Coevolution study of mitochondria respiratory chain proteins: 
Toward the understanding of protein–protein interaction

Ming Yang a,b, Yan Ge c, Jiayan Wu a, Jingfa Xiao a,*, Jun Yu a,*

a CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100029, China
b Graduate University of Chinese Academy of Sciences, Beijing 100049, China
c College of Biology, China Agricultural University, Beijing 100091, China

Received 20 January 2011; revised 8 April 2011; accepted 9 April 2011

Abstract

Coevolution can be seen as the interdependency between evolutionary histories. In the context of protein evolution, functional correlation proteins are ever-present coordinated evolutionary characters without disruption of organismal integrity. As to complex system, there are two forms of protein–protein interactions in vivo, which refer to inter-complex interaction and intra-complex interaction. In this paper, we studied the difference of coevolution characters between inter-complex interaction and intra-complex interaction using “Mirror tree” method on the respiratory chain (RC) proteins. We divided the correlation coefficients of every pairwise RC proteins into two groups corresponding to the binary protein–protein interaction in intra-complex and the binary protein–protein interaction in inter-complex, respectively. A dramatical discrepancy is detected between the coevolution characters of the two sets of protein interactions (Wilcoxon test, p-value = 4.4 × 10^{-6}). Our finding reveals some critical information on coevolutionary study and assists the mechanical investigation of protein–protein interaction. Furthermore, the results also provide some unique clue for supramolecular organization of protein complexes in the mitochondrial inner membrane. More detailed binding sites map and genome information of nuclear encoded RC proteins will be extraordinary valuable for the further mitochondria dynamics study.

Keywords: Coevolution; Respiratory chain proteins; “Mirror tree” method; Supercomplex; Protein–protein interaction

1. Introduction

The protein evolutionary process is affected by many factors, such as temperature, gene localization in the genome, gene expression, and function of proteins. Due to selection pressure, a change in one protein would necessitate compensatory changes in others (Pazos et al., 1997; Jespers et al., 1999; Goh et al., 2000; Fraser et al., 2004), otherwise, the interaction among proteins is lost as well as its function. This evolutionary process we called “coevolution”, which is critical to almost all biological processes, such as metabolic pathways, signaling cascades and transcription control networks, undergoes adaptive or constructive change without disruption of organism integrity (Pazos and Valencia, 2001; Fraser et al., 2004). Functional correlation proteins (i.e., proteins involve the same metabolic pathway or biological process, or proteins belong to the same structural complex or molecular machine) are ever-present coordinated evolutionary characters.

In a living cell, proteins mainly combine with other protein(s) to form protein complex to carry out their functions, especially for structural proteins. As to protein complex system, there are two forms of protein–protein interactions (Phizicky and Fields, 1995), which refer to binary protein–protein interaction in intra-complex and the binary protein–protein interaction in inter-complex. However, people tend to leave out the existence of the two kinds of interactions in their coevolutional proteins detecting. They prefer to employ one
parametric criterion to capture protein coevolution characters. Actually, it is an inappropriate performance, especially when the knowledge of protein–protein locations is lacking. In this work we intend to reveal the difference of protein coevolution characters under the two different forms of protein–protein interactions, based on vertebrate mitochondria self-encoding respiratory chain (RC) proteins training set. The relation of these proteins can be divided into two forms, intra-complex and inter-complex interactions. We aim to uncover whether protein–protein location affects their coevolutionary extent. This analysis may reveal some critical information on coevolution characters and provide new guidance on protein–protein interactions in the future.

2. Materials and methods

2.1. Data set

The data of vertebrate mitochondria self-encoding RC proteins are downloaded from NCBI database (June 24, 2009 update). These data are extracted from 267 species’ genomes. These species’ mitochondria encode 13 proteins, all of which are subunits of the respiratory chain complexes. These protein sequences are listed in Supplementary files.

2.2. Coevolution model

The sequence alignment of each homologous sequence cluster is performed by ClustalW program with default settings (Thompson et al., 1994). The distance matrix is calculated based on the sequence alignment and the phylogenetic trees of each protein family are generated by MEGA program (Tamura et al., 2007). Then the coevolutionary analysis is applied by “Mirror tree” method (Goh et al., 2000; Pazos and Valencia, 2001). The coevolutionary correlation coefficient \( r \) of the protein family trees is calculated with employing Pearson’s correlation coefficient (Press et al., 1992). The Pearson’s correlation coefficient is defined as:

\[
\begin{align*}
\rho &= \frac{\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (X_{ij} - \bar{X})(Y_{ij} - \bar{Y})}{\sqrt{\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (X_{ij} - \bar{X})^2 \sqrt{\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (Y_{ij} - \bar{Y})^2}}}.
\end{align*}
\]

In Equation (1), \( N \) is equal to the number of sequences in the multiple sequences alignments, \( \bar{X} \) and \( \bar{Y} \) is the mean of all \( X_{ij} \) values and all \( Y_{ij} \) values, respectively. \( X_{ij} \) is the pairwise distance between sequence \( i \) and sequence \( j \) of one protein family. \( Y_{ij} \) is defined the same as \( X_{ij} \).

2.3. Statistical analysis

The significance of the \( r \) value is assessed by bootstrap correlation analysis (Efron, 1979). In the bootstrap analysis, 1000 sets containing \( N \) pairwise distances are generated randomly drawn from the \( N \) pairwise distances in the original set. For every such set we computed the bootstrap correlation coefficient \( r_{\text{rand}} \). The \( p \)-value, which represents the probability of getting the observed \( r \) value by chance, is obtained from Equation (2).

\[
p = \frac{\text{erfc} \left( \frac{r - r_{\text{rand}}}{\sigma_{\text{rand}}} \right)}{\sqrt{2}}
\]

where \( r_{\text{rand}} \) is the mean of 1000 values of \( r_{\text{rand}} \) and \( \sigma \) is the standard deviation of \( r_{\text{rand}} \) in Equation (2).

In order to analyze the effect of protein interaction to the coevolution, the correlation coefficient is divided into two groups, one is the \( r \) values of protein pairs in one complex and the other is \( r \) values of protein pairs in different complexes. Ultimately Wilcoxon test in R software is used to test the significant difference between the two group values (Team RDC, 2009). In order to further analyze the possible supramolecular organization of protein complexes in the mitochondrial inner membrane, the amino acid composition method and Support Vector Machine (SVM) are used for classifying the RC protein pair as interacting or non-interacting with default parameters. SVM is an algorithm embodied in the webserver Proprint (ProPrInt, http://www.imtech.res.in/raghava/proprint/index.html). Proprint is a special website to predict protein–protein interaction, and the training set is from the PPI database provided on the website.

3. Results and discussion

3.1. The choice of coevolution model

Protein coevolution studies have been previously attempted for more than a decade (Altschuh et al., 1987). Various computational methods have been proposed to detect coevolution of proteins, such as phylogenetic profiles methods (Pellegrini et al., 1999), approaches based on protein coexpression patterns (Fraser et al., 2004; Ettwiller and Veitia, 2007), Bayesian methods (Dinnic et al., 2005; Burger and van Nimwegen, 2008), augmented continuous-time Markov process (Yeang and Haussler, 2007) and functional distances correlation analysis (e.g., “Mirror tree” method) (Goh et al., 2000; Pazos and Valencia, 2001; Jothi et al., 2005; Craig and Liao, 2007). After a comprehensive analysis of the favorable conditions and restraining factors of various methods, “Mirror tree” method is considered to be able to portray coevolution characters best. Thus we employ “Mirror tree” method here. The similarity among the phylogenetic trees of all possible pairs of proteins (or domains) is interpreted as an indication of their coordinated evolution and a direct consequence of the similar evolutionary pressure. And the extent of coevolution for each pair can be determined by measuring the correlation of their underlying distance matrices of phylogenetic trees (Pazos and Valencia, 2001). The “Mirror tree” method needs a considerable amount of homologous protein sequences to calculate the distance matrix. Both matrices must contain distances between the same numbers of homologous proteins, from the same set of species.
3.2. The sensitivity of “Mirror tree” method

In order to verify the sensitivity of this method to capture coevolution character, cyclin-dependent kinases (CDK) family is involved in the preliminary model system. CDK possess two independent domains (i.e., N-terminal domain and C-terminal domain) with the enzyme active site formed by the interface between the two domains (Fig. 1A). According to the coevolution theory, the two domains must coevolve to maintain the enzyme activity. Any change in the N-terminal domain which affects the activity of the enzyme must be selected against or subsequently compensated for a correlated change in the C-terminal domain, and vice versa.

One hundred and ten CDK sequences are retrieved from UniProt database. Each sequence is cut off to get N-terminal domain sequence and C-terminal domain sequence based on structure alignment information. The 3D structures of CDK family are got directly from PDB database and the 3D structures of CDK4, CDK8, CDK9, CDK10 and CDK11 are built using MODELLER program (Sali and Blundell, 1993; Eswar et al., 2006). The short linking regions of these proteins, which are not directly involved in forming the active site, are left out of the two domains. The N and C-terminal domain trees are built based on multiple sequence alignment (Fig. 1B), and the correlation coefficients of two domain trees are calculated. The correlation coefficient for the two trees is 0.92 ($p$-value $< 2.2 \times 10^{-16}$), which indicates the evolution of the C-terminal domain and N-terminal domain is highly correlated. To exclude the probability of false positive, we recalculated the correlation coefficients using randomly chosen incorrect pairings between the two domains. The correlation coefficient among the trees of these non-binding pairs is 0.02, which means these pairs are scarcely correlated. The results of CDK coevolution research indicate that the “Mirror tree” method has the capability to quantitatively recognize the coevolutionary parts.

3.3. The coevolution of the RC proteins

The mitochondrial respiratory chain located in the inner mitochondrial membrane ensures the process of electron transport from reducing equivalents (e.g., NADH, succinate) to molecular oxygen with a very large loss of free energy. Much of the mitochondrial respiratory chain is conserved in the process of oxidative phosphorylation (Rustin et al., 1994). The vertebrate RC consists of five multimeric protein complexes: complex I (reduced nicotinamide adenine dinucleotide (NADH) dehydrogenase—ubiquinone oxidoreductase, approximately 46 subunits), complex II (succinate dehydrogenase—ubiquinone oxidoareductase, 4-8 subunits), complex III (ubiquinone—cytochrome c oxidoreductase, 11 subunits), complex IV (cytochrome c oxidase, 13 subunits), and complex V (ATP synthase, approximately 16 subunits). The RC also requires two small electron carriers, ubiquinone (coenzyme Q10) and cytochrome c. The RC complexes are encoded by both the mitochondrial and nuclear genome. Due to a limited amount of sequenced vertebrate genomes, we only collected mitochondria self-encoding RC proteins in our research. These proteins (13 in all) are located in four respiratory multimeric protein complexes—7 subunits of NADH dehydrogenase—ubiquinone oxidoreductase (complex I), 1 subunit of ubiquinone—cytochrome c oxidoreductase (complex III), 3 subunits of cytochrome c oxidase (complex IV) and 2 subunits of ATP synthase (complex V) (Andersson et al., 2003; DiMauro and Schon, 2003).

Fig. 1. Coevolution analysis of N-terminal domain and C-terminal domain of CDK protein. A: the 3D structure of the human CDK2 (PDB 1w8c). It is mainly composed of two domains: small N-terminal domain (residues 1-81) in red and large C-terminal domain (residues 86-298) in green. A ligand is shown as a stick model, located deep within the active-site cleft between the two domains; B: the phylogenetic trees of the N-terminal and C-terminal domains of CDK. There are 11 groups of proteins represented by different colors.
The phylogenetic trees of RC proteins are generated and the correlation coefficient of each pair is also calculated. Here we take the phylogenetic trees of ATP6 and ATP8 for example (Fig. 2). The similarity of the trees’ shapes indicates ATP6 and ATP8 sustain analogous evolutionary pressures. The correlation coefficient \( r \) value of each pair between mitochondria self-encoding RC proteins is shown in Table 1 and the \( p \)-value of each \( r \) value is less than \( 2.2 \times 10^{-16} \). To better visualize the relationship of these values of correlation coefficient, a heatmap is drawn up (Fig. 3A). It should be noted that ND1, ND2, ND3, ND4, ND4L, ND5 and ND6 (ND stands for NADH dehydrogenase) are subunits of complex I; CYTB (cytochrome b) is sole in complex III; COX1, COX2 and COX3 (COX represents cytochrome c oxidase) jointly reside in complex IV; ATP6 and ATP8 belong to complex V. The maximum value of correlation coefficient between ND4L and ND6 is 0.93, demonstrating that the two subunits preserve strong function dependence. The minimal \( r \) value between COX1 and ATP8 is 0.58, which indicates that the protein may only display weak function correlation. In general, these proteins all have sharp coevolution characters. Comparatively, the coevolution degree between COX1 and other 12 proteins is lower, which shows that COX1 tends to be rather conservative. The potential reason is that COX1 is the essential subunit in the process of complex III assembly, and also forms the catalytic core of the complex (Fontanesi et al., 2008). This may be the reason why COX1 emerges weak correlation with other proteins, even though they jointly participate in one metabolic pathway.

To address the issue of high rate of false positives with the “Mirror tree” method, the opinion is widely acceptable that if a set of proteins belong to the same KEGG-base network, then the minimal \( r \) value between COX1 and ATP8 is 0.58, which indicates that the protein may only display weak function correlation. In general, these proteins all have sharp coevolution characters. Comparatively, the coevolution degree between COX1 and other 12 proteins is lower, which shows that COX1 tends to be rather conservative. The potential reason is that COX1 is the essential subunit in the process of complex III assembly, and also forms the catalytic core of the complex (Fontanesi et al., 2008). This may be the reason why COX1 emerges weak correlation with other proteins, even though they jointly participate in one metabolic pathway.

Using the coevolution character estimation to evaluate protein–protein interaction has its own limitation. For instance, interactions between free proteins and complexes and interactions among different complexes may only occur under specific conditions, which mean their coevolution characters are more complicated and difficult to identify. The scope of \( r \) values of proteins consisting in inter-complexes is wide, from 0.58 to 0.93, underlined numbers listed in Table 1. So it poses a big challenge to identify the coevolution characters effectively using current proteins coevolution predicted methods.

The selection pressure of intra-complex is also variation distinctly. For example, the \( r \) values of subunits in complex I vary from 0.80 to 0.93, which maybe correspond to these subunits’ relative locations. This opinion needs experimental interaction information to verify. However, knowledge on how subunits organize in complex remains to be elucidated. The known information for the organization of mitochondrial respiratory chain complex is only that all the 7 mitochondrial
encoded subunits of complex I reside in the membrane arm (Radermacher et al., 2006).

3.4. The evidence for formation of RC supercomplexes

The protein complexes of the mitochondrial oxidative phosphorylation system are recently reported to form supramolecular assemblies termed respiratory supercomplexes or respirasomes, such as supercomplex I1III2 (i.e., a supercomplex including complex I and dimeric complex III) from Arabidopsis thaliana and Bos Taurus, supercomplex I1II2IV2 (formed by monomeric complex I, dimeric complex II and two copies of complex IV) from Saccharomyces cerevisiae, and supercomplex I1III2IV1 (consisting of complex I, dimeric complex III, and monomeric complex IV) from B. Taurus (Vonck and Schafer, 2009). The supramolecular organization of the RC proteins plays an important function, which could enhance electron transfer rates and increase the protein insertion capacity of the inner mitochondrial membrane (Boekema and Braun, 2007). Based on single particle electron microscopy and X-ray crystallography results, the distribution of subunits in these supercomplexes is complicated in despite of the binding sites among these peptide chains (Vonck and Schafer, 2009). The r values in our study are exceptionally high in CYTB with ND1, ND5, COX1, COX2, and COX3, (Table 1 and Fig. 3A) but they distribute in different

![Fig. 3. The distribution of correlation coefficient values of paired mitochondria self-encoding RC proteins. A: heatmap of correlation coefficient values of paired mitochondria self-encoding RC proteins. It should be noted that ND1, ND2, ND3, ND4, ND4L, ND5 and ND6 are subunits of complex I; CYTB is sole in complex III; COX1, COX2 and COX3 jointly reside in complex IV; ATP6 and ATP8 belong to complex V; B: distribution of correlation coefficient values of inter-complex interaction and intra-complex interaction.](image-url)
complexes. If we considered the $r$ values as the only standard, all these proteins would have been classified into the same one complex. Coincidentally, this presumption is accordant with the fact that the proteins are really assembled into supercomplexes (Schägger and Pfeiffer, 2000, 2001; Dudkina et al., 2005). We believe that CYTB, ND1 and ND5 are highly correlated to binding sites inside supercomplexes consisting of complex I and complex III; CYTB, COX1, COX2, COX3 are also critical for the formation of supercomplexes including complex III and complex IV. To further verify our opinion for the supercomplex organization, the Amino Acid composition and Support Vector Machine (SVM) are used for classifying the protein—protein pair as interacting or non-interacting. In general, the higher SVM score represents the high probability for the protein pair interaction. The SVM scores of CYTB with other 12 proteins are shown in Fig. 4. CYTB has higher SVM scores with subunits of complex I (ND1, ND5) and complex IV (COX1, COX2 and COX3) than with subunits of complex V (ATP6 and ATP8), which are in accordance with the corresponding $r$ values in Fig. 3. Besides, the formation of supercomplex is universal phenomenon in the vertebrate mitochondria, according to mathematical model of the Pearson’s correlation coefficient. But, all of this information needs further validation by experimental work.

We have classified protein—protein interaction into two forms: intra-complex and inter-complex, and confirmed different coevolution characters between them. Our results also provide some unique clues for supramolecular organization of protein complexes in the mitochondrial inner membrane.

On the basis of the above, we conclude that the protein coevolution researches throw new sight on protein—protein interaction. A more detailed binding sites map and nuclear genome information encoded by nuclear RC proteins will be most valuable for the further mitochondria dynamics study.

Acknowledgments

This study was supported by the grants from the National Science Foundation of China (No. 31071163) and the National Basic Research Program (973 Program) (No. 2010CB126604), the Ministry of Science and Technology of the People’s Republic of China.

Supplementary data

Supplementary data related to this article can be found online at doi:10.1016/j.jgg.2011.04.003.

References


