Available online at www.sciencedirect.com
ScienceDirect
Journal of Genetics and Genomics 38 (2011) 201−207
JOURNAL OF GENETICS AND GENOMICS
www.jgenetgenomics.org

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 $= 4.4 \times 10^{-6}$). 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

Abbreviations: RC, respiratory chain; CDK, cyclin-dependent kinases; *r*, correlation coefficient; NADH, reduced nicotinamide adenine dinucleotide; CYTB, cytochrome b; COX, cytochrome c oxidase; SVM, support vector machine.

* Corresponding authors. Tel: +86 10 8299 5384, fax: +86 10 8299 5401 (J. Xiao); Tel: +86 10 8299 5357 (J. Yu).
E-mail addresses: xiaojingfa@big.ac.cn (J. Xiao), junyu@big.ac.cn (J. Yu).

doi:10.1016/j.jgg.2011.04.003

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:

$$r = \frac{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\left(X_{ij}-\overline{X}\right)\left(Y_{ij}-\overline{Y}\right)}{\sqrt{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\left(X_{ij}-\overline{X}\right)^2}\sqrt{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\left(Y_{ij}-\overline{Y}\right)^2}} \quad (1)$$

In Equation (1), N is equal to the number of sequences in the multiple sequences alignments, $\overline{X}$ and $\overline{Y}$ 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 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_{rand}. The p-value, which represents the probability of getting the observed r value by chance, is obtained from Equation (2).

$$p = \frac{erfc\left(\left|\frac{r-\overline{r}_{\text{rand}}}{\sigma_{\text{rand}}}\right|\right)}{\sqrt{2}} \quad (2)$$

where $\overline{r}_{\text{rand}}$ is the mean of 1000 values of r_{rand} and σ is the standard deviation of r_{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 (Dimmic et al., 2005; Burger and van Nimwegen, 2008), augmented continuous-time Markov process (Yeang and Haussler, 2007) and evolutional 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.