Available online at www.sciencedirect.com
ScienceDirect
Biochemical and Biophysical Research Communications 362 (2007) 233–237
BBRC
www.elsevier.com/locate/ybbrc

Mini Review

How similar are amino acid mutations in human genetic diseases and evolution

Hao Wu [a], Bin-Guang Ma [a,b], Ji-Tao Zhao [a], Hong-Yu Zhang [a,*]

[a] *Shandong Provincial Research Center for Bioinformatic Engineering and Technique, Center for Advanced Study, Shandong University of Technology, Zibo 255049, PR China*
[b] *College of Chemistry and Chemical Engineering, Suzhou University, Suzhou 215006, PR China*

Received 3 July 2007
Available online 2 August 2007

Abstract

Accumulating evidence indicates that some deleterious mutations responsible for genetic diseases may offer benefits for human to prevent other diseases. Therefore, human genetic diseases and evolution were tentatively regarded as the two sides of the same coin, which stimulated our interest to explore how similar are amino acid mutations in human genetic diseases and evolution. Through a large-scale analysis on amino acid mutation patterns of genetic diseases and evolution of Hominidae (*Homo sapiens* and *Pan troglodytes*), it was found that there exist significant correlations between two mutation patterns. Besides, there also exist some evident differences between both mutations, especially those associated with four amino acids C, G, R, and L. These findings are of significance to understanding the subtle connections between human genetic diseases and evolution.

Keywords: Genetic diseases; Deleterious mutation; Evolution

One of the most intriguing aspects in genetic diseases is that some of them can help the hosts to fight against other diseases. For instance, deleterious mutations associated with components in red blood cells and immune systems (e.g., human leukocyte antigen, α and β-globin, Duffy factor, tumor necrosis factor and glucose-6-phosphate dehydrogenase) are likely to benefit malarial prevention [1,2]. Besides, cystic fibrosis brings resistance to cholera toxin [3] and hereditary deafness may protect the body from infection and help to heal the wounds [4,5]. Therefore, it seems that some mutations causing genetic diseases are beneficial to the hosts to survive in other adverse environments and thus genetic diseases and evolution are likely to be the two sides of the same coin [6,7], which is supported by the similar nucleotide mutation patterns in human genetic diseases and evolutionary gene/pseudogene [6,7]. However, as proteins are the major players in the drama of life, it is interesting to compare the mutation patterns of human genetic diseases and evolution on the level of amino acids.

To perform the comparison, first, 1535 genetic disease-associated genes were collected from Human Gene Mutation Database (HGMD, http://www.hgmd.org/) [8] in Jan 2006 to constitute the disease-relevant protein dataset, which contains 21,395 amino acid missense mutations. Then, Jordan et al.'s dataset for substitutions in orthologous proteins of 15 taxa (ftp://ftp.ncbi.nih.gov/pub/koonin/Jordan/Cysteine) [9] was used as the data source of evolution-relevant proteins. 6976 amino acid substitutions contained in 7552 orthologous protein sequences corresponding to Hominidae (*Homo sapiens* and *Pan troglodytes*) evolution within the last 10 million years were selected to constitute the evolutionary protein dataset.

By counting the mutation (or substitution) numbers of 20 types of amino acids in the above two datasets, the amino acid mutation (or substitution) matrix (20×20)

* Corresponding author. Fax: +86 533 2780271.
E-mail address: zhanghy@sdut.edu.cn (H.-Y. Zhang).

0006-291X/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2007.07.141

Table 1
Amino acid mutation matrix derived from disease-relevant protein dataset[a]

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	0	0	147	76	0	50	0	0	0	0	0	0	206	0	0	76	370	356	0	0
C	0	0	0	0	138	112	0	0	0	0	0	0	0	0	342	152	0	0	98	412
D	39	0	0	98	0	195	126	0	0	0	0	318	0	0	0	0	0	129	0	138
E	53	0	108	0	0	143	0	0	529	0	0	0	0	79	0	0	0	54	0	0
F	0	75	0	0	0	0	0	41	0	193	0	0	0	0	0	152	0	58	0	12
G	149	149	441	344	0	0	0	0	0	0	0	0	0	0	797	427	0	333	44	0
H	0	0	44	0	0	0	0	0	0	38	0	28	73	72	172	0	0	0	0	129
I	0	0	0	0	67	0	0	0	22	32	81	105	0	0	13	61	268	100	0	0
K	0	0	0	156	0	0	0	15	0	0	17	111	0	39	74	0	41	0	0	0
L	0	0	0	0	195	0	37	22	0	0	36	0	743	67	221	89	0	122	20	0
M	0	0	0	0	0	0	0	132	59	48	0	0	0	0	92	0	174	178	0	0
N	0	0	88	0	0	0	42	62	178	0	0	0	0	0	0	182	46	0	0	45
P	84	0	0	0	0	0	50	0	0	483	0	0	0	39	152	234	94	0	0	0
Q	0	0	0	47	0	0	69	0	46	26	0	0	100	0	136	0	0	0	0	0
R	0	651	0	0	0	235	527	11	60	180	9	0	254	574	0	135	43	0	565	0
S	19	76	0	0	163	49	0	61	0	155	0	95	216	0	148	0	43	0	21	43
T	122	0	0	0	0	0	0	241	49	0	223	51	90	0	65	36	0	0	0	0
V	140	0	85	78	109	117	0	136	0	122	284	0	0	0	0	0	0	0	0	0
W	0	128	0	0	0	40	0	0	0	28	0	0	0	0	191	47	0	0	0	0
Y	0	336	71	0	19	0	117	0	0	0	0	44	0	0	0	54	0	0	0	0

[a]Each element of the matrix indicates the number of amino acid mutation (from column to row). The cells are colored according to the mutation frequency. The top 2.5% highest frequencies are in red, the next 2.5–12.5% are in orange, and the remaining non-zero cells are in yellow.

was constructed (Table 1 for genetic disease and Table 2 for Hominidae evolution). Each element of the matrix indicates the number of amino acid mutation or substitution (from column to row). There are totally 400 elements for each matrix which is further normalized by dividing each component by the largest component of that matrix. Then linear correlation analysis was made between the elements of both normalized matrixes (mutation matrix for genetic disease and substitution matrix for Hominidae evolution). A significant linear correlation was observed between them ($r = 0.57$, $p < 0.0001$, Fig. 1), suggesting that there exist certain similarity between the amino acid mutation patterns of human genetic disease and evolution. To shed more light on this similarity, further analysis of amino acid mutation patterns characterized by mutation contribution ratio were performed.

Mutation contribution ratio (R_i) of each type of amino acid is defined as: $R_i = (N_i/N_t) \times 100\%$, where N_i is the number of mutation from amino acid i to other amino acids, and N_t is the total number of mutations. By analyzing the correlation between mutation contribution ratios of amino acids for genetic diseases (R_i^{d}) and Hominidae evolution (R_i^{e}), we can get some deeper insights into the similarity observed above.

R_i^{d} and R_i^{e} for 20 amino acids are presented in Supplementary Tables 1 and 2. As shown in Fig. 2, there is certain correlation ($r = 0.49$, $p = 0.029$) between both parameters, but the correlation is not very strong. However, if 4 outliers, C, G, R, and L (labeled in red in Fig. 2), are removed, the correlation gets much more significant ($r = 0.79$, $p < 0.0001$). As the four outliers are located in the upper part of the figure, their deviations imply that the mutations associated with the four amino acids make relatively more contribution in genetic diseases or relatively less contribution in Hominidae evolution.

To pinpoint which is the dominant factor, the correlations between N_i and initial amino acid numbers were analyzed (Fig. 3). From Fig. 3A and B, one can find that C, G, and R contribute more than others in inducing genetic diseases and L contributes less than others in driving evolution. This phenomenon may be explained by the following facts: (i) protein structures are more sensitive to mutations from G and C than others, because G frequently occurs at the turns of α-helices and C is indispensable in forming disulphide bond [10,11]; (ii) CpG dinucleotides in Arg(R) codons contribute significantly to the incidence of human genetic diseases, because of the frequent deamination of