3D protein structure prediction of influenza A virus based on optimization genetic algorithm

Jie Gao* **Pei-Xuan Jin and Hong-xing Xu**
School of Science, Jiangnan University, Wuxi, China

Abstract: The 3D structure of close polymer is constituted by the interaction of close contact couples among amino acid residues. In this paper, 3D protein structure of influenza A virus was predicted. Twenty kinds of amino acid residues were divided into four categories according to the number of close contact couples. The stable structure with minimum energy was obtained by using optimization genetic algorithm. The HNXP 3D lattice model was established to predict the 3D protein structure. It can be concluded that the two kinds of structures are significantly similar by computing the similarity.

Keywords: influenza A virus; HNXP 3D lattice model; 3D protein structure prediction; close contact couples; optimization genetic algorithm (OGA)

INTRODUCTION

The protein structure of influenza virus has strong variability, and its spatial structure can be regarded as a fairly complex close polymer structure (Chan, 1989). The spatial structure and change of structural folding should be considered to analyze the morphologicalvariation. At present, it is a very complicated process to establish 3D structure based on primary protein structure, so the model of the protein structure needs to be simplified. The *HP* lattice model is the most common simplified model of protein structure (Lau, 1989), 20 kinds of amino acid residues are separated into two groups, e.g., hydrophilic amino acids (polar-*P*) and hydrophobic amino acids (hydrophobic-*H*). Optimization genetic algorithm was applied to get a self-avoiding structure with lowest energy in the process of building structure model, treated as a spatial structure in natural state. However, the amino acid residues are only divided into two categories, which seem to be too simple in the practical application. In recent years, many scholars focused on the classification problem of amino acid residues. Miyazawa and Jernigan adopted the statistical method to calculate the interaction energy among 20 kinds of amino acid residues, which were classified into three groups (Miyazawa, 1985). Vincent A *et al* further researched the principle of spatial structure folding on the basis of molecular dynamics method, and comprehensively analyzed the factors that affect the three dimensional spatial structure in 2010. In 2011, Ivan Dotu *et al* constructed an elaborate lattice model by employing Large Neighborhood Search (*LNS*). Md. Kamrul Islam *et al* made use of memetic searching algorithm to optimize 3DHP lattice model of protein structure. The optimized HP lattice model spatial structure was more similar to the actual situation.

**Corresponding author:* e-mail: ezhun6669@sina.com

The spatial structure of protein was treated as a 3D spatial structure of close polymer, constituted by the interaction of close contact couples among amino acid residues (Gromiha, 2001 and Gromiha, 2001). Twenty kinds of amino acid residues were divided into four categories according to the number of close contact couples. Ten energy values among the four types of amino acid residues were calculated. The HNXP 3D lattice model for protein structure prediction was built. The comparison between predicted protein structures and real structures provided by the PDB database was performed by computing the similarity. The results show that the two kinds of structures are significantly similar.

METHODS

Interaction energy computation

In the detailed HP lattice model, 20 kinds of amino acids were divided into four categories, e.g., nonpolar amino acid (H), negatively charged polar amino acids (N), uncharged polar amino acids (X) and positively charged polar amino acids (P). H={A, V, L, I , P , F , W , M}, N={D , E} , X={G , S , T , C , Y , N , Q} , P={K , R , H}(Yu, 2004). The spatial protein structure was formed by the interaction of the energy among the amino acid residues. Close contact couples among amino acid residues were defined to describe the energy relationship among amino acid residues.

In the spatial protein structure, if the distance between the center carbon atom Cα of the amino acid residue i and j is less than a fixed value □, amino acid residue i and amino acid residue j are recorded as a close contact couples, usually □=6.5Å. The interaction energy value of close contact couples between amino acid residue i and amino acid residue j is defined (Zhang, 2000 and Jiang, 2002),

$$e_{ij}=-\ln\frac{N_{ij}^2 C_{ii} C_{ij}}{N_{ii} N_{jj} C_{ij}^2}.\quad i,j=1,2,\ldots,20. \tag{1}$$

Where N_{ii}, N_{jj}, N_{ij} are total numbers of close contact couples between amino acid residue i and i, amino acid residue j and j and amino acid residue i and j, respectively. Take RT as the e_{ij}'s unit.

$$C_{ij}=\sum_{p=1}^{M} n_{rr,p}\frac{(q_i n_{i,p})(q_j n_{j,p})}{(\sum_{k=1}^{20} q_k n_{k,p})^2}.\ i,j=1,2,\ldots,20. \quad (2)$$

Where $w^{i,j}=\sum_{p=1}^{M} w^{i,j}_{p}=\sum_{p=1}^{M}\sum_{p=1}^{M} w^{i,j}_{p}$ and M is total sample number, qi is the coordination number of amino acid residue i. $n_{i,p}$ is the number of close contact couples about amino acid residue i of protein p. $n_{ij,p}$ is the number of close contact couples between amino acid residue i and amino acid residue j of protein p. The interaction energy value can be calculated among the 20 kinds of amino acid residues according to the equation (1) and (2).

Optimization genetic algorithm

The optimization genetic algorithm (OGA) is the combination of the genetic algorithm (GA) and simulated annealing algorithm (SA), which mainly adopts the local optimization strategy.

The algorithm implementation has six steps: 1) One hundred legal HNXP 3D lattice models are randomly generated firstly and respectively calculated the energy value of every spatial structure app: Addword: Respectively. 2) Two structures are randomly selected as algorithm optimization subjects from the 100 structures and then calculate their energies. 3) Randomly select a node as a cross site in the couple of structures, which are selected as the step 2 according to cross probability. Two new structures can be acquired through the interchange of structures on the cross site and then record the new node coordinates. 4) Change the overlapped nodes. The overlapped nodes should be changed by altering their folding directions to ensure that the two new individuals have legal spatial protein structures, and also calculate their energies. 5) The structure with the lowest energy is chosen app: Addword: Elected as the next optimization target among the four structures, the one with the largest energy is deleted. The SA is adopted to filter the structures of remaining two structures according to their energy values. Two new optimization targets are finally obtained. 6) The spatial protein structure will be optimized through above steps, and repeats step 3-5 until the minimum energy spatial structure is found, which is regarded as the stable spatial protein structure under natural condition.

RESULTS

Interaction energy among amino acid residues

Sixty groups of influenza *A* (*H1N1*) virus protein were selected from *PDB* database, *3HTO, 4F15, 3B7E, 3BEQ, 3HTP, 3HTQ, 3HTT, 4EDB, 4B7N, 4B7M, 3M5R, 3M8A, 3LKN, 3QQ4, 3QQ3, 3TI3, 3LZF, 3M6S, 3UBE, 3UBJ, 3UBN-A, 4D9J, 3GBN, 3KHW, 4EDA, 2ZKO, 2LWA, 1RUZ, 2IQH, 4BBL, 1RUY-H, 3A1G, 2HN8, 2YMN-A, 3RVC, 4DYA, 4DYB, 4DYN, 4DYP, 4DYT, 4HKX, 4IRY, 3SM5, 2ZNL, 2ZTT, 3L4Q, 3TG6, 3VDX, 4AVG, 4AVQ, 4AWF, 4AWH, 4AWK, 4AWM, 1RU7, 2WRG, 3RO5, 4EEF, 1RVX, 1RVZ.*20 kinds of amino acid residues produce 200 (1/2□20□20) different interaction energy values according to (1) and (2) and amino acid residues are divided into *H, N, X, P*. The averages of all interaction energy values are defined as the interaction energy values among the four amino acid residues (table 1). The energy value of *H-H*, $E_{H\text{-}H}$*=-0.38RT*, is the lowest; the energy value of *H-X*, $E_{H\text{-}X}$*=0.76RT*, is the largest; the proportions of close contact couples of *H-H* (19%), *H-X* (26%), are the largest; the proportion of close contact couples of *N-N*, only 1%, is the smallest. The proportions in 60 proteins are given in table 2. The proportion of *H* and *X* is 74%. It is easily found out that these data of 60 influenza A (H1N1) viruses conform to population data (Dill, 2008).

Table 1: Interaction energies and proportion of *H, N, X, P* in the 3D structure of influenza A (H1N1) virus

Contact	Energy	Percentage (%)
H-H	-0.38	0.19
H-P	-0.31	0.14
N-X	-0.16	0.06
N-P	-0.13	0.03
H-N	-0.1	0.07
N-N	-0.04	0.01
X-X	0.23	0.12
P-P	0.38	0.03
X-P	0.48	0.09
H-X	0.76	0.26

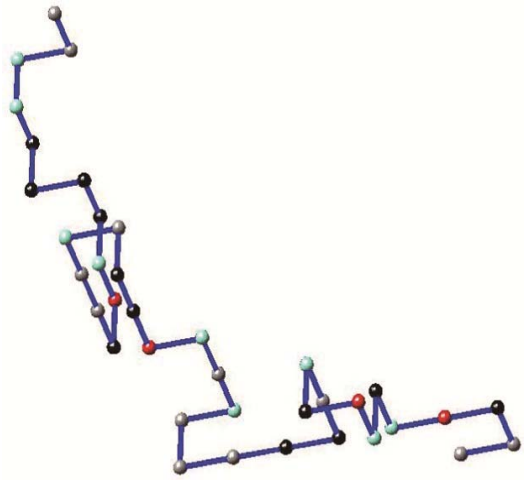

Fig. 1: 2ZTT-B's HNXP 3D structure model

HNXP 3D structure model

The HNXP 3D structure of influenza A (*H1N1*) virus protein can be established according to the energy values

Table 2: Proportion of amino acids residues in the influenza A (H1N1) virus (%)

Category	Residue	Percentage1	Percentage2	Category	Residue	Percentage1	Percentage2
H	A	0．064	0．367	X	G	0．078	0．373
	V	0．055			S	0．080	
	L	0．073			T	0．062	
	I	0．057			C	0．019	
	P	0．040			Y	0．039	
	F	0．038			N	0．063	
	W	0．017			Q	0．033	
	M	0．024		P	K	0．054	0．141
N	D	0．049	0．119		R	0．062	
	E	0．070			H	0．025	

Table 3: 2ZTT-B's coordinates of center carbon atoms (C_a) in PDB (Å)

Residue	X	Y	Z	Residue	X	Y	Z	Residue	X	Y	Z
GLY	3.102	4.897	-4.911	SER	4.236	12.167	13.842	VAL	3.993	-8.205	13.322
SER	6.274	4.184	-2.903	GLN	1.421	11.717	16.373	ASP	3.197	-11.337	15.432
MSE	6.282	7.872	-2.174	SER	1.845	10.045	19.661	HIS	0.674	-9.345	17.510
GLU	2.669	8.069	-0.930	ARG	-0.999	7.595	19.576	MSE	-1.348	-8.153	14.495
ARG	3.214	5.106	1.489	THR	-0.235	6.000	16.182	ALA	-2.631	-11.746	14.126
ILE	6.418	6.353	3.161	ARG	3.587	6.233	16.876	ILE	-3.546	-11.900	17.834
LYS	4.909	9.836	3.503	GLU	3.124	4.155	20.100	ILE	-5.593	-8.673	17.444
GLU	1.886	8.391	5.243	ILE	1.100	1.490	18.232	LYS	-7.239	-9.762	14.162
LEU	4.202	6.440	7.592	LEU	3.436	1.249	15.270	LYS	-8.639	-12.858	15.912
ARG	6.141	9.672	8.446	THR	6.473	1.028	17.645	TYR	-10.627	-10.618	18.331
ASN	2.727	11.197	9.200	LYS	5.063	-1.382	20.256	THR	-11.788	-7.794	16.094

(table 1). As fig. 1 shows, *H, N, X, P* are expressed as black, red, gray and cyan nodes, respectively. Hamiltonian energy function can be calculated, the two spatial adjacent amino acid residue nodes are defined as an energy value e_{ij}. Then the total energy value of spatial protein structure can be obtained, $E=\sum_{i<j-1,i<j} E_{ij}$, where $N$is the length of the amino acid sequence, E_{ij} is the energy value between the two spatial adjacent amino acid residue and (table 1).

3D Structure Analysis of influenza A (H1N1) Virus

The interaction energy e_{ij} among 20 kinds of amino acids can be calculated based on the three dimensional spatial structures of 60 types of influenza *A* (H1N1) viruses. The minimum total energy value can be obtained by OSA, and the degeneracy of the *HNXP* 3D lattice model equals to 1. Hence, the model was thought to be the spatial protein structure of influenza *A* (H1N1) virus in natural state.

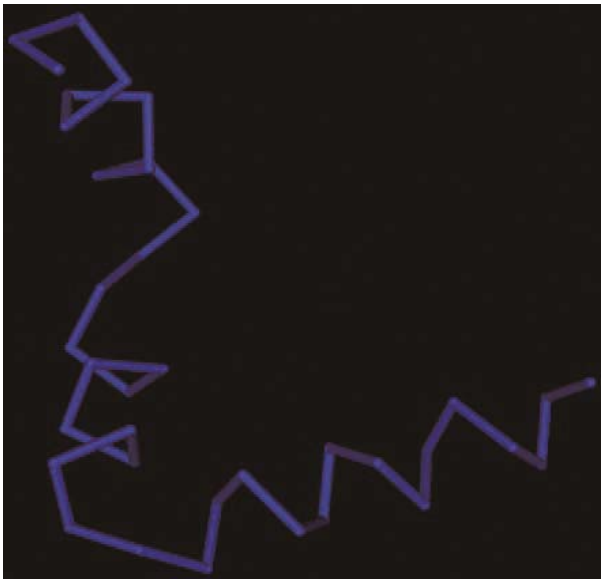

Fig. 2: 2ZTT-B's actual structure in the NCBI

Table 4: 2ZTT-B's coordinates of the nodes on the HNXP 3D structure model

Residue	X	Y	Z	Residue	X	Y	Z	Residue	X	Y	Z
GLY	0	0	1	SER	-1	4	1	VAL	-2	6	3
SER	0	-1	1	GLN	-1	5	1	ASP	-2	6	4
MSE	-1	-1	1	SER	-1	5	2	HIS	-3	6	4
GLU	-1	0	1	ARG	-1	4	2	MSE	-3	6	5
ARG	-1	1	1	THR	-2	4	2	ALA	-4	6	5
ILE	-2	1	1	ARG	-3	4	2	ILE	-4	7	5
LYS	-2	1	0	GLU	-3	5	2	ILE	-4	7	6
GLU	-3	1	0	ILE	-4	5	2	LYS	-5	7	6
LEU	-3	2	0	LEU	-5	5	2	LYS	-5	7	7
ARG	-3	2	1	THR	-5	5	3	TYR	-5	6	7
ASN	-2	2	1	LYS	-5	6	3	THR	-6	6	7
LEU	-1	2	1	THR	-4	6	3				
MSE	-1	3	1	THR	-3	6	3				

Take *2HN8, 2LWA-A, 2ZTT-A, 2ZTT-B, 3A1G-A, 3A1G-D, 3LKN-C* and *3QQ4-C*as representative samples to build 3D structures, respectively. For example, the coordinates of C_α for 2ZTT-A in table 3 were selected from PDB. The HNXP 3D spatial lattice model was used to predict the 3D structure of *2ZTT-A* (fig. 1), and its corresponding coordinates of amino acid nodes in table 4. By comparing the *HNXP* prediction model with Error! Hyperlink reference not valid. structure in the *NCBI* (fig. 2), it can be easily observed that they are similar.

Distance matrices were built by using the node coordinates in table 4 and the coordinates of Cα in fig. 2 (Taylor *et al*., 1989). The two distance matrices were considered to be the characteristic matrices of the two structures. Then the similarity of the two kinds of structures is calculated. And the similarities of the other 7 sequences are also calculated (table 5).

Table 5: The similarity and energy values of the eight sequences

Sequence	S	E
2HN8	0.9204	-3.69
2LWA-A	0.8930	-1.83
2ZTT-A	0.8335	-3.49
2ZTT-B	0.9053	-2.35
3A1G-A	0.8576	-3.24
3A1G-D	0.8938	-1.88
3LKN-C	0.7727	-1.93
3QQ4-C	0.8041	-1.14

CONCLUSIONS

On the basis of tight polymer structure of the influenza *A* (H1N1) virus globular protein, amino acid residues can be divided into four categories according to the number of close contact couples. The HP lattice model was replaced by the HNXP 3D lattice model. In the new model, the energy relationship among amino acids can be precisely divided into 10 groups. The HNXP 3D lattice model can be figured out by using the optimization genetic algorithm. The most-sparsest canonical correlation coefficients are higher than 80%, so the predicted structures are highly similar to the actual structures. The result suggests that HNXP 3D lattice model can predict 3D structure of influenza *A* (H1N1) virus reliably. The approach we proposed can be applied to predict 3D spatial structure of influenza A (H1N1) virus protein.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (Grant No. 11271163) and the Fundamental Research Funds for the Central Universities of Ministry of Education of China (Grant No. JUSRP21117).

REFERENCES

Chan HS and Dill KA (1989). Compact polymers. *Macromolecules,* **22**: 4559-4573.

Dill KA, Ozkan SB, Shell MS and Weikl TR (2008). The protein folding problem. *Annual review of biophysics,* **37**: 289-316.

Dotu I, Cebrian M, Van Hentenryck P and Clote P (2011). On lattice protein structure prediction revisited. *IEEE Acm T Comput Bi,* **8**: 1620-1632.

Gromiha MM (2001). Important inter-residue contacts for enhancing the thermal stability of thermophilic proteins. *Biophys. chem.,* **91**: 71-77.

Gromiha MM and Selvaraj S (2001). Role of medium-and long-range interactions in discriminating globular and membrane proteins. *Int. J. Biol. Macromol.,* **29**: 25-34.

Islam M and Chetty M (2013). Clustered Memetic Algorithm with local heuristics for ab initio protein structure prediction. *IEEE T Evolut Comput,* **17**: 558-576.

Jiang Z, Zhang L, Chen J, Xia A and Zhao D (2002). Effect of amino acid on forming residue–residue contacts in proteins. *Polymer,* **43**: 6037-6047.

Lau KF and Dill KA (1989). A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules,* **22**: 3986-3997.

Miyazawa S and Jernigan RL (1985). Estimation of effective interresidue contact energies From protein crystal structures: Quasi-chemical approximation. *Macromolecules,* **18**: 534-552.

Taylor WR and Orengo CA (1989). Protein structure alignment. *J. Membr. Biol.,* **208**: 1-22.

Voelz VA, Bowman GR, Beauchamp KP and VS (2010). Molecular simulation of abinitio protein folding for a millisecond folder NTL9 (1-39). *J. Am. Chem. Soc.,* **132**: 1526-1528.

Yu ZG, Anh V and Lau KS (2004). Fractal analysis of measure representation of large proteins based on the detailed HP model. *Physica A,* **337**: 171-184.

Zhang C and Kim SH (2000). Environment-dependent residue contact energies for proteins. *Proc. Natl. Acad .Sci.,* **97**: 2550-2555.