamine (PolyQ) disease-related proteins and predicted CCDs and oligomeric states using various predictors. Prediction results for CCDs were highly variable among different predictors. Only two peptides from two proteins were confirmed to be CCDs by majority voting. Both domains were predicted to form dimeric coiled-coils using oligomeric state prediction. It is anticipated that this comprehensive analysis will be an insightful resource for structural biologists with limited prior experience in bioinformatics tools, and for bioinformaticians who are interested in designing novel approaches for coiled-coil and its oligomeric state prediction.

4.1 Introduction

First described in 1953 by Pauling and Crick [178], the proliferation of studies of coiled-coil domains (CCDs) in proteins has driven continued computational prediction in the past few decades. CCDs can be summarized as at least two or more helices that wrap around each other, which can be defined as a repeat X_n of residues, where X can be denoted as (*a-b-c-d-e-f-g*) and n can be described as the number of helices. It is estimated that nearly 10% of eukaryotic proteins harbour CCDs [72, 179]. CCDs exhibit a preference for hydrophobic residues at positions *a* and *d*, charged residues at positions *e* and *g*, and hydrophilic residues at positions *b*, *c* and *f* [70, 180], which serve to stabilize helix oligomerization according to the "Peptide Velcro" (PV) hypothesis [73]. This repeating X_n motif enables the prediction of CCDs and their oligomeric states based on protein sequences.

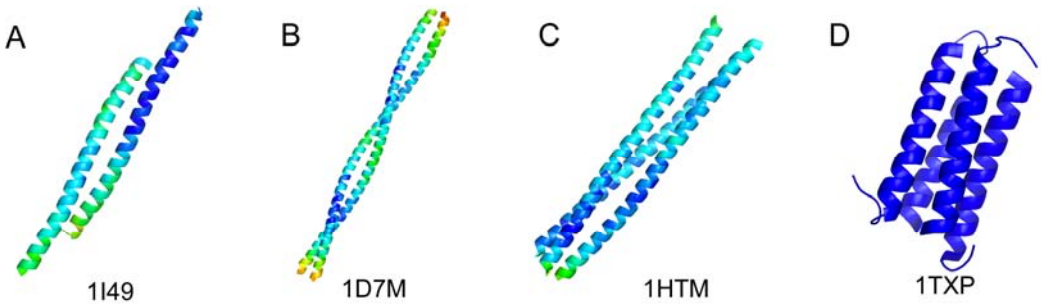


Figure 4.1 Examples of coiled-coil oligomeric states. (A) antiparallel dimer (PDB Accession: 1I49 [181] – crystal structure of arfaptin), (B) parallel dimer (PDB Accession: 1D7M [182] – coiled coil dimerization domain from cortexillin I), (C) trimer (PDB Accession: 1HTM [183] – structure of influenza haemagglutinin at the PH of membrane fusion) and (D) tetramer (PDB Accession: 1TXP [184] – heterogeneous nuclear ribonucleoprotein C oligomerization domain tetramer).

Experimental studies have confirmed that CCDs play a fundamental role in subcellular infrastructure and controlling trafficking of eukaryotic cells [74, 75]. The relatively high stability of CCDs has led to their promising use as delivery systems for a range of molecules. For example cartilage oligomeric matrix protein (COMP) [10, 185] and Right-handed (RH) protein [186] from *Staphylothermus marinus* have been used as drug delivery systems in anticancer therapies [179, 187, 188]. The 5 α-helix CCDs in COMP are capable of binding and carrying some important signaling molecules, including vitamins A and D_3. Other successful applications of CCDs, peptides and motifs employed in drug delivery systems have also been reported [189-193].

Sequence and structural analysis of CCDs have enabled the development of computational approaches for the prediction of CCDs from sequence alone [70, 73, 180, 194]. For example Vincent *et al.* performed coiled-coil prediction for proteins from tenascins and thrombospondins families, analysed the motif conservation of different coiled-coil oligomeric states and revealed that sequence conservation allows trimer and pentamers of CCDs to be distinguished, providing useful insights for future coiled-coil prediction [194]. However, the rapid growth in prediction approaches since the last comprehensive comparison which was reported almost a decade ago [195] creates an urgent need to critically assess and compare the now-large and diverse prediction methods. In this article, therefore, we present a comprehensive review of 12 sequence-based methods for coiled-coil prediction, offering insights into the nature of different predictors and facilitating potential improvement of coiled-coil domain prediction. All predictors are critically reviewed in terms of input, model construction and outcome (i.e., prediction performance) [196, 197]. To evaluate the performance of coiled-coil predictors, independent tests were conducted with new test datasets (http://lightning.med.monash.edu/coiledcoil/) carefully collected and curated from

different resources. Finally, as CCDs have been extensively found in disease-associated human polyglutamine (PolyQ) proteins [198], we applied various predictors to a dataset of nine human proteins containing PolyQ repeats and discussed our findings.

4.2 Materials and methods

4.2.1 Predictors evaluated in this study

Table 4.1 summarises the details of the tools of coiled-coil and its oligomeric state prediction that are evaluated in this article. These are COILS [90], PCOILS [199], Paircoil2 [91], SOSUIcoil [200], MARCOIL [94], CCHMM_PROF [93], SpiriCoil [95], SCORER 2.0 [98], LOGICOIL [97], PrOCoil [100], RFCoil [99] and Multicoil2 [71].

Table 4.1 A comprehensive list of coiled-coil and oligomeric state predictors reviewed in this study.

Task	Tool Publication Date	Input format	Model highlight [a]	Evaluation		Service			
				Strategy	Output format	Web service[b]	Availability	Speed[d]	Reliability[e]
Coiled-coil region prediction	COILS [90] 1997	Raw sequence or SwissProt IDs	Pairwise residue probabilities	Algorithm tested with sequences of known globular proteins, randomly generated sequences and all the sequences in GenBank	Residue score and probability to located in coiled-coil domain	Yes	Yes (with third-party implementations)	Fast	Consistent
	PCOILS [199] 2005	Raw sequence / FASTA sequence	Pairwise profile comparison using protein evolution profile	Case study	Residue score and probability to located in coiled-coil domain	Yes	Yes	Moderate	Results vary depending on BLAST database
	Paircoil2 [91] 2006	FASTA sequence	Pairwise residue probabilities	Leave-family-out cross-validation	Residue score and probability to located in coiled-coil domain	Yes	Yes	Fast	Unknown[f]
	MARCOIL [94] 2002	FASTA Sequence	HMM based on MTIDK and other matrices	150-fold cross-validation	Residue score and probability to located in coiled-coil domain	Yes	Yes	Fast	Consistent
	CCHMM_PROF [93] 2009	Raw sequence/FASTA sequence	HMM based on multiple sequence alignment	Overall accuracy, Segment overlap and case study	Overall probability of containing CCDs and Binary decision to (not) be in coiled-coil domain	Yes	Yes	Moderate	Results vary depending on the BLAST database
	SpiriCoil[c] [95] 2010	FASTA sequence	Structurally informed homology-based multiple HMMs	Independent test	Binary decision to (not) be in coiled-coil domain	Yes	No	Fast	-
	SOSUIcoil [200] 2008	One-letter symbol or multiple FASTA sequences	Canonical discriminant analysis	Independent test and case study	-	No	No	-	-

Coiled-coil oligomeric state prediction	SCORER 2.0 [98] 2011	Raw sequence and/or heptad register	Log-likelihood ratio with new defined score function	Independent test	Predicted scorer to be parallel dimeric and trimeric coiled-coil	Yes	Yes	Fast	Consistent
	LOGICOIL [97] 2013	Raw sequence and/or heptad register	Bayesian variable selection and multinomial probit regression	10-fold cross-validation and leave-one-out cross-validation	Predicted score to be parallel dimer, antiparallel dimer, trimer and tetramer	Yes	Yes	Fast	Consistent
	PrOCoil [100] 2011	Raw sequence and/or heptad register	SVM and coiled-coil Kernel	10-fold cross-validation, nested cross-validation and case study	Predicted scorer to be parallel dimeric and trimeric coiled-coil	Yes	Yes	Fast	Consistent
	RFCoil [99] 2014	Raw sequence and heptad register	Random forest with effective amino acid indices	10-fold cross-validation and independent tests	Predicted probability to be parallel dimeric and trimeric coiled-coil	Yes	Yes	Fast	Consistent
Coiled-coil region and oligomeric state prediction	Multicoil2 [71] 2011	FASTA sequence	Pairwise residue correlation and HMM	Leave-family-out cross-validation	Residue probability to be located in non-coiled-coil, dimer or trimer	Yes	Yes	Fast	Consistent

Note. [a] HMM – Hidden Markov Model; SVM – Support Vector Machines.

[b] The URLs of predictors listed are: COILS - http://embnet.vital-it.ch/software/COILS_form.html; PCOILS - http://toolkit.tuebingen.mpg.de/pcoils; PairCoil2 - http://groups.csail.mit.edu/cb/paircoil2/paircoil2.html; MARCOIL - http://toolkit.tuebingen.mpg.de/marcoil; SOSUIcoil - http://harrier.nagahama-i-bio.ac.jp/sosui/coil/submit.html (not available); CCHMM_PROF - http://gpcr.biocomp.unibo.it/cgi/predictors/cchmmprof/pred_cchmmprof.cgi; SpiriCoil – http://supfam.cs.bris.ac.uk/SUPERFAMILY/spiricoil/; SCORER 2.0 - http://coiledcoils.chm.bris.ac.uk/Scorer/; LOGICOIL - http://coiledcoils.chm.bris.ac.uk/LOGICOIL/; PrOCoil - http://www.bioinf.jku.at/software/procoil/; RFCoil - http://protein.cau.edu.cn/RFCoil/index.php?page=introduction; Multicoil2 - http://groups.csail.mit.edu/cb/multicoil2/cgi-bin/multicoil2.cgi.

[c] In [95], SpiriCoil was also applied for oligomeric state prediction. The prediction performance was comparable with that of MULTICOIL, which is the previous version of Multicoil2.

[d] Speed refers to the response time after submitting the sequence to the web server.

[e] Reliability refers to whether the outputs of the predictor's web server and its local executable are consistent.

[f] Paircoil2 is not runnable on our local machine.

4.2.2 Model input

The training dataset is used to build a computational model to learn potential patterns hidden in the dataset. Prior to model construction, data collection and pre-processing of the training dataset were performed. Datasets with too much noise or imbalanced distribution may lead to unsatisfactory prediction performance of the model. There are two main ways to collect coiled-coil domain data to build the model. In some studies, the CCDs were extracted with SCOP [201] and SOCKET [88], while other studies extracted the data directly from a publicly available database regarding experimentally verified CCDs, for example CC+ [89]. The CCDs in the CC+ database were annotated manually and with SOCKET, which has been widely used to extract reliable CCDs from protein structures. A cut-off value of 7.0Å was usually used for extracting coiled-coils from protein structures. Removal of sequence redundancy, an important step prior to model construction, was performed using CD-HIT [202].

4.2.3 Models construction and development

Relatively simple classification methods predict if a protein sequence contains a CCD or not. More sophisticated predictors perform multiclass classifications that categorise coiled-coil regions into different forms of α-helical assembly, such as dimer, trimer and tetramer. We discuss below the different algorithms used in the predictors (Table 1).

COILS, the first reported algorithm for CCD prediction, is a statistically controlled predictor based on the amino acid profile-based method. The similarity of a protein sequence with a structurally known protein is computed using a sliding window. The recommended window length for COILS is 28 in order to help remove false positives. PCOILS is an updated version of COILS that predicts coiled-coils through comparing

pairwise protein evolution profiles based on user-provided multiple sequence alignment or PSI-BLAST [154]. Paircoil2 is the latest development of PAIRCOIL [92]. These predictors use pairwise residue correlations or probabilities to detect the coiled-coil motif in a protein sequence. The training dataset of Paircoil2 is larger than that used for training PAIRCOIL due to the dramatically increased number of known coiled-coil sequences. SOSUIcoil uses amino acid physical properties to help determine an appropriate heptad register, followed by canonical discriminant analysis to discriminate coiled-coils.

Hidden Markov Models (HMM) has been employed in a number of coiled-coil predictors. These include MARCOIL, CCHMM_PROF and SpiriCoil. CCHMM_PROF is an improved version of CCHMM [203], which used multiple sequence alignments instead of single sequence-based HMM. MARCOIL also uses single sequence-based HMMs whereas SpiriCoil uses a large library of HMMs to predict coiled-coils that fall into known superfamilies. The application of SpiriCoil is limited to sequences that have reasonably high similarity to known families due to use of the training dataset for constructing SpiriCoil. On the other hand, MARCOIL, which uses explicit knowledge of existing coiled-coils to train a single HMM, possesses a more complicated algorithm to efficiently search for a variable length subsequence of high probability for coiled-coil formation. According to the HMM parameter *t*, MARCOIL model has two variations, MARCOIL-L (t=0.001) and MARCOIL-H (t=0.01).

MultiCoil [96], a predictor developed based on the PAIRCOIL algorithm, extends the dimeric coiled-coil prediction in PAIRCOIL to trimeric coiled-coils, using a multi-dimensional scoring approach. Multicoil2 further extends the algorithm to include pairwise correlations with HMM in a Markov Random Field (MRF). Multicoil2 also contains eight sequence-based features (including dimer probability, trimer probability, non-coiled probability, dimer correlations at distance 1-7, trimer correlations at distance

1-7, non-coiled correlations at distance 1-7, the hydrophobicity at the *a* and *d* positions) that are used to train the model (pairwise correlation HMM). The resulting algorithm integrated the sequence features and the pairwise interactions into a multinomial logistic regression to formulate an optimized scoring function for the classification of coiled-coil oligomeric state.

SCORER [204] employs a log-odd-based scoring system for the classification of coiled-coil sequences into parallel dimeric and trimeric coiled-coils. SCORER 2.0 combines an expanded and updated training set and a Bayes factor method, which takes into consideration the possible uncertainty in the profile tables. LOGICOIL [97] is a predictor based on the combined and concurrent application of Bayesian variable selection and multinomial probit regression. The application of Bayesian paradigm can provide informative posterior distributions on the selected parameters as well as offering a framework to apply this useful information based on biological data and expert knowledge. Traditional machine-learning techniques, including support vector machine (SVM) [205] and random forest [206] have also been applied to coiled-coil oligomeric state prediction. For example PrOCoil adopts an SVM based on identified rules converted into weighted amino acid patterns. In addition to PrOCoil, PrOCoil-BA (PrOCoil-Balanced Accuracy) is an alternative model, which is optimized for balanced accuracy, i.e., the average of sensitivity and specificity. RFCoil uses random forest combined with effective amino acid indices selected by Gini (a decision tree split function) decrease [207] and Kendall rank correlation coefficient [208].

4.2.4 Model evaluation

A variety of methods were used to assess the prediction performance of coiled-coil predictors listed in Table 1, including cross-validation, leave-one-out cross-validation, leave-family-out cross-validation, independent test and case study. Normally, cross-

validation can avoid over-fitting caused by the training dataset. The nature of cross-validation is to split the dataset into *N* folds and combine *N*-1 folds as the training dataset, leaving the remaining fold as the test dataset. Leave-one-out cross-validation and leave-family-out cross-validation are variations of cross-validation. Given a dataset with *D* data samples, leave-one-out cross-validation combines *D*-1 samples as the training dataset and leaves the remaining one sample as the test sample. In this cross-validation, all samples in the dataset are treated as a test sample once. If the dataset is collected from different species/families, each subset from the same species/family are regarded as test datasets once and other subsets from other families/species will be combined to form the training dataset. The final performance for cross-validation is often averaged from the results of different combinations of the training datasets. The independent test is another method to assess the performance of bioinformatics tools. To test the performance of an algorithm on a new dataset with a different data distribution, it is important to ensure that there is no overlap between the training dataset and the independent test dataset. Finally, the case study is as an effective way to test the performance of a method in real-world applications, providing useful insights into the method scalability and usefulness with unknown data.

4.2.5 Predictor utility

An important aspect of predictors in the biological research community is to provide a user-friendly web interface or a local tool to enable non-bioinformaticians to apply the model directly to their research. The usefulness of bioinformatics tools depends on three factors, i.e. the web interface, the output and interpretation of prediction results, and the availability of locally runnable software. A user-friendly interface can provide appropriate guidance and instructions to avoid potential mistakes when using the web server. This is especially important when parameter settings are required before

conducting prediction tasks. Among the predictors we tested, those predictors aimed at discriminating coiled-coils from non-coiled-coils (e.g. COILS, PCOILS, Paircoil2 and MARCOIL) require parameter settings before sequence submission. Documents are available online regarding the description of the parameters and their potential effect on the prediction performance. On the other hand, the predictors for coiled-coil oligomeric states are mostly parameter-free. For coiled-coil oligomeric state prediction, only sequence and its heptad register are required as the input (for example, SCORER 2.0, PrOCoil, RFCoil and LOGICOIL). Furthermore, SCORER 2.0, PrOCoil and LOGICOIL are also able to predict sequences without the prerequisite of knowing the coiled-coils/heptad registers by combing coiled-coil prediction and extracting heptad register from MARCOIL, without the necessity of performing a two-stage prediction.

Stand-alone software allows users to perform predictions for a large amount of sequences on local machines, offering an advantage over web servers. Among the coiled-coil predictors reviewed in this article, SpiriCoil and SOSUIcoil do not have available locally runnable tools. The local versions of SCORER 2.0, PrOCoil, RFCoil and LOGICOIL were written using the R package (http://www.r-project.org/). PrOCoil has been integrated with R so it can be downloaded and installed with the R console. Users should be aware of the difference in the length of the coiled-coils in the training datasets of different frameworks especially for the oligomeric state prediction. For SCORER 2.0, MultiCoil2, PrOCoil, RFCoil and LOGICOIL, the minimum lengths of their training coiled-coils are 15, 21, 8, 8 and 15, respectively. This means that one should take into consideration the length of the sequence when choosing appropriate predictors in order to obtain better prediction results. Although coiled-coil predictors recommend the preferable sequence lengths of coiled-coils, they can still predict the oligomeric state of the coiled-coils shorter than the specified length thresholds. Under such circumstance, it is the users'

responsibility to choose an appropriate predictor according to the length of query sequence before its submission.

Understandable and visualizable interpretation of the output is also important for better understanding the prediction results and their significance. The output of the coiled-coil predictors we reviewed is often organized in two ways, based on either a residue or a sequence basis. Most of the predictors for discrimination of coiled-coils from non-coiled-coils provide prediction outputs on a residue basis, which allows users to gain a detailed insight into each amino acid and its predicted score/probability. Moreover, COILS, PCOILS, Paircoil2 and MARCOIL also provide the visible plots of predicted score/probability for each amino acid and enable users to obtain an overview of predicted scores for the entire sequence. On the other hand, the predictors of coiled-coil oligomeric state (including SCORER 2.0 and LOGICOIL) provide only a final decision and an overall prediction score. These scores are not easy to interpret and understand. PrOCoil provides both prediction scores and visible plots for each amino acid. RFCoil, on the other hand, provides a matrix showing the probability of the query sequence forming a dimeric coiled-coil or a trimeric coiled-coil, which is relatively easy to understand.

Table 4.2 The list of nine human disease-related PolyQ proteins

Protein	Protein length	PolyQ tract	UniProt identifier	Associated Disease
TATA binding protein	339	58-95	P20226	Spinocerebellar ataxia 17 [138-140]
Huntingtin	3142	18-38	P42858	Huntington Disease [137]
Ataxin-1	815	197-208 212-225	P54253	Spinocerebellar ataxia 1 [141, 142]
Ataxin-2	1313	166-188	Q99700	Spinocerebellar ataxia 2 [209-211] and Amyotrophic lateral sclerosis 13 [212]
Voltage-dependent P/Q-type calcium channel subunit alpha-1A (Brain calcium channel I)	2505	2314-2324	O00555	Spinocerebellar ataxia 6 [213-216]
Atrophin-1	1190	484-502	P54259	Dentatorubro-pallidoluysian atrophy [217]

Ataxin 7	892	30-39	O15265	Spinocerebellar ataxia 7 [218]
Androgen receptor	919	58-78	P10275	Spinocerebellar muscular atrophy or Kennedy Disease [219]
Ataxin-3	364	296-305	P54252	Spinocerebellar ataxia 3 or Machado-Joseph Disease [136]

4.2.6 A case study of coiled-coil prediction for human PolyQ proteins

As an extended test of the reviewed coiled-coil predictors we examined the prediction consistency for nine disease-associated Polyglutamine (PolyQ) proteins. We submitted their sequences to the corresponding web servers and obtained the prediction results. PolyQ proteins contain a stretch of repeated glutamine residues (termed the "PolyQ tract"). PolyQ repeats with more than seven residues are abundant in 128 proteins in the human proteome [120]. These repeats have important biological functions especially in transcription regulation, and proteins harbouring expanded PolyQ repeats are involved in neurodegenerative diseases [220]. The PolyQ diseases are caused in part by a gain-of-function mechanism of neuronal toxicity involving protein conformational changes that result in the formation and deposition of β-sheet rich aggregates [144]. Since PolyQ repeats are highly aggregation-prone [144], it is difficult to determine their structure by X-ray crystallography [147]. The widely accepted model of β-sheet-mediated aggregation has been recently challenged by experimental and bioinformatics studies showing that disease-associated PolyQ proteins contain CCDs largely overlapping with their PolyQ repeats [198]. We therefore investigated the prediction of CCDs in human proteins containing PolyQ repeats, using the dataset containing the most updated nine disease-associated PolyQ proteins from UniProt database studied by Fiumara *et al.* [198], which is also available in the PolyQ database [120] (http://pxgrid.med.monash.edu.au/polyq/; Table 4.2).

4.3 Results and discussion

4.3.1 Independent test and performance evaluation

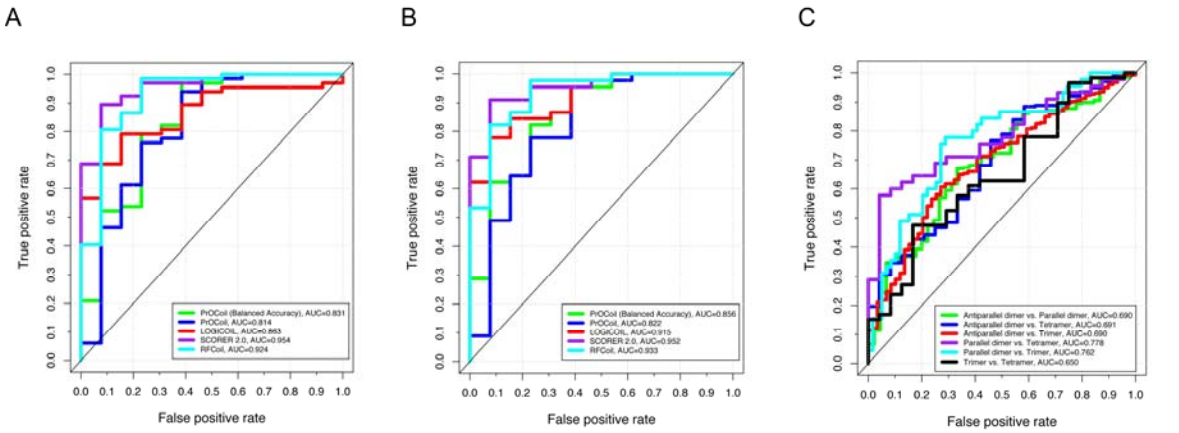
In this section, to assess the prediction performance of the reviewed coiled-coil tools in an objective and fair manner, we assembled two independent test datasets (discussed below) and measured the performance (in terms of AUC) of all tested tools on these two datasets. In particular, since the previous versions of CCHMM, SCORER and MultiCoil have been upgraded as CCHMM_PROF, SCORER 2.0 and Multicoil2, respectively, we only evaluated the advanced versions in the independent test. In addition, as SOSUIcoil and SpiriCoil did not provide local executables, and it was not possible to run Paircoil2 without execution errors, these three predictors were not included in this test. According to the nature of the prediction tasks, we performed independent tests for two different types of tasks, namely, coiled-coil oligomeric state prediction and CCD prediction. Coiled-coil oligomeric state prediction usually requires CCDs and their heptad registers (i.e. *a*-*g*) as the input, while CCD prediction often takes protein sequences as input. For the first type, we evaluated the performance of coiled-coil oligomeric state predictors, including RFCoil, PrOCoil, SCORER 2.0, LOGICOIL and Multicoil2. For the second type, we compared the prediction performance of COILS, PCOILS, MARCOIL, CCHMM_PROF and Multicoil2.

(1) Coiled-coil oligomeric state prediction

Test dataset construction. We carefully prepared two different test datasets. For the first dataset, CCDs and their respective heptad assignments were extracted from the PDB using SOCKET [88]. Only X-ray crystal structures were selected to ensure the quality of the dataset (downloaded on 6-May-2014). SOCKET was applied to annotate the coiled-coils in a given structure with a default packing cutoff of 7.0Å, which was the same as that specified in the dataset collection procedure of previous studies [99, 100]. In

addition, to improve the quality of the dataset, we further removed those structures with a resolution of worse than 4.0Å. Meanwhile, the structures with unnatural residues were also removed. For the second dataset, we first culled coiled-coil class (h class) proteins from SCOPe [221] (the extended version of SCOP) and then verified the CCDs with SOCKET. Only the consensus sequences assigned by both SCOPe and SOCKET analysis that contained coiled-coils were retained to constitute the second dataset, whereas the coiled-coil and heptad annotations were obtained by SOCKET. We subsequently examined the overlap between the second dataset and the training datasets of RFCoil, PrOCoil, SCORER 2.0 and LOGICOIL. Our analysis showed that the majority of entries in the second dataset were covered by the training datasets of the four predictors, suggesting that the second dataset was not sufficiently large enough to be an independent test dataset. Therefore, to address this, we first removed all the training data of investigated predictors from our datasets and then combined the first dataset, second dataset and other training datasets of the four predictors, and used CD-HIT to reduce the sequence redundancy of the resulting dataset to ensure that the sequence identity of any two sequences in the dataset was no more than 50%. For each cluster generated by CD-HIT, if all sequences in this cluster were from our first and second datasets, the representative sequence was collected. Although sequence redundancy can be reduced by other alternative ways, 50% has been commonly used as the preferred threshold for CCDs, since any threshold lower than 50% is deemed to be too strict for coiled-coil oligomeric state prediction [97]. Finally, the independent test dataset contained 509 antiparallel dimers, 88 parallel dimers, 94 trimers and 36 tetramers (Table S4.1; Additional file 1 - http://lightning.med.monash.edu/coiledcoil/).

Performance comparison. Among the four reviewed predictors, RFCoil and PrOCoil were trained using coiled-coils with length equal to or longer than 8 amino acids, while SCORER 2.0 and LOGICOIL were developed using coiled-coils with length longer than 14 residues. In addition, RFCoil, PrOCoil and SCORER 2.0 were designed to classify parallel dimeric and trimeric coiled-coils. LOGICOIL is the only currently available predictor that can be used to predict four types of coiled-coil oligomeric states including parallel/antiparallel dimers, trimers and tetramers. Therefore, to comprehensively evaluate the performance of these tools for predicting the two different types of coiled-coils, we first split the independent test dataset into two subsets, one with coiled-coils longer than 7 residues and the other with coiled-coils longer than 14 amino acids. For each subset, we evaluated the prediction performance using AUC (Area Under the Curve) values. This included the performance comparison of parallel dimer and parallel trimer between the four predictors, as well as pairwise performance comparison of LOGICOIL. The ROC (Receiver Operating Characteristic) curves of these different predictors are shown in Figure 2. We also notice that certain heptad registers for CCDs from SOCKET are non-canonical, which means that the heptad registers (i.e. *a*-*g*) are interrupted according to SOCKET annotations. In view of this, we further removed the coiled-coils with non-canonical heptad assignments and repeated our tests (Additional file 2 downloadable at http://lightning.med.monash.edu/coiledcoil/). The corresponding ROC curves of all predictors for predicting these coiled-coils without non-canonical heptad registers are shown in Figure 3. For Figures 3.1A, 3.1B, 3.2A and 3.2B, "positive" and "negative" indicate parallel dimeric and trimeric coiled-coils, respectively.



Predictor	95% Confidence Interval
PrOCoil-BA	0.689-0.974
PrOCoil	0.654-0.974
LOGICOIL	0.774-0.953
SCORER 2.0	0.901-1.0
RFCoil	0.830-1.0

Predictor	95% Confidence Interval
PrOCoil-BA	0.731-0.982
PrOCoil	0.667-0.978
LOGICOIL	0.837-0.992
SCORER 2.0	0.898-1.0
RFCoil	0.855-1.0

Task	95% Confidence Interval
Antiparallel dimer vs. Parallel dimer	0.610-0.770
Antiparallel dimer vs. Tetramer	0.585-0.798
Antiparallel dimer vs. Trimer	0.620-0.761
Parallel dimer vs. Tetramer	0.669-0.886
Parallel dimer vs. Trimer	0.669-0.856
Trimer vs. Tetramer	0.520-0.780

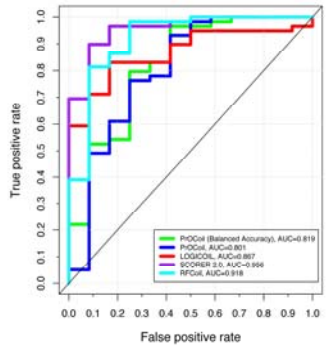
Figure 4.2 Performance comparison of coiled-coils with non-canonical heptad registers between RFCoil, SCORER 2.0, PrOCoil and LOGICOIL on the independent test. (A) ROC curves and the 95% Confidence Intervals for parallel dimeric and trimeric coiled-coils with length ≥ 8 amino acids. (B) ROC curves and the 95% Confidence Intervals for parallel dimeric and trimeric coiled-coils with length ≥15 amino acids. (C) ROC curves and the 95% Confidence Intervals of LOGICOIL for pairwise oligomeric state prediction with coiled-coils with length ≥15 residues.

We note that generally, when testing with parallel dimeric and trimeric coiled-coils with only canonical heptads, LOGICOIL and RFCoil achieved the highest AUC values (see Figures 4.1A, 4.1B, 4.2A and 4.2B). Although LOGICOIL was trained using longer coiled-coil sequences most of which contained canonical heptads, it was able to predict shorter coiled-coils with non-canonical heptads. Pairwise AUC values can be observed in Figure 3.1C and Figure 3.2C, where LOGICOIL achieved the highest AUC values when predicting parallel dimer and tetramer (with AUC values of 0.771 and 0.794, respectively). However, distinguishing tetramer from trimer appears to be the most challenging task. PrOCoil-BA performed constantly better than PrOCoil when tested with

both short and long coiled-coils (see Figures 4.1A, 4.1B, 4.2A and 4.2B). In addition to AUC values, we also computed the 95% Confidence Interval using the 'pROC' package [222]. The 95% Confidence Intervals are shown for each ROC curve in the corresponding tables in Figures 4.1 and 4.2. It can be seen that most of the 95% Confidence Intervals are overlapped. This suggests that even though the compared predictors achieved different AUC values, it is difficult to determine which predictor is the 'statistically significant' best model. For each of the parallel dimeric and trimeric testing samples, we also applied majority voting to generate consensus results and compared the performance of majority voting with other individual predictors (Tables S4.2 and S4.3). It is clear that majority voting could indeed improve the prediction accuracy when testing oligomeric state prediction of coiled-coils with length ≥15 amino acids that contained both canonical and non-canonical heptad registers. Since dimeric coiled-coils are more prevalent than trimer and tetramer, all these predictors were trained with imbalanced training datasets. Accordingly, some predictors are highly biased. For example, when testing RFCoil, we noticed that RFCoil could readily predict dimeric coiled-coils with high confidence, but often wrongly predicted many trimers as dimers. This is probably because of the limited number of trimers included in the training dataset and hence the trained RFCoil model did not generalize and perform well on trimer prediction. Therefore, to address this problem in future work, we recommend that certain techniques for imbalanced data processing and mining be applied (e.g. oversampling or undersampling) to enrich the imbalanced samples. Oversampling and undersampling [223] are both basic (opposite but equivalent) methodologies for sampling the data with imbalanced class distribution. Oversampling is a technique that randomly selects samples from the class where the number of samples is quite small in order to enrich the samples in this class, while undersampling randomly selects samples from the class where the number of samples in this class is large in order

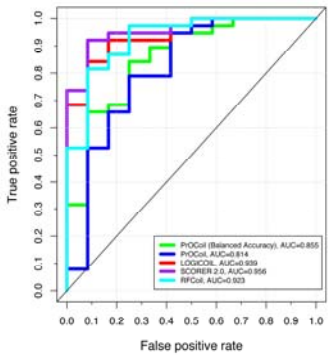
to reduce the number of samples in this class. These two techniques are basic and easy to implement. More complex and advanced techniques for imbalanced biological/medical data mining tasks also exist [224-226].

A



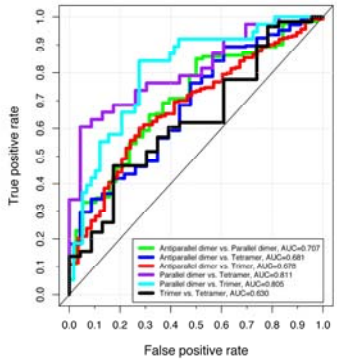
Predictor	95% Confidence Interval
PrOCoil-BA	0.668-0.970
PrOCoil	0.630-0.971
LOGICOIL	0.777-0.958
SCORER 2.0	0.901-1.0
RFCoil	0.815-1.0

B



Predictor	95% Confidence Interval
PrOCoil-BA	0.726-0.985
PrOCoil	0.648-0.980
LOGICOIL	0.872-1.0
SCORER 2.0	0.902-1.0
RFCoil	0.841-1.0

C

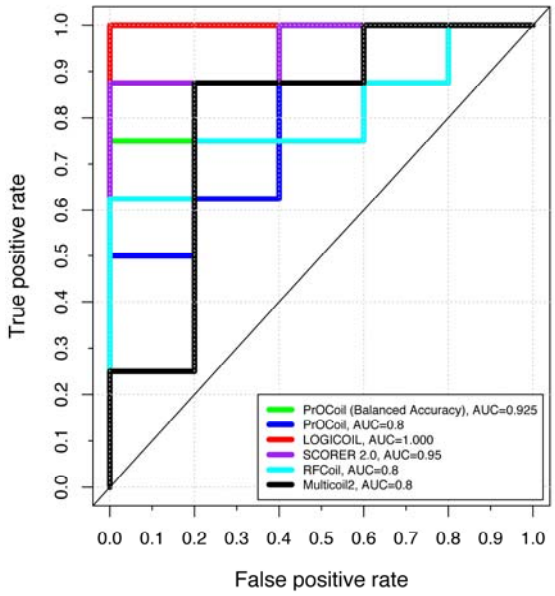


Task	95% Confidence Interval
Antiparallel dimer vs. Parallel dimer	0.623-0.792
Antiparallel dimer vs. Tetramer	0.569-0.793
Antiparallel dimer vs. Trimer	0.605-0.750
Parallel dimer vs. Tetramer	0.705-0.918
Parallel dimer vs. Trimer	0.716-0.895
Trimer vs. Tetramer	0.496-0.764

Figure 4.3 Performance comparison of coiled-coils without non-canonical heptad registers between RFCoil, SCORER 2.0, PrOCoil and LOGICOIL on the independent test. (A) ROC curves and the 95% Confidence Intervals for parallel dimeric and trimeric coiled-coils with length ≥8 amino acids. (B) ROC curves and the 95% Confidence Intervals for parallel dimeric and trimeric coiled-coils with length ≥15 amino acids. (C) ROC curves and the 95% Confidence Intervals of LOGICOIL for pairwise oligomeric state prediction with coiled-coils with length ≥15 residues.

We next compared the prediction performance of Multicoil2 and other predictors. Multicoil2 accepts the full-length protein sequences as the input rather than coiled-coil sequences and their respective heptad registers. Instead of providing an overall score for the input sequence, Multicoil2 generates predicted probabilities for each individual residue in the sequence to form parallel dimers, parallel trimers or non-coiled-coils. Here,

to compare with other methods, we calculated the average of the predicted probabilities by Multicoil2, normalized the value into the range of [0,1] and removed the predicted non-coiled-coils from the results (the prediction threshold was set as 0.5). We combined the parallel dimeric and trimeric coiled-coils with length longer than 21 amino acids (given that Multicoil2 can only predict CCDs with length longer than 21 amino acids) in the dataset used in our independent test with the dimers and trimers sequences in the Multicoil2 training dataset and applied CD-HIT to remove the sequence redundancy, ensuring that the identity between any two sequences in the resulting dataset was no more than 50%. As a result, only 22 CCDs remained in the resulting dataset. For the remaining CCDs, we downloaded their complete protein sequences so that we could use them as the input to Multicoil2. Multicoil2 predicted only 11 out of 22 (50.0%) sequences that contained CCDs that overlapped with SOCKET annotation. Therefore, we compared only the prediction performance of different predictors on these 11 'valid' CCDs (Figure 4.3; Additional file 3 - http://lightning.med.monash.edu/coiledcoil/). In Figure 4.3, "positive" and "negative" represent parallel dimeric and trimeric coiled-coils, respectively. LOGICOIL correctly classified all the parallel dimeric and trimeric coiled-coils, while Multicoil2 and PrOCoil obtained the lowest AUC value. Consistent with the results in Figures 4.1 and 4.2, PrOCoil-BA performed better than PrOCoil (greater by 0.2), followed by RFCoil and SCORER 2.0. In addition, the 95% Confidence Intervals suggest that LOGICOIL was the best predictor based on this independent testing dataset. Consistent with the AUC values shown in Figure 4.3, LOGICOIL correctly classified all the test samples. It is noteworthy that the majority voting strategy achieved an accuracy of 90.9%, which was ranked as the second best accuracy according to the accuracies of other individual predictors (Table S4.4).



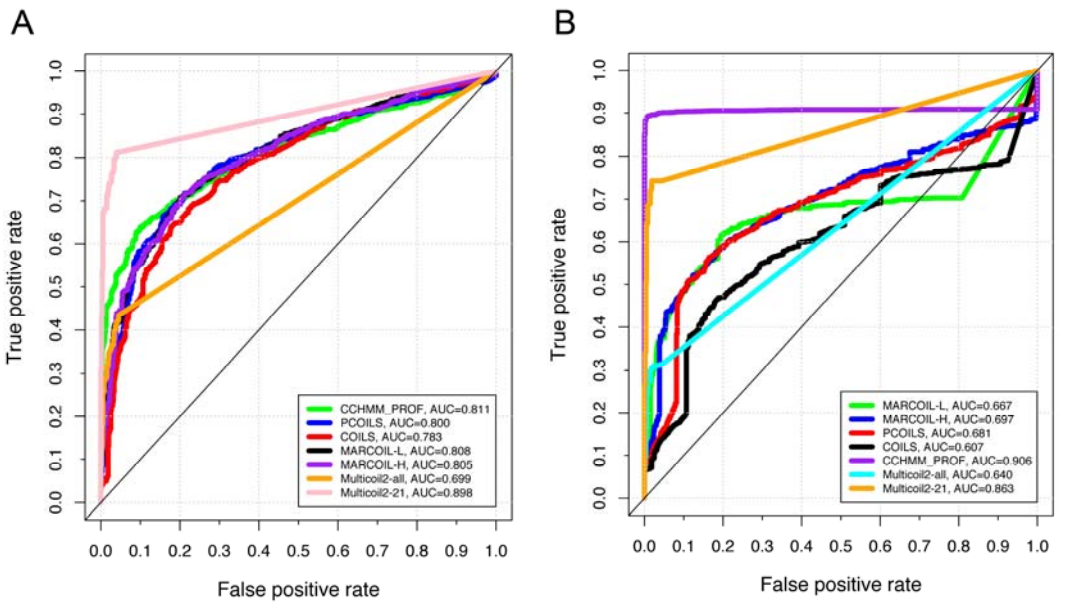
Predictor	95% Confidence Interval
PrOCoil-BA	0.783-1.0
PrOCoil	0.541-1.0
LOGICOIL	1.0-1.0
SCORER 2.0	0.835-1.0
RFCoil	0.545-1.0
Multicoil2	0.497-1.0

Figure 4.4 ROC curves and the 95% Confidence Intervals of Multcoil2 and other predictors for parallel dimeric and trimeric coiled-coil prediction.

(2) CCD prediction

Testing dataset construction. The positive dataset comprised protein sequences containing annotated CCDs based on SOCKET. For the negative dataset, we extracted protein sequences of alpha and beta classes (a/b; i.e. c class) from the SCOPe database, except for superfamilies c.37.1, c.49.2, c.67.1 and c.93.1 which are annotated to contain CCDs [195]. Protein sequences were extracted from PDB and those sequences that contain unnatural amino acids were removed. These sequences were further validated by SOCKET with a loosened threshold of 7.4Å [93] to ensure they did not contain any CCDs. After removing all the available training data of investigated predictors from our testing

dataset, we combined our testing datasets with the available training datasets of CCHMM_PROF, MARCOIL and Multicoil2. We then applied CD-HIT to remove the sequence redundancy so that the sequence identity between any two sequences was not greater than 30%. Similar to the construction process of the independent test dataset for CCD oligomeric state prediction, for each cluster generated by CD-HIT, only representative sequences from the clusters where there were no samples from the training datasets of the compared predictors in this cluster were collected. After this procedure, the final dataset included a total of 1,643 sequences, 601 of which did not contain any CCDs and 1,042 containing 2,176 CCDs (Additional files 4 and 5 - http://lightning.med.monash.edu/coiledcoil/). CCHMM_PROF and PCOILS require the PSSM (position-specific scoring matrix) generated by PSI-BLAST as the input to make the prediction. Accordingly, we used the Uniref90 database to generate the PSSM profiles of all the tested sequences and conduct the comparison, which was also used as the search database by CCHMM_PROF [93]. The parameters for PSI-BLAST was preliminarily set by the PCOILS program; for CCHMM_PROF, we used the same parameters described in [93].



Predictor	95% Confidence Interval
CCHMM_PROF	0.790-0.831
PCOILS	0.779-0.822
COILS	0.760-0.805
MARCOIL-L	0.787-0.829
MARCOIL-H	0.784-0.826
Multicoil-all	0.682-0.716
Multicoil-21	0.867-0.928

Predictor	95% Confidence Interval
CCHMM_PROF	0.894-0.919
PCOILS	0.660-0.703
COILS	0.584-0.629
MARCOIL-L	0.648-0.687
MARCOIL-H	0.677-0.717
Multicoil-all	0.628-0.652
Multicoil-21	0.835-0.891

Figure 4.5 Performance comparison of coiled-coil domains predictors. (A) ROC curves and the 95% Confidence Intervals of different predictors for identifying coiled-coil domains. (B) ROC curves and the 95% Confidence Intervals of different predictors, showing the consistency between the predicted coiled-coil domains and those annotated by SOCKET based on the protein structures.

Performance comparison. Firstly, we evaluated the effectiveness of different predictors for identifying CCDs by calculating the averaged probability score for each protein. If a protein was predicted to contain coiled-coil residues, the probability was calculated as the averaged score of all predicted coiled-coil residues; otherwise if a protein was not predicted to have CCDs, then the calculated probability was the averaged score of all residues of the whole protein. The ROC curves and corresponding AUC

values of the compared predictors are shown in Figure 4.4A, where the "positive" represents the sequences containing CCDs while the "negative" indicates the sequences without CCDs. Since Multicoil2 can only predict protein sequences with CCDs longer than 21 amino acids, we provided the results of Multicoil2 on both the entire test dataset (termed "Multicoil2-all") and a subset that only contained proteins with coiled-coils longer than 21 amino acids (termed "Multicoil2-21"). It is apparent that Multicoil2-21 identified the majority of coiled-coils and achieved the highest AUC value of 0.898, followed by CCHMM_PROF (AUC=0.811). The AUC value of PCOILS was higher than COILS by 0.017, presumably due to the incorporation of evolutionary information in the form of PSSM generated by PSI-BLAST. Next, we examined whether the identified CCDs were identical to those annotated by SOCKET. To do so, we compared all 2,176 CCDs and their corresponding prediction scores of all reviewed predictors. A domain was predicted as a coiled-coil domain if its probability was larger than 0.5. For the negative protein (i.e., proteins without CCDs), if it was predicted to have a coiled-coil domain, the average score would be calculated; otherwise the average prediction score for each residue in this protein would be calculated. The results are shown in Figure 4.4B, where the "positive" denotes CCDs while the "negative" indicates the sequences without CCDs. Similar to Figure 4.4A, CCHMM_PROF and Multicoil2-21 again achieved the highest and second highest AUC values (AUC=0.906 and 0.863, respectively), suggesting that the majority of their predicted CCDs were consistent with the SOCKET assignment. COILS obtained the lowest performance with an AUC score of only 0.607. We also note that Multicoil2-all achieved a lower AUC score, possibly due to its restriction of having a length requirement of coiled-coils during the model training. The performance comparison results between individual predictors and majority voting are shown in Table S5. Since the minimum length of coiled-coils used for training Multicoil2 is 21, we

further filtered the testing dataset with different thresholds of coiled-coil lengths to perform the CCD coverage test. Although majority voting did not actually improve the overall prediction accuracy, the performance of majority voting was still competitive compared with individual predictors (Table S4.5).

4.3.2 CCD and CCD oligomeric state prediction for human PolyQ proteins

(1) Identification of CCDs

We first made a consensus-based decision for CCD prediction based on the predictors that are capable of discriminating coiled-coils from non-coiled-coils. The predictors used in this step were COILS, PCOILS, Paircoil2 (the p-score version with different window sizes and probability score version), MARCOIL, CCHMM_PROF, SpiriCoil and Multicoil2. Strikingly, the results are largely inconsistent between different predictors (Tables S4.6 to S4.13), making it difficult to generate a consensus prediction. Only a small portion of the proteins was predicted to harbor CCDs according to the prediction results of PCOILS, Paircoil2 (both p-score and probability score versions), SpiriCoil and Multicoil2. In contrast, COILS, MARCOIL and CCHMM_PROF predicted several CCDs within the nine PolyQ proteins. Most of the predicted coiled-coils overlapped or flanked the PolyQ tract. Based on the prediction results, the final decisions of predicted CCDs were made through majority voting (i.e., the CCD peptides need to be predicted by at least four predictors; the results are listed in Table 4.3). In the prediction of CCDs in nine disease-associated PolyQ proteins by Fiumara *et al* [198], only two relatively old CCDs predictors were used (COILS and Paircoil2). We note that the results of Fiumara *et al.* are inconsistent with our predictions in this study based on several state-of-the-art predictors. This discrepancy highlights that it remains a challenging task to develop reliable and consistent CCD prediction methods, and that attention should be paid

when only a few specific methods are used to make the prediction, especially when these methods are used to guide and interpret experimental investigations such as the studies by Fiumara *et al* [198].

Table 4.3 The consensus CCDs predicted by at least four predictors

Protein	Predicted coiled-coils	Protein structure	Sequence	Overlapping PolyQ tract	Agreed by
Voltage-dependent P/Q-type calcium channel subunit alpha-1A (Brain calcium channel I)	720-747	3BXK (B/D=1955-1975)	AQELTKDEQEEEEAANQKLALQKAKEVA	No	COILS, PCOILS, Paircoil2 (P-score version), CCHMM_PROF, Multicoil2 and MARCOIL
Atrophin-1	793-819	-	AKKRADLVEKVRREAEQRAREEKERER	No	COILS, PCOILS, Paircoil2 (P-score version), CCHMM_PROF, Multicoil2 (cut-off=0.5) and MARCOIL

(2) Prediction of oligomeric state of PolyQ proteins

To examine the potential oligomeric states of the peptides listed in Table 3, we performed the prediction using RFCoil, SCORER 2.0, PrOCoil and LOGICOIL (Tables S4.14 and S4.15). Since COILS, MARCOIL, PCOILS, Paircoil2 and Multicoil2 all provided heptad registers, we used these heptads to facilitate the oligomeric state prediction. As we can see, with different heptad registers, RFCoil, SCORER 2.0 and PrOCoil produced consistent prediction results (dimer formation); while the oligomeric state predictions from LOGICOIL were variable.

4.4 Conclusions

Given the functional significance of coiled-coil domains, computational biologists are motivated to develop more accurate and reliable predictors for coiled-coil domain prediction. Aiming at providing a comprehensive review of coiled-coil predictors to non-bioinformaticians, this article describes and compares a number of widely used coiled-coil predictors in terms of their input, model construction and model evaluation. Independent

tests reveal that LOICOIL achieved the overall highest AUC value when used to predict parallel dimeric and trimeric coiled-coils. For coiled-coil domain prediction, Multicoil2 achieved the highest AUC value when detecting long CCDs in proteins, while CCHMM_PROF achieved the highest AUC value for the coverage of detected CCDs without the length limitation of CCDs. A case study of nine PolyQ proteins demonstrated that coiled-coil predictions were quite different among different predictors, which could further confound the consensus prediction analysis. We conclude that coiled-coil prediction is still a challenging task and we expect that more powerful algorithms with improved prediction performance will emerge with the increasing availability of coiled-coil data.

Chapter 5 Structural Capacitance in Protein Evolution and Human Diseases

Disordered regions of proteins play crucial roles in many biological processes. A number of studies have demonstrated that the disorder-order structural transition (*i.e.* disorder-to-order) can be mediated through the binding of other molecules (so called folding upon binding). I propose an additional mechanism, termed 'structural capacitance', which results in the *de novo* generation of microstructure in previously disordered regions. In accordance with this hypothesis, this chapter has examined the disorder-order structural transitions caused by single point mutations through employing multiple structural algorithms for protein disordered region prediction and online knowledge-base resources. As a result, two tables with experimental candidates have been provided with detailed annotations, thereby facilitating the experimental investigation of the proposed 'structural capacitance' hypothesis.

5.1 Introduction

The loss-of-function paradigm has been widely established to explain the relationships between human diseases and protein function. The tumour suppressor p53 is a good example to explain the loss-of-function paradigm. The somatic mutations occurring within the p53 sequence interrupt the protein-DNA binding, thereby inactivating its function [227]. Consequently, this inactivation eventually leads to a variety of cancers [228]. Following a survey of the human mutation dataset [38], a bioinformatics analysis was performed using structural algorithms to identify mutations predicted to generate localized regions of microstructure in previously disordered regions of target proteins. A new mechanism of protein evolution, termed 'structural capacitance', is proposed to suggest that structural and functional changes of proteins may be achieved through the introduction of point mutations that increase the hydrophobicity of key nucleating amino acids located in predicted structural disordered regions. Once mutated, these residues are predicted to generate new elements of microstructure in previously disordered regions of the protein that are functionally distinct from the parent fold. 'Structural capacitance' focuses on the other paradigm, 'gain-of-function', via generation of microstructures caused by D→O transition upon mutations. The analysis of function and Eukaryotic Linear Motif (ELM) generally agree that in the D→O transition, there seems to be little evidence to indicate that the wild-type residues are functional (Tables S5.1-S5.8). It is possible that the new generation of microstructure could bring novel functions via the 'gain-of-function' scheme (Figure 5.1). While the traditional loss-of-function paradigm has been well established, as its complementary scheme, the gain-of-function via new microstructure remained understudied. Some experimental studies, however, have revealed this scheme with some proteins [229-231].

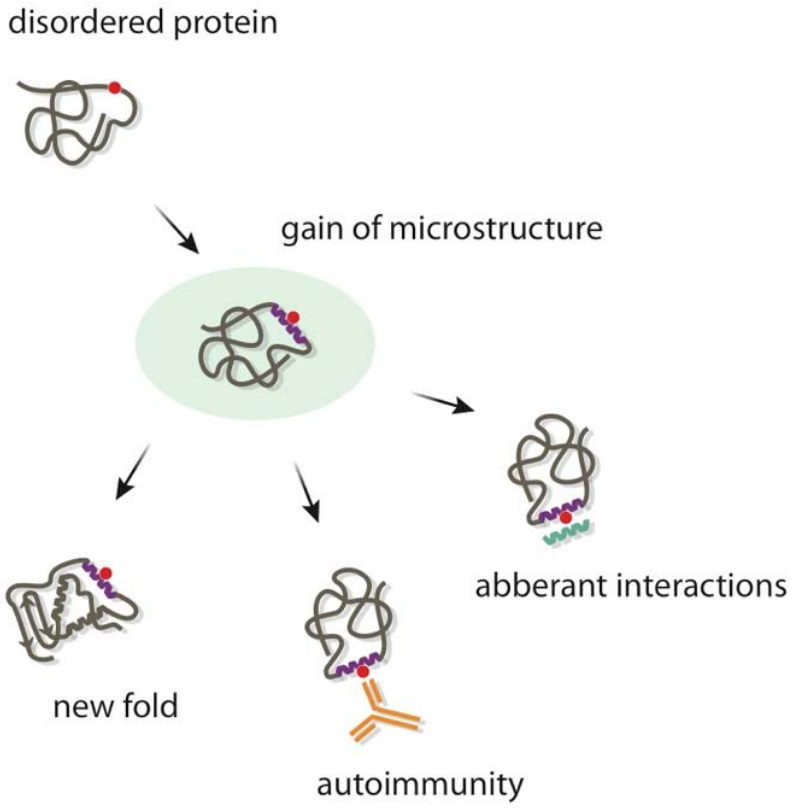


Figure 5.1 Disease-causing mutations may result in gain-of-function through the mechanism of structural capacitance. A D→O mutation (red circle) in a disordered protein results in the generation of local microstructure (purple helix). This may be a key nucleating factor in the evolution of a new adaptive fold, but may also have the potential to generate inappropriate interactions that are pathological, through stimulating inflammatory and autoimmune responses. Aberrant interactions may, furthermore, promote other pathogenic processes such as aggregate formation, which may result in the formation of toxic fibrils.

It has been suggested that the highly disordered proteins tend to easily evolve with new folds [232]. There are mainly two way of evolving: co-evolution of folds and functions through conformational selection from a repertoire of disordered polypeptides,

or the emergence of secondary structure elements followed by the evolution of fully folded proteins [232]. Both scenarios, however, require the prior formation of local structure from an essentially random and disordered population. It is suggested that the formation of local structure are controlled by the 'structural capacitance elements', which are the key mutation causing structuralization (D→O structural transition).

Here, it is necessary to make a distinction among this work and those two works (previously published by Vacic V. *et al.* [23, 36]) regarding disorder-order structural changes. First of all, this work and their published works are independent in parallel. Secondly, rather than focusing on the traditional 'loss-of-function' paradigm, research provided in this thesis is to provide experimental candidates for 'structural capacitance' hypothesis, which explains the potential acquirement of novel functions via the new generation of microstructure caused by single point mutations. Last but not least, the research provided in this thesis extracted the most complete disordered prediction results from currently available databases, to provide reliable protein disordered region predictions, in order to guarantee the quality of selected candidates causing disorder-order structural changes.

In this chapter, with help of currently available databases harbouring protein disordered region prediction results, the disordered/ordered regions predictions were investigated for the manually annotated human disease-associated mutations and polymorphisms dataset, and further provided the statistics for four structural transitions: Disorder-to-Order (D→O), Order-to-Disorder (O→D), Disorder-to-Disorder (D→D) and Order-to-Order (O→O). The analyses regarding to the chemico-physical properties, sequence conservation and functional annotations have also been conducted. To help validate the proposed 'structural capacitance' hypothesis, focus was then shifted to the D→O structural transition and results were further filtered by employing third-party

computational tools to guarantee the quality of selected D→O causing candidates. The two resulting tables provide useful candidates for validating the 'structural capacitance' hypothesis in the laboratory.

5.2 Materials and methods

5.2.1 Databases for protein disordered regions

Two main databases, D^2P^2 and DisProt, were used in this study. D^2P^2 is a knowledge base harbouring disordered region prediction results of a huge amount of proteins from nine computational approaches. Given that some of these predictors are not freely available, the prediction results provided enable the computational investigation of protein disordered regions. Meanwhile, D^2P^2 provides a mapping file so users can search protein disordered region prediction using the UniProt IDs.

The other database, DisProt, provides experimental verified data of protein disordered regions. For each disordered region in DisProt, experimental method and corresponding reference are also available. Compared with the protein disordered region prediction results, the entries stored in DisProt are more reliable and can be used as experimental materials in the laboratory directly.

5.2.2 Computational approaches for protein disordered region prediction

The proteins harboured in the D^2P^2 database are wild type forms. However, this study focuses on investigating the structural changes upon mutations. Therefore, several sequence-based disordered region predictors were employed for the mutated proteins. These computational approaches include: VSL2B, IUPred-S, IUPred-L and DynaMine. To perform a fair comparison, prediction results from DynaMine were also added for wild-type proteins. As a result, there are a total number of 10 predictors for wild-type

protein disordered regions prediction and 4 predictors for mutated protein disordered region prediction, respectively.

5.2.3 Majority voting for consensus decision of protein disordered region prediction

Majority voting is a widely employed strategy for ensemble learning in data mining research area [233]. Here, this strategy was used to obtain consensus decisions on protein disordered region prediction. According to VSL2B and IUPred, amino acids with predicted scores greater than 0.5 are considered to be located within disordered regions, while those with scores less than 0.5 are in ordered regions. For DynaMine, the residues with predicted scores smaller 0.69 are considered to be located in disordered regions, while those with scores larger than 0.8 are predicted to be in the structured regions. According to DynaMine, any amino acids with scores within 0.69 and 0.8 are context dependent, which means it is hard to determine whether the region is disordered or ordered. For the mutated proteins, mutations with DynaMine scores >0.69 were considered to locate in ordered regions, given that the voting would be made by only three predictors if only scores larger than 0.8 are considered to be ordered. A majority voting strategy was employed to decide the prediction results for each mutation. Given that there are multiple predictors for protein disorder prediction, the residues are predicted to locate in disordered regions if the number of predictors that agree the residues to be located in the disordered regions is equal to or larger than the number of predictors that agree the residues to be located in the ordered regions. As a result, and based on the consensus decision by the majority voting, four types of transitions have been explored: Disorder-to-Order (D→O), Order-to-Disorder (O→D), Disorder-to-Disorder (D→D) and Order-to-Order (O→O), respectively.

5.2.4 Human disease-associated mutations and polymorphisms dataset

In this study, the human disease mutations and polymorphisms dataset [38] was used given that the data entries and disease associations harboured in this dataset are manually annotated. Please see Section 1.1.5 for more detailed of this database.

5.2.5 Third-party computational tools for validating protein disordered regions

According to the 'structural capacitance' hypothesis, D→O structural transition is the key to enable protein to obtain novel function via the generation of microstructure. Therefore, multiple third-party computational tools were applied in order to guarantee the quality of protein disordered region prediction for both wild-type and mutated proteins.

Predictor for aggregation propensity upon mutation

Tango [41]. Tango predicts the aggregation propensity of both wild-type and mutated protein sequences. Tango is a computational model for prediction of aggregation nucleating regions in proteins and the effect of mutation on aggregation based on physico-chemical and other sequence-based properties. For disease-causing mutations predicted to involve a D→O structural transition, the aggregation scores for both wild-type and mutated proteins were collected and calculated the changes of the prediction scores. Mutations that are predicted to increase the protein aggregation propensity will be removed.

Predictor for protein transmembrane helices prediction

TMHMM [234]. TMHMM employs hidden Markov model for membrane protein topology prediction. Given the fact the protein transmembrane domains are structurally

stable and ordered, TMHMM was used to further control the predicted disordered regions. If the mutation predicted to locate in disordered region but is also predicted to be in the transmembrane domain, this disordered region prediction will be discarded.

Protein structure BLAST

In order to ensure that wild-type proteins with predicted disordered regions do not have structures or homologues structures current available, a BLAST search against the PDB database (http://www.rcsb.org/pdb/software/rest.do) using the protein sequences was performed. Any proteins with predicted disordered regions and BLAST hits against the PDB database will be removed.

5.2.6 Amino acid hydrophobicity indices used for characterizing amino acid properties in predicted disordered and ordered regions

In order to examine the hydropbobivity changes for the wild-type amino acids and their corresponding mutations for the four structural transitions, three widely used indices were chosen in this study: Eisenberg hydrophobicity index [235], Hopp-Woods hydrophilicity index [236] and Kyte-Doolittle hydropathy index [237].

1. **Eisenberg hydrophobicity index.** This index in a normalized consensus value hydrophobicity index, of which the mean and standard deviation are 0.00 and 1.00, respectively.
2. **Hopp-Woods hydrophilicity index.** Actually this index is hydrophilicity index, which means the higher the score is, the more hydrophilic the amino acid is. This index was used to predict potential antigenic sites of globular proteins
3. **Kyte-Doolittle hydropathy index**. This index is the most commonly used scale that is formed by taking both hydrophilic and hydrophobic properties of 20 amino acids. It can be used to identify hydrophobic regions for surface-exposed regions in

protein sequence. Hydrophobic regions are usually indicated by positive index values.

Table 5.1 lists the value of three indices we used in our analysis. Bold values are considered as hydrophilic amino acids.

Table 5.1 Values of three indices used for characterizing amino acid properties

AA	Kyte-Doolittle scale	Hopp-Woods scale	Eisenberge
A	1.80	-0.50	0.62
C	2.50	-1.00	0.29
D	-3.50	3.00	-0.9
E	-3.50	3.00	-0.74
F	2.80	-2.50	1.19
G	-0.40	0.00	0.48
H	-3.20	-0.50	-0.4
I	4.50	-1.80	1.38
K	-3.90	3.00	-1.5
L	3.80	-1.80	1.06
M	1.90	-1.30	0.64
N	-3.50	0.20	-0.78
P	-1.60	0.00	0.12
Q	-3.50	0.20	-0.85
R	-4.50	3.00	-2.53
S	-0.80	0.30	-0.18
T	-0.70	-0.40	-0.05
V	4.20	-1.50	1.08
W	-0.90	-3.40	0.81
Y	-1.30	-2.30	0.26

5.3 Results

5.3.1 Four types of transitions between protein disordered regions and ordered regions

The proteins with uncommon residues in their sequences have been removed from the dataset. Note that there is a chance that the sequences used for disordered region prediction in D^2P^2 database are not consistent with the protein sequences from the UniProt database, given the fact the UniProt updates its entries regularly. In this case, the prediction results are not valid since the sequences have changed. In light of this, all the sequences from the dataset have been validated with the sequences from the D^2P^2 database. Any inconsistent sequences and their associated mutations have been removed from the dataset. As a result, the resulting dataset remains 11,735 proteins with 63,287 single point mutations.

After mapping the disease-associated mutations and polymorphisms dataset to the D^2P^2 database and employing the protein disordered region predictors mentioned above, I obtained four tables with candidates of D→O, O→D, O→O and D→D transitions based on the majority voting strategy, respectively. Table 5.2 illustrates the statistics of number of mutations causing the four different types of structural transitions.

Table 5.2 Disorder prediction on human disease and polymorphisms dataset

	Mutations in Ordered Regions	**O→O**	**O→D**	**Mutations in Disordered Regions**	**D→D**	**D→O**
Disease	23,139	22,277 (96.3%)	862 (3.7%)	3,584	3,017 (84.2%)	567 (15.8%)
Non-disease	25,088	24,077 (96.0%)	1,011 (4.0%)	11,476	9,990 (87.1%)	1,486 (12.9%)

[1]O = predicted ordered;

[2]D = predicted disordered.

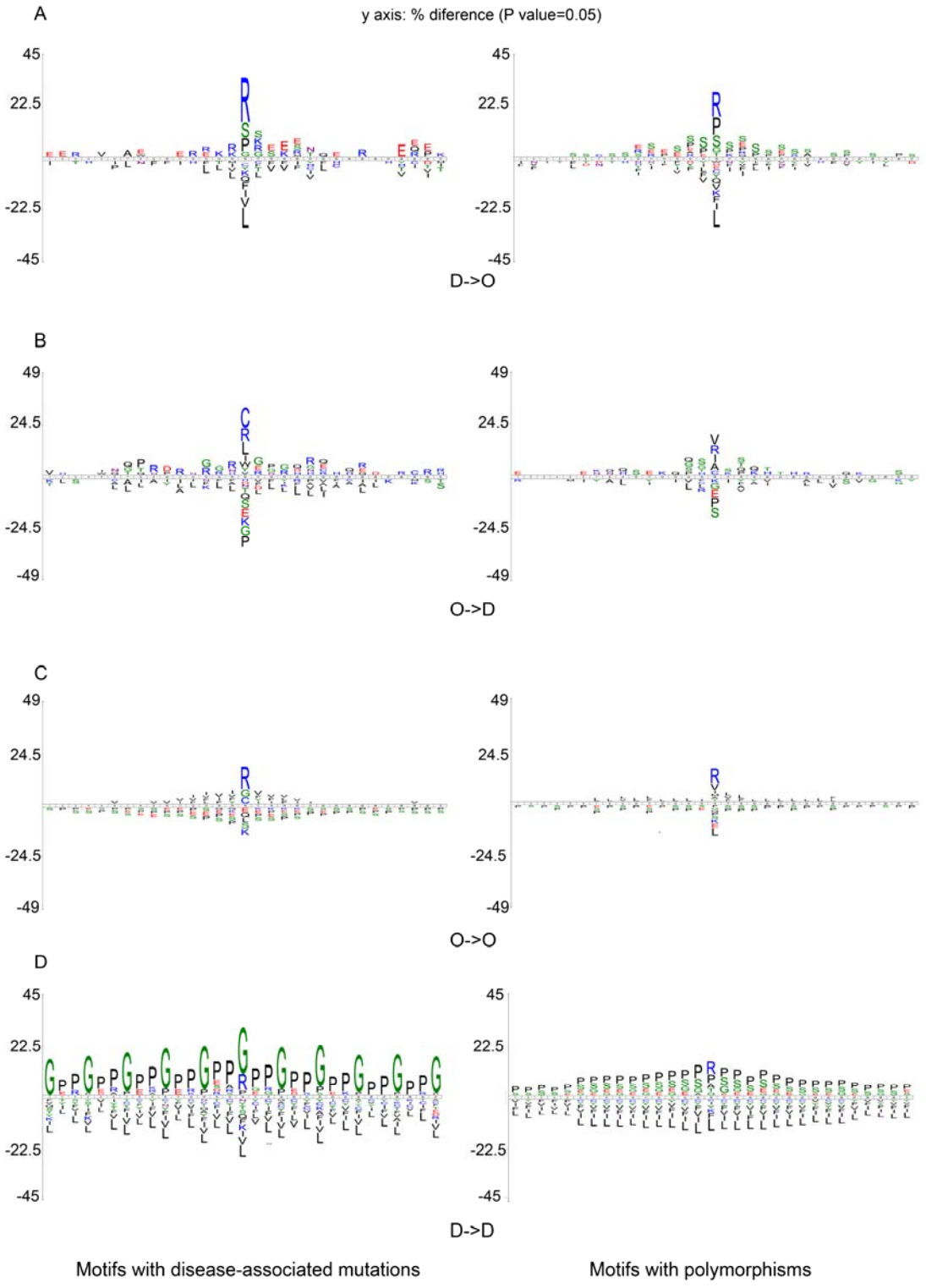


Figure 5.2 IceLogo [238] charts showing the residue conservation around the mutation site against a reference set (human Swiss-Prot proteome) for (A) D→O, (B) O→D, (C) O→O and (D) D→D structural transitions with wild-type residue in the central position. Amino

acids residues on top of the *x* axis are significantly conserved, while those underneath it are non-preferred or unfavored according to the reference set.

Then the sequence motifs using a window (size=31) with the residues to mutate in the central were extracted for both disease-associated mutations and polymorphisms. IceLogo [238] was employed to generate the sequence logos in Figure 5.2 and to calculate the sequence conversation scores. The types of mutations in each class (D→O, O→D, D→D and O→O) appear to be non-random. For all documented disease mutations Arginine is favourable mutated amino acid (Figure 5.2). The most common classes of disease mutation for D→O and O→D transitions are R→W (62 mutations) and L→P (97 mutations), respectively. For D→D and O→O transitions, the mutation patterns are more evenly distributed. This is consistent with a recent comparison of mutation frequencies in intrinsically disordered regions of proteins in both disease and healthy datasets that highlights the previously unappreciated role of mutations in disordered regions [36].

5.3.2 Hydrophobicity changes upon mutations in four transitions

Figure 5.3 shows the mean hydrophobicity change between wild type and mutated residues based on three different hydrophobicity indices (Eisenberg [235], Hopp-Woods [236] and Kyte-Doolittle [237] indices). Regardless of which hydrophobicity index chosen, it is clear that in D→O structural transition, for the wild-type, the majority of amino acids which will mutate are hydrophilic, while for the mutant, the majority of mutant residues are hydrophobic. In O→D transition, opposite trend can be observed. However, in O→O and D→D transitions, this trend is not obvious.

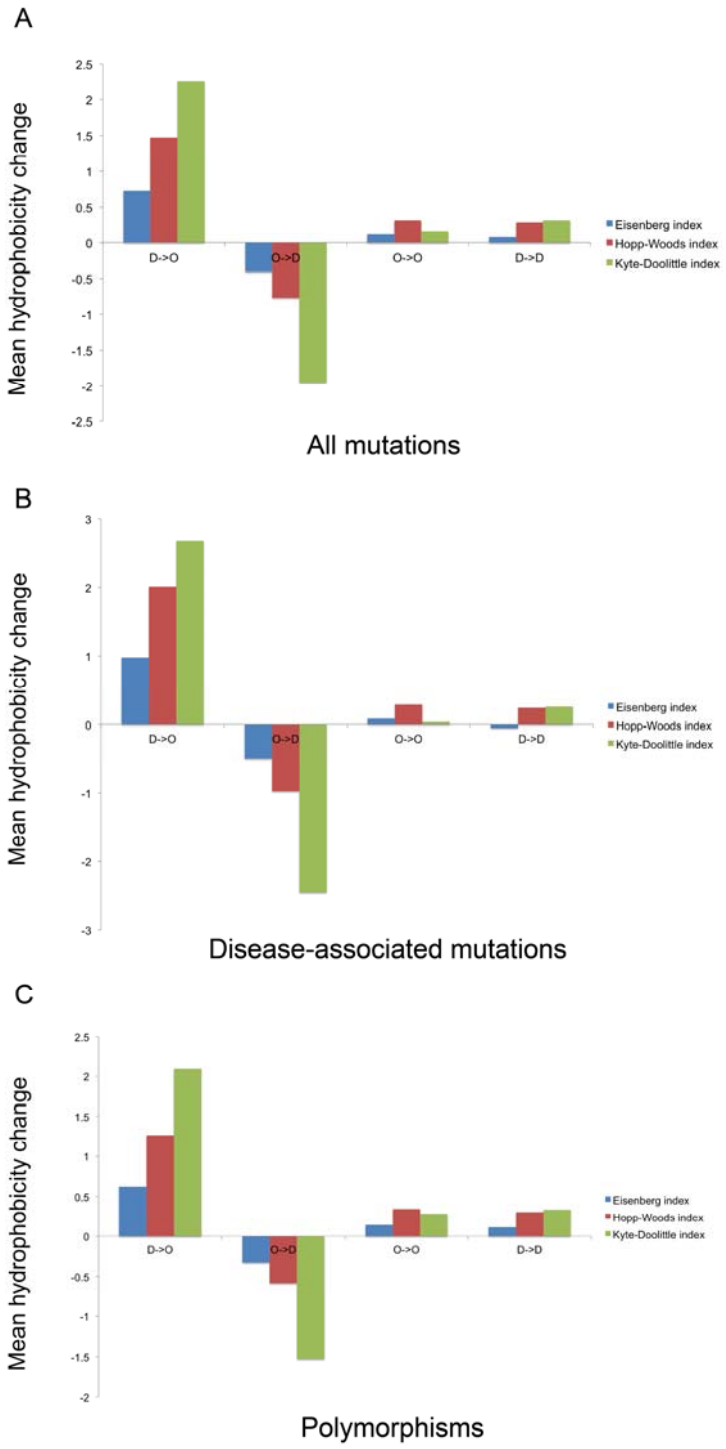


Figure 5.3 Mean hydrophobicity changes for (A) all mutations, (B) disease-causing mutations and (C) polymorphisms for four different classes of structure-altering (i.e., D→O, O→D, O→O and D→D) mutations predicted using D^2P^2 database and multiple sequence-based predictors for intrinsically disordered regions. Bars are shown for the three different

hydrophobicity indices used: Eisenberg hydrophobicity index [235] (Blue), Hopp-Woods hydrophilicity index [236] (Ochre) and Kyte-Doolittle hydropathy index [237] (Green).

5.3.3 Functional analysis of mutations for four structural transitions

Based on the four structural transitions, I further analyzed the function annotations for those amino acids to mutate. According to the UniProt database classification scheme, these functional features include active site, binding site, disulfide bond, glycosylation site, metal-binding site and modified residue. Here, protein active sites refer to the residues that are directly involved in catalysis. While binding sites are the residues interacting with another chemical entity. The comparison results are listed in Figure 5.3.

Generally, due to the fact that the O→O structural transition harbours a larger number of mutations compared with other three structural transitions, lots of functional sites can be found for the mutations in O→O structural transition. It is also noticeable that D→O does not tend to contain functional sites according to the annotations from the UniProt database.

5.3.4 Long disordered regions harbouring D→O causing mutations

In this section, long disordered regions (LDRs) were particularly focused where the mutations cause the D→O transition, in accordance with the proposed 'structural capacitance' hypothesis. Based on the candidates in D→O transition, third-party computational tools were employed to further verify the protein disordered region prediction by applying relatively rigorous standards. These computational tools include: TMHMM for protein transmembrane helices prediction, Tango for prediction of protein aggregation propensity and protein structure BLAST for homologous protein structures

for given protein sequences. In order to ensure the high quality of selected LDRs with D→O causing mutations, only the mutations satisfying the following standards will be remained:

(1) located in the LDRs but are not predicted to be in transmembrane domains;

(2) not predicted to increase the protein aggregation propensity;

(3) protein sequence do not have any structures or homologous structures.

The resulting candidates with detailed mutation and disease association annotations are shown in Table 5.3. The detailed sequence functional and contextual annotations including modified residues, protein superfamily domains [153] and Pfam domains [121] for proteins listed in Table 5.3 are listed in Table S5.9.

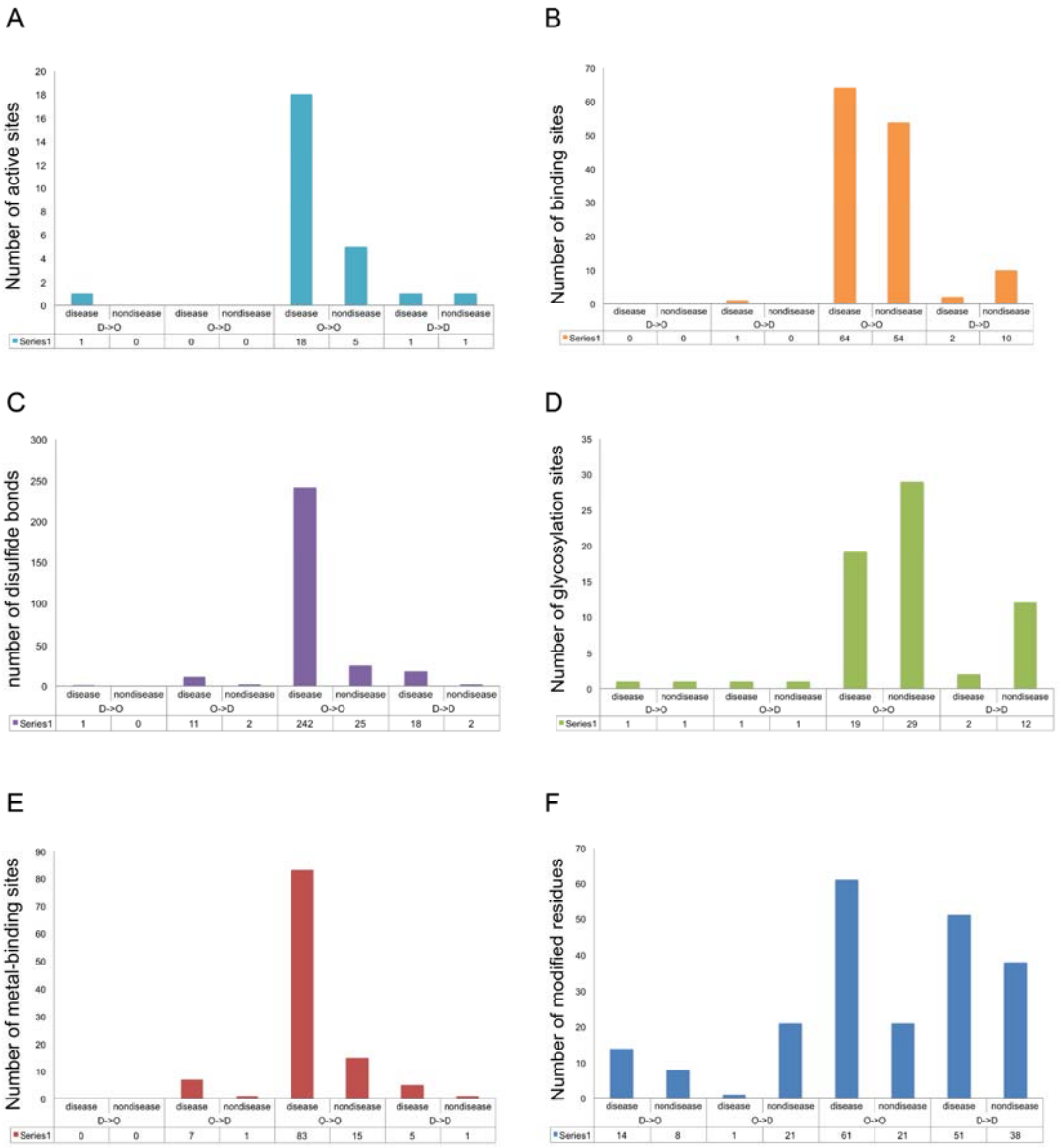


Figure 5.4 Statistics of function sites for both disease and non-disease mutations of four structural transitions in terms of (A) active site, (B) binding site, (C) disulfide bond, (D) glycosylation site, (E) metal-binding site and (F) modified residue.

Table 5.3 Disease-causing mutations and polymorphisms in LDRs of human proteins predicted to produce disorder-to-order transition

UniProt/dbSNP	Protein	Mutation	Disease/Phenotype	# disorder predictors[b]	# order predictors[c]	Average length of DR[d]
O95990/-	Protein FAM107A	PL19	Renal cell carcinoma cell line	6	4	63
Q69YN2/rs7073610	CWF19-like protein 1	PL259	-	7	4	47
Q8TBZ0/rs9683564	Coiled-coil domain-containing protein 110	SL817	-	7	4	33
A0JNW5/rs58214704	UHRF1-binding protein 1-like	ML1111	-	5	3	116
A4D1E1/rs801841	Zinc finger protein 804B	VI1195	-	6	3	31
A5PLN7/rs2276922	Protein FAM149A	PL532	-	5	3	65
A6H8Y1/rs1961760	Transcription factor TFIIIB component B" homolog	FI1244	-	5	3	804
A6NC98/rs1318165	Coiled-coil domain-containing protein 88B	DA886	-	5	3	443
O43303/rs3751821	Centriolar coiled-coil protein of 110 kDa	PL171	-	8	3	51
O60269/rs4445576	G protein-regulated inducer of neurite outgrowth 2	SC328	-	6	3	45
O75691/rs1061436	Small subunit processome component 20 homolog	EQ2612	-	5	3	38
O75952/rs3786417	Calcium-binding tyrosine phosphorylation-regulated protein	TM74	-	5	3	112
O95163/rs1538660	Elongator complex protein 1	PL1158	-	9	3	61
P01286/rs4988492	Somatoliberin	LF75	-	6	3	45
P07498/rs1048152	Kappa-casein	RL110	-	5	3	45
P19823/rs3740217	Inter-alpha-trypsin inhibitor heavy chain H2	PA674	-	6	3	34
P48745/rs2279112	Protein NOV homolog	RQ42	-	5	3	35
P55327/rs35099105	Tumor protein D52	DY52	-	8	3	63
Q08648/rs2853658	Sperm-associated antigen 11B	RQ77	-	5	3	30
Q0VG06/rs11552304	Fanconi anemia-associated protein	PL660	-	5	3	50

	of 100 kDa					
Q13111/rs35651457	Chromatin assembly factor 1 subunit A	DV167	-	5	3	80
Q13111/rs9352	Chromatin assembly factor 1 subunit A	AV923	-	5	3	87
Q14207/rs35095430	Protein NPAT	VA608	-	5	3	194
Q15361/rs1752676	Transcription termination factor 1	AV885	-	5	3	31
Q15572/rs4150167	TATA box-binding protein-associated factor RNA polymerase I subunit C	GR523	-	6	3	39
Q16534/-	Hepatic leukemia factor	IF253	-	5	3	174
Q17RF5/rs2306175	Uncharacterized protein C4orf26	PL30	-	8	3	34
Q3B820/rs17513722	Protein FAM161A	IV236	-	5	3	119
Q3MHD2/rs59168537	Protein LSM12 homolog	VL173	-	5	3	30
Q49AG3/rs2232920	Zinc finger BED domain-containing protein 5	PS77	-	5	3	43
Q4G0U5/rs2272058	Primary ciliary dyskinesia protein 1	VI637	-	5	3	46
Q52M75/rs17366761	Putative uncharacterized protein C5orf27	RC85	-	6	3	55
Q567U6/-	Coiled-coil domain-containing protein 93	HR315	A colorectal cancer sample; somatic mutation	6	3	74
Q569K6/rs12167903	Coiled-coil domain-containing protein 157	PL191	-	5	3	54
Q5FWF5/rs13381941	N-acetyltransferase ESCO1	TM221	-	6	3	262
Q5JSZ5/rs10736851	Protein PRRC2B	ST1630	-	8	3	244
Q5SQ13/rs11787585	Proline-rich protein 31	LF8	-	6	3	116
Q5SZD1/rs9473588	Uncharacterized protein C6orf141	PL235	-	8	3	38
Q5T752/rs41268490	Late cornified envelope protein 1D	RH78	-	7	3	83
Q5TA76/rs16834245	Late cornified envelope protein 3A	RC59	-	6	3	80
Q5TAP6/rs3742289	U3 small nucleolar RNA-associated protein 14 homolog C	GV85	-	5	3	76
Q5TAP6/rs3742290	U3 small nucleolar RNA-associated protein 14 homolog C	TA101	-	5	3	61
Q5VWN6/rs56856085	Protein FAM208B	SY724	-	8	3	158

Q63HN1/rs524512	Protein FAM205B	DE203	-	5	3	76
Q68BL7/rs7874348	Olfactomedin-like protein 2A	TA309	-	5	3	93
Q6L8H2/rs7129002	Keratin-associated protein 5-3	GS27	-	5	3	117
Q6L8H2/rs7108370	Keratin-associated protein 5-3	YC28	-	5	3	117
Q6L8H2/rs7125826	Keratin-associated protein 5-3	GV76	-	5	3	121
Q6L8H2/rs7113784	Keratin-associated protein 5-3	SC83	-	6	3	102
Q6P2C0/rs7163367	WD repeat-containing protein 93	ST254	-	7	3	32
Q6PK04/rs11150805	Coiled-coil domain-containing protein 137	RW177	-	9	3	134
Q6ZVD7/rs41278532	Storkhead-box protein 1	NI825	Pre-eclampsia/eclampsia 4 (PEE4) [MIM:609404]	6	3	78
Q7Z570/rs12476147	Zinc finger protein 804A	QL261	-	5	3	84
Q7Z570/rs12105159	Zinc finger protein 804A	GR1152	-	5	3	69
Q86UC2/rs3756987	Radial spoke head protein 3 homolog	GD518	-	6	3	73
Q86V48/rs12066671	Leucine zipper protein 1	SN1034	-	5	3	78
Q86WS4/rs58302581	Uncharacterized protein C12orf40	IL13	-	5	3	33
Q86X51/rs1875755	Uncharacterized protein CXorf67	RK470	-	6	3	380
Q86YV5/-	Tyrosine-protein kinase SgK223	AT1111	-	6	3	36
Q8IVM0/rs35380043	Coiled-coil domain-containing protein 50	LF121	-	5	3	125
Q8IXS0/rs10485172	Protein FAM217A	MV442	-	5	3	68
Q8IYE0/rs1109968	Coiled-coil domain-containing protein 146	NS345	-	6	3	87
Q8IYI0/rs237422	Uncharacterized protein C20orf196	AV23	-	6	3	45
Q8IZ63/rs3745640	Proline-rich protein 22	PL118	-	5	3	124
Q8N1H7/rs1033734	Protein SIX6OS1	SL309	-	6	3	134
Q8N4Y2/rs4075289	EF-hand calcium-binding domain-containing protein 4A	SI248	-	6	3	112
Q8N6Y0/rs9676419	Usher syndrome type-1C protein-binding protein 1	MV439	-	5	3	110
Q8N715/rs17852896	Coiled-coil domain-containing protein 185	RL380	-	5	3	180

Q8N7X0/rs1052445	Androglobin	TA1637	-	5	3	76
Q8N9H9/rs1281018	Uncharacterized protein C1orf127	AV530	-	6	3	273
Q8N9K7/rs2272624	Uncharacterized protein KIAA1456	QH18	-	6	3	73
Q8NEF3/rs34457718	Coiled-coil domain-containing protein 112	HL32	-	5	3	50
Q8NEM2/rs6598679	SHC SH2 domain-binding protein 1	MT21	-	5	3	48
Q8NEV8/rs3741046	Exophilin-5	RL118	-	5	3	85
Q8NEV8/rs17108127	Exophilin-5	ML512	-	6	3	129
Q8TC99/rs12952106	Fibronectin type III domain-containing protein 8	AT127	-	5	3	64
Q8TD31/rs2073720	Coiled-coil alpha-helical rod protein 1	KR546	-	6	3	148
Q8TD31/rs130079	Coiled-coil alpha-helical rod protein 1	GC575	-	7	3	131
Q8TF40/rs12109782	Folliculin-interacting protein 1	VL738	-	5	3	78
Q8WTT2/rs12572897	Nucleolar complex protein 3 homolog	PL194	-	6	3	98
Q8WXE1/rs35240314	ATR-interacting protein	PL240	-	5	3	124
Q8WYQ9/rs11648852	Zinc finger CCHC domain-containing protein 14	IV54	-	5	3	100
Q92665/rs1854421	28S ribosomal protein S31; mitochondrial	TM241	-	6	3	52
Q969Z0/rs2304693	Protein TBRG4	PL57	-	5	3	33
Q96GE4/rs9910506	Centrosomal protein of 95 kDa	MI165	-	5	3	83
Q96JM3/rs12428067	Chromosome alignment-maintaining phosphoprotein 1	PR604	-	5	3	295
Q96JM3/rs35564629	Chromosome alignment-maintaining phosphoprotein 1	KR591	-	5	3	338
Q96KD3/rs6949056	Protein FAM71F1	SL228	-	7	3	38
Q96LP6/rs7484376	Uncharacterized protein C12orf42	PR182	-	5	3	113
Q96NL8/rs36096184	Protein C8orf37	PA19	-	6	3	68
Q96PI1/rs16834786	Small proline-rich protein 4	PS45	-	6	3	73
Q9BW71/rs11643314	HIRA-interacting protein 3	GW521	-	10	3	54

Q9BWW9/rs2076672	Apolipoprotein L5	TM323	-	6	3	65
Q9H0A9/rs884134	Speriolin-like protein	PL113	-	8	3	105
Q9H0B3/rs12462974	Uncharacterized protein KIAA1683	TA524	-	5	3	225
Q9H0B3/rs2277921	Uncharacterized protein KIAA1683	PL835	-	6	3	91
Q9H4K1/rs2142661	RIB43A-like with coiled-coils protein 2	RC180	-	6	3	67
Q9H501/rs34414644	ESF1 homolog	IL824	-	7	3	85
Q9H8E8/rs6081011	Cysteine-rich protein 2-binding protein	PL214	-	7	3	79
Q9H9L4/rs3741628	KAT8 regulatory NSL complex subunit 2	PT445	-	5	3	98
Q9HAW4/rs34390044	Claspin	PT892	-	6	3	150
Q9HBH7/rs709036	Protein BEX1	AV40	-	7	3	69
Q9NSI2/rs3737075	Protein FAM207A	VL212	-	5	3	77
Q9NVL1/rs57679800	Protein FAM86C1	PL135	-	5	3	34
Q9NXF7/rs34085539	DDB1- and CUL4-associated factor 16	NS45	-	7	3	43
Q9NYF0/-	Dapper homolog 1	SL682	A colorectal cancer sample; somatic mutation	6	3	305
Q9NZM5/rs1804994	Glioma tumor suppressor candidate region gene 2 protein	QR389	-	5	3	144
Q9P0W8/rs17124677	Spermatogenesis-associated protein 7	GE324	-	5	3	46
Q9UHV2/rs268687	SERTA domain-containing protein 1	TA31	-	5	3	30
Q9UHY8/rs1544655	Fasciculation and elongation protein zeta-2	PL50	-	6	3	58
Q9UJX5/rs11550697	Anaphase-promoting complex subunit 4	EG800	-	6	3	43
Q9Y238/rs9840172	Deleted in lung and esophageal cancer protein 1	ND1150	-	6	3	35
Q9Y2X0/rs34859566	Mediator of RNA polymerase II transcription subunit 16	LF770	-	5	3	33
Q9Y5P3/rs6527818	Retinoic acid-induced protein 2	MV252	-	5	3	108

Q9Y6X0/rs77518617	SET-binding protein	VL1377	-	5	3	204

[a]Columns 1 and 2 describe the protein accession numbers in the UniProt database/dbSNP database and protein names, respectively. Column 3 indicates the D→O mutations, which can be described as *XY?*, where *X* is the wild-type residue, *Y* is the mutated residue and *?* is the position. The disease annotations of mutations are shown in column 4. Column 5 and 6 list the numbers of predictors that agree the mutations to located in disordered (column 5)/ordered (column 6) regions. The last column shows the lengths of predicted LDRs are the averaged length of predicted disordered regions from all the 10 predictors.

[b]Number of predictors that agree that the wild-type residues are located in disordered region.

[c]Number of predictors that agree that the mutations are located in ordered region.

[d]Averaged length of disordered regions from different predictors.

5.3.5 D→O mutations located in experimentally verified disordered regions

As mentioned in 'Materials and methods', DisProt is a database providing experimentally verified disordered regions of wild-type proteins. Therefore, to take advantage of these annotations, the human disease mutations and polymorphisms dataset has been mapped to DisProt to narrow down the list of mutations that are located in the experimentally verified disordered regions. In other words, the disordered region prediction results from D2P2 were not used any more. The disordered region annotations were acquired from DisProt database directly. Then 4 protein disordered region predictors were used to predict the structural changes using majority voting. Tango was constantly used to guarantee that the mutations do not increase the protein aggregation propensity. The list of candidates of D→O disease causing mutations and polymorphisms in LDRs are then shown in Table 5.4. The detailed sequence functional and contextual annotations for proteins in Table 5.4 including modified residues, protein superfamily domains and Pfam domains are listed in Table S5.10.

Table 5.4 List of candidates of D→O disease causing mutations and polymorphisms located in experimentally verified LDRs

UniProt/dbSNP	Protein	DR[b]	Mutation[b]	Disease/Phenotype	#OR predictors[d]
P38936/ rs4986867	Cyclin-dependent kinase inhibitor 1	1-164	FL63	Polymorphism	4
P35869/ rs2066853	Aryl hydrocarbon receptor	545-713	RK554	Polymorphism	4
P35869/ rs4986826		545-713	VI570	Polymorphism	4
P04234/ rs45510201	T-cell surface glycoprotein CD3 delta chain	127-171	QR147	Polymorphism	4
P38398/-	Breast cancer type 1 susceptibility protein	170-1649	PL798	In breast cancer; unknown pathological significance; functionally neutral in vitro. [MIM:	4

				114480]	
P38398/-		170-1649	NY810	In breast cancer; unknown pathological significance; functionally neutral in vitro. [MIM: 114480]	4
P38936/rs1801270	Cyclin-dependent kinase inhibitor 1	1-164	SR31	Polymorphism	3
P13569/rs1800103	Cystic fibrosis transmembrane conductance regulator	708-832	IM807	CBAVD (Congenital bilateral absence of the vas deferens) [MIM: 277180]	3
P04150/-	Glucocorticoid receptor	1-500	FL29	Polymorphism	3
P38398/rs56046357	Breast cancer type 1 susceptibility protein	170-1649	FL461	Breast cancer [MIM: 114480]	4
P38398/-		170-1649	GD960	Breast cancer [MIM: 114480]	3
P38398/-		170-1649	FL1226	BROVCA1 (Breast-ovarian cancer, familial, 1) [MIM: 604370]	3
P38398/-		170-1649	GC778	In a breast cancer sample; somatic mutation	3
P38398/-		170-1649	RW170	In breast cancer; unknown pathological significance; functionally neutral in vitro. [MIM: 114480]	3
P38398/-		170-1649	SY186	In breast cancer; unknown pathological significance; functionally neutral in vitro. [MIM: 114480]	3
P38398/-		170-1649	RC866	Polymorphism	3
P01106/rs4645959	Myc proto-oncogene protein	1-88	NS11	Polymorphism	3
Q9NR00/rs6474226	Uncharacterized protein C8orf4	1-106	VI10	Polymorphism	3
P30291/rs34412975	Wee1-like protein kinase	1-292	GC210	Polymorphism	3
Q13569/rs2888805	G/T mismatch-specific thymine DNA glycosylase	340-410	VM367	Polymorphism	3
P49918/-	Cyclin-dependent kinase inhibitor 1C	1-316	FV276	IMAGE (Intrauterine growth retardation, metaphyseal dysplasia, adrenal hypoplasia congenita, and	3

				genital anomalies) [MIM: 614732]	

[a]Mutations located by mapping the protein sequences extracted from the UniProt database to the DisProt database, which contains experimentally verified disordered regions.

[b]OR = Disordered region

[c]Mutations in same format as Table 5.3.

5.4 Conclusion

In this chapter, based on the computational prediction experimental evidence for protein disordered regions, four structural transitions between disordered regions and ordered regions, namely D→O (Disorder-to-Order), O→D (Order-to-Disorder), O→O (Order-to-Order) and D→D (Disorder-to-Disorder) have been investigated. Based on the four different transitions, hydrophobicity changes upon mutations, mutation function and the distribution of protein ELM have been analysed. It is important to note that all the analyses here were based on current dataset and annotations. The annotations from UniProt database update frequently therefore the analyses performed in this chapter expect to vary. According to the proposed 'structural capacitance' hypothesis, the mutations located in the LDRs are the 'structural capacitance elements'. In light of this, the resulting two tables, with experimental candidates causing D→O structural transition, have been provided for future laboratory investigation and validation.

Chapter 6 Discussion

The motivation of this thesis is to provide useful knowledge-bases and conduct insightful data analyses for protein structural and sequence features related to protein function and human disease. From a bioinformatics perspective, two biological databases have been constructed. Computational analyses and evaluation of computational approaches that are developed based on a variety of protein sequence and structural features have also been conducted.

With respect to protein sequence features that are strongly related to protein function and disease, two comprehensive knowledge-bases have been implemented, namely PolyQ 2.0 and KineotchoreDB, by integrating multifaceted protein sequence, structural, functional and disease-associated annotations. It is envisaged that these two databases will greatly facilitate in-depth functional studies of polyQ repeat-containing proteins and kinetochore related proteins that are associated with disease.

With respect to protein structural features, protein coiled-coil domains and disordered regions have been investigated in this thesis, which have important implications for protein function. For protein coiled-coil domain, this thesis focused on examining and benchmarking the state-of-art predictors of coiled-coil domains and oligomeric states, in order to provide insightful performance evaluation and existing challenges for future development of improved approaches. In addition, a case study of nine human pathogenic polyQ proteins has been performed for showcasing the performance of state-of-the-art CCD and its oligomeric state predictors using a complementary independent test. For the protein disordered regions, structural changes induced by single point disease-associated mutations and polymorphisms has been interrogated using a variety of data resources and algorithms for predicting protein disordered regions. Based on the results, we proposed a new mechanism, termed “structural capacitance”, to explain the underlying mechanism of disorder to order transitions that lead to human diseases. The experimental candidates are

also provided for the purpose of validating this hypothesis after a careful examination of all candidates.

This chapter will briefly discuss the major findings and highlight future directions for each topic covered in this thesis.

6.1 Database for human PolyQ proteins

Based on the previously published polyQ database, in this thesis, an updated database, PolyQ 2.0, has been implemented by enriching the data entries with protein structural and functional annotations, polyQ domain context information, protein signaling/metabolic pathway and multiple sequence alignment. A number of different types of annotation have been extracted and collected from multiple publicly available databases and have been carefully reviewed and collated before being made publicly available in the PolyQ 2.0 database.

In the future, more experimental results regarding structural and functional features of human polyQ proteins will be added to the PolyQ 2.0 database. With the advances of text mining and information retrieval techniques (such as Boolean model, Vector space model, semantic network, query expansion and etc.), massive online text extraction from the literature is possible. The auto extraction of human polyQ protein structural and functional annotations from the literature will be set up in the PolyQ 2.0 database, thereby providing up-to-date, cutting-edge experimental results relevant to structure, function and disease-associations of human polyQ proteins.

6.2 KinetochoreBD for kineotchore and its related proteins

KinetochoreDB, a biological database, for the kinetochore and its related proteins has been implemented. This comprehensive database focuses on detailed annotations for

proteins that have been experimentally verified to locate in and/or be functionally related to the kinetochore.

In the future, attempts to improve and update the annotations and analysis of data entries in KinetochoreDB will be made using the following approaches: (1) the database will be kept updated and the up-to-date information to reflect the progress of research on the kinetochore and its related proteins will be continually added; (2) genomic information from publicly available information or bioinformatics programs will be integrated into the database. These include coding sequence, transcription factor binding sites (TFBS), enhancers, promoters and other upstream or downstream regulatory information; (3) other state-of-the-art predictors will be combined to annotate the natively disordered regions of all entries in the database, with the consensus used as the final prediction. Meanwhile, experimentally verified disordered regions will also be collected from DisProt [239], a reliable resource for disordered region annotations in proteins; (4) Experimental biologists are encouraged to contribute to the development of KinetochoreDB by submitting their recent findings, which will be made available in the database after careful review. In addition, annotations and analysis of all entries in KinetochoreDB will be continually updated by implementing secondary analysis functions of the database, by integrating high-throughput experimental data.

6.3 Protein CCDs and their oligomeric state prediction

CCDs are a special type of protein tertiary structure that have been revealed to be related to disease and, in addition, are useful as drug delivery systems. From a bioinformatics perspective, Chapter 3 provided a comprehensive performance evaluation using current computational approaches for two related prediction tasks in structural bioinformatics, *i.e.*, coiled-coil domain prediction and coiled-coil oligomeric state prediction. This critical evaluation serves as a useful guide for researchers in the

community who would like to gain a better understanding of state-of-the-art computational approaches in this area and aim to develop their own methods with improved prediction performance.

Prediction performance of all currently available and executable algorithms and tools has been systematically assessed for the two prediction tasks by benchmarking them against rigorously prepared independent test datasets. The results highlight that the Multicoil2 and CCHMM_PROF algorithms achieved the overall best AUC values for coiled-coil domain prediction, while the LOGICOIL algorithm achieved the best performance for coiled-coil oligomeric state prediction.

In addition, an ensemble web server, namely Waggawagga [240], containing several CCD and oligomeric state predictors has been developed to enable the comparison of prediction results from different predictors. The user-friendly interface allows fast and straightforward comparisons of different prediction results.

A common problem during the training process of CCD oligomeric state predictors is the limited size of the training set. Compared to dimeric CCDs, the number of samples of other oligomeric states, including trimer and tetramer, is relatively small, which means the training datasets are not large enough for constructing reliable predictors. To address this problem, in the future, it is suggested that techniques used for mining skewed datasets can be employed. Oversampling and undersampling techniques have been proven to be effective when mining imbalanced datasets. These techniques are easy to employ during the data pre-process stage, before training computational models.

PolyQ domain-containing proteins have important biological functions and play a role in neurodegenerative diseases. Independent tests with nine human disease-associated polyQ proteins as a complementary independent test to evaluate the performance of currently available algorithms for predicting coiled-coiled regions and their oligomeric

states have also been performed. The CCD prediction results for nine polyQ proteins show inconsistencies, which should be borne in mind when using prediction methods to make meaningful and reliable biological inferences.

6.4 Structural capacitance in human diseases

Another biologically important protein structural feature, the protein disordered region, has been investigated in my thesis. Chapter 5 performed detailed analyses on protein disordered region prediction and the structural changes introduced by mutations, using a variety of computational approaches and data resources for protein disordered regions. Consequently, four structural changes between disordered and ordered state, including D→O, O→D, O→O and D→D, have been defined according to the computational prediction results. Based on the four transitions, I further investigated the hydrophobicity change upon mutations and motif conservation. In addition, I also conducted an analysis for residue functional annotations.

The data analysis in this chapter led to the proposal of a new hypothesis referred to as 'structural capacitance'. Traditional loss-of-function has been well established to illustrate the relationship between protein function disruption and human diseases. However, gain-of-function, on the other hand, has not been well recognized. Via a new generation of microstructure caused by mutations resulting in D→O structural change, it is possible that novel functions may be acquired that will lead to human diseases. Based on my computational work in this thesis, I have provided two tables with experimental candidates that cause D→O transition. The candidates listed in Table 5.2, were extracted purely from protein disordered regions prediction using majority voting for the consensus decision. However, as rigorous standards were used to ensure the quality of the selected candidates, the number of disease-associated mutations causing D→O transition is relatively small. In addition, the disease mutations dataset has been mapped to the DisProt

database, where experimentally verified disordered regions can be found. As a result, another table (Table 5.3) identifies several D→O causing disease-associated mutations. In order to test this hypothesis, the Buckle laboratory are currently using circular dichroism (CD) spectroscopy to detect structural changes between the wild-type and mutated motif (within 30-mer peptides with the mutation in the central position).

Given the nature of this work, there are three independent problems to be addressed: (1) to keep the analysis up-to-date; (2) to augment the current candidate table and (3) to obtain more somatic mutations. A challenge for this data analysis is that the protein sequences and related functional annotations are updated frequently. This makes it difficult to ensure that the analyses are performed with the 'most updated' sequences and related functional annotations. Therefore it is strongly suggested that an automatic and systematic framework should be constructed to facilitate the real time analyses. Once the sequences or functional annotations are updated by other databases (*e.g.* the UniProt database), this framework will be able to immediately update the analyses accordingly.

A suggested alternative way to augment the current candidate list is to conduct protein disorder prediction on significantly larger datasets. GWAS (Genome-Wide Association Studies) project (http://jjwanglab.org/gwasdb) provides a huge number of genetic variations. By translating genome sequences to protein sequences, it is possible to find more mutations and their disease-associations.

Given that the annotated mutation types of the mutations (i.e. germline *vs* somatic) in the human disease-associated mutation and polymorphism database are incomplete, it is suggested that other databases containing somatic mutations, for example, the COSMIC database [241], which contains somatic mutations in cancer, could be a good resource for investigation of structural changes upon mutation. Following the procedure deployed in

this thesis, it is believed that more testable D→O candidates will be found that can be used to validate the 'structural capacitance' hypothesis in the lab.

References

1. Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, Oh JS, Oldfield CJ, Campen AM, Ratliff CM, Hipps KW *et al*: **Intrinsically disordered protein**. *Journal of molecular graphics & modelling* 2001, **19**(1):26-59.
2. Pauling L, Corey RB, Branson HR: **The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain**. *Proceedings of the National Academy of Sciences of the United States of America* 1951, **37**(4):205-211.
3. Lupas A, Van Dyke M, Stock J: **Predicting coiled coils from protein sequences**. *Science* 1991, **252**(5009):1162-1164.
4. Janin J, Bahadur RP, Chakrabarti P: **Protein-protein interaction and quaternary structure**. *Quarterly reviews of biophysics* 2008, **41**(2):133-180.
5. Henrick K, Thornton JM: **PQS: a protein quaternary structure file server**. *Trends in biochemical sciences* 1998, **23**(9):358-361.
6. Roversi P, Johnson S, Caesar JJ, McLean F, Leath KJ, Tsiftsoglou SA, Morgan BP, Harris CL, Sim RB, Lea SM: **Structural basis for complement factor I control and its disease-associated sequence polymorphisms**. *Proceedings of the National Academy of Sciences of the United States of America* 2011, **108**(31):12839-12844.
7. Horton JR, Sawada K, Nishibori M, Zhang X, Cheng X: **Two polymorphic forms of human histamine methyltransferase: structural, thermal, and kinetic comparisons**. *Structure* 2001, **9**(9):837-849.
8. Hazes B, Sastry PA, Hayakawa K, Read RJ, Irvin RT: **Crystal structure of Pseudomonas aeruginosa PAK pilin suggests a main-chain-dominated mode of receptor binding**. *Journal of molecular biology* 2000, **299**(4):1005-1017.
9. Bakolitsa C, Schwarzenbacher R, McMullan D, Brinen LS, Canaves JM, Dai X, Deacon AM, Elsliger MA, Eshagi S, Floyd R *et al*: **Crystal structure of an orphan protein (TM0875) from Thermotoga maritima at 2.00-A resolution reveals a new fold**. *Proteins* 2004, **56**(3):607-610.
10. Ozbek S, Engel J, Stetefeld J: **Storage function of cartilage oligomeric matrix protein: the crystal structure of the coiled-coil domain in complex with vitamin D(3)**. *The EMBO journal* 2002, **21**(22):5960-5968.
11. Nakamura T, Mine S, Hagihara Y, Ishikawa K, Ikegami T, Uegaki K: **Tertiary structure and carbohydrate recognition by the chitin-binding domain of a hyperthermophilic chitinase from Pyrococcus furiosus**. *Journal of molecular biology* 2008, **381**(3):670-680.
12. Anfinsen CB: **Principles that govern the folding of protein chains**. *Science* 1973, **181**(4096):223-230.
13. Vendruscolo M, Zurdo J, MacPhee CE, Dobson CM: **Protein folding and misfolding: a paradigm of self-assembly and regulation in complex biological systems**. *Philosophical transactions Series A, Mathematical, physical, and engineering sciences* 2003, **361**(1807):1205-1222.
14. Wolynes PG, Onuchic JN, Thirumalai D: **Navigating the folding routes**. *Science* 1995, **267**(5204):1619-1620.
15. Zwanzig R, Szabo A, Bagchi B: **Levinthal's paradox**. *Proceedings of the National Academy of Sciences of the United States of America* 1992, **89**(1):20-22.
16. Calosci N, Chi CN, Richter B, Camilloni C, Engstrom A, Eklund L, Travaglini-Allocatelli C, Gianni S, Vendruscolo M, Jemth P: **Comparison of successive**

transition states for folding reveals alternative early folding pathways of two homologous proteins. *Proceedings of the National Academy of Sciences of the United States of America* 2008, **105**(49):19241-19246.
17. Karplus M: **Behind the folding funnel diagram**. *Nature chemical biology* 2011, **7**(7):401-404.
18. Mollapour M, Neckers L: **Post-translational modifications of Hsp90 and their contributions to chaperone regulation**. *Biochimica et biophysica acta* 2012, **1823**(3):648-655.
19. Romero P, Obradovic Z, Li X, Garner EC, Brown CJ, Dunker AK: **Sequence complexity of disordered protein**. *Proteins* 2001, **42**(1):38-48.
20. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT: **Prediction and functional analysis of native disorder in proteins from the three kingdoms of life**. *Journal of molecular biology* 2004, **337**(3):635-645.
21. Midic U, Oldfield CJ, Dunker AK, Obradovic Z, Uversky VN: **Protein disorder in the human diseasome: unfoldomics of human genetic diseases**. *BMC genomics* 2009, **10 Suppl 1**:S12.
22. Uversky VN, Oldfield CJ, Dunker AK: **Intrinsically disordered proteins in human diseases: introducing the D2 concept**. *Annual review of biophysics* 2008, **37**:215-246.
23. Vacic V, Iakoucheva LM: **Disease mutations in disordered regions--exception to the rule?** *Molecular bioSystems* 2012, **8**(1):27-32.
24. Xue B, Dunker AK, Uversky VN: **Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life**. *Journal of biomolecular structure & dynamics* 2012, **30**(2):137-149.
25. Tompa P: **Intrinsically disordered proteins: a 10-year recap**. *Trends in biochemical sciences* 2012, **37**(12):509-516.
26. Lobley A, Swindells MB, Orengo CA, Jones DT: **Inferring function using patterns of native disorder in proteins**. *PLoS computational biology* 2007, **3**(8):e162.
27. He B, Wang K, Liu Y, Xue B, Uversky VN, Dunker AK: **Predicting intrinsic disorder in proteins: an overview**. *Cell research* 2009, **19**(8):929-949.
28. Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradovic Z: **Intrinsic disorder and protein function**. *Biochemistry* 2002, **41**(21):6573-6582.
29. Kurotani A, Tokmakov AA, Kuroda Y, Fukami Y, Shinozaki K, Sakurai T: **Correlations between predicted protein disorder and post-translational modifications in plants**. *Bioinformatics* 2014.
30. Wright PE, Dyson HJ: **Intrinsically disordered proteins in cellular signalling and regulation**. *Nature reviews Molecular cell biology* 2015, **16**(1):18-29.
31. Iakoucheva LM, Radivojac P, Brown CJ, O'Connor TR, Sikes JG, Obradovic Z, Dunker AK: **The importance of intrinsic disorder for protein phosphorylation**. *Nucleic acids research* 2004, **32**(3):1037-1049.
32. Radivojac P, Vacic V, Haynes C, Cocklin RR, Mohan A, Heyen JW, Goebl MG, Iakoucheva LM: **Identification, analysis, and prediction of protein ubiquitination sites**. *Proteins* 2010, **78**(2):365-380.
33. Goh GK, Dunker AK, Uversky VN: **Protein intrinsic disorder and influenza virulence: the 1918 H1N1 and H5N1 viruses**. *Virology journal* 2009, **6**:69.
34. Dunker AK, Oldfield CJ, Meng J, Romero P, Yang JY, Chen JW, Vacic V, Obradovic Z, Uversky VN: **The unfoldomics decade: an update on intrinsically disordered proteins**. *BMC genomics* 2008, **9 Suppl 2**:S1.

35. Uversky VN, Oldfield CJ, Midic U, Xie H, Xue B, Vucetic S, Iakoucheva LM, Obradovic Z, Dunker AK: **Unfoldomics of human diseases: linking protein intrinsic disorder with diseases**. *BMC genomics* 2009, **10 Suppl 1**:S7.
36. Vacic V, Markwick PR, Oldfield CJ, Zhao X, Haynes C, Uversky VN, Iakoucheva LM: **Disease-associated mutations disrupt functionally important regions of intrinsic protein disorder**. *PLoS computational biology* 2012, **8**(10):e1002709.
37. Cooper DN, Ball EV, Krawczak M: **The human gene mutation database**. *Nucleic acids research* 1998, **26**(1):285-287.
38. Yip YL, Famiglietti M, Gos A, Duek PD, David FP, Gateau A, Bairoch A: **Annotating single amino acid polymorphisms in the UniProt/Swiss-Prot knowledgebase**. *Human mutation* 2008, **29**(3):361-366.
39. UniProt C: **UniProt: a hub for protein information**. *Nucleic acids research* 2015, **43**(Database issue):D204-212.
40. Reumers J, Schymkowitz J, Ferkinghoff-Borg J, Stricher F, Serrano L, Rousseau F: **SNPeffect: a database mapping molecular phenotypic effects of human non-synonymous coding SNPs**. *Nucleic acids research* 2005, **33**(Database issue):D527-532.
41. Fernandez-Escamilla AM, Rousseau F, Schymkowitz J, Serrano L: **Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins**. *Nature biotechnology* 2004, **22**(10):1302-1306.
42. Oliveberg M: **Waltz, an exciting new move in amyloid prediction**. *Nature methods* 2010, **7**(3):187-188.
43. Van Durme J, Maurer-Stroh S, Gallardo R, Wilkinson H, Rousseau F, Schymkowitz J: **Accurate prediction of DnaK-peptide binding via homology modelling and experimental data**. *PLoS computational biology* 2009, **5**(8):e1000475.
44. Guerois R, Nielsen JE, Serrano L: **Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations**. *Journal of molecular biology* 2002, **320**(2):369-387.
45. Luu TD, Rusu AM, Walter V, Ripp R, Moulinier L, Muller J, Toursel T, Thompson JD, Poch O, Nguyen H: **MSV3d: database of human MisSense Variants mapped to 3D protein structure**. *Database : the journal of biological databases and curation* 2012, **2012**:bas018.
46. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Edgar R, Federhen S *et al*: **Database resources of the National Center for Biotechnology Information**. *Nucleic acids research* 2008, **36**(Database issue):D13-21.
47. Dunker AK, Cortese MS, Romero P, Iakoucheva LM, Uversky VN: **Flexible nets. The roles of intrinsic disorder in protein interaction networks**. *The FEBS journal* 2005, **272**(20):5129-5148.
48. Kim PM, Lu LJ, Xia Y, Gerstein MB: **Relating three-dimensional structures to protein networks provides evolutionary insights**. *Science* 2006, **314**(5807):1938-1941.
49. Higurashi M, Ishida T, Kinoshita K: **Identification of transient hub proteins and the possible structural basis for their multiple interactions**. *Protein science : a publication of the Protein Society* 2008, **17**(1):72-78.
50. Fong JH, Panchenko AR: **Intrinsic disorder and protein multibinding in domain, terminal, and linker regions**. *Molecular bioSystems* 2010, **6**(10):1821-1828.

51. Ramirez J, Recht R, Charbonnier S, Ennifar E, Atkinson RA, Trave G, Nomine Y, Kieffer B: **Disorder-to-order transition of MAGI-1 PDZ1 C-terminal extension upon peptide binding: thermodynamic and dynamic insights**. *Biochemistry* 2015, **54**(6):1327-1337.
52. Zhang Y, Tan H, Chen G, Jia Z: **Investigating the disorder-order transition of calmodulin binding domain upon binding calmodulin using molecular dynamics simulation**. *Journal of molecular recognition : JMR* 2010, **23**(4):360-368.
53. Hyre DE, Klevit RE: **A disorder-to-order transition coupled to DNA binding in the essential zinc-finger DNA-binding domain of yeast ADR1**. *Journal of molecular biology* 1998, **279**(4):929-943.
54. Lobanov MY, Shoemaker BA, Garbuzynskiy SO, Fong JH, Panchenko AR, Galzitskaya OV: **ComSin: database of protein structures in bound (complex) and unbound (single) states in relation to their intrinsic disorder**. *Nucleic acids research* 2010, **38**(Database issue):D283-287.
55. Dosztanyi Z, Meszaros B, Simon I: **ANCHOR: web server for predicting protein binding regions in disordered proteins**. *Bioinformatics* 2009, **25**(20):2745-2746.
56. Dosztanyi Z, Csizmok V, Tompa P, Simon I: **IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content**. *Bioinformatics* 2005, **21**(16):3433-3434.
57. Disfani FM, Hsu WL, Mizianty MJ, Oldfield CJ, Xue B, Dunker AK, Uversky VN, Kurgan L: **MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins**. *Bioinformatics* 2012, **28**(12):i75-83.
58. Dosztanyi Z, Csizmok V, Tompa P, Simon I: **The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins**. *Journal of molecular biology* 2005, **347**(4):827-839.
59. Rose PW, Beran B, Bi C, Bluhm WF, Dimitropoulos D, Goodsell DS, Prlic A, Quesada M, Quinn GB, Westbrook JD *et al*: **The RCSB Protein Data Bank: redesigned web site and web services**. *Nucleic acids research* 2011, **39**(Database issue):D392-401.
60. Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z: **Length-dependent prediction of protein intrinsic disorder**. *BMC bioinformatics* 2006, **7**:208.
61. Ghalwash MF, Dunker AK, Obradovic Z: **Uncertainty analysis in protein disorder prediction**. *Molecular bioSystems* 2012, **8**(1):381-391.
62. Cilia E, Pancsa R, Tompa P, Lenaerts T, Vranken WF: **From protein sequence to dynamics and disorder with DynaMine**. *Nature communications* 2013, **4**:2741.
63. Walsh I, Martin AJ, Di Domenico T, Tosatto SC: **ESpritz: accurate and fast prediction of protein disorder**. *Bioinformatics* 2012, **28**(4):503-509.
64. Yang ZR, Thomson R, McNeil P, Esnouf RM: **RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins**. *Bioinformatics* 2005, **21**(16):3369-3376.
65. Sickmeier M, Hamilton JA, LeGall T, Vacic V, Cortese MS, Tantos A, Szabo B, Tompa P, Chen J, Uversky VN *et al*: **DisProt: the Database of Disordered Proteins**. *Nucleic acids research* 2007, **35**(Database issue):D786-793.
66. Fukuchi S, Amemiya T, Sakamoto S, Nobe Y, Hosoda K, Kado Y, Murakami SD, Koike R, Hiroaki H, Ota M: **IDEAL in 2014 illustrates interaction networks**

composed of intrinsically disordered proteins and their binding partners. *Nucleic acids research* 2014, **42**(Database issue):D320-325.

67. Fukuchi S, Sakamoto S, Nobe Y, Murakami SD, Amemiya T, Hosoda K, Koike R, Hiroaki H, Ota M: **IDEAL: Intrinsically Disordered proteins with Extensive Annotations and Literature**. *Nucleic acids research* 2012, **40**(Database issue):D507-511.
68. Oates ME, Romero P, Ishida T, Ghalwash M, Mizianty MJ, Xue B, Dosztanyi Z, Uversky VN, Obradovic Z, Kurgan L *et al*: **D(2)P(2): database of disordered protein predictions**. *Nucleic acids research* 2013, **41**(Database issue):D508-516.
69. Li X, Romero P, Rani M, Dunker AK, Obradovic Z: **Predicting Protein Disorder for N-, C-, and Internal Regions**. *Genome informatics Workshop on Genome Informatics* 1999, **10**:30-40.
70. Mason JM, Arndt KM: **Coiled coil domains: stability, specificity, and biological implications**. *Chembiochem : a European journal of chemical biology* 2004, **5**(2):170-176.
71. Trigg J, Gutwin K, Keating AE, Berger B: **Multicoil2: predicting coiled coils and their oligomerization states from sequence in the twilight zone**. *PloS one* 2011, **6**(8):e23519.
72. Grigoryan G, Keating AE: **Structural specificity in coiled-coil interactions**. *Current opinion in structural biology* 2008, **18**(4):477-483.
73. Arndt KM, Pelletier JN, Muller KM, Pluckthun A, Alber T: **Comparison of in vivo selection and rational design of heterodimeric coiled coils**. *Structure* 2002, **10**(9):1235-1248.
74. Gillingham AK, Munro S: **Long coiled-coil proteins and membrane traffic**. *Biochim Biophys Acta* 2003, **1641**(2-3):71-85.
75. Rose A, Schraegle SJ, Stahlberg EA, Meier I: **Coiled-coil protein composition of 22 proteomes--differences and common themes in subcellular infrastructure and traffic control**. *BMC Evol Biol* 2005, **5**:66.
76. Burkhard P, Stetefeld J, Strelkov SV: **Coiled coils: a highly versatile protein folding motif**. *Trends in cell biology* 2001, **11**(2):82-88.
77. Kuhn M, Hyman AA, Beyer A: **Coiled-coil proteins facilitated the functional expansion of the centrosome**. *PLoS computational biology* 2014, **10**(6):e1003657.
78. Magin TM, Reichelt J, Hatzfeld M: **Emerging functions: diseases and animal models reshape our view of the cytoskeleton**. *Experimental cell research* 2004, **301**(1):91-102.
79. Mounkes L, Kozlov S, Burke B, Stewart CL: **The laminopathies: nuclear structure meets disease**. *Current opinion in genetics & development* 2003, **13**(3):223-230.
80. Puls I, Jonnakuty C, LaMonte BH, Holzbaur EL, Tokito M, Mann E, Floeter MK, Bidus K, Drayna D, Oh SJ *et al*: **Mutant dynactin in motor neuron disease**. *Nature genetics* 2003, **33**(4):455-456.
81. Hirokawa N, Takemura R: **Molecular motors in neuronal development, intracellular transport and diseases**. *Current opinion in neurobiology* 2004, **14**(5):564-573.
82. Chigira S, Sugita K, Kita K, Sugaya S, Arase Y, Ichinose M, Shirasawa H, Suzuki N: **Increased expression of the Huntingtin interacting protein-1 gene in cells from Hutchinson Gilford Syndrome (Progeria) patients and aged donors**. *The journals of gerontology Series A, Biological sciences and medical sciences* 2003, **58**(10):B873-878.

83. Mounkes LC, Stewart CL: **Aging and nuclear organization: lamins and progeria**. *Current opinion in cell biology* 2004, **16**(3):322-327.
84. Raff JW: **Centrosomes and cancer: lessons from a TACC**. *Trends in cell biology* 2002, **12**(5):222-225.
85. McClatchey AI: **Merlin and ERM proteins: unappreciated roles in cancer development?** *Nature reviews Cancer* 2003, **3**(11):877-883.
86. Lo Giudice M, Neri M, Falco M, Sturnio M, Calzolari E, Di Benedetto D, Fichera M: **A missense mutation in the coiled-coil domain of the KIF5A gene and late-onset hereditary spastic paraplegia**. *Archives of neurology* 2006, **63**(2):284-287.
87. Alsaadi MM, Erzurumluoglu AM, Rodriguez S, Guthrie PA, Gaunt TR, Omar HZ, Mubarak M, Alharbi KK, Al-Rikabi AC, Day IN: **Nonsense mutation in coiled-coil domain containing 151 gene (CCDC151) causes primary ciliary dyskinesia**. *Human mutation* 2014, **35**(12):1446-1448.
88. Walshaw J, Woolfson DN: **Socket: a program for identifying and analysing coiled-coil motifs within protein structures**. *Journal of molecular biology* 2001, **307**(5):1427-1450.
89. Testa OD, Moutevelis E, Woolfson DN: **CC+: a relational database of coiled-coil structures**. *Nucleic acids research* 2009, **37**(Database issue):D315-322.
90. Lupas A: **Predicting coiled-coil regions in proteins**. *Current opinion in structural biology* 1997, **7**(3):388-393.
91. McDonnell AV, Jiang T, Keating AE, Berger B: **Paircoil2: improved prediction of coiled coils from sequence**. *Bioinformatics* 2006, **22**(3):356-358.
92. Berger B, Wilson DB, Wolf E, Tonchev T, Milla M, Kim PS: **Predicting coiled coils by use of pairwise residue correlations**. *Proc Natl Acad Sci U S A* 1995, **92**(18):8259-8263.
93. Bartoli L, Fariselli P, Krogh A, Casadio R: **CCHMM_PROF: a HMM-based coiled-coil predictor with evolutionary information**. *Bioinformatics* 2009, **25**(21):2757-2763.
94. Delorenzi M, Speed T: **An HMM model for coiled-coil domains and a comparison with PSSM-based predictions**. *Bioinformatics* 2002, **18**(4):617-625.
95. Rackham OJ, Madera M, Armstrong CT, Vincent TL, Woolfson DN, Gough J: **The evolution and structure prediction of coiled coils across all genomes**. *Journal of molecular biology* 2010, **403**(3):480-493.
96. Wolf E, Kim PS, Berger B: **MultiCoil: a program for predicting two- and three-stranded coiled coils**. *Protein science : a publication of the Protein Society* 1997, **6**(6):1179-1189.
97. Vincent TL, Green PJ, Woolfson DN: **LOGICOIL--multi-state prediction of coiled-coil oligomeric state**. *Bioinformatics* 2013, **29**(1):69-76.
98. Armstrong CT, Vincent TL, Green PJ, Woolfson DN: **SCORER 2.0: an algorithm for distinguishing parallel dimeric and trimeric coiled-coil sequences**. *Bioinformatics* 2011, **27**(14):1908-1914.
99. Li C, Wang XF, Chen Z, Zhang Z, Song J: **Computational characterization of parallel dimeric and trimeric coiled-coils using effective amino acid indices**. *Molecular bioSystems* 2014.
100. Mahrenholz CC, Abfalter IG, Bodenhofer U, Volkmer R, Hochreiter S: **Complex networks govern coiled-coil oligomerization--predicting and profiling by means of a machine learning approach**. *Mol Cell Proteomics* 2011, **10**(5):M110 004994.
101. Apgar JR, Gutwin KN, Keating AE: **Predicting helix orientation for coiled-coil dimers**. *Proteins* 2008, **72**(3):1048-1065.

102. Woolfson DN: **The design of coiled-coil structures and assemblies**. *Advances in protein chemistry* 2005, **70**:79-112.
103. Fletcher JM, Boyle AL, Bruning M, Bartlett GJ, Vincent TL, Zaccai NR, Armstrong CT, Bromley EH, Booth PJ, Brady RL *et al*: **A basis set of de novo coiled-coil peptide oligomers for rational protein design and synthetic biology**. *ACS synthetic biology* 2012, **1**(6):240-250.
104. Xu C, Liu R, Mehta AK, Guerrero-Ferreira RC, Wright ER, Dunin-Horkawicz S, Morris K, Serpell LC, Zuo X, Wall JS *et al*: **Rational design of helical nanotubes from self-assembly of coiled-coil lock washers**. *Journal of the American Chemical Society* 2013, **135**(41):15565-15578.
105. Potapov V, Kaplan JB, Keating AE: **Data-driven prediction and design of bZIP coiled-coil interactions**. *PLoS computational biology* 2015, **11**(2):e1004046.
106. Wood CW, Bruning M, Ibarra AA, Bartlett GJ, Thomson AR, Sessions RB, Brady RL, Woolfson DN: **CCBuilder: an interactive web-based tool for building, designing and assessing coiled-coil protein assemblies**. *Bioinformatics* 2014, **30**(21):3029-3035.
107. Faux NG, Huttley GA, Mahmood K, Webb GI, de la Banda MG, Whisstock JC: **RCPdb: An evolutionary classification and codon usage database for repeat-containing proteins**. *Genome research* 2007, **17**(7):1118-1127.
108. Willadsen K, Cao MD, Wiles J, Balasubramanian S, Boden M: **Repeat-encoded poly-Q tracts show statistical commonalities across species**. *BMC genomics* 2013, **14**:76.
109. Li H, Liu J, Wu K, Chen Y: **Insight into role of selection in the evolution of polyglutamine tracts in humans**. *PloS one* 2012, **7**(7):e41167.
110. Schaefer MH, Wanker EE, Andrade-Navarro MA: **Evolution and function of CAG/polyglutamine repeats in protein-protein interaction networks**. *Nucleic acids research* 2012, **40**(10):4273-4287.
111. Gemayel R, Chavali S, Pougach K, Legendre M, Zhu B, Boeynaems S, van der Zande E, Gevaert K, Rousseau F, Schymkowitz J *et al*: **Variable Glutamine-Rich Repeats Modulate Transcription Factor Activity**. *Molecular cell* 2015, **59**(4):615-627.
112. Kratter IH, Finkbeiner S: **PolyQ disease: too many Qs, too much function?** *Neuron* 2010, **67**(6):897-899.
113. Khare SD, Ding F, Gwanmesia KN, Dokholyan NV: **Molecular origin of polyglutamine aggregation in neurodegenerative diseases**. *PLoS computational biology* 2005, **1**(3):230-235.
114. Petrakis S, Schaefer MH, Wanker EE, Andrade-Navarro MA: **Aggregation of polyQ-extended proteins is promoted by interaction with their natural coiled-coil partners**. *BioEssays : news and reviews in molecular, cellular and developmental biology* 2013, **35**(6):503-507.
115. Faux NG, Bottomley SP, Lesk AM, Irving JA, Morrison JR, de la Banda MG, Whisstock JC: **Functional insights from the distribution and role of homopeptide repeat-containing proteins**. *Genome research* 2005, **15**(4):537-551.
116. Goto J, Watanabe M, Ichikawa Y, Yee SB, Ihara N, Endo K, Igarashi S, Takiyama Y, Gaspar C, Maciel P *et al*: **Machado-Joseph disease gene products carrying different carboxyl termini**. *Neuroscience research* 1997, **28**(4):373-377.
117. Padiath QS, Srivastava AK, Roy S, Jain S, Brahmachari SK: **Identification of a novel 45 repeat unstable allele associated with a disease phenotype at the MJD1/SCA3 locus**. *American journal of medical genetics Part B,*

Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics 2005, **133B**(1):124-126.
118. Li W, Serpell LC, Carter WJ, Rubinsztein DC, Huntington JA: **Expression and characterization of full-length human huntingtin, an elongated HEAT repeat protein**. *The Journal of biological chemistry* 2006, **281**(23):15916-15922.
119. Shao J, Diamond MI: **Polyglutamine diseases: emerging concepts in pathogenesis and therapy**. *Human molecular genetics* 2007, **16 Spec No. 2**:R115-123.
120. Robertson AL, Bate MA, Androulakis SG, Bottomley SP, Buckle AM: **PolyQ: a database describing the sequence and domain context of polyglutamine repeats in proteins**. *Nucleic acids research* 2011, **39**(Database issue):D272-276.
121. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J *et al*: **Pfam: the protein families database**. *Nucleic acids research* 2014, **42**(Database issue):D222-230.
122. Akiyoshi B, Gull K: **Discovery of unconventional kinetochores in kinetoplastids**. *Cell* 2014, **156**(6):1247-1258.
123. Cleveland DW, Mao Y, Sullivan KF: **Centromeres and kinetochores: from epigenetics to mitotic checkpoint signaling**. *Cell* 2003, **112**(4):407-421.
124. Bakhoum SF, Compton DA: **Kinetochores and disease: keeping microtubule dynamics in check!** *Current opinion in cell biology* 2012, **24**(1):64-70.
125. Bakhoum SF, Genovese G, Compton DA: **Deviant kinetochore microtubule dynamics underlie chromosomal instability**. *Current biology : CB* 2009, **19**(22):1937-1942.
126. Tolmie J, Boyd E, Batstone P, Ferguson-Smith M, Al Roomi L, Connor J: **Siblings with chromosome mosaicism, microcephaly, and growth retardation: the phenotypic expression of a human mitotic mutant?** *Human genetics* 1988, **80**(2):197-200.
127. Scheres J, Hustinx T, Madan K, Beltman J, Lindhout D: **A mitotic mutant causing non-disjunction in man**. In: *7th International Congress of Human Genetics, Berlin: 1986*. 163.
128. Tomonaga T, Matsushita K, Ishibashi M, Nezu M, Shimada H, Ochiai T, Yoda K, Nomura F: **Centromere protein H is up-regulated in primary human colorectal cancer and its overexpression induces aneuploidy**. *Cancer research* 2005, **65**(11):4683-4689.
129. Spruck CH, Won KA, Reed SI: **Deregulated cyclin E induces chromosome instability**. *Nature* 1999, **401**(6750):297-300.
130. Kasiappan R, Shih HJ, Chu KL, Chen WT, Liu HP, Huang SF, Choy CO, Shu CL, Din R, Chu JS *et al*: **Loss of p53 and MCT-1 overexpression synergistically promote chromosome instability and tumorigenicity**. *Molecular cancer research : MCR* 2009, **7**(4):536-548.
131. Huang Z, Ma L, Wang Y, Pan Z, Ren J, Liu Z, Xue Y: **MiCroKiTS 4.0: a database of midbody, centrosome, kinetochore, telomere and spindle**. *Nucleic acids research* 2015, **43**(Database issue):D328-334.
132. Fan HC, Ho LI, Chi CS, Chen SJ, Peng GS, Chan TM, Lin SZ, Harn HJ: **Polyglutamine (PolyQ) diseases: genetics to treatments**. *Cell transplantation* 2014, **23**(4-5):441-458.
133. Perutz MF, Johnson T, Suzuki M, Finch JT: **Glutamine repeats as polar zippers: their possible role in inherited neurodegenerative diseases**. *Proceedings of the National Academy of Sciences of the United States of America* 1994, **91**(12):5355-5358.

134. Chen S, Berthelier V, Hamilton JB, O'Nuallain B, Wetzel R: **Amyloid-like features of polyglutamine aggregates and their assembly kinetics**. *Biochemistry* 2002, **41**(23):7391-7399.
135. Robertson AL, Horne J, Ellisdon AM, Thomas B, Scanlon MJ, Bottomley SP: **The structural impact of a polyglutamine tract is location-dependent**. *Biophysical journal* 2008, **95**(12):5922-5930.
136. Kawaguchi Y, Okamoto T, Taniwaki M, Aizawa M, Inoue M, Katayama S, Kawakami H, Nakamura S, Nishimura M, Akiguchi I *et al*: **CAG expansions in a novel gene for Machado-Joseph disease at chromosome 14q32.1**. *Nature genetics* 1994, **8**(3):221-228.
137. Lin B, Nasir J, MacDonald H, Hutchinson G, Graham RK, Rommens JM, Hayden MR: **Sequence of the murine Huntington disease gene: evidence for conservation, alternate splicing and polymorphism in a triplet (CCG) repeat [corrected]**. *Human molecular genetics* 1994, **3**(1):85-92.
138. Zuhlke C, Hellenbroich Y, Dalski A, Kononowa N, Hagenah J, Vieregge P, Riess O, Klein C, Schwinger E: **Different types of repeat expansion in the TATA-binding protein gene are associated with a new form of inherited ataxia**. *European journal of human genetics : EJHG* 2001, **9**(3):160-164.
139. Nakamura K, Jeong SY, Uchihara T, Anno M, Nagashima K, Nagashima T, Ikeda S, Tsuji S, Kanazawa I: **SCA17, a novel autosomal dominant cerebellar ataxia caused by an expanded polyglutamine in TATA-binding protein**. *Human molecular genetics* 2001, **10**(14):1441-1448.
140. Silveira I, Miranda C, Guimaraes L, Moreira MC, Alonso I, Mendonca P, Ferro A, Pinto-Basto J, Coelho J, Ferreirinha F *et al*: **Trinucleotide repeats in 202 families with ataxia: a small expanded (CAG)n allele at the SCA17 locus**. *Archives of neurology* 2002, **59**(4):623-629.
141. Banfi S, Servadio A, Chung MY, Kwiatkowski TJ, Jr., McCall AE, Duvick LA, Shen Y, Roth EJ, Orr HT, Zoghbi HY: **Identification and characterization of the gene causing type 1 spinocerebellar ataxia**. *Nature genetics* 1994, **7**(4):513-520.
142. Quan F, Janas J, Popovich BW: **A novel CAG repeat configuration in the SCA1 gene: implications for the molecular diagnostics of spinocerebellar ataxia type 1**. *Human molecular genetics* 1995, **4**(12):2411-2413.
143. Stefani M, Dobson CM: **Protein aggregation and aggregate toxicity: new insights into protein folding, misfolding diseases and biological evolution**. *J Mol Med (Berl)* 2003, **81**(11):678-699.
144. Saunders HM, Bottomley SP: **Multi-domain misfolding: understanding the aggregation pathway of polyglutamine proteins**. *Protein engineering, design & selection : PEDS* 2009, **22**(8):447-451.
145. Ellisdon AM, Thomas B, Bottomley SP: **The two-stage pathway of ataxin-3 fibrillogenesis involves a polyglutamine-independent step**. *The Journal of biological chemistry* 2006, **281**(25):16888-16896.
146. DiFiglia M, Sapp E, Chase KO, Davies SW, Bates GP, Vonsattel JP, Aronin N: **Aggregation of huntingtin in neuronal intranuclear inclusions and dystrophic neurites in brain**. *Science* 1997, **277**(5334):1990-1993.
147. Kim MW, Chelliah Y, Kim SW, Otwinowski Z, Bezprozvanny I: **Secondary structure of Huntingtin amino-terminal region**. *Structure* 2009, **17**(9):1205-1212.

148. Robertson AL, Bate MA, Buckle AM, Bottomley SP: **The rate of polyQ-mediated aggregation is dramatically affected by the number and location of surrounding domains**. *Journal of molecular biology* 2011, **413**(4):879-887.
149. **UniProt: a hub for protein information**. *Nucleic acids research* 2015, **43**(Database issue):D204-212.
150. Rose PW, Prlic A, Bi C, Bluhm WF, Christie CH, Dutta S, Green RK, Goodsell DS, Westbrook JD, Woo J *et al*: **The RCSB Protein Data Bank: views of structural biology for basic and applied research and education**. *Nucleic acids research* 2015, **43**(Database issue):D345-356.
151. Chatr-Aryamontri A, Breitkreutz BJ, Oughtred R, Boucher L, Heinicke S, Chen D, Stark C, Breitkreutz A, Kolas N, O'Donnell L *et al*: **The BioGRID interaction database: 2015 update**. *Nucleic acids research* 2015, **43**(Database issue):D470-478.
152. Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M: **Data, information, knowledge and principle: back to metabolism in KEGG**. *Nucleic acids research* 2014, **42**(Database issue):D199-205.
153. Wilson D, Pethica R, Zhou Y, Talbot C, Vogel C, Madera M, Chothia C, Gough J: **SUPERFAMILY--sophisticated comparative genomics, data mining, visualization and phylogeny**. *Nucleic acids research* 2009, **37**(Database issue):D380-386.
154. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic acids research* 1997, **25**(17):3389-3402.
155. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Soding J *et al*: **Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega**. *Molecular systems biology* 2011, **7**:539.
156. Mukhyala K, Masselot A: **Visualization of protein sequence features using JavaScript and SVG with pViz.js**. *Bioinformatics* 2014, **30**(23):3408-3409.
157. Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ: **Jalview Version 2--a multiple sequence alignment editor and analysis workbench**. *Bioinformatics* 2009, **25**(9):1189-1191.
158. Brinkley BR, Tousson A, Valdivia MM: **The kinetochore of mammalian chromosomes: structure and function in normal mitosis and aneuploidy**. *Basic life sciences* 1985, **36**:243-267.
159. Chan GK, Liu ST, Yen TJ: **Kinetochore structure and function**. *Trends in cell biology* 2005, **15**(11):589-598.
160. McAinsh AD, Tytell JD, Sorger PK: **Structure, function, and regulation of budding yeast kinetochores**. *Annual review of cell and developmental biology* 2003, **19**:519-539.
161. Westermann S, Drubin DG, Barnes G: **Structures and functions of yeast kinetochore complexes**. *Annual review of biochemistry* 2007, **76**:563-591.
162. Stankovic A, Jansen LE: **Reductionism at the vertebrate kinetochore**. *The Journal of cell biology* 2013, **200**(1):7-8.
163. Rago F, Cheeseman IM: **Review series: The functions and consequences of force at kinetochores**. *The Journal of cell biology* 2013, **200**(5):557-565.
164. Yang Y, Wu F, Ward T, Yan F, Wu Q, Wang Z, McGlothen T, Peng W, You T, Sun M *et al*: **Phosphorylation of HsMis13 by Aurora B kinase is essential for assembly of functional kinetochore**. *The Journal of biological chemistry* 2008, **283**(39):26726-26736.

165. Wan X, O'Quinn RP, Pierce HL, Joglekar AP, Gall WE, DeLuca JG, Carroll CW, Liu ST, Yen TJ, McEwen BF *et al*: **Protein architecture of the human kinetochore microtubule attachment site**. *Cell* 2009, **137**(4):672-684.
166. Sakuno T, Tada K, Watanabe Y: **Kinetochore geometry defined by cohesion within the centromere**. *Nature* 2009, **458**(7240):852-858.
167. Tanaka TU, Desai A: **Kinetochore-microtubule interactions: the means to the end**. *Current opinion in cell biology* 2008, **20**(1):53-63.
168. Bakhoum SF, Thompson SL, Manning AL, Compton DA: **Genome stability is ensured by temporal control of kinetochore-microtubule dynamics**. *Nature cell biology* 2009, **11**(1):27-35.
169. Huang H, Mahler-Araujo BM, Sankila A, Chimelli L, Yonekawa Y, Kleihues P, Ohgaki H: **APC mutations in sporadic medulloblastomas**. *The American journal of pathology* 2000, **156**(2):433-437.
170. Miyaki M, Nishio J, Konishi M, Kikuchi-Yanoshita R, Tanaka K, Muraoka M, Nagato M, Chong JM, Koike M, Terada T *et al*: **Drastic genetic instability of tumors and normal tissues in Turcot syndrome**. *Oncogene* 1997, **15**(23):2877-2881.
171. Stella A, Montera M, Resta N, Marchese C, Susca F, Gentile M, Romio L, Pilia S, Prete F, Mareni C *et al*: **Four novel mutations of the APC (adenomatous polyposis coli) gene in FAP patients**. *Human molecular genetics* 1994, **3**(9):1687-1688.
172. van der Luijt RB, Khan PM, Vasen HF, Tops CM, van Leeuwen-Cornelisse IS, Wijnen JT, van der Klift HM, Plug RJ, Griffioen G, Fodde R: **Molecular analysis of the APC gene in 105 Dutch kindreds with familial adenomatous polyposis: 67 germline mutations identified by DGGE, PTT, and southern analysis**. *Human mutation* 1997, **9**(1):7-16.
173. Kops GJ, Weaver BA, Cleveland DW: **On the road to cancer: aneuploidy and the mitotic checkpoint**. *Nature reviews Cancer* 2005, **5**(10):773-785.
174. Huntley RP, Sawford T, Mutowo-Meullenet P, Shypitsyna A, Bonilla C, Martin MJ, O'Donovan C: **The GOA database: gene Ontology annotation updates for 2015**. *Nucleic acids research* 2015, **43**(Database issue):D1057-1063.
175. Consortium TU: **UniProt: a hub for protein information**. *Nucleic acids research* 2015, **43**(Database issue):D204-212.
176. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA: **Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders**. *Nucleic acids research* 2005, **33**(Database issue):D514-517.
177. Chatr-Aryamontri A, Breitkreutz BJ, Heinicke S, Boucher L, Winter A, Stark C, Nixon J, Ramage L, Kolas N, O'Donnell L *et al*: **The BioGRID interaction database: 2013 update**. *Nucleic acids research* 2013, **41**(Database issue):D816-823.
178. Lupas A: **Coiled coils: new structures and new functions**. *Trends Biochem Sci* 1996, **21**(10):375-382.
179. McFarlane AA, Orriss GL, Stetefeld J: **The use of coiled-coil proteins in drug delivery systems**. *European journal of pharmacology* 2009, **625**(1-3):101-107.
180. Gromiha MM, Parry DA: **Characteristic features of amino acid residues in coiled-coil protein structures**. *Biophys Chem* 2004, **111**(2):95-103.
181. Tarricone C, Xiao B, Justin N, Walker PA, Rittinger K, Gamblin SJ, Smerdon SJ: **The structural basis of Arfaptin-mediated cross-talk between Rac and Arf signalling pathways**. *Nature* 2001, **411**(6834):215-219.

182. Burkhard P, Kammerer RA, Steinmetz MO, Bourenkov GP, Aebi U: **The coiled-coil trigger site of the rod domain of cortexillin I unveils a distinct network of interhelical and intrahelical salt bridges**. *Structure* 2000, **8**(3):223-230.
183. Bullough PA, Hughson FM, Skehel JJ, Wiley DC: **Structure of influenza haemagglutinin at the pH of membrane fusion**. *Nature* 1994, **371**(6492):37-43.
184. Whitson SR, LeStourgeon WM, Krezel AM: **Solution structure of the symmetric coiled coil tetramer formed by the oligomerization domain of hnRNP C: implications for biological function**. *Journal of molecular biology* 2005, **350**(2):319-337.
185. Guo Y, Bozic D, Malashkevich VN, Kammerer RA, Schulthess T, Engel J: **All-trans retinol, vitamin D and other hydrophobic compounds bind in the axial pore of the five-stranded coiled-coil domain of cartilage oligomeric matrix protein**. *The EMBO journal* 1998, **17**(18):5265-5272.
186. Stetefeld J, Jenny M, Schulthess T, Landwehr R, Engel J, Kammerer RA: **Crystal structure of a naturally occurring parallel right-handed coiled coil tetramer**. *Nature structural biology* 2000, **7**(9):772-776.
187. Eriksson M, Hassan S, Larsson R, Linder S, Ramqvist T, Lovborg H, Vikinge T, Figgemeier E, Muller J, Stetefeld J *et al*: **Utilization of a right-handed coiled-coil protein from archaebacterium Staphylothermus marinus as a carrier for cisplatin**. *Anticancer research* 2009, **29**(1):11-18.
188. Boulikas T, Vougiouka M: **Recent clinical trials using cisplatin, carboplatin and their combination chemotherapy drugs (review)**. *Oncology reports* 2004, **11**(3):559-595.
189. Deacon SP, Apostolovic B, Carbajo RJ, Schott AK, Beck K, Vicent MJ, Pineda-Lucena A, Klok HA, Duncan R: **Polymer coiled-coil conjugates: potential for development as a new class of therapeutic "molecular switch"**. *Biomacromolecules* 2011, **12**(1):19-27.
190. Hodges RS: **Boehringer Mannheim award lecture 1995. La conference Boehringer Mannheim 1995. De novo design of alpha-helical proteins: basic research to medical applications**. *Biochemistry and cell biology = Biochimie et biologie cellulaire* 1996, **74**(2):133-154.
191. Kakizawa Y, Furukawa S, Ishii A, Kataoka K: **Organic-inorganic hybrid-nanocarrier of siRNA constructing through the self-assembly of calcium phosphate and PEG-based block aniomer**. *Journal of controlled release : official journal of the Controlled Release Society* 2006, **111**(3):368-370.
192. Wu K, Liu J, Johnson RN, Yang J, Kopecek J: **Drug-free macromolecular therapeutics: induction of apoptosis by coiled-coil-mediated cross-linking of antigens on the cell surface**. *Angew Chem Int Ed Engl* 2010, **49**(8):1451-1455.
193. Pechar M, Pola R: **The coiled coil motif in polymer drug delivery systems**. *Biotechnology advances* 2013, **31**(1):90-96.
194. Vincent TL, Woolfson DN, Adams JC: **Prediction and analysis of higher-order coiled-coils: insights from proteins of the extracellular matrix, tenascins and thrombospondins**. *The international journal of biochemistry & cell biology* 2013, **45**(11):2392-2401.
195. Gruber M, Soding J, Lupas AN: **Comparative analysis of coiled-coil prediction methods**. *Journal of structural biology* 2006, **155**(2):140-145.
196. Chang CC, Song J, Tey BT, Ramanan RN: **Bioinformatics approaches for improved recombinant protein production in Escherichia coli: protein solubility prediction**. *Briefings in bioinformatics* 2014, **15**(6):953-962.

197. Chang CC, Tey BT, Song J, Ramanan RN: **Towards more accurate prediction of protein folding rates: a review of the existing web-based bioinformatics approaches**. *Briefings in bioinformatics* 2014.
198. Fiumara F, Fioriti L, Kandel ER, Hendrickson WA: **Essential role of coiled coils for aggregation and activity of Q/N-rich prions and PolyQ proteins**. *Cell* 2010, **143**(7):1121-1135.
199. Gruber M, Soding J, Lupas AN: **REPPER--repeats and their periodicities in fibrous proteins**. *Nucleic acids research* 2005, **33**(Web Server issue):W239-243.
200. Tanizawa H, Ghimire GD, Mitaku S: **A hight performance prediction system of coiled coil domains containing heptad breaks: SOSUIcoil**. *Chem-Bio Informatics Journal* 2008, **8**(3):16.
201. Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJ, Chothia C, Murzin AG: **Data growth and its impact on the SCOP database: new developments**. *Nucleic acids research* 2008, **36**(Database issue):D419-425.
202. Fu L, Niu B, Zhu Z, Wu S, Li W: **CD-HIT: accelerated for clustering the next-generation sequencing data**. *Bioinformatics* 2012, **28**(23):3150-3152.
203. Fariselli P, Molinini D, Casadio R, Krogh A: **Prediction of structurally-determined coiled-coil domains with hidden Markov models**. *Lect Notes Comput Sc* 2007, **4414**:292-302.
204. Woolfson DN, Alber T: **Predicting oligomerization states of coiled coils**. *Protein science : a publication of the Protein Society* 1995, **4**(8):1596-1607.
205. Cortes C, Vapnik V: **Support-Vector Networks**. *Mach Learn* 1995, **20**(3):273-297.
206. Breiman L: **Random forests**. *Mach Learn* 2001, **45**(1):5-32.
207. Raileanu LE, Stoffel K: **Theoretical comparison between the Gini Index and Information Gain criteria**. *Ann Math Artif Intel* 2004, **41**(1):77-93.
208. Kendall M: **A new measure of rank correlation**. *Biometrika* 1938, **30**((1-2)):9.
209. Pulst SM, Nechiporuk A, Nechiporuk T, Gispert S, Chen XN, Lopes-Cendes I, Pearlman S, Starkman S, Orozco-Diaz G, Lunkes A *et al*: **Moderate expansion of a normally biallelic trinucleotide repeat in spinocerebellar ataxia type 2**. *Nature genetics* 1996, **14**(3):269-276.
210. Sanpei K, Takano H, Igarashi S, Sato T, Oyake M, Sasaki H, Wakisaka A, Tashiro K, Ishida Y, Ikeuchi T *et al*: **Identification of the spinocerebellar ataxia type 2 gene using a direct identification of repeat expansion and cloning technique, DIRECT**. *Nature genetics* 1996, **14**(3):277-284.
211. Imbert G, Saudou F, Yvert G, Devys D, Trottier Y, Garnier JM, Weber C, Mandel JL, Cancel G, Abbas N *et al*: **Cloning of the gene for spinocerebellar ataxia 2 reveals a locus with high sensitivity to expanded CAG/glutamine repeats**. *Nature genetics* 1996, **14**(3):285-291.
212. Elden AC, Kim HJ, Hart MP, Chen-Plotkin AS, Johnson BS, Fang X, Armakola M, Geser F, Greene R, Lu MM *et al*: **Ataxin-2 intermediate-length polyglutamine expansions are associated with increased risk for ALS**. *Nature* 2010, **466**(7310):1069-1075.
213. Zhuchenko O, Bailey J, Bonnen P, Ashizawa T, Stockton DW, Amos C, Dobyns WB, Subramony SH, Zoghbi HY, Lee CC: **Autosomal dominant cerebellar ataxia (SCA6) associated with small polyglutamine expansions in the alpha 1A-voltage-dependent calcium channel**. *Nature genetics* 1997, **15**(1):62-69.
214. Jodice C, Mantuano E, Veneziano L, Trettel F, Sabbadini G, Calandriello L, Francia A, Spadaro M, Pierelli F, Salvi F *et al*: **Episodic ataxia type 2 (EA2) and spinocerebellar ataxia type 6 (SCA6) due to CAG repeat expansion in the**

CACNA1A gene on chromosome 19p. *Human molecular genetics* 1997, **6**(11):1973-1978.
215. Tonelli A, D'Angelo MG, Salati R, Villa L, Germinasi C, Frattini T, Meola G, Turconi AC, Bresolin N, Bassi MT: **Early onset, non fluctuating spinocerebellar ataxia and a novel missense mutation in CACNA1A gene**. *Journal of the neurological sciences* 2006, **241**(1-2):13-17.
216. Romaniello R, Zucca C, Tonelli A, Bonato S, Baschirotto C, Zanotta N, Epifanio R, Righini A, Bresolin N, Bassi MT *et al*: **A wide spectrum of clinical, neurophysiological and neuroradiological abnormalities in a family with a novel CACNA1A mutation**. *Journal of neurology, neurosurgery, and psychiatry* 2010, **81**(8):840-843.
217. Nagafuchi S, Yanagisawa H, Ohsaki E, Shirayama T, Tadokoro K, Inoue T, Yamada M: **Structure and expression of the gene responsible for the triplet repeat disorder, dentatorubral and pallidoluysian atrophy (DRPLA)**. *Nature genetics* 1994, **8**(2):177-182.
218. David G, Abbas N, Stevanin G, Durr A, Yvert G, Cancel G, Weber C, Imbert G, Saudou F, Antoniou E *et al*: **Cloning of the SCA7 gene reveals a highly unstable CAG repeat expansion**. *Nature genetics* 1997, **17**(1):65-70.
219. Echaniz-Laguna A, Rousso E, Anheim M, Cossee M, Tranchant C: **A family with early-onset and rapidly progressive X-linked spinal and bulbar muscular atrophy**. *Neurology* 2005, **64**(8):1458-1460.
220. Bilen J, Bonini NM: **Drosophila as a model for human neurodegenerative disease**. *Annual review of genetics* 2005, **39**:153-171.
221. Fox NK, Brenner SE, Chandonia JM: **SCOPe: Structural Classification of Proteins--extended, integrating SCOP and ASTRAL data and classification of new structures**. *Nucleic acids research* 2014, **42**(Database issue):D304-309.
222. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Muller M: **pROC: an open-source package for R and S+ to analyze and compare ROC curves**. *BMC bioinformatics* 2011, **12**:77.
223. Chawla NV: **Data Mining for Imbalanced Datasets: An Overview**. *Data Mining and Knowledge Discovery Handbook, Second Edition* 2010:875-886.
224. Munkhdalai T, Namsrai OE, Ryu K: **Self-training in significance space of support vectors for imbalanced biomedical event data**. *BMC bioinformatics* 2015, **16 Suppl 7**:S6.
225. Wu K, Edwards A, Fan W, Gao J, Zhang K: **Classifying Imbalanced Data Streams via Dynamic Feature Group Weighting with Importance Sampling**. *Proceedings of the SIAM International Conference on Data Mining SIAM International Conference on Data Mining* 2014, **2014**:722-730.
226. Yang P, Xu L, Zhou BB, Zhang Z, Zomaya AY: **A particle swarm based hybrid system for imbalanced medical data sampling**. *BMC genomics* 2009, **10 Suppl 3**:S34.
227. Joerger AC, Fersht AR: **Structure-function-rescue: the diverse nature of common p53 cancer mutants**. *Oncogene* 2007, **26**(15):2226-2242.
228. Lane DP: **Cancer. p53, guardian of the genome**. *Nature* 1992, **358**(6381):15-16.
229. Lemma V, D'Agostino M, Caporaso MG, Mallardo M, Oliviero G, Stornaiuolo M, Bonatti S: **A disorder-to-order structural transition in the COOH-tail of Fz4 determines misfolding of the L501fsX533-Fz4 mutant**. *Scientific reports* 2013, **3**:2659.
230. Dembinski H, Wismer K, Balasubramaniam D, Gonzalez HA, Alverdi V, Iakoucheva LM, Komives EA: **Predicted disorder-to-order transition**

mutations in IkappaBalpha disrupt function. *Physical chemistry chemical physics : PCCP* 2014, **16**(14):6480-6485.
231. Mittal J, Yoo TH, Georgiou G, Truskett TM: **Structural ensemble of an intrinsically disordered polypeptide**. *The journal of physical chemistry B* 2013, **117**(1):118-124.
232. Tokuriki N, Tawfik DS: **Protein dynamism and evolvability**. *Science* 2009, **324**(5924):203-207.
233. Polikar R: **Ensemble based systems in decision making**. *Circuits and Systems Magazine, IEEE* 2006, **6**(3):21-45.
234. Krogh A, Larsson B, von Heijne G, Sonnhammer EL: **Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes**. *Journal of molecular biology* 2001, **305**(3):567-580.
235. Eisenberg D, Schwarz E, Komaromy M, Wall R: **Analysis of membrane and surface protein sequences with the hydrophobic moment plot**. *Journal of molecular biology* 1984, **179**(1):125-142.
236. Hopp TP, Woods KR: **Prediction of protein antigenic determinants from amino acid sequences**. *Proceedings of the National Academy of Sciences of the United States of America* 1981, **78**(6):3824-3828.
237. Kyte J, Doolittle RF: **A simple method for displaying the hydropathic character of a protein**. *Journal of molecular biology* 1982, **157**(1):105-132.
238. Colaert N, Helsens K, Martens L, Vandekerckhove J, Gevaert K: **Improved visualization of protein consensus sequences by iceLogo**. *Nature methods* 2009, **6**(11):786-787.
239. Vucetic S, Obradovic Z, Vacic V, Radivojac P, Peng K, Iakoucheva LM, Cortese MS, Lawson JD, Brown CJ, Sikes JG *et al*: **DisProt: a database of protein disorder**. *Bioinformatics* 2005, **21**(1):137-140.
240. Simm D, Hatje K, Kollmar M: **Waggawagga: comparative visualization of coiled-coil predictions and detection of stable single alpha-helices (SAH domains)**. *Bioinformatics* 2015, **31**(5):767-769.
241. Bamford S, Dawson E, Forbes S, Clements J, Pettett R, Dogan A, Flanagan A, Teague J, Futreal PA, Stratton MR *et al*: **The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website**. *British journal of cancer* 2004, **91**(2):355-358.
242. Dinkel H, Van Roey K, Michael S, Davey NE, Weatheritt RJ, Born D, Speck T, Kruger D, Grebnev G, Kuban M *et al*: **The eukaryotic linear motif resource ELM: 10 years and counting**. *Nucleic acids research* 2014, **42**(1):D259-266.
243. Fuxreiter M, Tompa P, Simon I: **Local structural disorder imparts plasticity on linear motifs**. *Bioinformatics* 2007, **23**(8):950-956.

Appendices

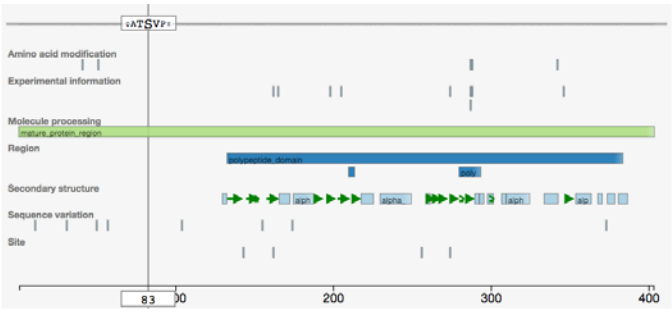
Appendices for Chapter 3

Table S3.1. GO terms selected in KinetochoreDB using the keyword 'kinetochore' from the QuickGO database

Aspect	GO ID	Name
Component	GO:0000776	kinetochore
Component	GO:0000777	condensed chromosome kinetochore
Component	GO:0005828	kinetochore microtubule
Component	GO:0000939	condensed chromosome inner kinetochore
Component	GO:0000940	condensed chromosome outer kinetochore
Component	GO:0000778	condensed nuclear chromosome kinetochore
Component	GO:0000941	condensed nuclear chromosome inner kinetochore
Component	GO:0000942	condensed nuclear chromosome outer kinetochore
Component	GO:0031617	NMS complex
Component	GO:0042729	DASH complex
Component	GO:0005818	aster
Component	GO:1990423	RZZ complex
Component	GO:0000817	COMA complex
Component	GO:0031518	CBF3 complex
Component	GO:0031262	Ndc80 complex
Component	GO:0033551	monopolin complex
Component	GO:0044816	Nsk1-Dlc1 complex
Component	GO:1990298	bub1-bub3 complex
Component	GO:0000444	MIS12/MIND type complex
Component	GO:0000818	nuclear MIS12/MIND complex
Component	GO:0005868	cytoplasmic dynein complex
Component	GO:0061638	CENP-A containing chromatin
Component	GO:0032133	chromosome passenger complex
Component	GO:0000779	condensed chromosome, centromeric region
Component	GO:0000780	condensed nuclear chromosome, centromeric region
Function	GO:0043515	kinetochore binding
Function	GO:0003777	microtubule motor activity
Process	GO:0051382	kinetochore assembly
Process	GO:0051383	kinetochore organization
Process	GO:0090234	regulation of kinetochore assembly
Process	GO:0034501	protein localization to kinetochore
Process	GO:1990299	Bub1-Bub3 complex localization to kinetochore
Process	GO:0008608	attachment of spindle microtubules to kinetochore
Process	GO:0072356	chromosome passenger complex localization to kinetochore
Process	GO:0051315	attachment of mitotic spindle microtubules to kinetochore
Process	GO:0051988	regulation of attachment of spindle microtubules to kinetochore
Process	GO:1903394	protein localization to kinetochore involved in kinetochore

		assembly
Process	GO:0051987	positive regulation of attachment of spindle microtubules to kinetochore
Process	GO:0051316	attachment of spindle microtubules to kinetochore involved in meiotic chromosome segregation
Process	GO:0051455	attachment of spindle microtubules to kinetochore involved in homologous chromosome segregation
Process	GO:0051456	attachment of spindle microtubules to kinetochore involved in meiotic sister chromatid segregation
Process	GO:2000751	histone H3-T3 phosphorylation involved in chromosome passenger complex localization to kinetochore
Process	GO:1902423	regulation of attachment of spindle microtubules to kinetochore involved in mitotic sister chromatid segregation
Process	GO:2000817	regulation of histone H3-T3 phosphorylation involved in chromosome passenger complex localization to kinetochore
Process	GO:1902424	negative regulation of attachment of spindle microtubules to kinetochore involved in mitotic sister chromatid segregation
Process	GO:1902425	positive regulation of attachment of spindle microtubules to kinetochore involved in mitotic sister chromatid segregation
Process	GO:0098653	centromere clustering
Process	GO:0031134	sister chromatid biorientation
Process	GO:2000574	regulation of microtubule motor activity
Process	GO:0072766	centromere clustering at the nuclear periphery
Process	GO:2000575	negative regulation of microtubule motor activity
Process	GO:2000576	positive regulation of microtubule motor activity
Process	GO:0034508	centromere complex assembly

A



B

C

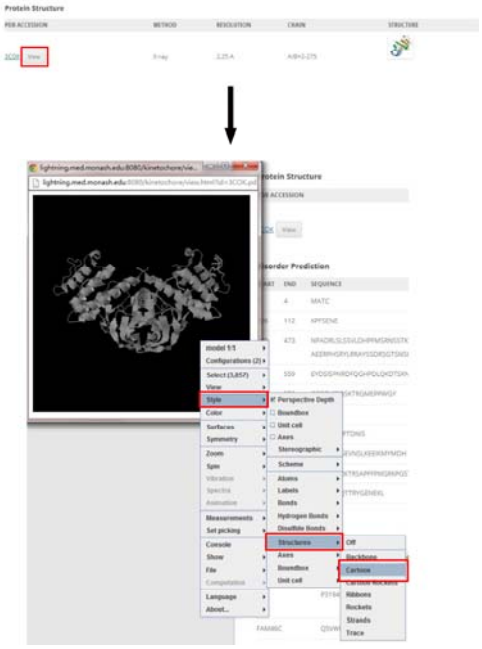


Figure S3.1. JQuery implementation for protein entries in KinetochoreDB. (A) Protein overview. (B) Protein structure view in an ensemble way with pViz. (C) Protein single structure view with Jmol.

Figure S3.2. Submission page for the users to add a new protein entry.

Appendices for Chapter 4

Table S4.1: Statistics of independent test dataset for coiled-coil oligomeric state prediction according to different lengths

Oligomeric state	Length of coiled-coil region	Number of coiled-coils
Parallel dimer	≥ 8	67
	≥15	45
Antiparallel Dimer	≥ 8	509
	≥15	302
Trimer	≥ 8	94
	≥15	63
Tetramer	≥ 8	36
	≥15	29

Table S4.2: Predictive performance of coiled-coils with non-canonical heptad registers between RFCoil, SCORER 2.0, PrOCoil and LOGICOIL on the independent test

(A) Accuracy and precision for parallel dimeric and trimeric coiled-coils with length ≥ 8 amino acids.

Predictor	True positive	False positive	True negative	False negative	Accuracy	Precision
LOGICOIL	45	1	12	22	71.3%	0.98
RFCoil	67	7	6	0	91.3%	0.91
SCORER2.0	46	1	12	21	72.5%	0.98
PrOCoil	66	6	7	1	91.3%	0.92
PrOCoil-BA	56	5	8	11	80.0%	0.92
Majority Voting	62	3	10	5	90.0%	0.95

(B) Accuracy and precision for parallel dimeric and trimeric coiled-coils with length ≥ 15 amino acids.

Predictor	True positive	False positive	True negative	False negative	Accuracy	Precision
LOGICOIL	35	1	12	10	81.0%	0.97
RFCoil	45	7	6	0	87.9%	0.87
SCORER2.0	32	1	12	13	75.9%	0.97
PrOCoil	44	6	7	1	87.9%	0.88
PrOCoil-BA	39	5	8	6	81.0%	0.89
Majority Voting	43	3	10	2	**91.4%**	0.93

Table S4.3: Predictive performance of coiled-coils with only canonical heptad registers between RFCoil, SCORER 2.0, PrOCoil and LOGICOIL on the independent test

(A) Accuracy and precision for parallel dimeric and trimeric coiled-coils with length ≥ 8 amino acids.

Predictor	True positive	False positive	True negative	False negative	Accuracy	Precision
LOGICOIL	41	1	11	18	73.2%	0.98
RFCoil	59	6	6	0	91.5%	0.91
SCORER2.0	41	1	11	18	73.2%	0.98
PrOCoil	58	6	6	1	90.1%	0.91
PrOCoil-BA	50	5	7	9	80.3%	0.91
Majority Voting	55	3	9	4	90.1%	0.95

(B) Accuracy and precision for parallel dimeric and trimeric coiled-coils with length ≥ 15 amino acids.

Predictor	True positive	False positive	True negative	False negative	Accuracy	Precision
LOGICOIL	32	1	11	6	86.0%	0.97
RFCoil	38	6	6	0	88.0%	0.86
SCORER2.0	28	1	11	10	78.0%	0.97
PrOCoil	37	6	6	1	86.0%	0.86
PrOCoil-BA	34	5	7	4	82.0%	0.87
Majority Voting	37	3	9	1	**92.0%**	0.93

Table S4.4: Predictive performance of Multcoil2 and other predictors for parallel dimeric and trimeric coiled-coil prediction

Predictor	True positive	False positive	True negative	False negative	Accuracy	Precision
LOGICOIL	8	0	5	0	100%	8
RFCoil	8	4	1	0	69.2%	0.67
SCORER2.0	7	1	4	1	84.6%	0.88
PrOCoil	8	3	2	0	76.9%	0.73
PrOCoil-BA	7	2	3	1	76.9%	0.78
Multicoil2	8	4	1	0	69.2%	0.67
Majority Voting	8	2	3	0	84.6%	0.80

Table S4.5: Comparison of accuracy and precision of coiled-coil domains predictors

(A) Accuracy and precision of different predictors for identifying coiled-coil domains.

Predictor	True positive	False positive	True negative	False negative	Accuracy	Precision
CCHMM_PROF	216	0	601	826	49.7%	1.0
MARCOIL-H	404	23	578	638	59.8%	0.95
MARCOIL-L	270	9	592	772	52.5%	0.97
COILS	512	64	537	530	63.8%	0.89
PCOILS	536	49	552	506	66.2%	0.92
Multicoil2-all	149	2	599	893	45.5%	0.99
Majority Voting	383	13	588	659	59.1%	0.97

(B) Accuracy and precision of different predictors, showing the consistency between the predicted coiled-coil domains and those annotated by SOCKET based on the protein structures

CCD > 8

Predictor	True positive	False positive	True negative	False negative	Accuracy	Precision
CCHMM_PROF	556	0	601	1620	41.7%	1.0
MARCOIL-H	427	5	596	1746	36.8%	0.99
MARCOIL-L	318	3	598	1858	33.0%	0.99
COILS	435	64	537	1741	35.0%	0.87
PCOILS	492	49	552	1684	37.6%	0.91
Multicoil2	230	2	599	1946	29.9%	0.99
Majority Voting	399	5	596	1777	35.8%	0.99

CCD >= 21

Predictor	True positive	False positive	True negative	False negative	Accuracy	Precision
CCHMM_PROF	112	0	601	149	82.7%	1.0
MARCOIL-H	150	5	596	111	86.5%	0.97
MARCOIL-L	128	3	598	133	84.2%	0.98
COILS	124	64	537	137	76.7%	0.66
PCOILS	135	49	552	126	79.7%	0.73
Multicoil2	94	2	599	167	80.4%	0.98
Majority Voting	140	5	596	121	85.4%	0.97

Table S4.6: CCDs predicted in disease associated PolyQ proteins using COILS

Protein	Predicted coiled-coil regions			PolyQ tract	Overlapping PolyQ tract		
	W=14	W=21	W=28		W=14	W=21	W=28
Ataxin-7	None	None	None	30-39	N/A	N/A	N/A
Ataxin-1	194-209	193-227 310-333	192-230 303-333	197-208 212-225	Yes	Yes	Yes
TATA binding protein	49-70 83-96	47-100	44-104	58-95	Yes	Yes	Yes
Ataxin-2	378-395 436-451 786-799 891-913	166-188 891-915	810-837	166-188	No	Yes	No
Ataxin-3	194-207 227-254 290-305	221-255 278-299	221-249 277-307	296-305	Yes	Yes	Yes
Voltage-dependent P/Q-type calcium channel subunit alpha-1A	362-379 717-745 1195-1212 1968-1981	362-387 710-750	710-749	2314-2324	No	No	No
Atrophin-1	800-844 1145-1158	482-504 791-845	788-847	484-502	No	Yes	No
Androgen receptor	57-70	54-81	52-81	58-78	Yes	Yes	Yes
Huntingtin	14-27 2554-2568 2633-2646	12-38 1441-1462	1-37	18-38	Yes	Yes	Yes

Table S4.7: CCDs predicted in disease associated PolyQ proteins using PCOILS

Protein	Predicted coiled-coil regions			PolyQ tract	Overlapping PolyQ tract		
	W=14	W=21	W=28		W=14	W=21	W=28
Ataxin-7	None	None	None	30-39	N/A	N/A	N/A
Ataxin-1	None	None	None	197-208 212-225	N/A	N/A	N/A
TATA binding protein	None	None	None	58-95	N/A	N/A	N/A
Ataxin-2	None	None	None	166-188	N/A	N/A	N/A
Ataxin-3	None	None	None	296-305	N/A	N/A	N/A
Voltage-dependent P/Q-type calcium channel subunit alpha-1A	None	383-403 721-742 1903-1925	720-750 1903-1930	2314-2324	N/A	No	No
Atrophin-1	None	None	791-819	484-502	N/A	N/A	N/A
Androgen receptor	None	None	None	58-78	N/A	N/A	N/A
Huntingtin	1444-1458	1441-1463	1441-1468	18-38	No	No	No

Table S4.8: CCDs predicted in disease-associated PolyQ proteins using Paircoil2 (P-score version, cut-off=0.025)

Protein	Predicted coiled-coil regions		PolyQ tract	Overlapping PolyQ tract	
	W=21	W=28		W=21	W=28
Ataxin-7	None	None	30-39	N/A	N/A
Ataxin-1	None	None	197-208 212-225	N/A	N/A
TATA binding protein	None	None	58-95	N/A	N/A
Ataxin-2	None	None	166-188	N/A	N/A
Ataxin-3	None	None	296-305	N/A	N/A
Voltage-dependent P/Q-type calcium channel subunit alpha-1A	719-750	719-750	2314-2324	No	No
Atrophin-1	790-819	790-819	484-502	No	No
Androgen receptor	None	None	58-78	N/A	N/A
Huntingtin	None	None	18-38	N/A	N/A

Table S4.9: CCDs predicted in disease associated PolyQ proteins using Paircoil2 (Probability score version, cut-off = 0.5)

Protein	Predicted coiled-coil regions	PolyQ tract	Overlapping PolyQ tract
Ataxin-7	None	30-39	N/A
Ataxin-1		197-208 212-225	
TATA binding protein		58-95	
Ataxin-2		166-188	
Ataxin-3		296-305	
Voltage-dependent P/Q-type calcium channel subunit alpha-1A		2314-2324	
Atrophin-1		484-502	
Androgen receptor		58-78	
Huntingtin		18-38	

Table S4.10: CCDs predicted in disease-associated PolyQ proteins using SpiriCoil

Protein	Predicted coiled-coil regions	PolyQ tract	Overlapping PolyQ tract
Ataxin-7	None	30-39	N/A
Ataxin-1	None	197-208 212-225	N/A
TATA binding protein	None	58-95	N/A
Ataxin-2	None	166-188	N/A
Ataxin-3	None	296-305	N/A
Voltage-dependent P/Q-type calcium channel subunit alpha-1A	None	2314-2324	N/A
Atrophin-1	None	484-502	N/A
Androgen receptor	None	58-78	N/A
Huntingtin	992-1003 1032-1043 1109-1120 1166-1177	18-38	No

Table S4.11: CCDs predicted in disease-associated PolyQ proteins using CCHMM_PROF

Protein	Predicted coiled-coil regions	PolyQ tract	Overlapping PolyQ tract
Ataxin-7	329-336 385-411	30-39	No
Ataxin-1	174-184 195-230 311-328 584-593	197-208 212-225	Yes
TATA binding protein	49-105 227-258 318-332	58-95	Yes
Ataxin-2	166-187 434-451 787-839	166-188	Yes
Ataxin-3	29-46 176-213 222-253 278-306	296-305	Yes
Voltage-dependent P/Q-type calcium channel subunit alpha-1A	125-138 182-189 202-230 364-447 514-520 590-618 710-747 767-798 1259-1283 1353-1381 1515-1535 1590-1601 1640-1696 1810-1817 1909-1942 1946-1953 1963-1987 2318-2324	2314-2324	Yes
Atrophin-1	68-76 485-495 793-843 1140-1157	484-502	Yes
Androgen receptor	54-80 179-200 631-715 730-737 772-804 824-898	58-78	Yes
Huntingtin	4-36 325-332 377-383 866-875 880-890 900-910 1135-1149 1270-1281 1394-1405 1441-1458 1564-1613 1744-1768 1816-1822 1921-1927 2021-2031 2035-2043 2173-2183 2221-2232 2254-2264 2339-2374 2490-2496 2554-2568 2697-2710 2887-2894	18-38	Yes

Table S4.12: CCDs predicted in disease-associated PolyQ proteins using Multicoil2

Protein	Predicted coiled-coil regions	PolyQ tract	Overlapping PolyQ tract
TATA binding protein	None	58-95	N/A
Huntingtin	None	197-208 212-225	N/A
Ataxin 1	None	197-208 212-225	N/A
Ataxin 2	None	166-188	N/A
Voltage-dependent P/Q-type calcium channel subunit alpha-1A	710-812	2314-2324	No
Atrophin 1	791-843	484-502	No
Ataxin 7	None	30-39	N/A
Androgen receptor	None	58-78	N/A
Ataxin-3	None	296-305	N/A

Table S4.13: CCDs predicted in disease-associated PolyQ proteins using MARCOIL (threshold: 50.0)

Protein	Predicted coiled-coil regions	PolyQ tract	Overlapping PolyQ tract
TATA binding protein	49-104	58-95	Yes
Huntingtin	3-37	197-208 212-225	No
Ataxin 1	194-227	197-208 212-225	Yes
Ataxin 2	166-187 441-450 788-837	166-188	Yes
Voltage-dependent P/Q-type calcium channel subunit alpha-1A	714-793	2314-2324	No
Atrophin 1	486-500 793-844	484-502	Yes
Ataxin 7	None	30-39	N/A
Androgen receptor	57-87	58-78	Yes
Ataxin-3	221-305	296-305	Yes

Table S4.14: Oligomeric state prediction for Voltage-dependent P/Q-type calcium channel subunit alpha-1A with the heptad registers from MARCOIL, Multicoil2, COILS, Paircoil2 and PCOILS

Sequence	Heptad register from MARCOIL / Multicoil2	Predictor	Antiparallel dimer	Parallel dimer	Trimer	Tetramer
AQELTKDEQEEEEAA NQKLALQKAKEVA	gabcdefgabcdefgabcdefgabcd ef	RFCoil	Parallel dimer (0.988)			N/A
		SCORER 2.0	Parallel dimer (7.370933)			N/A
		PrOCoil	Parallel dimer (-1.31802926054814)			N/A
		LOGICOIL	0.98	1.03	0.97	1.2

Sequence	Heptad register from COILS	Predictor	Antiparallel dimer	Parallel dimer	Trimer	Tetramer
AQELTKDEQEEEEAA NQKLALQKAKEVA	abcdebcdefefgabcdefgabcdef cd	RFCoil	Parallel dimer (0.964)			N/A
		SCORER 2.0	Parallel dimer (2.821595)			N/A
		PrOCoil	Parallel dimer (-0.852586401490681)			N/A
		LOGICOIL	1.03	0.92	0.89	1.09

Sequence	Heptad register from Paircoil2 / PCOILS	Predictor	Antiparallel dimer	Parallel dimer	Trimer	Tetramer
AQELTKDEQEEEEAA NQKLALQKAKEVA	efgabcdefgabcdefgabcdefgab cd	RFCoil	Parallel dimer (0.974)			N/A
		SCORER 2.0	Parallel dimer (7.439582)			N/A
		PrOCoil	Parallel dimer (-1.62351890840628)			N/A
		LOGICOIL	1	1.07	0.84	0.89

Table S4.15: Oligomeric state prediction for Atrophin-1 with the heptad registers from MARCOIL, Multicoil2, Paircoil2 and COILS

Sequence	Heptad register from MARCOIL	Predictor	Antiparallel dimer	Parallel dimer	Trimer	Tetramer
AKKRADLVEKVRRE AEQRAREEKERER	defgabcdefgabcdefgabcdefga b	RFCoil	Parallel dimer (0.885)			N/A
		SCORER 2.0	Parallel dimer (1.236905)			N/A
		PrOCoil	Parallel dimer (-1.13446490431804)			N/A
		LOGICOIL	1.05	0.89	0.86	1.06

Sequence	Heptad register from Multicoil2 / Paircoil2	Predictor	Antiparallel dimer	Parallel dimer	Trimer	Tetramer
AKKRADLVEKVRRE AEQRAREEKERER	efgabcdefgabcdefgabcdefgab c	RFCoil	Parallel dimer (0.846)			N/A
		SCORER 2.0	Parallel dimer (5.279424)			N/A
		PrOCoil	Parallel dimer (-1.32662956957516)			N/A
		LOGICOIL	0.94	1.22	0.9	0.85

Sequence	Heptad register from COILS	Predictor	Antiparallel dimer	Parallel dimer	Trimer	Tetramer
AKKRADLVEKVRRE AEQRAREEKERER	effgabcdefgabcdefgaabcdefg a	RFCoil	Parallel dimer (0.921)			N/A
		SCORER 2.0	Parallel dimer (4.434048)			N/A
		PrOCoil	Parallel dimer (-1.30943561523379)			N/A
		LOGICOIL	1.06	0.89	0.91	0.94

Appendices for Chapter 5

(1) ELM database mapping

Both disease and non-disease mutations in D→O, O→D, D→D and O→O transitions have been mapped to ELM [242] (Eukaryotic Linear Motif - http://elm.eu.org/search/) database. Eukaryotic linear motifs are short linear motifs in eukaryotic proteins, which are predominantly functional modules found in intrinsically disordered regions [243]. The Tables S5.1-S5.8 show the mapping results of our mutation of both pathogenic and nonsense for four transitions. All results listed in Table S5.1-S5.8 are experimentally verified (i.e., true positive).

Table S5.1 ELM mapping results for D→O disease mutations

UniProt	Mutation	Disease	ELM Type	Start	End	Motif
P04637	K24N	A sporadic cancer	DEG	19	26	LSQETF*FSDLWKLL*PENNVL
P04637	P34L	A sporadic cancer	MOD	30	37	LLPENN*NVLSPLPS*QAMDDL
P04637	P34L	A sporadic cancer	DOC	30	35	LLPENN*NVLSPL*PSQAMD
P35222	S37F	Pilomatrixoma (PTR) [MIM:132600]	MOD	30	37	WQQQSY*YLDSGIHS*GATTTA
P35222	S37C	Pilomatrixoma (PTR) [MIM:132600]	MOD	30	37	WQQQSY*YLDSGIHS*GATTTA
P35222	S33L	Hepatocellular carcinoma	DEG	32	37	QQSYLD*DSGIHS*GATTTA
P35222	S37Y	Hepatocellular carcinoma	MOD	30	37	WQQQSY*YLDSGIHS*GATTTA
P35222	S37F	Pilomatrixoma (PTR) [MIM:132600]	DEG	32	37	QQSYLD*DSGIHS*GATTTA
P35222	S33L	Hepatocellular carcinoma	MOD	30	37	WQQQSY*YLDSGIHS*GATTTA
P35222	S37Y	Hepatocellular carcinoma	DEG	32	37	QQSYLD*DSGIHS*GATTTA
P35222	S33F	Pilomatrixoma (PTR) [MIM:132600]	MOD	30	37	WQQQSY*YLDSGIHS*GATTTA
P35222	S33F	Pilomatrixoma (PTR) [MIM:132600]	DEG	32	37	QQSYLD*DSGIHS*GATTTA
P35222	D32Y	Pilomatrixoma (PTR) [MIM:132600]	DEG	32	37	QQSYLD*DSGIHS*GATTTA
P35222	S33Y	Pilomatrixoma (PTR) [MIM:132600]	MOD	30	37	WQQQSY*YLDSGIHS*GATTTA
P35222	D32Y	Pilomatrixoma (PTR) [MIM:132600]	MOD	30	37	WQQQSY*YLDSGIHS*GATTTA
P35222	S37C	Pilomatrixoma (PTR) [MIM:132600]	DEG	32	37	QQSYLD*DSGIHS*GATTTA
P35222	S33Y	Pilomatrixoma (PTR) [MIM:132600]	DEG	32	37	QQSYLD*DSGIHS*GATTTA
Q02548	G183S	Leukemia; acute lymphoblastic; 3 (ALL3) [MIM:613065]	LIG	178	186	DSAGSS*SYSISGILG*ITSPSA
Q99814	M535V	Erythrocytosis; familial; 4 (ECYT4) [MIM:611783]	DEG	529	542	LDLETL*LAPYIPMDGEDFQL*SPICPE

Table S5.2 ELM mapping results for D→O polymorphisms

UniProt	Mutation	ELM Type	Start	End	Motif
P35222	G34V	DEG	32	37	QQSYLD*DSGIHS*GATTTA
P35222	G34V	MOD	30	37	WQQQSY*YLDSGIHS*GATTTA
Q8WXE1	P240L	LIG	238	242	IKPEAC*CSPQF*GKTSFP
P09619	N718Y	LIG	716	719	PSAELY*YSNA*LPVGLP
P52701	K13T	LIG	4	13	MSRQ*QSTLYSFFPK*SPALSD
Q02548	G183V	LIG	178	186	DSAGSS*SYSISGILG*ITSPSA

Table S5.3 ELM mapping results for O→D disease mutations

UniProt	Mutation	Disease	ELM Type	Start	End	Motif
P00740	C108S	Hemophilia B (HEMB) [MIM:306900]	MOD	108	119	LNGGSC*CKDDINSYECWC*PFGFEG
P04156	F198S	Gerstmann-Straussler disease (GSD) [MIM:137440]	MOD	196	201	TTTKGE*ENFTET*DVKMME
P04156	E196K	Creutzfeldt-Jakob disease (CJD) [MIM:123400]	MOD	196	201	TTTKGE*ENFTET*DVKMME
P04637	F341C	Sporadic cancers	TRG	339	352	GRERFE*EMFRELNEALELKD*AQAGKE
P04637	R342P	Sporadic cancers	TRG	339	352	GRERFE*EMFRELNEALELKD*AQAGKE
P04637	R342Q	Sporadic cancers	TRG	339	352	GRERFE*EMFRELNEALELKD*AQAGKE

Table S5.4 ELM mapping results for O→D polymorphisms

UniProt	Mutation	ELM Type	Start	End	Motif
Q9UQB8	Q519R	LIG	516	521	SRNPFA*AHVQLK*PTVTND

Table S5.5 ELM mapping results for D→D disease mutations

UniProt	Mutation	Disease	ELM Type	Start	End	Motif
P04637	L323R	A sporadic cancer	TRG	305	323	PPGSTK*KRALPNNTSSSPQPKKKPL*DGEYFT
P04637	T312S	Sporadic cancers	TRG	305	323	PPGSTK*KRALPNNTSSSPQPKKKPL*DGEYFT
P04637	S314F	A sporadic cancer	DOC	312	317	ALPNNT*TSSSPQ*PKKKPL
P04049	P261A	Noonan syndrome 5 (NS5) [MIM:611553]	MOD	254	262	GSLSQR*RQRSTSTPN*VHMVST
P78314	P418L	Cherubism (CRBM) [MIM:118400]	DOC	414	421	QLPHLQ*QRSPPDGQ*SFRSFS
P04637	L323G	A sporadic cancer	TRG	305	323	PPGSTK*KRALPNNTSSSPQPKKKPL*DGEYFT
P49918	F276S	Intrauterine growth retardation; metaphyseal dysplasia; adrenal hypoplasia congenita; and genital anomalies (IMAGE) [MIM:614732]	DEG	270	282	KKLSGP*PLISDFFAKRKRS*APEKSS
P04637	L344P	Li-Fraumeni syndrome (LFS) [MIM:151623]	TRG	339	352	GRERFE*EMFRELNEALELKD*AQAGKE
P04637	A79V	Sporadic cancers	DOC	78	83	VAPAPA*AAPTPA*APAPAP
P04637	A83V	Sporadic cancers	DOC	78	83	VAPAPA*AAPTPA*APAPAP
P04637	S313C	A sporadic cancer	DOC	312	317	ALPNNT*TSSSPQ*PKKKPL
P04637	K305R	Sporadic cancers	TRG	305	323	PPGSTK*KRALPNNTSSSPQPKKKPL*DGEYFT
P35222	D32G	Pilomatrixoma (PTR) [MIM:132600]	MOD	30	37	WQQQSY*YLDSGIHS*GATTTA

P04637	A364P	A sporadic cancer	DOC	364	368	PGGSRA*AHSSH*LKSKKG
P04637	E17D	A sporadic cancer	MOD	12	18	DPSVEP*PPLSQET*FSDLWK
P04637	A307P	A sporadic cancer	TRG	305	323	PPGSTK*KRALPNNTSSSPQPKKKPL*DGEYFT
P04637	H365Y	A familial cancer not matching LFS	DOC	364	368	PGGSRA*AHSSH*LKSKKG
P04637	R306P	Li-Fraumeni syndrome (LFS) [MIM:151623]	TRG	305	323	PPGSTK*KRALPNNTSSSPQPKKKPL*DGEYFT
P04637	Q317L	A sporadic cancer	DOC	312	317	ALPNNT*TSSSPQ*PKKKPL
P04637	N310T	Sporadic cancers	TRG	305	323	PPGSTK*KRALPNNTSSSPQPKKKPL*DGEYFT
P04637	N311T	Sporadic cancers	TRG	305	323	PPGSTK*KRALPNNTSSSPQPKKKPL*DGEYFT
P04637	P309R	A sporadic cancer	TRG	305	323	PPGSTK*KRALPNNTSSSPQPKKKPL*DGEYFT
P04637	P80S	A sporadic cancer	DOC	78	83	VAPAPA*AAPTPA*APAPAP
P04637	T312I	Sporadic cancers	TRG	305	323	PPGSTK*KRALPNNTSSSPQPKKKPL*DGEYFT
P49918	K278E	Intrauterine growth retardation; metaphyseal dysplasia; adrenal hypoplasia congenita; and genital anomalies (IMAGE) [MIM:614732]	DEG	270	282	KKLSGP*PLISDFFAKRKRS*APEKSS
P04637	S315P	A sporadic cancer	DOC	312	317	ALPNNT*TSSSPQ*PKKKPL
P49918	F276V	Intrauterine growth retardation; metaphyseal dysplasia; adrenal hypoplasia congenita; and genital anomalies (IMAGE) [MIM:614732]	DEG	270	282	KKLSGP*PLISDFFAKRKRS*APEKSS
P04637	T312I	Sporadic cancers	DOC	312	317	ALPNNT*TSSSPQ*PKKKPL
P35222	G34R	Hepatocellular carcinoma	DEG	32	37	QQSYLD*DSGIHS*GATTTA
P04637	P82S	Sporadic cancers	DOC	78	83	VAPAPA*AAPTPA*APAPAP
P04049	S257L	Noonan syndrome 5 (NS5) [MIM:611553]	LIG	256	261	LSQRQR*RSTSTP*NVHMVS
P78314	P418R	Cherubism (CRBM) [MIM:118400]	DOC	414	421	QLPHLQ*QRSPPDGQ*SFRSFS
P04049	P261S	Noonan syndrome 5 (NS5) [MIM:611553]	MOD	254	262	GSLSQR*RQRSTSTPN*VHMVST
P04637	A307T	Sporadic cancers	TRG	305	323	PPGSTK*KRALPNNTSSSPQPKKKPL*DGEYFT
P35222	S37A	Medulloblastoma (MDB) [MIM:155255]	DEG	32	37	QQSYLD*DSGIHS*GATTTA
Q495M9	D458V	Usher syndrome 1G (USH1G) [MIM:606943]	LIG	456	461	ERPPAL*LEDTEL*
P04637	Q317R	Sporadic cancers	DOC	312	317	ALPNNT*TSSSPQ*PKKKPL
P04637	P80L	A sporadic cancer	DOC	78	83	VAPAPA*AAPTPA*APAPAP
P04637	K305T	A sporadic cancer	TRG	305	323	PPGSTK*KRALPNNTSSSPQPKKKPL*DGEYFT
P04637	L308V	A sporadic cancer	TRG	305	323	PPGSTK*KRALPNNTSSSPQPKKKPL*DGEYFT
P04637	S37T	A sporadic cancer	MOD	30	37	LLPENN*NVLSPLPS*QAMDDL
P51168	Y620H	Liddle syndrome (LIDDS) [MIM:177200]	LIG	617	620	IPGTPP*PPNY*DSLRLQ
P04637	P36L	A sporadic cancer	MOD	30	37	LLPENN*NVLSPLPS*QAMDDL
P04637	A79T	A sporadic cancer	DOC	78	83	VAPAPA*AAPTPA*APAPAP
P04637	Q16L	A sporadic cancer	MOD	12	18	DPSVEP*PPLSQET*FSDLWK
P04637	K305E	A sporadic cancer	TRG	305	323	PPGSTK*KRALPNNTSSSPQPKKKPL*DGEYFT
P04637	E346A	A sporadic cancer	TRG	339	352	GRERFE*EMFRELNEALELKD*AQAGKE
P04637	N311K	A sporadic cancer	TRG	305	323	PPGSTK*KRALPNNTSSSPQPKKKPL*DGEYFT
Q99814	F540L	Erythrocytosis; familial; 4 (ECYT4) [MIM:611783]	DEG	529	542	LDLETL*LAPYIPMDGEDFQL*SPICPE
P04637	R363K	A sporadic cancer	DOC	359	363	QAGKEP*PGGSR*AHSSHL
P04637	N311S	A sporadic cancer	TRG	305	323	PPGSTK*KRALPNNTSSSPQPKKKPL*DGEYFT
P04637	K305N	Sporadic cancers	TRG	305	323	PPGSTK*KRALPNNTSSSPQPKKKPL*DGEYFT
P04637	E349D	A sporadic cancer	TRG	339	352	GRERFE*EMFRELNEALELKD*AQAGKE

P04637	P309S	Li-Fraumeni syndrome (LFS) [MIM:151623]	TRG	305	323	PPGSTK*KRALPNNTSSSPQPKKKPL*DGEYFT
P04637	S37P	A sporadic cancer	MOD	30	37	LLPENN*NVLSPLPS*QAMDDL
P04637	H365R	A sporadic cancer	DOC	364	368	PGGSRA*AHSSH*LKSKKG
P04637	A364T	A sporadic cancer	DOC	364	368	PGGSRA*AHSSH*LKSKKG
P04637	A347T	Sporadic cancers	TRG	339	352	GRERFE*EMFRELNEALELKD*AQAGKE
P04637	E17D	A sporadic cancer	MOD	15	21	VEPPLS*SQETFSD*LWKLLP
P35222	G34E	Pilomatrixoma (PTR) [MIM:132600]	DEG	32	37	QQSYLD*DSGIHS*GATTTA
P04049	T260R	Noonan syndrome 5 (NS5) [MIM:611553]	LIG	256	261	LSQRQR*RSTSTP*NVHMVS
P04637	Q317R	Sporadic cancers	TRG	305	323	PPGSTK*KRALPNNTSSSPQPKKKPL*DGEYFT
P04637	P82L	Li-Fraumeni syndrome (LFS) [MIM:151623]	DOC	78	83	VAPAPA*AAPTPA*APAPAP
P04637	N311H	Sporadic cancers	TRG	305	323	PPGSTK*KRALPNNTSSSPQPKKKPL*DGEYFT
Q15583	S157C	Holoprosencephaly 4 (HPE4) [MIM:142946]	LIG	153	157	DSMDIP*PLDLS*SSAGSG
P04637	K305M	A familial cancer not matching LFS	TRG	305	323	PPGSTK*KRALPNNTSSSPQPKKKPL*DGEYFT
Q99814	M535T	Erythrocytosis; familial; 4 (ECYT4) [MIM:611783]	DEG	529	542	LDLETL*LAPYIPMDGEDFQL*SPICPE
P04637	D352H	A sporadic cancer	TRG	339	352	GRERFE*EMFRELNEALELKD*AQAGKE
P04637	S315F	A sporadic cancer	TRG	305	323	PPGSTK*KRALPNNTSSSPQPKKKPL*DGEYFT
P04637	V31I	Sporadic cancers	MOD	30	37	LLPENN*NVLSPLPS*QAMDDL
P04637	A307S	Sporadic cancers	TRG	305	323	PPGSTK*KRALPNNTSSSPQPKKKPL*DGEYFT
P04049	P261S	Noonan syndrome 5 (NS5) [MIM:611553]	LIG	256	261	LSQRQR*RSTSTP*NVHMVS
P04637	L323V	A sporadic cancer	TRG	305	323	PPGSTK*KRALPNNTSSSPQPKKKPL*DGEYFT
P35222	G34E	Pilomatrixoma (PTR) [MIM:132600]	MOD	30	37	WQQQSY*YLDSGIHS*GATTTA
P04637	G360V	A sporadic cancer	DOC	359	363	QAGKEP*PGGSR*AHSSHL
P04637	L308M	Sporadic cancers	TRG	305	323	PPGSTK*KRALPNNTSSSPQPKKKPL*DGEYFT
P04637	S315C	A sporadic cancer	TRG	305	323	PPGSTK*KRALPNNTSSSPQPKKKPL*DGEYFT
P04637	K321R	A sporadic cancer	TRG	305	323	PPGSTK*KRALPNNTSSSPQPKKKPL*DGEYFT
P04637	S314F	A sporadic cancer	TRG	305	323	PPGSTK*KRALPNNTSSSPQPKKKPL*DGEYFT
P04637	S315F	A sporadic cancer	DOC	312	317	ALPNNT*TSSSPQ*PKKKPL
P35222	G34R	Hepatocellular carcinoma	MOD	30	37	WQQQSY*YLDSGIHS*GATTTA
P04049	S257L	Noonan syndrome 5 (NS5) [MIM:611553]	MOD	254	262	GSLSQR*RQRSTSTPN*VHMVST
P04637	L348F	A sporadic cancer	TRG	339	352	GRERFE*EMFRELNEALELKD*AQAGKE
P30518	R247H	A breast cancer sample	TRG	247	249	ERPGGR*RRR*GRRTGS
P04637	Q317L	A sporadic cancer	TRG	305	323	PPGSTK*KRALPNNTSSSPQPKKKPL*DGEYFT
P04637	L35F	Sporadic cancers	DOC	30	35	LLPENN*NVLSPL*PSQAMD
Q9Y458	V16A	A colorectal cancer sample	LIG	9	17	SSRARA*AFSVEALVG*RPSKRK
P04637	K319E	Sporadic cancers	TRG	305	323	PPGSTK*KRALPNNTSSSPQPKKKPL*DGEYFT
P78314	R415Q	Cherubism (CRBM) [MIM:118400]	DOC	414	421	QLPHLQ*QRSPPDGQ*SFRSFS
P04637	S33T	A sporadic cancer	DOC	30	35	LLPENN*NVLSPL*PSQAMD
P04637	N310I	A sporadic cancer	TRG	305	323	PPGSTK*KRALPNNTSSSPQPKKKPL*DGEYFT
P51168	P617S	Liddle syndrome (LIDDS) [MIM:177200]	LIG	617	620	IPGTPP*PPNY*DSLRLQ
P04637	K320N	Sporadic cancers	TRG	305	323	PPGSTK*KRALPNNTSSSPQPKKKPL*DGEYFT
P04637	K319N	A sporadic cancer	TRG	305	323	PPGSTK*KRALPNNTSSSPQPKKKPL*DGEYFT
P35222	I35S	Hepatocellular carcinoma	MOD	30	37	WQQQSY*YLDSGIHS*GATTTA
P04049	P261L	Noonan syndrome 5 (NS5) [MIM:611553]	LIG	256	261	LSQRQR*RSTSTP*NVHMVS
P04637	S313I	A sporadic cancer	TRG	305	323	PPGSTK*KRALPNNTSSSPQPKKKPL*DGEYFT
P04637	S15R	A sporadic cancer	MOD	12	18	DPSVEP*PPLSQET*FSDLWK

Q96J92	Q565E	Pseudohypoaldosteronism 2B (PHA2B) [MIM:614491]	DEG	557	566	SVFPPE*EPEEPEADQH*QPFLFR
P04637	K321E	Kidney cancer	TRG	305	323	PPGSTK*KRALPNNTSSSPQPKKKPL*DGEYFT
P35222	D32G	Pilomatrixoma (PTR) [MIM:132600]	DEG	32	37	QQSYLD*DSGIHS*GATTTA
P04637	V31I	Sporadic cancers	DOC	30	35	LLPENN*NVLSPL*PSQAMD
P04637	P316T	A sporadic cancer	TRG	305	323	PPGSTK*KRALPNNTSSSPQPKKKPL*DGEYFT
P04049	S259A	An ovarian serous carcinoma sample	MOD	254	262	GSLSQR*RQRSTSTPN*VHMVST
P35222	D32A	Hepatocellular carcinoma	MOD	30	37	WQQQSY*YLDSGIHS*GATTTA
P25963	S32I	Ectodermal dysplasia; anhidrotic; with T-cell immunodeficiency autosomal dominant (ADEDAID) [MIM:612132]	DEG	31	36	LDDRHD*DSGLDS*MKDEEY
P04637	S313C	A sporadic cancer	TRG	305	323	PPGSTK*KRALPNNTSSSPQPKKKPL*DGEYFT
P04637	Q317H	A kidney cancer with no family history	TRG	305	323	PPGSTK*KRALPNNTSSSPQPKKKPL*DGEYFT
P04637	S313R	A sporadic cancer	TRG	305	323	PPGSTK*KRALPNNTSSSPQPKKKPL*DGEYFT
P04637	S315P	A sporadic cancer	TRG	305	323	PPGSTK*KRALPNNTSSSPQPKKKPL*DGEYFT
P04637	P318L	Sporadic cancers	TRG	305	323	PPGSTK*KRALPNNTSSSPQPKKKPL*DGEYFT
P04637	E343G	Sporadic cancers	TRG	339	352	GRERFE*EMFRELNEALELKD*AQAGKE
P04637	Q16L	A sporadic cancer	MOD	15	21	VEPPLS*SQETFSD*LWKLLP
P04049	P261A	Noonan syndrome 5 (NS5) [MIM:611553]	LIG	256	261	LSQRQR*RSTSTP*NVHMVS
P04637	L35F	Sporadic cancers	MOD	30	37	LLPENN*NVLSPLPS*QAMDDL
P04637	S33T	A sporadic cancer	MOD	30	37	LLPENN*NVLSPLPS*QAMDDL
P04637	K319R	A sporadic cancer	TRG	305	323	PPGSTK*KRALPNNTSSSPQPKKKPL*DGEYFT
Q96J92	E562K	Pseudohypoaldosteronism 2B (PHA2B) [MIM:614491]	DEG	557	566	SVFPPE*EPEEPEADQH*QPFLFR
P04637	S15R	A sporadic cancer	MOD	15	21	VEPPLS*SQETFSD*LWKLLP
P04637	Q317K	Sporadic cancers	DOC	312	317	ALPNNT*TSSSPQ*PKKKPL
Q96J92	D564A	Pseudohypoaldosteronism 2B (PHA2B) [MIM:614491]	DEG	557	566	SVFPPE*EPEEPEADQH*QPFLFR
P04637	S313N	A sporadic cancer	DOC	312	317	ALPNNT*TSSSPQ*PKKKPL
P04637	S315C	A sporadic cancer	DOC	312	317	ALPNNT*TSSSPQ*PKKKPL
P04637	Q317K	Sporadic cancers	TRG	305	323	PPGSTK*KRALPNNTSSSPQPKKKPL*DGEYFT
P04637	A364V	A sporadic cancer	DOC	364	368	PGGSRA*AHSSH*LKSKKG
P04637	P322R	Sporadic cancers	TRG	305	323	PPGSTK*KRALPNNTSSSPQPKKKPL*DGEYFT
P04049	T260R	Noonan syndrome 5 (NS5) [MIM:611553]	MOD	254	262	GSLSQR*RQRSTSTPN*VHMVST
P49918	R279P	Intrauterine growth retardation; metaphyseal dysplasia; adrenal hypoplasia congenita; and genital anomalies (IMAGE) [MIM:614732]	DEG	270	282	KKLSGP*PLISDFFAKRKRS*APEKSS
P35222	S37A	Medulloblastoma (MDB) [MIM:155255]	MOD	30	37	WQQQSY*YLDSGIHS*GATTTA
P04637	T312S	Sporadic cancers	DOC	312	317	ALPNNT*TSSSPQ*PKKKPL
P04637	Q317P	A sporadic cancer	TRG	305	323	PPGSTK*KRALPNNTSSSPQPKKKPL*DGEYFT
P04637	A83E	A sporadic cancer	DOC	78	83	VAPAPA*AAPTPA*APAPAP
P04637	P316T	A sporadic cancer	DOC	312	317	ALPNNT*TSSSPQ*PKKKPL
P78314	R415P	Cherubism (CRBM) [MIM:118400]	DOC	414	421	QLPHLQ*QRSPPDGQ*SFRSFS
P04637	Q317P	A sporadic cancer	DOC	312	317	ALPNNT*TSSSPQ*PKKKPL
P04637	T81I	Sporadic cancers	DOC	78	83	VAPAPA*AAPTPA*APAPAP

P51168	P618R	Liddle syndrome (LIDDS) [MIM:177200]	LIG	617	620	IPGTPP*PPNY*DSLRLQ
P04637	P322L	Sporadic cancers	TRG	305	323	PPGSTK*KRALPNNTSSSPQPKKKPL*DGEYFT
P12644	R287H	Non-syndromic orofacial cleft 11 (OFC11) [MIM:600625]	CLV	287	293	HALTRR*RRRAKRS*PKHHSQ
P04637	F385L	A sporadic cancer	MOD	385	388	HKKLMF*FKTE*GPDSD
P04049	S259A	An ovarian serous carcinoma sample	LIG	256	261	LSQRQR*RSTSTP*NVHMVS
P78314	G420R	Cherubism (CRBM) [MIM:118400]	DOC	414	421	QLPHLQ*QRSPPDGQ*SFRSFS
P78314	G420E	Cherubism (CRBM) [MIM:118400]	DOC	414	421	QLPHLQ*QRSPPDGQ*SFRSFS
P35222	I35S	Hepatocellular carcinoma	DEG	32	37	QQSYLD*DSGIHS*GATTTA
P04637	L348S	A sporadic cancer	TRG	339	352	GRERFE*EMFRELNEALELKD*AQAGKE
P04049	R256S	Noonan syndrome 5 (NS5) [MIM:611553]	MOD	254	262	GSLSQR*RQRSTSTPN*VHMVST
P04637	R306Q	Sporadic cancers	TRG	305	323	PPGSTK*KRALPNNTSSSPQPKKKPL*DGEYFT
P04637	A347G	A sporadic cancer	TRG	339	352	GRERFE*EMFRELNEALELKD*AQAGKE
P04049	S259F	Noonan syndrome 5 (NS5) [MIM:611553]	LIG	256	261	LSQRQR*RSTSTP*NVHMVS
P04637	S313R	A sporadic cancer	DOC	312	317	ALPNNT*TSSSPQ*PKKKPL
P04049	P261L	Noonan syndrome 5 (NS5) [MIM:611553]	MOD	254	262	GSLSQR*RQRSTSTPN*VHMVST
P78314	P418H	Cherubism (CRBM) [MIM:118400]	DOC	414	421	QLPHLQ*QRSPPDGQ*SFRSFS
Q06187	P190K	A lung large cell carcinoma sample	LIG	186	192	HRKTKK*KPLPPTP*EEDQIL
P49918	D274N	Intrauterine growth retardation; metaphyseal dysplasia; adrenal hypoplasia congenita; and genital anomalies (IMAGE) [MIM:614732]	DEG	270	282	KKLSGP*PLISDFFAKRKRS*APEKSS
P35222	D32A	Hepatocellular carcinoma	DEG	32	37	QQSYLD*DSGIHS*GATTTA
P04637	L323P	A sporadic cancer	TRG	305	323	PPGSTK*KRALPNNTSSSPQPKKKPL*DGEYFT
P04637	S366A	A familial cancer not matching LFS	DOC	364	368	PGGSRA*AHSSH*LKSKKG
P04637	P316L	A sporadic cancer	TRG	305	323	PPGSTK*KRALPNNTSSSPQPKKKPL*DGEYFT
P04637	A79G	A sporadic cancer	DOC	78	83	VAPAPA*AAPTPA*APAPAP
P04637	Q317H	A kidney cancer with no family history	DOC	312	317	ALPNNT*TSSSPQ*PKKKPL
P04637	A78V	Sporadic cancers	DOC	78	83	VAPAPA*AAPTPA*APAPAP
P04637	L344R	A sporadic cancer	TRG	339	352	GRERFE*EMFRELNEALELKD*AQAGKE
P04637	F385L	A sporadic cancer	DOC	381	385	STSRHK*KKLMF*KTEGPD
P04049	R256S	Noonan syndrome 5 (NS5) [MIM:611553]	LIG	256	261	LSQRQR*RSTSTP*NVHMVS
P04049	S259F	Noonan syndrome 5 (NS5) [MIM:611553]	MOD	254	262	GSLSQR*RQRSTSTPN*VHMVST
P04637	S313N	A sporadic cancer	TRG	305	323	PPGSTK*KRALPNNTSSSPQPKKKPL*DGEYFT
P04637	L323M	A sporadic cancer	TRG	305	323	PPGSTK*KRALPNNTSSSPQPKKKPL*DGEYFT
P04637	P316L	A sporadic cancer	DOC	312	317	ALPNNT*TSSSPQ*PKKKPL
P04637	S313I	A sporadic cancer	DOC	312	317	ALPNNT*TSSSPQ*PKKKPL

Table S5.6 ELM mapping results for D→D polymorphisms

UniProt	Mutation	ELM Type	Start	End	Motif
P08047	T737A	MOD	736	742	GSEGSG*GTATPSA*LITTNM
P38398	R507I	TRG	503	508	LTNKLK*KRKRRP*TSGLHP
P04637	G360A	DOC	359	363	QAGKEP*PGGSR*AHSSHL

Q9Y5M8	V9L	TRG	7	9	ASADSR*RRV*ADGGGA
P06213	Y1361C	LIG	1361	1364	EEHIPY*YTHM*NGGKKN
P43630	T397M	LIG	396	401	QDPQEV*VTYAQL*DHCVFI
Q9NZC9	Y206D	LIG	202	217	ASPSGQ*QNISYIHSSSESVTPR*TEGRLQ
P04921	K124E	LIG	123	128	AGDSSR*RKEYFI*
Q13464	T1112P	CLV	1110	1114	SFPSAD*DETDG*NLPESR
P30559	A16S	MOD	14	19	WSAEAA*ANASAA*PPGAEG
P38398	R507I	TRG	501	508	RPLTNK*KLKRKRRP*TSGLHP
P38936	D149G	DEG	145	157	RKRRQT*TSMTDFYHSKRRL*IFSKRK
P38936	D149G	LIG	144	153	GRKRRQ*QTSMTDFYHS*KRRLIF
Q14108	E471G	TRG	470	476	DEGTAD*DERAPLI*RT
P48552	I441V	LIG	440	444	YSNCVP*PIDLS*CKHRTE
Q9H4A3	R1957H	DOC	1957	1961	ANKVGR*RFSVS*KTEDKI
Q13492	F641L	LIG	638	642	PVMRPP*PNPFG*PVSGAQ
P04049	T260I	LIG	256	261	LSQRQR*RSTSTP*NVHMVS
P13051	Q4R	LIG	4	13	MIGQ*QKTLYSFFSP*SPARKR
O94979	P841L	LIG	839	843	HGENPP*PPPGF*IMHGNV
P38398	R507I	TRG	502	507	PLTNKL*LKRKRR*PTSGLH
P38398	R507I	MOD	504	512	TNKLKR*RKRRPTSGL*HPEDFI
Q9NZC9	I207F	LIG	202	217	ASPSGQ*QNISYIHSSSESVTPR*TEGRLQ
P04049	T260I	MOD	254	262	GSLSQR*RQRSTSTPN*VHMVST
Q9NZC9	A22G	LIG	12	27	EEQRKK*KIEENRQKALARRAEK*LLAEQH
Q71U36	E447K	LIG	443	451	DSVEGE*EGEEEGEEY*
P38398	R507I	TRG	504	509	TNKLKR*RKRRPT*SGLHPE
P38936	D149G	TRG	142	158	SQGRKR*RRQTSMTDFYHSKRRLI*FSKRKP
O95081	T365N	LIG	365	369	AFGAFT*TNPFT*APAAQS
O43683	N534D	DEG	534	538	VFEDGN*NKENY*GLPQPK
O15151	T406I	MOD	400	406	LDLAHS*SSESQET*ISSMGE

Table S5.7 ELM mapping results for O→O disease mutations

UniProt	Mutation	Disease	ELM Type	Start	End	Motif
Q9NVV9	N136S	Dystonia 6; torsion (DYT6) [MIM:602629]	LIG	134	137	LSVFCD*DHNY*TVEDTM
P04156	V180I	Creutzfeldt-Jakob disease (CJD) [MIM:123400]	MOD	180	185	FVHDCV*VNITIK*QHTVTT
Q9GZX7	R24W	Immunodeficiency with hyper-IgM 2 (HIGM2) [MIM:605258]	MOD	24	30	RWAKGR*RRETYLC*YVVKRR
P78363	S100P	Stargardt disease 1 (STGD1) [MIM:248200]	MOD	97	102	GIVSNY*YNNSIL*ARVYRD
P04637	E339Q	A sporadic cancer	TRG	339	352	GRERFE*EMFRELNEALELKD*AQAGKE
Q99814	G537R	Erythrocytosis; familial; 4 (ECYT4) [MIM:611783]	DEG	529	542	LDLETL*LAPYIPMDGEDFQL*SPICPE
Q9NVV9	Y137C	Dystonia 6; torsion (DYT6) [MIM:602629]	LIG	134	137	LSVFCD*DHNY*TVEDTM
O76024	P504L	Wolfram syndrome 1 (WFS1) [MIM:222300]	MOD	499	504	GHLVVL*LNVSVP*CLLYVY
Q14524	Q1909R	Long QT syndrome 3 (LQT3) [MIM:603830]	LIG	1902	1921	RRKHEE*EVSAMVIQRAFRRHLLQRSL*KHASFL
Q99814	G537W	Erythrocytosis; familial; 4 (ECYT4) [MIM:611783]	DEG	529	542	LDLETL*LAPYIPMDGEDFQL*SPICPE
P04637	E339K	A sporadic cancer	TRG	339	352	GRERFE*EMFRELNEALELKD*AQAGKE
P78363	S445R	Stargardt disease 1 (STGD1) [MIM:248200]	MOD	443	448	IWYFFD*DNSTQM*NMIRDT

P00740	D110N	Hemophilia B (HEMB) [MIM:306900]	MOD	108	119	LNGGSC*CKDDINSYECWC*PFGFEG
Q14524	S1904L	Long QT syndrome 3 (LQT3) [MIM:603830]	LIG	1902	1921	RRKHEE*EVSAMVIQRAFRRHLLQRSL*KHASFL
Q9Y463	Q275R	A metastatic melanoma sample	MOD	267	277	SSCQLG*GQRIYQYIQSR*FYRSPE
P48048	V315G	Bartter syndrome 2 (BS2) [MIM:241200]	MOD	310	316	SATCQV*VRTSYVP*EEVLWG
Q9UH77	Y557C	Pseudohypoaldosteronism 2D (PHA2D) [MIM:614495]	DEG	557	566	ASVEYY*YNPVTDKWTL*LPTNMS
P00740	Y115C	Hemophilia B (HEMB) [MIM:306900]	MOD	108	119	LNGGSC*CKDDINSYECWC*PFGFEG
P78363	D1532N	Stargardt disease 1 (STGD1) [MIM:248200]	MOD	1528	1533	QDLTDR*RNISDF*LVKTYP
P48048	S219R	Bartter syndrome 2 (BS2) [MIM:241200]	MOD	216	222	IRVANL*LRKSLLI*GSHIYG
P00740	N113K	Hemophilia B (HEMB) [MIM:306900]	MOD	108	119	LNGGSC*CKDDINSYECWC*PFGFEG
P04156	E200K	Creutzfeldt-Jakob disease (CJD) [MIM:123400]	MOD	196	201	TTTKGE*ENFTET*DVKMME
P00740	P101R	Hemophilia B (HEMB) [MIM:306900]	MOD	97	102	VDGDQC*CESNPC*LNGGSC
P04275	C1149R	Von Willebrand disease 1 (VWD1) [MIM:193400]	MOD	1147	1149	CEWRYN*NSC*APACQV
P00740	C119R	Hemophilia B (HEMB) [MIM:306900]	MOD	108	119	LNGGSC*CKDDINSYECWC*PFGFEG
P00740	C102R	Hemophilia B (HEMB) [MIM:306900]	MOD	97	102	VDGDQC*CESNPC*LNGGSC
P00740	C119F	Hemophilia B (HEMB) [MIM:306900]	MOD	108	119	LNGGSC*CKDDINSYECWC*PFGFEG
P00740	I112S	Hemophilia B (HEMB) [MIM:306900]	MOD	108	119	LNGGSC*CKDDINSYECWC*PFGFEG
O14686	R5340L	Kabuki syndrome 1 (KABUK1) [MIM:147920]	LIG	5338	5344	INPTGC*CARSEPK*ILTHYK
P25054	S171I	Familial adenomatous polyposis (FAP) [MIM:175100]	TRG	163	176	YYAQLQ*QNLTKRIDSLPLTE*NFSLQT
P01130	Y828C	Familial hypercholesterolemia (FH) [MIM:143890]	LIG	822	829	NINSIN*NFDNPVYQ*KTTEDE
P06213	P997T	Rabson-Mendenhall syndrome (RMS) [MIM:262190]	LIG	993	999	LGPLYA*ASSNPEY*LSASDV
P04629	E492K	Congenital insensitivity to pain with anhidrosis (CIPA) [MIM:256800]	LIG	490	496	GLQGHI*IIENPQY*FSDACV
Q99814	P534L	Erythrocytosis; familial; 4 (ECYT4) [MIM:611783]	DEG	529	542	LDLETL*LAPYIPMDGEDFQL*SPICPE
P00740	C97S	Hemophilia B (HEMB) [MIM:306900]	MOD	97	102	VDGDQC*CESNPC*LNGGSC
Q9NVV9	N136K	Dystonia 6; torsion (DYT6) [MIM:602629]	LIG	134	137	LSVFCD*DHNY*TVEDTM
P04637	R342L	A sporadic cancer	TRG	339	352	GRERFE*EMFRELNEALELKD*AQAGKE
P01130	P826S	Familial hypercholesterolemia (FH) [MIM:143890]	LIG	822	829	NINSIN*NFDNPVYQ*KTTEDE

Table S5.8 ELM mapping results for O→O polymorphisms

UniProt	Mutation	ELM Type	Start	End	Motif
P24394	A492T	LIG	491	497	ETPLVI*IAGNPAY*RSFSNS
P02730	R646Q	MOD	641	646	DGFKVS*SNSSAR*GWVIHP
P24394	A492V	LIG	491	497	ETPLVI*IAGNPAY*RSFSNS
P01130	V827I	LIG	822	829	NINSIN*NFDNPVYQ*KTTEDE
Q99741	V441I	CLV	439	443	ISQVIS*SEVDG*NRMTLS

Q13698	R1539C	LIG	1523	1542	VTVGKF*FYATFLIQEHFRKFMKRQEE*YYGYRP
P28562	A56T	DOC	54	62	STIVRR*RRAKGAMGL*EHIVPN
P02730	R646W	MOD	641	646	DGFKVS*SNSSAR*GWVIHP
Q14145	D349N	DEG	347	352	LEAYNP*PSDGTW*LRLADL
P21452	I23T	MOD	18	23	SSGPES*SNTTGI*TAFSMP
P01266	G815R	MOD	815	820	REAASG*GNFSLF*IQSLYE
Q9NQ25	T302M	LIG	300	307	LKEDPA*ANTVYSTV*EIPKKM
P49792	V548L	LIG	536	549	WWDAVC*CTLIHRKAVPGNVA*KLRLLV
P04156	T183A	MOD	180	185	FVHDCV*VNITIK*QHTVTT
Q14145	G350S	DEG	347	352	LEAYNP*PSDGTW*LRLADL
P01266	S1222L	MOD	1219	1224	RCPLPF*FNASEV*VGGTIL
Q9HBG7	M602V	LIG	599	606	ESVVGE*ENTMYAQV*FNLQGK
Q14145	G350S	DEG	349	354	AYNPSD*DGTWLR*LADLQV
Q9GZX7	R25C	MOD	24	30	RWAKGR*RRETYLC*YVVKRR
P55211	R192C	MOD	191	199	DCEKLR*RRRFSSLHF*MVEVKG
P49137	A361S	TRG	354	368	WEDVKE*EEMTSALATMRVDYE*QIKIKK
Q13111	M239V	LIG	238	242	FKGKVP*PMVVL*QDILAV
P04234	Q147R	LIG	146	163	ALLRND*DQVYQPLRDRDDAQYSHL*GGNWAR
Q14145	D349N	DEG	349	354	AYNPSD*DGTWLR*LADLQV
Q14145	D236H	DEG	231	236	CQLVTL*LISRDD*LNVRCE

(2) The ensemble figures for Table 5.3 and Table 5.4 showing detailed function/domain annotations and sequence context for predicted disordered regions in the wild-type proteins. The legends have been listed as follows:

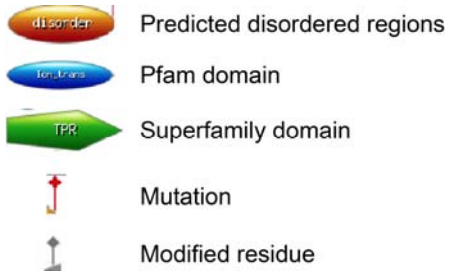
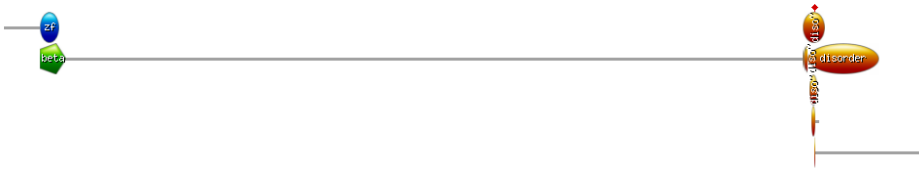
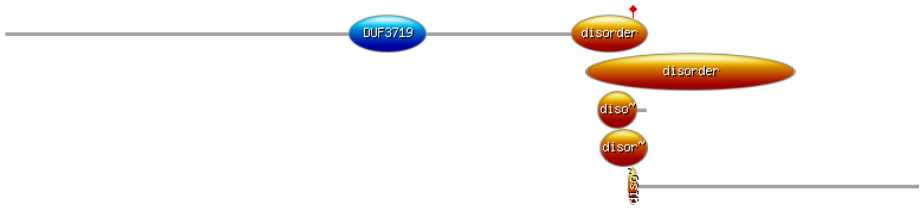
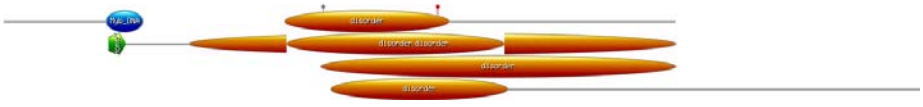
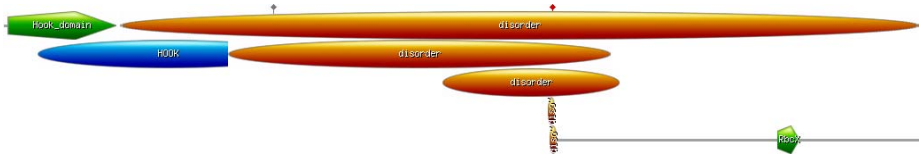
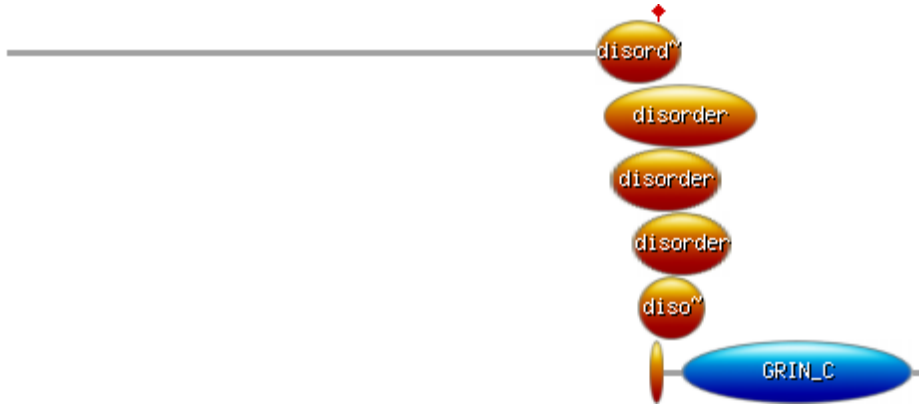


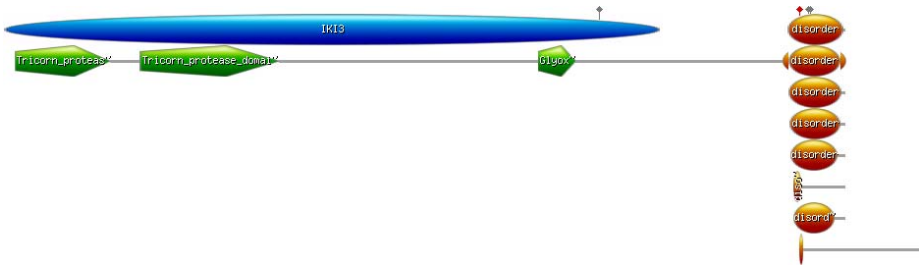
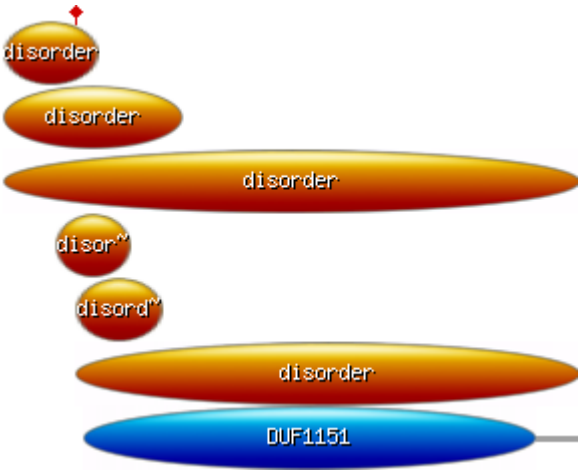
Table S5.9 Detailed functional annotations for proteins in Table 5.3 including modified residues, protein superfamily domains and Pfam domains

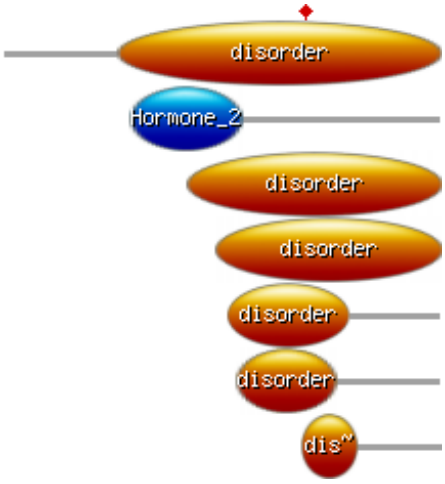
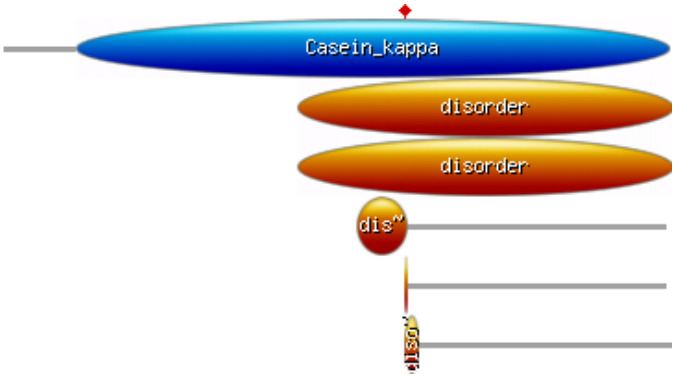
A0JNW5 UHRF1-binding protein 1-like	Superfamily domain	1420-1460; Ribosomal protein L29 (L29p)	
	Pfam domain (http://pfam.xfam.org/protein/A0JNW5)	1-103;Chorein_N	
	Post-translational modification site	414;Phosphoserine	
	Predicted disordered region	M1111L: -	1104-1119;1055-1190;1049-1132;1098-1135;880-1186;

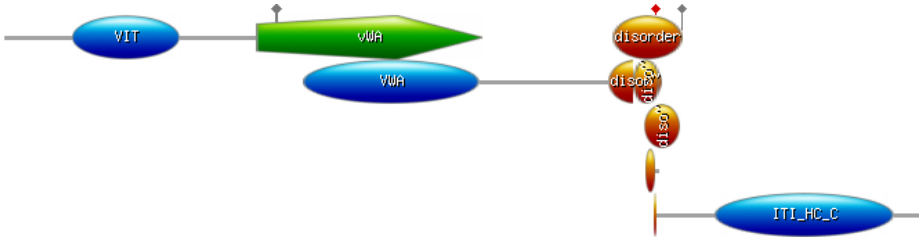
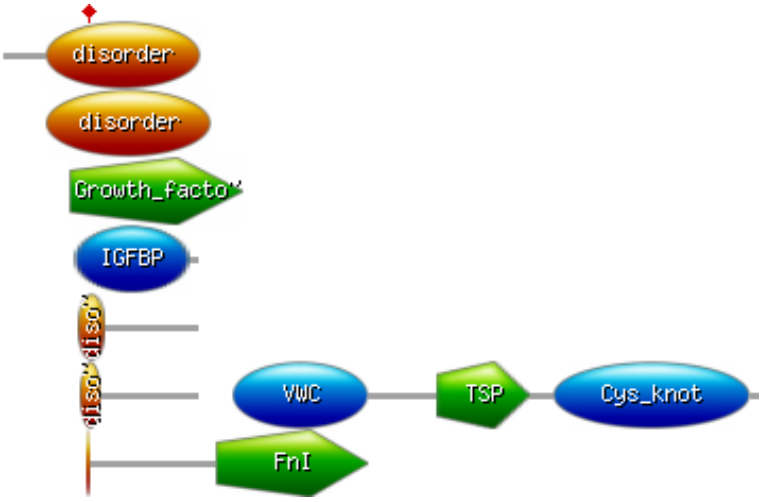
			
A4D1E1 Zinc finger protein 804B	Superfamily domain	54-91; beta-beta-alpha zinc fingers	
	Pfam domain (http://pfam.xfam.org/protein/A4D1E1)	54-82;zf-C2H2_jaz	
	Post-translational modification site		
	Predicted disordered region	V1195I: -	1192-1196;1195-1195;1178-1204;1179-1211;1188-1199;1185-1289;
			
A5PLN7 Protein FAM149A	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/A5PLN7)	292-357;DUF3719	
	Post-translational modification site		
	Predicted disordered region	P532L: -	492-669;480-544;502-535;504-544;528-535;
			
A6H8Y1 Protein FAM149A	Superfamily domain	301-350; Homeodomain-like	
	Pfam domain (http://pfam.xfam.org/protein/A6H8Y1)	293-399;Myb_DNA-bind_7	
	Post-translational modification site	915;Phosphothreonine	
	Predicted disordered region	F1244I: -	938-1444;812-1434;908-1927;805-1275;529-1926;
			
A6NC98 Coiled-coil domain-containing protein 88B	Superfamily domain	7-182; Hook domain	
		1249-1284; RbcX-like	
	Pfam domain (http://pfam.xfam.org/protein/A6NC98)	56-476;HOOK	
	Post-translational modification site	436;Phosphoserine	
	Predicted disordered region	D886A: -	879-890;882-892;363-979;709-993;188-1476;

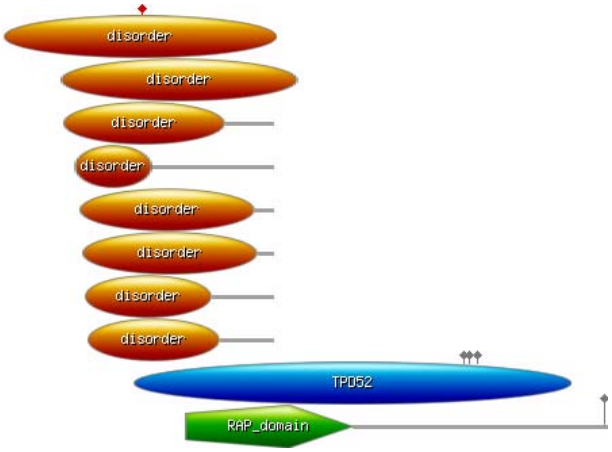
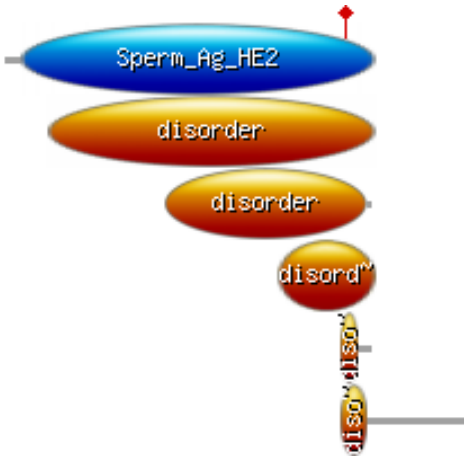
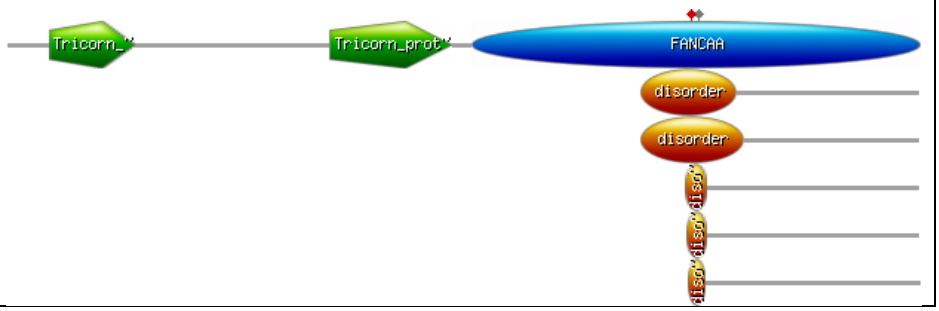
			
O43303 Centriolar coiled-coil protein of 110 kDa	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/O43303)	29-136;CALM_bind	
		589-709;CALM_bind	
	Post-translational modification site		
	Predicted disordered region	P171L: -	164-171;153-205;166-177;91-221;155-216;170-221;147-222;166-180;
			
O60269 G protein-regulated inducer of neurite outgrowth 2	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/O60269)	340-455;GRIN_C	
	Post-translational modification site		
	Predicted disordered region	S328C: -	297-339;324-330;315-364;301-377;318-351;304-359;
			
O75691 Small subunit processome component 20 homolog	Superfamily domain	1441-1555; ARM repeat	
		1630-1687; ARM repeat	
		1801-1891; ARM repeat	
		1937-1995; ARM repeat	
		2082-2502; ARM repeat	

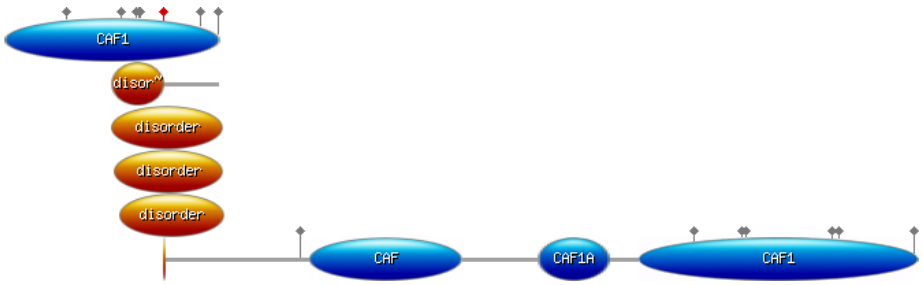
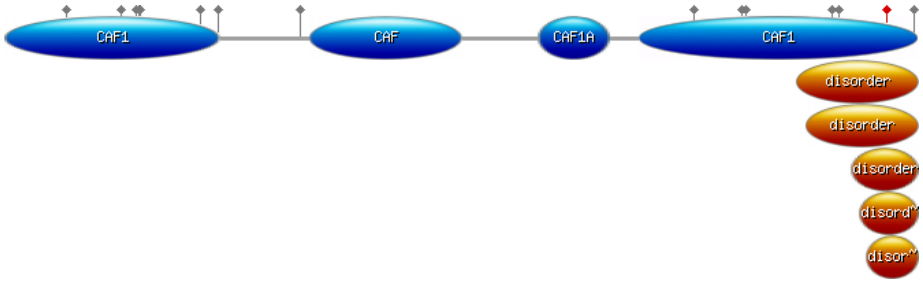
		77-455; ARM repeat	
		487-552; ARM repeat	
		520-675; ARM repeat	
		808-857; ARM repeat	
		906-993; ARM repeat	
		1099-1158; ARM repeat	
		1298-1420; ARM repeat	
		2437-2697; ARM repeat	
	Pfam domain (http://pfam.xfam.org/protein/O75691)	909-1534;DRIM	
	Post-translational modification site	788;Phosphoserine	
		1741;Phosphothreonine	
		2601;Phosphoserine	
	Predicted disordered region	E2612Q: -	2593-2614;2568-2616;2575-2617;2584-2615;2573-2616;
			
O75952 Calcium-binding tyrosine phosphorylation-regulated protein	Superfamily domain	6-47; Dimerization-anchoring domain of cAMP-dependent PK regulatory subunit	
	Pfam domain (http://pfam.xfam.org/protein/O75952)	12-49;RIIa	
	Post-translational modification site		
	Predicted disordered region	T74M: -	68-164;74-173;65-165;67-199;68-194;
			
O95163 Elongator complex protein 1	Superfamily domain	17-155; Tricorn protease domain 2	
		198-400; Tricorn protease domain 2	
		779-832; Glyoxalase/Bleomycin resistance protein/Dihydroxybiphenyl dioxygenase	
	Pfam domain (http://pfam.xfam.org/protein/O95163)	1-954;IKI3	
	Post-translational modification site	867;Phosphoserine	
		1171;Phosphoserine	
		1174;Phosphoserine	
	Predicted disordered region	P1158L: -	1141-1221;1144-1217;1150-1160;1150-1209;1146-1213;1143-1219;1134-1226;1143-

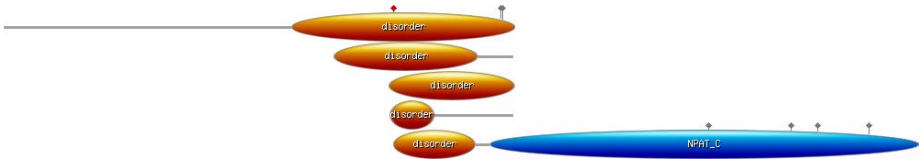
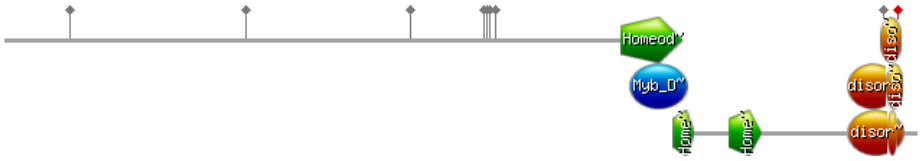
			1217;1158-1163;
			
O95990 Protein FAM107A	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/O95990)	21-133;DUF1151	
	Post-translational modification site		
	Predicted disordered region	P19L: Renal cell carcinoma cell line	1-144;1-24;1-45;19-40;14-32;19-144;
			
P01286 Somatolibe rin	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/P01286)	32-59;Hormone_2	
	Post-translational modification site	75;Leucine amide	
	Predicted disordered region	L75F: -	29-108;58-82;56-85;53-108;74-87;46-108;

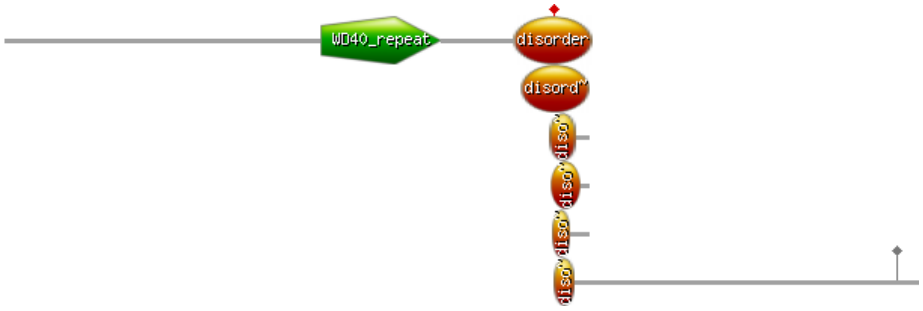
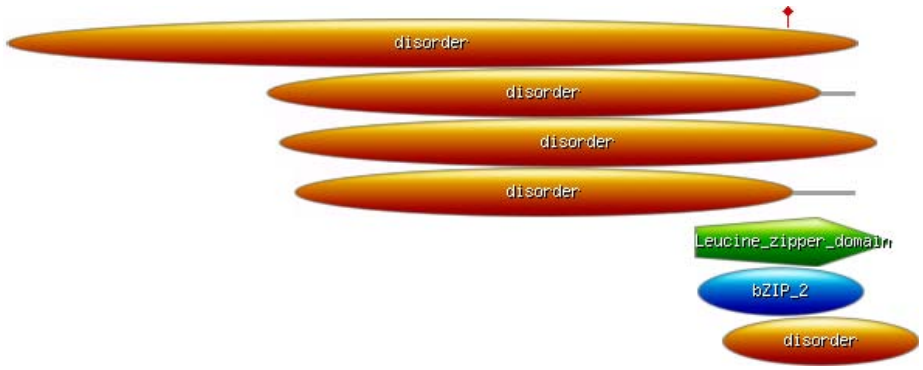
			
P07498 Kappa-casein	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/P07498)	21-181;Casein_kappa	
	Post-translational modification site		
	Predicted disordered region	R110L: -	110-113;110-110;97-110;81-182;81-182;
			
P19823 Inter-alpha-trypsin inhibitor heavy chain H2	Superfamily domain	261-494; vWA-like	
	Pfam domain (http://pfam.xfam.org/protein/P19823)	72-183;VIT	
		736-921;ITI_HC_C	
		310-492;VWA	
	Post-translational modification site	282;4-carboxyglutamate	
		283;4-carboxyglutamate	
		702;Aspartate 1-(chondroitin 4-sulfate)-ester	
	Predicted disordered region	P674A: -	673-675;630-702;625-679;663-699;652-678;665-674;

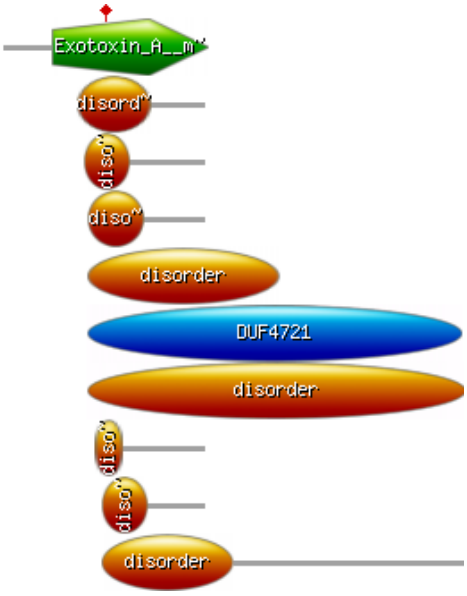
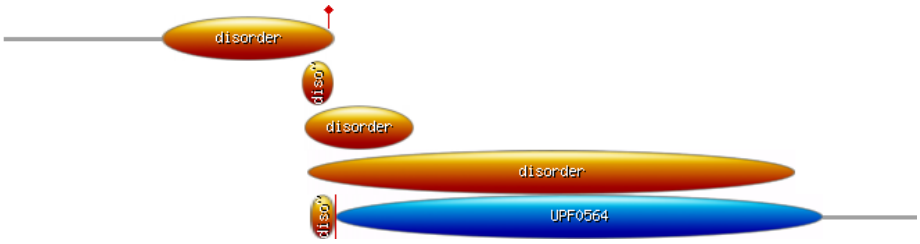
			
P48745 Protein NOV homolog	Superfamily domain	33-114; Growth factor receptor domain	
		102-173; FnI-like domain	
		206-249; TSP-1 type 1 repeat	
	Pfam domain (http://pfam.xfam.org/protein/P48745)	261-353;Cys_knot	
		110-173;VWC	
		35-89;IGFBP	
	Post-translational modification site		
	Predicted disordered region	R42Q: -	38-47;41-42;22-94;37-49;22-99;
			
P55327 Tumor protein D52	Superfamily domain	68-129; RAP domain-like	
	Pfam domain (http://pfam.xfam.org/protein/P55327)	49-210;TPD52	
	Post-translational modification site	171;Phosphoserine	
		173;Phosphothreonine	
		176;Phosphoserine	
		223;Phosphoserine	
	Predicted disordered region	D52Y: -	1-101;23-82;27-55;32-80;31-77;30-94;29-93;22-109;

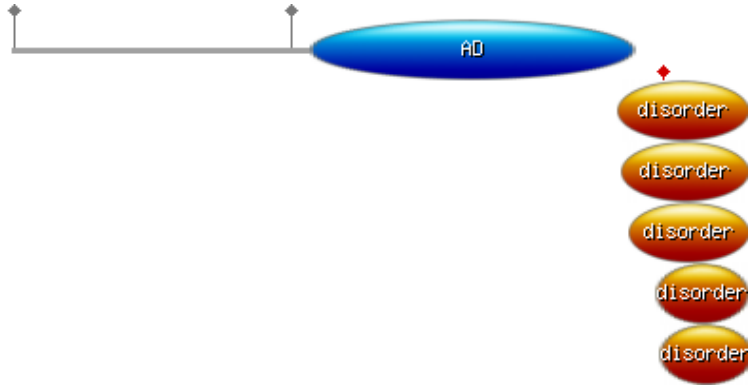
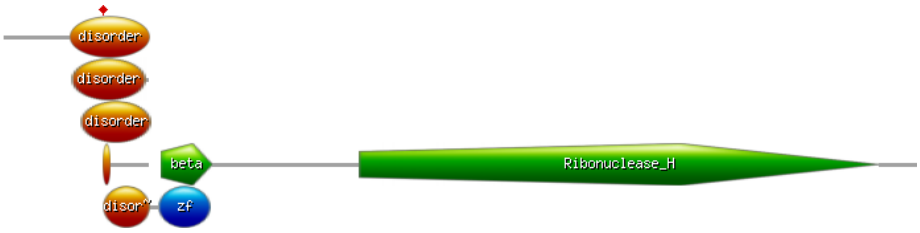
<table>
<tr><td></td><td colspan="3"> 
</td></tr>
<tr><td rowspan="5">Q08648
Sperm-associated antigen 11B</td><td>Superfamily domain</td><td></td><td></td></tr>
<tr><td>Pfam domain (http://pfam.xfam.org/protein/Q08648)</td><td>5-83;Sperm_Ag_HE2</td><td></td></tr>
<tr><td>Post-translational modification site</td><td></td><td></td></tr>
<tr><td>Predicted disordered region</td><td>R77Q: -</td><td>37-81;76-81;62-83;11-83;76-79;</td></tr>
<tr><td colspan="3"> 
</td></tr>
<tr><td rowspan="5">Q0VG06
Fanconi anemia-associated protein of 100 kDa</td><td rowspan="2">Superfamily domain</td><td>41-123; Tricorn protease domain 2</td><td></td></tr>
<tr><td>312-430; Tricorn protease domain 2</td><td></td></tr>
<tr><td>Pfam domain (http://pfam.xfam.org/protein/Q0VG06)</td><td>449-881;FANCAA</td><td></td></tr>
<tr><td>Post-translational modification site</td><td>667;Phosphoserine</td><td></td></tr>
<tr><td>Predicted disordered region</td><td>P660L: -</td><td>654-675;656-675;612-711;612-704;657-673;</td></tr>
<tr><td colspan="3"> 
</td></tr>
</table>

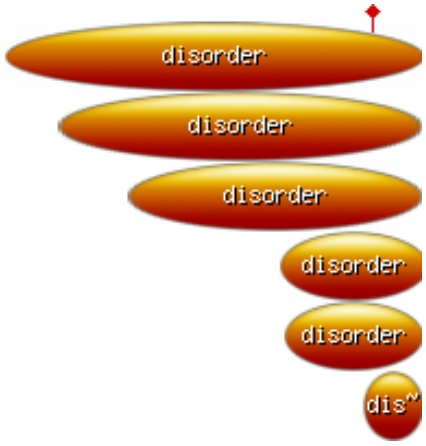
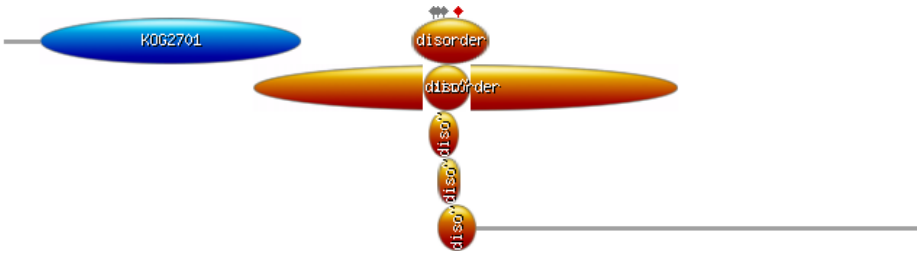
Q13111 Chromatin assembly factor 1 subunit A	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/Q13111)	558-632;CAF1A	
		320-479;CAF-1_p150	
		665-956;CAF1-p150_C2	
		1-226;CAF1-p150_N	
	Post-translational modification site	65;Phosphoserine	
		123;Phosphoserine	
		138;Phosphoserine	
		141;Phosphoserine	
		143;Phosphoserine	
		206;Phosphoserine	
		224;Phosphoserine	
		310;Phosphoserine	
		722;Phosphothreonine	
		772;Phosphoserine	
		775;Phosphoserine	
		865;Phosphothreonine	
		873;Phosphoserine	
		951;Phosphoserine	
	Predicted disordered region	D167V: -	167-169;113-167;121-230;113-229;116-229;
			
		A923V: -	902-956;886-957;829-956;894-956;839-956;
			
Q14207 Protein NPAT	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/Q14207)	758-1427;NPAT_C	
	Post-translational modification site	775;Phosphoserine	
		779;Phosphoserine	

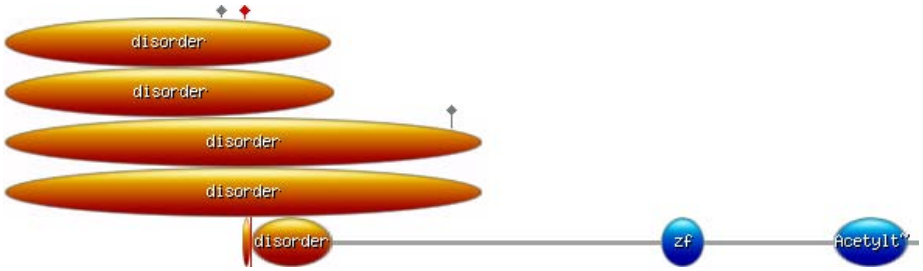
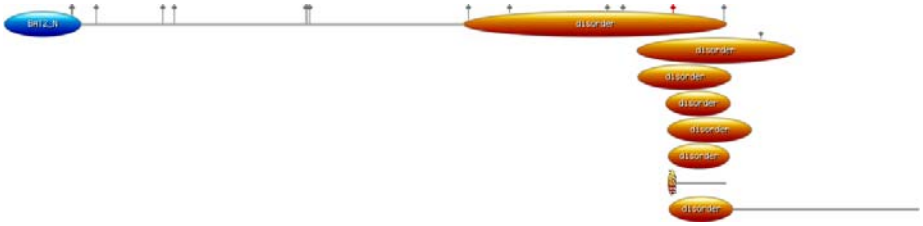
		1100;Phosphoserine	
		1228;N6-acetyllysine	
		1270;Phosphothreonine	
		1350;Phosphothreonine	
	Predicted disordered region	V608A: -	608-736;515-740;601-797;603-672;450-797;
			
Q15361 Transcription termination factor 1	Superfamily domain	663-684; Homeodomain-like	
		719-751; Homeodomain-like	
		612-673; Homeodomain-like	
	Pfam domain (http://pfam.xfam.org/protein/Q15361)	620-677;Myb_DNA-bind_6	
	Post-translational modification site	65;Phosphoserine	
		240;Phosphoserine	
		403;Phosphoserine	
		476;Phosphotyrosine	
		478;Phosphoserine	
		481;Phosphoserine	
		487;Phosphoserine	
		872;Phosphoserine	
	Predicted disordered region	A885V: -	876-885;869-889;835-891;874-890;836-886;
			
Q15572 TATA box-binding protein-associated factor RNA polymerase I subunit C	Superfamily domain	301-414; WD40 repeat-like	
	Pfam domain (http://pfam.xfam.org/protein/Q15572)		
	Post-translational modification site	848;Phosphoserine	
	Predicted disordered region	G523R: -	520-547;523-542;484-558;518-542;492-557;522-538;

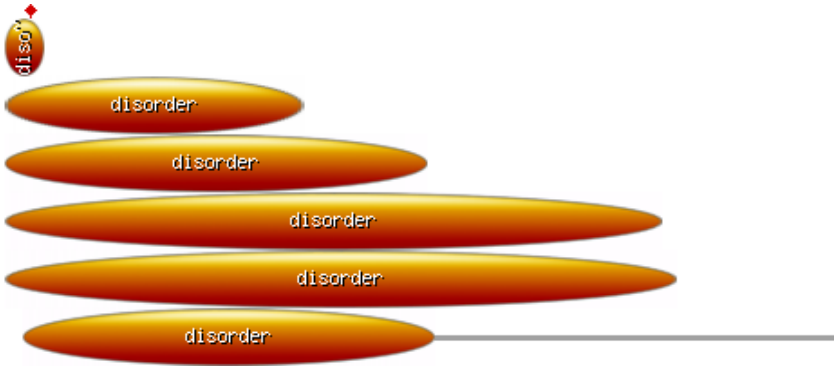
			
Q16534 Hepatic leukemia factor	Superfamily domain	223-286; Leucine zipper domain	
	Pfam domain (http://pfam.xfam.org/protein/Q16534)	224-277;bZIP_2	
	Post-translational modification site		
	Predicted disordered region	I253F: -	232-295;94-254;85-263;89-281;1-275;
			
Q17RF5 Uncharacterized protein C4orf26	Superfamily domain	15-58; Exotoxin A; middle domain	
	Pfam domain (http://pfam.xfam.org/protein/Q17RF5)	25-129;DUF4721	
	Post-translational modification site		
	Predicted disordered region	P30L: -	29-41;25-40;27-34;29-65;22-42;25-78;24-36;25-130;

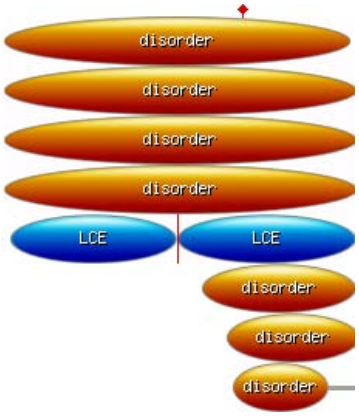
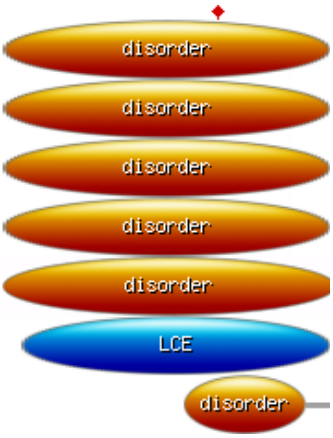
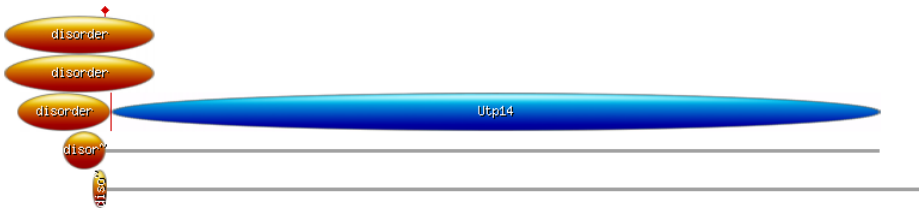
			
Q3B820 Protein FAM161A	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/Q3B820)	234-592;UPF0564	
	Post-translational modification site		
	Predicted disordered region	I236V: -	223-240;116-240;219-297;217-239;221-572;
			
Q3MHD2 Protein LSM12 homolog	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/Q3MHD2)	80-165;AD	
	Post-translational modification site	2;N-acetylalanine	
		75;Phosphothreonine	
	Predicted disordered region	V173L: -	171-195;172-196;162-195;161-195;164-195;

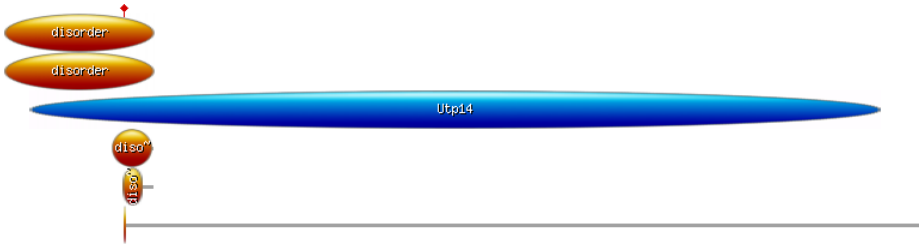
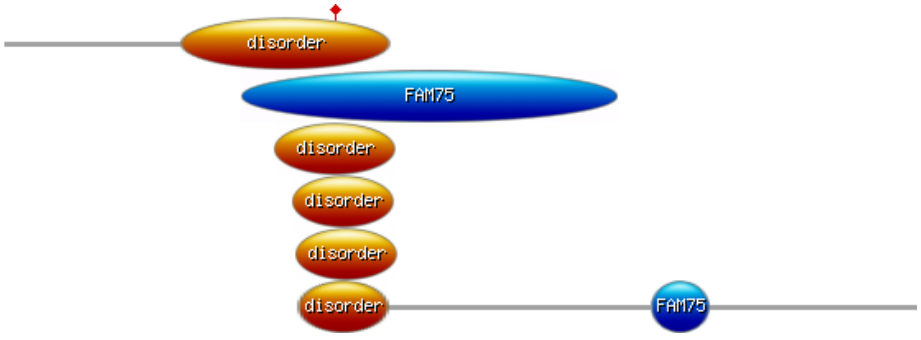
			
Q49AG3 Zinc finger BED domain-containing protein 5	Superfamily domain	271-664; Ribonuclease H-like	
		121-159; beta-beta-alpha zinc fingers	
	Pfam domain (http://pfam.xfam.org/protein/Q49AG3)	119-158;zf-BED	
	Post-translational modification site		
	Predicted disordered region	P77S: -	77-112;53-109;60-113;77-82;52-112;
			
Q4G0U5 Cilia- and flagella-associated protein 221	Superfamily domain	201-232; PapD-like	
	Pfam domain (http://pfam.xfam.org/protein/Q4G0U5)		
	Post-translational modification site		
	Predicted disordered region	V637I: -	623-652;614-650;589-648;600-643;589-645;
			
Q52M75 Putative uncharacterized protein encoded by LINC01554	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/Q52M75)		
	Post-translational modification site		
	Predicted disordered region	R85C: -	1-96;64-97;29-96;65-96;13-96;83-96;

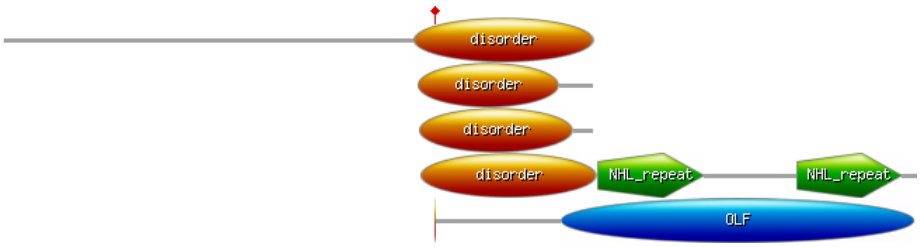
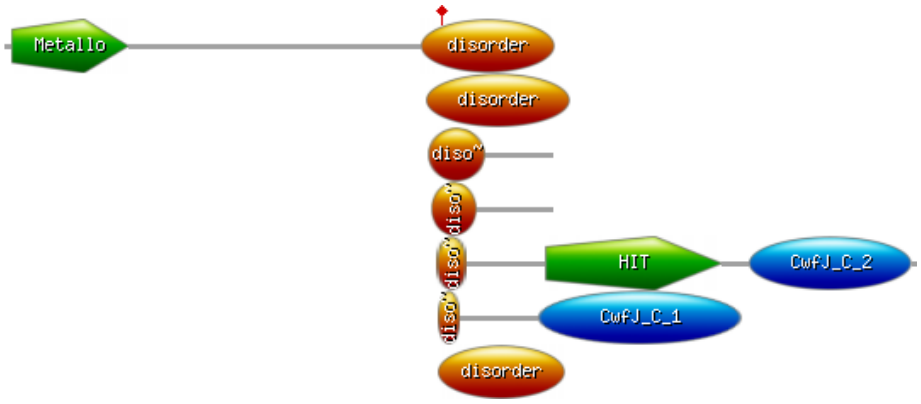
			
Q567U6 Coiled-coil domain-containing protein 93	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/Q567U6)	27-206;KOG2701	
	Post-translational modification site	298;Phosphoserine	
		301;Phosphoserine	
		305;Phosphoserine	
	Predicted disordered region	H315R: A colorectal cancer sample	295-315;301-316;291-323;283-336;301-327;174-466;
			
Q569K6 Coiled-coil domain-containing protein 157	Superfamily domain	496-615; Prefoldin	
	Pfam domain (http://pfam.xfam.org/protein/Q569K6)		
	Post-translational modification site		
	Predicted disordered region	P191L: -	191-191;134-215;128-214;178-193;128-209;
			
Q5FWF5 N-acetyltransferase ESCO1	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/Q5FWF5)	604-643;zf-C2H2_3	
		763-831;Acetyltransf_13	
	Post-translational modification site	200;Phosphoserine	
		412;Phosphoserine	

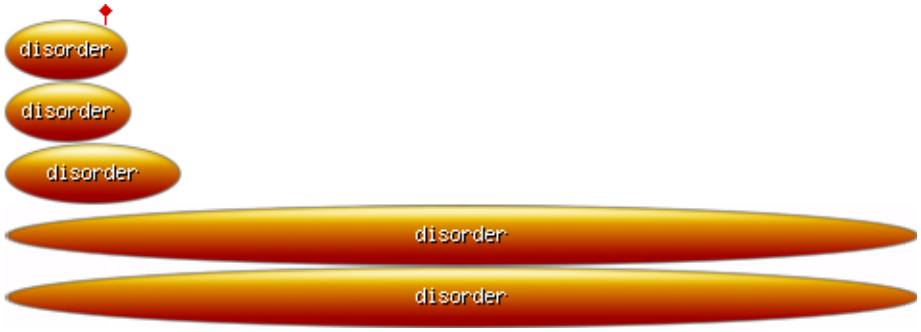
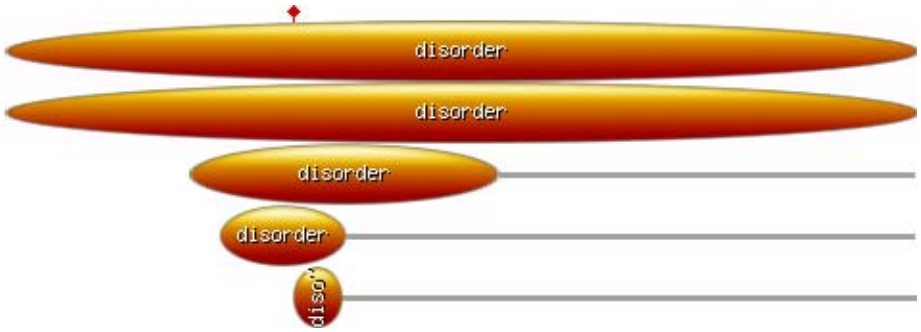
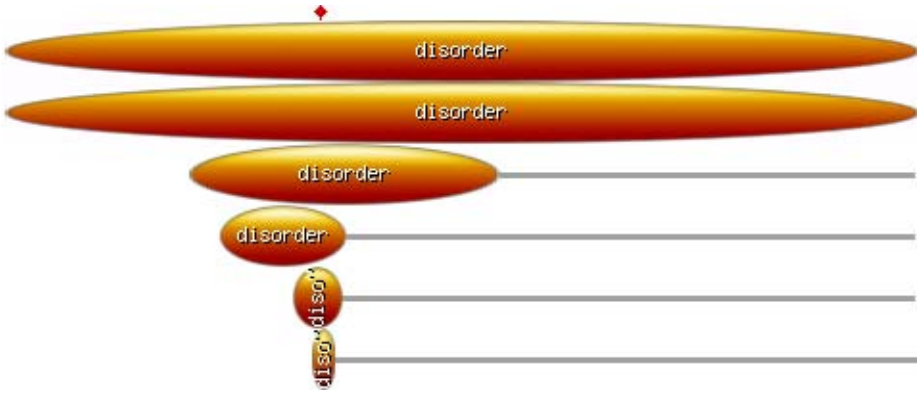
	Predicted disordered region	T221M: -	221-301;1-303;1-300;1-439;1-439;220-226;
			
Q5JSZ5 Protein PRRC2B	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/Q5JSZ5)	1-190;BAT2_N	
	Post-translational modification site	166;Phosphoserine	
		168;Phosphoserine	
		226;Phosphoserine	
		388;Phosphoserine	
		416;Phosphoserine	
		736;Phosphothreonine	
		740;Phosphoserine	
		745;Phosphoserine	
		1132;Phosphoserine	
		1231;Phosphoserine	
		1470;Phosphoserine	
		1507;Phosphoserine	
		1754;Phosphoserine	
		1843;Phosphoserine	
	Predicted disordered region	S1630T: -	1120-1761;1617-1767;1612-1771;1620-1777;1542-1926;1544-1773;1618-1637;1616-1823;
			
Q5SQ13 Proline-rich protein 31	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/Q5SQ13)		
	Post-translational modification site		
	Predicted disordered region	L8F: -	1-83;1-117;6-119;1-11;1-182;1-186;

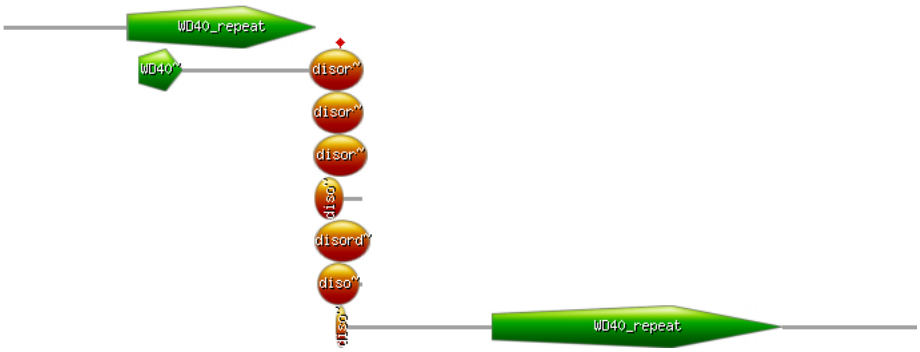
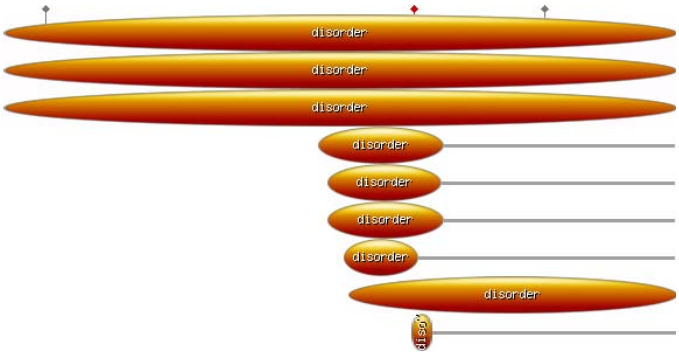
			
Q5SZD1 Uncharacterized protein C6orf141	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/Q5SZD1)		
	Post-translational modification site		
	Predicted disordered region	P235L: -	200-244;219-244;235-238;191-245;183-244;207-240;179-244;227-239;
			
Q5T752 Late cornified envelope protein 1D	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/Q5T752)	54-114;LCE	
		3-56;LCE	
	Post-translational modification site		
	Predicted disordered region	R78H: -	1-112;65-114;1-114;73-115;1-114;75-105;1-114;

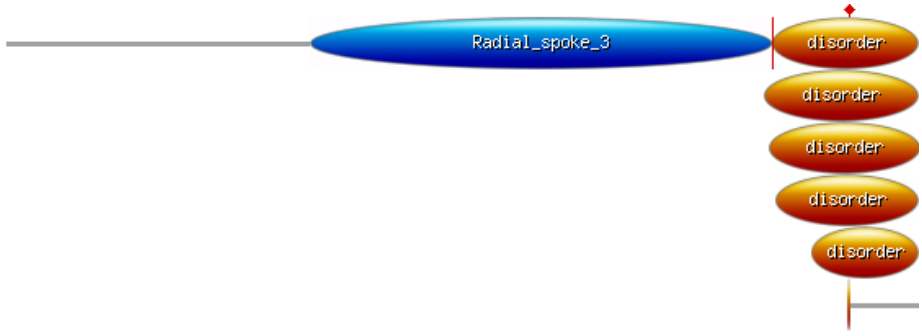
			
Q5TA76 Late cornified envelope protein 3A	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/Q5TA76)	6-89;LCE	
	Post-translational modification site		
	Predicted disordered region	R59C: -	1-89;1-89;1-90;1-89;50-82;1-89;
			
Q5TAP6 U3 small nucleolar RNA-associated protein 14 homolog C	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/Q5TAP6)	22-734;Utp14	
	Post-translational modification site		
	Predicted disordered region	G85V: -	75-86;50-85;1-126;12-89;1-126;
			
		T101A: -	101-102;91-124;1-126;100-116;1-126;

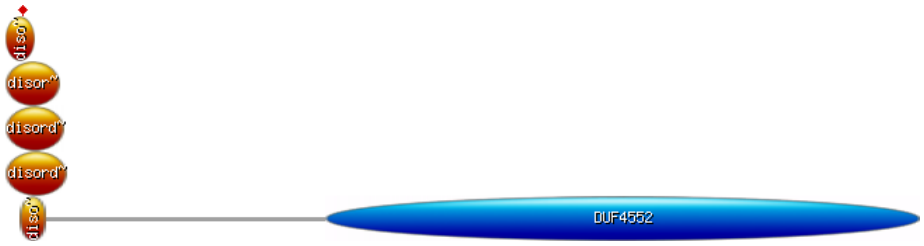
			
Q5VWN6 Protein FAM208B	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/Q5VWN6)	91-167;DUF3699	
		1052-1110;DUF3699	
		538-702;DUF3715	
		855-1033;DUF3715	
	Post-translational modification site	19;Phosphoserine	
		384;Phosphoserine	
		1087;Phosphoserine	
		2009;Phosphoserine	
		2037;Phosphoserine	
	Predicted disordered region	S724Y: -	542-731;669-724;671-727;701-730;301-740;709-734;303-744;703-726;
			
Q63HN1 Putative protein FAM205B	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/Q63HN1)	146-374;FAM75	
		395-430;FAM75	
	Post-translational modification site		
	Predicted disordered region	D203E: -	180-235;177-238;179-240;166-239;109-236;
			
Q68BL7	Superfamily domain	425-500; NHL repeat	

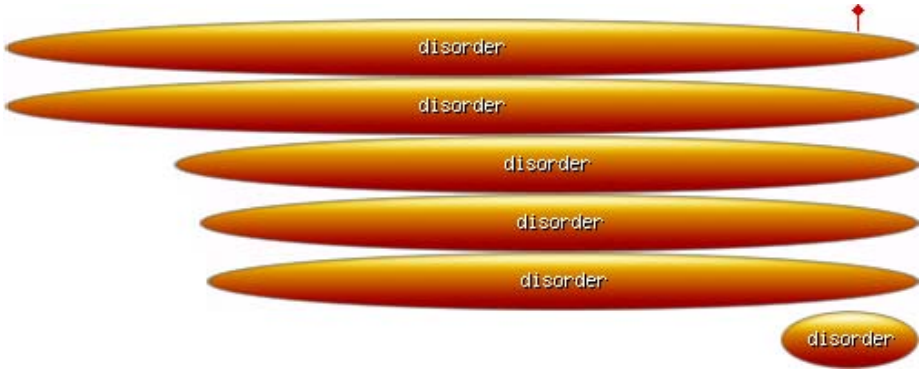
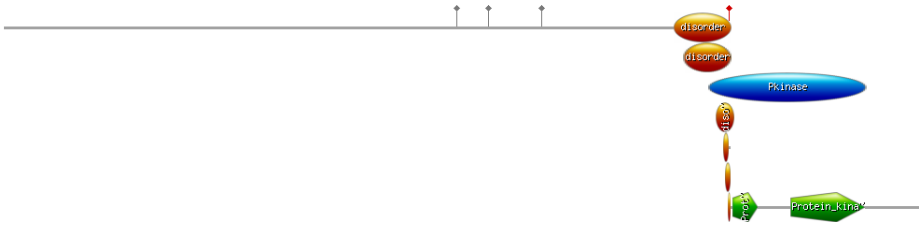
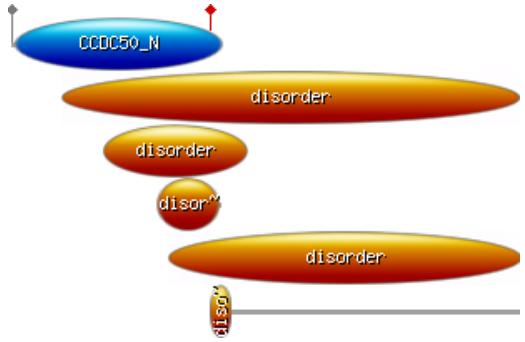
		567-641; NHL repeat	
	Pfam domain (http://pfam.xfam.org/protein/Q68BL7)	399-650;OLF	
	Post-translational modification site		
	Predicted disordered region	T309A: -	299-424;294-422;298-407;297-397;309-309;
			
Q69YN2 CWF19-like protein 1	Superfamily domain	320-423; HIT-like	
		6-74; Metallo-dependent phosphatases	
	Pfam domain (http://pfam.xfam.org/protein/Q69YN2)	440-535;CwfJ_C_2	
		316-435;CwfJ_C_1	
	Post-translational modification site		
	Predicted disordered region	P259L: -	256-273;251-284;257-269;247-325;250-334;253-279;257-331;
			
Q6L8H2 Keratin-associated protein 5-3	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/Q6L8H2)		
	Post-translational modification site		
	Predicted disordered region	G27S: -	1-238;1-33;1-32;1-46;1-238;

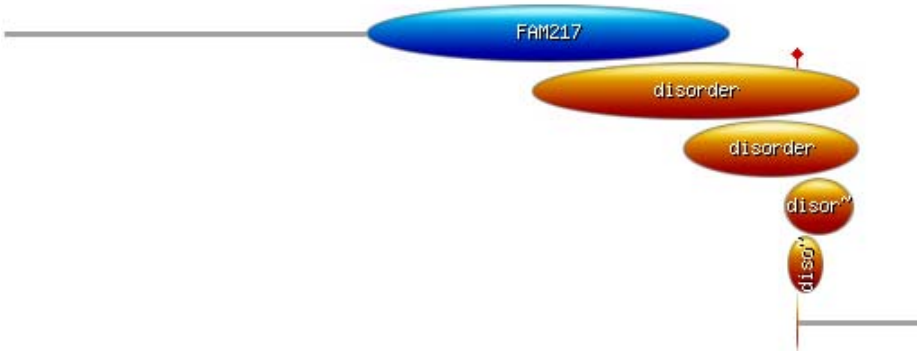
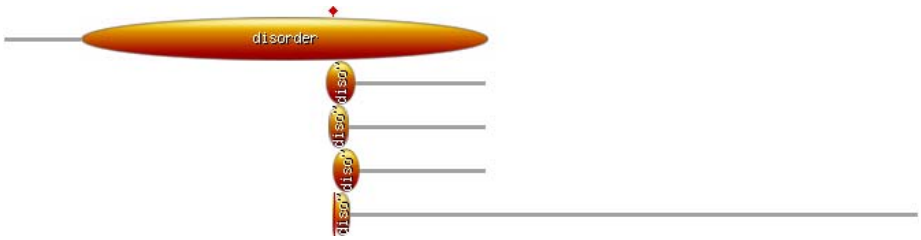
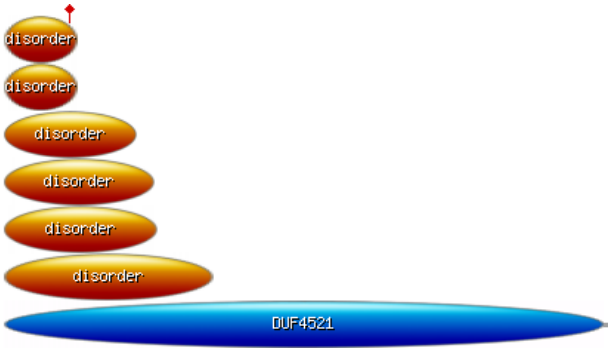
		
	Y28C: -	1-238;1-33;1-32;1-46;1-238;
		
	G76V: -	1-238;76-88;57-89;49-129;1-238;
		
	S83C: -	1-238;76-88;57-89;49-129;1-238;81-86;
		

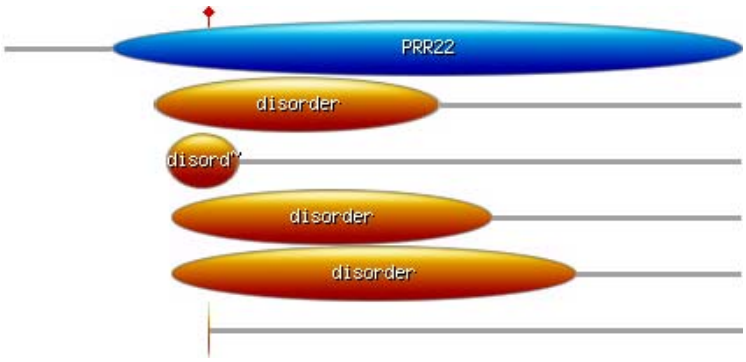
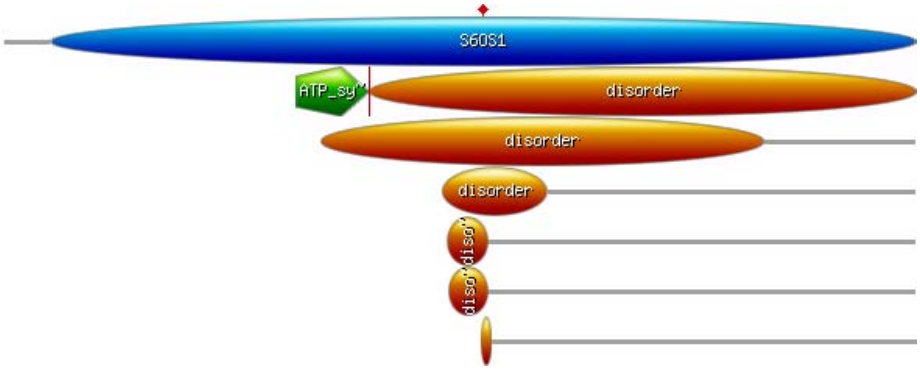
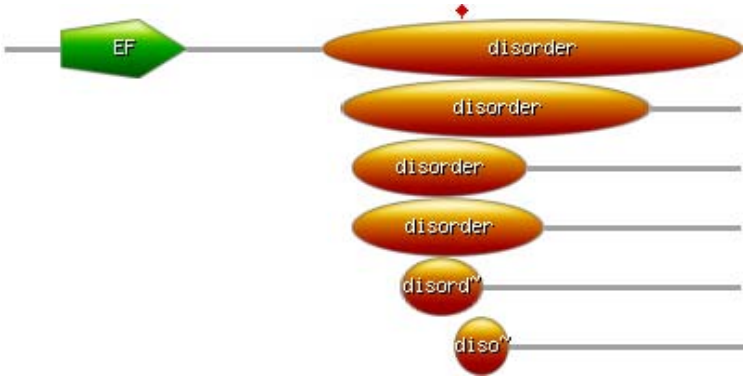
Q6P2C0 WD repeat-containing protein 93	Superfamily domain	103-135; WD40 repeat-like	
		368-585; WD40 repeat-like	
		94-235; WD40 repeat-like	
	Pfam domain (http://pfam.xfam.org/protein/Q6P2C0)		
	Post-translational modification site		
	Predicted disordered region	S254T: -	235-256;234-274;235-276;231-271;237-268;233-271;251-258;
			
Q6PK04 Coiled-coil domain-containing protein 137	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/Q6PK04)		
	Post-translational modification site	19;Phosphoserine	
		233;Phosphoserine	
	Predicted disordered region	R177W: -	1-289;176-184;140-188;136-189;147-178;140-189;1-289;149-289;1-289;
			
Q6ZVD7 Storkhead-box protein 1	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/Q6ZVD7)	111-189;Stork_head	
	Post-translational modification site		
	Predicted disordered region	N825I: Pre-eclampsia/eclampsia 4 (PEE4) [MIM:609404]	798-827;758-855;682-826;804-828;684-843;823-829;

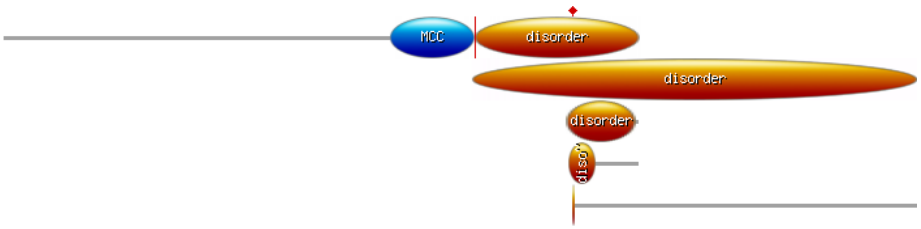
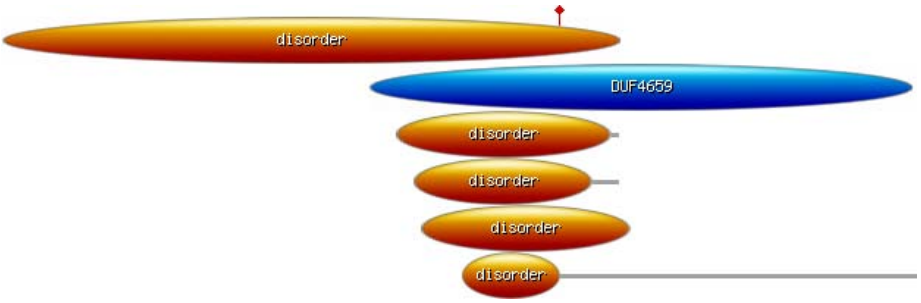
			
Q7Z570 Zinc finger protein 804A	Superfamily domain	56-93; beta-beta-alpha zinc fingers	
	Pfam domain (http://pfam.xfam.org/protein/Q7Z570)	56-84;zf-C2H2_jaz	
	Post-translational modification site		
	Predicted disordered region	Q261L: -	261-281;232-423;219-308;219-330;261-264;
			
		G1152R: -	1148-1154;1063-1209;1141-1177;1063-1209;1147-1154;
			
Q86UC2 Radial spoke head protein 3 homolog	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/Q86UC2)	188-470;Radial_spoke_3	
	Post-translational modification site		
	Predicted disordered region	G518D: -	495-560;473-560;517-518;469-561;466-560;466-560;
			
Q86V48 Leucine zipper protein 1	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/Q86V48)		
	Post-translational modification site	2;N-acetylalanine	

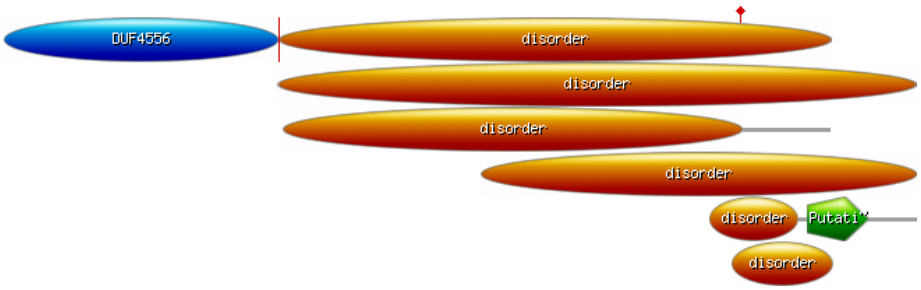
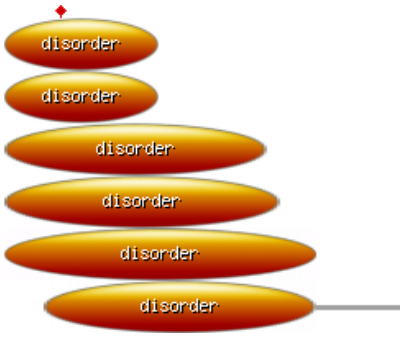
		394;Phosphoserine	
		570;Phosphoserine	
		574;Phosphoserine	
		611;Phosphoserine	
		659;Phosphoserine	
		679;Phosphothreonine	
		690;Phosphoserine	
		745;Phosphoserine	
	Predicted disordered region	S1034N: -	1023-1076;1008-1076;927-1077;1022-1076;1017-1076;
			
Q86WS4 Uncharacterized protein C12orf40	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/Q86WS4)	229-652;DUF4552	
	Post-translational modification site		
	Predicted disordered region	I13L: -	1-21;1-44;1-39;11-29;1-42;
			
Q86X51 Uncharacterized protein CXorf67	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/Q86X51)		
	Post-translational modification site		
	Predicted disordered region	R470K: -	1-503;1-503;94-503;108-503;428-503;112-503;

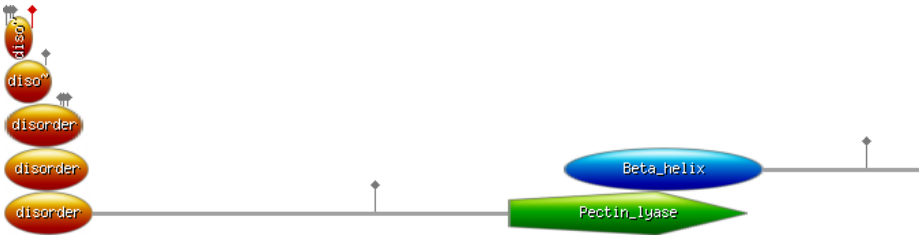
			
Q86YV5 Tyrosine-protein kinase SgK223	Superfamily domain	1117-1154; Protein kinase-like (PK-like)	
		1206-1319; Protein kinase-like (PK-like)	
	Pfam domain (http://pfam.xfam.org/protein/Q86YV5)	1080-1322;Pkinase	
	Post-translational modification site	694;Phosphoserine	
		743;Phosphoserine	
		824;Phosphoserine	
	Predicted disordered region	A1111T: -	1110-1113;1092-1120;1105-1113;1042-1115;1103-1111;1027-1115;
			
Q8IVM0 Coiled-coil domain-containing protein 50	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/Q8IVM0)	4-128;CCDC50_N	
	Post-translational modification site	2;N-acetylalanine	
	Predicted disordered region	L121F: -	121-133;96-307;57-143;89-126;32-306;
			
Q8IXS0 Protein FAM217A	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/Q8IXS0)	203-404;FAM217	
	Post-translational modification site		

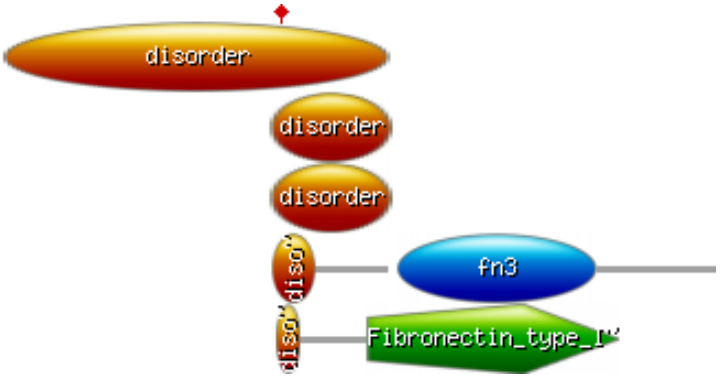
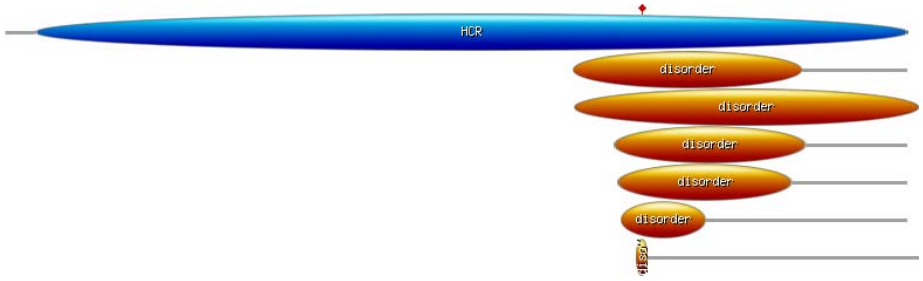
	Predicted disordered region	M442V: -	437-456;435-473;379-476;295-476;442-442;
			
Q8IYE0 Coiled-coil domain-containing protein 146	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/Q8IYE0)		
	Post-translational modification site		
	Predicted disordered region	N345S: -	345-362;345-345;340-360;337-367;344-371;81-505;
			
Q8IYI0 Uncharacterized protein C20orf196	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/Q8IYI0)	1-203;DUF4521	
	Post-translational modification site		
	Predicted disordered region	A23V: -	1-51;1-52;1-25;1-25;1-45;1-71;
			
Q8IZ63 Proline-rich protein 22	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/Q8IZ63)	63-422;PRR22	
	Post-translational modification site		
	Predicted disordered region	P118L: -	87-250;118-

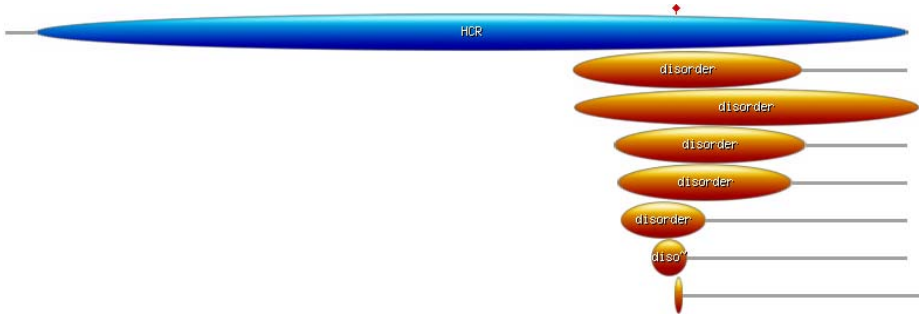
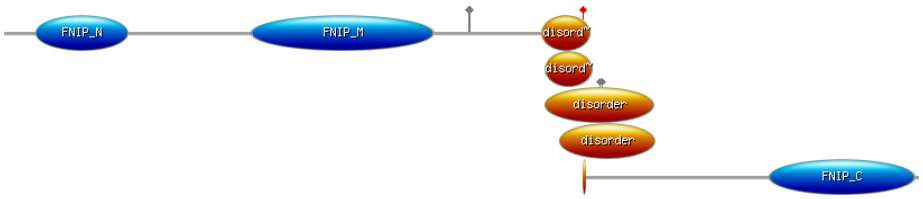
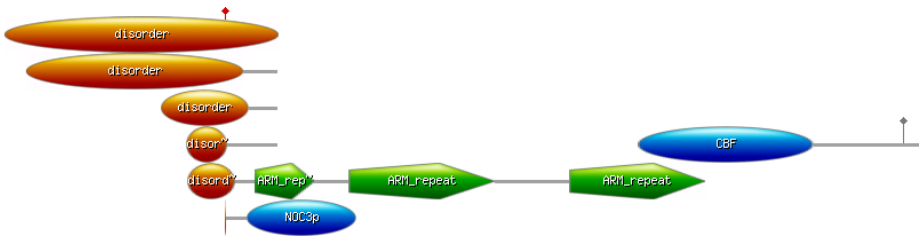
			118;97-279;94-135;97-327;
			
Q8N1H7 Protein SIX6OS1	Superfamily domain	189-235; ATP synthase B chain-like	
	Pfam domain (http://pfam.xfam.org/protein/Q8N1H7)	31-587;S6OS1	
	Post-translational modification site		
	Predicted disordered region	S309L: -	287-312;308-314;283-350;205-489;286-312;198-587;
			
Q8N4Y2 EF-hand calcium-binding domain-containing protein 4A	Superfamily domain	32-99; EF-hand	
	Pfam domain (http://pfam.xfam.org/protein/Q8N4Y2)		
	Post-translational modification site		
	Predicted disordered region	S248I: -	189-283;215-259;189-292;183-350;244-273;173-399;
			

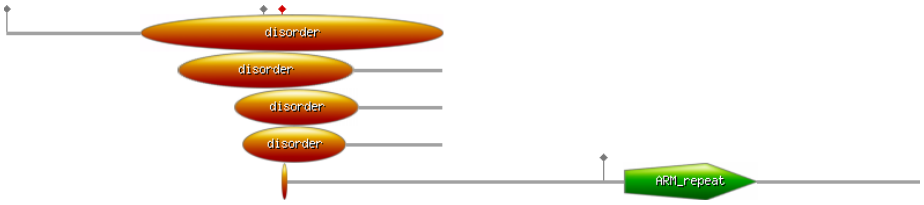
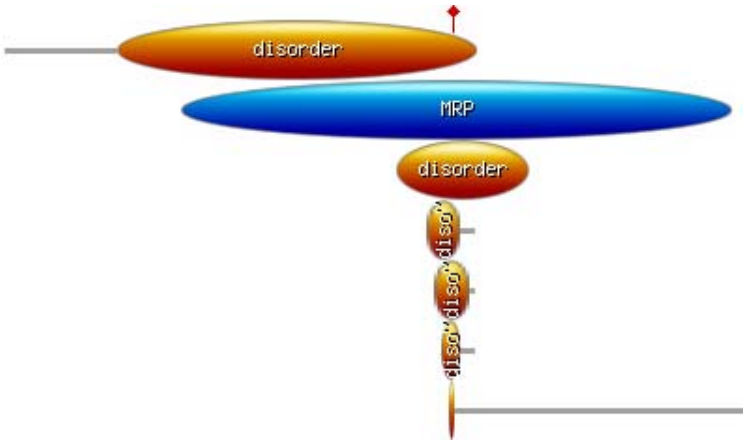
Q8N6Y0 Usher syndrome type-1C protein-binding protein 1	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/Q8N6Y0)	299-363;MCC-bdg_PDZ	
	Post-translational modification site		
	Predicted disordered region	M439V: -	436-456;439-440;362-490;434-487;362-703;
			
Q8N715 Coiled-coil domain-containing protein 185	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/Q8N715)	251-620;DUF4659	
	Post-translational modification site		
	Predicted disordered region	R380L: -	314-380;286-428;281-402;269-415;1-421;
			
Q8N7X0 Androglobin	Superfamily domain	37-294; Cysteine proteinases	
		784-854; Globin-like	
		627-659; Cysteine proteinases	
	Pfam domain (http://pfam.xfam.org/protein/Q8N7X0)	178-319;Peptidase_C2	
	Post-translational modification site		
	Predicted disordered region	T1637A: -	1636-1667;1633-1668;1635-1667;1636-1638;1394-1667;
			
Q8N9H9 Uncharacterized protein C1orf127	Superfamily domain	578-621; Putative DNA-binding domain	
	Pfam domain (http://pfam.xfam.org/protein/Q8N9H9)	2-198;DUF4556	
	Post-translational modification site		
	Predicted disordered region	A530V: -	344-656;524-596;202-531;198-656;508-571;198-595;

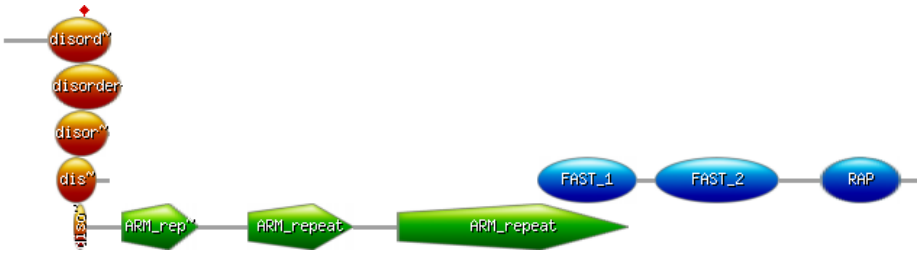
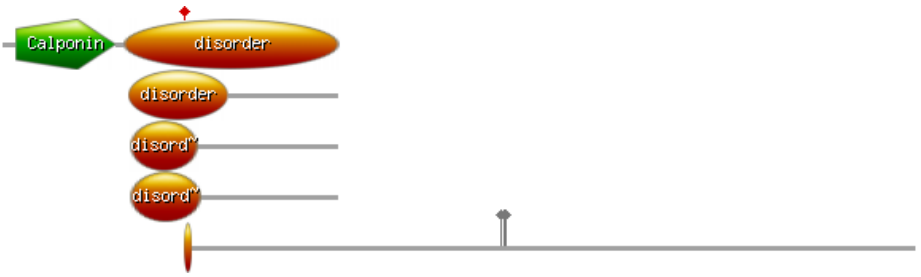
			
Q8N9K7 Uncharacterized protein KIAA1456 isoform	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/Q8N9K7)		
	Post-translational modification site		
	Predicted disordered region	Q18H: -	13-95;1-80;1-47;1-47;1-84;1-95;
			
Q8NEF3 Coiled-coil domain-containing protein 112	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/Q8NEF3)		
	Post-translational modification site		
	Predicted disordered region	H32L: -	1-34;1-49;30-55;21-48;1-111;
			
Q8NEM2 SHC SH2 domain-binding protein 1	Superfamily domain	371-546; Pectin lyase-like	
	Pfam domain (http://pfam.xfam.org/protein/Q8NEM2)	412-558;Beta_helix	
	Post-translational modification site	2;N-acetylalanine	
		5;Phosphoserine	
		7;Phosphothreonine	
		31;Phosphoserine	

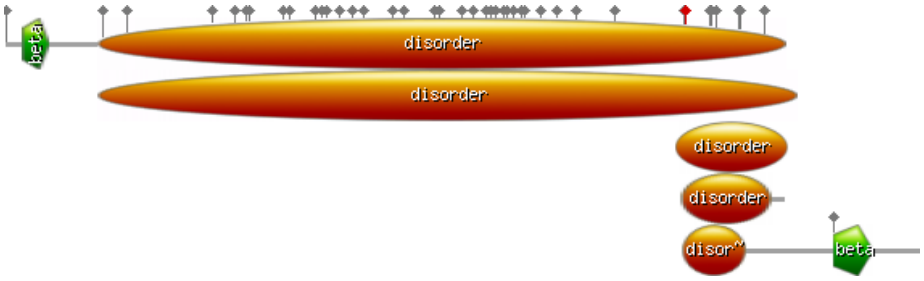
		42;Phosphoserine	
		44;Phosphoserine	
		47;Phosphoserine	
		273;Phosphoserine	
		634;Phosphoserine	
	Predicted disordered region	M21T: -	1-21;1-65;1-58;1-35;1-62;
			
Q8NEV8 Exophilin-5	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/Q8NEV8)		
	Post-translational modification site		
	Predicted disordered region	R118L: -	85-271;93-118;73-123;80-162;80-159;
			
		M512L: -	463-581;512-513;512-581;496-556;508-519;323-833;
			
Q8TBZ0 Coiled-coil domain-containing protein 110	Superfamily domain	708-792; 4-helical cytokines	
	Pfam domain (http://pfam.xfam.org/protein/Q8TBZ0)		
	Post-translational modification site		
	Predicted disordered region	S817L: -	793-833;803-821;801-818;817-820;802-834;779-833;775-833;

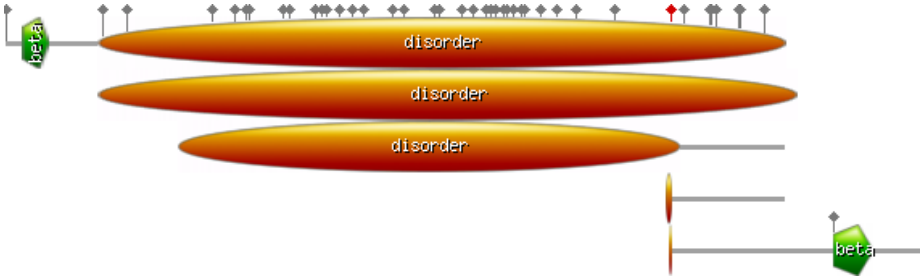
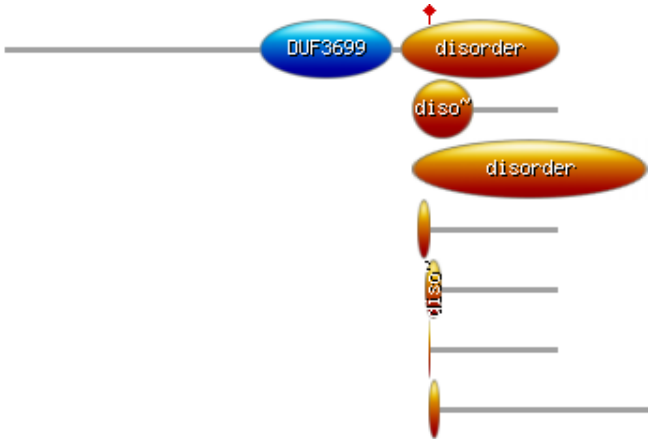
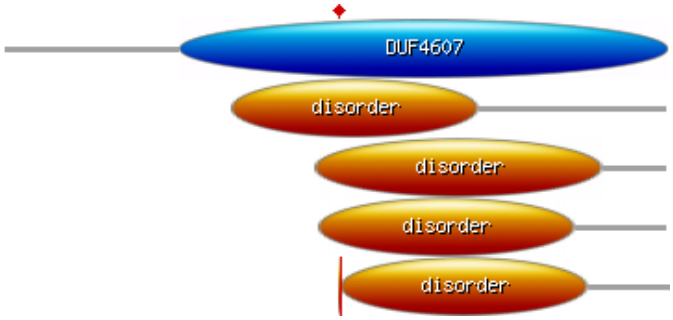
			
Q8TC99 Fibronectin type III domain-containing protein 8	Superfamily domain	166-280; Fibronectin type III	
	Pfam domain (http://pfam.xfam.org/protein/Q8TC99)	180-270;fn3	
	Post-translational modification site		
	Predicted disordered region	A127T: -	125-135;122-177;122-177;123-142;1-176;
			
Q8TD31 Coiled-coil alpha-helical rod protein 1	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/Q8TD31)	27-773;HCR	
	Post-translational modification site		
	Predicted disordered region	K546R: -	528-600;541-550;525-674;487-682;522-686;488-782;
			
		G575C: -	528-600;554-584;525-674;487-682;522-686;488-782;574-580;

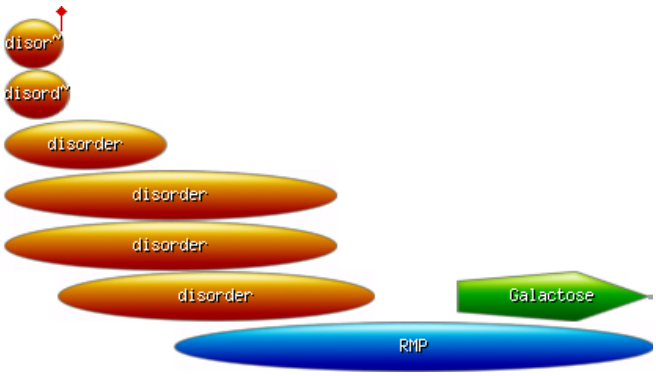
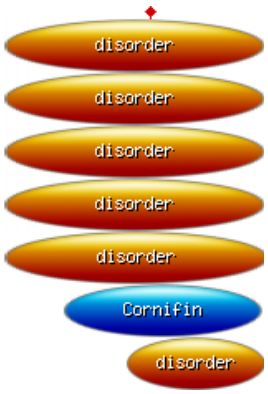
			
Q8TF40 Folliculin-interacting protein 1	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/Q8TF40)	41-159;FNIP_N	
		316-549;FNIP_M	
		976-1162;FNIP_C	
	Post-translational modification site	593;Phosphoserine	
		594;Phosphoserine	
		760;Phosphoserine	
		763;Phosphoserine	
	Predicted disordered region	V738L: -	709-831;738-741;686-748;690-829;690-750;
			
Q8WTT2 Nucleolar complex protein 3 homolog	Superfamily domain	220-271; ARM repeat	
		302-429; ARM repeat	
		495-613; ARM repeat	
	Pfam domain (http://pfam.xfam.org/protein/Q8WTT2)	555-708;CBF	
		213-308;NOC3p	
	Post-translational modification site	787;Phosphoserine	
	Predicted disordered region	P194L: -	194-194;160-195;20-209;138-214;161-202;1-240;
			
Q8WXE1 ATR-interacting protein	Superfamily domain	536-650; ARM repeat	
	Pfam domain (http://pfam.xfam.org/protein/Q8WXE1)		
	Post-translational modification site	2;N-acetylalanine	

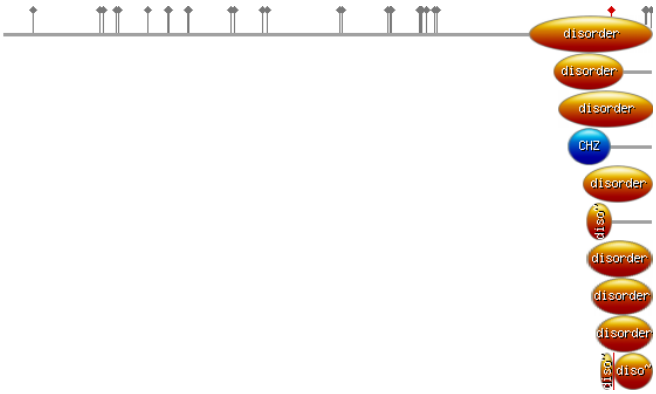
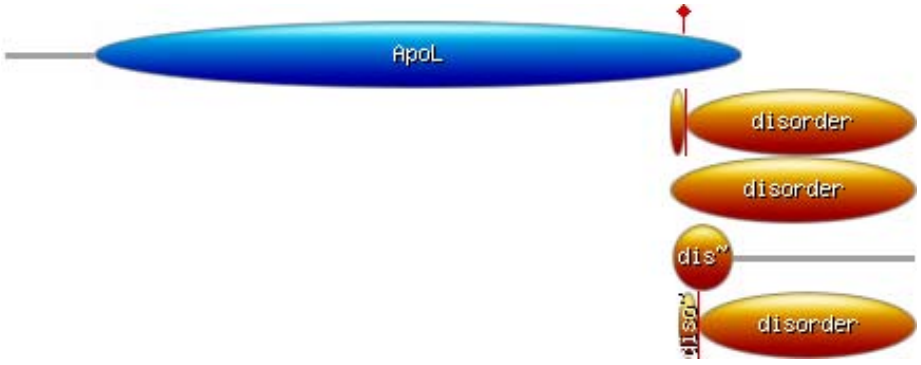
		224;Phosphoserine	
		518;Phosphoserine	
	Predicted disordered region	P240L: -	206-295;199-306;150-302;118-379;240-244;
			
Q8WYQ9	Superfamily domain	97-216; PX domain	
		298-352; SAM/Pointed domain	
		895-925; Retrovirus zinc finger-like domains	
	Pfam domain (http://pfam.xfam.org/protein/Q8WYQ9)	906-923;zf-CCHC	
	Post-translational modification site		
	Predicted disordered region	I54V: -	1-95;1-95;24-100;1-114;1-117;
			
Q92665 Zinc finger CCHC domain-containing protein 14	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/Q92665)	96-389;MRP-S31	
	Post-translational modification site		
	Predicted disordered region	T241M: -	227-244;235-244;239-241;211-281;62-253;231-249;
			
Q969Z0 Protein TBRG4	Superfamily domain	83-135; ARM repeat	
		170-242; ARM repeat	

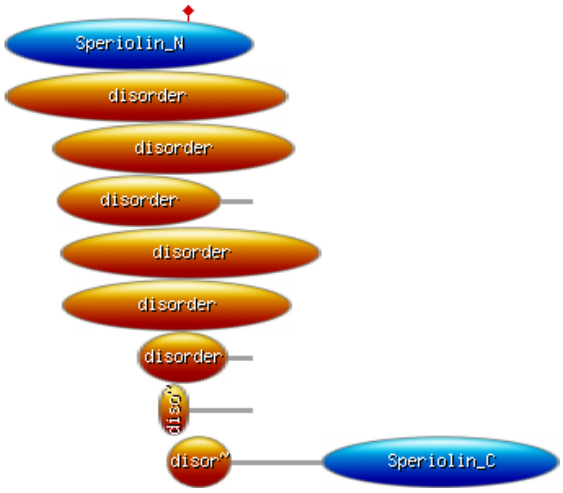
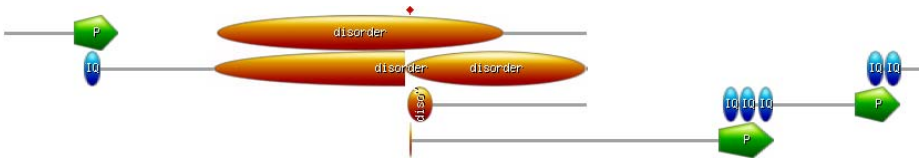
		273-432; ARM repeat	
	Pfam domain (http://pfam.xfam.org/protein/Q969Z0)	451-536;FAST_2	
		565-620;RAP	
		370-438;FAST_1	
	Post-translational modification site		
	Predicted disordered region	P57L: -	50-58;32-75;35-82;38-65;37-73;
			
Q96GE4 Centrosomal protein of 95 kDa	Superfamily domain	13-102; Calponin-homology domain; CH-domain	
	Pfam domain (http://pfam.xfam.org/protein/Q96GE4) Post-translational modification site		
		449;Phosphoserine	
		451;Phosphoserine	
		453;Phosphoserine	
	Predicted disordered region	M165I: -	114-203;115-178;115-175;110-303;164-170;
			
Q96JM3 Chromosome alignment-maintaining phosphoprotein 1	Superfamily domain	736-771; beta-beta-alpha zinc fingers	
		15-38; beta-beta-alpha zinc fingers	
	Pfam domain (http://pfam.xfam.org/protein/Q96JM3)		
	Post-translational modification site	1;N-acetylmethionine	
		87;Phosphoserine	
		108;Phosphoserine	
		184;Phosphoserine	
		204;Phosphoserine	
		214;Phosphoserine	
		217;Phosphoserine	
		247;Phosphoserine	
		253;Phosphoserine	
		275;Phosphoserine	
		282;Phosphoserine	

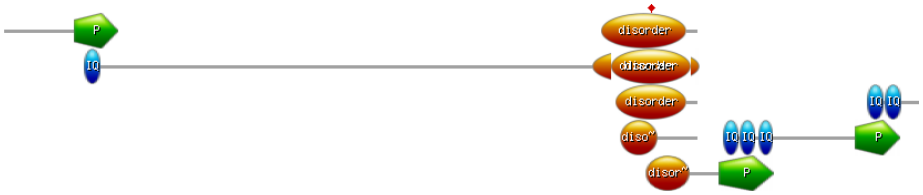
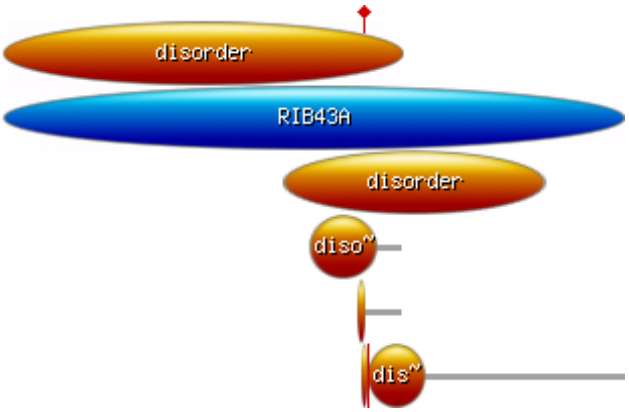
		286;Phosphoserine	
		297;Phosphoserine	
		308;Phosphoserine	
		319;Phosphoserine	
		344;Phosphoserine	
		355;Phosphoserine	
		382;Phosphoserine	
		386;Phosphoserine	
		405;Phosphoserine	
		416;Phosphoserine	
		427;Phosphoserine	
		432;Phosphoserine	
		436;Phosphoserine	
		443;Phosphoserine	
		445;Phosphoserine	
		452;Phosphoserine	
		459;Phosphoserine	
		462;Phosphoserine	
		476;Phosphoserine	
		490;N6-acetyllysine	
		507;Phosphoserine	
		542;Phosphoserine	
		603;Phosphoserine	
		626;Phosphoserine	
		627;Phosphoserine	
		632;Phosphoserine	
		651;Phosphoserine	
		652;Phosphoserine	
		653;Phosphoserine	
		675;Phosphoserine	
		736;Phosphoserine	
	Predicted disordered region	P604R: -	601-657;600-681;83-694;595-694;83-704;
			
		K591R: -	154-599;587-591;83-694;590-592;83-704;

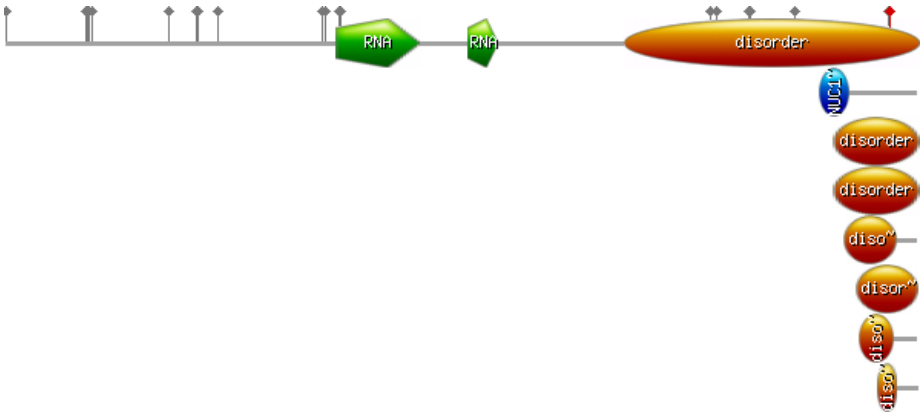
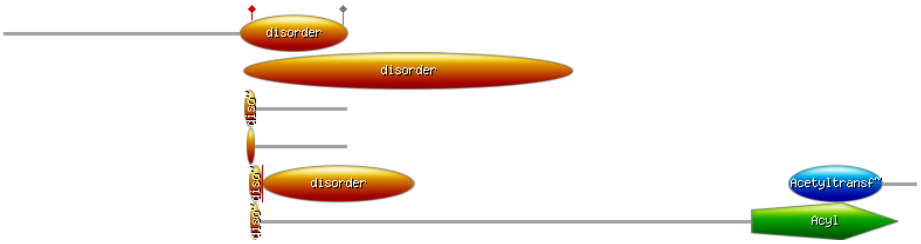
			
Q96KD3 Protein FAM71F1	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/Q96KD3)	138-208;DUF3699	
	Post-translational modification site		
	Predicted disordered region	S228L: -	222-228;228-233;228-228;213-297;219-251;226-234;219-344;
			
Q96LP6 Uncharacterized protein C12orf42	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/Q96LP6)	96-359;DUF4607	
	Post-translational modification site		
	Predicted disordered region	P182R: -	124-257;169-324;182-316;182-182;171-309;
			
Q96NL8 Protein C8orf37	Superfamily domain	145-205; Galactose-binding domain-like	
	Pfam domain (http://pfam.xfam.org/protein/Q96NL8)	55-207;RMP	
	Post-translational modification site		
	Predicted disordered region	P19A: -	1-19;1-52;18-118;1-106;1-21;1-106;

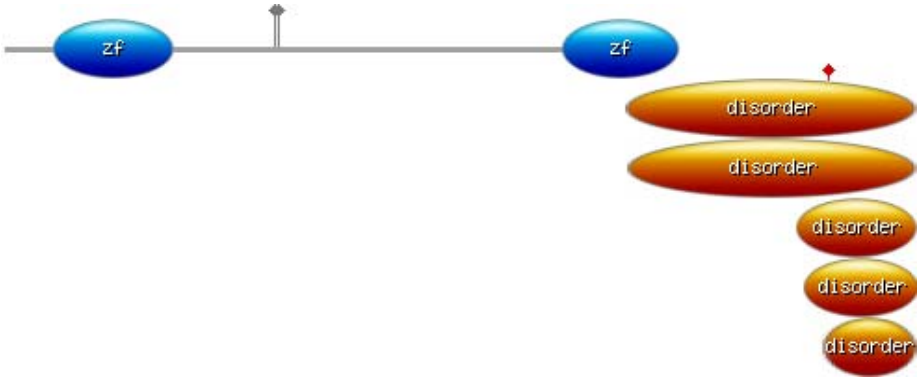
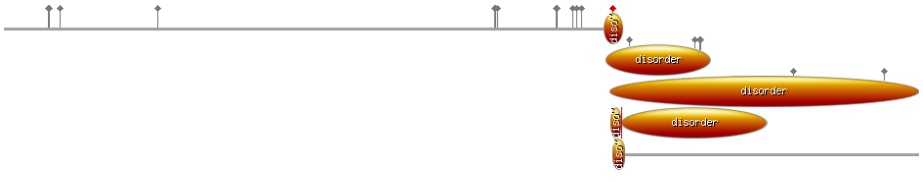
			
Q96PI1	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/Q96PI1)	19-78;Cornifin	
	Post-translational modification site		
	Predicted disordered region	P45S: -	1-79;1-79;38-79;1-80;1-79;1-79;
			
Q9BW71 Small proline-rich protein 4	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/Q9BW71)	484-520;CHZ	
	Post-translational modification site	27;Phosphoserine	
		84;Phosphothreonine	
		87;Phosphoserine	
		98;Phosphoserine	
		100;Phosphoserine	
		125;Phosphoserine	
		142;Phosphoserine	
		143;Phosphoserine	
		159;Phosphoserine	
		160;Phosphoserine	
		196;Phosphoserine	
		199;Phosphoserine	
		223;Phosphoserine	
		227;Phosphoserine	
		289;Phosphoserine	

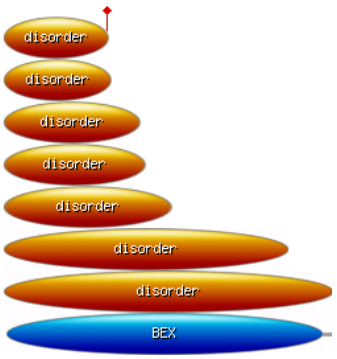
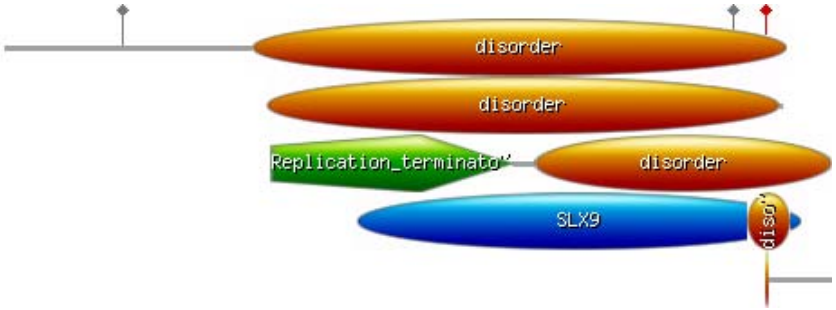
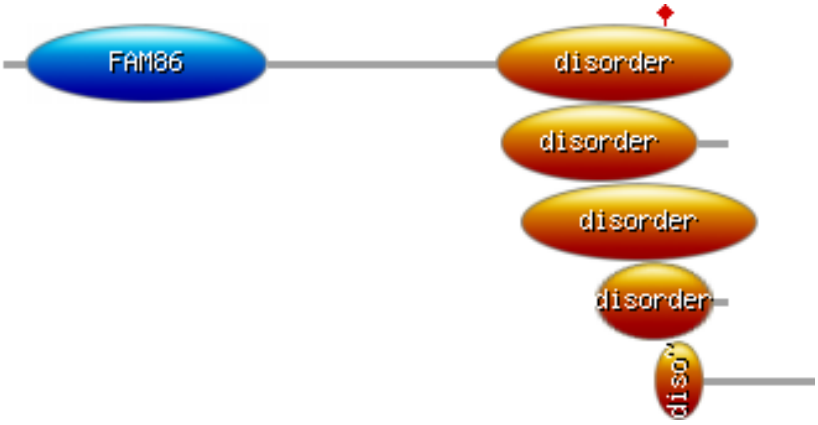
		291;Phosphoserine	
		330;Phosphoserine	
		332;Phosphoserine	
		333;Phosphoserine	
		357;Phosphoserine	
		358;Phosphothreonine	
		359;Phosphoserine	
		363;Phosphoserine	
		370;Phosphoserine	
		372;Phosphoserine	
		550;Phosphoserine	
		551;Phosphoserine	
		555;Phosphoserine	
	Predicted disordered region	G521W: -	500-556;508-556;500-521;504-556;521-556;476-557;497-556;472-531;451-556;512-522;
			
Q9BWW9 Apolipoprotein L5	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/Q9BWW9)	44-350;ApoL	
	Post-translational modification site		
	Predicted disordered region	T323M: -	317-323;322-433;321-329;317-433;318-346;317-433;
			

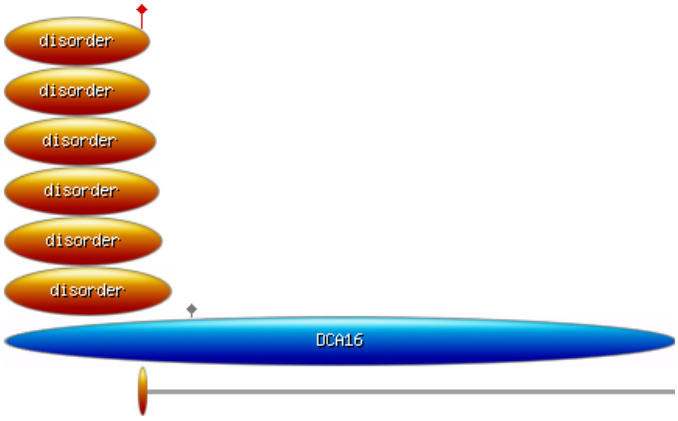
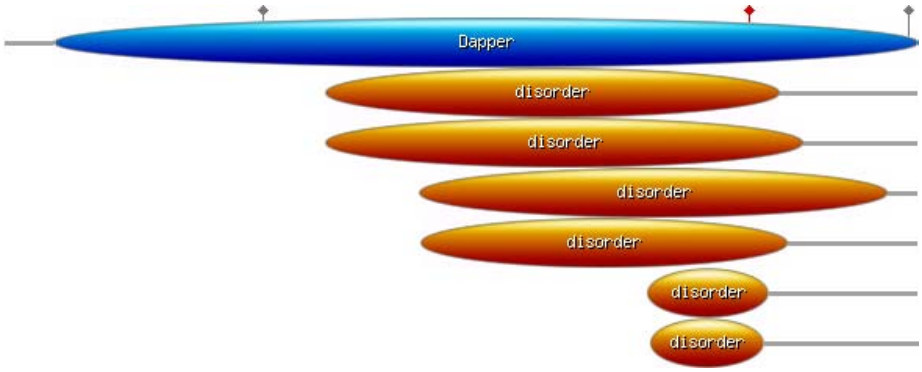
Q9H0A9 Speriolin-like protein	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/Q9H0A9)	1-153;Speriolin_N	
		195-340;Speriolin_C	
	Post-translational modification site		
	Predicted disordered region	P113L: -	33-133;36-176;82-137;100-139;35-194;30-178;1-174;95-113;
			
Q9H0B3 Uncharacterized protein KIAA1683	Superfamily domain	923-993; P-loop containing nucleoside triphosphate hydrolases	
		1099-1157; P-loop containing nucleoside triphosphate hydrolases	
		91-147; P-loop containing nucleoside triphosphate hydrolases	
	Pfam domain (http://pfam.xfam.org/protein/Q9H0B3)	928-947;IQ	
		104-124;IQ	
		950-969;IQ	
		1114-1134;IQ	
		974-992;IQ	
		1137-1157;IQ	
	Post-translational modification site		
	Predicted disordered region	T524A: -	524-526;518-751;272-754;521-553;276-645;
			
		P835L: -	829-885;784-886;790-881;772-881;795-842;760-897;

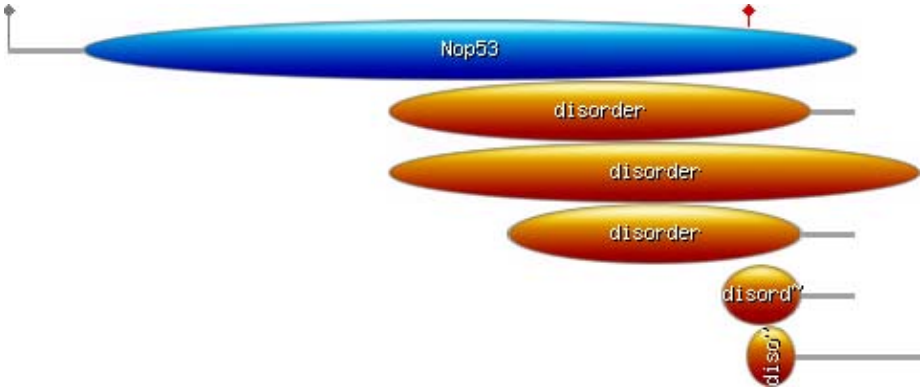
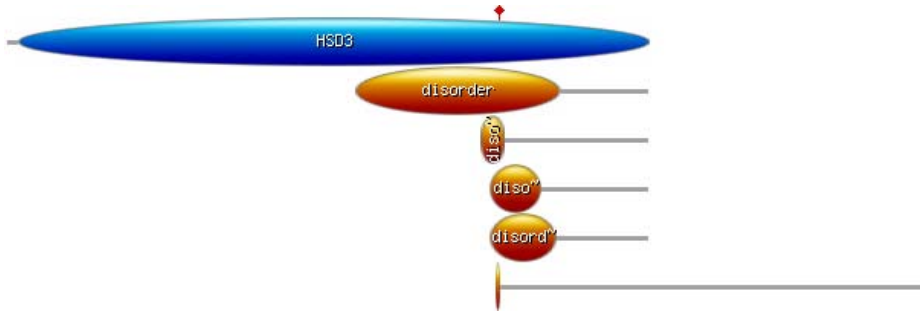
			
Q9H4K1 RIB43A-like with coiled-coils protein 2	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/Q9H4K1)	1-309;RIB43A	
	Post-translational modification site		
	Predicted disordered region	R180C: -	179-210;177-180;179-181;140-270;153-186;1-199;
			
Q9H501 ESF1 homolog	Superfamily domain	309-386; RNA-binding domain; RBD	
		432-459; RNA-binding domain; RBD	
	Pfam domain (http://pfam.xfam.org/protein/Q9H501)	759-787;NUC153	
	Post-translational modification site	2;N-acetylserine	
		75;Phosphoserine	
		77;Phosphoserine	
		79;Phosphoserine	
		82;Phosphoserine	
		153;Phosphoserine	
		179;Phosphoserine	
		180;Phosphoserine	
		198;Phosphoserine	
		296;Phosphoserine	
		298;Phosphoserine	
		311;Phosphothreonine	
		312;Phosphoserine	
		313;Phosphoserine	
		657;Phosphoserine	
		663;Phosphoserine	
		693;Phosphothreonine	
		694;Phosphoserine	
		735;Phosphoserine	

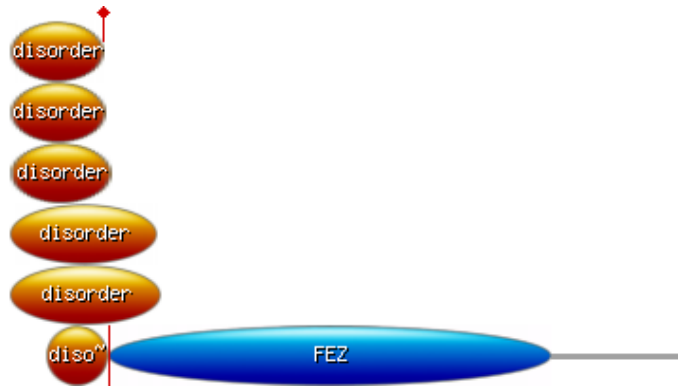
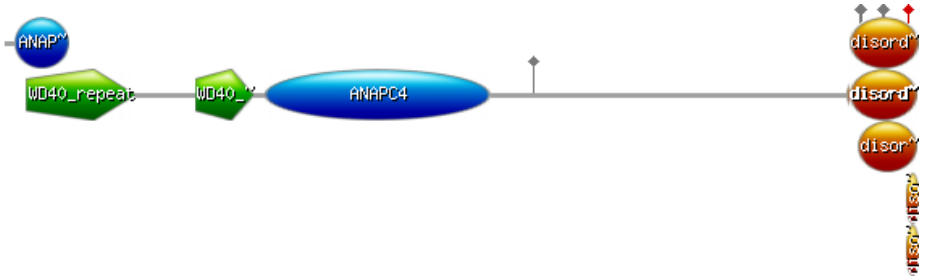
		823;Phosphoserine	
	Predicted disordered region	I824L: -	577-851;793-851;795-827;782-831;771-851;813-830;771-851;
			
Q9H8E8 Cysteine-rich protein 2-binding protein	Superfamily domain	641-766; Acyl-CoA N-acyltransferases (Nat)	
	Pfam domain (http://pfam.xfam.org/protein/Q9H8E8)	673-753;Acetyltransf_7	
	Post-translational modification site	292;N6-acetyllysine	
	Predicted disordered region	P214L: -	208-217;212-222;213-220;212-353;204-296;207-488;210-216;
			
Q9H9L4 KAT8 regulatory NSL complex subunit 2	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/Q9H9L4)	28-92;zf-C3Hc3H	
		302-364;zf-C3Hc3H	
	Post-translational modification site	147;Phosphoserine	
		149;Phosphoserine	
	Predicted disordered region	P445T: -	442-492;432-493;337-492;428-492;336-492;

			
Q9HAW4 Claspin	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/Q9HAW4)		
	Post-translational modification site	65;Phosphoserine	
		67;Phosphoserine	
		83;Phosphoserine	
		225;Phosphoserine	
		718;Phosphoserine	
		720;Phosphoserine	
		723;Phosphoserine	
		808;Phosphoserine	
		810;Phosphoserine	
		833;Phosphoserine	
		839;Phosphoserine	
		846;Phosphoserine	
		891;N6-acetyllysine	
		916;Phosphothreonine	
		1012;Phosphoserine	
		1018;Phosphoserine	
		1020;Phosphoserine	
		1156;Phosphoserine	
		1289;Phosphoserine	
	Predicted disordered region	P892T: -	881-1034;888-903;878-905;888-1117;891-908;887-1339;
			
Q9HBH7	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/Q9HBH7)	2-121;BEX	
	Post-translational modification site		
	Predicted disordered region	A40V: -	1-108;1-54;1-41;1-52;1-125;1-64;1-40;

			
Q9NSI2 Protein FAM207A	Superfamily domain	75-141; Replication terminator protein (Tus)	
	Pfam domain (http://pfam.xfam.org/protein/Q9NSI2)	99-221;SLX9	
	Post-translational modification site	34;Phosphothreonine	
		203;Phosphoserine	
	Predicted disordered region	V212L: -	148-230;212-212;70-217;207-218;74-215;
			
Q9NVL1 Protein FAM86C1	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/Q9NVL1)	6-54;FAM86	
	Post-translational modification site		
	Predicted disordered region	P135L: -	121-144;106-153;102-141;101-148;133-142;
			
Q9NXF7 DDB1- and CUL4-	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/Q9NXF7)	1-216;DCA16	

associated factor 16	Post-translational modification site	61;N6-acetyllysine	
	Predicted disordered region	N45S: -	1-54;1-50;44-46;1-47;1-51;1-47;1-49;
			
Q9NYF0 Dapper homolog 1	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/Q9NYF0)	46-836;Dapper	
	Post-translational modification site	237;Phosphoserine	
		827;Phosphoserine	
	Predicted disordered region	S682L: A colorectal cancer sample	589-700;381-716;380-807;294-730;591-694;294-708;
			
Q9NZM5 Glioma tumor suppressor candidate region gene 2 protein	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/Q9NZM5)	41-445;Nop53	
	Post-translational modification site	2;N-acetylalanine	
	Predicted disordered region	Q389R: -	388-413;263-417;201-421;375-416;201-478;

			
Q9P0W8 Spermatogenesis-associated protein 7	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/Q9P0W8)	9-422;HSD3	
	Post-translational modification site		
	Predicted disordered region	G324E: -	312-327;230-364;318-351;318-361;322-324;
			
Q9UHV2 SERTA domain-containing protein 1	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/Q9UHV2)	45-82;SERTA	
	Post-translational modification site		
	Predicted disordered region	T31A: -	1-37;26-32;1-43;1-51;30-43;
			
Q9UHY8 Fasciculation and elongation protein zeta-2	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/Q9UHY8)	51-286;FEZ	
	Post-translational modification site		
	Predicted disordered region	P50L: -	1-54;1-50;1-51;1-80;20-52;1-78;

			
Q9UJX5 Anaphase-promoting complex subunit 4	Superfamily domain	20-116; WD40 repeat-like	
		170-220; WD40 repeat-like	
	Pfam domain (http://pfam.xfam.org/protein/Q9UJX5)	10-57;ANAPC4_WD40	
		232-431;ANAPC4	
	Post-translational modification site	469;Phosphotyrosine	
		757;Phosphoserine	
		758;Phosphoserine	
		777;Phosphoserine	
		779;Phosphoserine	
	Predicted disordered region	E800G: -	756-808;749-808;799-809;746-808;798-808;748-808;
			
Q9Y238 Deleted in lung and esophageal cancer protein 1	Superfamily domain	756-831; PapD-like	
	Pfam domain (http://pfam.xfam.org/protein/Q9Y238)		
	Post-translational modification site		
	Predicted disordered region	N1150D: -	1135-1151;1136-1185;1138-1177;1144-1174;1122-1177;1137-1151;
			
Q9Y2X0 Mediator of RNA	Superfamily domain	25-141; WD40 repeat-like	
		62-113; Soluble quinoprotein glucose	

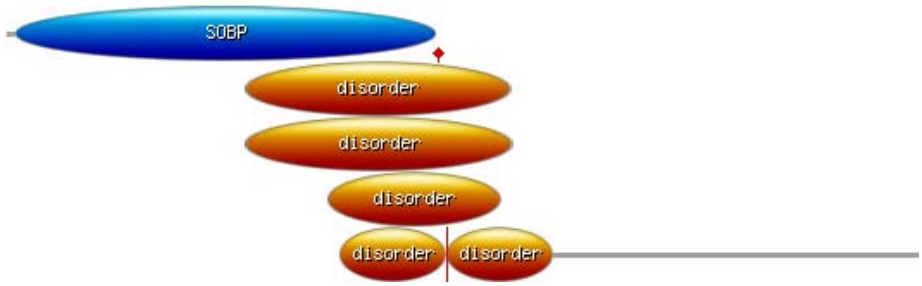
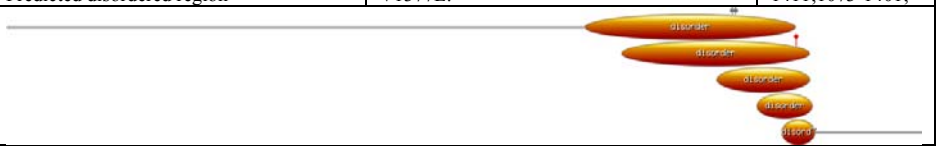
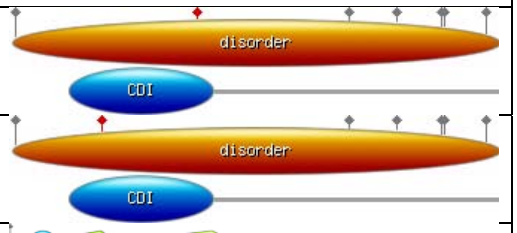
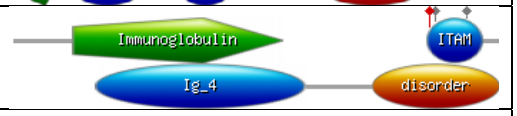
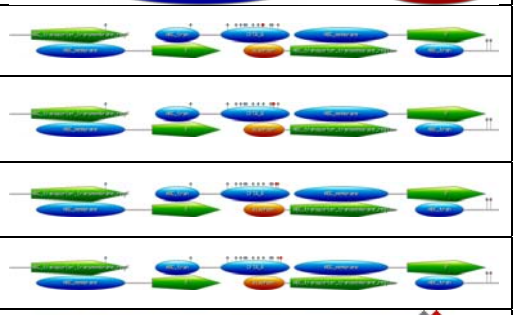
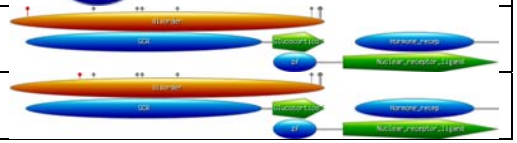
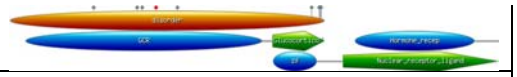
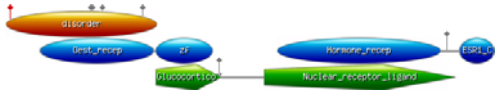
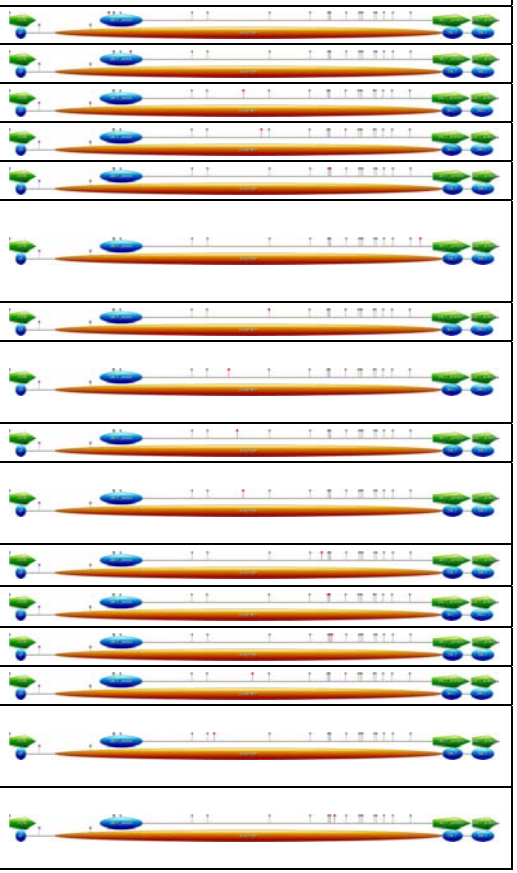
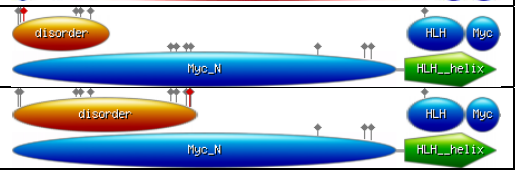
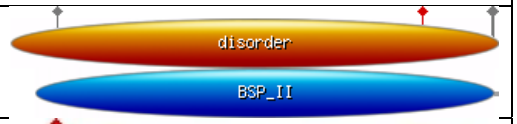
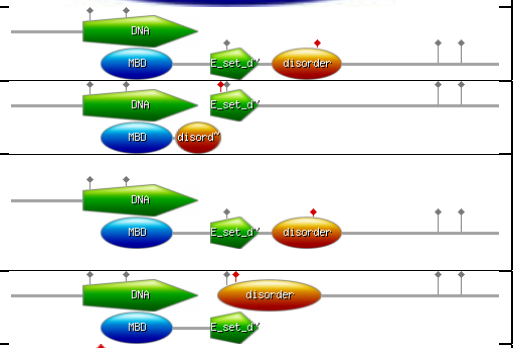
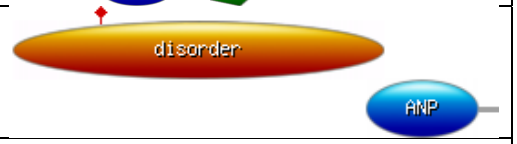
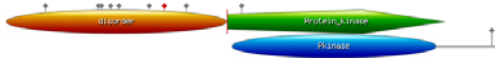
polymerase II transcription subunit 16		dehydrogenase	
		146-235; Soluble quinoprotein glucose dehydrogenase	
		265-324; Soluble quinoprotein glucose dehydrogenase	
		357-395; Soluble quinoprotein glucose dehydrogenase	
	Pfam domain (http://pfam.xfam.org/protein/Q9Y2X0)	119-829;Med16	
	Post-translational modification site		
	Predicted disordered region	L770F: -	769-770;733-779;744-788;763-772;726-788;
			
Q9Y5P3 Retinoic acid-induced protein 2	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/Q9Y5P3)	7-250;SOBP	
	Post-translational modification site		
	Predicted disordered region	M252V: -	188-288;252-318;140-294;195-256;140-295;
			
Q9Y6X0 SET-binding protein	Superfamily domain		
	Pfam domain (http://pfam.xfam.org/protein/Q9Y6X0)		
	Post-translational modification site	1266;Phosphoserine	
		1272;Phosphoserine	
	Predicted disordered region	V1377L: -	1009-1377;1308-1406;1238-1401;1352-1411;1073-1401;
			

Table S5.10 Detailed functional annotations for proteins in Table 5.4 including modified residues, protein superfamily domains and Pfam domains

UniProt	Name	Disorder region	Mutation	Disease /Phenotype	Superfamily domain	Pfam domain	Modified residue	Ensemble figure
P38936	Cyclin-dependent kinase inhibitor 1	#1-164	Phe63Leu	polymophism		20-68;CDI	2;N-acetylserine 114;Phosphoserine 130;Phosphoserine 145;Phosphothreonine 146;Phosphoserine 160;Phosphoserine#	
			Ser31Arg	polymophism				
P35869	Aryl hydrocarbon receptor	#545-713	Arg554Lys	polymophism	293-388; PYP-like sensor domain (PAS domain) 122-183; PYP-like sensor domain (PAS domain) 37-78; HLH; helix-loop-helix DNA-binding domain	297-383;PAS_3 113-227;PAS 35-80;HLH	1;N-acetylmethionine	
			Val570Ile	polymophism				
P04234	T-cell surface glycoprotein CD3 delta chain	#127-171	Gln147Arg	polymophism	22-95; Immunoglobulin	146-165;ITAM 30-103;Ig_4	149;Phosphotyrosine 160;Phosphotyrosine	
P13569	Cystic fibrosis transmembrane conductance regulator	#708-831	Arg766Met	CBAVD	1203-1440; P-loop containing nucleoside triphosphate hydrolases 432-640; P-loop containing nucleoside triphosphate hydrolases 850-1175; ABC transporter transmembrane region 67-361; ABC transporter transmembrane region	639-849;CFTR_R 862-1147;ABC_membrane 81-350;ABC_membrane 1227-1374;ABC_tran 441-576;ABC_tran	291;Phosphothreonine 549;Phosphoserine 660;Phosphoserine 686;Phosphoserine 700;Phosphoserine 712;Phosphoserine 717;Phosphothreonine 737;Phosphoserine 753;Phosphoserine 768;Phosphoserine 790;Phosphoserine 795;Phosphoserine 813;Phosphoserine 1444;Phosphoserine 1456;Phosphoserine	
		#708-832	Ala800Gly	CBAVD				
		#708-832	Ile807Met	CBAVD				
		#708-832	Glu822Lys	CF				
P49918	Cyclin-dependent kinase inhibitor 1C	#1-316	Phe276Val	IMAGE		32-82;CDI	268;Phosphoserine	
P04150	Glucocorticoid receptor	#1-500	Phe29Leu	polymophism	530-776; Nuclear receptor ligand-binding domain 418-500; Glucocorticoid receptor-like (DNA-binding domain)	26-401;GCR 549-738;Hormone_recep 419-488;zf-C4	134;Phosphoserine 203;Phosphoserine 211;Phosphoserine 267;Phosphoserine 480;N6-acetyllysine 492;N6-acetyllysine	
		#1-500	Leu112Phe	polymophism				

		#1-500	Asp233Asn	polymophism			494;N6-acetyllysine 495;N6-acetyllysine	
P03372	Estrogen receptor	#1-184	His6Tyr	in a breast cancer sample; somatic mutation	315-547; Nuclear receptor ligand-binding domain 183-262; Glucocorticoid receptor-like (DNA-binding domain)	552-595;ESR1_C 42-181;Oest_recep 331-529;Hormone_recep 183-252;zf-C4	104;Phosphoserine 106;Phosphoserine 118;Phosphoserine 167;Phosphoserine 260;Asymmetric dimethylarginine 537;Phosphotyrosine#	
P38398	Breast cancer type 1 susceptibility protein	#170-1649	Ile379Met	polymophism	1-102; RING/U-box 1608-1753; BRCT domain 1759-1857; BRCT domain	345-508;BRCT_assoc 1756-1842;BRCT 1644-1723;BRCT 24-64;zf-C3HC4	1;N-acetylmethionine 114;Phosphoserine 308;Phosphoserine 395;Phosphoserine 398;Phosphoserine 423;Phosphoserine 694;Phosphoserine 753;Phosphoserine 988;Phosphoserine 1143;Phosphoserine 1211;Phosphoserine 1217;Phosphoserine 1218;Phosphoserine 1280;Phosphoserine 1328;Phosphoserine 1336;Phosphoserine 1342;Phosphoserine 1387;Phosphoserine 1394;Phosphothreonine 1423;Phosphoserine 1457;Phosphoserine 1524;Phosphoserine	
			Phe461Leu	BC				
			Leu892Ser	BC				
			Gly960Asp	BC				
			Glu1219Asp	polymophism				
			Thr1561Ile	Found in breast cancer; unknown pathological significance.				
			Phe989Ser	polymophism				
			His835Tyr	BROVCA1; unknown pathological significance.				
			Arg866Gln	polymophism				
			His888Tyr	in BC; unknown pathological significance.				
			Ser1187Ile	BC and BROVCA1.				
			Ser1217Tyr	BC and BROVCA1.				
			Phe1226Leu	BROVCA1				
			Ile925Leu	polymophism				
			Gly778Cys	in a breast cancer sample; somatic mutation				
			Asn1236Lys	in BC; unknown pathological significance;				

				functionally neutral in vitro.				
			Glu1250Lys	polymophism				
			Arg170Trp	in BC; unknown pathological significance; functionally neutral in vitro.				
			Ser186Tyr	in BC; unknown pathological significance; functionally neutral in vitro.				
			Arg866Cys	polymophism				
			Leu1267Ser	in BC; unknown pathological significance; functionally neutral in vitro.				
			Glu1282Val	in BC; unknown pathological significance; functionally neutral in vitro.				
			Ser1301Arg	in BC; unknown pathological significance; functionally neutral in vitro.				
			Pro798Leu	in BC; unknown pathological significance; functionally neutral in vitro.				
			Asn810Tyr					
P01106	Myc proto-oncogene protein	#1-88	Asn11Ser	polymophism	353-435; HLH; helix-loop-helix DNA-binding domain	1-345;Myc_N 408-438;Myc-LZ 355-407;HLH	6;Phosphoserine 8;Phosphothreonine 58;Phosphothreonine 62;Phosphoserine 71;Phosphoserine 143;N6-acetyllysine 148;N6-acetyllysine	
		#1-167	Gly160Cys	polymophism				

		#1-88	Asn86Thr	in a Burkitt lymphoma sample			157;N6-acetyllysine 161;Phosphoserine 275;N6-acetyllysine 317;N6-acetyllysine 323;N6-acetyllysine 371;N6-acetyllysine	
P21815	Bone sialoprotein 2	#1-317	Ala268Val	polymophism		17-314;BSP_II	31;Phosphoserine 313;Sulfotyrosine 314;Sulfotyrosine	
Q9NR00	Uncharacterized protein C8orf4	#1-106	Val10Ile	polymophism		10-85;TC1		
P51608	Methyl-CpG-binding protein 2	#261-330	Arg306Cys	RTT	73-187; DNA-binding domain 200-247; E set domains	91-162;MBD	80;Phosphoserine 116;Phosphoserine 216;Phosphoserine 426;Phosphoserine 449;N6-acetyllysine	
		#165-210	Lys210Ile	RTT				
		#261-330	Pro302Ala	RTT				
		#261-330	Pro302His	RTT				
		#261-330	Pro302Arg	RTT				
		#207-310	Pro225Leu	RTT				
P16860	Natriuretic peptides B	#1-102	Arg25Leu	polymophism		98-128;ANP		
P30291	Wee1-like protein kinase	#1-292	Gly210Cys	polymophism	279-580; Protein kinase-like (PK-like)	299-569;Pkinase	53;Phosphoserine 123;Phosphoserine 127;Phosphoserine 139;Phosphoserine 150;Phosphoserine 190;Phosphothreonine 239;Phosphothreonine 312;Phosphoserine 642;Phosphoserine#	

Q13569	G/T mismatch-specific thymine DNA glycosylase	#340-410	Val367Met	polymophism	123-297; Uracil-DNA glycosylase-like	125-296;UDG		

Publications

OXFORD UNIVERSITY PRESS | DATABASE

PolyQ 2.0: an updated database of human polyglutamine proteins

Journal:	*DATABASE*
Manuscript ID:	DATABASE-2015-0119
Manuscript Type:	Database Update
Date Submitted by the Author:	13-Aug-2015
Complete List of Authors:	Li, Chen; Monash University, Department of Biochemistry and Molecular Biology Nagel, Jeremy; Monash University, Department of Biochemistry and Molecular Biology Androulakis, Steve; Monash University, Monash Bioinformatics Platform Song, Jiangning; Monash University, Department of Biochemistry and Molecular Biology Buckle, Ashley; Monash University, Department of Biochemistry and Molecular Biology
Keywords:	Polyglutamine repeats, neurodegenerative diseases, database, functional annotation, multiple sequence alignment

PolyQ 2.0: an updated database of human polyglutamine proteins

Chen Li[1], Jeremy Nagel[1], Steve Androulakis[2], Jiangning Song[2,3,4,*] and Ashley M. Buckle[1,*]

[1]Department of Biochemistry and Molecular Biology, Faculty of Medicine, Monash University, Melbourne, Victoria 3800, Australia

[2]Monash Bioinformatics Platform, Monash University, Melbourne, Victoria 3800, Australia

[3]Department of Microbiology, Faculty of Medicine, Melbourne, Victoria 3800, Australia

[4]National Engineering Laboratory of Industrial Enzymes and Key Laboratory of Systems Microbial Biotechnology, Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, Tianjin 300308, China

*To whom correspondence should be addressed to [redacted]

Abstract

Proteins with expanded polyglutamine (PolyQ) repeats are involved in human neurodegenerative diseases, via a gain-of-function mechanism of neuronal toxicity involving protein conformational changes that result in the formation and deposition of β-sheet-rich aggregates. Aggregation is dependent on the context and the properties of the host protein, such as domain architecture and location of the repeat tract. In order to explore this relationship in greater detail, here we describe PolyQ 2.0, an updated database that provides a comprehensive knowledgebase for human polyQ proteins. The database details domain context information, protein structural and functional annotation, single point mutations, predicted disordered regions, protein-protein interaction partners, metabolic/signaling pathways, post-translational modification sites and evolutionary information. Several new database functionalities have also been added, including search with multiple keywords, and new data entry submission. Currently the database contains nine reviewed disease-associated polyQ proteins, 105 reviewed non-disease polyQ proteins and 146 un-reviewed polyQ proteins. We envisage that this updated database will be a useful resource for functional and structural investigation of human polyQ proteins.

Database URL: http://lightning.med.monash.edu/polyq2/

Introduction

The polyglutamine (PolyQ) repeat family of proteins contain a stretch of multiple consecutive glutamines (1). Expansion of the polyQ tract can lead to a toxic gain-of-function via a conformational change within the protein and the deposition of β-sheet-rich amyloid-like fibrils (2-4). As such, polyQ repeats are implicated in several neurodegenerative diseases, including Huntington disease and spinocerebellar ataxia (5-11). While the length of the polyQ repeat is critical to the pathogenesis, the polyQ domain context (i.e. the domains flanking the polyQ tract) is also important (12-15). Although there are many human polyQ-containing proteins (16), only nine proteins are implicated in pathogenesis, with the precise repeat threshold to pathogenesis varying within the disease subset (17-19).

Given the importance of polyQ repeats and their domain context information, we recently performed a bioinformatics investigation of the protein context of polyglutamine repeats (20), and constructed a web-accessible database of all human proteins containing a polyQ repeat greater than seven glutamines in length (21). Although the PolyQ database provides basic information for each entry, it lacks in both depth and breadth of annotation as well as functionality. Here, we present PolyQ 2.0, a substantially updated knowledgebase for human polyQ proteins. PolyQ 2.0 contains a variety of structural and functional annotations, broad protein information, and domain context of polyQ repeats. In addition, the usability of the web interface has been improved, which now offers database search with multiple keywords as well as user data submission. PolyQ updates the MySQL relational database that stores entries, and enhances the web interface through the use of modern Javascript tools for visualization and interaction. Apache Tomcat mediates users access to the database through Java Servlets and JavaServer Pages (JSP).

Update of database entries

Whereas PolyQ contained two types of polyQ proteins, namely disease and non-disease-associated, in PolyQ 2.0 all entries are categorized into three groups according to the annotation of disease involvement and review completeness. Here disease-associated proteins refer to those proteins causing neurodegenerative diseases due to the abnormal expansion of polyQ repeats rather than other proteins with common disease-associated mutations. These groups are: reviewed disease-associated

polyQ proteins, reviewed non-disease polyQ proteins and un-reviewed polyQ proteins. We first validated all the data entries in the previous PolyQ database with their UniProt annotation in order to ensure that only high quality data entries are included in PolyQ 2.0. Proteins were included as reviewed entries according to their annotation in the UniProt database. We incorporated polyQ proteins that have not been manually verified from UniProt as un-reviewed polyQ proteins for potential future reference. As a result, we obtained nine reviewed disease-associated polyQ proteins, 105 reviewed non-disease polyQ proteins and 146 un-reviewed polyQ proteins, respectively (Figure 1A).

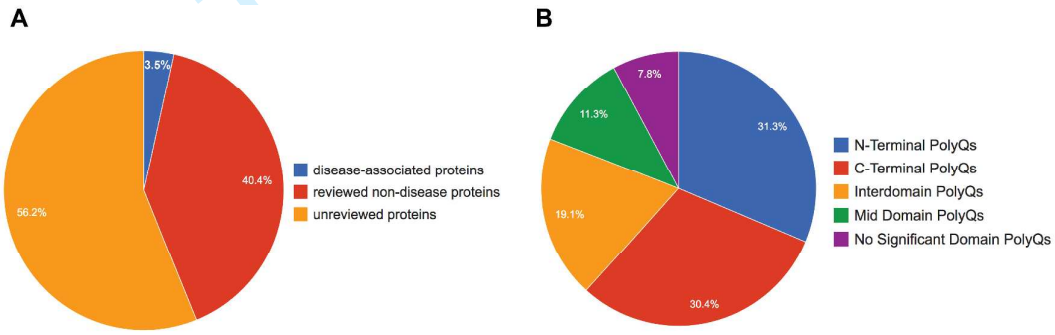


Figure 1. Statistics of data entries in PolyQ 2.0. (**A**) Distribution of disease-associated proteins, reviewed non-disease proteins and un-reviewed proteins; (**B**) Distribution of the sequence context of different types of polyQ domains for reviewed entries only.

Following the classification system set out previously in PolyQ, we further classified all reviewed 114 sequences into six categories based on the locations and context of polyQ repeats relative to Pfam domains (22): (1) *N-Terminal PolyQs* – the first polyQ repeat appears before all Pfam domains; (2) *C-Terminal PolyQs* – the last polyQ repeat appears after all the Pfam domains; (3) *Interdomain PolyQs* – the polyQ tracts appear between the first Pfam and last Pfam domain; (4) *Mid Domain PolyQs* – the polyQ repeat appears in the middle of a Pfam domain or overlaps with a Pfam domain; (5) *No Significant Domain PolyQs* – sequences that do not contain any significant Pfam domains; (6) *Unclassified PolyQs* – sequences that do not fit into any of the above categories. The majority of polyQ domains are either *N-* or *C-Terminal PolyQs* while only 7.8% of the reviewed polyQ containing entries do not harbor any significant Pfam domains (Figure 1B).

Update of content and annotation

For PolyQ 2.0, the information content and annotations for entries have been significantly improved and expanded. The updated content includes basic protein information, protein structural information, predicted disordered regions, protein-protein interaction partners, metabolic/signaling pathways, single point disease- and non-disease associated mutations, and protein post-translational modification sites. In addition, we also performed BLAST search and generated multiple sequence alignments (MSA) in order to provide evolutionary information for each protein entry. A comparison of protein annotations provided in PolyQ and PolyQ 2.0 is shown in Table 1.

(Table 1)

Annotations were extracted and reviewed from a variety of different publicly available resources, including UniProt (23), Protein Data Bank (24), BioGrid (25), KEGG (26), SUPERFAMILY (27) and Pfam (22). We employed VSL2B (28) to annotate predicted disordered regions. Homologous sequence search was conducted using PSI-BLAST (29) (with an *E*-value of 0.001) against the Swiss-Prot database (http://www.uniprot.org/downloads). Multiple sequence alignments were generated using Clustal Omega (30). A summary of the database contents and annotations is shown in Table 2.

(Table 2)

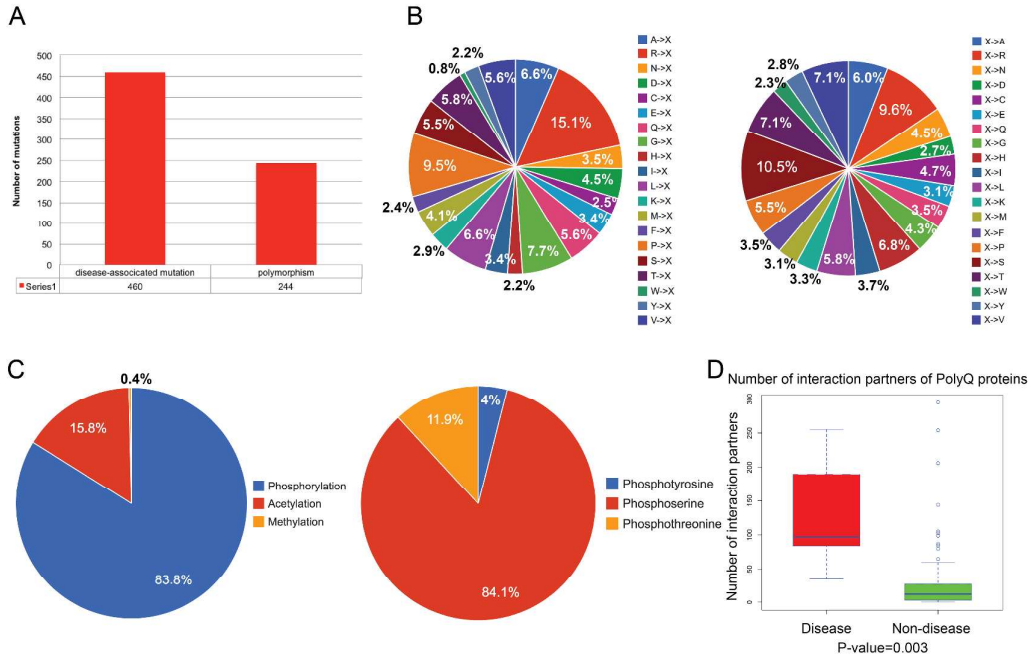


Figure 2. Statistical analysis of database content in terms of distributions of disease-associated mutations, post-translational modification site and number of protein-protein interaction partners. (**A**) Distribution of disease-associated mutation and polymorphism; (**B**) Distribution of the number of mutations with respect to two mutation patterns (where *X* means any amino acid); (**C**) Distribution of types of protein post-translational modification with detailed distribution of sub-types of phosphorylation; (**D**) Number of protein-protein interaction partners of reviewed polyQ disease-associated proteins and non-disease proteins.

We analyzed the database content in terms of distribution of disease-associated mutations, post-translational modification sites and number of protein-protein interaction partners. From a total of 704 single point mutations within the 260 data entries, 460 (65.3%) mutations are disease-associated, while 244 (34.7%) mutations are polymorphisms (Figure 2A). By analyzing the distribution of different types of mutations associated with polyQ proteins, we found that arginine is the most frequently mutated amino acid (approximately 15% of the mutated residues; Figure 2B). Phosphorylation is the most frequently observed post-translational modification (Figure 2C). Disease-associated polyQ proteins have significantly more protein interaction partners than non-disease polyQ proteins (*p*-value = 0.003; Figure 2D).

Database functionality and web interface improvements

PolyQ 2.0 features several important improvements of the user interface as well as new functionality, including database search with multiple types of keywords and new entry submission. A comparison of database functionality between PolyQ and PolyQ 2.0 is listed in Table 3.

(Table 3)

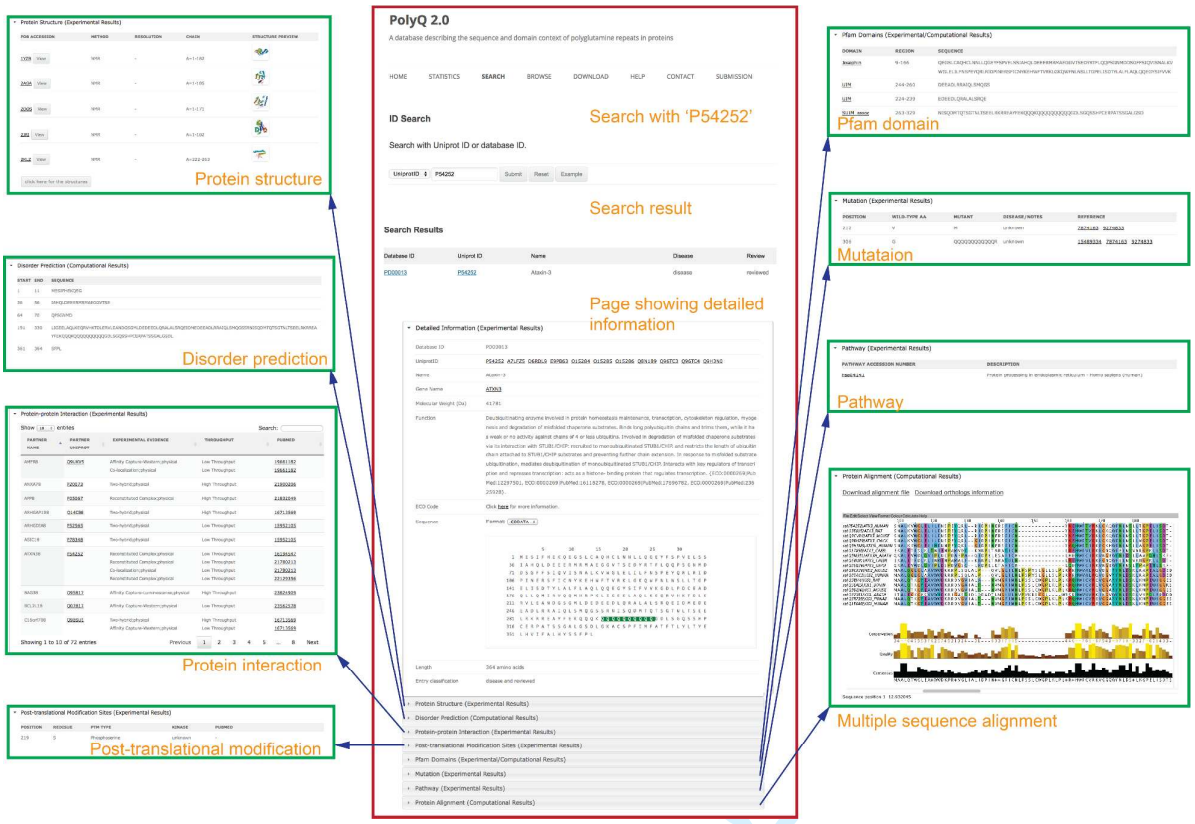


Figure 3. Typical search results in PolyQ 2.0 using the UniProt ID P54252 as an example. The results are summarized and displayed in nine main sections, including protein information, protein structure, metabolic/signalling pathway, protein interaction, post-translational modification site, Pfam domain, disorder region prediction, protein mutation and multiple sequence alignment.

The search functionality in PolyQ 2.0 has been considerably improved, with search options available using multiple keywords, in addition to the options of protein name and Pfam domain offered by the previous version. The database can be searched by PolyQ/UniProt ID, protein name, Pfam domain, disease, type of protein post-translational modification sites/kinase and protein-protein interaction partner name.

The PolyQ ID is composed of "PD" followed by five digits. As there are in total 260 entries in PolyQ 2.0, the PolyQ ID ranges from "PD00001" to "PD00260". An example of the result of database search with UniProt ID=P54252 (Ataxin-3) is shown in Figure 3, comprising nine main sections related to different annotations.

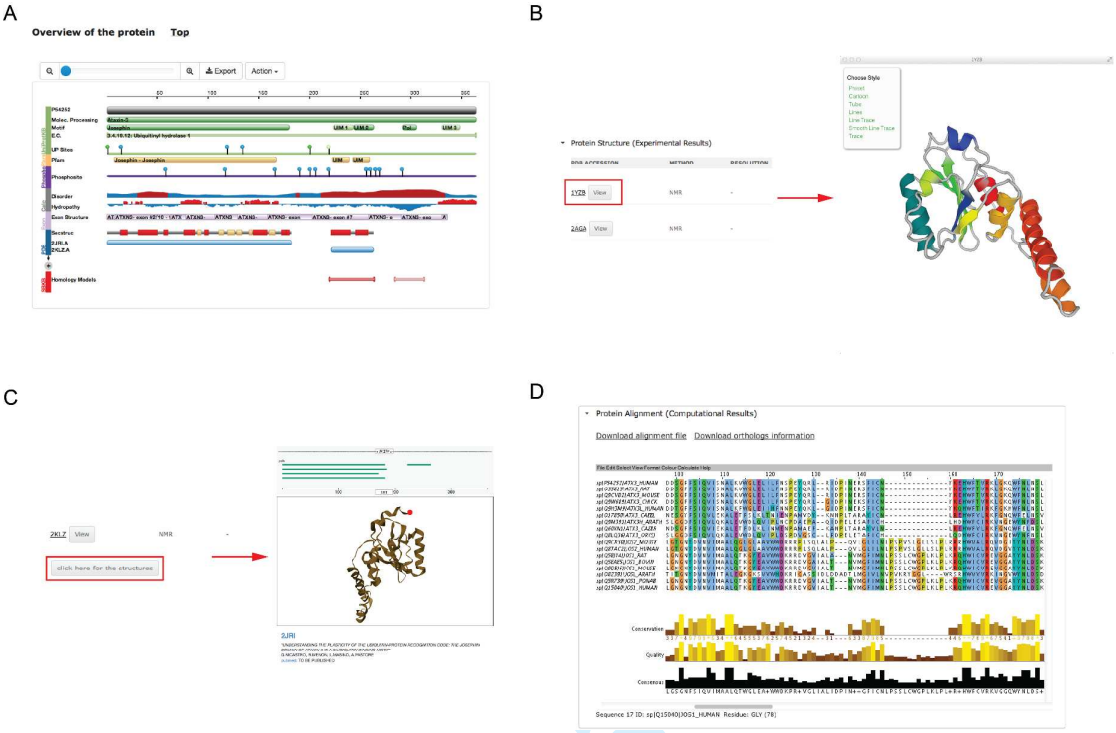


Figure 4. Plug-ins in PolyQ 2.0 to enhance database visualization. (**A**) Protein feature plug-in; (**B**) PV showing protein structure; (**C**) pViz for visualizing multiple structures; (**D**) Jalview displaying MSAs.

Several plug-ins were employed to enhance visualization of database entries. In the protein basic information section, we embedded a protein feature view plug-in in order to show protein functional sites/domains and basic structural information (Figure 4A). PV (http://biasmv.github.io/pv/) and pViz (31) were also used to allow detailed examination of protein structures (Figure 4BC). Multiple sequence alignment is displayed using JalView (32) to visualize sequence conservation (Figure 4D).

Browsing of data entries has also been improved. The entries can now be categorized in terms of disease involvement and completeness of review and annotation. In addition, detailed context annotations, which show the distribution of polyQ domain, protein superfamily domain and protein post-translational

modification sites are available. A webpage showing database statistics is available, giving users a one-page snapshot of database contents as well as convenient navigation around the database. Detailed user help and instructions are also provided. Finally, we have built a data submission page, enabling users to deposit data in the database, which are made publically available after checking, curation and approval by the site administrator.

Conclusions

Based on our previous PolyQ database for human polyQ proteins, in the present study we have developed an updated database, PolyQ 2.0, to provide comprehensive protein functional, structural and evolutional annotations together with domain context information for human polyQ proteins. Integrating publicly available annotations and computational resources, PolyQ 2.0 offers a variety of annotations in terms of protein basic information, protein structure, predicted intrinsically disordered domain, protein-protein interaction, protein functional site/domain, single point mutation, metabolic/signaling pathway and multiple sequence alignment. We anticipate that this updated knowledgebase will benefit functional and structural studies of human polyQ proteins and their role in neurodegenerative diseases.

Acknowledgements

We would like to thank Dr. Andreas Prlić for his help with the protein feature view plug-in.

Funding

This work was supported by grants from the National Natural Science Foundation of China (61202167, 61303169), the Hundred Talents Program of the Chinese Academy of Sciences (CAS), and the Knowledge Innovative Program of CAS (KSCX2-EW-G-8) of CAS. JS is a recipient of the Hundred Talents Program of CAS. AMB is an NHMRC Senior Research Fellow.

Conflict of Interest: none declared.

References

1. Fan, H.C., Ho, L.I., Chi, C.S., *et al.* (2014) Polyglutamine (PolyQ) diseases: genetics to treatments. *Cell Transplant.*, 23, 441-458.
2. Perutz, M.F., Johnson, T., Suzuki, M., *et al.* (1994) Glutamine repeats as polar zippers: their possible role in inherited neurodegenerative diseases. *Proc. Natl. Acad. Sci. U. S. A.*, 91, 5355-5358.
3. Chen, S., Berthelier, V., Hamilton, J.B., *et al.* (2002) Amyloid-like features of polyglutamine aggregates and their assembly kinetics. *Biochemistry*, 41, 7391-7399.
4. Robertson, A.L., Horne, J., Ellisdon, A.M., *et al.* (2008) The structural impact of a polyglutamine tract is location-dependent. *Biophys. J.*, 95, 5922-5930.
5. Kawaguchi, Y., Okamoto, T., Taniwaki, M., *et al.* (1994) CAG expansions in a novel gene for Machado-Joseph disease at chromosome 14q32.1. *Nat. Genet.*, 8, 221-228.
6. Lin, B., Nasir, J., MacDonald, H., *et al.* (1994) Sequence of the murine Huntington disease gene: evidence for conservation, alternate splicing and polymorphism in a triplet (CCG) repeat [corrected]. *Hum. Mol. Genet.*, 3, 85-92.
7. Zuhlke, C., Hellenbroich, Y., Dalski, A., *et al.* (2001) Different types of repeat expansion in the TATA-binding protein gene are associated with a new form of inherited ataxia. *Eur. J. Hum. Genet.*, 9, 160-164.
8. Nakamura, K., Jeong, S.Y., Uchihara, T., *et al.* (2001) SCA17, a novel autosomal dominant cerebellar ataxia caused by an expanded polyglutamine in TATA-binding protein. *Hum. Mol. Genet.*, 10, 1441-1448.
9. Silveira, I., Miranda, C., Guimaraes, L., *et al.* (2002) Trinucleotide repeats in 202 families with ataxia: a small expanded (CAG)n allele at the SCA17 locus. *Arch. Neurol*, 59, 623-629.
10. Banfi, S., Servadio, A., Chung, M.Y., *et al.* (1994) Identification and characterization of the gene causing type 1 spinocerebellar ataxia. *Nat. Genet.*, 7, 513-520.
11. Quan, F., Janas, J., Popovich, B.W. (1995) A novel CAG repeat configuration in the SCA1 gene: implications for the molecular diagnostics of spinocerebellar ataxia type 1. *Hum. Mol. Genet.*, 4, 2411-2413.
12. Stefani, M., Dobson, C.M. (2003) Protein aggregation and aggregate toxicity: new insights into protein folding, misfolding diseases and biological evolution. *J. Mol. Med .(Berl.)*, 81, 678-699.
13. Saunders, H.M., Bottomley, S.P. (2009) Multi-domain misfolding: understanding the aggregation pathway of polyglutamine proteins. *Protein Eng. Des. Sel.*, 22, 447-451.
14. Ellisdon, A.M., Thomas, B., Bottomley, S.P. (2006) The two-stage pathway of ataxin-3 fibrillogenesis involves a polyglutamine-independent step. *J. Biol. Chem.*, 281, 16888-16896.
15. DiFiglia, M., Sapp, E., Chase, K.O., *et al.* (1997) Aggregation of huntingtin in neuronal intranuclear inclusions and dystrophic neurites in brain. *Science*, 277, 1990-1993.
16. Faux, N.G., Bottomley, S.P., Lesk, A.M., *et al.* (2005) Functional insights from the distribution and role of homopeptide repeat-containing proteins. *Genome Res.*, 15, 537-551.

17. Goto, J., Watanabe, M., Ichikawa, Y., *et al.* (1997) Machado-Joseph disease gene products carrying different carboxyl termini. *Neurosci. Res.*, 28, 373-377.
18. Padiath, Q.S., Srivastava, A.K., Roy, S., *et al.* (2005) Identification of a novel 45 repeat unstable allele associated with a disease phenotype at the MJD1/SCA3 locus. *Am. J. Med. Genet. B Neuropsychiatr. Genet.*, 133B, 124-126.
19. Li, W., Serpell, L.C., Carter, W.J., *et al.* (2006) Expression and characterization of full-length human huntingtin, an elongated HEAT repeat protein. *J. Biol. Chem.*, 281, 15916-15922.
20. Robertson, A.L., Bate, M.A., Buckle, A.M., *et al.* (2011) The rate of polyQ-mediated aggregation is dramatically affected by the number and location of surrounding domains. *J. Mol. Biol.*, 413, 879-887.
21. Robertson, A.L., Bate, M.A., Androulakis, S.G., *et al.* (2011) PolyQ: a database describing the sequence and domain context of polyglutamine repeats in proteins. *Nucleic Acids Res.*, 39, D272-276.
22. Finn, R.D., Bateman, A., Clements, J., *et al.* (2014) Pfam: the protein families database. *Nucleic Acids Res.*, 42, D222-230.
23. (2015) UniProt: a hub for protein information. *Nucleic Acids Res.*, 43, D204-212.
24. Rose, P.W., Prlic, A., Bi, C., *et al.* (2015) The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. *Nucleic Acids Res.*, 43, D345-356.
25. Chatr-Aryamontri, A., Breitkreutz, B.J., Oughtred, R., *et al.* (2015) The BioGRID interaction database: 2015 update. *Nucleic Acids Res.*, 43, D470-478.
26. Kanehisa, M., Goto, S., Sato, Y., *et al.* (2014) Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.*, 42, D199-205.
27. Wilson, D., Pethica, R., Zhou, Y., *et al.* (2009) SUPERFAMILY--sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Res.*, 37, D380-386.
28. Peng, K., Radivojac, P., Vucetic, S., *et al.* (2006) Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics*, 7, 208.
29. Altschul, S.F., Madden, T.L., Schaffer, A.A., *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25, 3389-3402.
30. Sievers, F., Wilm, A., Dineen, D., *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, 7, 539.
31. Mukhyala, K., Masselot, A. (2014) Visualization of protein sequence features using JavaScript and SVG with pViz.js. *Bioinformatics*, 30, 3408-3409.
32. Waterhouse, A.M., Procter, J.B., Martin, D.M., *et al.* (2009) Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, 25, 1189-1191.

Table 1. The comparison of protein annotation in PolyQ and PolyQ 2.0

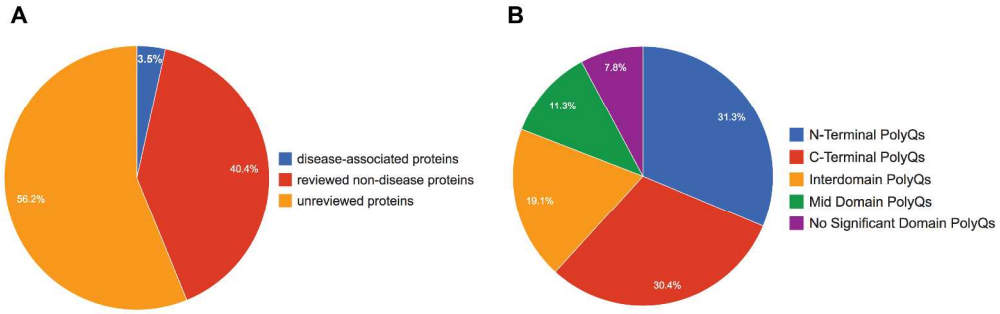
Content	**PolyQ**	**PolyQ 2.0**
Protein information	Sequence and unstructured FASTA headers	Structured protein information (function, gene name, protein accession…)
Protein 3D structure	No	Yes
Pfam domain	Yes	Yes
Protein disordered regions	No	Yes
Protein interaction partner	No	Yes
Metabolic/signaling pathway	No	Yes
Single point mutation	No	Yes, incorporating both disease-associated and nonsense mutations
Post-translational modification sites	No	Yes
Multiple sequence alignment	No	Yes

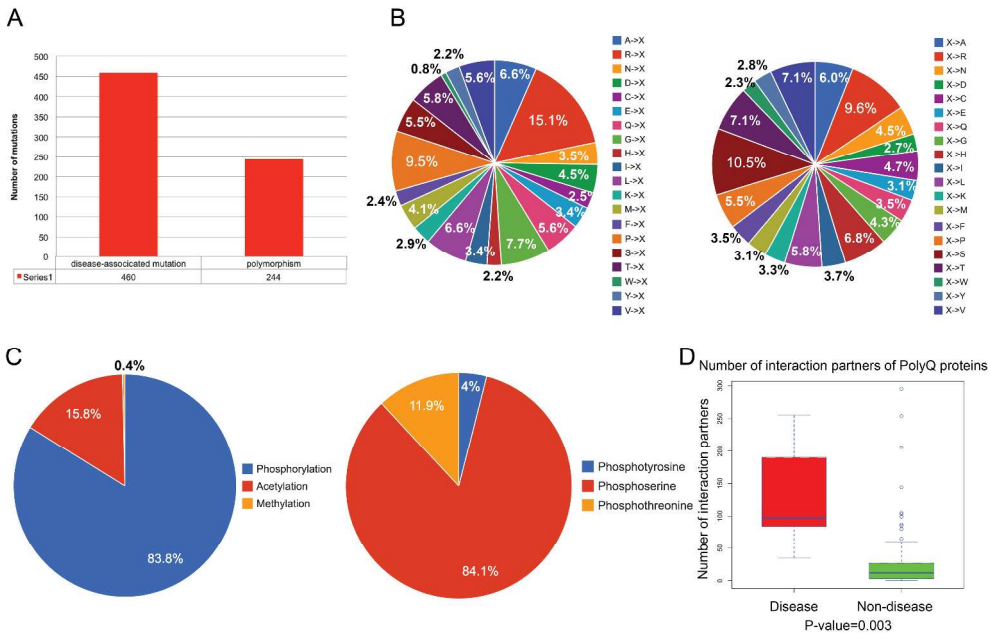
Table 2. Summary of the database contents and annotations of PolyQ 2.0.

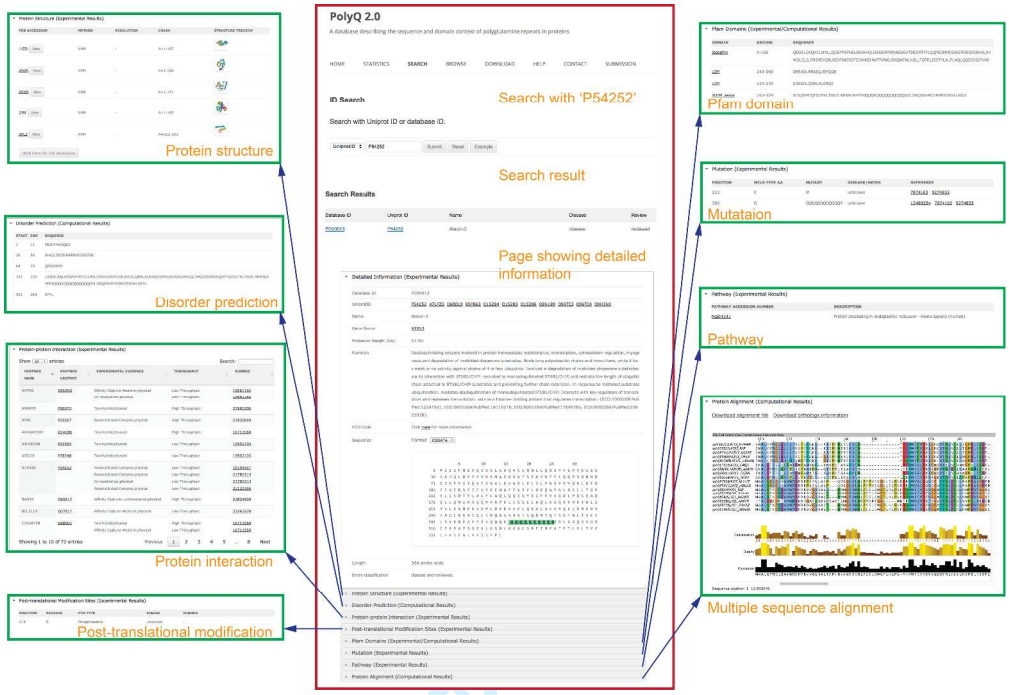
Number of protein structures	356
Number of protein interactions	4,081
Number of single point mutations	704
Number of KEGG pathways	41
Number of Pfam domains	498
Number of post-translational modification sites	569

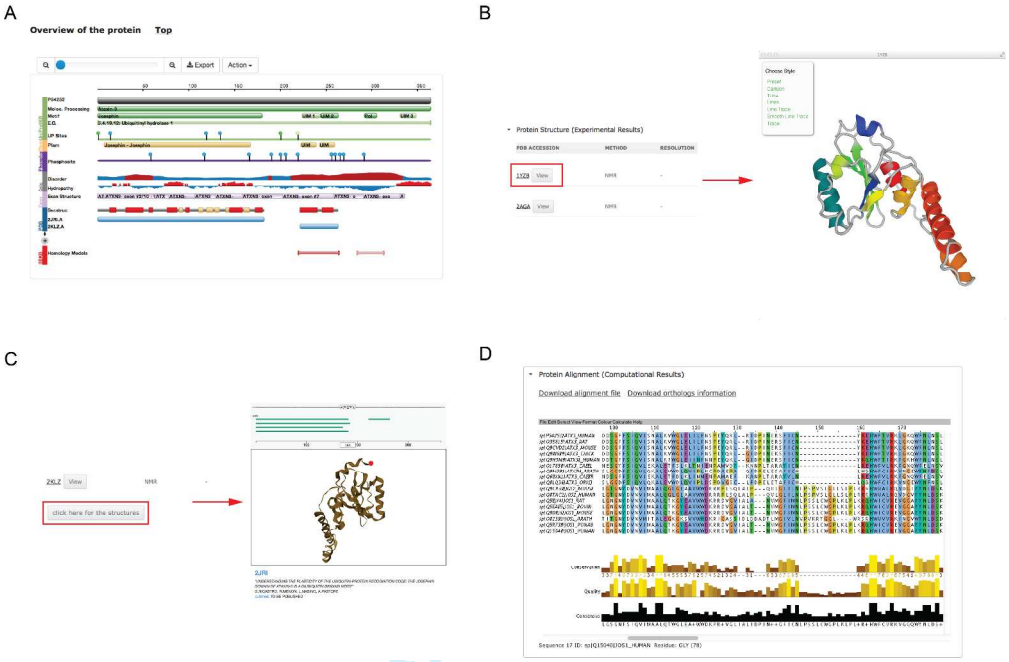
Table 3. Database functionality comparison between PolyQ and PolyQ 2.0

Functionality		**PolyQ**	**PolyQ 2.0**
Database search	Database ID/UniProt ID	No	Yes
	Protein name	Yes	Yes
	Pfam domain	Yes	Yes
	Disease	No	Yes
	PTM	No	Yes
	PTM kinase	No	Yes
	Interaction partner	No	Yes
User submission		No	Yes









OXFORD UNIVERSITY PRESS | DATABASE

KinetochoreDB: a comprehensive online resource for the kinetochore and its related proteins

Journal:	*DATABASE*
Manuscript ID:	DATABASE-2015-0074
Manuscript Type:	Original Article
Date Submitted by the Author:	05-Jun-2015
Complete List of Authors:	Li, Chen; Monash University, Department of Biochemistry and Molecular Biology Androulakis, Steve; Monash University, Monash Bioinformatics Platform Buckle, Ashley; Monash University, Department of Biochemistry and Molecular Biology Song, Jiangning; Monash University, Biochemistry and Molecular Biology; Jiangning Song,
Keywords:	Kinetochore, Database, Functional annotation, Protein structure, Multiple sequence alignment

KinetochoreDB: a comprehensive online resource for the kinetochore and its related proteins

Chen Li[1], Steve Androulakis[2], Ashley M. Buckle[1,*] and Jiangning Song[2,3,4,*]

[1]Department of Biochemistry and Molecular Biology, Faculty of Medicine, Monash University, Melbourne, Victoria 3800, Australia

[2]Monash Bioinformatics Platform, Monash University, Melbourne, Victoria 3800, Australia

[3]Department of Microbiology, Faculty of Medicine, Melbourne, Victoria 3800, Australia

[4]National Engineering Laboratory of Industrial Enzymes and Key Laboratory of Systems Microbial Biotechnology, Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, Tianjin 300308, China

*To whom correspondence should be addressed to [redacted]

Abstract

KinetochoreDB is an online resource for the kinetochore and its related proteins. It provides comprehensive annotations on 1,554 related protein entries in terms of their amino acid sequence, protein 3D structure, predicted intrinsically disordered region, protein-protein interaction, post-translational modification site, functional domain and key metabolic/signaling pathways, integrating several public databases, computational annotations and experimental results. KinetochoreDB provides interactive and customizable search and data display functions that allow users to interrogate the database in an efficient and user-friendly manner. It uses PSI-BLAST searches to retrieve the homologs of all entries and generate multiple sequence alignments that contain important evolutionary information. This knowledgebase also provides annotations of single point mutations for entries with respect to their pathogenicity, which may be useful for generation of new hypotheses on their functions, as well as follow-up studies of human diseases.

Database URL: http://lightning.med.monash.edu/kinetochoreDB2/

Introduction

During cell mitosis and meiosis, the kinetochore plays a critical role of locating the attachments on chromosomes and pulling sister chromatids apart. It is assembled on centromeric chromatin through complex pathways and functions during the cell cycle (1-6). During the last few decades, numerous studies of the kinetochore and its related proteins have characterized its function, architecture and the repertoire of its related proteins using biochemistry, structural biology and cell biology techniques (4,7-11). Both the stability of the kinetochore–microtubule interface and mutations occurring in the kinetochore and its related proteins are associated with a number of human diseases (12-15). Dynamics studies of the kinetochore have also shown that deregulation of the kinetochore-microtubule dynamics frequently results in chromosome instability, leading to the development of cancer (10,11,16). Other experimental studies reveal that mutations of the kinetochore and its related proteins are closely linked to human diseases. For example, the adenomatous polyposis coli protein, found in both centrosome and kinetochore, contains approximately 30 disease mutations that cause Familial Adenomatous Polyposis (FAP) (14,15) and Medulloblastoma (MDB) (12).

Despite its biological significance and our increasing awareness of its potential roles in human diseases, there is currently a paucity of publically available databases or resources that focus on providing comprehensive functional annotations of the kinetochore and its related proteins. The only available database is MiCroKiTS (17), an integrated online resource for kinetochore, midbody, telomere, centrosome and spindle proteins. However, important annotations of entries in MiCroKiTS are not available in terms of protein 3D structure, protein interaction partners, metabolic/signaling pathways etc., all of which are crucial aspects for follow-up functional studies of these proteins.

(Table 1)

In an effort to address this knowledge gap, we created KinetochoreDB, which integrates several public databases, computational annotations and experimental results for currently 1,554 related entries. KinetochoreDB contains several important features, the majority of which are not available in MiCroKiTS (Table 1):

(1) It provides annotations of protein 3D structure when structural information is available. For protein entries with available structural information, the corresponding PDB IDs and their related information are provided. In addition, information on predicted intrinsic disorder is provided, which is particularly important for providing structural insights into those entries in KinetochoreDB whose 3D structures have not been solved.

(2) It provides comprehensive annotations of single point mutations and their pathogenic effects. These mutations are classified as either pathogenic or nonsense mutations in KinetochoreDB. For disease-associated pathogenic mutations, KinetochoreDB provides details of the disease caused by the mutation, allows users to search the entire database with the disease name of interest, and provides user-friendly options to browse the related kinetochore proteins that harbor such disease-associated mutations.

(3) It provides metabolic/signaling pathway information for each entry by cross-referencing the KEGG database. Such information is important for understanding the functions of kinetochore proteins from a biochemical network perspective. In particular, the pathway information and the link to KEGG will be provided if an entry has pathway information available in KEGG.

(4) It provides multiple sequence alignments (MSAs) for all included entries, thereby allowing users to readily identify evolutionarily conserved regions within the family of a kinetochore protein. In addition, the visualization of MSAs implemented by Jalview is user-friendly and customizable.

(5) It provides convenient user enquiry and new entry submission options by enabling users to automatically upload their newly discovered sequences into the online database.

Database construction and features

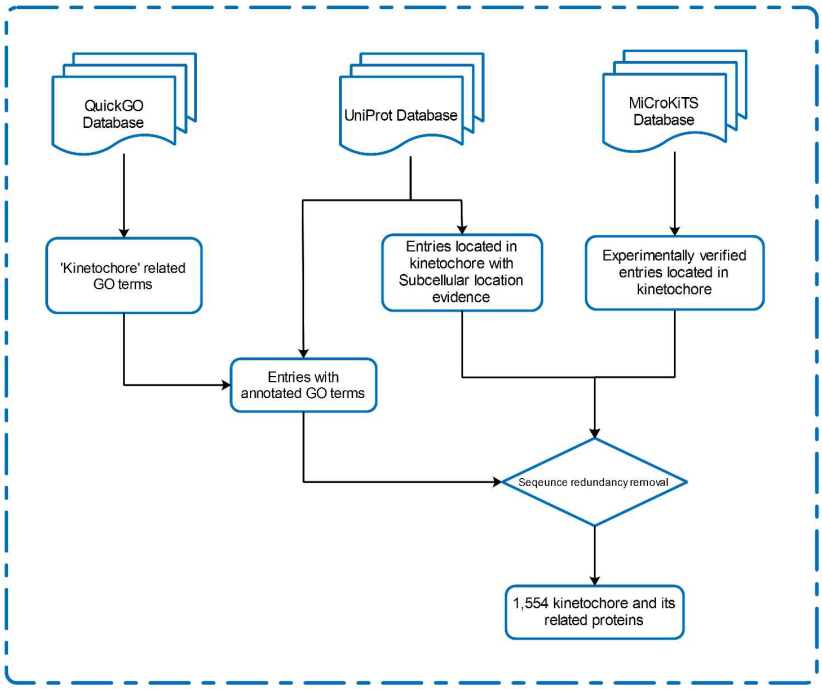


Figure 1. The schema of database construction and data collection processes.

We define 'kinetochore and its related proteins' with respect to protein subcellular location and Gene Ontology. The entries of KinetochoreDB originate from three major resources; QuickGo database (18), UniProt database (19) and MiCroKiTS, and the database was populated as follows. From MiCroKiTS, we obtained data entries that have been experimentally verified to be located in kinetochore. By searching GO terms from QuickGo with the keyword 'kinetochore', we obtained 64 GO terms related to kinetochore. For each GO term, we searched and filtered the reviewed entries from the UniProt database to ensure that all the downloaded entries contain the GO annotation. Applying this procedure resulted in 53 GO terms including 25 cellular component terms, 2 molecular function terms and 26 biological process terms (Table S1). In addition, we queried 'subcellular localization' with the keyword 'kinetochore' in UniProt and downloaded the entries with published experimental evidence from the search results. After the removal of redundant entries, the resulting dataset contained a total of 1,554 carefully reviewed entries. The detailed procedures of database construction and data collection are illustrated in Figure 1 and a statistical summary can be found in Figure 2 and Table 2, respectively.

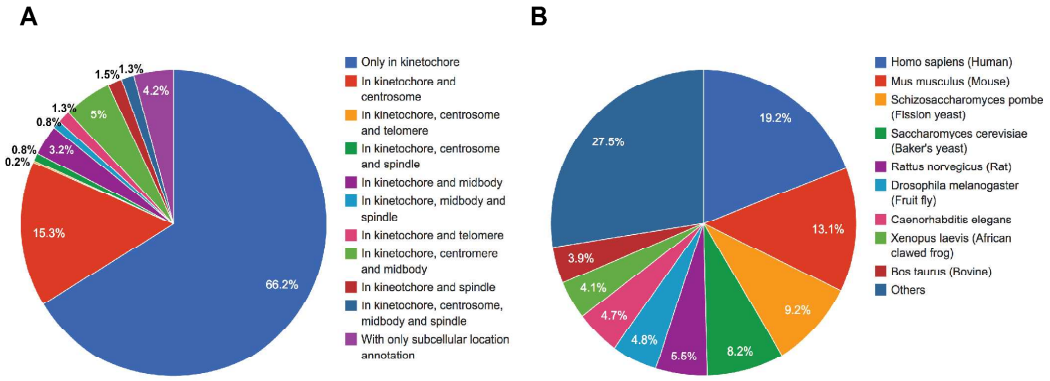


Figure 2. Statistical summary of locations and species of KinetochoreDB entries. (**A**) Distribution of protein locations according to MiCroKiTS and protein subcellular location annotations from UniProt. (**B**) Distribution of species of all KinetochoreDB entries.

(Table 2)

For each entry, KinetochoreDB integrates several public resources, including the UniProt database, RCSB Protein Data Bank (PDB) (20), OMIM (21), BioGRID 3.2 (22), Pfam database (23) and KEGG (24), in order to provide a comprehensive description with respect to basic protein information, protein structure, function, mutation and evolutionary conservation. An important feature of KinetochoreDB is the provision of 3D structure. To achieve this, we manually searched all the entries against the Protein Data Bank database using their corresponding UniProt identifiers and protein names. For protein complex structures, we identified the PDB chain for each entry and annotated the entry with that chain. In addition, we also generated MSAs using all homologous sequences for each protein entry. Homologous sequences were retrieved by PSI-BLAST (25) search against the Swiss-Prot dataset obtained from UniProt. The alignments were generated using Clustal Omega (26). We also predicted natively disordered regions for all protein entries using one of the most widely used disorder predictors, namely VSL2B (27). A residue is annotated as disordered by VSL2B if its prediction score was greater than 0.5.

We used Jmol (http://jmol.sourceforge.net/) and pViz (28) for visualization of protein structures, and Jalview (29) for customizable editing and display of MSAs for each protein entry. The information stored in KinetochoreDB resides in a MySQL relational database. A highly interactive web front-end to the data was implemented using the Javascript framework, JQuery. Apache Tomcat handles serving of data to

users on the web, utilizing a set of Java Servlets and JavaServer Pages (JSP) for data searching and viewing.

Database utility

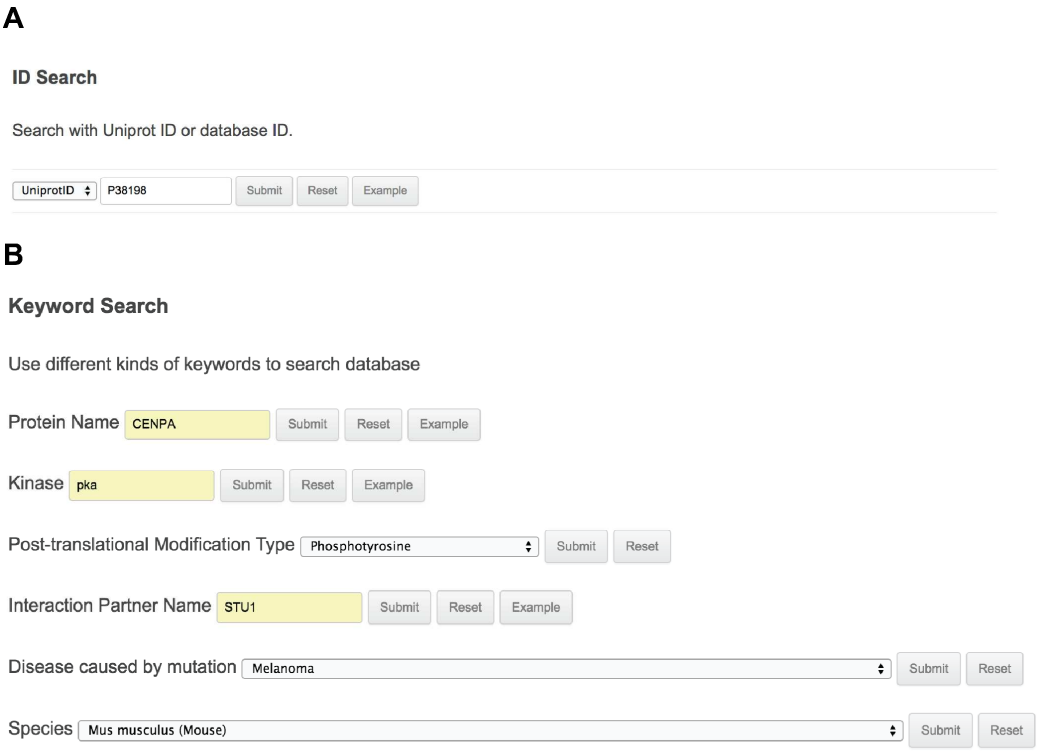


Figure 3. The search interface and options provided by KinetochoreDB. (**A**) Protein ID search with either UniProt ID or KinetochoreDB ID. (**B**) Keyword search with the protein name, kinase, PTM type, interaction partner name, disease or species.

The 'Search' page (http://lightning.med.monash.edu/kinetochoreDB2/Search.jsp) (Figure 3) allows users to search the database in several different ways. These search options can be generally classified into two groups: search with ID or search with keywords. Examples are provided below to assist users to understand how to perform the search. When searching the database with IDs, KinetochoreDB provides two different kinds of IDs to facilitate the search: UniProt ID and KinetochoreDB ID. The latter is composed of 'KD' and five digits, e.g. KD00095. As there are a total of 1,554 entries in the database, the database ID ranges from KD00001 to KD01554. In addition KinetochoreDB offers alternative search options with keywords. These

include protein name, kinase name, PTM type, name of protein interaction partner and name of diseases caused by single point mutations (Figure 3B). After selecting the 'Submit' button, the corresponding search results will be shown on the webpage. For each entry, there are generally nine sections of structural and functional categories, including general information, protein structure, disordered regions prediction, interaction partner, PTMs, Pfam domain, protein mutation, metabolic/signaling pathway and protein alignment with homologs. To provide an illustration of the annotations for each entry in KinetochoreDB, we use 'UniProt ID = O14965' (KinetochoreDB ID = 'KD01531') as an example query. The resulting page with nine sections is shown in Figure 4.

Figure 4. An example of search results in KinetochoreDB using the UniProtID O14965 as the query. The results are summarized and displayed in nine sections including protein information, protein structure, metabolic/signaling pathway, protein interaction, PTMs, Pfam domain, disorder region prediction, protein mutation and multiple sequence alignment.

For protein overview, KinetochoreDB uses pViz (28) to facilitate the general description of protein entries including functional sites and domains (Figure S1A), allowing a more detailed inspection of the constituent domains of the protein. For protein structure overview, KinetochoreDB provides two different ways to inspect the

3D structures. A single structure for the current protein entry can be examined by clicking the 'View' button to launch Jmol, a Java applet for displaying 3D structures (Figure S1C). Multiple structures can also be viewed together as an ensemble using pViz (Figure S1B).

With respect to protein-protein interaction, the interaction partner is highlighted if this protein is also an entry of KinetochoreDB. As a result of our search strategy (see 'database construction and features' for details), certain proteins in MSAs might not be included in the current KinetochoreDB. To facilitate the comparison between entries in KinetochoreDB and their homologs, we archived the homologs by extracting protein UniProt IDs. Detailed information for these files is available in the 'Protein alignment' section of the webpage.

KinetochoreDB will be updated on a regular basis to include newly available entries from various databases, in order to allow an up-to-date archive of recent results of the kinetochore and its related proteins. To this end, we allow users to submit new sequences and their structural and functional annotations to KinetochoreDB (Figure S2). After careful review and verification, new data will be included in KinetochoreDB and made publically available.

Discussion

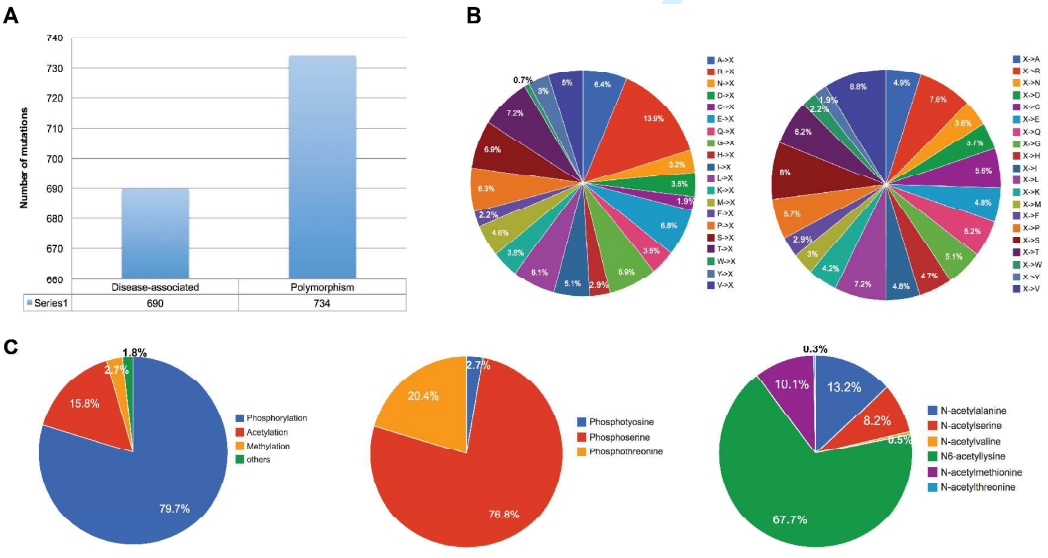


Figure 5. Statistical analysis of single point mutations and protein PTM types in KinetochoreDB. (**A**) Distribution of disease-associated mutations and polymorphisms. (**B**) Distribution of the number of mutations according to two mutation patterns (i.e. (A…V)->X

and X-> (A…V), where X denotes any type of amino acid). (**C**) Distribution of different major types of protein PTM, e.g. phosphorylation, acetylation, methylation and others. The distribution of sub-types of phosphorylation and acetylation is also shown.

Some of the protein entries in KinetochoreDB harbour mutations. We therefore provide a brief statistical analysis of the mutations. There are 1,424 mutations occurring in the 206 entries in KinetochoreDB. Among these, 690 (48.5%) mutations cause diseases while 1,283 (51.5%) are nonsense mutations (Figure 5A). We plotted the distributions of different types of mutations in Figure 5B. It can be noticed that there is no apprarent difference between the two types of mutation patterns (Figure 5B). PTMs, on the other hand, extend the chemical repertoire of amino acids by attaching new chemical groups and small molecules to the side chains of amino acids. Based on the available PTM annotations in KinetochoreDB, we further analysed the distribution of different types of PTMs for all the entries in KinetochoreDB. We noticed that kinetochore and its related proteins possess many PTM sites, the top three of which are phosphorylation, acetylation and methylation (Figure 5C). The distribution of different sub-types of acetylation and phosphorylation is also shown in Figure 5C.

In addition, with the comprehensive dataset from KinetochoreDB, we conducted a statistical analysis of the number of proteins involved in different GO terms including cellular component, molecular function and biological process. The results are shown in Figure 6. For cellular component, the 10 top ranked GO terms are condensed chromosome kinetochore (GO:0000777), kinetochore (GO:0000776), condensed nucler choromosome kinetochore (GO:0000778), cytoplasmic dynein complex (GO:0005868), condensed nuclear chromosome, centromeric region (GO:0000780), Ndc80 complex (GO:0031262), DASH complex (GO:0042729), Chromosome passenger complex (GO:0032133), Condensed chromosome outer kinetochore (GO:0000940) and kinetochore microtubule (GO:00005828). For molecular functioin, the top ranked GO terms are microtube motor activity (GO:0003777) and kinetochore binding (GO:0043515). For biological process, the 10 top ranked GO terms are protein localization to kinetochore (GO:0034501), attachment of spindle microtubules to kinetochore (GO:0008608), kinetochore assembly (GO:0051382), attachment of mitotic spindle microtubules to kinetochore (GO:0051315), attachment of spindle microtubules to kinetochore involved in homologous chromsome segregation

(GO:0051455), centromere complexe assembly (GO:0034508), positive regulation of attachment of spindle microtubules to kinetochore (GO:0051987), sister chromatid biorientation (GO:0031134), regulation of attachment of spindle microtubules to kinetochore (GO:0051988) and kinetochore organization (GO:0051383). More specifically, 414, 285 and 72 proteins contain the annotation of condensed chromosome kinetochore (GO:0000777), microtubule motor activity (GO:0003777) and protein localization to kinetochore (GO:0034501).

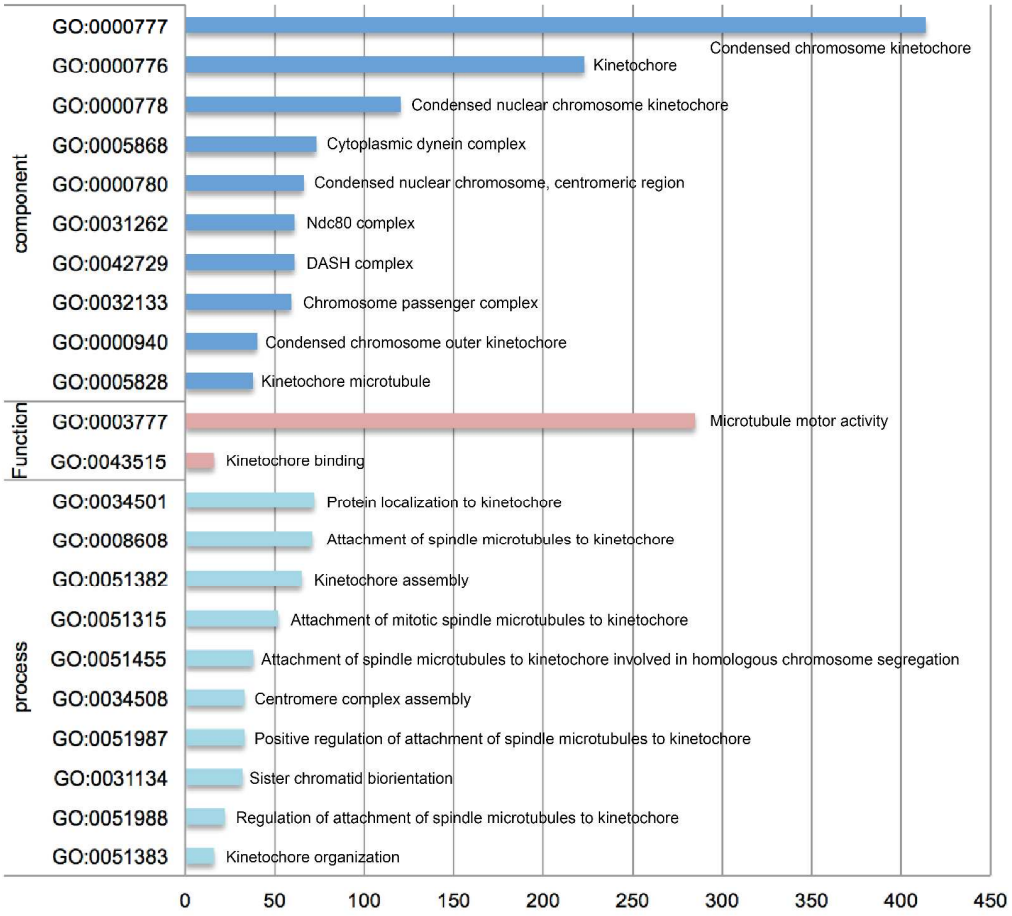


Figure 6. Statistical analysis of the number of entries from KinetochoreDB involved in different GO terms, which were grouped according to three categories: cellular component, molecular function and cellular process.

The kinetochore and its related proteins play extremely important roles during cell division and mitosis. In the past few decades, research on this topic has attracted a great deal of interest, not only because they are important for the cell cycle, mitosis

and meiosis (1-6), but also because they harbor mutations that can cause human diseases (12-15). In this context, databases such as KinetochoreDB that provide comprehensive annotations on the repertoire of kinetochore-related proteins will greatly facilitate in-depth functional investigation of these proteins and their relationships with human diseases. Through effective data integration from multiple public resources, KinetochoreDB has collected large amounts of information for related protein entries with respect to their amino acid sequence, protein 3D structure, biological function and evolutionary conservation. By providing comprehensive functional annotations of all available kinetochore-related proteins, we believe that this online resource will be used as a powerful tool to bridge functional characterization and disease-associated mutation studies of this important class of proteins.

In the future, we will endeavor to improve and update the annotations and analysis of data entries in KinetochoreDB by the following means: (1) We will keep the database updated and provide up-to-date information to synchronize with the research progress in the kinetochore and its related proteins; (2) We will integrate genomic information into our database and source these data from publicly available information or bioinformatics programs. These include coding sequence, transcription factor binding site, enhancer, promoter and other upstream or downstream regulatory information; (3) We will combine other state-of-the-art predictors to annotate the natively disordered regions of all entries in the database, while highlighting the consensus prediction. Meanwhile, we will also collect experimentally verified disordered regions from DisProt (30), the most comprehensive resource dedicated to annotating disordered region of proteins; (4) We will encourage experimental biologists to contribute to the development of KinetochoreDB by submitting their recent findings by making available newly added entries in the database after careful review.

In addition we will continue to improve and update the annotations and analysis of all entries in KinetochoreDB by implementing secondary analysis functions of the database and by integrating high-throughput experimental data. In particular, we will explore gene expression microarray data, transcriptomics and proteomics and functional pathway data, so as to provide a comprehensive useful resource for the wider research community.

Funding

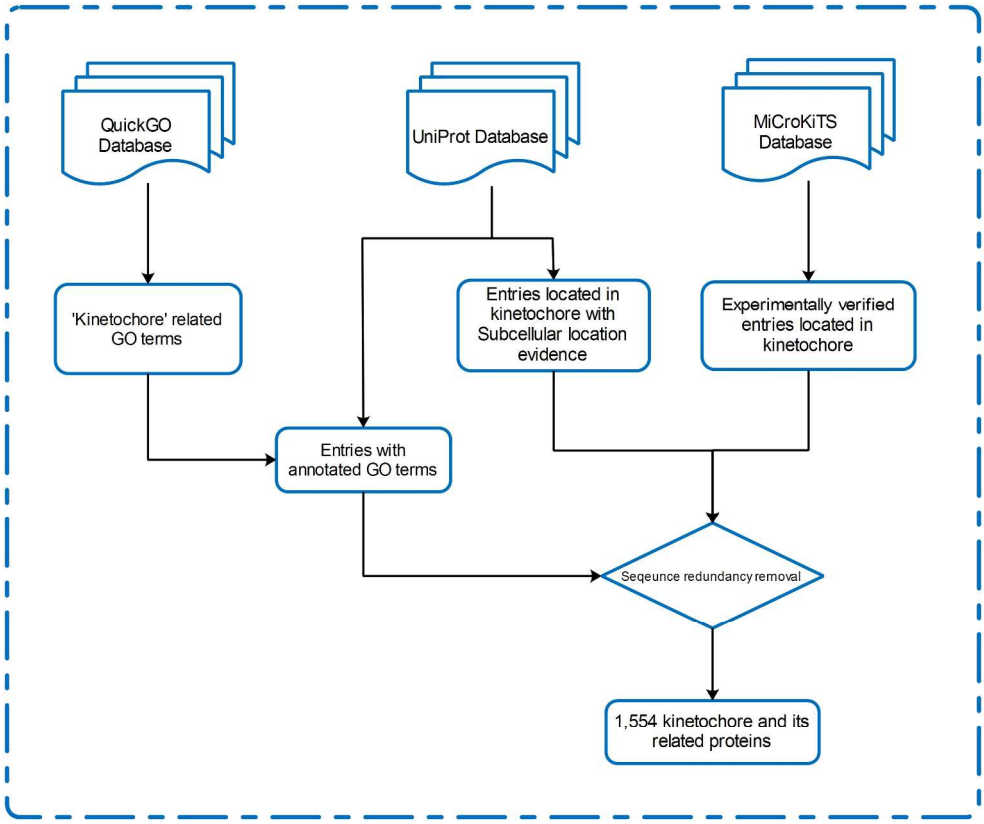
This work was supported by grants from the National Natural Science Foundation of China (61202167, 61303169), the Hundred Talents Program of the Chinese Academy of Sciences (CAS), the Knowledge Innovative Program of CAS (KSCX2-EW-G-8) of CAS, and the National Health and Medical Research Council of Australia (NHMRC). JS is a Recipient of the Hundred Talents Program of CAS. AMB is an NHMRC Senior Research Fellow (1022688).

Conflict of Interest: none declared.

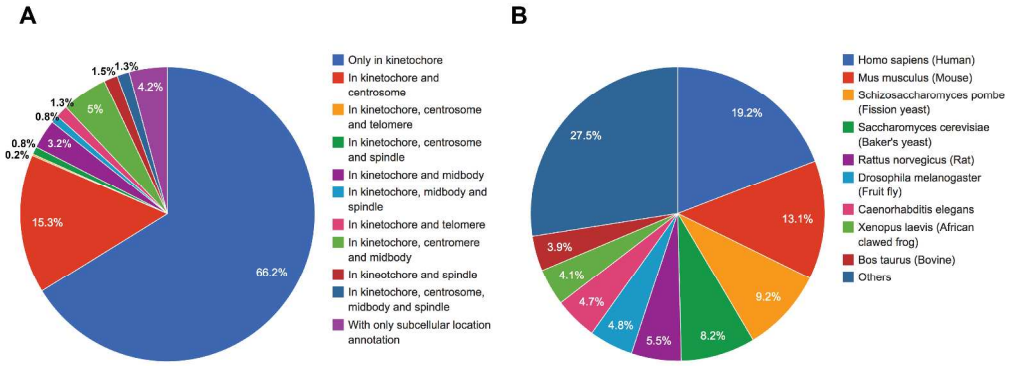
References

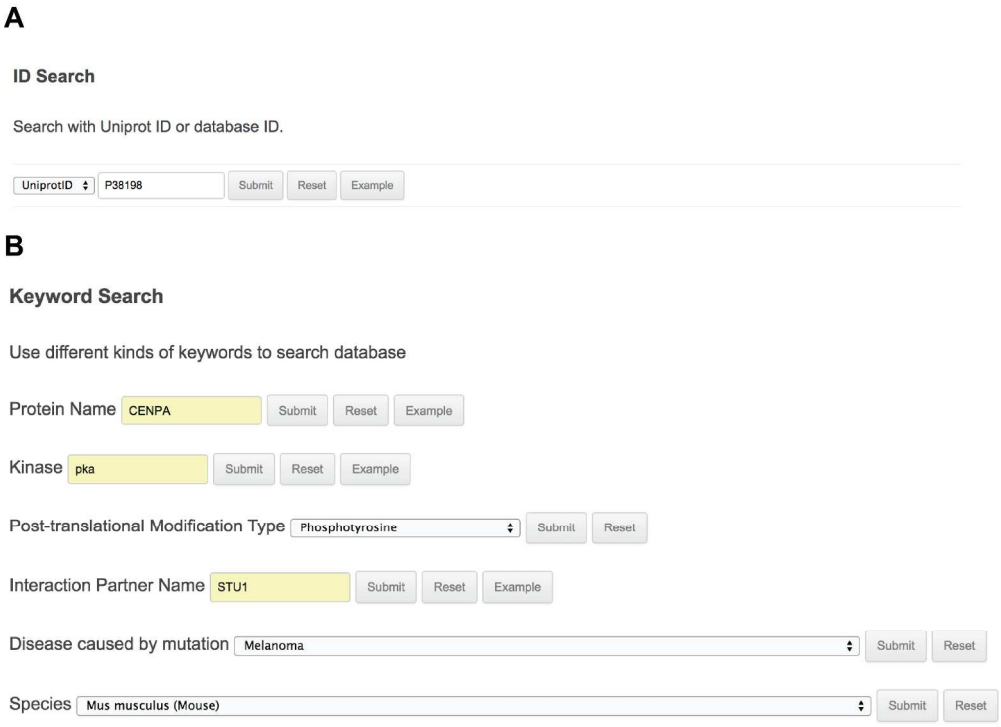
1. Brinkley, B.R., Tousson, A., Valdivia, M.M. (1985) The kinetochore of mammalian chromosomes: structure and function in normal mitosis and aneuploidy. *Basic Life Sci.*, **36**, 243-267.
2. Chan, G.K., Liu, S.T., Yen, T.J. (2005) Kinetochore structure and function. *Trends Cell Biol.*, **15**, 589-598.
3. McAinsh, A.D., Tytell, J.D., Sorger, P.K. (2003) Structure, function, and regulation of budding yeast kinetochores. *Annu. Rev. Ccell Dev. Biol.*, **19**, 519-539.
4. Westermann, S., Drubin, D.G., Barnes, G. (2007) Structures and functions of yeast kinetochore complexes. *Annu. Rev. Biochem.*, **76**, 563-591.
5. Stankovic, A., Jansen, L.E. (2013) Reductionism at the vertebrate kinetochore. *J. Cell Biol.*, **200**, 7-8.
6. Rago, F., Cheeseman, I.M. (2013) Review series: The functions and consequences of force at kinetochores. *J. Cell Biol.*, **200**, 557-565.
7. Yang, Y., Wu, F., Ward, T.*, et al.* (2008) Phosphorylation of HsMis13 by Aurora B kinase is essential for assembly of functional kinetochore. *J. Biol. Chem.*, **283**, 26726-26736.
8. Wan, X., O'Quinn, R.P., Pierce, H.L.*, et al.* (2009) Protein architecture of the human kinetochore microtubule attachment site. *Cell*, **137**, 672-684.
9. Sakuno, T., Tada, K., Watanabe, Y. (2009) Kinetochore geometry defined by cohesion within the centromere. *Nature*, **458**, 852-858.
10. Tanaka, T.U., Desai, A. (2008) Kinetochore-microtubule interactions: the means to the end. *Curr. Opin Cell Biol.*, **20**, 53-63.
11. Bakhoum, S.F., Thompson, S.L., Manning, A.L.*, et al.* (2009) Genome stability is ensured by temporal control of kinetochore-microtubule dynamics. *Nat. Cell Biol.*, **11**, 27-35.
12. Huang, H., Mahler-Araujo, B.M., Sankila, A.*, et al.* (2000) APC mutations in sporadic medulloblastomas. *Am. J. Pathol.*, **156**, 433-437.
13. Miyaki, M., Nishio, J., Konishi, M.*, et al.* (1997) Drastic genetic instability of tumors and normal tissues in Turcot syndrome. *Oncogene*, **15**, 2877-2881.
14. Stella, A., Montera, M., Resta, N.*, et al.* (1994) Four novel mutations of the APC (adenomatous polyposis coli) gene in FAP patients. *Hum. Mol. Genet.*, **3**, 1687-1688.

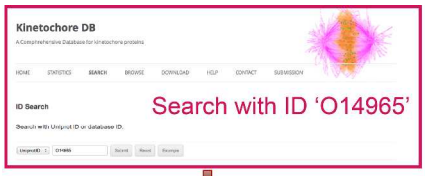
15. van der Luijt, R.B., Khan, P.M., Vasen, H.F., *et al.* (1997) Molecular analysis of the APC gene in 105 Dutch kindreds with familial adenomatous polyposis: 67 germline mutations identified by DGGE, PTT, and southern analysis. *Hum. Mut.*, **9**, 7-16.
16. Kops, G.J., Weaver, B.A., Cleveland, D.W. (2005) On the road to cancer: aneuploidy and the mitotic checkpoint. *Nat. Rev. Cancer*, **5**, 773-785.
17. Huang, Z., Ma, L., Wang, Y., *et al.* (2015) MiCroKiTS 4.0: a database of midbody, centrosome, kinetochore, telomere and spindle. *Nucleic Acids Res.*, **43**, D328-334.
18. Huntley, R.P., Sawford, T., Mutowo-Meullenet, P., *et al.* (2015) The GOA database: gene Ontology annotation updates for 2015. *Nucleic Acids Res.*, **43**, D1057-1063.
19. Consortium, T.U. (2015) UniProt: a hub for protein information. *Nucleic Acids Res.*, **43**, D204-212.
20. Rose, P.W., Beran, B., Bi, C., *et al.* (2011) The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Res.*, **39**, D392-401.
21. Hamosh, A., Scott, A.F., Amberger, J.S., *et al.* (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **33**, D514-517.
22. Chatr-Aryamontri, A., Breitkreutz, B.J., Heinicke, S., *et al.* (2013) The BioGRID interaction database: 2013 update. *Nucleic Acids Res.*, **41**, D816-823.
23. Finn, R.D., Bateman, A., Clements, J., *et al.* (2014) Pfam: the protein families database. *Nucleic Acids Res.*, **42**, D222-230.
24. Kanehisa, M., Goto, S., Sato, Y., *et al.* (2014) Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.*, **42**, D199-205.
25. Altschul, S.F., Madden, T.L., Schaffer, A.A., *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389-3402.
26. Sievers, F., Wilm, A., Dineen, D., *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, **7**, 539.
27. Peng, K., Radivojac, P., Vucetic, S., *et al.* (2006) Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics*, **7**, 208.
28. Mukhyala, K., Masselot, A. (2014) Visualization of protein sequence features using JavaScript and SVG with pViz.js. *Bioinformatics*, **30**, 3408-3409.
29. Waterhouse, A.M., Procter, J.B., Martin, D.M., *et al.* (2009) Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**, 1189-1191.
30. Vucetic, S., Obradovic, Z., Vacic, V., *et al.* (2005) DisProt: a database of protein disorder. *Bioinformatics*, **21**, 137-140.

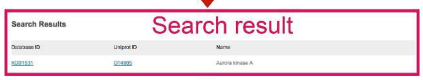


250x210mm (300 x 300 DPI)

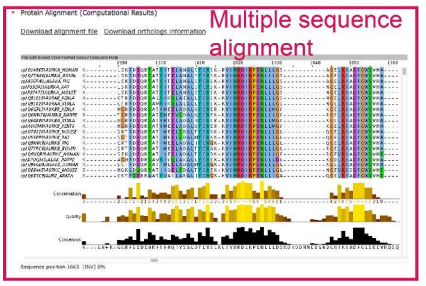



A
ID Search
Search with Uniprot ID or database ID.
UniprotID
P38198
Submit
Reset
Example
B
Keyword Search
Use different kinds of keywords to search database
Protein Name
CENPA
Submit
Reset
Example
Kinase
pka
Submit
Reset
Example
Post-translational Modification Type
Phosphotyrosine
Submit
Reset
Interaction Partner Name
STU1
Submit
Reset
Example
Disease caused by mutation
Melanoma
Submit
Reset
Species
Mus musculus (Mouse)
Submit
Reset

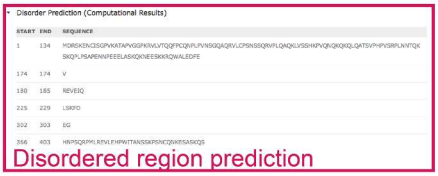

Kinetochore DB
Search with ID 'O14965'

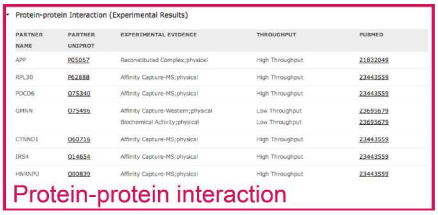

Search Results
Search result

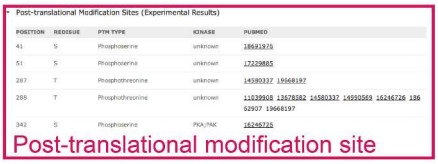

Detailed Information (Experimental Results)
Detailed information

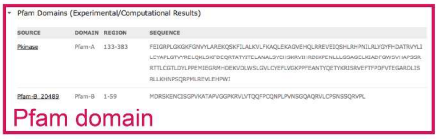

Protein Alignment (Computational Results)
Multiple sequence alignment

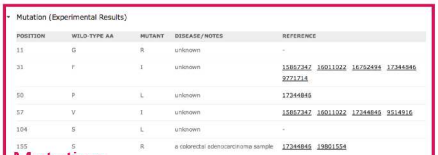

Protein Structure (Experimental Results)
Protein structure


Disorder Prediction (Computational Results)
Disordered region prediction

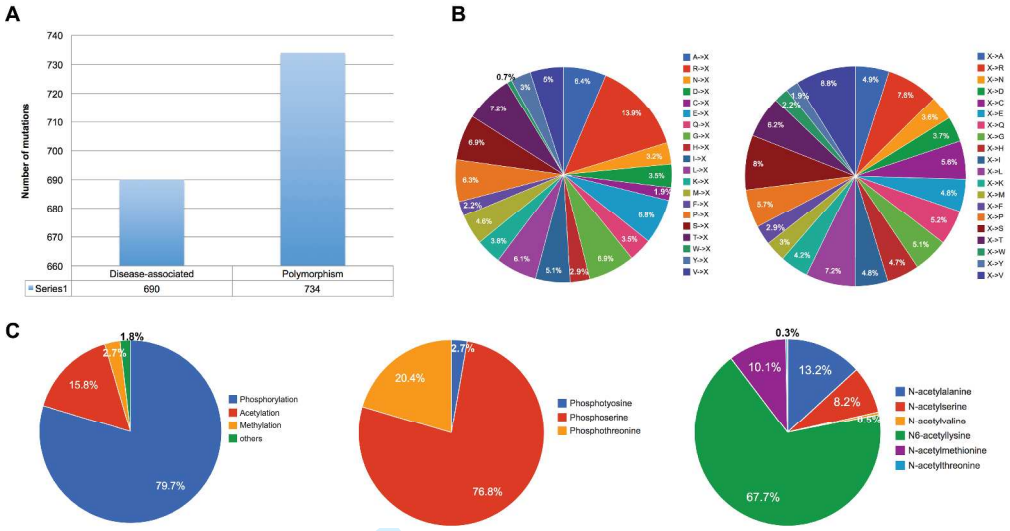

Protein-protein Interaction (Experimental Results)
Protein-protein interaction


Post-translational Modification Sites (Experimental Results)
Post-translational modification site


Pfam Domains (Experimental/Computational Results)
Pfam domain


Mutation (Experimental Results)
Mutation


Pathway (Experimental Results)
Pathway


A
Number of mutations
740
730
720
710
700
690
680
670
660
Disease-associated
Polymorphism
Series1
690
734
B
A->X
R->X
N->X
D->X
C->X
E->X
Q->X
G->X
H->X
I->X
L->X
K->X
M->X
F->X
P->X
S->X
T->X
W->X
Y->X
V->X
X->A
X->R
X->N
X->D
X->C
X->E
X->Q
X->G
X->H
X->I
X->L
X->K
X->M
X->F
X->P
X->S
X->T
X->W
X->Y
X->V
C
79.7%
15.8%
2.7%
1.8%
Phosphorylation
Acetylation
Methylation
others
76.8%
20.4%
2.7%
Phosphotyosine
Phosphoserine
Phosphothreonine
67.7%
10.1%
13.2%
8.2%
0.3%
N-acetylalanine
N-acetylserine
N-acetylvaline
N6-acetyllysine
N-acetylmethionine
N-acetylthreonine

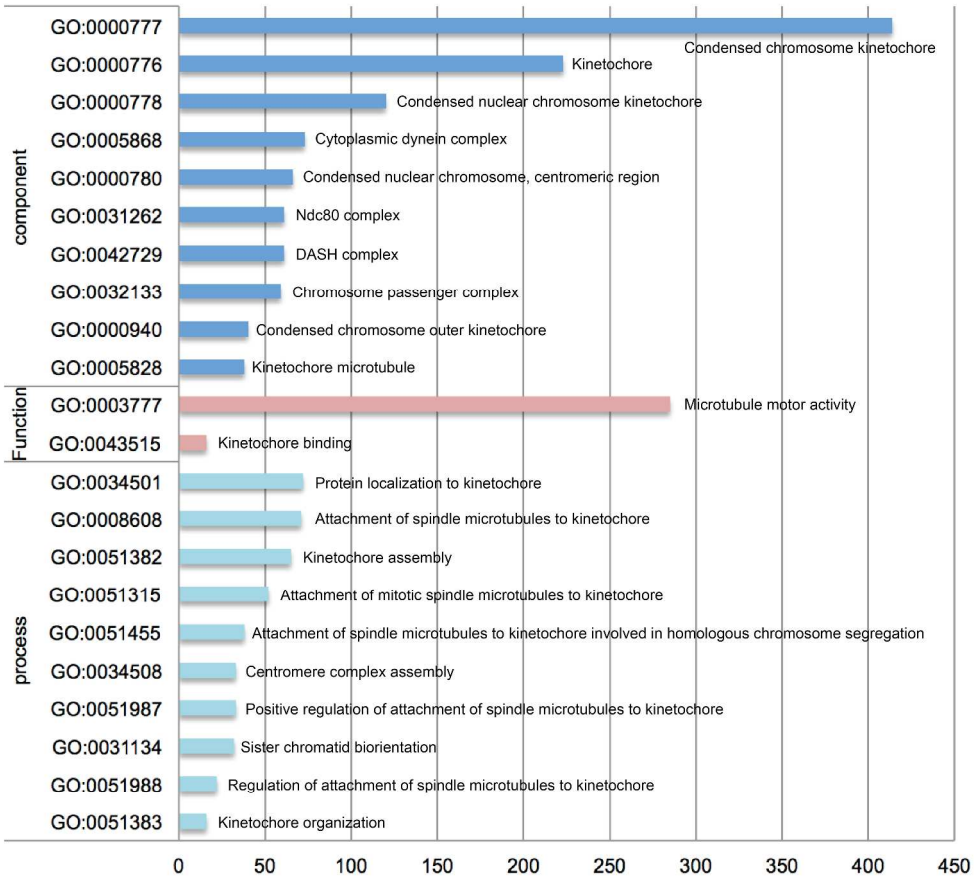

component
GO:0000777
Condensed chromosome kinetochore
GO:0000776
Kinetochore
GO:0000778
Condensed nuclear chromosome kinetochore
GO:0005868
Cytoplasmic dynein complex
GO:0000780
Condensed nuclear chromosome, centromeric region
GO:0031262
Ndc80 complex
GO:0042729
DASH complex
GO:0032133
Chromosome passenger complex
GO:0000940
Condensed chromosome outer kinetochore
GO:0005828
Kinetochore microtubule
Function
GO:0003777
Microtubule motor activity
GO:0043515
Kinetochore binding
process
GO:0034501
Protein localization to kinetochore
GO:0008608
Attachment of spindle microtubules to kinetochore
GO:0051382
Kinetochore assembly
GO:0051315
Attachment of mitotic spindle microtubules to kinetochore
GO:0051455
Attachment of spindle microtubules to kinetochore involved in homologous chromosome segregation
GO:0034508
Centromere complex assembly
GO:0051987
Positive regulation of attachment of spindle microtubules to kinetochore
GO:0031134
Sister chromatid biorientation
GO:0051988
Regulation of attachment of spindle microtubules to kinetochore
GO:0051383
Kinetochore organization
0
50
100
150
200
250
300
350
400
450

Table 1. Comparison between KinetochoreDB and MiCroKiTS.

Annotation category	KinetochoreDB	MiCroKiTS
Target	Kinetochore and its related proteins	Kinetochore, centrosome, midbody, telomere and spindle proteins
Protein 3D structure	Yes, detailed structural information available; customizable display	No
Protein intrinsic disorder	Yes, predicted by VSL2B	No
Protein interaction partner	Yes, detailed information available	No
Metabolic/signaling pathway	Yes	No
Disease-associated mutations	Yes, incorporating both disease-associated and nonsense mutations	No
Evolutionary conservation	Yes, multiple sequence alignments curated and visualized using Jalview	No
User enquiry and submission	Yes	No

Table 2. Statistical summary of the information contained in KinetochoreDB.

Number of entries	1,554
Number of protein structures	1,232
Number of protein interactions	49,931
Number of mutations	1,424
Number of KEGG pathways	452
Number of Pfam domains	4,145
Number of post-translational modification sites (PTMs)	4,027

KinetochoreDB: a comprehensive online resource for the kinetochore and its related proteins

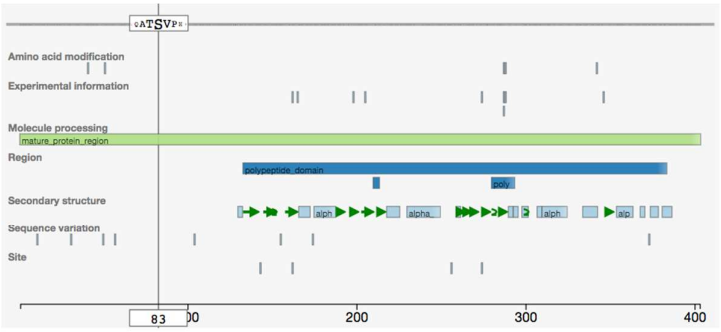
Supplementary Material

Table S1. GO terms selected in KinetochoreDB using the keyword 'kinetochore' from the QuickGO database

Aspect	GO ID	Name
Component	GO:0000776	kinetochore
Component	GO:0000777	condensed chromosome kinetochore
Component	GO:0005828	kinetochore microtubule
Component	GO:0000939	condensed chromosome inner kinetochore
Component	GO:0000940	condensed chromosome outer kinetochore
Component	GO:0000778	condensed nuclear chromosome kinetochore
Component	GO:0000941	condensed nuclear chromosome inner kinetochore
Component	GO:0000942	condensed nuclear chromosome outer kinetochore
Component	GO:0031617	NMS complex
Component	GO:0042729	DASH complex
Component	GO:0005818	aster
Component	GO:1990423	RZZ complex
Component	GO:0000817	COMA complex
Component	GO:0031518	CBF3 complex
Component	GO:0031262	Ndc80 complex
Component	GO:0033551	monopolin complex
Component	GO:0044816	Nsk1-Dlc1 complex
Component	GO:1990298	bub1-bub3 complex
Component	GO:0000444	MIS12/MIND type complex
Component	GO:0000818	nuclear MIS12/MIND complex
Component	GO:0005868	cytoplasmic dynein complex
Component	GO:0061638	CENP-A containing chromatin
Component	GO:0032133	chromosome passenger complex
Component	GO:0000779	condensed chromosome, centromeric region
Component	GO:0000780	condensed nuclear chromosome, centromeric region
Function	GO:0043515	kinetochore binding
Function	GO:0003777	microtubule motor activity
Process	GO:0051382	kinetochore assembly
Process	GO:0051383	kinetochore organization
Process	GO:0090234	regulation of kinetochore assembly
Process	GO:0034501	protein localization to kinetochore
Process	GO:1990299	Bub1-Bub3 complex localization to kinetochore
Process	GO:0008608	attachment of spindle microtubules to kinetochore
Process	GO:0072356	chromosome passenger complex localization to kinetochore

Process	GO:0051315	attachment of mitotic spindle microtubules to kinetochore
Process	GO:0051988	regulation of attachment of spindle microtubules to kinetochore
Process	GO:1903394	protein localization to kinetochore involved in kinetochore assembly
Process	GO:0051987	positive regulation of attachment of spindle microtubules to kinetochore
Process	GO:0051316	attachment of spindle microtubules to kinetochore involved in meiotic chromosome segregation
Process	GO:0051455	attachment of spindle microtubules to kinetochore involved in homologous chromosome segregation
Process	GO:0051456	attachment of spindle microtubules to kinetochore involved in meiotic sister chromatid segregation
Process	GO:2000751	histone H3-T3 phosphorylation involved in chromosome passenger complex localization to kinetochore
Process	GO:1902423	regulation of attachment of spindle microtubules to kinetochore involved in mitotic sister chromatid segregation
Process	GO:2000817	regulation of histone H3-T3 phosphorylation involved in chromosome passenger complex localization to kinetochore
Process	GO:1902424	negative regulation of attachment of spindle microtubules to kinetochore involved in mitotic sister chromatid segregation
Process	GO:1902425	positive regulation of attachment of spindle microtubules to kinetochore involved in mitotic sister chromatid segregation
Process	GO:0098653	centromere clustering
Process	GO:0031134	sister chromatid biorientation
Process	GO:2000574	regulation of microtubule motor activity
Process	GO:0072766	centromere clustering at the nuclear periphery
Process	GO:2000575	negative regulation of microtubule motor activity
Process	GO:2000576	positive regulation of microtubule motor activity
Process	GO:0034508	centromere complex assembly

A



B

C

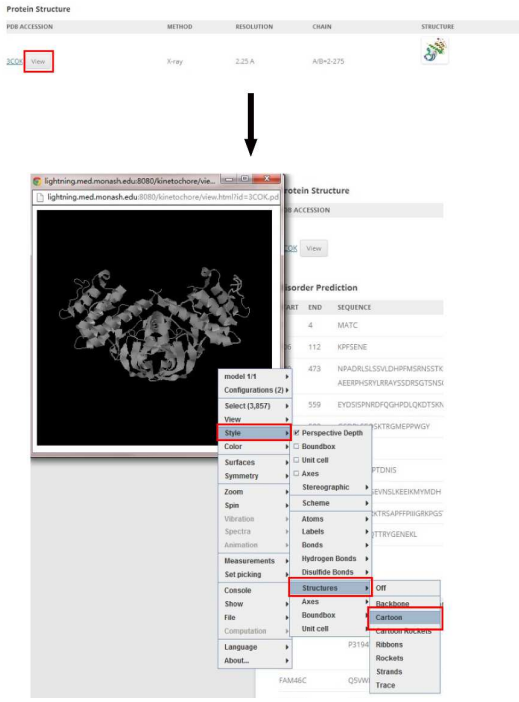


Figure S1. JQuery implementation for protein entries in KinetochoreDB. (**A**) Protein overview. (**B**) Protein structure view in an ensemble way with pViz. (**C**) Protein single structure view with Jmol.

Figure S2. Submission page for the users to add a new protein entry.

Briefings in Bioinformatics, 2015, 1–13

doi: 10.1093/bib/bbv047
Paper

Critical evaluation of *in silico* methods for prediction of coiled-coil domains in proteins

Chen Li, Catherine Ching Han Chang, Jeremy Nagel, Benjamin T. Porebski, Morihiro Hayashida, Tatsuya Akutsu, Jiangning Song and Ashley M. Buckle

Corresponding authors. Jiangning Song, Department of Biochemistry and Molecular Biology, Faculty of Medicine, Monash University, Melbourne, Victoria 3800, Australia. [redacted] Ashley M. Buckle, Department of Biochemistry and Molecular Biology, Faculty of Medicine, Monash University, Melbourne, Victoria 3800, Australia. [redacted]

Abstract

Coiled-coils refer to a bundle of helices coiled together like strands of a rope. It has been estimated that nearly 3% of protein-encoding regions of genes harbour coiled-coil domains (CCDs). Experimental studies have confirmed that CCDs play a fundamental role in subcellular infrastructure and controlling trafficking of eukaryotic cells. Given the importance of coiled-coils, multiple bioinformatics tools have been developed to facilitate the systematic and high-throughput prediction of CCDs in proteins. In this article, we review and compare 12 sequence-based bioinformatics approaches and tools for coiled-coil prediction. These approaches can be categorized into two classes: coiled-coil detection and coiled-coil oligomeric state prediction. We evaluated and compared these methods in terms of their input/output, algorithm, prediction performance, validation methods and software utility. All the independent testing data sets are available at http://lightning.med.monash.edu/coiledcoil/. In addition, we conducted a case study of nine human polyglutamine (PolyQ) disease-related proteins and predicted CCDs and oligomeric states using various predictors. Prediction results for CCDs were highly variable among different predictors. Only two peptides from two proteins were confirmed to be CCDs by majority voting. Both domains were predicted to form dimeric coiled-coils using oligomeric state prediction. We anticipate that this comprehensive analysis will be an insightful resource for structural biologists with limited prior experience in bioinformatics tools,

Chen Li received his M.Eng. in Computer Science from Northwest A&F University, China. He is currently pursuing his PhD in the Department of Biochemistry and Molecular Biology, Faculty of Medicine, Monash University. His research interests are structural bioinformatics, systems biology, data mining and machine learning.

Catherine Ching Han Chang received her degree in Chemical Engineering from Monash University Sunway Campus. She is currently pursuing PhD in Chemical Engineering in the Chemical Engineering Discipline, School of Engineering, Monash University, Malaysia. Her research interests include modelling of soluble recombinant protein expression in *Escherichia coli*.

Jeremy Nagel received his Bachelor in Environmental Science (Honours) from Monash University in 2011.

Benjamin T. Porebski received a BSc (Honours) in biochemistry and molecular biology from Monash University and is presently pursing a PhD in biochemistry in the Department of Biochemistry and Molecular Biology, Monash University, Australia. His research interests involve protein engineering using a wide spectrum of techniques including biophysics, protein crystallography and computational biology.

Morihiro Hayashida received his PhD degree in Informatics in 2005 from Kyoto University, Japan. He is an assistant professor in Bioinformatics Center, Institute for Chemical Research, Kyoto University, Japan. His research interests include functional analysis of proteins and development of computational methods.

Tatsuya Akutsu received his Dr. Eng. degree in Information Engineering in 1989 from University of Tokyo, Japan. Since 2001, he has been a professor in Bioinformatics Center, Institute for Chemical Research, Kyoto University, Japan. His research interests include bioinformatics and discrete algorithms.

Jiangning Song received his PhD degree in Bioinformatics in 2005 from Jiangnan University Wuxi China. He is a senior research fellow at the Monash Bioinformatics Platform, Faculty of Medicine, Monash University, Australia. He is also a principal investigator at the Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences. His research interests are bioinformatics, systems biology, machine learning, systems pharmacology and enzyme engineering.

Ashley M Buckle completed his PhD in Biochemistry in 1994 in the laboratory of Prof. Sir Alan Fersht at the University of Cambridge, UK. He is an NHMRC senior research fellow and group leader in the Department of Biochemistry and Molecular Biology, Monash University, Australia. His laboratory uses a multidisciplinary approach to understand protein structure, dynamics and function, using protein crystallography, biophysics and molecular simulation.

Submitted: 10 April 2015; **Received (in revised form):** 29 May 2015

and for bioinformaticians who are interested in designing novel approaches for coiled-coil and its oligomeric state prediction.

Key words: coiled-coil; prediction; oligomeric state; polyglutamine

Introduction

First described in 1953 by Pauling and Crick [1], the proliferation of studies of coiled-coil domains (CCDs) in proteins has driven continued computational prediction in the past few decades. CCDs can be summarized as at least two or more helices that wrap around each other, which can be defined as a repeat X_n of residues, where X can be denoted as (a-b-c-d-e-f-g) and n can be described as the number of helices. It is estimated that nearly 10% of eukaryotic proteins harbour CCDs [2, 3]. Depending on the value of n, CCDs can be categorized into several groups, including antiparallel dimer, parallel dimer, trimer and tetramer (Figure 1). The colour scheme in Figure 1 is based on the B-factor values using PyMOL. CCDs exhibit a preference for hydrophobic residues at positions a and d, charged residues at positions e and g and hydrophilic residues at positions b, c and f [8, 9], which serve to stabilize helix oligomerization according to the 'Peptide Velcro' hypothesis [10]. This repeating X_n motif enables the prediction of CCDs and their oligomeric states based on protein sequences.

Experimental studies have confirmed that CCDs play a fundamental role in subcellular infrastructure and controlling trafficking of eukaryotic cells [11, 12]. The relatively high stability of CCDs has led to their promising use as delivery systems for a range of molecules. For example, cartilage oligomeric matrix protein (COMP) [13, 14] and right-handed protein [15] from *Staphylothermus marinus* have been used as drug delivery systems in anticancer therapies [3, 16, 17]. The five α-helix CCDs in COMP are capable of binding and carrying some important signalling molecules, including vitamins A and D_3. Other successful applications of CCDs, peptides and motifs used in drug delivery systems have also been reported [18–22].

Sequence and structural analysis of CCDs have enabled the development of computational approaches for the prediction of CCDs from sequence alone [8–10, 23]. For example, Vincent *et al.* performed coiled-coil prediction for proteins from tenascins and thrombospondins families, analysed the motif conservation of different coiled-coil oligomeric states and revealed that sequence conservation allows trimers and pentamers of CCDs to be distinguished, providing useful insights for future coiled-coil prediction [23]. However, the rapid growth in prediction approaches since the last comprehensive comparison, which was reported almost a decade ago [24], creates an urgent need to critically assess and compare the now-large and diverse prediction methods. In this article, therefore, we present a comprehensive review of 12 sequence-based methods for coiled-coil prediction, offering insights into the nature of different predictors and facilitating potential improvement of CCD prediction. All predictors are critically reviewed in terms of input, model construction and outcome (i.e. prediction performance) [25, 26]. To evaluate the performance of coiled-coil predictors, independent tests were conducted with new test data sets (http://lightning.med.monash.edu/coiledcoil/) carefully collected and curated from different resources. Finally, as CCDs have been extensively found in disease-associated human polyglutamine (PolyQ) proteins [27], we applied various predictors to a data set of nine human proteins containing PolyQ repeats and discussed our findings.

Materials and methods

Predictors reviewed in this study

Table 1 summarizes the details of the tools of coiled-coil and its oligomeric state prediction that are evaluated in this article. These are COILS [28], PCOILS [29], Paircoil2 [30], SOSUIcoil [31], MARCOIL [32], CCHMM_PROF [33], SpiriCoil [34], SCORER 2.0 [35], LOGICOIL [36], PrOCoil [37], RFCoil [38] and Multicoil2 [39].

Model input

The training data set is used to build a computational model to learn potential patterns hidden in the data set. Before model construction, data collection and preprocessing of the training data set were performed. Data sets with too much noise or imbalanced distribution may lead to unsatisfactory prediction performance of the model. There are two main ways to collect the CCD data to build the model. In some studies, the CCDs were extracted with SCOP [40] and SOCKET [41], while other studies extracted the data directly from a publicly available database regarding experimentally verified CCDs, for example, CC+ [42]. The CCDs in the CC+ database were annotated manually and with SOCKET, which has been widely used to extract

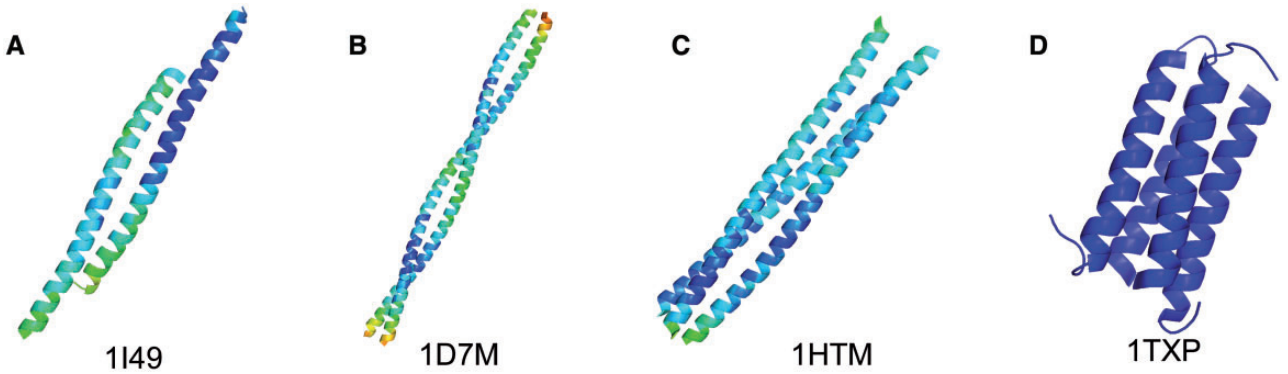


Figure 1. Examples of coiled-coil oligomeric states. (**A**) Antiparallel dimer (PDB Accession: 1I49 [4]). (B) Parallel dimer (PDB Accession: 1D7M [5]). (**C**) Trimer (PDB Accession: 1HTM [6]). (**D**) Tetramer (PDB Accession: 1TXP [7]).

Table 1. A comprehensive list of coiled-coil and oligomeric state predictors reviewed in this study

Task	Tool publication date	Input format	Model highlight[a]	Evaluation		Service			
				Strategy	Output format	Web service[b]	Availability	Speed[c]	Reliability[d]
Coiled-coil region prediction	COILS [28] 1997	Raw sequence or SwissProt IDs	Pairwise residue probabilities	Algorithm tested with sequences of known globular proteins, randomly generated sequences and all the sequences in GenBank	Residue score and probability located in CCD	Yes	Yes (with third-party implementations)	Fast	Consistent
	PCOILS [29] 2005	Raw sequence/ FASTA sequence	Pairwise profile comparison using protein evolution profile	Case study	Residue score and probability to located in CCD	Yes	Yes	Moderate	Results vary depending on BLAST database
	Paircoil2 [30] 2006	FASTA sequence	Pairwise residue probabilities	Leave-family-out cross-validation	Residue score and probability to located in CCD	Yes	Yes	Fast	Unknown[e]
	MARCOIL [32] 2002	FASTA Sequence	HMM based on MTIDK and other matrices	150-fold cross-validation	Residue score and probability to located in CCD	Yes	Yes	Fast	Consistent
	CCHMM_PROF [33] 2009	Raw sequence/ FASTA sequence	HMM based on multiple sequence alignment	Overall accuracy, Segment overlap and case study	Overall probability of containing CCDs and Binary decision to (not) be in CCD	Yes	Yes	Moderate	Results vary depending on the BLAST database
	SpiriCoil[f] [34] 2010	FASTA sequence	Structurally informed homology-based multiple HMMs	Independent test	Binary decision to (not) be in CCD	Yes	No	Fast	–
	SOSUIcoil [31] 2008	One-letter symbol or multiple FASTA sequences	Canonical discriminant analysis	Independent test and case study	–	No	No	–	–
Coiled-coil oligomeric	SCORER 2.0 [35] 2011	Raw sequence and/ or heptad register		Independent test	Predicted scorer to be parallel dimeric	Yes	Yes	Fast	Consistent

(continued)

Table 1. Continued

Task	Tool publication date	Input format	Model highlight[a]	Evaluation		Service			
				Strategy	Output format	Web service[b]	Availability	Speed[c]	Reliability[d]
state prediction			Log-likelihood ratio with new defined score function		and trimeric coiled-coil				
	LOGICOIL [36] 2013	Raw sequence and/or heptad register	Bayesian variable selection and multinomial probit regression	10-fold cross-validation and leave-one-out cross-validation	Predicted score to be parallel dimer, antiparallel dimer, trimer and tetramer	Yes	Yes	Fast	Consistent
	PrOCoil [37] 2011	Raw sequence and/or heptad register	SVM and coiled-coil Kernel	10-fold cross-validation, nested cross-validation and case study	Predicted scorer to be parallel dimeric and trimeric coiled-coil	Yes	Yes	Fast	Consistent
	RFCoil [38] 2014	Raw sequence and heptad register	Random forest with effective amino acid indices	10-fold cross-validation and independent tests	Predicted probability to be parallel dimeric and trimeric coiled-coil	Yes	Yes	Fast	Consistent
Coiled-coil region and oligomeric state prediction	Multicoil2 [39] 2011	FASTA sequence	Pairwise residue correlation and HMM	Leave-family-out cross-validation	Residue probability to be located in non-coiled-coil, dimer or trimer	Yes	Yes	Fast	Consistent

Note. [a]HMM—Hidden Markov Model; SVM—Support Vector Machines.

[b]The URLs of predictors listed are: COILS—http://embnet.vital-it.ch/software/COILS_form.html; PCOILS—http://toolkit.tuebingen.mpg.de/pcoils; PairCoil2—http://groups.csail.mit.edu/cb/paircoil2/paircoil2.html; MARCOIL—http://toolkit.tuebingen.mpg.de/marcoil; SOSUIcoil—http://harrier.nagahama-i-bio.ac.jp/sosui/coil/submit.html (not available); CCHMM_PROF—http://gpcr.biocomp.unibo.it/cgi/predictors/cchmmprof/pred_cchmmprof.cgi; SpiriCoil – http://supfam.cs.bris.ac.uk/SUPERFAMILY/spiricoil/; SCORER 2.0—http://coiledcoils.chm.bris.ac.uk/Scorer/; LOGICOIL—http://coiledcoils.chm.bris.ac.uk/LOGICOIL/; PrOCoil—http://www.bioinf.jku.at/software/procoil/; RFCoil—http://protein.cau.edu.cn/RFCoil/index.php?page=introduction; Multicoil2—http://groups.csail.mit.edu/cb/multicoil2/cgi-bin/multicoil2.cgi.

[c]Speed refers to the response time after submitting the sequence to the web server.

[d]Reliability refers to whether the outputs of the predictor's web server and its local executable are consistent.

[e]Paircoil2 is not runnable on our local machine.

[f]In [34], SpiriCoil was also applied for oligomeric state prediction. The prediction performance was comparable with that of MULTICOIL, which is the previous version of Multicoil2.

reliable CCDs from protein structures. A cut-off value of 7.0Å was usually used for extracting coiled-coils from protein structures. Removal of sequence redundancy, an important step before model construction, was performed using CD-HIT [43].

Models construction and development

Relatively simple classification methods predict whether a protein sequence contains a CCD. More sophisticated predictors perform multiclass classifications that categorize coiled-coil regions into different forms of α-helical assembly, such as dimer, trimer and tetramer. We discuss below the different algorithms used in the predictors (Table 1).

COILS, the first reported algorithm for CCD prediction, is a statistically controlled predictor based on the amino-acid profile-based method. The similarity of a protein sequence with a structurally known protein is computed using a sliding window. The recommended window length for COILS is 28 to help remove false positives. PCOILS is an updated version of COILS that predicts coiled-coils through comparing pairwise protein evolution profiles based on user-provided multiple sequence alignment or PSI-BLAST [44]. Paircoil2 is the latest development of PAIRCOIL [45]. These predictors use pairwise residue correlations or probabilities to detect the coiled-coil motif in a protein sequence. The training data set of Paircoil2 is larger than that used for training PAIRCOIL because of the dramatically increased number of known coiled-coil sequences. SOSUIcoil uses amino acid physical properties to help determine an appropriate heptad register, followed by canonical discriminant analysis to discriminate coiled-coils.

Hidden Markov Models (HMM) has been used in a number of coiled-coil predictors. These include MARCOIL, CCHMM_PROF and SpiriCoil. CCHMM_PROF is an improved version of CCHMM [46], which used multiple sequence alignments instead of single sequence-based HMM. MARCOIL also uses single sequence-based HMMs, whereas SpiriCoil uses a large library of HMMs to predict coiled-coils that fall into known superfamilies. The application of SpiriCoil is limited to sequences that have reasonably high similarity to known families because of the use of the training data set for constructing SpiriCoil. On the other hand, MARCOIL, which uses explicit knowledge of existing coiled-coils to train a single HMM, possesses a more complicated algorithm to efficiently search for a variable length subsequence of high probability for coiled-coil formation. According to the HMM parameter t, MARCOIL model has two variations, MARCOIL-L ($t = 0.001$) and MARCOIL-H ($t = 0.01$).

MultiCoil [47], a predictor developed based on the PAIRCOIL algorithm, extends the dimeric coiled-coil prediction in PAIRCOIL to trimeric coiled-coils, using a multidimensional scoring approach. Multicoil2 further extends the algorithm to include pairwise correlations with HMM in a Markov Random Field. Multicoil2 also contains eight sequence-based features (including dimer probability, trimer probability, non-coiled probability, dimer correlations at distance 1–7, trimer correlations at distance 1–7, non-coiled correlations at distance 1–7, the hydrophobicity at the *a* and *d* positions) that are used to train the model (pairwise correlation HMM). The resulting algorithm integrated the sequence features and the pairwise interactions into a multinomial logistic regression to formulate an optimized scoring function for the classification of coiled-coil oligomeric state.

SCORER [48] uses a log-odd-based scoring system for the classification of coiled-coil sequences into parallel dimeric and trimeric coiled-coils. SCORER 2.0 combines an expanded and updated training set and a Bayes factor method, which takes into consideration the possible uncertainty in the profile tables. LOGICOIL is a predictor based on the combined and concurrent application of Bayesian variable selection and multinomial probit regression. The application of Bayesian paradigm can provide informative posterior distributions on the selected parameters, as well as offering a framework to apply this useful information based on biological data and expert knowledge. Traditional machine learning techniques, including support vector machine (SVM) [49] and random forest [50], have also been applied to coiled-coil oligomeric state prediction. For example, PrOCoil adopts an SVM based on identified rules converted into weighted amino-acid patterns. In addition to PrOCoil, PrOCoil-BA (PrOCoil-Balanced Accuracy) is an alternative model, which is optimized for balanced accuracy, i.e. the average of sensitivity and specificity. RFCoil uses random forest combined with effective amino-acid indices selected by Gini (a decision tree split function) decrease [51] and Kendall rank correlation coefficient [52].

Model evaluation

A variety of methods were used to assess the prediction performance of coiled-coil predictors listed in Table 1, including cross-validation, leave-one-out cross-validation, leave-family-out cross-validation, independent test and case study. Normally, cross-validation can avoid over-fitting caused by the training data set. The nature of cross-validation is to split the data set into N folds and combine $N-1$ folds as the training data set, leaving the remaining fold as the test data set. Leave-one-out cross-validation and leave-family-out cross-validation are variations of cross-validation. Given a data set with D data samples, leave-one-out cross-validation combines $D-1$ samples as the training data set and leaves the remaining one sample as the test sample. In this cross-validation, all samples in the data set are treated as a test sample once. If the data set is collected from different species/families, each subset from the same species/family is regarded as test data sets once, and other subsets from other families/species will be combined to form the training data set. The final performance for cross-validation is often averaged from the results of different combinations of the training data sets. The independent test is another method to assess the performance of bioinformatics tools. To test the performance of an algorithm on a new data set with a different data distribution, it is important to ensure that there is no overlap between the training data set and the independent test data set. Finally, the case study is as an effective way to test the performance of a method in real-world applications, providing useful insights into the method scalability and usefulness with unknown data.

Predictor utility

An important aspect of predictors in the biological research community is to provide a user-friendly web interface or a local tool to enable non-bioinformaticians to apply the model directly to their research. The usefulness of bioinformatics tools depends on three factors, i.e. the web interface, the output and interpretation of prediction results and the availability of locally runnable software. A user-friendly interface can provide appropriate guidance and instructions to avoid potential mistakes when using the web server. This is especially important when parameter settings are required before conducting prediction tasks. Among the predictors we tested, those predictors aimed

at discriminating coiled-coils from non-coiled-coils (e.g. COILS, PCOILS, Paircoil2 and MARCOIL) require parameter settings before sequence submission. Documents are available online regarding the description of the parameters and their potential effect on the prediction performance. On the other hand, the predictors for coiled-coil oligomeric states are mostly parameter-free. For coiled-coil oligomeric state prediction, only sequence and its heptad register are required as the input (for example, SCORER 2.0, PrOCoil, RFCoil and LOGICOIL). Furthermore, SCORER 2.0, PrOCoil and LOGICOIL are also able to predict sequences without the prerequisite of knowing the coiled-coils/heptad registers by combing coiled-coil prediction and extracting heptad register from MARCOIL, without the necessity of performing a two-stage prediction.

Stand-alone software allows users to perform predictions for a large amount of sequences on local machines, offering an advantage over web servers. Among the coiled-coil predictors reviewed in this article, SpiriCoil and SOSUIcoil do not have available locally runnable tools. The local versions of SCORER 2.0, PrOCoil, RFCoil and LOGICOIL were written using the R package (http://www.r-project.org/). PrOCoil has been integrated with R so it can be downloaded and installed with the R console. Users should be aware of the difference in the length of the coiled-coils in the training data sets of different frameworks especially for the oligomeric state prediction. For SCORER 2.0, MultiCoil2, PrOCoil, RFCoil and LOGICOIL, the minimum lengths of their training coiled-coils are 15, 21, 8, 8 and 15, respectively. This means that one should take into consideration the length of the sequence when choosing appropriate predictors to obtain better prediction results. Although coiled-coil predictors recommend the preferable sequence lengths of coiled-coils, they can still predict the oligomeric state of the coiled-coils shorter than the specified length thresholds. Under such circumstance, it is the users' responsibility to choose an appropriate predictor according to the length of query sequence before its submission.

Understandable and visualizable interpretation of the output is also important for better understanding the prediction results and their significance. The output of the coiled-coil predictors we reviewed is often organized in two ways, based on either a residue or a sequence basis. Most of the predictors for discrimination of coiled-coils from non-coiled-coils provide prediction outputs on a residue basis, which allows users to gain a detailed insight into each amino acid and its predicted score/probability. Moreover, COILS, PCOILS, Paircoil2 and MARCOIL also provide the visible plots of predicted score/probability for each amino acid and enable users to obtain an overview of predicted scores for the entire sequence. On the other hand, the predictors of coiled-coil oligomeric state (including SCORER 2.0 and LOGICOIL) provide only a final decision and an overall prediction score. These scores are not easy to interpret and understand. PrOCoil provides both prediction scores and visible plots for each amino acid. RFCoil, on the other hand, provides a matrix showing the probability of the query sequence forming a dimeric coiled-coil or a trimeric coiled-coil, which is relatively easy to understand.

A case study of coiled-coil prediction for human PolyQ proteins

As an extended test of the reviewed coiled-coil predictors, we examined the prediction consistency for nine disease-associated PolyQ proteins. We submitted their sequences to the corresponding web servers and obtained the prediction results. PolyQ proteins contain a stretch of repeated glutamine residues (termed the 'PolyQ tract'). PolyQ repeats with more than seven residues are abundant in 128 proteins in the human proteome [53]. These repeats have important biological functions especially in transcription regulation, and proteins harbouring expanded PolyQ repeats are involved in neurodegenerative diseases [54]. The PolyQ diseases are caused in part by a gain-of-function mechanism of neuronal toxicity involving protein conformational changes that result in the formation and deposition of β-sheet rich aggregates [55]. Because PolyQ repeats are highly aggregation-prone [55], it is difficult to determine their structure by X-ray crystallography [56]. The widely accepted model of β-sheet-mediated aggregation has been recently challenged by experimental and bioinformatics studies showing that disease-associated PolyQ proteins contain CCDs largely overlapping with their PolyQ repeats [27]. We therefore investigated the prediction of CCDs in human proteins containing PolyQ repeats, using the data set containing the most updated nine disease-associated PolyQ proteins from UniProt database studied by Fiumara *et al.* [27], which is also available in the PolyQ database [53] (http://pxgrid.med.monash.edu.au/polyq/; Table 2).

Results and discussion

Independent test and performance evaluation

In this section, to assess the prediction performance of the reviewed coiled-coil tools in an objective and fair manner, we

Table 2. The list of nine human disease-related PolyQ proteins

Protein	Protein length	PolyQ tract	UniProt identifier	Associated disease
TATA binding protein	339	58–95	P20226	Spinocerebellar ataxia 17 [57–59]
Huntingtin	3142	18–38	P42858	Huntington disease [60]
Ataxin-1	815	197–208 212–225	P54253	Spinocerebellar ataxia 1 [61, 62]
Ataxin-2	1313	166–188	Q99700	Spinocerebellar ataxia 2 [63–65] and Amyotrophic lateral sclerosis 13 [66]
Voltage-dependent P/Q-type calcium channel subunit alpha-1A (Brain calcium channel I)	2505	2314–2324	O00555	Spinocerebellar ataxia 6 [67–70]
Atrophin-1	1190	484–502	P54259	Dentatorubro-pallidoluysian atrophy [71]
Ataxin 7	892	30–39	O15265	Spinocerebellar ataxia 7 [72]
Androgen receptor	919	58–78	P10275	Spinocerebellar muscular atrophy or Kennedy disease [73]
Ataxin-3	364	296–305	P54252	Spinocerebellar ataxia 3 or Machado-Joseph disease [74]

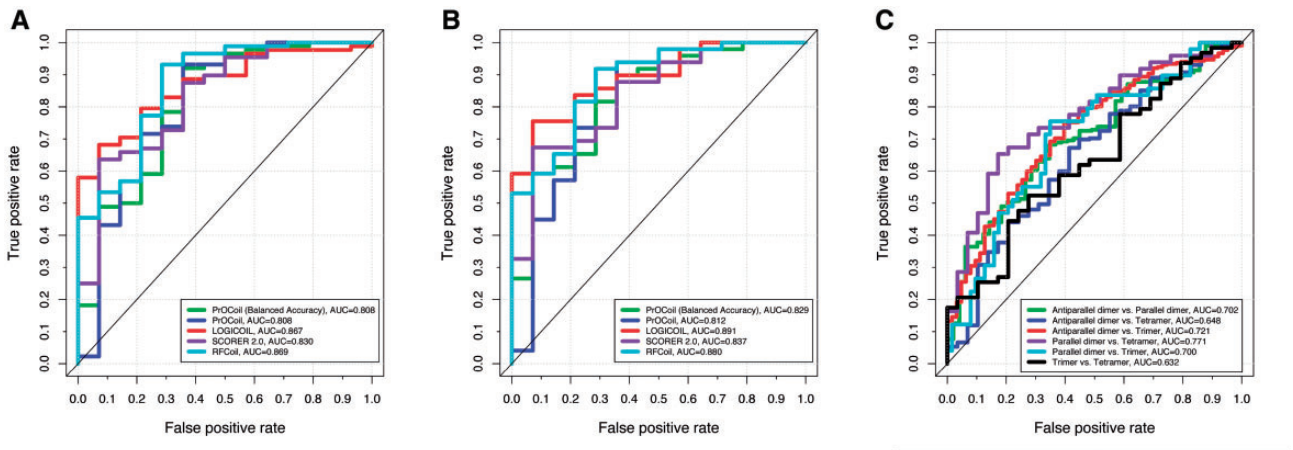
assembled two independent test data sets (discussed below) and measured the performance [in terms of area under curve (AUC)] of all tested tools on these two data sets. In particular, as the previous versions of CCHMM, SCORER and MultiCoil have been upgraded as CCHMM_PROF, SCORER 2.0 and Multicoil2, respectively, we only evaluated the advanced versions in the independent test. In addition, as SOSUIcoil and SpiriCoil did not provide local executables, and it was not possible to run Paircoil2 without execution errors, these three predictors were not included in this test. According to the nature of the prediction tasks, we performed independent tests for two different types of tasks, namely, coiled-coil oligomeric state prediction and CCD prediction. Coiled-coil oligomeric state prediction usually requires CCDs and their heptad registers (i.e. *a*-*g*) as the input, while CCD prediction often takes protein sequences as input. For the first type, we evaluated the performance of coiled-coil oligomeric state predictors, including RFCoil, PrOCoil, SCORER 2.0, LOGICOIL and Multicoil2. For the second type, we compared the prediction performance of COILS, PCOILS, MARCOIL, CCHMM_PROF and Multicoil2.

Coiled-coil oligomeric state prediction

Test data set construction. We carefully prepared two different test data sets. For the first data set, CCDs and their respective heptad assignments were extracted from the PDB using SOCKET [41]. Only X-ray crystal structures were selected to ensure the quality of the data set (downloaded on 6 May 2014). SOCKET was applied to annotate the coiled-coils in a given structure with a default packing cut-off of 7.0Å, which was the same as that specified in the data set collection procedure of previous studies [37, 38]. In addition, to improve the quality of the data set, we further removed those structures with a resolution of worse than 4.0Å. Meanwhile, the structures with unnatural residues were also removed. For the second data set, we first culled coiled-coil class (h class) proteins from SCOPe [75] (the extended version of SCOP) and then verified the CCDs with SOCKET. Only the consensus sequences assigned by both SCOPe and SOCKET analysis that contained coiled-coils were retained to constitute the second data set, whereas the coiled-coil and heptad annotations were obtained by SOCKET. We subsequently examined the overlap between the second data set and the training data sets of RFCoil, PrOCoil, SCORER 2.0 and LOGICOIL. Our analysis showed that the majority of entries in the second data set were covered by the training data sets of the four predictors, suggesting that the second data set was not sufficiently large enough to be an independent test data set. Therefore, to address this, we first removed all the training data of investigated predictors from our data sets and then combined the first, second and other training data sets of the four predictors, and used CD-HIT to reduce the sequence redundancy of the resulting data set to ensure that the sequence identity of any two sequences in the data set was no more than 50%. For each cluster generated by CD-HIT, if all sequences in this cluster were from our first and second data sets, the representative sequence was collected. Although sequence redundancy can be reduced by other alternative ways, 50% has been commonly used as the preferred threshold for CCDs, as any threshold lower than 50% is deemed to be too strict for coiled-coil oligomeric state prediction [36]. Finally, the independent test data set contained 509 antiparallel dimers, 88 parallel dimers, 94 trimers and 36 tetramers (Supplementary Table S1; Additional file 1—http://lightning.med.monash.edu/coiledcoil/).

Performance comparison. Among the four reviewed predictors, RFCoil and PrOCoil were trained using coiled-coils with length $\geq$8 amino acids, while SCORER 2.0 and LOGICOIL were developed using coiled-coils with length >14 residues. In addition, RFCoil, PrOCoil and SCORER 2.0 were designed to classify parallel dimeric and trimeric coiled-coils. LOGICOIL is the only currently available predictor that can be used to predict four types of coiled-coil oligomeric states, including parallel/antiparallel dimers, trimers and tetramers. Therefore, to comprehensively evaluate the performance of these tools for predicting the two different types of coiled-coils, we first split the independent test data set into two subsets, one with coiled-coils >7 residues and the other with coiled-coils >14 amino acids. For each subset, we evaluated the prediction performance using AUC values. This included the performance comparison of parallel dimer and parallel trimer between the four predictors, as well as pairwise performance comparison of LOGICOIL. The receiver operating characteristic (ROC) curves of these different predictors are shown in Figure 2. We also notice that certain heptad registers for CCDs from SOCKET are non-canonical, which means that the heptad registers (i.e. *a*-*g*) are interrupted according to SOCKET annotations. In view of this, we further removed the coiled-coils with non-canonical heptad assignments and repeated our tests (Additional file 2 downloadable at http://lightning.med.monash.edu/coiledcoil/). The corresponding ROC curves of all predictors for predicting these coiled-coils without non-canonical heptad registers are shown in Figure 3. For Figures 2A, B, 3A and B, 'positive' and 'negative' indicate parallel dimeric and trimeric coiled-coils, respectively.

We note that generally, when testing with parallel dimeric and trimeric coiled-coils, LOGICOIL and RFCoil achieved the highest AUC values (see Figures 2A, B, 3A and B). Although LOGICOIL was trained using longer coiled-coil sequences, most of which contained canonical heptads, it was able to predict shorter coiled-coils with non-canonical heptads. Pairwise AUC values can be observed in Figures 2C and 3C, where LOGICOIL achieved the highest AUC values when predicting parallel dimer and tetramer (with AUC values of 0.771 and 0.794, respectively). However, distinguishing tetramer from trimer appears to be the most challenging task. PrOCoil-BA performed constantly better than PrOCoil when tested with both short and long coiled-coils (see Figures 2A, B, 3A and B). In addition to AUC values, we also computed the 95% confidence interval using the 'pROC' package [76]. The 95% confidence intervals are shown for each ROC curve in the corresponding tables in Figures 2 and 3. It can be seen that most of the 95% confidence intervals are overlapped. This suggests that even though the compared predictors achieved different AUC values, it is difficult to determine which predictor is the 'statistically significant' best model. For each of the parallel dimeric and trimeric testing samples, we also applied majority voting to generate consensus results and compared the performance of majority voting with other individual predictors (Supplementary Tables S2 and S3). It is clear that majority voting could indeed improve the prediction accuracy when testing oligomeric state prediction of coiled-coils with length $\geq$15 amino acids that contained both canonical and non-canonical heptad registers. Because dimeric coiled-coils are more prevalent than trimer and tetramer, all these predictors were trained with imbalanced training data sets. Accordingly, some predictors are highly biased. For example, when testing RFCoil, we noticed that RFCoil could readily predict dimeric coiled-coils with high confidence, but often wrongly predicted many trimers as dimers. This is probably because of the limited number of trimers included in the training data set, and hence the trained RFCoil model did not generalize and perform well on trimer prediction. Therefore, to address this problem in future work, we

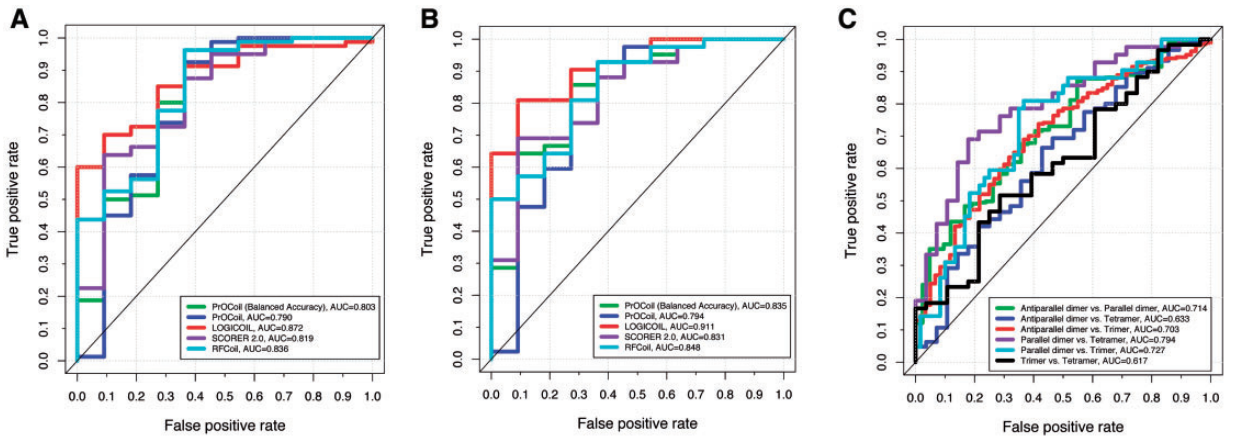


Predictor	95% Confidence Interval
PrOCoil-BA	0.667-0.950
PrOCoil	0.653-0.964
LOGICOIL	0.784-0.950
SCORER 2.0	0.709-0.954
RFCoil	0.759-0.978

Predictor	95% Confidence Interval
PrOCoil-BA	0.703-0.956
PrOCoil	0.657-0.967
LOGICOIL	0.807-0.974
SCORER 2.0	0.720-0.954
RFCoil	0.782-0.979

Task	95% Confidence Interval
Antiparallel dimer vs. Parallel dimer	0.628-0.776
Antiparallel dimer vs. Tetramer	0.540-0.756
Antiparallel dimer vs. Trimer	0.654-0.788
Parallel dimer vs. Tetramer	0.664-0.877
Parallel dimer vs. Trimer	0.601-0.798
Trimer vs. Tetramer	0.511-0.753

Figure 2. Performance comparison of coiled-coils with non-canonical heptad registers between RFCoil, SCORER 2.0, PrOCoil and LOGICOIL on the independent test. (**A**) ROC curves and the 95% confidence intervals for parallel dimeric and trimeric coiled-coils with length ≥8 amino acids. (**B**) ROC curves and the 95% confidence intervals for parallel dimeric and trimeric coiled-coils with length ≥15 amino acids. (**C**) ROC curves and the 95% confidence intervals of LOGICOIL for pairwise oligomeric state prediction with coiled-coils with length ≥15 residues.



Predictor	95% Confidence Interval
PrOCoil-BA	0.635-0.972
PrOCoil	0.596-0.983
LOGICOIL	0.779-0.964
SCORER 2.0	0.671-0.968
RFCoil	0.700-0.973

Predictor	95% Confidence Interval
PrOCoil-BA	0.693-0.978
PrOCoil	0.603-0.986
LOGICOIL	0.828-0.995
SCORER 2.0	0.692-0.970
RFCoil	0.723-0.974

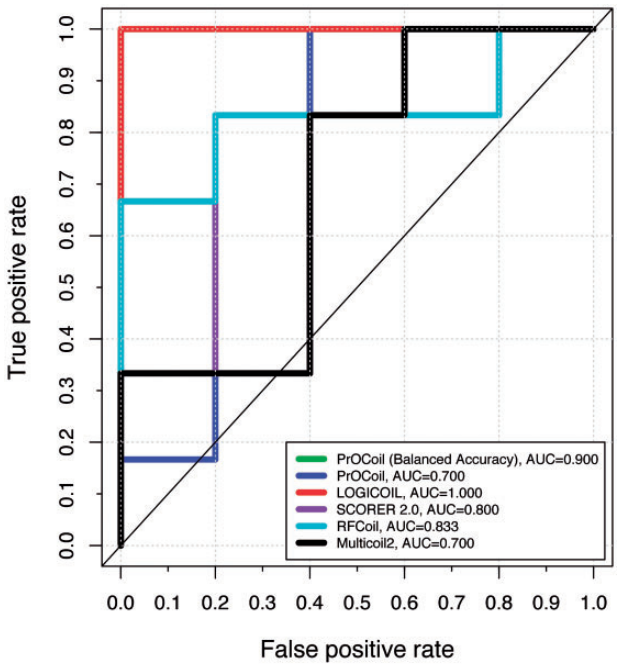
Task	95% Confidence Interval
Antiparallel dimer vs. Parallel dimer	0.636-0.793
Antiparallel dimer vs. Tetramer	0.520-0.746
Antiparallel dimer vs. Trimer	0.634-0.773
Parallel dimer vs. Tetramer	0.688-0.901
Parallel dimer vs. Trimer	0.628-0.827
Trimer vs. Tetramer	0.492-0.743

Figure 3. Performance comparison of coiled-coils without non-canonical heptad registers between RFCoil, SCORER 2.0, PrOCoil and LOGICOIL on the independent test. (**A**) ROC curves and the 95% confidence intervals for parallel dimeric and trimeric coiled-coils with length ≥8 amino acids. (**B**) ROC curves and the 95% confidence intervals for parallel dimeric and trimeric coiled-coils with length ≥15 amino acids. (**C**) ROC curves and the 95% confidence intervals of LOGICOIL for pairwise oligomeric state prediction with coiled-coils with length ≥15 residues.

recommend that certain techniques for imbalanced data processing and mining be applied (e.g. oversampling or undersampling) to enrich the imbalanced samples. Oversampling and undersampling [77] are both basic (opposite but equivalent) methodologies for sampling the data with imbalanced class distribution. Oversampling is a technique that randomly selects samples from the class where the number of samples is quite small to enrich the samples in this class, while undersampling randomly selects samples from the class where the number of samples in this class is large to reduce the number of samples in this class. These two techniques are basic and easy to implement. More complex and advanced techniques

for imbalanced biological/medical data mining tasks also exist [78–80].

We next compared the prediction performance of Multicoil2 and other predictors. Multicoil2 accepts the full-length protein sequences as the input rather than coiled-coil sequences and their respective heptad registers. Instead of providing an overall score for the input sequence, Multicoil2 generates predicted probabilities for each individual residue in the sequence to form parallel dimers, parallel trimers or non-coiled-coils. Here, to compare with other methods, we calculated the average of the predicted probabilities by Multicoil2, normalized the value into the range of [0, 1] and removed the predicted non-coiled-coils from the results (the prediction threshold was set as 0.5). We combined the parallel dimeric and trimeric coiled-coils with length $>=21$ amino acids (given that Multicoil2 can only predict CCDs with length $>=21$ amino acids) in the data set used in our independent test with the dimers and trimers sequences in the Multicoil2 training data set and applied CD-HIT to remove the sequence redundancy, ensuring that the identity between any two sequences in the resulting data set was no more than 50%. As a result, only 22 CCDs remained in the resulting data set. For the remaining CCDs, we downloaded their complete protein sequences so that we could use them as the input to Multicoil2. Multicoil2 predicted only 11 of 22 (50.0%) sequences that contained CCDs that overlapped with SOCKET annotation. Therefore, we compared only the prediction performance of different predictors on these 11 'valid' CCDs (Figure 4; Additional file 3—http://lightning.med.monash.edu/coiledcoil/). In Figure 4, 'positive' and 'negative' represent parallel dimeric and trimeric coiled-coils, respectively. LOGICOIL correctly classified all the parallel dimeric and trimeric coiled-coils, while Multicoil2 and PrOCoil obtained the lowest AUC value. Consistent with the results in Figures 2 and 3, PrOCoil-BA performed better than PrOCoil (greater by 0.2), followed by RFCoil and SCORER 2.0. In addition, the 95% confidence intervals suggest that LOGICOIL was the best predictor based on this independent testing data set. Consistent with the AUC values shown in Figure 4, LOGICOIL correctly classified all the test samples. It is noteworthy that the majority voting strategy achieved an accuracy of 90.9%, which was ranked as the second best accuracy according to the accuracies of other individual predictors (Supplementary Table S4).



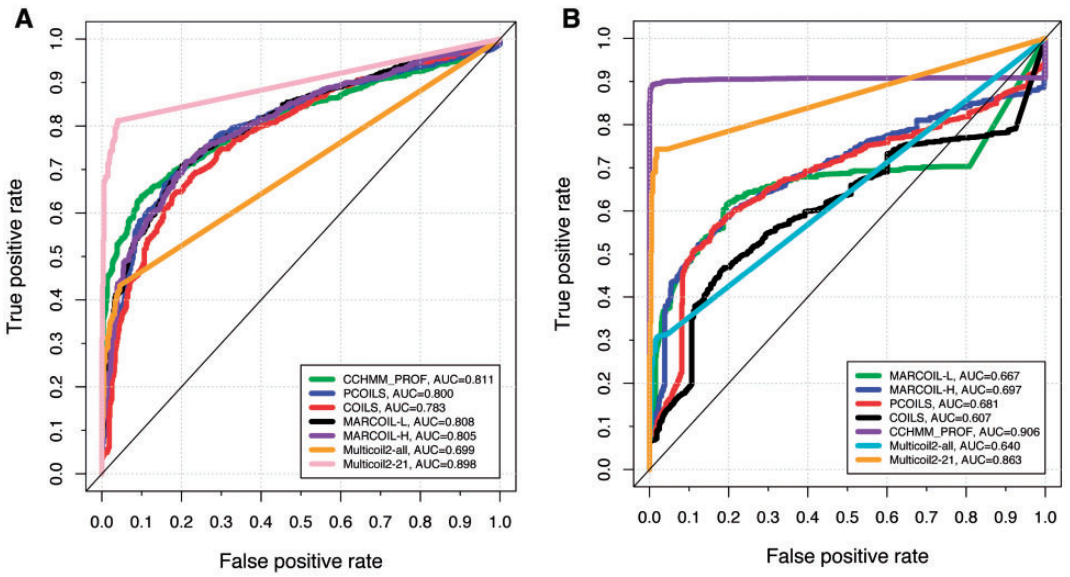
Predictor	95% Confidence Interval
PrOCoil-BA	0.713-1.0
PrOCoil	0.312-1.0
LOGICOIL	1.0-1.0
SCORER 2.0	0.503-1.0
RFCoil	0.56-1.0
Multicoil2	0.342-1.0

Figure 4. ROC curves and the 95% confidence intervals of Multcoil2 and other predictors for parallel dimeric and trimeric coiled-coil prediction.

CCD prediction

Testing data set construction. The positive data set comprised protein sequences containing annotated CCDs based on SOCKET. For the negative data set, we extracted protein entries of alpha and beta classes (a/b; i.e. c class) from the SCOPe database, except for superfamilies c.37.1, c.49.2, c.67.1 and c.93.1, which are annotated to contain CCDs [24]. Protein sequences were extracted from PDB, and those sequences that contain unnatural amino acids were removed. These sequences were further validated by SOCKET with a loosened threshold of 7.4Å [33] to ensure they did not contain any CCDs. After removing all the available training data of investigated predictors from our testing data set, we combined our testing data sets with the available training data sets of CCHMM_PROF, MARCOIL and Multicoil2. We then applied CD-HIT to remove the sequence redundancy, so that the sequence identity between any two sequences was not >30%. Similar to the construction process of the independent test data set for CCD oligomeric state prediction, for each cluster generated by CD-HIT, only representative sequences from the clusters where there were no samples from the training data sets of the compared predictors in this cluster were collected. After this procedure, the final data set included a total of 1643 sequences, 601 of which did not contain any CCDs and 1042 containing 2176 CCDs (Additional files 4 and 5—http://lightning.med.monash.edu/coiledcoil/). CCHMM_PROF and PCOILS require the position-specific scoring matrix (PSSM) generated by PSI-BLAST as the input to make the prediction. Accordingly, we used the Uniref90 database to generate the PSSM profiles of all the tested sequences and conduct the comparison, which was also used as the search database by CCHMM_PROF [33]. The parameters for PSI-BLAST was preliminarily set by the PCOILS program; for CCHMM_PROF, we used the same parameters described in [33].

Performance comparison. Firstly, we evaluated the effectiveness of different predictors for identifying CCDs by calculating the averaged probability score for each protein. If a protein was predicted to contain coiled-coil residues, the probability was calculated as the averaged score of all predicted coiled-coil residues; otherwise, if a protein was not predicted to have CCDs, then the calculated probability was the averaged score of all residues of the whole protein. The ROC curves and corresponding AUC values of the compared predictors are shown in Figure 5A, where 'positive' represents the sequences containing CCDs, while 'negative' indicates the sequences without CCDs. Because Multicoil2 can only predict protein sequences with CCDs >21 amino acids, we provided the results of Multicoil2 on



Predictor	95% Confidence Interval
CCHMM_PROF	0.790-0.831
PCOILS	0.779-0.822
COILS	0.760-0.805
MARCOIL-L	0.787-0.829
MARCOIL-H	0.784-0.826
Multicoil-all	0.682-0.716
Multicoil-21	0.867-0.928

Predictor	95% Confidence Interval
CCHMM_PROF	0.894-0.919
PCOILS	0.660-0.703
COILS	0.584-0.629
MARCOIL-L	0.648-0.687
MARCOIL-H	0.677-0.717
Multicoil-all	0.628-0.652
Multicoil-21	0.835-0.891

Figure 5. Performance comparison of CCD predictors. (**A**) ROC curves and the 95% confidence intervals of different predictors for identifying coiled-coil domains. (**B**) ROC curves and the 95% confidence intervals of different predictors, showing the consistency between the predicted CCDs and those annotated by SOCKET based on the protein structures.

both the entire test data set (termed 'Multicoil2-all') and a subset that only contained proteins with coiled-coils >= 21 amino acids (termed 'Multicoil2-21'). It is apparent that Multicoil2-21 identified the majority of coiled-coils and achieved the highest AUC value of 0.898, followed by CCHMM_PROF (AUC = 0.811). The AUC value of PCOILS was higher than COILS by 0.017, presumably owing to the incorporation of evolutionary information in the form of PSSM generated by PSI-BLAST. Next, we examined whether the identified CCDs were identical to those annotated by SOCKET. To do so, we compared all 2176 CCDs and their corresponding prediction scores of all reviewed predictors. A domain was predicted as a CCD if its probability was >0.5. For the negative protein (i.e. proteins without CCDs), if it was predicted to have a CCD, the average score would be calculated; otherwise, the average prediction score for each residue in this protein would be calculated. The results are shown in Figure 5B, where the 'positive' denotes CCDs while the 'negative' indicates the sequences without CCDs. Similar to Figure 5A, CCHMM_PROF and Multicoil2-21 again achieved the highest and second highest AUC values (AUC = 0.906 and 0.863, respectively), suggesting that the majority of their predicted CCDs were consistent with the SOCKET assignment. COILS obtained the lowest performance with an AUC score of only 0.607. We also note that Multicoil2-all achieved a lower AUC score, possibly owing to its restriction of having a length requirement of coiled-coils during the model training. The performance comparison results between individual predictors and majority voting are shown in Supplementary Table S5. Because the minimum length of coiled-coils used for training Multicoil2 is 21, we further filtered the testing data set with different thresholds of coiled-coil lengths to perform the CCD coverage test. Although majority voting did not improve the overall prediction accuracy, the performance of majority voting was still competitive compared with individual predictors (Supplementary Table S5).

CCD and CCD oligomeric state prediction for human PolyQ proteins

Identification of CCDs

We first made a consensus-based decision for CCD prediction based on the predictors that are capable of discriminating coiled-coils from non-coiled-coils. The predictors used in this step were COILS, PCOILS, Paircoil2 (the p-score version with different window sizes and probability score version), MARCOIL, CCHMM_PROF, SpiriCoil and Multicoil2. Strikingly, the results are largely inconsistent between different predictors (Supplementary Tables S6–S13), making it difficult to generate a consensus prediction. Only a small portion of the proteins was predicted to harbour CCDs according to the prediction results of PCOILS, Paircoil2 (both p-score and probability score versions), SpiriCoil and Multicoil2. In contrast, COILS, MARCOIL and CCHMM_PROF predicted several CCDs within the nine PolyQ proteins. Most of the predicted coiled-coils overlapped or flanked the PolyQ tract. Based on the prediction results, the final decisions of predicted CCDs were made through majority voting (i.e. the CCD peptides need to be predicted by at least four predictors; the results are listed in Table 3). In the prediction of CCDs in nine disease-associated PolyQ proteins by

Table 3. The consensus CCDs predicted by at least four predictors

Protein	Predicted coiled-coils	Protein structure	Sequence	Overlapping PolyQ tract	Agreed by
Voltage-dependent P/Q-type calcium channel subunit alpha-1A (Brain calcium channel I)	720–747	3BXK (B/D = 1955–1975)	AQELTKDEQEEEEAANQKLALQKAKEVA	No	COILS, PCOILS, Paircoil2 (P-score version), CCHMM_PROF, Multicoil2 and MARCOIL
Atrophin-1	793–819	–	AKKRADLVEKVRREAEQRAREEKERER	No	COILS, PCOILS, Paircoil2 (P-score version), CCHMM_PROF, Multicoil2 (cut-off = 0.5) and MARCOIL

Fiumara *et al.* [27], only two relatively old CCDs predictors were used (COILS and Paircoil2). We note that the results of Fiumara *et al.* are inconsistent with our predictions in this several state-of-the-art predictors. This discrepancy highlights that it remains a challenging task to develop reliable and consistent CCD prediction methods, and that attention should be paid when only a few specific methods are used to make the prediction, especially when these methods are used to guide and interpret experimental investigations such as the studies by Fiumara *et al.* [27].

Prediction of oligomeric state of PolyQ proteins. To examine the potential oligomeric states of the peptides listed in Table 3, we performed the prediction using RFCoil, SCORER 2.0, PrOCoil and LOGICOIL (Supplementary Tables S14 and S15). Because COILS, MARCOIL, PCOILS, Paircoil2 and Multicoil2 all provided heptad registers, we used these heptads to facilitate the oligomeric state prediction. As we can see, with different heptad registers, RFCoil, SCORER 2.0 and PrOCoil produced consistent prediction results (dimer formation), while the oligomeric state predictions from LOGICOIL were variable.

Conclusions

Given the functional significance of CCDs, computational biologists are motivated to develop more accurate and reliable predictors for CCD prediction. Aiming at providing a comprehensive review of coiled-coil predictors to non-bioinformaticians, this article describes and compares a number of widely used coiled-coil predictors in terms of their input, model construction and model evaluation. Independent tests reveal that LOICOIL achieved the overall highest AUC value when used to predict parallel dimeric and trimeric coiled-coils. For CCD prediction, Multicoil2 achieved the highest AUC value when detecting long CCDs in proteins, while CCHMM_PROF achieved the highest AUC value for the coverage of detected CCDs without the length limitation of CCDs. A case study of nine PolyQ proteins demonstrated that coiled-coil predictions were quite different among different predictors, which could further confound the consensus prediction analysis. We conclude that coiled-coil prediction is still a challenging task, and we expect that more powerful algorithms with improved prediction performance will emerge with the increasing availability of coiled-coil data.

Supplementary Data

Supplementary data are available online at http://bib.oxfordjournals.org/.

Key Points

- This article provides a comprehensive review on the current progress of computational approaches for coiled-coil domain (CCD) prediction and coiled-coil oligomeric state prediction.
- Independent tests using rigorously prepared data sets highlight that Multicoil2 (tested with long coiled-coils) and CCHMM_PROF achieved the highest area under curve (AUC) values for coiled-coil domain prediction, while LOGICOIL achieved the highest AUC value for parallel dimeric and trimeric prediction.
- The CCD prediction results on nine PolyQ proteins show inconsistencies of CCD prediction, which should be borne in mind when using prediction methods to make meaningful and reliable biological inferences.
- This review serves as a useful guide for researchers who want to gain a better understanding of state-of-the-art approaches in this area and aim to develop their own methods with improved performance.

Acknowledgements

The authors would like to thank Derek N. Woolfson, Mauro Delorenzi and Piero Fariselli for providing the training data sets for LOGICOIL, MARCOIL and CCHMM_PROF, respectively.

Funding

China Scholarship Council (CSC) and Monash University Joint PhD Student Scholarship (to C.L.); Higher Degree by Research Scholarship (HDR) awarded by Monash University, Malaysia (to C.C.H.C); JSPS, Japan (grant-in-aid #26240034 to T.A.); Hundred Talents Program of the Chinese Academy of Sciences (CAS) (J.S.); NIH R01 (AI111990 to J.S.); Monash University Major Inter-Disciplinary Research Project grant (201402 to J.S.); and A.M.B. is an NHMRC Senior Research Fellow (1022688).

References

1. Lupas A. Coiled coils: new structures and new functions. *Trends Biochem Sci* 1996;**21**:375–82.
2. Grigoryan G, Keating AE. Structural specificity in coiled-coil interactions. *Curr Opin Struct Biol* 2008;**18**:477–83.
3. McFarlane AA, Orriss GL, Stetefeld J. The use of coiled-coil proteins in drug delivery systems. *Eur J Pharmacol* 2009;**625**:101–7.
4. Tarricone C, Xiao B, Justin N, *et al*. The structural basis of Arfaptin-mediated cross-talk between Rac and Arf signalling pathways. *Nature* 2001;**411**:215–19.
5. Burkhard P, Kammerer RA, Steinmetz MO, *et al*. The coiled-coil trigger site of the rod domain of cortexillin I unveils a distinct network of interhelical and intrahelical salt bridges. *Structure* 2000;**8**:223–30.
6. Bullough PA, Hughson FM, Skehel JJ, *et al*. Structure of influenza haemagglutinin at the pH of membrane fusion. *Nature* 1994;**371**:37–43.
7. Whitson SR, LeStourgeon WM, Krezel AM. Solution structure of the symmetric coiled coil tetramer formed by the oligomerization domain of hnRNP C: implications for biological function. *J Mol Biol* 2005;**350**:319–37.
8. Gromiha MM, Parry DA. Characteristic features of amino acid residues in coiled-coil protein structures. *Biophys Chem* 2004;**111**:95–103.
9. Mason JM, Arndt KM. Coiled coil domains: stability, specificity, and biological implications. *Chembiochem* 2004;**5**:170–6.
10. Arndt KM, Pelletier JN, Muller KM *et al*. Comparison of *in vivo* selection and rational design of heterodimeric coiled coils. *Structure* 2002;**10**:1235–48.
11. Gillingham AK, Munro S. Long coiled-coil proteins and membrane traffic. *Biochim Biophys Acta* 2003;**1641**:71–85.
12. Rose A, Schraegle SJ, Stahlberg EA, *et al*. Coiled-coil protein composition of 22 proteomes–differences and common themes in subcellular infrastructure and traffic control. *BMC Evol Biol* 2005;**5**:66.
13. Guo Y, Bozic D, Malashkevich VN, *et al*. All-trans retinol, vitamin D and other hydrophobic compounds bind in the axial pore of the five-stranded coiled-coil domain of cartilage oligomeric matrix protein. *EMBO J* 1998;**17**:5265–72.
14. Ozbek S, Engel J, Stetefeld J. Storage function of cartilage oligomeric matrix protein: the crystal structure of the coiled-coil domain in complex with vitamin D(3). *EMBO J* 2002;**21**:5960–8.
15. Stetefeld J, Jenny M, Schulthess T, *et al*. Crystal structure of a naturally occurring parallel right-handed coiled coil tetramer. *Nat Struct Biol* 2000;**7**:772–76.
16. Eriksson M, Hassan S, Larsson R, *et al*. Utilization of a right-handed coiled-coil protein from archaebacterium Staphylothermus marinus as a carrier for cisplatin. *Anticancer Res* 2009;**29**:11–18.
17. Boulikas T, Vougiouka M. Recent clinical trials using cisplatin, carboplatin and their combination chemotherapy drugs (review). *Oncol Rep* 2004;**11**:559–95.
18. Deacon SP, Apostolovic B, Carbajo RJ, *et al*. Polymer coiled-coil conjugates: potential for development as a new class of therapeutic "molecular switch". *Biomacromolecules* 2011;**12**:19–27.
19. Hodges RS. Boehringer Mannheim award lecture 1995. La conference Boehringer Mannheim 1995. De novo design of alpha-helical proteins: basic research to medical applications. *Biochem Cell Biol* 1996;**74**:133–54.
20. Kakizawa Y, Furukawa S, Ishii A, *et al*. Organic-inorganic hybrid-nanocarrier of siRNA constructing through the self-assembly of calcium phosphate and PEG-based block aniomer. *J Control Release* 2006;**111**:368–70.
21. Wu K, Liu J, Johnson RN, *et al*. Drug-free macromolecular therapeutics: induction of apoptosis by coiled-coil-mediated cross-linking of antigens on the cell surface. *Angew Chem Int Ed Engl* 2010;**49**:1451–5.
22. Pechar M, Pola R. The coiled coil motif in polymer drug delivery systems. *Biotechnol Adv* 2013;**31**:90–6.
23. Vincent TL, Woolfson DN, Adams JC. Prediction and analysis of higher-order coiled-coils: insights from proteins of the extracellular matrix, tenascins and thrombospondins. *Int J Biochem Cell Biol* 2013;**45**:2392–401.
24. Gruber M, Soding J, Lupas AN. Comparative analysis of coiled-coil prediction methods. *J Struct Biol* 2006;**155**:140–5.
25. Chang CC, Song J, Tey BT, *et al*. Bioinformatics approaches for improved recombinant protein production in *Escherichia coli*: protein solubility prediction. *Brief Bioinform* 2014;**15**:953–62.
26. Chang CC, Tey BT, Song J, *et al*. Towards more accurate prediction of protein folding rates: a review of the existing web-based bioinformatics approaches. *Brief Bioinform* 2014;**16**(2):314–24.
27. Fiumara F, Fioriti L, Kandel ER, *et al*. Essential role of coiled coils for aggregation and activity of Q/N-rich prions and PolyQ proteins. *Cell* 2010;**143**:1121–35.
28. Lupas A. Predicting coiled-coil regions in proteins. *Curr Opin Struct Biol* 1997;**7**:388–93.
29. Gruber M, Soding J, Lupas AN. REPPER–repeats and their periodicities in fibrous proteins. *Nucleic Acids Res* 2005;**33**:W239–43.
30. McDonnell AV, Jiang T, Keating AE, *et al*. Paircoil2: improved prediction of coiled coils from sequence. *Bioinformatics* 2006;**22**:356–58.
31. Tanizawa H, Ghimire GD, Mitaku S. A hight performance prediction system of coiled coil domains containing heptad breaks: SOSUIcoil. *Chem-Bio Inform J* 2008;**8**:16.
32. Delorenzi M, Speed T. An HMM model for coiled-coil domains and a comparison with PSSM-based predictions. *Bioinformatics* 2002;**18**:617–25.
33. Bartoli L, Fariselli P, Krogh A, *et al*. CCHMM_PROF: a HMM-based coiled-coil predictor with evolutionary information. *Bioinformatics* 2009;**25**:2757–63.
34. Rackham OJ, Madera M, Armstrong CT, *et al*. The evolution and structure prediction of coiled coils across all genomes. *J Mol Biol* 2010;**403**:480–93.
35. Armstrong CT, Vincent TL, Green PJ, *et al*. SCORER 2.0: an algorithm for distinguishing parallel dimeric and trimeric coiled-coil sequences. *Bioinformatics* 2011;**27**:1908–14.
36. Vincent TL, Green PJ, Woolfson DN. LOGICOIL–multi-state prediction of coiled-coil oligomeric state. *Bioinformatics* 2013;**29**:69–76.
37. Mahrenholz CC, Abfalter IG, Bodenhofer U, *et al*. Complex networks govern coiled-coil oligomerization–predicting and profiling by means of a machine learning approach. *Mol Cell Proteomics* 2011;**10**:M110.004994.
38. Li C, Wang XF, Chen Z, *et al*. Computational characterization of parallel dimeric and trimeric coiled-coils using effective amino acid indices. *Mol Biosyst* 2015;**1**:354–60.
39. Trigg J, Gutwin K, Keating AE, *et al*. Multicoil2: predicting coiled coils and their oligomerization states from sequence in the twilight zone. *PLoS One* 2011;**6**:e23519.
40. Andreeva A, Howorth D, Chandonia JM, *et al*. Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res* 2008;**36**:D419–25.

41. Walshaw J, Woolfson DN. Socket: a program for identifying and analysing coiled-coil motifs within protein structures. *J Mol Biol* 2001;**307**:1427–50.
42. Testa OD, Moutevelis E, Woolfson DN. CC+: a relational database of coiled-coil structures. *Nucleic Acids Res* 2009;**37**:D315–22.
43. Fu L, Niu B, Zhu Z, *et al*. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 2012;**28**:3150–2.
44. Altschul SF, Madden TL, Schaffer AA, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;**25**:3389–402.
45. Berger B, Wilson DB, Wolf E, *et al*. Predicting coiled coils by use of pairwise residue correlations. *Proc Natl Acad Sci USA* 1995;**92**:8259–63.
46. Fariselli P, Molinini D, Casadio R, *et al*. Prediction of structurally-determined coiled-coil domains with hidden Markov models. *Bioinform Res Dev Proc* 2007;**4414**:292–302.
47. Wolf E, Kim PS, Berger B. MultiCoil: a program for predicting two- and three-stranded coiled coils. *Protein Sci* 1997;**6**:1179–89.
48. Woolfson DN, Alber T. Predicting oligomerization states of coiled coils. *Protein Sci* 1995;**4**:1596–607.
49. Cortes C, Vapnik V. Support-vector networks. *Mach Learn* 1995;**20**:273–97.
50. Breiman L. Random forests. *Mach Learn* 2001;**45**:5–32.
51. Raileanu LE, Stoffel K. Theoretical comparison between the Gini Index and Information Gain criteria. *Anna Math Artif Intell* 2004;**41**:77–93.
52. Kendall M. A new measure of rank correlation. *Biometrika* 1938;**30**:9.
53. Robertson AL, Bate MA, Androulakis SG, *et al*. PolyQ: a database describing the sequence and domain context of polyglutamine repeats in proteins. *Nucleic Acids Res* 2011;**39**:D272–6.
54. Bilen J, Bonini NM. Drosophila as a model for human neurodegenerative disease. *Annu Rev Genet* 2005;**39**:153–71.
55. Saunders HM, Bottomley SP. Multi-domain misfolding: understanding the aggregation pathway of polyglutamine proteins. *Protein Eng Des Sel* 2009;**22**:447–51.
56. Kim MW, Chelliah Y, Kim SW, *et al*. Secondary structure of Huntingtin amino-terminal region. *Structure* 2009;**17**:1205–12.
57. Zuhlke C, Hellenbroich Y, Dalski A, *et al*. Different types of repeat expansion in the TATA-binding protein gene are associated with a new form of inherited ataxia. *Eur J Hum Genet* 2001;**9**:160–4.
58. Nakamura K, Jeong SY, Uchihara T, *et al*. SCA17, a novel autosomal dominant cerebellar ataxia caused by an expanded polyglutamine in TATA-binding protein. *Hum Mol Genet* 2001;**10**:1441–8.
59. Silveira I, Miranda C, Guimaraes L, *et al*. Trinucleotide repeats in 202 families with ataxia: a small expanded (CAG)n allele at the SCA17 locus. *Arch Neurol* 2002;**59**:623–9.
60. Lin B, Nasir J, MacDonald H, *et al*. Sequence of the murine Huntington disease gene: evidence for conservation, alternate splicing and polymorphism in a triplet (CCG) repeat [corrected]. *Hum Mol Genet* 1994;**3**:85–92.
61. Banfi S, Servadio A, Chung MY, *et al*. Identification and characterization of the gene causing type 1 spinocerebellar ataxia. *Nat Genet* 1994;**7**:513–20.
62. Quan F, Janas J, Popovich BW. A novel CAG repeat configuration in the SCA1 gene: implications for the molecular diagnostics of spinocerebellar ataxia type 1. *Hum Mol Genet* 1995;**4**:2411–13.
63. Pulst SM, Nechiporuk A, Nechiporuk T, *et al*. Moderate expansion of a normally biallelic trinucleotide repeat in spinocerebellar ataxia type 2. *Nat Genet* 1996;**14**:269–76.
64. Sanpei K, Takano H, Igarashi S, *et al*. Identification of the spinocerebellar ataxia type 2 gene using a direct identification of repeat expansion and cloning technique, DIRECT. *Nat Genet* 1996;**14**:277–84.
65. Imbert G, Saudou F, Yvert G, *et al*. Cloning of the gene for spinocerebellar ataxia 2 reveals a locus with high sensitivity to expanded CAG/glutamine repeats. *Nat Genet* 1996;**14**:285–91.
66. Elden AC, Kim HJ, Hart MP, *et al*. Ataxin-2 intermediate-length polyglutamine expansions are associated with increased risk for ALS. *Nature* 2010;**466**:1069–75.
67. Zhuchenko O, Bailey J, Bonnen P, *et al*. Autosomal dominant cerebellar ataxia (SCA6) associated with small polyglutamine expansions in the alpha 1A-voltage-dependent calcium channel. *Nat Genet* 1997;**15**:62–9.
68. Jodice C, Mantuano E, Veneziano L, *et al*. Episodic ataxia type 2 (EA2) and spinocerebellar ataxia type 6 (SCA6) due to CAG repeat expansion in the CACNA1A gene on chromosome 19p. *Hum Mol Genet* 1997;**6**:1973–8.
69. Tonelli A, D'Angelo MG, Salati R, *et al*. Early onset, non fluctuating spinocerebellar ataxia and a novel missense mutation in CACNA1A gene. *J Neurol Sci* 2006;**241**:13–17.
70. Romaniello R, Zucca C, Tonelli A, *et al*. A wide spectrum of clinical, neurophysiological and neuroradiological abnormalities in a family with a novel CACNA1A mutation. *J Neurol Neurosurg Psychiatry* 2010;**81**:840–3.
71. Nagafuchi S, Yanagisawa H, Ohsaki E, *et al*. Structure and expression of the gene responsible for the triplet repeat disorder, dentatorubral and pallidoluysian atrophy (DRPLA). *Nat Genet* 1994;**8**:177–82.
72. David G, Abbas N, Stevanin G, *et al*. Cloning of the SCA7 gene reveals a highly unstable CAG repeat expansion. *Nat Genet* 1997;**17**:65–70.
73. Echaniz-Laguna A, Rousso E, Anheim M, *et al*. A family with early-onset and rapidly progressive X-linked spinal and bulbar muscular atrophy. *Neurology* 2005;**64**:1458–60.
74. Kawaguchi Y, Okamoto T, Taniwaki M, *et al*. CAG expansions in a novel gene for Machado-Joseph disease at chromosome 14q32.1. *Nat Genet* 1994;**8**:221–8.
75. Fox NK, Brenner SE, Chandonia JM. SCOPe: Structural Classification of Proteins–extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res* 2014;**42**:D304–9.
76. Robin X, Turck N, Hainard A, *et al*. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 2011;**12**:77.
77. Chawla NV. *Data Mining for Imbalanced Datasets: An Overview, Data Mining and Knowledge Discovery Handbook*, 2nd edn., Springer, United States of America, 2010:875–86.
78. Munkhdalai T, Namsrai OE, Ryu K. Self-training in significance space of support vectors for imbalanced biomedical event data. *BMC Bioinformatics* 2015;**16**(Suppl 7):S6.
79. Wu K, Edwards A, Fan W, *et al*. Classifying imbalanced data streams via dynamic feature group weighting with importance sampling. *Proc SIAM Int Conf Data Min* 2014;**2014**:722–30.
80. Yang P, Xu L, Zhou BB, *et al*. A particle swarm based hybrid system for imbalanced medical data sampling. *BMC Genomics* 2009;**10**(Suppl 3):S34.

Molecular BioSystems

METHOD

View Article Online
View Journal | View Issue

Cite this: *Mol. BioSyst.*, 2015, **11**, 354

Computational characterization of parallel dimeric and trimeric coiled-coils using effective amino acid indices†

Chen Li,[a] Xiao-Feng Wang,[bc] Zhen Chen,[b] Ziding Zhang*[b] and Jiangning Song*[ad]

The coiled-coil, which consists of two or more α-helices winding around each other, is a ubiquitous and the most frequently observed protein–protein interaction motif in nature. The coiled-coil is known for its straightforward heptad repeat pattern and can be readily recognized based on protein primary sequences, exhibiting a variety of oligomer states and topologies. Due to the stable interaction formed between their α-helices, coiled-coils have been under close scrutiny to design novel protein structures for potential applications in the fields of material science, synthetic biology and medicine. However, their broader application requires an in-depth and systematic analysis of the sequence-to-structure relationship of coiled-coil folding and oligomeric formation. In this article, we propose a new oligomerization state predictor, termed as *RFCoil*, which exploits the most useful and non-redundant amino acid indices combined with the machine learning algorithm – random forest (RF) – to predict the oligomeric states of coiled-coil regions. Benchmarking experiments show that *RFCoil* achieves an AUC (area under the ROC curve) of 0.849 on the 10-fold cross-validation test using the training dataset and 0.855 on the independent test using the validation dataset, respectively. Performance comparison results indicate that *RFCoil* outperforms the four existing predictors LOGICOIL, PrOCoil, SCORER 2.0 and Multicoil2. Furthermore, we extract a number of predominant rules from the trained RF model that underlie the oligomeric formation. We also present two case studies to illustrate the applicability of the extracted rules to the prediction of coiled-coil oligomerization state. The *RFCoil* web server, source codes and datasets are freely available for academic users at http://protein.cau.edu.cn/RFCoil/.

Received 25th September 2014,
Accepted 18th November 2014

DOI: 10.1039/c4mb00569d

www.rsc.org/molecularbiosystems

Introduction

The coiled-coil is a ubiquitous structural motif consisting of two or more α-helices, which wind around each other to form a rope-like structure. Nearly sixty years ago, Crick proposed the standard structure model of the coiled-coil, which is distinct from other protein structures. Dimeric and trimeric coiled-coils are the two most common types of coiled-coil structures. Coiled-coils can be found in all organisms and it is estimated that nearly 10% of eukaryotic proteins and 3% of all protein-encoding regions of genes harbour the coiled-coil domain,[1–4] respectively. Due to their ability to oligomerize, coiled-coils play crucial roles in many biological processes, such as transcription, intracellular trafficking, viral infection and cellular signaling.[5,6] The property of coiled-coils, which enables two proteins to interact with each other, also attracts a great deal of interest from protein designers.[7] Coiled-coils are among the first designed proteins,[8,9] with potential applications in material science, synthetic biology and medicine.[10,11] Accordingly, understanding the mechanism of coiled-coil oligomerization is critically important for researchers to design versatile proteins with different functions.

The rope-like structure of coiled coils enables them to generate an interesting heptad repeat sequence pattern. That is, the structure goes around two complete turns of the helix after 7 residues, rather than the regular 7.2 residues. The heptad repeat is often labeled as *abcdefg*. Residues at the register positions *a* and *d* are often hydrophobic, forming a buried hydrophobic surface and providing the driving force for oligomerization. In contrast, residues at positions *e* and *g* are often charged or polar, which form salt bridges and electrostatic interactions, helping specify the binding partners.[12] Despite the simple heptad repeat pattern at the sequence

[a] *Department of Biochemistry and Molecular Biology, Faculty of Medicine, Monash University, Melbourne, VIC 3800, Australia.* [redacted]
[b] *State Key Laboratory of Agrobiotechnology, College of Biological Sciences, China Agricultural University, Beijing 100193, China.* [redacted]
[c] *School of Mathematics and Computer Science, Shanxi Normal University, Linfen 041004, China*
[d] *National Engineering Laboratory for Industrial Enzymes and Key Laboratory of Systems Microbial Biotechnology, Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, Tianjin 300308, China*

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c4mb00569d

level, coiled-coils display a great variety of oligomerization states, including dimers, trimers, tetramers, pentamers, and even heptamers. In addition, they often vary in the helix orientation, parallel or anti-parallel. Most coiled-coils adopt left-handed super-coils; however, right-handed coiled-coils are also observed.[13] Accordingly, an important question to address is, how can this simple heptad sequence repeat pattern encode such diverse structures?

To answer this question, a number of computational methods have been developed to analyze coiled-coils, which can be generally grouped as sequence-based or structure-based methods. Sequence-based methods mainly use the frequencies of residues or residue pairs at specific register positions to predict coiled-coil regions,[14–21] oligomerization states[4,17,18,22,23] and helix orientations.[24] In contrast, structure-based methods usually utilise structural information to facilitate the prediction, including SOCKET[12] and Twister.[25] In particular, the SOCKET algorithm is able to recognize characteristic knobs-into-holes side-chain packing of coiled-coil structures, clearly define coiled-coil helix boundaries, oligomerization states and helix orientations and assign heptad registers. The CC+ database[26] is developed based on the SOCKET algorithm, which can be used to create training datasets for building coiled-coil classifiers. Twister is implemented to compute local structural parameters of coiled-coils, based on Crick's parameterization.[27]

Regarding the prediction of coiled-coil oligomerization state, two early-stage algorithms SCORER[28] and Multicoil[29] exist. More recently, two new versions, SCORER 2.0[23] and Multicoil2[17] have been developed, and have been shown to perform better than their respective older versions. Almost at the same time, another two predictors for the coiled-coil oligomerization state, PrOCoil[22] and LOGICOIL,[4] were published. Multicoil2 employs a Markov Random Field method to integrate sequence features. It assigns the probability of a residue in a sequence to be non-coiled-coil, dimeric or trimeric. SCORER 2.0 and PrOCoil classify parallel dimeric and trimeric coiled-coils, given a coiled-coil sequence with known heptad registers. SCORER 2.0 uses statistically significant amino acid frequencies at seven heptad registers in combination with a Bayes factor method to distinguish parallel dimers from trimers. PrOCoil designs a new kernel function and uses the SVM (Support Vector Machine) algorithm to classify parallel dimers and trimers.[22] LOGICOIL, trained with coiled-coil regions larger than 14 amino acids using Bayesian variable selection response probabilities, can predict multiple oligomerization states for coiled-coil regions such as parallel dimer, antiparallel dimer, trimer and tetramer.[4] Therefore, LOGICOIL is currently considered as the state-of-the-art predictor for oligomerization states of coiled-coils.

In this article, we address the same classification task of SCORER 2.0 and PrOCoil by developing a novel tool *RFCoil*, which uses a sequence-based approach to distinguish parallel dimeric from trimeric coiled-coils (see Fig. 1 for examples of parallel dimer and trimer). More specifically, *RFCoil* employs the random forest (RF) algorithm to identify the most important and non-redundant amino acid indices and construct the classifiers to predict the oligomerization state of coiled-coils. We further compare the performance of *RFCoil* with four existing tools SCORER 2.0, PrOCoil, Multicoil2 and LOGICOIL by performing both 10-fold cross-validation and independent tests. The results show that *RFCoil* outperforms four existing tools LOGICOIL, SCORER 2.0, PrOCoil and Multicoil2 in the independent test. Moreover, we extract a number of important rules from the built RF models in an effort to provide biological insights into the underlying rules of the formation of oligomerization states of coiled-coils.

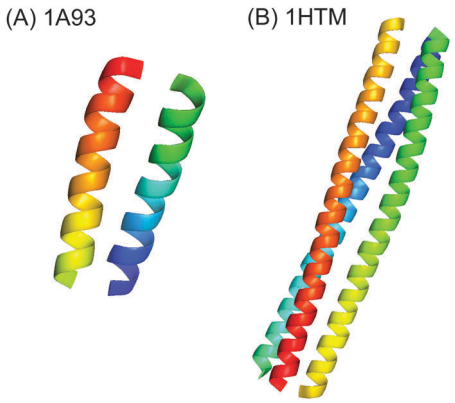


Fig. 1 Cartoon representations of parallel (A) dimeric (PDB ID: 1A93[30]) and (B) trimeric (PDB ID: 1HTM[31]) coiled-coils.

Materials and methods

Dataset

We used the benchmark dataset originally compiled by the developers of PrOCoil to train our models and assess the performance of our method. This benchmark dataset comprises 385 dimers and 92 trimers. The minimum length of the coiled-coils is 8 and nearly half of the coiled-coils have lengths longer than 14. This dataset was further divided into ten folds, and any two sequences from different folds have a sequence identity of no more than 60%. The methods were tested using the 10-fold cross-validation tests.

Moreover, apart from the benchmark dataset, we also constructed an independent test dataset to assess and compare the predictive performance of different methods. The procedures for constructing this independent test dataset are as follows: first, we used the SOCKET algorithm[12] to search the PDB database[32] for parallel coiled-coil dimers and trimers. For dimers, we selected those sharing a sequence identity of no more than 60% with the dimeric coiled-coil sequences in the training dataset. The selected dimers were further filtered to ensure that any two sequences shared a sequence identity of no more than 60%. The trimeric coiled-coils were filtered in a similar way to the dimers. Note that the sequence identity was calculated using the Needleman–Wunsch algorithm.[33] The final independent test set consists of 363 dimers and 48 trimers.

RFCoil

Our *RFCoil* approach includes four major steps, as shown in Fig. 2. The first step is to construct the training and independent test datsets extracted from the PDB database. The second step is to encode the input data, which was achieved by extracting the average amino acid index values for each heptad register.

The third step is to select the informative and non-redundant features for oligomerization state classification. We assumed no prior knowledge of the importance of each feature and this makes it possible for our feature selection method presented here to be applied to other questions. The final step is to use the selected features as the input to train *RFCoil* models. More details about the *RFCoil* approach are discussed in the following sections.

Sequence encoding. We attempted to capture the oligomerization state of the coiled-coil using its amino acid sequence information and each coiled-coil sequence using the physiochemical and biochemical properties of amino acids. To realize this, we extracted 529 amino acid indices that had no "NA" values in the AAindex database[34] (see Tables S1 and S2, ESI†). We encoded each coiled-coil sequence using the average amino acid index value at each heptad register, obtained using the following equation:

$$I(r,i) = \frac{\sum_{a \in r} \mathrm{AA}(a,i)}{n(r)} \quad (1)$$

where r represents a heptad register which can be a, b, c, d, e, f or g, i denotes the ith amino acid index amongst the 529 amino acid indices, a represents the amino acid residue in the coiled-coil sequence whose heptad register is r, AA(a, i) stands for the value of the ith amino acid index for the amino acid a, while $n(r)$ is the number of amino acid residues at the heptad register r. As there are a total of 7 heptad registers and 529 amino acid indices, a coiled-coil sequence is represented by a 3703-dimensional vector.

Random forest. Ensemble learning is a prevalent machine learning technique. Its underlying principle is based on the observation that the ensemble of some weak classifiers can usually achieve a better accuracy than a single classifier when using the same training information. RF[35] is an effective ensemble learning algorithm and has been widely applied in bioinformatics.[36–41] RF consists of many decision trees, each of which is grown as follows. Suppose that there are N instances and M variables in the training set. First, N instances are randomly selected from the training set with replacement. Second, at each node, $\sqrt{M}$ variables are randomly selected and the best is used to split the node. Finally, each tree is grown as large as possible. The RF chooses the classification of the most votes given by all the individual trees. In this work, the random forest algorithm was implemented using the 'RandomForest' R package.[42]

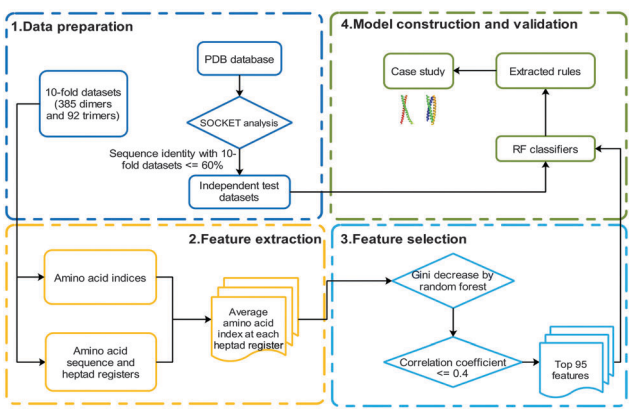


Fig. 2 Flowchart of *RFCoil*. Its development comprises four major steps, including data preparation, feature extraction, feature selection and RF model training and validation.

Feature selection and model training. As described above, a coiled-coil sequence was encoded by 3703 features. However, it is likely that some features were irrelevant or redundant, making little or no contribution to the prediction. We thus performed feature selection experiments to select and identify the most meaningful features for the classification of coiled-coil oligomerization states. For each feature, *i.e.* the variable in the RF, its importance is measured by the gini index of RF. When splitting the variable on a node in the process of growing a tree, the gini impurity criterion, which is a "goodness of split" criterion,[43] is less than the parent node for the two child nodes. Therefore, summing up the gini decrease for the variables over all trees gives the value for assessing the importance of the variable.

After evaluating the importance of each feature, another issue remains to be resolved. That is, the integration of individual best features does not necessarily lead to the best classification performance[44] and there still exists redundancy between different features. For example, there are many amino acid indices that describe the amino acid hydrophobicity in the AAindex database and some might be highly correlated with each other. To address this, we calculated the correlation coefficient between any two amino acid indices. If two features encode the same heptad register and the correlation coefficient of their representative amino acid indices has an absolute value of less than a threshold c, then the feature with a smaller gini decrease will be removed from the feature set. After this repetitive procedure, we select the top n features to build the final RF model.

In the above process, we used the Kendall rank correlation coefficient. Let $(X_1, X_2, \ldots, X_{20})$ and $(Y_1, Y_2, \ldots, Y_{20})$ be two sets of amino acid indices. A pair of amino acid index values (X_i, Y_i) and (X_j, Y_j) are defined to be concordant, if both $X_i > X_j$ and $Y_i > Y_j$ or both $X_i < X_j$ and $Y_i < Y_j$, and defined to be discordant, if $X_i > X_j$ and $Y_i < Y_j$ or $X_i < X_j$ and $Y_i > Y_j$. The Kendall correlation coefficient τ is defined as follows:

$$\tau = \frac{n_c - n_d}{\frac{1}{2} \times 20 \times (20-1)} \quad (2)$$

where n_c and n_d represent the numbers of concordant pairs and discordant pairs, respectively.

Extracting significant rules

Each tree in the RF can be represented by a set of rules. Each path from the root to a leaf node in a tree is a rule. A total of 4000 decision trees were grown in our work to build the RF model, resulting in the presence of many rules in the model. We devised a method to extract a rule set that contains as few rules as possible to correctly classify all the instances in the dataset: firstly, we extracted the rules without wrongly classifying any instance in the dataset and identified the rules that could classify the largest number of dimers or trimers; secondly, we

saved the rules found in the first step in the rule set and removed those instances that were correctly classified by the rule; thirdly, we repeated steps 1 and 2 until there were no instances in the dataset.

Accessing the prediction performance of the RF model

We used the receiver operating characteristic (ROC) curve[45] to assess the prediction performance of the RF model. The ROC curve is a plot of true positive rate (TPR) against false positive rate (FPR). TPR defines the ratio of correctly predicted positives to all the positive instances, while FPR stands for the ratio of incorrectly predicted positives to all the negative instances. In this study, we defined dimeric coiled-coils as positive instances and trimeric coiled-coils as negative instances. In addition, the area under the ROC curve (AUC) represents the probability of a classifier to rank a randomly selected positive instance higher than a randomly selected negative one. Hence, AUC was also used as an important performance measure in this study to compare the performance of different methods.

Performance comparison between *RFCoil* and four existing predictors

To evaluate the performance of *RFCoil*, we conducted two benchmarking experiments. In the first benchmarking experiment, we compare the performance of *RFCoil* with SCORER 2.0 and PrOCoil by performing 10-fold cross-validation tests on the PrOCoil dataset. In the second benchmarking experiment, we used the PrOCoil dataset as the training dataset to train the models of *RFCoil* and PrOCoil. Then the constructed independent test dataset was used to assess the performance of *RFCoil* in comparison with the other four tools SCORER 2.0, PrOCoil, Multicoil2 and LOGICOIL. In particular, the prediction outputs of SCORER 2.0, PrOCoil and LOGICOIL were generated by their local versions downloaded from the corresponding websites. In the case of Multicoil2, we instead submitted the test sequences to its online server and obtained the prediction results.

Results and discussion

In this section, we first report the prediction performance of *RFCoil* in comparison to SCORER 2.0 and PrOCoil on the 10-fold cross-validation tests. We then comprehensively assess the performance of *RFCoil*, PrOCoil, SCORER 2.0, LOGICOIL and Multicoil2 in the independent tests. Finally, we discuss the final features selected by our feature selection method and the extract significant rules on the PrOCoil benchmark dataset.

Prediction performance on the 10-fold cross-validation tests using the PrOCoil dataset

We performed 10-fold cross-validation tests to assess the performance of the predictive models of *RFCoil* using the PrOCoil dataset (Table 1). When using the average amino acid index values at each heptad as the input, the average AUC of *RFCoil* was 0.819, compared with 0.808 of PrOCoil and 0.789 of SCORER 2.0, respectively. After setting the Kendall correlation coefficient between the amino acid indices at ≤ 0.4 to select the 95 top features, the average AUC of *RFCoil* was further improved to 0.849. The authors of PrOCoil[22] found that the training set could be further augmented by blast search against the NCBI-NR database, which could provide an improved prediction performance in their study. Here, our results indicate that the AUC of PrOCoil on the augmented training dataset indeed reached 0.818, representing a better performance than that of the original PrOCoil. On the other hand, we find that *RFCoil* performed the best for certain folds and reasonably well for other folds during 10-fold cross-validation tests (Table 1). In summary, *RFCoil* achieved a better performance than the other two methods PrOCoil and SCORER 2.0 on the 10-fold cross-validation tests using the PrOCoil dataset. According to the 10-fold cross-validation tests, we implemented the final online web server of *RFCoil* using the selected feature set.

Table 1 The AUC scores of *RFCoil*, SCORER 2.0 and PrOCoil, evaluated using 10-fold cross-validation tests

Fold	*RFCoil* (all features)	*RFCoil* (selected features)	SCORER 2.0	PrOCoil	PrOCoil_blast[a]
1	0.612	0.691	0.773	0.882	0.882
2	0.801	0.817	0.776	0.967	0.935
3	0.750	0.835	0.625	0.581	0.681
4	0.885	0.875	0.810	0.830	0.850
5	0.971	0.957	0.833	0.848	0.867
6	0.869	0.865	0.808	0.741	0.842
7	0.908	0.961	0.875	0.809	0.724
8	0.803	0.769	0.735	0.744	0.744
9	0.698	0.825	0.651	0.738	0.702
10	0.890	0.895	1.000	0.943	0.957
Average	0.819	0.849	0.789	0.808	0.818

[a] PrOCoil_blast denotes the model trained using the augmented PrOCoil dataset using blast search against the NCBI-NR database.

Prediction performance on the independent tests

In addition to the performance evaluation using the PrOCoil benchmark dataset, we also curated an independent test dataset to comprehensively compare the performance of our method *RFCoil* for predicting the coiled-coil oligomerization state with four existing predictors SCORER 2.0, PrOCoil, Multicoil2 and LOGICOIL. In particular, we used the PrOCoil dataset as the training set to build the two types of predictive models for *RFCoil* (denoted as "*RFCoil* (all features)" and "*RFCoil* (selected features)" which used all features and final selected features as the respective inputs to build the models) to classify coiled-coil sequences in this independent test dataset. LOGICOIL and SCORER 2.0 were trained on the coiled-coil sequences no shorter than 15 amino acids, while Multicoil2 could only predict coiled-coil sequences longer than 21 amino acids. In the training dataset of PrOCoil, the minimum length of coiled-coil sequences is 8 amino acids. In this study, we reported the results by performing the independent test using our independent test dataset with the minimum length of coiled-coil sequences of 8 amino acids.

The output scores were selected from two prediction categories of LOGICOIL (*i.e.*, parallel dimer and trimer) and normalized to [0,1] before plotting the ROC curve. Instead of providing an overall

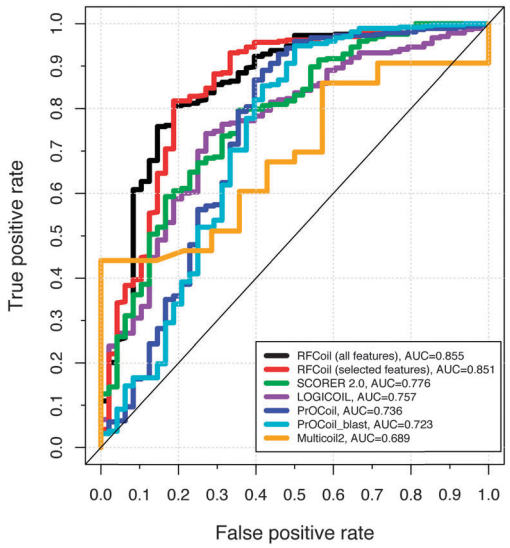


Fig. 3 The ROC curves of different methods on the independent test dataset.

prediction score for the input sequence, Multicoil2 provides predicted probabilities for each individual residue in the sequence of forming dimers, trimers or non-coiled-coils. Accordingly, to compare with other methods, we calculated the average of the predicted probabilities of Multicoil2, normalized them into the range of [0,1] and removed the predicted non-coiled-coils from the results (with the prediction threshold set at 0.5).

The ROC curves and the corresponding AUC values of *RFCoil*, SCORER 2.0, PrOCoil, LOGICOIL and Multicoil2 in the independent tests are shown in Fig. 3. The AUC values of the two types of *RFCoil* models that used all features and the final selected features as inputs were 0.855 and 0.851, respectively. These represent the overall best AUC scores among different predictors. In contrast, Multicoil2 achieved an AUC value of 0.689, while SCORER 2.0 achieved an AUC score of 0.776. PrOCoil achieved an AUC value of 0.736 and the PrOCoil_blast model trained using the augmented dataset achieved an AUC of 0.723, both of which decreased considerably compared to that upon the 10-fold cross validation. In contrast, LOGICOIL achieved an AUC value of 0.757. We also noted that augmenting the training set in this case did not help improve the performance of PrOCoil, as reflected by a lower AUC of 0.723 obtained using the latter model.

Analysis of final selected features based on the PrOCoil dataset

Application of the Kendall correlation coefficient set at ≤ 0.4 resulted in a subset of top 95 features selected (see Table S3, ESI†). The average AUC of the *RFCoil* model trained using this selected feature set reached its maximum value of 0.849 on the 10-fold cross-validation tests using the PrOCoil benchmark dataset (Table 1). We further calculated the number of features at each heptad register, as well as the sum of the gini decreases for the features at each heptad register. Table 2 shows that the position *a* is the most important position for the discrimination between parallel dimers and trimmers, as determined by the sum of the gini decreases. The other positions *d*, *e*, *c*, *g* are less important compared with the position *a*, while positions *f* and *b* are the least important positions.

Table 2 Statistics of the selected features

Heptad register	*a*	*b*	*c*	*d*	*e*	*f*	*g*
Number of features	13	5	8	10	9	5	8
Sum of the gini decrease	35.6	5.8	12.5	16.6	18.4	7.8	11.8

Significant rules extracted from the PrOCoil dataset

Using the method of rule extraction described in the Methods section, we extracted 10 significant rules covering all the 382 dimers, and another 10 significant rules covering all the 92 trimers in the PrOCoil dataset. The description of each specific rule and the numbers of dimers and trimers covered by the corresponding rule are given in Tables 3 and 4, respectively. Note that it is likely that a sample in the dataset may be identified by more than two rules, as shown in the tables.

Each rule is a combination of useful amino acid indices at certain heptad registers. The RF algorithm is particularly powerful in making use of the correlations between different heptad registers for efficient classification. In contrast, SCORER 2.0 only uses residue frequencies at each heptad register, failing to take into account the potential interactions between different heptad-repeat positions, while PrOCoil employs the frequencies of each amino acid pair in each pair of heptad registers. An important advantage of RF is that it can make use of the correlations between two or more heptad registers. This might explain why our method outperformed the other four methods PrOCoil, Multicoil 2, SCORER 2.0 and LOGICOIL.

Case studies

Using the selected 95 features on the PrOCoil dataset, we built the RF model and illustrated the performance of this model on two parallel coiled-coil structures from the independent test dataset (see Fig. S1 for structural information regarding these two proteins, ESI†). The first one is a coiled-coil parallel dimer from the Rho-associated protein kinase 1 (PDB ID: 3O0Z). This protein is involved in a variety of cellular processes including muscle contraction, cell migration and stress fiber formation.[46] Its predicted probability of being dimeric by the RF model was 0.872. The other is a trimer from the avian reovirus S1133 fibre (PDB ID: 2VRS), a minor component of the avian reovirus outer capsid.[47] Its probability of being a parallel trimer predicted by the RF model was 0.759. The coiled-coil oligomerization states of both proteins were correctly predicted by *RFCoil*.

In addition, we found that the dimeric coiled-coil in the Rho-associated protein kinase 1 conformed to the significant rules 1, 2, 5 and 10, as listed in Table 3. Further, the trimeric coiled-coil in 2VRS conformed to the significant rules 1 and 5 listed in Table 4. Altogether, these results showcase the predictive ability of the constructed *RFCoil* model and usefulness of the extracted rules based on the selected effective amino acid indices.

Table 3 The extracted rules for coiled-coil dimers

No.	Description of the rule[a]	Number of samples covered by the rule
1	$I(c, 260) \leq 0.2825$ & $I(d, 17) > 4.2435$ & $I(f, 16) > 7.213$ & $I(f, 240) > -3.3475$ & $I(a, 294) > -0.2925$ & $I(a, 400) \leq 14.183$	225
2	$I(c, 340) \leq 5.83$ & $I(d, 17) > 4.2315$ & $I(d, 195) > 2.1225$ & $I(e, 74) > -62.35$ & $I(f, 16) \leq 8.676$ & $I(f, 73) > 240.0835$ & $I(g, 50) \leq 0.088$ & $I(g, 201) \leq 1.654$ & $I(g, 408) > 1.1735$ & $I(b, 18) \leq 7.3085$ & $I(b, 273) > -0.375$	173
3	$I(c, 371) > 0.355$ & $I(d, 220) \leq 2.9275$ & $I(e, 372) \leq 2.202$ & $I(e, 495) \leq 0.9795$ & $I(g, 61) > 0.3625$ & $I(a, 386) \leq 0.388$	128
4	$I(e, 299) \leq 1.165$ & $I(a, 275) > 0.1125$ & $I(a, 374) \leq 0.7665$	58
5	$I(c, 194) > -1.4475$ & $I(c, 293) \leq 0.4225$ & $I(c, 340) \leq 4.169$ & $I(d, 342) > -1.2415$ & $I(d, 401) \leq 1.22$ & $I(f, 338) \leq 1.4625$ & $I(g, 155) > 107.1895$ & $I(g, 201) > 0.759$ & $I(a, 44) > 0.5575$ & $I(b, 529) > -3.1775$	201
6	$I(c, 303) \leq 1.2345$ & $I(d, 17) > 4.279$ & $I(d, 342) > -0.425$ & $I(e, 110) > 0.3625$ & $I(g, 336) \leq 0.8415$ & $I(a, 386) \leq 0.1705$ & $I(a, 506) > 1.4695$	34
7	$I(c, 18) \leq 6.9585$ & $I(c, 361) > -0.177$ & $I(e, 296) \leq 0.2385$ & $I(a, 400) \leq 16.35$ & $I(a, 506) \leq 1.7425$ & $I(b, 185) \leq 4.195$ & $I(a, 99) \leq 1.54$	183
8	$I(a, 107) > 0.7325$ & $I(d, 401) \leq 1.21$ & $I(a, 1) \leq 4.7025$ & $I(a, 294) > -0.335$ & $I(g, 370) \leq 0.773$ & $I(b, 185) \leq 4.195$	185
9	$I(c, 326) \leq 1.5165$ & $I(e, 296) \leq 0.28$ & $I(e, 495) \leq 0.9985$ & $I(g, 408) \leq 1.171$	60
10	$I(c, 18) > 6.89$ & $I(c, 141) > 0.45$ & $I(d, 94) > 0.8835$ & $I(d, 275) \leq 0.097$ & $I(e, 296) > 0.161$ & $I(f, 331) \leq 1.2875$ & $I(g, 61) \leq 1.056$ & $I(a, 337) > 0.7415$	9

[a] "&" denotes the conjunction word "and", while $I(r, n)$ represents the nth amino acid index at the heptad r.

Table 4 The extracted rules of coiled-coil trimers

No.	Description of the rule[a]	Number of samples covered by the rule
1	$I(c, 236) > 0.795$ & $I(c, 361) \leq 0.123$ & $I(d, 326) \leq 0.7415$ & $I(e, 219) > 0.945$ & $I(e, 299) > 1.1665$ & $I(g, 201) > 0.536$ & $I(g, 309) > 0.8665$ & $I(a, 400) > 14.1515$ & $I(a, 506) > 1.464$	44
2	$I(c, 293) > -0.324$ & $I(c, 361) \leq 0.123$ & $I(c, 405) \leq 1.2725$ & $I(d, 175) \leq 0.8575$ & $I(e, 110) > 0.3725$ & $I(f, 16) \leq 8.5555$ & $I(g, 408) > 0.655$ & $I(a, 374) \leq 0.826$ & $I(b, 529) \leq -3.167$ & $I(b, 273) > -0.1685$	43
3	$I(c, 340) > 0.096$ & $I(a, 176) > 0.675$ & $I(d, 195) > 5.3525$ & $I(f, 73) \leq 245.6$ & $I(f, 385) > -0.0975$ & $I(a, 386) \leq 0.1365$ & $I(b, 18) > 5.85$	17
4	$I(d, 94) \leq 1.154$ & $I(a, 18) > 5.125$ & $I(g, 336) > 0.8415$ & $I(a, 374) \leq 0.765$ & $I(a, 400) > 12.6765$ & $I(b, 18) \leq 7.7415$ & $I(b, 273) > -0.104$	30
5	$I(c, 361) \leq -0.176$ & $I(d, 94) \leq 1.2915$ & $I(a, 176) \leq 0.8375$ & $I(e, 360) \leq 0.2115$ & $I(b, 329) \leq 1.325$	19
6	$I(c, 141) > 0.655$ & $I(e, 296) > 0.2665$ & $I(g, 408) \leq 0.9935$ & $I(a, 374) > 0.7135$	8
7	$I(a, 107) > 0.7505$ & $I(d, 220) \leq 2.9275$ & $I(d, 240) \leq -2.141$ & $I(e, 495) > 0.9795$ & $I(a, 294) \leq -0.245$	5
8	$I(d, 195) \leq 9.1325$ & $I(d, 422) > -0.501$ & $I(e, 295) > -0.061$ & $I(e, 372) > 0.1965$ & $I(f, 73) \leq 267.9165$ & $I(a, 374) > 0.6285$ & $I(a, 400) > 14.385$ & $I(a, 99) \leq 1.2675$	16
9	$I(c, 340) > -0.0625$ & $I(c, 371) > 1.061$ & $I(f, 16) > 8.481$ & $I(g, 12) > -4.7165$ & $I(b, 284) > -0.06$	14
10	$I(b, 478) > 1.6165$ & $I(c, 361) > -0.1935$ & $I(d, 74) > -25.1175$ & $I(d, 422) > -0.3215$ & $I(g, 98) > 1.0125$ & $I(g, 370) \leq 0.773$	3

[a] See the footnote in Table 3 for the notations of each symbols in the rules.

Conclusions

In this article, we addressed the challenging task of distinguishing parallel dimeric from trimeric coiled-coils by developing an RF-based approach termed as *RFCoil*, which used effective amino acid indices to build the predictive models. To remove redundant and irrelevant features and improve the classification performance, we combined the gini index calculated by RF and the correlation coefficients between the amino acid indices at different positions of heptad registers to select the most meaningful features. The model trained using the selected features indeed improved the prediction performance. We further analyzed the selected features and proposed a rule extraction method to identify significant rules from the RF model to better understand the important rules that underlie the organization of dimeric and trimeric coiled-coils. The rules provide useful insights into the design of coiled-coil proteins. In addition, our method can be readily extended to predict coiled-coils of higher order oligomerization states, provided that more solved structures are available in the near future. Benchmarking experiments indicate that *RFCoil* outperforms the other four existing tools. It is expected to become an efficient tool to facilitate the studies of coiled-coil structures. Finally, as an implementation of our method, an online prediction server of *RFCoil* has been made freely available at http://protein.cau.edu.cn/RFCoil. The source code can be downloaded for interested users to build their specific models using their own datasets.

Acknowledgements

This work is supported by the Hundred Talents Program of the Chinese Academy of Sciences (CAS), the National Natural

Science Foundation of China (No. 61202167, 61303169, 11250110508, 31350110507), the Knowledge Innovation Program of CAS (No. KSCX2-EW-G-8), the Tianjin Municipal Science & Technology Commission (No. 10ZCKFSY05600), the Major Inter-Disciplinary Research (IDR) Project awarded by Monash University and the National Health and Medical Research Council of Australia (NHMRC) (No. 490989). JS is an NHMRC Peter Doherty Fellow and a recipient of the Hundred Talents Program of CAS.

References

1 G. Grigoryan and A. E. Keating, *Curr. Opin. Struct. Biol.*, 2008, **18**, 477–483.
2 T. L. Vincent, D. N. Woolfson and J. C. Adams, *Int. J. Biochem. Cell Biol.*, 2013, **45**, 2392–2401.
3 A. A. McFarlane, G. L. Orriss and J. Stetefeld, *Eur. J. Pharmacol.*, 2009, **625**, 101–107.
4 T. L. Vincent, P. J. Green and D. N. Woolfson, *Bioinformatics*, 2013, **29**, 69–76.
5 A. N. Lupas and M. Gruber, *Adv. Protein Chem.*, 2005, **70**, 37–78.
6 Y. Wang, X. Zhang, H. Zhang, Y. Lu, H. Huang, X. Dong, J. Chen, J. Dong, X. Yang, H. Hang and T. Jiang, *Mol. Biol. Cell*, 2012, **23**, 3911–3922.
7 N. R. Zaccai, B. Chi, A. R. Thomson, A. L. Boyle, G. J. Bartlett, M. Bruning, N. Linden, R. B. Sessions, P. J. Booth, R. L. Brady and D. N. Woolfson, *Nat. Chem. Biol.*, 2011, **7**, 935–941.
8 S. F. Betz, J. W. Bryson and W. F. DeGrado, *Curr. Opin. Struct. Biol.*, 1995, **5**, 457–463.
9 K. Chen and L. Kurgan, *Methods Mol. Biol.*, 2013, **932**, 63–86.
10 A. Lupas, *Trends Biochem. Sci.*, 1996, **21**, 375–382.
11 E. H. Bromley, K. Channon, E. Moutevelis and D. N. Woolfson, *ACS Chem. Biol.*, 2008, **3**, 38–50.
12 J. Walshaw and D. N. Woolfson, *J. Mol. Biol.*, 2001, **307**, 1427–1450.
13 P. B. Harbury, J. J. Plecs, B. Tidor, T. Alber and P. S. Kim, *Science*, 1998, **282**, 1462–1467.
14 L. Bartoli, P. Fariselli, A. Krogh and R. Casadio, *Bioinformatics*, 2009, **25**, 2757–2763.
15 O. J. Rackham, M. Madera, C. T. Armstrong, T. L. Vincent, D. N. Woolfson and J. Gough, *J. Mol. Biol.*, 2010, **403**, 480–493.
16 M. Delorenzi and T. Speed, *Bioinformatics*, 2002, **18**, 617–625.
17 J. Trigg, K. Gutwin, A. E. Keating and B. Berger, *PLoS One*, 2011, **6**, e23519.
18 E. Wolf, P. S. Kim and B. Berger, *Protein Sci.*, 1997, **6**, 1179–1189.
19 A. V. McDonnell, T. Jiang, A. E. Keating and B. Berger, *Bioinformatics*, 2006, **22**, 356–358.
20 B. Berger, D. B. Wilson, E. Wolf, T. Tonchev, M. Milla and P. S. Kim, *Proc. Natl. Acad. Sci. U. S. A.*, 1995, **92**, 8259–8263.
21 A. Lupas, M. Van Dyke and J. Stock, *Science*, 1991, **252**, 1162–1164.
22 C. C. Mahrenholz, I. G. Abfalter, U. Bodenhofer, R. Volkmer and S. Hochreiter, *Mol. Cell. Proteomics*, 2011, **10**, M110004994.
23 C. T. Armstrong, T. L. Vincent, P. J. Green and D. N. Woolfson, *Bioinformatics*, 2011, **27**, 1908–1914.
24 J. R. Apgar, K. N. Gutwin and A. E. Keating, *Proteins*, 2008, **72**, 1048–1065.
25 S. V. Strelkov and P. Burkhard, *J. Struct. Biol.*, 2002, **137**, 54–64.
26 O. D. Testa, E. Moutevelis and D. N. Woolfson, *Nucleic Acids Res.*, 2009, **37**, D315–D322.
27 F. H. Crick, *Acta Crystallogr.*, 1953, **6**, 689–697.
28 D. N. Woolfson and T. Alber, *Protein Sci.*, 1995, **4**, 1596–1607.
29 P. S. Kim, B. Berger and E. Wolf, *Protein Sci.*, 1997, **6**, 1179–1189.
30 P. Lavigne, M. P. Crump, S. M. Gagne, R. S. Hodges, C. M. Kay and B. D. Sykes, *J. Mol. Biol.*, 1998, **281**, 165–181.
31 P. A. Bullough, F. M. Hughson, J. J. Skehel and D. C. Wiley, *Nature*, 1994, **371**, 37–43.
32 H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, *Nucleic Acids Res.*, 2000, **28**, 235–242.
33 S. B. Needleman and C. D. Wunsch, *J. Mol. Biol.*, 1970, **48**, 443–453.
34 S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama and M. Kanehisa, *Nucleic Acids Res.*, 2008, **36**, D202–D205.
35 L. Breiman, *Mach. Learn.*, 2001, **45**, 5–32.
36 X. F. Wang, Z. Chen, C. Wang, R. X. Yan, Z. Zhang and J. Song, *PLoS One*, 2011, **6**, e26767.
37 C. Zheng, M. Wang, K. Takemoto, T. Akutsu, Z. Zhang and J. Song, *PLoS One*, 2012, **7**, e49716.
38 M. M. Dehmer, N. N. Barbarini, K. K. Varmuza and A. A. Graber, *BMC Struct. Biol.*, 2010, **10**, 18.
39 S. Hirose, K. Yokota, Y. Kuroda, H. Wako, S. Endo, S. Kanai and T. Noguchi, *BMC Struct. Biol.*, 2010, **10**, 20.
40 M. Wang, X. M. Zhao, K. Takemoto, H. Xu, Y. Li, T. Akutsu and J. Song, *PLoS One*, 2012, **7**, e43847.
41 Z. P. Liu, L. Y. Wu, Y. Wang, X. S. Zhang and L. Chen, *Bioinformatics*, 2010, **26**, 1616–1622.
42 A. Liaw and M. Wiener, *R News*, 2002, **2**, 18–22.
43 L. Raileanu and K. Stoffel, *Ann. Math. Artif. Intell.*, 2004, **41**, 77–93.
44 H. Peng, F. Long and C. Ding, *IEEE Trans. Pattern Anal. Mach. Intell.*, 2005, **27**, 1226–1238.
45 T. Fawcett, *Pattern Recognit. Lett.*, 2006, **27**, 861–874.
46 D. Tu, Y. Li, H. K. Song, A. V. Toms, C. J. Gould, S. B. Ficarro, J. A. Marto, B. L. Goode and M. J. Eck, *PLoS One*, 2011, **6**, e18080.
47 P. Guardado-Calvo, G. C. Fox, A. L. Llamas-Saiz and M. J. van Raaij, *J. Gen. Virol.*, 2009, **90**, 672–677.

