Reversible auto-encoding of amino-acid residues in reduced space: an application to predicting DNA-binding proteins

Shandar Ahmad

Abstract— **There have been a number of recent studies aiming to predict binding sites and other structural and sequence features of proteins using local amino acid sequence as inputs to a machine learning system. This requires representing amino acids in numerical space, which is typically 20 bits per residue. Number of trainable parameters significantly becomes large with the addition of each neighbor information and hence the application of the technique becomes restricted to the prediction of properties for which large amounts of data is available. Thus, there is a need to find alternatives to this type of sparse encoding. Here a method of auto encoding 20-dimensional sparse representation into lower dimensional space is developed with amino-acids in perspective- although the method is general. It is shown that 20-bit sparse encoding could be reduced to 6-dimensional real space without loss of information and to even lower dimensions with varying degrees of information loss. An application to predicting DNA-binding sites was tested to assess the validity of the proposed method and it was observed that auto-encoded neural network prediction was comparable or superior to sparse encoding system.**

Index Terms—**neural network, auto-encoding, DNA-binding proteins.**

I.Introduction

Biologists and bioinformatics scientists often speak of vast amounts of data resulting from experiments [1]. Yet, when it comes to discovering knowledge patterns from this data, we frequently encounter the complaints of insufficient data. One such paradox is seen in the case of machine learning methods, applied to predicting binding sites or even classical cases of secondary structure and solvent accessibility. There are millions of protein sequences, tens of thousand of solved three-dimensional structures and yet, we have not been able to reach a stage where structural and functional properties of full length proteins or individual residues are well-predicted from the available information. Part of the problem lies in the redundancy of representing this information. For example, in order to predict binding sites from single sequences, local sequence environments of amino acid residues are encoded by *20* binary vectors, expanding a seemingly few residue environment of *N* residues into a huge *20*N* vectors, thereby making the task of discovering a relationship between these inputs and target patterns (binding site or structural property) so much more complicated than what appears to be the real case in nature [2-5]. In this work, this particular example of over-sized amino acid representations has been tackled. Some existing methods such as bio-basis function rely on pattern matching and use residue similarity as a criterion to reduce residue fragments rather than a residue itself [6]. Attempts to look at the dimensionality of amni acid space were made earlier [4,7]. It has been recently shown that a neural network can be employed to reduce higher dimension data to lower dimensions [8]. This work attempts to implement the later strategy because it is more modular and does not depend on a definition of residue-residue similarity. A fully connected unbiased neural network is designed such that it tries to reproduce its 20 input vectors, representing amino acids in sparse encoding. An input to hidden layer section of the neural network serves as the encoder and the hidden to output layer section serves as the decoder. The size of hidden layer is changed to obtain encoding in different dimensions and it is observed that a six-dimensional encoding can reproduce original sparse encoding with a 0.999 correlation coefficient. Resulting six dimensional encoding is applied to the prediction of DNA-binding sites and it is observed that the prediction performance is slightly improved despite a reduced representation.

II.Methods

A.Sparse Coding:

Twenty amino acids are first encoded in a sparse system [2]. This scheme is trivial and consists of 20-bits for each amino acid, such that one of these bits identifying the corresponding amino acid is set to 1 and all other bits are set to zero. There are only 20 possibilities, one each for an amino acid and 20 pattern vectors are created, which forms the entire data set used for training. Although some non-standard amino acids and terminal positions also need to be encoded in real systems, they have been ignored for the current study. It may be noted that the assignment of a given position to an amino acid identity is purely arbitrary and may be interchanged if desired. This will also result in an arbitrary encoding in lower dimensions and hence encoded lower dimensional space can be arbitrarily assigned to any amino acid residue without any loss of information and physical similarities between amino acids have no meanings in terms of coincidental similarities that

.

Shandar Ahmad. (corresponding author), is with National Institute of Biomedical Innovation, Saito Asagi, Ibaraki, Osaka 5670085, Japan (e-mail:shandar@nibio.go.jp).

M. Chetty, S. Ahmad, A. Ngom, S. W. Teng (Eds.): PRIB2008,
Supplementary Conf. Proc., pp.239-241, 2008.
(c) PRIB 2008

may be seen between the representations of two amino acids at a given position of the coded vector.

B.Auto-encoding Neural network

Auto-encoder neural network consists of an input layer of 20 dimensions, an output layer of the same size and a hidden layer of *N* units where *N* is the dimensionality of the reduced space. *N* has been changed from 2 to 6 and no further because 6-dimensions were found sufficient to encode residues with nearly 100\% recovery rate (correlation=0.999). Neural network is optimized using gradient descent algorithm and fine tuning is performed by iteratively optimizing each connection weight. Transfer function between the input and hidden layer is *arctan* and between hidden and output is sigmoidal.

C.Performance of auto-encoder

Auto-encoder performance is measured by calculating the coefficient of correlation between the inputs and neural network output from the trained system. Coefficient of correlation is defined as the squared root of the ratio of the covariance between output and input vectors to the product of their variances.

D.Auto-encoder Prediction of DNA-binding sites

Earlier works [2] reporting prediction of DNA-binding sites from single sequences have used 62 proteins for the purpose (more recent methods have a better performance and use PSSM [3], but since this problem relates to single sequences, only the former is used). Classification capacity of the neural network is measured by estimating the area under the ROC curve, taking average of sensitivity and specificity and has been termed as net prediction in this work.

E.Predictor neural network

Neural network with auto-encoded and sparse representations of amino acid residues are used to input five residue window (two neighbors on each terminal) and trained in a three-fold cross-validation manner by dividing the list of 62 proteins into 20, 20 and 22 members and training one set, using the second to stop training and then using the third set to evaluate the prediction performance. Average prediction performance from three cycles are used to compare performances. The window size, neural network architecture and other features of the neural networks are the same and hence any difference in performance can be attributed to the representation scheme.

III.RESULTS AND DISCUSSION

A.Performance for sparse code recovery

Recovery of sparse 20 bit vectors from the encoded hidden layer representations may be measured in terms of the mean squared errors (MSE) between the input and output values of the encoding neural network. However, since 95% units are zero MSE may not give us the best estimate of utility as an encoder. Coefficient of correlation between the original and decoded values of sparse code are calculated. Coefficient of correlation calculated in this way systematically improved on

increasing dimensions from 2 to six. Specifically these values were 0.071232, 0.596588, 0.695329, 0.851083 and 0.999209 for 2, 3, 4, 5 and 6-dimensional encoders. As observed, the correlation reaches almost 1.0 for six-dimensional codes and were treated as the most suitable for loss-less encoding of residues. Finally observed codes in six dimensions are shown in Table [1].

Table 1. Auto-encoded representations of amino acid residues in six dimensions. Please note that assignments of vectors to amino acids are arbitrary and have no physical meaning, and hence can be interchanged in a given application

Res	Encoding					
A	-0.8150	-1.3692	-0.6694	-0.5711	-0.4128	0.6792
C	1.2255	-1.2678	0.1365	0.9430	-0.9403	-1.1437
D	0.0667	-1.3598	-0.0880	-1.0533	-0.5977	-1.1676
E	0.7301	-1.2223	1.3012	1.1965	0.8793	-1.0832
F	-0.7280	-1.3438	0.0123	1.2536	-0.8933	0.7177
G	-1.2353	-1.2011	0.8486	1.2496	-0.1525	-1.1145
H	-1.2862	-1.1155	-1.0789	1.1405	0.9190	0.4537
I	-0.0739	-1.3481	1.2169	0.1018	-0.7625	-0.2137
K	0.5846	-1.2485	-1.2777	1.2545	-0.4704	0.2620
L	-0.7421	-1.2797	1.1115	0.8551	1.1953	1.0698
M	1.2076	-1.2813	-1.2526	-0.3312	0.5651	-0.8745
N	1.1035	-1.3494	0.2796	0.7013	-0.1026	1.0707
P	0.7540	-1.1920	0.6023	-0.8019	1.2290	-1.2808
Q	-0.2493	-0.9882	-1.0013	1.1071	1.1717	-1.3074
R	-1.0776	-1.2836	-1.0708	0.3800	-0.9817	-1.1639
S	-0.9652	-1.2965	-1.1533	-1.1025	1.1816	-0.9427
T	0.4597	-1.3558	0.5566	-1.0594	1.0608	0.5842
V	-1.2203	-1.3210	1.0374	-0.8318	0.7967	-0.8234
W	0.2345	-1.3015	-1.1936	0.0089	1.2623	1.0817
Y	1.1045	-1.0025	-0.3280	1.2697	1.2590	0.1844

B.Comparison with sparse encoding

Neural networks using auto-encoded and sparse encoding were separately trained to predict DNA-binding sites in the manner as reported in earlier works [2]. A window size of five residue inputs was used and prediction performance of the sensitivity and specificity was monitored. Average of the two scores was found to be 61% for the sparse encoding and 62% for the auto-encoded system. The difference is small, therefore the only claim that can be made is that the new encoding system is at least as efficient as the sparse encoding system, used earlier. Further validity tests with other types of prediction are underway.

IV Conclusion

An auto-encoder can represent amino acid residues in 6 dimensions without information loss. Multiple residue windows using this type of encoding to predict DNA-binding sites perform as efficiently or better than the sparse encoding system. The encoding does not rely on any sequence or residue-residue similarity and is therefore robust and universally applicable.

References

[1] Stein LD., Integrating biological databases. Nat Rev Genet, 4(5) (2003) 337--45.

[2] Ahmad S, Gromiha MM, Sarai A. Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. Bioinformatics. 20(4) (2004) 477--486

[3] Ahmad S, Sarai A. PSSM-based prediction of DNA binding sites in proteins. BMC Bioinformatics. 6 (2005) 33.

[4] Araúzo-Bravo MJ, Ahmad S, Sarai A. Dimensionality of amino acid space and solvent accessibility prediction with neural networks. Comput Biol Chem. 30(2) (2006) 160--168

[5] Rost B, Sander C. Progress of 1D protein structure prediction at last. Proteins. 23(3) (1995) 295--300.

[6] Yang ZR, Johnson FC. Prediction of T-cell epitopes using biosupport vector machines. J Chem Inf Model. 45(5) (2005) 1424--1428.

[7] Coghlan A, Mac Donail KDA, Buttimore NH, Representations of amino acids as five-bit or three-bit patterns for filtering protein databases. Bioinformatics 17(8) (2001) 676--685.

[8] Hinton GE and Salakhutdinov RR Reducing the Dimensionality of Data with Neural Networks Science 313 (5786) (2006) 504--507