International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970)
Volume-3 Number-1 Issue-8 March-2013

An Unsupervised approach for augmentation of Protein Structure

Geetika S. Pandey[1], R.C Jain[2]
Assistant Professor, Computer Science and Engineering, Samrat Ashok Technological Institute, Vidisha (M.P)[1]
Director, Samrat Ashok Technological Institute, Vidisha (M.P)[2]

Abstract

Among several possible effects of mutations, protein stability is the most critical factor related to protein function. The destabilization of a protein is due to misfolding or the glassy state of a protein which occurs due to the wrong amino acid substitutions. This destabilization is cause of many prion diseases. A check on stability of the protein could be an efficient solution for this problem, in this paper such an unsupervised approach is proposed.

Keywords

Protein folding, destabilization, unsupervised, ART.

I. Introduction

The protein structure is the ultimate factor responsible for the proper functionality. Due to certain physical changes the structure of the protein changes. The structures can be stable or unstable. There are many properties of amino acids which could help in predicting the stability. Many machine learning approaches are being developed, which can be specified as : given the amino acid sequence of a protein and a single amino acid substitution, the task is to predict whether the substitution may alter protein stability, these classifiers can be constructed for predicting either the free energy change of protein stability upon mutations or the direction of the change. Teng et. al.[1] used the SVM classifiers, which is based on supervised learning with associated learning algorithms. Various biological features are also analysed by Teng et. al. for sequence-based prediction of protein stability changes upon single amino acid substitutions. Shakir et. al. used Fuzzy ARTMAP as the classifier, which uses the features which are selected using genetic algorithm[4]. Fuzzy ARTMAP again is a supervised learning method with fuzzy operators. Here the free energy change ΔG, as mentioned by Zhang et.al. –
$\Delta G(folding)=G(folded)-G(unfolded)$ [5],
is taken as the key feature. This feature is discussed in detail later in this paper.

II. Background

A. Experimental Dataset

The protein community has over the years established many publicly available protein-information related databases. Some of these include the well known Protein Data Bank (PDB) and UniProt among others, with each of these databases serving a particular segment of the protein analysis community. The experimental dataset could be retrieved from the ProTherm website (Thermodynamic Database for Proteins and Mutants: http://gibk26.bio.kyutech.ac.jp/jouhou/Protherm/protherm_search.html) [5].Shakir et. al. used Structural Classification of Proteins (SCOP) version 1.69 to retrieve protein sequences[4]. One can use these sequences by directly retrieving it through these databases, or can download the sequences in various file formats like FASTA or XML, as per the requirement. Many alignment based techniques have been developed, most notably are the Basic Local Alignment and Search Tool – BLAST, FASTA and position specific weight matrices [4]. These databases also provide various feature values as well as the secondary structures of the sequences. While concentrating more on augmentation and optimization of the folding, the secondary structures could also be retrieved from these databases , and further (as per our work) calculations could be made for free energy change. Otherwise, while including the prognostication and prediction of the mutations, the structures could be generated through the sequence retrieved by the databases.

B. Features

As mentioned by Teng et.al.[1] about twenty biological features of amino acid could be used for research purposes. These features were obtained from Protscale http://expasy.org/tools/protscale.html) [9] and AAindex (http://www.genome.jp/aaindex/) [10]. The biological features fall into the following four classes:

Biochemical features – includes M, molecular weight, this is related to volume of space that a residue occupies in protein structure. K, side chain pka value, which is related to the ionization state of a

residue and thus plays a key role in pH dependent protein stability. H, hydrophobicity index, which is important for amino acid side chain packing and protein folding. The hydrophobic interactions make non-polar side chains to pack together inside proteins and disruption of these interactions may cause protein destabilization. P, polarity, which is the dipole-dipole intermolecular interactions between the positively and negatively charged residues. Co, overall amino acid composition , which is related to the evolution and stability of small proteins.

Structural features- this includes A, alpha-helix. B, beta-sheet. C, coil. Aa, average area buried on transfer from standard state to folded protein. Bu, bulkiness, the ratio of the side chain volume to the length of the amino acid.

Empirical Features- this includes, S1, protein stability scale based on atom atom potential of mean force based on Distance Scaled Finite Ideal-gas Reference (DFIRE). S2, relative protein stability scale derived from mutation experiments. S3, side-chain contribution to protein stability based on data from protein denaturation experiments.

Other biological features- F, average flexibility index. Mc, mobility of an amino acid on chromatography paper. No, number of codons for an amino acid. R, refractivity, protein density and folding characteristics. Rf, recognition factor, average of stabilization energy for an amino acid. Rm, relative mutability of an amino acid. Relative mutability indicates the probability that a given amino acid can be changed to others during evolution. Tt, transmembrane tendency scale. F, average flexibility index of an amino acid derived from structures of globular proteins.

Name	3-Letter Symbol	1-Letter Symbol	Molecular weight	Molecular Formula	Residue Formula	Residue Weight (-H_2O)	pK_a [1]	pK_b [2]	pK_x [3]	pI [4]
Alanine	Ala	A	89.10	$C_3H_7NO_2$	C_3H_5NO	71.08	2.34	9.69	—	6.00
Arginine	Arg	R	174.20	$C_6H_{14}N_4O_2$	$C_6H_{12}N_4O$	156.19	2.17	9.04	12.48	10.76
Asparagine	Asn	N	132.12	$C_4H_8N_2O_3$	$C_4H_6N_2O_2$	114.11	2.02	8.80	—	5.41
Aspartic acid	Asp	D	133.11	$C_4H_7NO_4$	$C_4H_5NO_3$	115.09	1.88	9.60	3.65	2.77
Cysteine	Cys	C	121.16	$C_3H_7NO_2S$	C_3H_5NOS	103.15	1.96	10.28	8.18	5.07
Glutamic acid	Glu	E	147.13	$C_5H_9NO_4$	$C_5H_7NO_3$	129.12	2.19	9.67	4.25	3.22
Glutamine	Gln	Q	146.15	$C_5H_{10}N_2O_3$	$C_5H_8N_2O_2$	128.13	2.17	9.13	—	5.65
Glycine	Gly	G	75.07	$C_2H_5NO_2$	C_2H_3NO	57.05	2.34	9.60	—	5.97
Histidine	His	H	155.16	$C_6H_9N_3O_2$	$C_6H_7N_3O$	137.14	1.82	9.17	6.00	7.59
Hydroxyproline	Hyp	O	131.13	$C_5H_9NO_3$	$C_5H_7NO_2$	113.11	1.82	9.65	—	—
Isoleucine	Ile	I	131.18	$C_6H_{13}NO_2$	$C_6H_{11}NO$	113.16	2.36	9.60	—	6.02
Leucine	Leu	L	131.18	$C_6H_{13}NO_2$	$C_6H_{11}NO$	113.16	2.36	9.60	—	5.98
Lysine	Lys	K	146.19	$C_6H_{14}N_2O_2$	$C_6H_{12}N_2O$	128.18	2.18	8.95	10.53	9.74
Methionine	Met	M	149.21	$C_5H_{11}NO_2S$	C_5H_9NOS	131.20	2.28	9.21	—	5.74
Phenylalanine	Phe	F	165.19	$C_9H_{11}NO_2$	C_9H_9NO	147.18	1.83	9.13	—	5.48
Proline	Pro	P	115.13	$C_5H_9NO_2$	C_5H_7NO	97.12	1.99	10.60	—	6.30
Pyroglutamatic	Glp	U	139.11	$C_5H_7NO_3$	$C_5H_5NO_2$	121.09	—	—	—	5.68
Serine	Ser	S	105.09	$C_3H_7NO_3$	$C_3H_5NO_2$	87.08	2.21	9.15	—	5.68
Threonine	Thr	T	119.12	$C_4H_9NO_3$	$C_4H_7NO_2$	101.11	2.09	9.10	—	5.60
Tryptophan	Trp	W	204.23	$C_{11}H_{12}N_2O_2$	$C_{11}H_{10}N_2O$	186.22	2.83	9.39	—	5.89
Tyrosine	Tyr	Y	181.19	$C_9H_{11}NO_3$	$C_9H_9NO_2$	163.18	2.20	9.11	10.07	5.66
Valine	Val	V	117.15	$C_5H_{11}NO_2$	C_5H_9NO	99.13	2.32	9.62	—	5.96

[1] pK_a is the negative of the logarithm of the dissociation constant for the -COOH group
[2] pK_b is the negative of the logarithm of the dissociation constant for the $-NH_3^+$ group
[3] pK_x is the negative of the logarithm of the dissociation constant for any other group in the molecule
[4] pI is the pH at the isoelectric point
References: D. R. Lide, *Handbook of Chemistry and Physics, 72nd Edition*, CRC Press, Boca Raton, FL, 1991.

Fig. 1: features of amino acids

C. Protein folding

Interactions giving rise to folding is a consequence of intermolecular forces , including

i. Pure ionic interactions
ii. Dipole interactions
iii. Hydrogen bonds
iv. Vander waals forces
v. Hydrophobic interactions

All the above except the hydrophobic interactions are electrostatic in origin and contribute to the enthalpy of protein folding. The Hydrophobic effect is an indirect effect resulting from a peculiarity of water structure. Water molecules exchange hydrogen bonds with neighbors at a rate of about 10^{11} s^{-1}. At the interface between water and a non- H-bonding group such as CH_3, water molecules have fewer opportunities for H-bond exchange, leading to longer than usual lifetime of H-bonds, an ice-like state at the interface, and consequent decrease in entropy.

$$\Delta G = \Delta H - T\Delta S$$

Negative enthalpy change and positive entropy change give negative ,i.e. stabilizing, contributions to the free energy of protein folding, i.e. the lower the ΔG ,the more stable the protein structure is. Any situation that minimizes the area of contact between H_2O and non-polar, i.e. , hydrocarbon, regions of the protein results in an increase in entropy.[2]

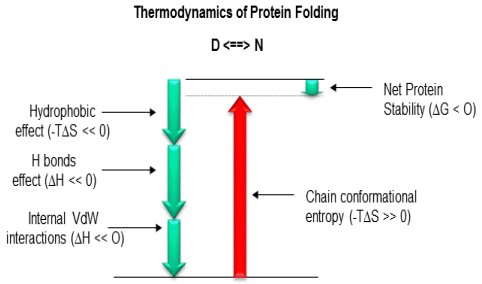

Fig. 2: thermodynamics of protein folding

D. Learning

Machine learning has been used in various phases in proteomics. For this approach we prefer unsupervised learning because of the following reasons [6]:

In practice , models for supervised learning often leave the probability for inputs undefined , whereas in unsupervised learning , all the observations are assumed to be caused by latent variables.

With the unsupervised learning it is possible to learn larger and more complex models than with supervised learning. This is because in supervised learning one is trying to find the connection between two sets of observations, whereas the difficulty of the learning task increases exponentially in the number of steps between the two sets and that is why supervised learning cannot, in practice, learn models with deep hierarchies.

In unsupervised learning, the learning can proceed hierarchically from observations into ever more abstract levels of representation. Each additional hierarchy needs to learn only one step and therefore the learning time increases (approximately) linearly in the number of levels in the model hierarchy. For this approach we prefer ART1 as the basic model, which classifies the protein structure, as stable and unstable.

ART1: Real world problems, like the one we are discussing, face situations where data is continuously changing. In such situation every intelligent learning system faces plasticity-stability dilemma, i.e ,a learning system should be plastic, or adaptive in reacting to changing environments, and should be stable to preserve knowledge acquired previously. ART learning system are competitive learning networks, and its architecture can self organize in real time producing stable recognition while getting input patterns beyond those originally stored, thus a solution to the dilemma.[3]

The Adaptive Resonance Theory (ART) networks are self organizing competitive neural network. There is a wide variety of these networks which follow both supervised and unsupervised algorithms. Here we prefer ART1 i.e the basic unsupervised learning network. The basic architecture of ART1 is as follows:

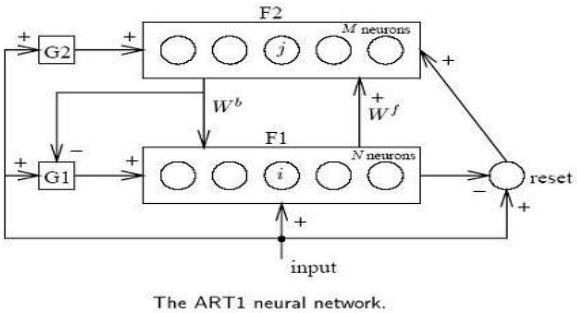

Fig. 3: ART1 model

III. Experimental Procedure

Now we discuss the proposed approach, in this we use the ART model for clustering. The output layer consists of two clusters C1: if Y= 0 (stable) and C2: if Y=1 (unstable).
In case of the vigilance parameter we use the G(unfolded) value, i.e the free energy value of the unfolded amino acid sequence.
The inputs will be the a set of free energy values of all the protein structures of the given amino acid sequence, i.e G(folded) = {G_1 , G_2 , G_3 …….} .

A. Algorithm

1. Retrieve the protein sequence from an online database.
2. Predict the various structures of the sequence. Calculate the free energy change of all the structures as
$$\Delta G = \Delta H - T\Delta S$$
3. Taking these values as the binary input for the ART model as I(B)={B_1 , B_2 , B_3….} and B_1= {x1, x2, x3….xn}. Initially, control gains G1 =0 & G2=0 , when input vector I is empty and the nodes at layer F1 & F2 are set to 0.
4. Weight matrix W(t)=w_{ir}(t), is the bottom up weight matrix of size n*m where i=1 , n and r=1,m. Initially, w_{ir}=(1/n+1).
   Weight matrix $V(t) = v_{ri}(t)$ is the top down weight matrix of size m*n, which is initially a unit vector.
5. Set vigilance parameter 'ρ' as the free energy change of the unfolded state and learning parameter $\alpha \in (0,1)$.
6. Now with the first input I≠0 , therefore G1=1 and thus activates all nodes of F1.

Again since I≠0 , and O1=0 i.e no output from F2 , therefore G2=1 and thus activates all the nodes in F2, means recognition in F2 is allowed.

7. Compute input for each node in F2, using $y_r = \sum_{i=1}^{n} Ii \times w_{ir}$.
8. Select the winning node $= \sum_{r=1}^{p} \max(y_r)$, where p is the no. of nodes on layer F2, which in this case is 2.
9. Perform the vigilance test $\frac{<vk,xk>}{\|xk\|} >$ ρ here in this case if the u=$\frac{<vk,xk>}{\|xk\|}$ is lower than the vigilance factor which is the free energy of the unfolded protein then it means ΔG is lower i.e the structure is stable .as per the

following equation , the lower the ΔG the more stable is the structure

$$\Delta G = G(folded) - G(unfolded)$$

10. Now check for the similarity between xk and the input, if 1 then update the weights. Weight vector

$vki(new) = vki(t) \times xi$ And

$$wki(new) = \frac{vki(new)}{0.5 + \|vki(new)\|}$$

11. Repeat steps 3 to 10 for the rest of the input instance.

IV. Expected Outcome

By following the above mentioned algorithm the input structures would be classified into two clusters C1 (stable) and C2(unstable). Thus the ultimate output for the particular primary amino acid sequence would be all the structures in the cluster C1, which all are stable, thus a stable secondary structure is assured. The efficiency of machine learning and other soft computing approaches in this field has already been proven. Till date, as per the research done none of these approaches follow unsupervised learning coupled with the same features. So along with the previously mentioned benefits of unsupervised learning, the proposed approach is under the observation, and is expected to be more efficient than the existing ones.

V. Discussion and Future work

By this approach a generalized solution could be generated for problem of destabilization. As per the research work done for this approach the most promising features are being used but as mentioned in this paper there are various other features of amino acid, which could also be considered. In this work we concentrate more on optimization and augmentation of the protein structure, further the prognostication could also be included, by making the prediction more generalized, thus the whole protein folding problem could be solved more efficiently.

References

[1] Shaolei Teng, Anand K. Srivastava, and Liangjiang Wang, "Biological Features for Sequence-Based Prediction of Protein Stability Changes upon Amino Acid Substitutions", 2009 International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing.

[2] 76-456/731 Biophysical Methods – Protein structure component , Lecture 2: Protein interations leading to folding http://www.chembio.uoguelph.ca/educmat/phy456/456lec02.htm.

[3] Adaptive Resonance Theory : Soft Computing Course Lecture 25-28,notes,slides, www.myreaders.info/html/soft_computing.html.

[4] Shakir Mohammed, David Rubin and Tshilidzi Marwala, " Multi-class Protein Sequence Classification Using Fuzzy ARTMAP", 2006 IEEE International Conference on Systems, Man and Cybernetics.

[5] Zhe Zhang, Lin Wang, Daquan Gao, Jie Zhang, Maxim Zhenirovskyy and Emil Alexov, "Predicting folding free energy changes upon single point mutations".Bioinformatics Advance Access published January ,2012.

[6] Supervised vs. unsupervised learning, http://users.ics.aalto.fi/harri/thesis/valpola_thesis/node34.html.

[7] B. Gassend, C. W. O'Donnell, W. Thies, A. Lee, M. van Dijk and S. Devadas, " Secondary Stucture Prediction of All-Helical Proteins Using Hidden Markov Support Vector Machines".

[8] Zikrija Avdagic, Elvir Purisevic, Emir Biza, Zlatan Coralic, "Neural Network Algorithm for Prediction of Secondary Protein Structure".

[9] H.C. Gasteiger E., Gattiker A., Duvaud S., Wilkins M.R., Appel R.D. and Bairoch A., The Proteomics Protocols Handbook, Humana Press,2005.

[10] S. Kawashima and M. Kanehisa, "AAindex: amino acid index database," Nucleic Acids Res, vol. 28, Jan 1. 2000, pp. 374.

[11] "Neural Network, Fuzzy Logic, and Genetic Algorithms – Synthesis and Applications", by S. Rajasekaran and G.A. Vijayalaksmi Pai, (2005), Prentice Hall, Chapter 5, page 117-154.

[12] "Elements of Artificial Neural Networks", by Kishan Mehrotra, Chilukuri K. Mohan and Sanjay Ranka, (1996), MIT Press, Chapter 5 , page 157-197.

[13] "Pattern Recognition Using Neural and Functional Networks", by Vasantha Kalyani David, Sundaramoorthy Rajasekaran, (2008), Springer, Chapter 4, page 27-49.

[14] Rost B, Sander C. "Prediction of protein secondary structure at better than 70% accuracy". J Mol. Biology, 1993;232: 584-99.

[15] Cuff JA. Barton G.J. "Evaluation and Improvement of Multiple Sequence Methods for Protein Secondary Structure Prediction, PROTEINS: Structure, Function, and Genetics, 1999; 34: 508-19, Available from: httop://binf.gmu.edu/vaisman/csi731/pr99-cuff.pdf.

R. C. Jain, M.Sc., M. Tech., Ph. D., is a Director of S.A.T.I. (Engg. College) Vidisha (M. P.) India. He has 37 years of teaching experience. He is actively involved in Research with area of interest as Soft Computing, Fuzzy Systems, DIP, Mobile Computing, Data Mining and Adhoc Networks. He has published more than 125 research papers, produced 7 Ph. Ds. and 10 Ph. Ds are under progress.

Geetika S. Pandey obtained her B.E degree in Computer Science and Engineering from University Institute of Technology, B.U, Bhopal in 2006. She obtained Mtech degree in Computer Science from Banasthali Vidyapith, Rajasthan in 2008. She worked as Assistant Professor in Computer Science and Engineering Department in Samrat Ashok Technological Institute, Vidisha (M.P). She is currently pursuing Ph.D. under the supervision of Dr. R.C Jain, Director, SATI, Vidisha. Her research is centered on efficient prognostication and augmentation of protein structure using soft computing techniques.