BMC Evolutionary Biology

RESEARCH ARTICLE Open Access

Genome wide exploration of the origin and evolution of amino acids

Xiaoxia Liu[1], Jingxian Zhang[2], Feng Ni[1], Xu Dong[2], Bucong Han[2], Daxiong Han[1], Zhiliang Ji[1,2*], Yufen Zhao[1,3*]

Abstract

Background: Even after years of exploration, the terrestrial origin of bio-molecules remains unsolved and controversial. Today, observation of amino acid composition in proteins has become an alternative way for a global understanding of the mystery encoded in whole genomes and seeking clues for the origin of amino acids.

Results: In this study, we statistically monitored the frequencies of 20 alpha-amino acids in 549 taxa from three kingdoms of life: archaebacteria, eubacteria, and eukaryotes. We found that the amino acids evolved independently in these three kingdoms; but, conserved linkages were observed in two groups of amino acids, (A, G, H, L, P, Q, R, and W) and (F, I, K, N, S, and Y). Moreover, the amino acids encoded by GC-poor codons (F, Y, N, K, I, and M) were found to "lose" their usage in the development from single cell eukaryotic organisms like *S. cerevisiae* to *H. sapiens*, while the amino acids encoded by GC-rich codons (P, A, G, and W) were found to gain usage. These findings further support the co-evolution hypothesis of amino acids and genetic codes.

Conclusion: We proposed a new chronological order of the appearance of amino acids (L, A, V/E/G, S, I, K, T, R/D, P, N, F, Q, Y, M, H, W, C). Two conserved evolutionary paths of amino acids were also suggested: A→G→R→P and K→Y.

Background

The origin of life arising from either proteins or nucleic acids has been argued for nearly half century. Putting the "Chicken or Egg" question aside, there exist some unsolved problems. Which amino acid(s) appeared first in the prebiotic environment? What cause the different usage of amino acids in modern organisms? To address these questions, a number of hypotheses and theories, e.g. mutation drifts and natural selection, have been proposed. Multiple factors, such as genetic codes, physicochemical properties, mutation-selection equilibrium, amino acid biosynthesis, etc, are likely related to the variation of amino acid usage in organisms [1,2]. Since there is no way to trace geological evidence in the way scientists normally use in chronicling the evolution of organisms, an alternative path is needed to seek a clue from current living organisms.

Observation of amino acid composition in proteins was recently applied as a statistical approach in facilitating various investigations of the evolution of genetic codes [3], the origin of amino acids [1,2,4-6], the co-evolution of amino acids and genetic codes [7], the evolution of protein families [8-10], the conservation of subcellular location [11], the prediction of protein secondary structure [12-14], the natural selection of protein charge [15], the correlation between gene expression level and protein function [16], the kinship of different taxa [17], the molecular mechanisms of dinosaur extinction [18], the lifestyles of organisms [19], and even the tracing of the Latest Universal Ancestor (LUA) of life [4-6]. Recently, some research groups have successfully applied genomic information on monitoring amino acid composition linked with various biological phenomena [1,5,11,17,20]. It is beyond question that an insight into the evolution of amino acids on a genomic scale can extend our knowledge about molecular evolution and the origin of life. In this study, 549 genomes from three kingdoms of life were adopted to investigate statistically the patterns of amino acid usage during evolution. Also, clues for the origin of amino acids in prebiotic environment and their co-evolution with genetic codes were explored.

* Correspondence: appo@bioinf.xmu.edu.cn; yfzhao@xmu.edu.cn
[1] The Key Laboratory for Chemical Biology of Fujian Province, Department of Chemistry, College of Chemistry and Chemical Engineering, Xiamen University, Xiamen 361005, Fujian, PR China

Results and Discussion

Chronological order of appearance of amino acids

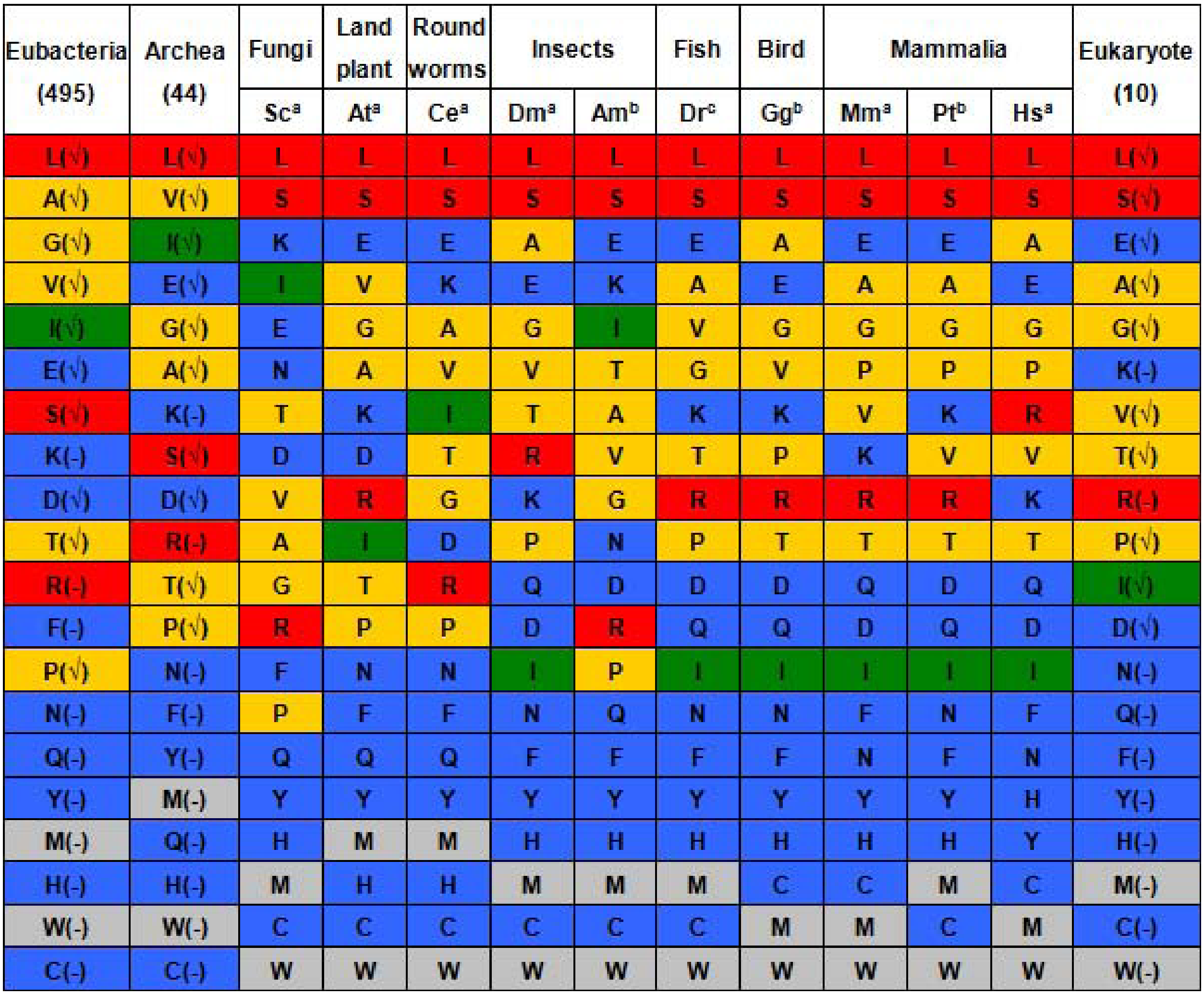
Which amino acid(s) appeared first in the prebiotic environment? To address this question, we might go back to the first life form in the world. When the first simple life was formed, most amino acid biosynthesis processes had not become fully functional. The environment was the only source to acquire amino acids and other fundamental bio-molecules for life. As a consequence, the amino acid composition of the early life was mainly determined by the amino acid content in the "prebiotic soup" with no or little bias on selection of amino acids. It was assumed that the "early" amino acids had higher concentration in the primitive environment than that of "late" amino acids, thus had higher composition in early life form. Retrospectively, if the amino acid composition of the early life form was estimated, it could be used to determine the amino acid concentration in the environment and further deduced the chronological order of amino acid appearance.

According to this assumption, we estimated the amino acid composition of early life form by genome-wide monitoring of amino acid usage in modern organisms. As observation, amino acid compositions are substantially varied not only inter-species between three kingdoms of life but also inter-species within a kingdom (Additional Files 1). This caused difficulty in deducing a consensus amino acid composition for the LUA. In an additional construction of taxa kinship hierarchy based on amino acid composition in the three kingdoms of life (data not presented), we found that taxa from same life kingdom tended to gather together. Therefore, an ancestral amino acid usage was determined separately by kingdoms as follows: Amino acids were first scored from 20 to 1 in terms of descending order of their frequencies in each designated species. The sum of the scores for each amino acid was then calculated and ranked by kingdom of life (Figure 1). Integrating the data of three life domains, a generally-agreed rank of amino acid frequency was achieved. This rank was considered as the estimated amino acid composition of the early life form. Accordingly, a possible chronological order of amino acid appearance was thus proposed in descending order: L, A, V/E/G, S, I, K, T, R/D, P, N, F, Q, Y, M, H, W, C. This order agrees well with the previous findings of Miller's experiments [21] that ten "early" amino acids (A, D, E, G, I, L, P, S, T, V) rank in the top 12. It is slightly different from Trifonov's study (G/A, V/D, P, S, E/L, T, R, N, K, Q, I, C, H, F, M, Y, W) which was determined by comparison of forty different single-factor criteria and multi-factor hypotheses [22]. The assignment of "early or late" amino acids was also supported by additional correlation analyses between physiochemical properties of amino acids and their genetic codes (Additional File 2). Both this study (Figure 1 & Additional File 2) and previous evidence [21,22] supported the assignment of aromatic amino acids (F, W, and Y) as "late" amino acids. Moreover, the effects of a high pH primitive environment on genetic codes in early earth environment determined that there were more early basic amino acids than early acidic amino acids [15].

Co-evolution of amino acids and genetic codes

It has been suggested earlier that amino acid composition was determined largely by existing genetic codes [23]. In our study, the relationship between amino acids and codons has also been studied. As shown in Figure 1, the amino acids with more codons are "favored" by proteins. This phenomenon was observed not only in eukaryotes, but also in most representatives of eubacteria and archaebacteria. Two six-codon owners, leucine and serine, are the most frequently-used amino acids in all selective eukaryotic species. Arginine is also a six-codon amino acid, but its frequency of use is much lower than expected (averagely ranking 9th in eukaryotes, 10th in archaebacteria, and 11th in eubacteria). The under-utilization of arginine is as yet mechanistically unclear, but it may be related to its physiochemical properties and roles in protein functions. All the four-codon amino acids (A, G, V, T, and P) are positioned in the middle zone, and most of the two-codon amino acids and all the one-codon amino acids are used less often.

Previous research has proposed that all amino acids with declining frequencies were the first to be incorporated into the genetic code [1]. To examine this finding, 3D charts of amino acid frequency-codon relationship were prepared (Figure 2), including 5 selected eukaryotic representatives in the same branch of Darwin Evolution (*S. cerevisiae*, *D. rerio*, *M. musculus*, *P. troglodytes*, and *H. sapiens*). Interestingly, all the amino acids encoded by GC-rich codons (definitively, CCX/GCX/GGX/UGG [24,25]), i.e. (P, A, G, and W), increased their frequencies from *S. cerevisiae* to *H. sapiens*; while all the amino acids encoded by GC-poor codons (definitively, AAX/AUX/UAX/UUX), i.e. (F, Y, N, K, I, M), decreased. These results conflict with the previous findings of Jordan and his team [1] that P, A, G, and E 'lose' in protein evolution. The disagreement may be caused by different protein data sets adopted by these two studies and different evolutionary history of amino acids in three kingdoms of life (Figure 1 & Additional file 1). Further statistics of codons showed that the numbers of GC-rich codons (CCX/GCX/GGX/UGG) increased from *S. cerevisiae* to *H. sapiens*, while the GC-poor codons (AAX/AUX/UAY/UUY) decreased (Figure 2c, 2d & Additional



Eubacteria (495)	Archea (44)	Fungi	Land plant	Round worms	Insects		Fish	Bird	Mammalia			Eukaryote (10)
		Sc^a	At^a	Ce^a	Dm^a	Am^b	Dr^c	Gg^b	Mm^a	Pt^b	Hs^a	
L(√)	L(√)	L	L	L	L	L	L	L	L	L	L	L(√)
A(√)	V(√)	S	S	S	S	S	S	S	S	S	S	S(√)
G(√)	I(√)	K	E	E	A	E	E	A	E	E	A	E(√)
V(√)	E(√)	I	V	K	E	K	A	E	A	A	E	A(√)
I(√)	G(√)	E	G	A	G	I	V	G	G	G	G	G(√)
E(√)	A(√)	N	A	V	V	T	G	V	P	P	P	K(-)
S(√)	K(-)	T	K	I	T	A	K	K	V	K	R	V(√)
K(-)	S(√)	D	D	T	R	V	T	P	K	V	V	T(√)
D(√)	D(√)	V	R	G	K	G	R	R	R	R	K	R(-)
T(√)	R(-)	A	I	D	P	N	P	T	T	T	T	P(√)
R(-)	T(√)	G	T	R	Q	D	D	D	Q	D	Q	I(√)
F(-)	P(√)	R	P	P	D	R	Q	Q	D	Q	D	D(√)
P(√)	N(-)	F	N	N	I	P	I	I	I	I	I	N(-)
N(-)	F(-)	P	F	F	N	Q	N	N	F	N	F	Q(-)
Q(-)	Y(-)	Q	Q	Q	F	F	F	F	N	F	N	F(-)
Y(-)	M(-)	Y	Y	Y	Y	Y	Y	Y	Y	Y	H	Y(-)
M(-)	Q(-)	H	M	M	H	H	H	H	H	H	Y	H(-)
H(-)	H(-)	M	H	H	M	M	M	C	C	M	C	M(-)
W(-)	W(-)	C	C	C	C	C	C	M	M	C	M	C(-)
C(-)	C(-)	W	W	W	W	W	W	W	W	W	W	W(-)

Figure 1 Rankings of amino acid composition in three kingdoms of life. Frequency rankings for 20 alpha amino acids in eubacteria, archaebacteria and 10 selected eukaryotic representatives: *Saccharomyces cerevisiae* (Sc), *Abrabidopsis thaliana* (At), *Caenorhaditis elegans* (Ce), *Drosophila melanogaster* (Dm), *Apis mellifera* (Am), *Danio rerio* (Dr), *Gullus gallus* (Gg), *Mus musculus* (Mm), *Pan troglodytes* (Pt), and *Homo sapiens* (Hs). The status of genome was labeled as: $^{(a)}$ complete annotation, $^{(b)}$ draft assembly, $^{(c)}$ in progress. The numbers of genomes from the three life kingdoms of archaebacteria, eubacteria and eukaryotes are presented after the kingdom names. For each taxa or life kingdom, the amino acids were ranked in descending order of their frequencies in the whole proteome. Amino acids are colored by the number of codons: red, orange, green, blue, and gray stands for 6, 4, 3, 2, and 1 codons respectively. An amino acid is marked with "√ " if it was detected by Miller's spark tube experiments, otherwise, marked with "-".

file 3). Similar results have also been obtained in previous studies using different approaches [26,27].

Additionally, we calculated the correlation coefficients between "random" amino acid frequencies following from a uniform usage of codons of the universal genetic code and amino acid compositions of the modern organisms (Additional File 1). As in previous findings [17], all eukaryotic representatives showed a higher correlation coefficient, indicating the small selection of amino acid composition of proteins in eukaryotes. However in eubacteria and in archaea, correlation coefficients varied from 0.05 to 0.9, suggesting that some microbials show a significant selection of amino acids for their proteins. The substantial variety of selection pressure in microbials may be explained by factors such as particular living environments, frequent mutation, rapid generation, etc. To have an overview of how GC content could affect amino acid usage, we compared the GC% of both coding regions and non-coding regions in the whole genomes of eight organisms. Statistically, the coding regions in lower eukaryotes have rather higher net GC content than the non-coding regions, but this is manifestly reversed in higher organisms (*A. mellifera*, *D. rerio*, *M. musculus*, and *H. sapiens*), where it can be

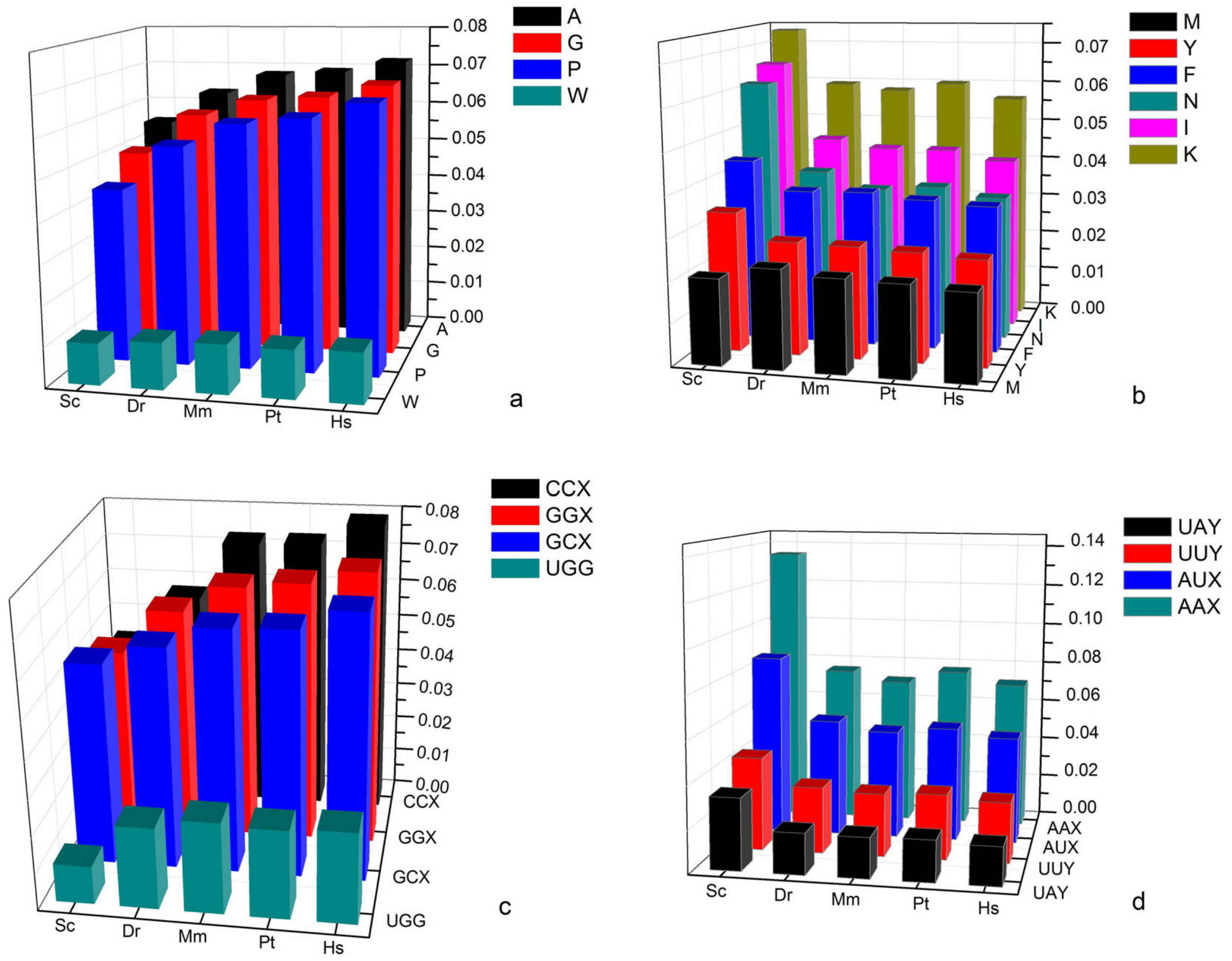


Figure 2 Trends in usage of GC-rich/poor codons and their encoded amino acids. Trends in usage of amino acids encoded by GC-rich codons (2a), amino acids encoded by GC-poor codons (2b), GC-rich codons (2c), and GC-poor codons (2d) over five eukaryotic organisms, *Saccharomyces cerevisiae* (Sc), *Danio rerio* (Dr), *Mus musculus* (Mm), *Pan troglodytes* (Pt), and *Homo sapiens* (Hs). Amino acids encoded by GC-rich codons (P, A, G, W) increase their usage from lower organisms to higher organisms while amino acids encoded by GC-poor codons (F, Y, N, K, I, M) in general decrease. All GC-rich codons (CCX, GCX, GGX, and UGG) increase their usage over the five eukaryotic organisms, while GC-poor codons (AAX, AUX, UUY and UAY) decrease their usage over eukaryotic species.

seen that the net GC content of the coding regions decreases from lower eukaryotes to higher eukaryotes (Figure 3 & Additional file 4). But our previous finding (Figure 2c) indicates that the usage of GC-rich codons increases from *S. cerevisiae* to *H. sapiens*. So the decrease in G and C content in coding region in higher eukaryotic species might come from the decrease in the usage of intermediate-GC codons (defined in ref 24). All these suggest the GC rich condons are favorable in proteins even under the pressure of the decrease in GC content.

To seek a connection between the frequency changes of amino acids and genetic codes, correlation analyses were established for eukaryotes and eubacteria (Figure 4). It was shown that all amino acids encoded by GC-rich codons (P, A, G, W) clustered together both in the eukaryote and eubacteria. All the amino acids encoded by GC-poor codons (F, Y, N, K, I, M) in bacteria gathered into a cluster except methionine. These results further support the co-evolution of amino acids and genetic codes.

Kinship of amino acids

It is challenging to describe how amino acids develop from "early" to "late". A plausible approach is to seek hints from the correlation of amino acid composition. This is based on the assumption that two amino acids are evolutionarily connected if they are correlated in frequencies across species. In this study, Pearson Correlation Coefficients (*r*) of amino acid compositions were calculated separately within

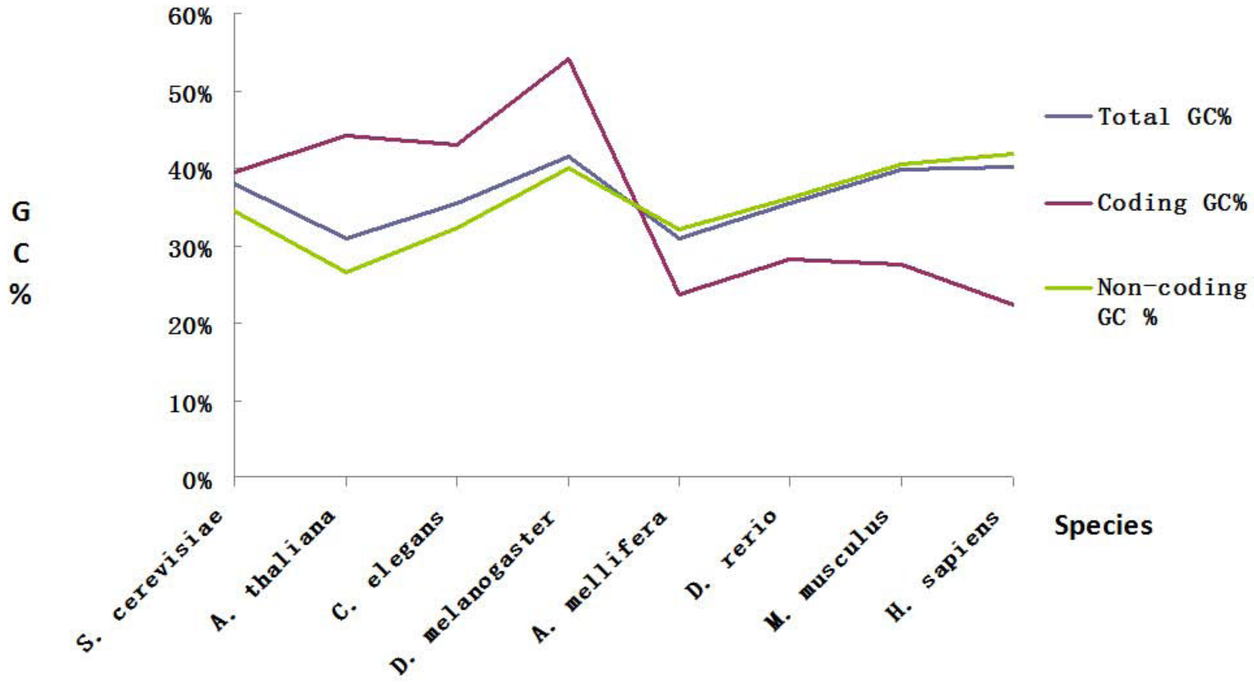


Figure 3 Comparison of GC% in coding regions and non-coding regions. The GC% of both coding regions and non-coding regions in the whole genomes of eight organisms (*S. cerevisiae*, *A. thaliana*, *C. elegans*, *D. melanogaster*, *A. mellifera*, *D. rerio*, *M. musculus*, and *H. sapiens*) were compared. The GC% in the coding regions is higher than that of the non-coding regions in lower organisms, while reversed in higher organisms.

three kingdoms of life, and the results were arranged and presented in triangular tables (Figure 5). It was observed that amino acids were gathered into several zones: two zones of (A, G, P, R) and (K, I, N, Y) in eubacteria (Figure 5a), two zones of (P, R, V) and (N, I, K) in archaebacteria (Figure 5b), and two zones of (D, K, I, N, Y) and (A, G, P, R) in eukaryotes (Figure 5c). Within a zone, the amino acid pairs show highly positive or negative correlation (colored in red or deep blue, respectively) in frequency. The correlations was further analyzed and illustrated in kinship maps (Figure 6a and Figure 6b). As illustration, 20 alpha-amino acids were assigned into two distinct clusters: (cluster 1 (D, F, I, K, M, N, S, T, V and Y) and cluster 2 (A, C, E, G, H, L, P, Q, R and W)) for eukaryotes and cluster 1 (A, D, G, H, L, M, P, Q, R, T, V and W) and cluster 2 (C, E, F, I, K, N, S and Y) for eubacteria. The amino acids are connected by lines of different correlations. It is evident that the positively correlated amino acids normally locate in the same cluster, which suggests a common evolutionary history or functional connection. In contrast, those negatively correlated amino acids were separated into different clusters. It indicates a distinct evolutionary history or functional competition. As the evolution of amino acids may have proceeded independently in the three kingdoms of life, it is understandable that amino acids show different kinships in eukaryote and eubacteria. However, conserved linkages were observed that some amino acids are always gathered together: (A, G, H, L, P, Q, R, W) and (F, I, K, N, S, Y). This suggests that amino acids may have evolved mainly in two separate paths.

It was also found (Figure 2) that amino acids with similar codons are inclined towards having similar usage during evolution, e.g. P/R, and N/I/Y. These amino acids may have a common evolutionary origin. Accordingly, the potential evolutional paths of amino acids were drawn using the following criteria. Firstly, amino acids with a Pearson correlation coefficient above 0.8 were designated as lineal consanguinity, either paternity or fraternity. Secondly, the assignment of kinship should agree with the chronological order of amino acids. Thirdly, the transition between amino acids is favored by one-codon mutation, especially the last codon in the codon triplet. These results are illustrated in Figure 7. It can be seen that the evolutionary paths of amino acids in eukaryotes and eubacteria are not always coincident. However, two independent and conserved evolutional paths were found: A→G→R→P and K→Y.

Conclusion

Our study agrees with previous research that statistical analysis of amino acid composition in proteins is a feasible route to global understanding of the physiological

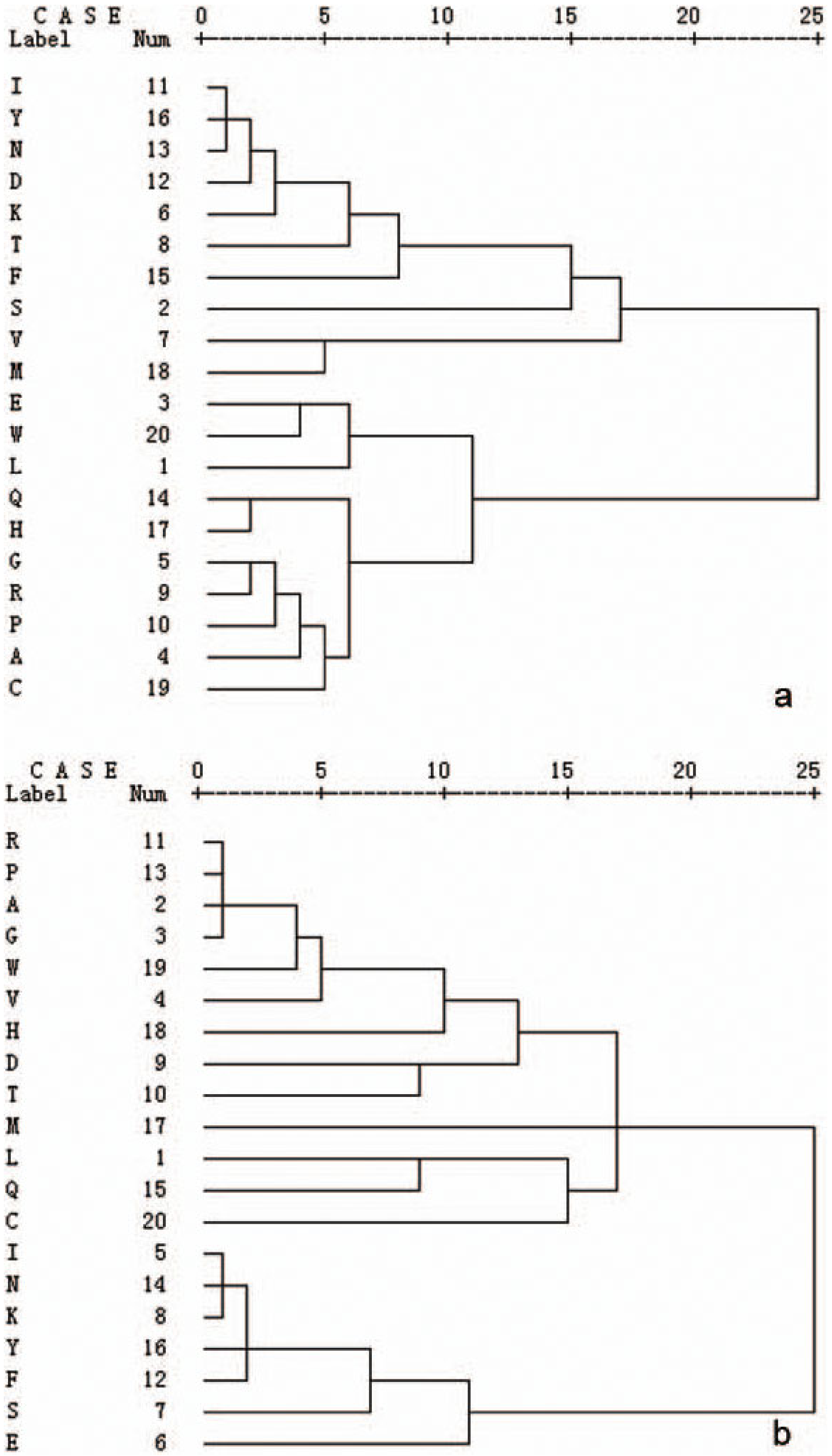
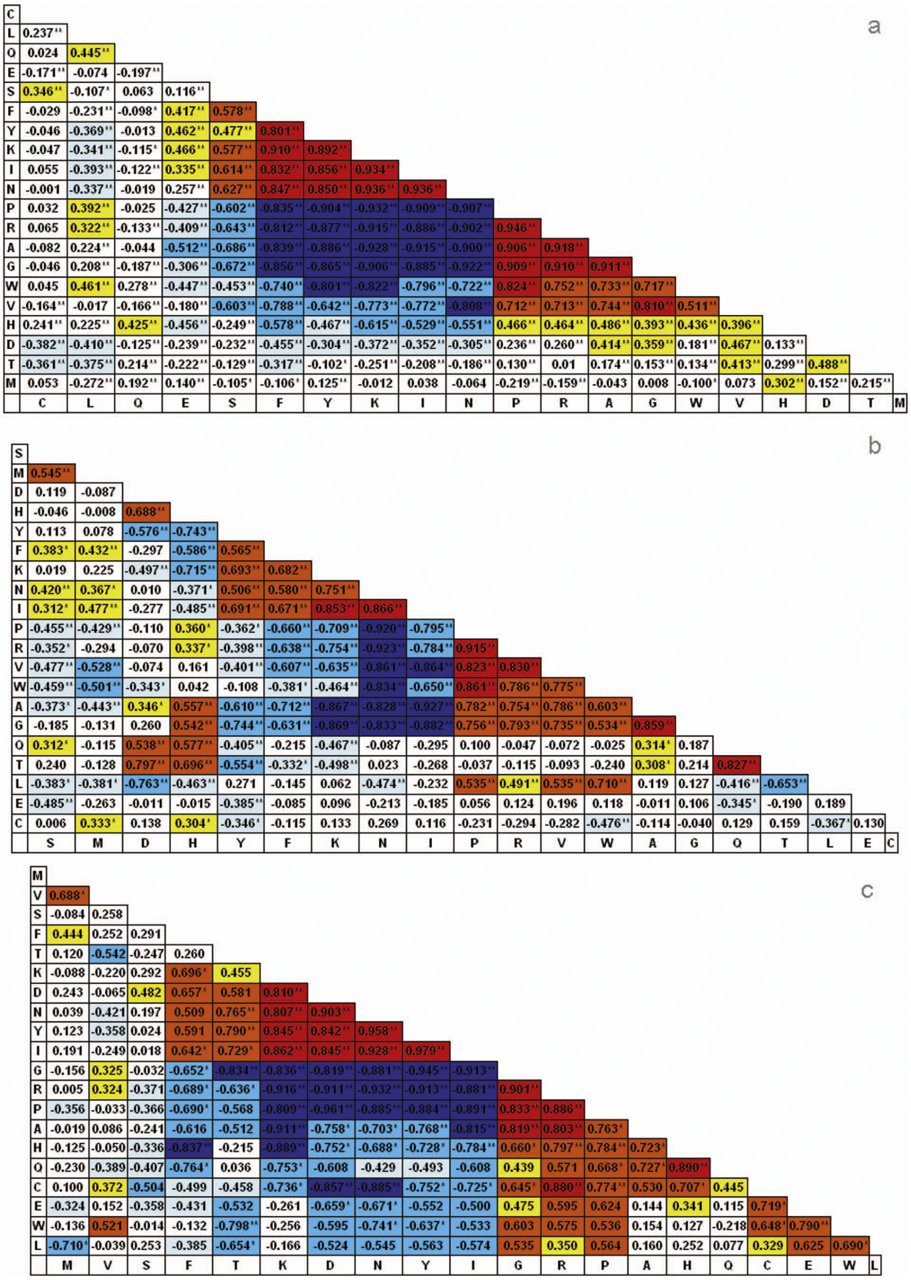


Figure 4 Kinship clusters of amino acids. Kinships of amino acids based on their frequencies over eukaryotes (4a) and eubacteria (4b). Amino acids were clustered using SPSS11.0 software by calculating the Pearson correlation coefficient of their frequencies over 10 eukaryotic organisms or 495 eubacteria. As illustrated in 4a and 4b, the 20 alpha-amino acids were divided into two clusters: cluster 1 (A, C, G, H, P, Q and R) and cluster 2 (D, I, K, N, and Y) for eukaryotes and cluster 1 (A, G, P, R, V, and W) and cluster 2 (F, K, I, N, and Y) for eubacteria. The number besides the amino acids indicates their frequency order in Figure 1.



a

C																				
L	0.237**																			
Q	0.024	0.445**																		
E	-0.171**	-0.074	-0.197**																	
S	0.346**	-0.107*	0.063	0.116**																
F	-0.029	-0.231**	-0.098*	0.417**	0.578**															
Y	-0.046	-0.369**	-0.013	0.462**	0.477**	0.801**														
K	-0.047	-0.341**	-0.115*	0.466**	0.577**	0.910**	0.892**													
I	0.055	-0.393**	-0.122**	0.335**	0.614**	0.832**	0.856**	0.934**												
N	-0.001	-0.337**	-0.019	0.257**	0.627**	0.847**	0.850**	0.936**	0.936**											
P	0.032	0.392**	-0.025	-0.427**	-0.602**	-0.835**	-0.904**	-0.932**	-0.909**	-0.907**										
R	0.065	0.322**	-0.133**	-0.409**	-0.643**	-0.812**	-0.877**	-0.915**	-0.886**	-0.902**	0.946**									
A	-0.082	0.224**	-0.044	-0.512**	-0.686**	-0.839**	-0.886**	-0.928**	-0.915**	-0.900**	0.906**	0.918**								
G	-0.046	0.208**	-0.187**	-0.306**	-0.672**	-0.856**	-0.865**	-0.906**	-0.885**	-0.922**	0.909**	0.910**	0.911**							
W	0.045	0.461**	0.278**	-0.447**	-0.453**	-0.740**	-0.801**	-0.822**	-0.796**	-0.722**	0.824**	0.752**	0.733**	0.717**						
V	-0.164**	-0.017	-0.166**	-0.180**	-0.603**	-0.788**	-0.642**	-0.773**	-0.772**	-0.808**	0.712**	0.713**	0.744**	0.810**	0.511**					
H	0.241**	0.225**	0.425**	-0.456**	-0.249**	-0.578**	-0.467**	-0.615**	-0.529**	-0.551**	0.466**	0.464**	0.486**	0.393**	0.436**	0.396**				
D	-0.382**	-0.410**	-0.125**	-0.239**	-0.232**	-0.455**	-0.304**	-0.372**	-0.352**	-0.305**	0.236**	0.260**	0.414**	0.359**	0.181**	0.467**	0.133**			
T	-0.361**	-0.375**	0.214**	-0.222**	-0.129**	-0.317**	-0.102*	-0.251**	-0.208**	-0.186**	0.130**	0.01	0.174**	0.153**	0.134**	0.413**	0.299**	0.488**		
M	0.053	-0.272**	0.192**	0.140**	-0.105*	-0.106*	0.125**	-0.012	0.038	-0.064	-0.219**	-0.159**	-0.043	0.008	-0.100*	0.073	0.302**	0.152**	0.215**	
	C	L	Q	E	S	F	Y	K	I	N	P	R	A	G	W	V	H	D	T	M

b

S																				
M	0.545**																			
D	0.119	-0.087																		
H	-0.046	-0.008	0.688**																	
Y	0.113	0.078	-0.576**	-0.743**																
F	0.383*	0.432**	-0.297	-0.586**	0.565**															
K	0.019	0.225	-0.497**	-0.715**	0.693**	0.682**														
N	0.420**	0.367*	0.010	-0.371*	0.506**	0.580**	0.751**													
I	0.312*	0.477**	-0.277	-0.485**	0.691**	0.671**	0.853**	0.866**												
P	-0.455**	-0.429**	-0.110	0.360*	-0.362*	-0.660**	-0.709**	-0.920**	-0.795**											
R	-0.352*	-0.294	-0.070	0.337*	-0.398**	-0.638**	-0.754**	-0.923**	-0.784**	0.915**										
V	-0.477**	-0.528**	-0.074	0.161	-0.401**	-0.607**	-0.635**	-0.861**	-0.864**	0.823**	0.830**									
W	-0.459**	-0.501**	-0.343*	0.042	-0.108	-0.381*	-0.464**	-0.834**	-0.650**	0.861**	0.786**	0.775**								
A	-0.373*	-0.443**	0.346*	0.557**	-0.610**	-0.712**	-0.867**	-0.828**	-0.927**	0.782**	0.754**	0.786**	0.603**							
G	-0.185	-0.131	0.260	0.542**	-0.744**	-0.631**	-0.869**	-0.833**	-0.882**	0.756**	0.793**	0.735**	0.534**	0.859**						
Q	0.312*	-0.115	0.538**	0.577**	-0.405**	-0.215	-0.467**	-0.087	-0.295	0.100	-0.047	-0.072	-0.025	0.314*	0.187					
T	0.240	-0.128	0.797**	0.696**	-0.554**	-0.332*	-0.498**	0.023	-0.268	-0.037	-0.115	-0.093	-0.240	0.308*	0.214	0.827**				
L	-0.383*	-0.381*	-0.763**	-0.463**	0.271	-0.145	0.062	-0.474**	-0.232	0.535**	0.491**	0.535**	0.710**	0.119	0.127	-0.416**	-0.653**			
E	-0.485**	-0.263	-0.011	-0.015	-0.385**	-0.085	0.096	-0.213	-0.185	0.056	0.124	0.196	0.118	-0.011	0.106	-0.345*	-0.190	0.189		
C	0.006	0.333*	0.138	0.304*	-0.346*	-0.115	0.133	0.269	0.116	-0.231	-0.294	-0.282	-0.476**	-0.114	-0.040	0.129	0.159	-0.367*	0.130	
	S	M	D	H	Y	F	K	N	I	P	R	V	W	A	G	Q	T	L	E	C

c

M																				
V	0.688*																			
S	-0.084	0.258																		
F	0.444	0.252	0.291																	
T	0.120	-0.542	-0.247	0.260																
K	-0.088	-0.220	0.292	0.696*	0.455															
D	0.243	-0.065	0.482	0.657*	0.581	0.810**														
N	0.039	-0.421	0.197	0.509	0.765**	0.807**	0.903**													
Y	0.123	-0.358	0.024	0.591	0.790**	0.845**	0.842**	0.958**												
I	0.191	-0.249	0.018	0.642*	0.729*	0.862**	0.845**	0.928**	0.979**											
G	-0.156	0.325	-0.032	-0.652*	-0.834**	-0.836**	-0.819**	-0.881**	-0.945**	-0.913**										
R	0.005	0.324	-0.371	-0.689*	-0.636*	-0.916**	-0.911**	-0.932**	-0.913**	-0.881**	0.901**									
P	-0.356	-0.033	-0.366	-0.690*	-0.568	-0.809**	-0.961**	-0.885**	-0.884**	-0.891**	0.833**	0.886**								
A	-0.019	0.086	-0.241	-0.616	-0.512	-0.911**	-0.758*	-0.703*	-0.768**	-0.815**	0.819**	0.803**	0.763*							
H	-0.125	-0.050	-0.336	-0.837**	-0.215	-0.889**	-0.752*	-0.688*	-0.728*	-0.784**	0.660*	0.797**	0.784**	0.723*						
Q	-0.230	-0.389	-0.407	-0.764*	0.036	-0.753*	-0.608	-0.429	-0.493	-0.608	0.439	0.571	0.668*	0.727*	0.890**					
C	0.100	0.372	-0.504	-0.499	-0.458	-0.736*	-0.857**	-0.885**	-0.752*	-0.725*	0.645*	0.880**	0.774**	0.530	0.707*	0.445				
E	-0.324	0.152	-0.358	-0.431	-0.532	-0.261	-0.659*	-0.671*	-0.552	-0.500	0.475	0.595	0.624	0.144	0.341	0.115	0.719*			
W	-0.136	0.521	-0.014	-0.132	-0.798**	-0.256	-0.595	-0.741*	-0.637*	-0.533	0.603	0.575	0.536	0.154	0.127	-0.218	0.648*	0.790**		
L	-0.710*	-0.039	0.253	-0.385	-0.654*	-0.166	-0.524	-0.545	-0.563	-0.574	0.535	0.350	0.564	0.160	0.252	0.077	0.329	0.625	0.690*	
	M	V	S	F	T	K	D	N	Y	I	G	R	P	A	H	Q	C	E	W	L

Figure 5 Triangle tables of correlation coefficients between amino acids. Triangle tables of correlation coefficients between amino acids in kingdoms of eubacteria (5a), archaebacteria (5b), and eukaryotes (5c). The correlations between 20 amino acids were determined by calculating Pearson correlation coefficients (*r*) of amino acid frequencies over 495 eubacteria, 44 archaebacteria, and 10 eukaryotic representatives respectively. Each correlation between a pair of amino acids was colored: red (*r* > 0.8), orange (0.5 <*r* < 0.8), yellow (0.3 <*r* < 0.5), blue (*r* < -0.8), light blue (-0.8 <*r* < -0.5), light green (-0.5 <*r* < -0.3) and white for others. The significance (2-tailed) was also listed in the table: ** indicates that the correlation is significant at the 0.05 level (2-tailed), and * indicates that the correlation is significant at the 0.01 level (2-tailed).

function of living organisms and the mystery encoded in whole genomes. However, proper evaluation of "real" amino acid usage in a modern taxa may be affected by a series of factors, including, time scale of evolution, frequency of organism generation, diverse living environments, chronological order of amino acid appearance, bias of genetic codes, gene mutation frequency, mutation-selection equilibrium, preference of physico-chemical properties, difficulty of biosynthesis, co-evolution of amino acids and genetic codes, incomplete annotation of genomes, existence of "retired" genes and pseudogenes in genomes, and other as yet unrecognized reasons. Many

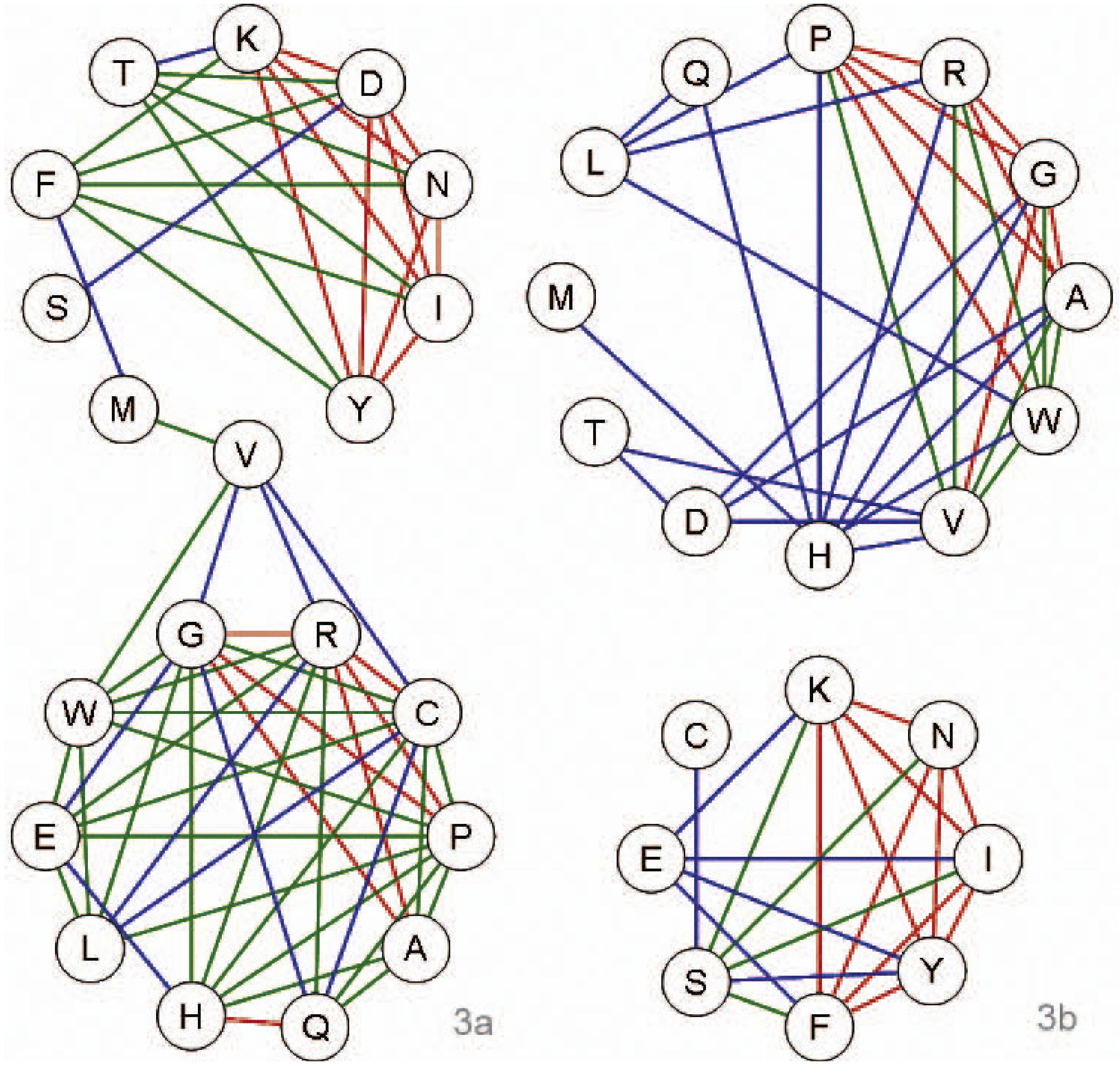


Figure 6 Correlated clusters of amino acids. Clusters determined by correlation analyses of amino acid composition in eukaryotes (6a) and eubacteria (6b). Amino acids with correlated frequencies are connected by lines and colored according to Pearson correlation coefficients (*r*): red ($r > 0.8$), green ($0.5 < r < 0.8$), and blue ($0.3 < r < 0.5$). It is noted that 20 alpha-amino acids group into two clusters: cluster 1 (D, F, I, K, M, N, S, T, V and Y) and cluster 2 (A, C, E, G, H, L, P, Q, R and W) for eukaryotes (6a) and cluster 1 (A, D, G, H, L, M, P, Q, R, T, V and W) and cluster 2 (C, E, F, I, K, N, S and Y) for eubacteria (6b). Amino acids in the same cluster are suggested to have common evolutionary history.

of these factors are currently unpredictable and incalculable and thus have been ignored in this study. It can be concluded that statistical observation of amino acid composition in modern proteomes is an alternative means for broadening our current knowledge on the origin of life.

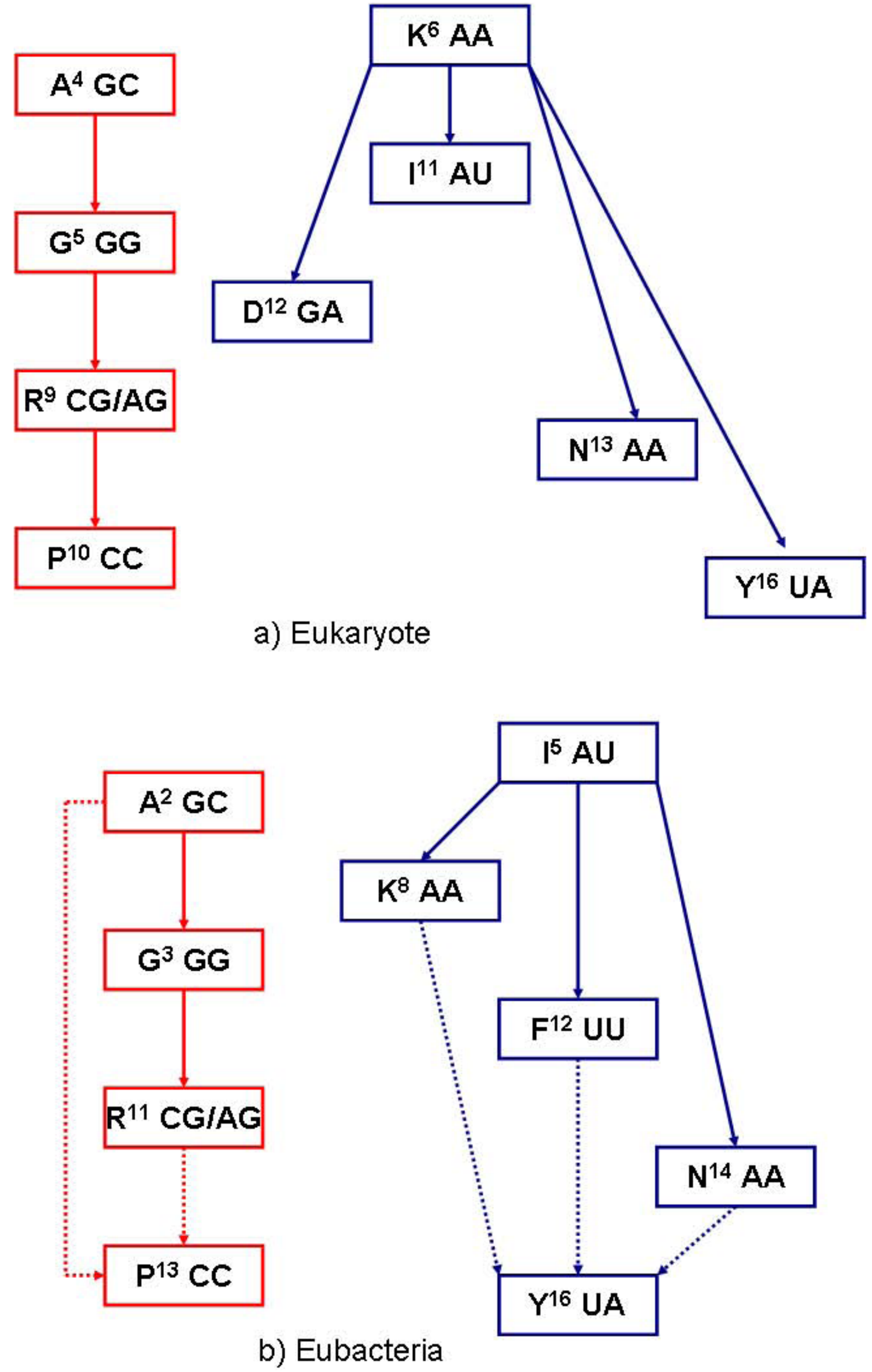
Methods

Data

Whole genome information of 549 prokaryotes (including 495 eubacteria and 44 archaebacteria) and 10 eukaryotic representatives (*Saccharomyces cerevisiae, Abrabidopsis thaliana, Caenorhaditis elegans, Drosophila melanogaster, Apis mellifera, Danio rerio, Gullus gallus, Mus musculus, Pan troglodytes, and Homo sapiens*) were derived from NCBI genome resource. Taxonomy of these selected organisms, their unique NCBI entry IDs and annotation versions were listed in the Additional File 1.

Methods

The composition of amino acids in each genome was measured by calculating the frequencies of amino acids



a) Eukaryote

b) Eubacteria

Figure 7 Putative evolutionary paths of amino acids. Putative evolutionary paths of amino acids based on data of eukaryotes (7a) and eubacteria (7b). Each block contains an amino acid, its ranking of frequency (the number beside the amino acid), and first two characters of its codons. The continuous arrow indicates the direction of evolution and the dotted arrow indicates more than one possible direction of evolution. It seems that new amino acids tend to derive from "parent" amino acids by one-base codon change. It is noted that evolutionary paths of amino acids in eukaryotes (7a) and eubacteria (7b) are not exactly the same, however, conserved evolutionary paths are observed: A→G→R→P and K→Y.

against all open reading frames (ORFs) in the whole genome. The frequency of each amino acid was determined by

$$Fi = \frac{N_i}{N_{total}} \tag{1}$$

where N_i is the number of amino acid i in genome ORFs, and N_{total} is the sum of all 20 amino acids in genome ORFs. This calculation is subjected to the assumption that the ORF assignments in selected genomes are correct. The Pearson correlation coefficients (r) of amino acids frequency data from the three kingdoms was computed by Bivariate Correlations procedure of software SPSS 13.0, at significance level 0.01 and 0.05.

$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E((X-\mu_X)(Y-\mu_Y))}{\sigma_X \sigma_Y} \tag{2}$$

Where X and Y are the two random amino acids; $\rho_{X,Y}$ is the correlation coefficient between X and Y; μ_X and μ_Y are expected values; σ_X and σ_Y are standard deviations; E is the expected value operator; cov means covariance.

Additional file 1: Table S1. Frequencies of 20 alpha-amino acids in 495 eubacteria, 44 archaebacteria and 10 eukaryotic representatives. The archaebacteria were prefixed with "*". NCBI entry IDs and annotation versions were provided with taxa names. The correlation coefficients between "random" amino acid frequencies following from uniform codon usage and amino acid compositions of the modern organisms were also given.

Additional file 2: Table S2. Average frequencies of genetic codes for basic or acidic amino acids in three kingdoms of life. Average frequencies of genetic codes for basic or acidic amino acids in kingdoms of archaebacteria, eubacteria, and eukaryotes respectively. The basic amino acids are Histidine (H), Lysine (K), and Arginine (R), and the acidic amino acids are Aspartic acid (D) and Glutamic acid (E). AFGC stands for Average frequency of genetic codes.

Additional file 3: Table S3. Frequencies of 64 genetic codes in 495 eubacteria, 44 archaebacteria and 10 eukaryote representatives.

Additional file 4: Table S4. The GC% of both coding regions and non-coding regions in the whole genomes of eight organisms (*S. cerevisiae, A. thaliana, C. elegans, D. melanogaster, A. mellifera, D. rerio, M. musculus, and H. sapiens*).

Acknowledgements
The project is supported by the National Natural Science Foundation of China (NO. 20572061 and No.20732004). Support from the Program for New Century Excellent Talents in University (NCET) of MOE (2006 to Ji ZL) is gratefully acknowledged.

Author details
[1]The Key Laboratory for Chemical Biology of Fujian Province, Department of Chemistry, College of Chemistry and Chemical Engineering, Xiamen University, Xiamen 361005, Fujian, PR China. [2]School of Life Science, Xiamen University, Xiamen 361005, Fujian, PR China. [3]The Key Laboratory for Bioorganic Phosphorus Chemistry and Chemical Biology, Ministry of Education, Department of Chemistry, School of Life Sciences and Engineering, Tsinghua University, Beijing 100084, PR China.

Authors' contributions
ZJ is the designer and supervisor of this project. YZ is the co-supervisor of this project and helped to revise the manuscript. XL is the major researcher and prepared the manuscript. JZ, FN, XD, BH, DH were involved in data collection and analyses at different stages of this project. All authors have read and approved the final manuscript.

Received: 15 February 2009 Accepted: 15 March 2010
Published: 15 March 2010

References

1. Jordan IK, Kondrashov FA, Adzhubei IA, Wolf YI, Koonin EV, Kondrashov AS, Sunyaev S: **A universal trend of amino acid gain and loss in protein evolution.** *Nature* 2005, **433(7026)**:633-638.
2. Hurst LD, Feil EJ, Rocha EP: **Protein evolution: causes of trends in amino-acid gain and loss.** *Nature* 2006, **442(7105)**:E11-12, discussion E12.
3. Swire J, Judson OP, Burt A: **Mitochondrial genetic codes evolve to match amino acid requirements of proteins.** *J Mol Evol* 2005, **60(1)**:128-139.
4. Brooks DJ, Fresco JR, Lesk AM, Singh M: **Evolution of amino acid frequencies in proteins over deep time: inferred order of introduction of amino acids into the genetic code.** *Mol Biol Evol* 2002, **19(10)**:1645-1655.
5. Brooks DJ, Fresco JR, Singh M: **A novel method for estimating ancestral amino acid composition and its application to proteins of the Last Universal Ancestor.** *Bioinformatics* 2004, **20(14)**:2251-2257.
6. Brooks DJ, Fresco JR: **Increased frequency of cysteine, tyrosine, and phenylalanine residues since the last universal ancestor.** *Mol Cell Proteomics* 2002, **1(2)**:125-131.
7. Biro JC, Benyo B, Sansom C, Szlavecz A, Fordos G, Micsik T, Benyo Z: **A common periodic table of codons and amino acids.** *Biochemical and biophysical research communications* 2003, **306(2)**:408-415.
8. Graur D: **Amino acid composition and the evolutionary rates of protein-coding genes.** *J Mol Evol* 1985, **22(1)**:53-62.
9. Tourasse NJ, Li WH: **Selective constraints, amino acid composition, and the rate of protein evolution.** *Mol Biol Evol* 2000, **17(4)**:656-664.
10. Chirpich TP: **Rates of protein evolution: a function of amino acid composition.** *Science* 1975, **188(4192)**:1022-1023.
11. Julenius K, Pedersen AG: **Protein evolution is faster outside the cell.** *Mol Biol Evol* 2006, **23(11)**:2039-2048.
12. Chou KC: **Using pair-coupled amino acid composition to predict protein secondary structure content.** *J Protein Chem* 1999, **18(4)**:473-480.
13. Pilizota T, Lucic B, Trinajstic N: **Use of variable selection in modeling the secondary structural content of proteins from their composition of amino acid residues.** *J Chem Inf Comput Sci* 2004, **44(1)**:113-121.
14. Lee S, Lee BC, Kim D: **Prediction of protein secondary structure content using amino acid composition and evolutionary information.** *Proteins* 2006, **62(4)**:1107-1114.
15. Jukes TH, Holmquist R, Moise H: **Amino acid composition of proteins: Selection against the genetic code.** *Science* 1975, **189(4196)**:50-51.
16. Raghava GP, Han JH: **Correlation and prediction of gene expression level from amino acid and dipeptide composition of its protein.** *BMC Bioinformatics* 2005, **6**:59.
17. Bogatyreva NS, Finkelstein AV, Galzitskaya OV: **Trend of amino acid composition of proteins of different taxa.** *J Bioinform Comput Biol* 2006, **4(2)**:597-608.
18. Wang GZ, Ma BG, Yang Y, Zhang HY: **Unexpected amino acid composition of modern Reptilia and its implications in molecular mechanisms of dinosaur extinction.** *Biochem Biophys Res Commun* 2005, **333(4)**:1047-1049.
19. Tekaia F, Yeramian E, Dujon B: **Amino acid composition of genomes, lifestyles of organisms, and evolutionary trends: a global picture with correspondence analysis.** *Gene* 2002, **297(1-2)**:51-60.
20. Cutter AD, Wasmuth JD, Blaxter ML: **The evolution of biased codon and amino acid usage in nematode genomes.** *Mol Biol Evol* 2006, **23(12)**:2303-2315.
21. Miller SL: **A production of amino acids under possible primitive earth conditions.** *Science* 1953, **117(3046)**:528-529.
22. Trifonov EN: **Consensus temporal order of amino acids and evolution of the triplet code.** *Gene* 2000, **261(1)**:139-151.
23. Ota T, Kimura M: **Amino acid composition of proteins as a product of molecular evolution.** *Science* 1971, **174(5)**:150-153.

24. Gu X, Hewett-Emmett D, Li WH: **Directional mutational pressure affects the amino acid composition and hydrophobicity of proteins in bacteria.** *Genetica* 1998, **102-103(1-6)**:383-391.
25. Banerjee T, Gupta SK, Ghosh TC: **Role of mutational bias and natural selection on genome-wide nucleotide bias in prokaryotic organisms.** *Biosystems* 2005, **81(1)**:11-18.
26. Jimenez-Sanchez A: **On the origin and evolution of the genetic code.** *Journal of molecular evolution* 1995, **41(6)**:712-716.
27. Ikehara K, Omori Y, Arai R, Hirose A: **A novel theory on the origin of the genetic code: a GNC-SNS hypothesis.** *Journal of molecular evolution* 2002, **54(4)**:530-538.

doi:10.1186/1471-2148-10-77
Cite this article as: Liu *et al.*: Genome wide exploration of the origin and evolution of amino acids. *BMC Evolutionary Biology* 2010 **10**:77.