Recognition of Protease Inhibition Pattern using Topological Interaction Matrix and Genetic Algorithm-optimized Support Vector Machines

Michael Fernandez, Shandar Ahmad, and Akinori Sarai

***Abstract*— Proteases are among the most studied diagnostic and therapeutic targets for a variety of human diseases. Intensive research in computational design of protease inhibitors has been done by molecular dynamics simulations, docking and Quantitative Structure-Activity Relationships. In this work, proteochemometrics was applied to the recognition of stable and unstable protease inhibition complexes from a large dataset (>1700) using Genetic Algorithm-optimized Support Vector Machines classifiers. Genetic Algorithm optimized SVM were trained with Topological Autocorrelation Interaction vectors, computed on the protease sequences and the inhibitor 2D structure sketches. Optimum classifier with 10 inputs correctly predicted more than 80% in both training and test sets.**

***Index Terms*—kernel-based methods, feature selection, peptidases, structure-activity relationship.**

I. Introduction

Proteases are a family of enzymes representing approximately 2% of an organism proteosome which take part in bioregulation, matrix remodeling, digestion, and immune response processes.[1] These enzymes occur in virtually all biological processes and their ability to alter catalytic turnover, makes them ideal biomarkers for diseases diagnostic and therapy, resulting in about 5–10% of the targets in the current pharmaceutical market.[2]

The development of novel diagnostic and therapeutic protease-active compounds depends on the understanding of protease's inhibition mechanism and specificity. Identification of protease substrates and hydrolytic products will contribute in elucidating the mechanisms behind the progression of disease and increase opportunities for the development of drug candidates and their interventional use. One experimental approach is to screen a protease against a synthetic combinatorial library of small molecules, in a format that systematically probes each subsite and collectively generates a positional profile. This information can be used to design a specificity element, or a compound with similar spatial and electronic properties such as a protease's natural substrate.[3]

Computational modeling constitutes an alternative and parallel way of gaining insights into target-ligand interactions. Enzyme-substrate interactions have been studied in-silico by means of molecular dynamics simulations, Quantum Mechanical/Molecular Mechanical (QM-MM), docking, and Quantitative-Structure Activity Relationships (QSAR).[4] In the current paper, we applied 2D autocorrelation methodology to a dataset of ligands with reported inhibitory activities towards 32 proteases. Proteochemometrics (PCMs)[5] was employed for developing classification models of proteases inhibition. The structural information of the target proteases was encoded in Amino Acids Sequence Autocorrelation (AASA) vectors, a structure encoding scheme of protein sequence previously reported by us in proteometrics studies.[6] Afterwards, the target-ligand Topological Autocorrelation Interaction (TAI) matrix was computed as the matrix product of the proteases AASA vectors and inhibitor 2D autocorrelation descriptors. Support Vector Machines (SVMs) were optimized by Genetic Algorithm (GA).

II. Material and Methods

A. DATA SET

Protease inhibition data was collected from the literature, annotation includes inhibitor structures, inhibitory activity, activity type and unit as well as protein code for protease tested in each report. Protease sequences were retrieved from PIR database[7] and added to the protease inhibition data. Instant JChem software[8] was used for chemical database management. A dataset of 1706 proteases inhibition complexes, including peptide and nonpeptide ligands, was used for training and testing SVM classifiers. Affinity threshold was selected as Ki=0.1 μM and complexes with Ki < 0.1 were considered "unstable", while those with Ki > 0.1 were considered "stable" class. According to this criterion, we get 718 stable and 988 unstable inhibition complexes. TAI matrix was calculated as the matrix product between 2D autocorrelation vectors and AASA vectors. TAI data was divided into training (80% dataset) and test sets (20% dataset) by using a k-means clustering algorithm. Five clusters were generated and cases were added homogeneously to training and test sets by selecting instances from each cluster according to cluster sizes.

M. Fernandez is with the Department of Bioscience and Bioinformatics, Kyushu Institute of Technology (KIT), 680-4 Kawazu, Iizuka, 820-8502 Japan.; (corresponding author to provide phone: 0948-29-7811; fax: 0948-29-7841; e-mail: michael_llamosa@yahoo.com)

S. Ahmad is with the National Institute of Biomedical Innovation, 7-6-8, Saito-Asagi, Ibaraki-shi, Osaka 5670085, Japan (e-mail: shandar@nibio.go.jp).

A. Sarai is with the Department of Bioscience and Bioinformatics, Kyushu Institute of Technology (KIT), 680-4 Kawazu, Iizuka, 820-8502 Japan.; (e-mail: sarai@bio.kyutech.ac.jp)

M. Chetty, S. Ahmad, A. Ngom, S. W. Teng (Eds.): PRIB2008, Supplementary Conf. Proc., pp.246-249, 2008.

B. Proteochemometrics (PCMs) modeling

PCMs are typically described by three descriptor blocks; the ligand descriptor (DL*)*, protein descriptor (DP), and ligand-protein cross-term (DLP) blocks. A vector of variables for ligands, DL, characterizes each ligand *L*. Similarly, each protein *P* has its *DP*. Depending on the faced problem one or more descriptors blocks can be discarded.[5] In our study, a DLP block, called TAI matrix was calculated as the matrix product of the inhibitor 2D autocorrelation vectors and the AASA vectors of the proteases. Moreau's autocorrelation vectors were employed for encoding the topological structure of the protease inhibitors.

Broto-Moreau's autocorrelation coefficient:[9]

$$ATSlp_k = \sum_i \delta_{ij} p_{ki} p_{kj} \qquad (1)$$

where $ATSlp_k$ is Broto-Moreau's autocorrelation coefficient at spatial lag *l* respectively; p_{ki} and p_{kj} are the values of property *k* of atom *i* and *j* respectively and $\delta(l, d_{ij})$ is a Dirac-delta function defined as

$$\delta(l, d_{ij}) = \begin{Bmatrix} 1 \;\; if \;\; d_{ij} = l \\ 0 \;\; if \;\; d_{ij} \neq l \end{Bmatrix} \qquad (2)$$

where d_{ij} is the topological distance or spatial lag between atoms *i* and *j*. 2D autocorrelation vectors at spatial lags ranging from 1 to 8 were weighted by 3 atomic properties: atomic van der Waals volumes, atomic Sanderson electronegativities and atomic polarizabilities, thus a total of 24 (8×3) 2D autocorrelation vectors were computed.

AASA vectors of lag *l* are calculated as follows:[6]

$$AASAlp_k = \frac{1}{L} \sum_i \delta_{ij} p_{ki} p_{kj} \qquad (3)$$

where $AASAlp_k$ is the *AASA* at spatial lag *l* weighted by the p_k property; *L* is the number of elements in the sum; p_{ki} and p_{kj} are the values of property *k* of amino acids *i* and *j* in the sequence respectively and $\delta(l, d_{ij})$ is the Dirac-delta function in Eq. 2.

Seven physicochemical and conformational amino acid/ residues properties and spatial lag, *l*, ranging from 1 to 5, were used as weights for sequence residues. Computational code for AASA vector calculation was written in Matlab environment.[10] A data matrix of 35 AASA vectors, 7 properties×5 different lags, were generated with the autocorrelation vectors calculated for each protease.

The TAI matrix was calculated as the matrix product of the AASA vectors of the five proteases and the inhibitor 2D autocorrelation vectors, resulting in 480 TAI descriptors, 35 AASA vectors×24 2D autocorrelation vectors. TAI descriptors are calculated as follows:

$$TAIl_1 p^1{}_k l_2 p^2{}_o = AASAl_1 p^1{}_k \times ATSl_2 p^2{}_o \qquad (4)$$

where $TAIl_1 p^1{}_k l_2 p^2{}_o$ is the topological autocorrelation interaction at spatial lag l_1 in the protein sequence weighted by the amino acid and/or residue property $p^1{}_k$ and at spatial lag l_2 in the ligand topological structure weighted by atomic property $p^2{}_o$; $AASAl_1 p^1{}_k$ is the AASA at spatial lag l_1 in the protein sequence weighted by the amino acid and/or residue property $p^1{}_k$; $ATSl_2 p^2{}_o$ is Broto-Moreau's autocorrelation coefficient at spatial lag l_2 weighted by the atomic property $p^2{}_o$.

C. Support Vector Machine (SVM)

SVM is a machine learning method, which has been used for many kinds of pattern recognition problems.[11] First, the input vectors are mapped onto one feature space (possible with a higher dimension). Secondly, a hyperplane, which can separate two classes, is constructed within this feature space. Only relatively low-dimensional vectors in the input space and dot products in the feature space will involve by a mapping function. SVM was designed to minimize structural risk whereas previous techniques were usually based on minimization of empirical risk. The mapping into the feature space is performed by a kernel function. There are several parameters in the SVM, including the kernel function and regularization parameter. GA-based SVM (GA-SVM) algorithm was implemented for selecting optimum subset of input training vectors and setting the two SVM parameters, regularization parameter and width of the RBF kernel, to optimum values. The toolbox used to implement the SVM with RBF kernel (RBF-SVM) was LIBSVM for Matlab by Chang and Lin[12] that can be downloaded from: http://www.csie.ntu.edu.tw/cjlin/libsvm/.

D. Genetic Algorithm based feature selection and hyperparameter optimization

GA was applied for selection of the optimum subset of variables and the optimization of regularization parameter and width of an RBF kernel. We simply concatenated a representation of the parameters to a chromosome encoding the subset of variables used for SVM training.[13] Usually it is not necessary to consider any arbitrary value but only certain discrete values with the form: $n \times 10^k$, where n=1…9 and k=-4…4. So, these values can be calculated by randomly generating *n* and *k* values as integers between (1…9) and (-4…4), respectively. In this way, GA optimized regularization parameter and the width of an RBF kernel.

A three-fold-out crossvalidation assessed model's quality throughout the GA search. Three data subsets were created, two subsets are generated in the crossvalidation process for training the SVM and another subset is then predicted. This process is repeated until all subsets have been predicted. The GA routine minimized the misclassification percent of three-fold-out (MCP_{TFO}) crossvalidation. GA-SVM was implemented in Matlab environment.[10]

III. RESULTS AND DISCUSSION

We applied a topological framework for obtaining a feature data matrix for SVM training. Hyperparameters and optimum training TAI descriptors for SVMs were optimized by GA. The procedure yielded optimum subsets of *TAI* descriptors calculated over the weighted linear graph representations of the protease sequences and ligands 2D topological representation.

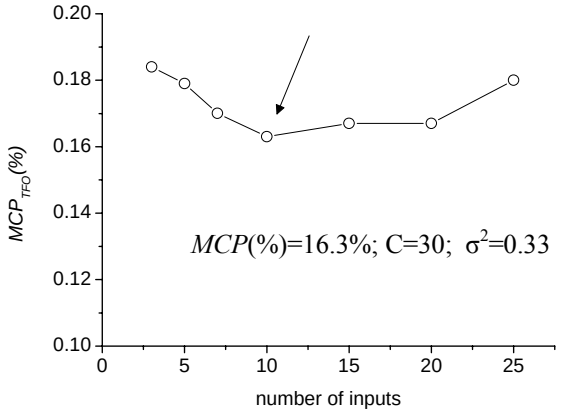

Figure 1. Plots of MCP_{TFO} in the GA-SVM search vs. the number of inputs in the models for discrimination between stable and unstable protease-inhibition complexes. Arrows point out the optimum number of inputs with lowest MCP(%)=16.3% for 10 inputs and hyperparameters values of C=30 and σ^2=0.33.

GA-SVMs were implemented in a first attempt with the simpler linear kernel but the highest crossvalidation accuracy was not higher than 70% for the classification of the stability of inhibition complexes. Afterwards, nonlinear RBF kernel was used inside the SVM framework. The GA algorithm optimized the hyperparameters, the kernel regularization parameter C and the width of an RBF kernel σ^2. Nonlinear subspace in the dataset was searched varying problem dimension from 3 to 25. Figure 1 depicts the behaviors of the minimum MCP_{TFO} values yielded throughout the GA search vs. the number of SVM inputs. The best model with 10 inputs misclassified only 16.3% of the cases in three-fold-out crossvalidation experiments. Optimum values of regularization parameter and width of the RBF kernel were 30 and 0.33, respectively. Table 1 reflects that the accuracies for predicting stable and unstable inhibition complexes separately were 84%. According to this result the predictor recognized both classes with similar accuracies in the internal validation. Although, the internal validation is a measure of the robustness of the model, a more realistic measurement of the predictive power of a model should be evaluated on an independent test set. As can be observed in Table 1, the accuracies for prediction of stable and unstable inhibition complexes were 80% and 82%, respectively with an overall accuracy about 81%. The performance of the classifier on the independent test set is about 80% of data correctly predicted.

Table 1. Hyperparameters and statistics of crossvalidation and test set prediction of optimum model for the classification of protease inhibition complexes

	Q^2	*Q(+)*	*Q(-)*	*P(+)*	*P(-)*
Training set crossvalidation	0.84	0.84	0.84	0.80	0.87
Test set prediction	0.81	0.80	0.82	0.74	0.87

+ and - : the indexes account for stable and unstable. Q^2 is the number of correct predictions/number of examples and *Q(s)* is the number of correct prediction for class s/observed in class s. *Q(+)* and *Q(-)* are sensitivity and specificity of stable class prediction and *P(+)* and *P(-)* are precision scores.

Inputs in the optimum model are: *TAI4R$_\alpha$6p*, is the topological autocorrelation interaction of lag 4 weighted by solvent-accessible reduction ratio in the target sequence and lag 6 weighted by polarizability in the ligand; *TAI2Ht1v* is the topological autocorrelation interaction of lag 2 weighted by thermodynamic transfer hydrophobicity in the target sequence and lag 1 weighted by van der Waals volume in the ligand; *TAI1pK'1e* is the topological autocorrelation interaction of lag 5 weighted by equilibrium constant with reference to the ionization property of COOH group in the target sequence and lag 1 weighted by electronegativity in the ligand; *TAI5ASA$_N$6p* is the topological autocorrelation interaction of lag 5 weighted by solvent-accessible surface area for native protein in the target sequence and lag 6 weighted by polarizability in the ligand; *TAI4s8p* is the topological autocorrelation interaction of lag 3 weighted by shape (position of branch point in a side chain) in the target sequence and lag 8 weighted by polarizability in the ligand; *TAI3V^{0}5v* is the topological autocorrelation interaction of lag 3 weighted by partial specific volume in the target sequence and lag 5 weighted by van der Waals volume in the ligand; *TAI3pK'1v* is the topological autocorrelation interaction of lag 3 weighted by equilibrium constant with reference to the ionization property of COOH group in the target sequence and lag 1 weighted by van der Waals volume in the ligand; *TAI3pK'4v* is the topological autocorrelation interaction of lag 3 weighted by equilibrium constant with reference to the ionization property of COOH group in the target sequence and lag 4 weighted by van der Waals volume in the ligand; *TAI1Ht6e* is the topological autocorrelation interaction of lag 1 weighted by thermodynamic transfer hydrophobicity in the target sequence and lag 6 weighted by electronegativity in the ligand and *TAI1ASA$_N$4v* is the topological autocorrelation interaction of lag 1 weighted solvent-accessible surface area for native protein in the target sequence and lag 4 weighted by van der Waals volume in the ligand.

According to Figure 2, thermodynamic transfer hydrophobicity, equilibrium constant with reference to the ionization property of COOH group and solvent-accessible surface area for native protein, were the most relevant properties used for weighting the sequence linear graphs according to the GA feature selection. This fact suggests that hydrophobic and electrostatic

natures of the residues along the sequence rule protease-ligand inhibitions.

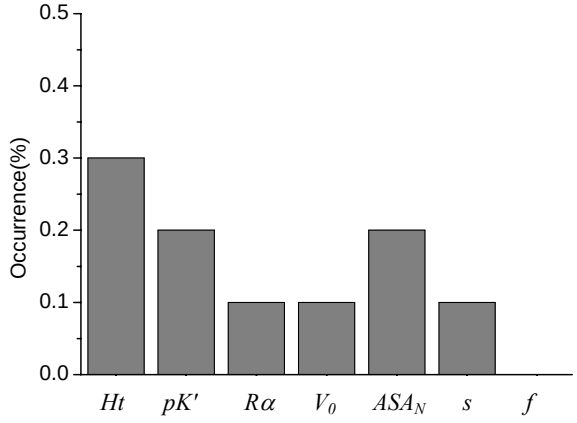

Figure 2. Relative occurrences of the seven properties in the optimum 10 inputs variables. H_t, thermodynamic transfer hydrophobicity; *pK',* equilibrium constant with reference to the ionization property of COOH group; R_a, solvent-accessible reduction ratio; $V°$, partial specific volume; ASA_N, solvent-accessible surface area for native protein; *s,* shape (position of branch point in a side chain); *f*, flexibility (number of side-chain dihedral angles).

In turn, the relevance order of the atomic properties according to the GA optimization was: polarizability>van der Waals volume>electronegativity (Figure 3). Hydrophobicity, resembled as polarizability and van der Waals properties, are the most relevant ligand feature ruling the interaction with protease ligands. Furthermore, polarizability also accounts for the deformability of the ligand for interacting with the active site. This result also suggests that the higher variability of the dataset corresponds to the hydrophobic moieties on the ligands structure.

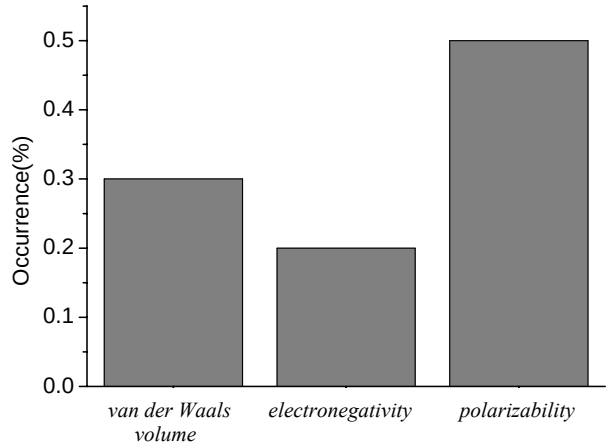

Figure 3. Relative occurrences of the three atomic properties in the optimum 10 input variables: van der Waals volume, electronegativity and polarizability.

CONCLUSIONS

Protease-inhibitor binding stability, characterized by their k_i values has been predicted using SVM optimized by genetic algorithm. Models developed on systematically selected features of proteases, inhibitors and their combination could successfully classify about 80% of protease-inhibitor pairs in terms of their being stable or unstable at a given threshold. This method will be useful in the understanding and applications of protease-inhibitor interactions.

REFERENCES

[1] N. D. Rawlings, F. R. Morton, A. J. Barrett, "MEROPS: the peptidase database," *Nucleic Acids Res*., vol. 34, pp. D270–D272, 2006

[2] C. M. Salisbury; J. A. Ellman, "Rapid Identification of Potent Nonpeptidic Serine Protease Inhibitors," *ChemBioChem*, vol. 7, pp. 1034-1037, 2006.

[3] M.D. Lim,C.S. Craik,"Using specificity to strategically target proteases," *Bioorg. Med. Chem.*, doi:10.1016/j.bmc.2008.03.068 , 2008.

[4] S. Bjelic, M. Nervall,H. Gutierrez-de-Teran, K. Ersmark, A. Hallberg, J. Åqvist,"Computational inhibitor design against malaria plasmepsins," *Cell. Mol. Life Sci*. vol. 64, pp. 2285-2305, 2007.

[5] J.E.S. Wikberg, M. Lapinsh, P. Prusis, "Proteochemometrics: A tool for modelling the molecular interaction space," In: *H. Kubinyi, G. Müller, eds. Chemogenomics in Drug Discovery. A Medicinal Chemistry Perspective.* Weinheim: Wiley-VCH, 2004, pp. 289-309.

[6] L. Fernández, J. Caballero, J.I. Abreu, M. Fernández, "Amino Acid Sequence Autocorrelation Vectors And Bayesian-Regularized Genetic Neural Networks For Modeling Protein Conformational Stability: Gene V Protein Mutants," *Proteins* vol. 67, pp. 834–852, 2007.

[7] Instant JChem Version: 2.1.1, 2007, ChemAxon, web: http://www.chemaxon.com.

[8] H Huang, ZZ Hu, BE Suzek, CH Wu. The PIR integrated protein databases and data retrieval system. *Data Sci. J.* vol. 3 pp. 163-174, 2004.

[9] G . Moreau, P Broto, "Autocorrelation of a topological structure: A new molecular descriptor," *Nouv. J. Chim.* vol. 4, pp. 359-360, 1980.

[10] MATLAB 7.0. program, available from The Mathworks Inc., Natick, MA. *http://www.mathworks.com*.

[11] C . Cortes, V. Vapnik, "Support-Vector Networks*," Mach. Learn.* vol. 20, pp. 273-297, 1995.

[12] C. Chih-Chung, L. Chih-Jen. LIBSVM: a library for support vector machines, Software available at *http://www.csie.ntu.edu.tw/~cjlin/libsvm*, 2001.

[13] H . Fröhlich, O . Chapelle, B. Schölkopf, "Feature Selection for Support Vector Machines by Means of Genetic Algorithms," in *Proc. 15th IEEE Int. Conf. on Tools with AI*. 2003, pp. 142-148.