Expectation – Maximization algorithm for protein – ligand complex of HFE gene

O.S. Deepa[1], and Ani. R[2]

[1]Department of Mathematics, Amrita School of Engineering, Coimbatore

[2]Department of Computer Science and applications, Amrita School of Engineering, Amritapuri

Amrita Vishwa Vidyapeetham, Amrita University, India.

***Corresponding author: E-Mail: os_deepa@cb.amrita.edu**

ABSTRACT

In the field of pharmaceutical sciences and biomedicine, the issue of protein stabilization presumes meticulous importance. It plays a significant role in purification, formulation, and storage. Suitably folded proteins are usually stable during expression and purification. The interaction between ligands and proteins generally produces changes in protein thermal stability with changes in the midpoint denaturation temperature, enthalpy of unfolding, and heat capacity. The stability of eleven mutations of the proteins corresponding to HFE gene are identified using Random forest and Support vector machine. Various parameters like Half-life period, aliphatic index and GRAVY are computed using online webservers. Based on the machine learning techniques and the computed parameters, the ligands for HFE proteins are obtained. The contact surface area between ligand atom and protein atoms are also identified. The expectation – maximization algorithm was done on the contact surface area to test whether there exist any change in the destabilizing contact between the ligand atom and protein atoms.

KEY WORDS: Hemochromatosis, Random forest, Support vector machine, expectation–maximization algorithm.

1. INTRODUCTION

The HFE gene gets interacted with other proteins on the surface of the cell to detect the amount of iron in the body and is most commonly found in the liver and intestinal cells. The HFE protein also helps in production of the hepcidin hormone and determines the percentage of iron that should be absorbed from the diet and the percentage of iron that should be released from storage sites in the body. Research studies shows that the HFE gene lays the foundation for the hereditary disease hemochromatosis of type with almost 20 mutations. Each of these mutations changes one of the amino acid in the protein building blocks of HFE protein. One of the mutation is due to the replacement of the amino acid cysteine with the amino acid tyrosine at position 282 in the protein chain as C282Y. Similarly other mutation would be due to the replacement of the amino acid Glycine with the amino acid Histidine at position 105. The Cys282Tyr mutation prevents the changed HFE protein from reaching the cell surface, so it cannot interact with hepcidin receptors. Hence, iron regulation is interrupted, and too much iron is absorbed from the diet. This increase in the absorption of iron leads to the iron overload characteristic of type 1 hemochromatosis. The Cys282Tyr mutation is considered as a common cause for hereditary hemochromatosis which increases the iron overload in X-linked side roblastic anemia which is inherited along with a mutation in the ALAS2 gene. The amalgamation of HFE and ALAS2 mutations gave rise to symptoms of X-linked side roblastic anemia by increasing the absorption of iron, leading to greater iron overload. Eleven mutations are given in Table.1.

Table.1. Amino acid change for the HFE protein

Entry name - HFE	Amino acid change	Condon change
1A6Z	Gly-Asp	G-D
1A6Z	Ser-Cys	S-C
1A6Z	Giy-Arg	G-R
1A6Z	Gln-His	Q-H
1A6Z	Ala-Val	A-V
1A6Z	Arg-Gly	R-G
1A6Z	Cys-Tyr	C-Y
1A6Z	Gln-Pro	Q-P
1DE4	Ary-Cys	R-C
1DE4	IIE-Thr	I-T
1DE4	Val-Ala	V-A

For the 11 mutations in the HFE protein the percentage of each amino acid is noted. It is found out that percentage of Leucine in the mutations is found to be more followed by Gln and Glu. The amino acid Cystine has less role in the mutations of HFE.

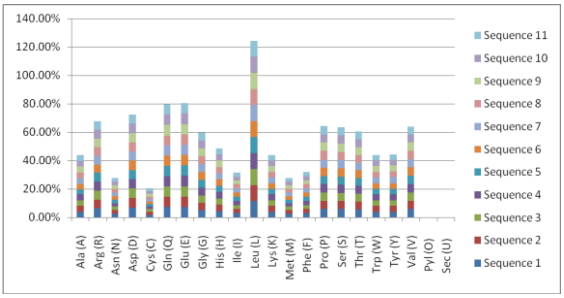

Figure.1. Percentage of amino acid for each mutation

The aliphatic index value was found to be greater than 30. This predicts a high thermodynamic stability of these protein molecules. In all mutations the instability index is found to be greater than 40 predicting unstable nature of molecules. As the half-life period of the protein increases, the possibility of accumulation of the protein in the corresponding sub-cellular location increases.

Table.2. Half-life Instability index, aliphatic index and GRAVY

	Half -life period	Instability index	Aliphatic index	Grand average of hydropathicity
G-D	1	43.99	76.13	-.649
S-C	1	44.19	76.18	-.637
G-R	1	47.12	76.18	-.653
Q-H	1	45.65	76.18	-.637
A-V	1	43.99	76.87	-.629
R-G	1	43.29	76.18	-.623
C-Y	1	44.27	76.18	-.652
Q-P	1	44.20	76.18	-.653
R-C	1	42.40	76.18	-.612
I-T	1	44.12	75.04	-.657
V-A	1	44.30	75.49	-.647

For all the 11 mutations the half-life period, aliphatic index, and grand average of hydropathicity and instability index, is obtained by PROTPRAM.

Random forest method and support vector machine are most commonly used methods for classification. Random forest ensemble classifier to predict the coronary heart disease using risk factors was studied by Ani (2015). Based on the mutations and to understand more on the stability of the mutated proteins, an online web server Auto-Mute is used. The stability of the proteins (Table.2) is obtained by various machine learning techniques like Random forest, Support Vector Machine, Support Vector Regression and Tree Regression. It is found that the stability is increased for mutations A-V using Random forest and mutations S-C in Support vector machine. Only the location of G-D and S-C of the muted protein is found on the surface and other mutations are found buried. Number of edge contact with the surface positions is also obtained. The secondary structure of the muted protein is obtained. The change in the amino acid may cause a disease or could remain neutral. It is found that S-C, Q-H, A-V, C-Y, Q-P and I-T are related with disease. Among these change only I-T in 1DE4 is found to be affected by the mutant activity. Also support vector regression and tree regression give the predicted value of ddG in Table.3.

Table.3. Stability, location, number of edge contact, secondary structure, disease affected and mutant activity of proteins

	Random forest Stability	SVM Stability	Location	Number of edge contact	Secondary structure	Disease affected	Mutant activity of proteins
G-D	Decreased	Decreased	Surface	10	C	Neutral	Unaffected
S-C	Decreased	Increased	Surface	6	C	Disease	Unaffected
Q-H	Decreased	Decreased	Buried	0	H	Disease	Unaffected
A-V	Increased	Decreased	Buried	0	C	Disease	Unaffected
R-G	Decreased	Decreased	Buried	0	H	Neutral	Unaffected
C-Y	Decreased	Decreased	Buried	0	S	Disease	Unaffected
Q-P	Decreased	Decreased	Buried	0	S	Disease	Unaffected
R-C	Decreased	Decreased	Buried	0	C	Neutral	Affected
I-T	Decreased	Decreased	Buried	0	C	Disease	Affected
V-A	Decreased	Decreased	Buried	0	H	Neutral	Unaffected

Table.4. Predicted value of ddG by SVM Regression and tree Regression

Codon change	Predicated value of ddG by SVM Regression	Predicated value of ddG by Tree Regression	Location
G-D	-1.21	-1.35	Surface
S-C	-0.30	-0.64	Surface
G-R	-0.25	-0.61	Buried
Q-H	0.00	-0.61	Buried
A-V	-1.15	-0.14	Buried
R-G	-1.15	-0.53	Buried
C-Y	-1.13	-2.58	Buried
Q-P	-1.15	-0.35	Buried
R-C	-0.86	-1.29	Buried
I-T	-1.75	-3.94	Buried
V-A	-1.24	-2.61	Under surface

Sobolev (1999), had studied on the interatomic contact in proteins. Victor and Marta (2015), had studied on the structural and functional stabilization of protein entities. Computational prediction of ligand entry is possible for buried active site of proteins. Hence the number of ligand is extracted from the LPU/CSU online server for both the PDB entry 1A6Z and 1DE4. It is found that 1A6Z has no ligand and there are 7 ligands in 1DE4 PDB entry. The names of the ligand are N-Acetyl-D-Glucosamine (with chain C, F and I of different residue), Calcium ion (with chain C, F and I of different residue) and Glycerol (with chain C). The contact surface area between ligand and protein atoms are obtained for all the 7 ligands and it is found that only N-Acetyl-D-Glucosamine has destabilizing contacts with residue number 900, 901 and 902 of chains C, F and I. Table.5, gives the contact surface area $(A^2)$of N-Acetyl-D-Glucosamine with residue number 900 of chain C. Similarly contact surface area $(A^2)$of N-Acetyl-D-Glucosamine with residue number 901 and 902 of chain F and I are also obtained. Only residues connecting ligands by side chain are alone considered.

Table.5. Surface contact between Ligand atom and Protein atom (NAG 900C)

Ligand atom-Name	Protein atom		Surface Contacts		
	Residue	Name	I	II	III
N2	ASN	CG	0.6	1.2	0.2*
C1	ASN	ND2	41.1	40.6	39.9
C1	ASN	CG	6.1	9.0	8.1
C1	PHE	CD1	5.2	3.6	2.7
C2	ASN	ND2	1.1	0.4	0.7
C3	PHE	CE1	8.1	9.9	9.6
C3	PHE	CZ	3.8	2.5	2.0
C5	PHE	CG	11.7	12.6	12.3
C6	GLU	0E1	29.4	21.3	20.4
C6	GLU	CD	8.7	8.3	4.5
C7	ASN	ND2	6.1	5.2	6.1
C7	ASN	CG	0.4	0.9	0.4
C8	ASN	CG	0.4	2.9	0.4
C8	ASN	CB	5.2	2.2	2.9
C8	ASN	ND2	0	0.2*	0.2*

*- indicates destabilizing contact

The EM algorithm: The algorithm is as follows:

a) Start with guess for values of the model parameters
b) E-step: For each data point that has destabilized values (considered as missing values), the model equation is used to solve for the distribution of the destabilized values given the current assumption of the model parameters and given the observed data. Based on distribution for each destabilized values, we can calculate the expectation of the likelihood function with respect to the unobserved variables. If the estimate for the model parameter was correct, this expected likelihood will be the actual likelihood of our observed data; if the parameters were not correct, it will just be a lower bound.
c) M-step: The expected likelihood function with no unobserved variables in it is obtained. Now maximize the function as in the fully observed case, to get a new estimate of the model parameters.

d) Repeat until convergence.

Table.6. Iterated values of destabilized contacts

	Original values	I - Iteration	II – Iteration	III Iteration	IV- Iteration	V - Iteration	VI- Iteration	VII Iteration
N2-ASN-CG	I - 0.2	1.167307	0.471695	0.399699	0.392396	0.392396	0.391662	0.391587
C8-ASN-ND2	I - 0.2	0.932652	0.049670	0.00123	0	0	0	0
C8-ASN-ND2	II - 0.2	1.237152	0.676841	0.661033	0.660588	0.660588	0.660575	0.660575

Table.7. The variance covariance matrix for the I iteration

N2-ASN-CG	58.8086	17.18889	38.6399
C8-ASN-ND2	17.18889	38.6031	33.2809
C8-ASN-ND2	38.6399	33.2809	47.0375

Table.8. The variance covariance matrix for the VII iteration

N2-ASN-CG	49.04826	17.4134	39.1368
C8-ASN-ND2	17.41348	35.4243	33.2808
C8-ASN-ND2	39.1368	33.2808	47.0375

2. CONCLUSION

The expected maximization algorithm is implemented on the surface contact area between ligand atoms of NAG 900C, 901F and 902I and protein atoms in 1DE4 entry. The selected values of the surface contact area of NAG with corresponding ligand atom and residue atom are entered in a matrix. The destabilized values are assumed as missing data because the intention is to test whether there exists a value which is different from current destabilized value and would remain constant after some iteration. The values converged approximately in the seventh iteration. It is found that the surface contact area of ligand (atom N_2) and the residue ASN (atom CG) of NAG901F has a minor change from 0.2 to 0.39 and atoms with a destabilizing contacts for ligand atom C8 of ASN (atom ND_2) of NAG 902I has a drastic change from 0.2 to 0.66. Computational changes or experimental changes can be done on these atoms to test existence of stabilized contact by rotating the side chains or balancing the medium pH or by making the changes in the location of the expression etc.

REFERENCES

Ani R, Aneesh Augustine, Akhil N.C, Deepa O.S, Random forest ensemble classifier to predict the coronary heart disease using risk factors, proceeding of the International conference on soft computing systems, Springer, 397, 2016.

Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins M.R, Appel R.D, Bairoch A, Protein Identification and Analysis Tools on the ExPASy Server (In) John M. Walker (ed), The Proteomics Protocols Handbook, Humana Press, 2005, 571-607.

Maria Soledad Celej, Guillermo G. Montich and Gerardo D Fidelio, Protein stability induced by ligand binding correlates with changes in protein flexibility, Protein Sci. Jul., 12 (7), 2003, 1496-1506.

Masso M, & Vaisman I.I, AUTO-MUTE: web-based tools for predicting stability changes in proteins due to single amino acid replacements, Protein Eng. Des. Sel., 23, 2010, 683-687.

Masso M, & Vaisman I.I, Knowledge-based computational mutagenesis for predicting the disease potential of human non-synonymous single nucleotide polymorphisms, J. Theor. Biol, 266, 2010, 560-568.

Masso M, & Vaisman I.I, Structure based prediction of protein activity changes: assessing the impact of single residue replacements, Proc. 3rd IEEE EMBC, 2011, 3221-3224.

Masso M, & Vaisman I.I, Structure-based machine learning models for computational mutagenesis, in Protein Structure Methods and Algorithms (eds: H. Rangwala and G. Karypis), Wiley Book Series on Bioinformatics, 2010.

Sobolev V, Sorokine A, Prilusky J, Abola E.E, Edelman M, Automated analysis of interatomic contacts in proteins, Bioinformatics, Pub Med, 15 (4), 1999, 327-332.

Victor M. Balcao, and Marta M.D.C. Vila, Structural and functional stabilization of protein entities: state of the art; Advanced drug delivary Reviews, 93, 2015, 25-41.