J. Biomedical Science and Engineering, 2010, 3, 262-267 **JBiSE**
doi:10.4236/jbise.2010.33035 Published Online March 2010 (http://www.SciRP.org/journal/jbise/).

Mutation pattern in human adrenoleukodystrophy protein in terms of amino-acid pair predictability

Shao-Min Yan[1], Guang Wu[2]*

[1]National Engineering Research Center for Non-food Biorefinery, Guangxi Academy of Sciences, Nanning, China;
[2]Computational Mutation Project, DreamSciTech Consulting, Shenzhen, China; *Corresponding author.
Email: hongguanglishibahao@yahoo.com

Received 28 December 2009; revised 5 January 2010; accepted 8 January 2010.

ABSTRACT

The mutation pattern in protein is a very important feature and is studied through various approaches including the study on mutation pattern in domains where amino acids are converted into numbers from letters. In this study, we converted the amino acids in human adrenoleukodystrophy protein with its 128 missense mutations into random domain using the amino-acid pair predictability, and then we studied their mutation patterns. The results show 1) the mutations are more likely to target the amino-acid pairs whose actual frequency is larger than their predicted one, 2) the mutations are more likely to form the amino-acid pairs whose actual frequency is smaller than their predicted frequency, 3) mutations are more likely to occur at unpredictable amino-acid pairs, and 4) mutations have the trend to narrow the difference between predicted and actual frequencies of amino-acid pairs.

Keywords: Adrenoleukodystrophy;
Amino-Acid Pair Predictability; Mutation Pattern

1. INTRODUCTION

The public-accessible protein databases provide us not only the possibility of tracking the history of protein family as well as other important topics, but also the possibility of analyzing the protein evolution from various angles. Of many facts that influence the protein evolution is the mutation, which has been the objective of many studies.

One way to study the mutation is to find out its pattern, for example, the “hotspot” sites in a protein have been defined as to be sensitive to endogenous and exogenous mutagens [1,2,3]. This approach and others such as multi-sequence comparison and alignment work in the domain of amino-acid sequence, that is, we directly analyze the mutation patterns in terms of the letters, which represent amino acids in a protein.

Another approach, which is extremely powerful and widely used in other research fields, is to analyze the issue of interest in numeric domains, which also paves the ways to use more sophisticated mathematical tools to analyze the mutation patterns. For example, we can use the physicochemical property of amino acids to represent a protein sequence, and analyze the protein sequence in physicochemical domain as a numeric sequence [4].

However, we should not stop here because the physicochemical property and other parameters borrowed from physics and chemistry were developed for the purpose of other types of studies, but may not for the purpose of studying mutations. Our group has developed three approaches since 1999 to study protein mutation in a random domain [5,6,7,8] not only because pure chance is now considered to lie at the very heart of nature [9] but also because our approaches are more sensitive to the protein length, amino-acid composition and position, neighboring amino acids etc. In particular, our approaches are sensitive to mutations because they can give different values before and after mutation [5,6,7,8].

In this study, we will analyze the mutation pattern in the adrenoleukodystrophy protein (ALDP), which is a transporter in the peroxisome membrane and belongs to the ATP-binding cassette transporter superfamily [10,11,12,13]. This protein is involved in the transport of coenzyme A esters of very-long-chain fatty acids from the cytoplasm into the peroxisomal lumen [14,15]. Mutations in the gene *ABCD*1 mapping to Xq28 can result in the defects of adrenoleukodystrophy protein that are the cause of a severe X-linked disease, called X-linked adrenoleukodystrophy [16,17]. This disease is the most common peroxisomal disorder [18] with a minimal incidence of 1:21000 males [19]. It is characterized by the characteristic accumulation of saturated very long-chain fatty acids and disorder of peroxisomal beta-oxidation [20,21,22,23,24]. The clinical outcomes of adrenoleukodystrophy vary strikingly and unpredictably [13,14,15, 16,17,18,19,20,21,22,23,24,25,26]. The phenotypes include the rapidly progressive childhood cerebral form (CCALD), the milder adult form, adrenomyeloneuropathy (AMN), and variants without neurologic involvement. There is no apparent correlation between genotype and phenotype [24,25,26,27].

Besides the importance mentioned above, there are currently 132 mutations documented in human adrenoleukodystrophy protein, which is statistically sufficient for pattern analysis.

2. MATERIALS AND METHODS

2.1. Data

The amino-acid sequence of human adrenoleukodystrophy protein and its 132 mutations were obtained from the UniProtKB/Swiss-Prot (accession number P33897). Among the mutations, 128 are missense point mutations and the rest are small deletions or insertions [28].

2.2. Conversion of Adrenoleukodystrophy Protein into Random Domain

We use the amino-acid pair predictability as a measurement for the randomness of adjacent amino-acid pairs in a protein, and this simple measure can serve as an indicator to the complicity of protein construction, which leads to many our studies [29,30,31,32,33,34,35,36,37, 38,39].

The human adrenoleukodystrophy protein consists of 745 amino acids. The first and second amino acids are an adjacent amino-acid pair, the second and third as another amino-acid pair, the third and fourth, until the 744th and 745th, thus there are 744 adjacent amino-acid pairs.

Then, we can use the permutation to define if an amino-acid pair is predictable. For example, there are 80 alanines (A) and 57 valines (V) in human adrenoleukodystrophy protein: if the permutation works, then the amino-acid pair AV would appear 6 times ($80/745 \times 57/744 \times 744 = 6.12$); actually we do find six AV pairs in this protein, so the appearance of AV is predictable, because the actual frequency is equal to its predicted one.

On the other hand, there are 59 arginines (R) in human adrenoleukodystrophy protein, if the permutation works, the predicted frequency of AR appearance would be 6 ($80/745 \times 59/744 \times 744 = 6.34$); however, the pair AR appears 9 times in realty, so the appearance of AR is unpredictable, because the actual frequency is larger than its predicted one.

Also, we can find the case, where the actual frequency is smaller than its predicted one. For example, there are 94 leucines (L) and 55 glutamic acids (G) in human adrenoleukodystrophy protein so that the predicted frequency of pair LG is 7 ($94/745 \times 55/744 \times 744 = 6.94$), however, its actual frequency is only 2.

2.3. Mutations in Terms of Predictable and Unpredictable Amino-Acid Pairs

A point missense mutation would lead to the change in two amino-acid pairs if the mutation would not occur at terminals. For instance, a mutation at position 484 substitutes proline (P) to arginine (R), which impairs the protein dimerization [40]. This mutation leads amino-acid pairs TP and PS to change to TR and RS, because threonine (T) is located at position 483 and serine (S) is located at position 485. Nevertheless, this mutation would be reflected in terms of predicted frequency, actual frequency and their difference (see Table 1).

This mutation changed the difference between predicted frequency (PF) and actual frequency (AF) of affected amino-acid pairs, $\sum(PF-AF)$. Before mutation, $\sum(PF - AF) = (1 - 1) + (2 - 4) = -2$ for TP and PS, and $\sum(PF - AF) = (2 - 0)+(4 - 4) = 2$ for TR and RS. After mutation, $\sum(PF - AF) = (1 - 0) + (2 - 3) = 0$ for TP and PS, and $\sum(PF - AF) = (2 - 1) + (4 - 5) = 0$ for TR and RS. Needless to say, there would construct a certain mutation pattern if we analyze sufficient mutations.

2.4. Statistics

The data were presented as median with an interquartile. The Mann-Whitney *U*-test and *Chi*-square test were used for comparisons, and $P < 0.05$ is considered statistically significant.

3. RESULTS

As there are 20 kinds of amino acids, they can theoretically construct 400 types of amino-acid pairs, which serve us as a reference for comparison with human adrenoleukodystrophy protein. Of 400 types of amino-acid pairs, 118 are absent including 44 predictable and 74 unpredictable. Consequently 744 amino-acid pairs in human adrenoleukodystrophy protein include only 282 types of amino-acid pairs (400 – 118 = 282), which means these 282 types would host all mutations occurred in the protein. Of those 282 types, 98 are predictable and 184 are unpredictable; while of 744 amino-acid pairs 175 and 569 pairs are predictable and unpredictable because some types of amino-acid pairs appear more than once.

The first mutation pattern is the amino-acid pairs tar-

Table 1. Predicted frequency, actual frequency and their difference of amino-acid pairs affected by T485S mutant of human adrenoleukodystrophy protein.(PF: Predicted frequency, AF: actual frequency; T: threonine, P: praline, R: arginine, S: serine.).

Amino-acid pair	Before mutation			After mutation		
	PF	AF	PF–AF	PF	AF	PF–AF
TP	1	1	0	1	0	1
PS	2	4	–2	2	3	–1
TR	2	0	2	2	1	1
RS	4	4	0	4	5	–1

geted by mutations, or we might call them as hotspots. Table 2 details the substituted amino-acid pairs in terms of the relationship between actual and predicted frequencies. Here, the mutations are more likely to target the amino-acid pairs whose actual frequency is larger than their predicted one, for example, 47 mutations targeted these amino-acid pairs; by contrast, the mutations are less likely to target the amino-acid pairs whose actual frequency is smaller than their predicted one, for example, only 5 mutations targeted these amino-acid pairs. The *Chi*-square test indicates remarkable statistical difference before and after mutation (*P*=<0.001).

The second mutation pattern is the amino-acid pairs formed after mutations. Table 3 details the substituting amino-acid pairs in terms of the relationship between actual and predicted frequencies. The *Chi*-square test indicates remarkable statistical difference before and after mutation (*P*=<0.001). Table 3 has the same format as those in Table 2, thus we can easily find out the second mutation pattern by comparing two tables. For example, the data in the fourth column in both tables are almost in totally opposite orders, that is, mutations are more likely to target the amino-acid pairs whose actual frequency is larger than their predicted frequency, while mutations are more likely to form the amino-acid pairs whose actual frequency is smaller than their predicted frequency. Again, the data in the seventh column in both tables appear somewhat similar.

Meanwhile, Tables 2 and 3 also reveal the third mutation pattern that is mutations are more likely to occur at unpredictable amino-acid pairs (lines 4–8 in both tables).

The above three mutation patterns are mainly related to the relationship in amino-acid pair between actual and predicted frequencies. In fact, the difference between predicted and actual frequencies can also provide us with further mutation patterns. Figure 1 shows the difference between predicted and actual frequencies related to amino-acid pairs identical to substituted and substituting amino-acid pairs before and after mutation, and their statistical results are shown in Figure 2, which highlights the fourth mutation pattern.

Before mutation, the median of difference between predicted and actual frequencies is –2 in substituted amino-acid pairs, suggesting that the mutations occur in the amino-acid pairs, which appear more than their predicted frequency. Meanwhile, the corresponding value is 1 in substituting amino-acid pairs, indicating that the mutations lead to the appearance of the amino-acid pairs that appear less than their predicted frequency.

After mutation, the median of difference between predicted and actual frequencies is 0 in substituted amino-acid pairs, suggesting that these amino-acid pairs are

Table 2. Amino-acid pairs identical to substituted amino-acid pairs before and after mutations. (AF: actual frequency; PF: predicted frequency. There is a remarkable statistical difference before and after mutation (*Chi*-square = 47.841 with 5 degrees of freedom, *P* = <0.001).

Amino-acid pairs			Before Mutation			After Mutation		
	Pair I	Pair II	Number	%	Total %	Number	%	Total %
Predictable	AF=PF	AF=PF	21	16.41	16.41	14	10.94	10.94
Unpredictable	AF>PF	AF>PF	47	36.72	83.59	14	10.94	89.06
	AF>PF	AF=PF	29	22.66		20	15.63	
	AF>PF	AF<PF	20	15.63		30	23.44	
	AF<PF	AF=PF	6	4.69		27	21.09	
	AF<PF	AF<PF	5	3.91		23	17.97	

Table 3. Amino-acid pairs identical to substituting amino-acid pairs before and after mutations (AF: actual frequency; PF: predicted frequency. There is a remarkable statistical difference before and after mutation (*Chi*-square = 54.114 with 5 degrees of freedom, *P* = <0.001).

Amino-acid pairs			Before Mutation			After Mutation		
	Pair I	Pair II	Number	%	Total %	Number	%	Total %
Predictable	AF=PF	AF=PF	11	8.59	8.59	9	7.03	7.03
Unpredictable	AF>PF	AF>PF	6	4.69	91.41	39	30.47	92.97
	AF>PF	AF=PF	24	18.75		40	31.25	
	AF>PF	AF<PF	29	22.66		25	19.53	
	AF<PF	AF=PF	34	26.56		8	6.25	
	AF<PF	AF<PF	24	18.75		7	5.47	

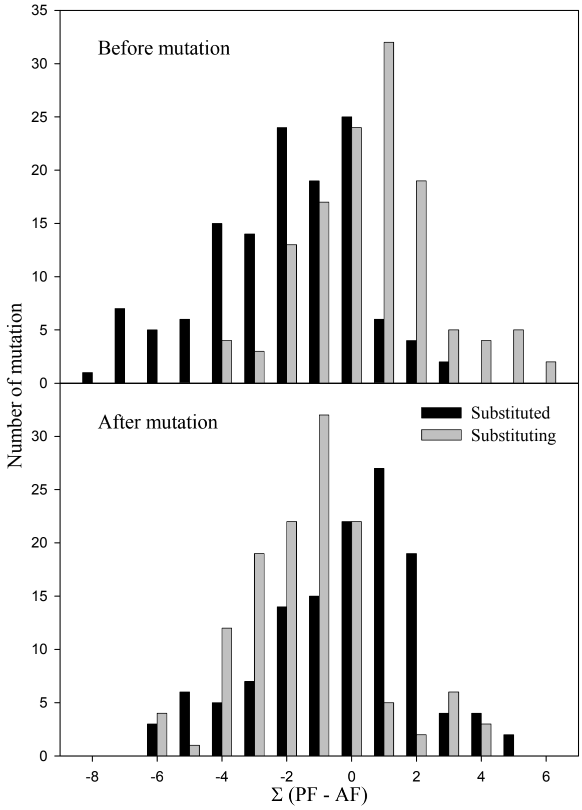

Figure 1. Difference between predicted frequency (PF) and actual frequency (AF) related to amino-acid pairs identical to substituted and substituting amino-acid pairs before and after mutation.

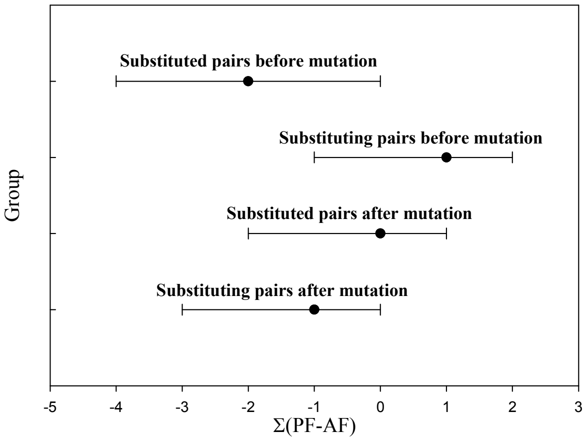

Figure 2. Sum of difference between predicted and actual frequencies [Σ(PF − AF)] of substituted and substituting amino-acid pairs before and after mutation in human adrenoleukodystrophy protein. The data are presented by median with an interquatile interval. There is a statistically significant difference between corresponding groups ($P < 0.001$, Mann-Whitney *U*-test).

more randomly constructed in the mutant adrenoleukodystrophy proteins, as their predicted and actual frequencies are about the same. However, the corresponding value is –1 in substituting amino-acid pairs, indicating that the mutations are favor the amino-acid pairs that already exist in the protein. Striking statistical difference is found between the corresponding groups ($P < 0.001$).

4. DISCUSSION

In this study, we analyze the mutation patterns in numerical domain rather than word descriptions because it is far much easy to find repeatable patterns in numerical domain. Actually there are many ways to analyze the mutation patterns in numerical domain, for example, we can use the physicochemical property to replace amino acids in a protein, and then we can analyze the mutation patterns in physicochemical property domain, which would be an interesting topic for pursuit.

The numerical domain in our approach is random whose rationale has been given in the Introduction, its biological implications would include the followings: 1) Nature follows parsimony, which suggests to construct a protein with minimal time and energy, thus the predictable amino-acid pairs in our approach confirm the nature parsimony, while the unpredictable amino-acid pairs suggest that nature deliberately spends more time and energy to construct them, which could be functional sites. 2) The difference between predicted and actual frequencies in amino-acid pairs can be regarded as a force driving mutation. Although there uncountable factors driving mutations, their effect in fact is the difference between predicted and actual frequencies. 3) The basic effect of mutation in our sense is to narrow the difference between predicted and actual frequencies in targeted amino-acid pairs: most mutations do have such effects. However, the new formed amino-acid pairs could create new unpredictable amino-acid pairs, which once again have the difference between predicted and actual frequencies leading to new mutation so the evolution can continue.

As the difference between predicted and actual frequencies is a measure of random construction of amino-acid pairs in a protein, thus the smaller the difference is, the more random the construction of amino-acid pairs is. In particular, a) the larger the positive difference is, the more randomly unpredictable amino-acid pairs are absent; and b) the larger the negative difference is, the more randomly unpredictable amino-acid pairs are present.

Therefore, this study highlights the mutation patterns in terms of amino-acid pair predictability in human adrenoleukodystrophy protein. In future, we hope to incorporate these mutation patterns in random domain into the changes in the secondary structure contents and consequently affect biological functions of the protein [41,42].

5. ACKNOWLEDGEMENTS

This study was partly supported by National Basic Research Program of China (2009CB724703), Guangxi Science Foundation (0991080) and Guangxi Academy of Sciences (09YJ17SW07).

REFERENCES

[1] Rideout, W.M., Coetzee, G.A., Olumi, A.F. and Jones, P.A. (1990) 5-Methylcytosine as an endogenous mutagen in human LL receptor and p53 genes. *Science*, **249**, 1288-1290.

[2] Montesano, R., Hainaut, P. and Wild, C.P. (1997) Hepatocellular carcinoma: From gene to public health. *Journal of the National Cancer Institute*, **89**, 1844-1851.

[3] Hainaut, P. and Pfeifer, G.P. (2001) Patterns of p53 G→T transversions in lung cancers reflect the primary mutagenic signature of DNA-damage by tobacco smoke. *Carcinogenesis*, **22**, 367-374.

[4] Fasman, G.D. (1976) Handbook of biochemistry: Section D physical chemical data. 3rd Edition, CRC Press, London and New York.

[5] Wu, G., and Yan, S. (2002) Randomness in the primary structure of protein: methods and implications. *Molecular Biology Today*, **3**, 55-69.

[6] Wu, G. and Yan, S. (2006) Mutation trend of hemagglutinin of influenza A virus: A review from computational mutation viewpoint. *Acta Pharmacologica Sinica*, **27**, 513-526.

[7] Wu, G. and Yan, S. (2006) Fate of influenza A virus proteins. *Protein and Peptide Letters*, **13**, 377-384.

[8] Wu, G. and Yan, S. (2008) Lecture Notes on Computational Mutation. Nova Science Publishers, New York.

[9] Everitt, B.S. (1999) Chance rules: An informal guide to probability, risk, and statistics. Springer, New York.

[10] Efferth, T. (2001) The human ATP-binding cassette transporter genes: From the bench to the bedside. *Current Molecular Medicine*, **1**, 45-65.

[11] Pohl, A., Devaux, P.F. and Herrmann, A. (2005) Function of prokaryotic and eukaryotic ABC proteins in lipid transport. *Biochimica and Biophysica Acta*, **1733**, 29-52.

[12] Oswald, C., Holland, I.B. and Schmitt, L. (2006) The motor domains of ABC-transporters. What can structures tell us? *Naunyn-Schmiedeberg's Archives of Pharmacology*, **372**, 385-399.

[13] Kim, J.H. and Kim, H.J. (2005) Childhood X-linked adrenoleukodystrophy: Clinical-pathologic overview and MR imaging manifestations at initial evaluation and follow-up. *Radiographics*, **25**, 619-631.

[14] Shimozawa, N. (2007) Molecular and clinical aspects of peroxisomal diseases. *Journal of inherited metabolic disease*, **30**, 193-197.

[15] Wanders, R.J., Visser, W.F., van Roermund, S., Kemp, C.W. and Waterham, H.R. (2007) The peroxisomal ABC transporter family. *Pflügers Archiv European Journal of Physiology*, **453**, 719-734.

[16] Moser, H., Dubey, P. and Fatemi, A. (2004) Progress in X-linked adrenoleukodystrophy. *Current opinion in neurology*, **17**, 263-269.

[17] Moser, H.W., Mahmood, A. and Raymond, G.V. (2007) X-linked adrenoleukodystrophy. *Nature Clinical Practice. Neurology*, **3**, 140-151.

[18] Wanders, R.J. and Waterham, H.R. (2005) Peroxisomal disorders I: Biochemistry and genetics of peroxisome biogenesis disorders. *Clinical Genetics*, **67**, 107-133.

[19] Bezman, L., Moser, A.B., Raymond, G.V., Rinaldo, P., Watkins, P.A., Smith, K.D., Kass, N.E. and Moser, H.W. (2001) Adrenoleukodystrophy: Incidence, new mutation rate, and results of extended family screening. *Annals of Neurol*, **49**, 512-517.

[20] Elgersma, Y. and Tabak, H.F. (1996) Proteins involved in peroxisome biogenesis and functioning. *Biochimica and Biophysica Acta*, **1286**, 269-283.

[21] Hettema, E.H. and Tabak, H.F. (2000) Transport of fatty acids and metabolites across the peroxisomal membrane. *Biochimica and Biophysica Acta*, **1486**, 18-27.

[22] Clayton, P.T. (2001) Clinical consequences of defects in peroxisomal beta-oxidation. *Biochemical Society Transactions*, **29**, 298-305.

[23] Hargrove, J.L., Greenspan, P. and Hartle, D.K. (2004) Nutritional significance and metabolism of very long chain fatty alcohols and acids from dietary waxes. *Experimental Biology and Medicine*, **229**, 215-226.

[24] Kemp, S. and Wanders, R.J. (2007) X-linked adrenoleukodystrophy: Very long-chain fatty acid metabolism, ABC half-transporters and the complicated route to treatment. *Molecular Genetics and Metabolism*, **90**, 268-276.

[25] Takano, H., Koike, R., Onodera, O. and Tsuji, S. (2000) Mutational analysis of X-linked adrenoleukodystrophy gene. *Cell Biochemistry and Biophysics*, **32**, 177-185.

[26] Berger, J. and Gärtner, J. (2006) X-linked adrenoleukodystrophy: Clinical, biochemical and pathogenetic aspects. *Biochimica and Biophysica Acta*, **1763**, 1721- 1732.

[27] Kemp, S., Pujol, A., Waterham, H.R., van Geel, B.M., Boehm, C.D., Raymond, G.V., Cutting, G.R., Wanders, R.J.A. and Moser, H.W. (2001) ABCD1 mutations and the X-linked adrenoleukodystrophy mutation database: Role in diagnosis and clinical correlations. *Human Mutation*, **18**, 499-515.

[28] Bairoch, A. and Apweiler, R. (2000) The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 2000. *Nucleic Acids Research*, **28**, 45-48.

[29] Wu, G. (1999) The first and second order Markov chain analysis on amino acids sequence of human haemoglobin -chain and its three variants with low O_2 affinity. *Comparative Haematology International*, **9**, 148-151.

[30] Wu, G. (2000) The first, second, third and fourth order Markov chain analysis on amino acids sequence of human dopamine-hydroxylase. *Molecular Psychiatry*, **5**, 448-451.

[31] Wu, G. and Yan, S.M. (2001) Prediction of presence and absence of two- and three-amino-acid sequence of human monoamine oxidase B from its amino acid composition according to the random mechanism. *Biomolecular Engineering*, **18**, 23-27.

[32] Wu, G. and Yan, S.M. (2002) Estimation of amino acid pairs sensitive to variants in human phenylalanine hydroxylase protein by means of a random approach. *Peptides,* **23**, 2085-2090.

[33] Wu, G. and Yan, S. (2003) Determination of amino acid pairs sensitive to variants in human-glucocerebrosidase by means of a random approach. *Protein Engineering Design and Selection*, **16**, 195-199.

[34] Wu, G. and Yan, S. (2004) Fate of 130 hemagglutinins from different influenza A viruses. *Biochemical and BioPhysical Research Communications*, **317**, 917-924.

[35] Wu, G. and Yan, S. (2005) Prediction of mutation trend in hemagglutinins and neuraminidases from influenza A viruses by means of cross-impact analysis. *Biochemical and Bio-Physical Research Communications*, **326**, 475-482.

[36] Wu, G. and Yan, S. (2006) Timing of mutation in hemagglutinins from influenza A virus by means of amino-acid distribution rank and fast Fourier transform. *Protein and Peptide Letters*, **13**, 143-148.

[37] Wu, G. and Yan, S. (2007) Prediction of mutations in H1 neuraminidases from North America influenza A virus engineered by internal randomness. *Molecular Diversity*, **11**, 131-140.

[38] Wu, G. and Yan, S. (2008) Prediction of mutations engineered by randomness in H5N1 neuraminidases from influenza A virus. *Amino Acids*, **34**, 81-90.

[39] Yan, S. and Wu, G. (2009) Describing evolution of hemagglutinins from influenza A viruses using a differential equation. *Protein and Peptide Letters*, **16**, 794-804.

[40] Zhou, H.X. (2004) Improving the understanding of human genetic diseases through predictions of protein structures and protein-protein interaction sites. *Current Medicinal Chemistry*, **11**, 539-549.

[41] Kuvaniemi, H., Tromp, G. and Prockop, D.J. (1997) Mutations in fibrillar collagens (types I, II, III, and IV), fibril-associated collagen (type IV), and network-forming collagen (type X) cause a spectrum of diseases of bone, cartilage, and blood vessels. *Human Mutation*, **9**, 300-315.

[42] Kashtan, C.E. (2000) Alport syndromes: Phenotypic heterogeneity of progressive hereditary nephritis. *Pediatric Nephrology*, **14**, 502-512.