Genomics xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Evolutionary rate heterogeneity between multi- and single-interface hubs across human housekeeping and tissue-specific protein interaction network: Insights from proteins' and its partners' properties

Kakali Biswas[a], Debarun Acharya[a], Soumita Podder[a,b], Tapash Chandra Ghosh[a,*]

[a] Bioinformatics Centre, Bose Institute, P-1/12, C.I.T. Scheme VII M, Kolkata 700 054, India
[b] Department of Microbiology, Raiganj University, Raiganj, Uttar Dinajpur 733134, India

ARTICLE INFO

Keywords:
Tissue-specific hubs
Housekeeping hubs
Multi-interface hubs
Single-interface hubs
dN/dS ratio
Functional divergence
Conformational diversity

ABSTRACT

Integrating gene expression into protein-protein interaction network (PPIN) leads to the construction of tissue-specific (TS) and housekeeping (HK) sub-networks, with distinctive TS- and HK-hubs. All such hub proteins are divided into multi-interface (MI) hubs and single-interface (SI) hubs, where MI hubs evolve slower than SI hubs. Here we explored the evolutionary rate difference between MI and SI proteins within TS- and HK-PPIN and observed that this difference is present only in TS, but not in HK-class. Next, we explored whether proteins' own properties or its partners' properties are more influential in such evolutionary discrepancy. Statistical analyses revealed that this evolutionary rate correlates negatively with protein's own properties like expression level, miRNA count, conformational diversity and functional properties and with its partners' properties like protein disorder and tissue expression similarity. Moreover, partial correlation and regression analysis revealed that both proteins' and its partners' properties have independent effects on protein evolutionary rate.

1. Introduction

Cells are the fundamental unit of life. Except for the unicellular ones, every living organism possesses diverse types of cells adapted to perform specialized functions. The functions of each cell are mediated by the molecular machinery, of which proteins play an essential part. Proteins interact with each other and perform almost all the fundamental life processes. Such interactions involve interfaces or domains, which execute the functions of the protein. Protein domains play a crucial part in molecular evolution since these are used as structural building blocks and may create proteins with discrete functions due to exon shuffling [1–3]. The advancement in high-throughput protein interaction data helps to analyze protein functions from the network perspective. Moreover, within the whole protein interaction network, there are some small, densely linked components formed by the interactions between proteins, nucleic acids, and other small molecules, and are weakly connected to the rest of the protein-interaction network. These components are termed as modules [4]. Recent advances in discovery and revision of the proteins in modules using computational biology have enabled us to model these protein-protein interactions as a network where proteins represent the nodes with interactions as links between the nodes.

Inside the protein-protein interactions network (PPIN), proteins with a high degree of connectivity are found to be essential and are likely to perturb the PPIN upon deletion, misfunction or misregulation [5]. These proteins, named as hub, are distinct from lesser connected proteins or non-hubs and are evolutionary more conserved. Although most of the earlier studies featuring hub proteins from evolutionary perspective compared hub and non-hub proteins in PPIN, more recent studies aim at detailed analysis of hub proteins. One such study by Han et al. classified hub proteins into two groups - multi-interface hub proteins (MI or party hubs) and single interface hub proteins (SI or date hubs) based on protein domain architecture and correlated expression of the interacting partners [6]. Comparing the evolutionary rate between these MI and SI hubs revealed discrete differences—MI proteins were found to be evolutionarily more conserved than SI proteins [7], which may be mainly due to selective constraint acting on a larger region in MI proteins, as it usually possesses more interacting surfaces. Additionally, the party hubs mediate within-module interactions (intra-module), whereas date hubs integrate between modules (inter-module) [7]. However, the SI proteins acting on various modules face stronger consequences when deleted than the less pervasive densely connected MI proteins, due to their association with diverse functions [8]. Besides, a few studies have been carried out to understand the structural

* Corresponding author.
E-mail address: tapash@jcbose.ac.in (T.C. Ghosh).

https://doi.org/10.1016/j.ygeno.2017.11.006
Received 23 June 2017; Received in revised form 10 November 2017; Accepted 29 November 2017

(conformational) and functional role of these hub proteins [9,10]. The functions of a protein are mediated mainly by its structure. Although each protein is thought to possess definite three-dimensional conformation determined by its amino acid sequence, that may not be the only conformation adopted by the protein within a cellular system [11]. The magnitude of conformational diversity encompasses structural changes like fluctuation of protein's side chains and the movement of loops and secondary structures, even to the global rearrangement in protein tertiary structure [12].

Further insights into human PPIN classified topological variation based on gene expression data. Based on gene expression breadth, all genes are grouped as either tissue-specific (TS), or housekeeping (HK). Previous studies revealed many differences between the HK and TS genes in humans. Human HK genes are more compact in structure, containing shorter intron length, 5′ UTR length and coding sequence length [13]. Consistent with this, HK genes are enriched in shorter repetitive sequences such as Alu-elements, but depleted in longer repetitive sequences like Long Interspersed Nuclear Element 1 (LINE-1) elements [14]. Additionally, elucidation of evolutionary rate differences among these two groups resulted in similar findings across organisms as diverse as unicellular fungi to humans, the housekeeping genes (HK) evolve slower than tissue-specific genes (TS) [15]. Accordingly, the whole PPI network was also grouped into tissue-specific or local network and housekeeping or global network, where TS hubs (TSH) evolve faster than HK hubs (HKH). These TSH also feature longer genes, less protein expression abundance, tight regulation and greater protein intrinsic disorder content than HKH [54]. Additionally, within the PPI network, HK genes are more central and are associated with core cellular processes whereas TS genes are more peripheral with modified core cellular processes as well as regulatory and developmental functions [16–18]. However, these findings remain confounding as some TS genes are reported to evolve slower than even this HK class of genes [19–21]. To address this issue, Podder et al. classified human proteins into MI and SI counterparts and analyzed the evolutionary rate of TS and HK genes between these two groups. They found that within MI proteins, both TS- and HK-genes show similar evolutionary rates, whereas, within SI proteins, HK genes evolve slower than TS genes [10]. Furthermore, recent studies based on PPI-network properties highlights the impact of the partner proteins on proteins' evolutionary rate [16,22], as the interacting partners also contribute to the central node evolution via the domain-domain interaction [23]. Such analysis on HK- and TS-hubs revealed that interacting partners of the TSH are more conserved than HKH with diverse subcellular localization [22]. However, these studies lack detailed insights into the protein interaction network-based properties and the influence of interacting partners on the evolutionary rate. Therefore, a detailed spatially resolved analysis is required to explain the evolutionary rate variation between these two hub classes.

In this study, we delved deep into the understanding of protein evolutionary rate based on their expression breadth (whether housekeeping or tissue-specific) and the contribution of domain number (whether single or multiple) to it. We tried to identify at which level the evolutionary conservation endures. Furthermore, we sought to explore which among the two: protein's own property or partner properties influence the evolutionary rate of proteins the most.

2. Materials and methods

2.1. Retrieval of dataset

We obtained tissue specific gene expression data from EMBL-EBI expression atlas (https://www.ebi.ac.uk) for “baseline” expression where the expression level of each gene in normal and untreated conditions. Then we calculated tissue specificity index τ [24] of each gene for tissue specificity using the following formula [10]−

$$\tau = \frac{\sum_{j=1}^{\eta H}\left(1 - \left[\frac{\log_2 S_H(i,j)}{\log_2 S_H(i,max)}\right]\right)}{\eta_H - 1}$$

(where, ηH = number of human tissues examined and S_H (i, max) = highest expression signal of gene i across the ηH tissues). The value ranges from 0 to 1, where genes with τ-values close to ‘0’ are considered to be more towards housekeeping and those with τ-values close to ‘1’ are considered as tissue-specific (TS). τ = 0 represents equal expression of the gene across all tissue, i.e. housekeeping (HK) genes. We sorted our dataset according to an increasing τ values and obtained genes from extreme 20% of the population from both ends. Thereby, we obtained 1198 HK and 7767 TS genes.

2.2. Protein connectivity data retrieval and interacting domain identification

Protein-protein interaction data was obtained from BioGRID (release 3.4.130) (https://thebiogrid.org/) [25]. Genes with at least five interacting partners were considered to be highly connected or hub proteins. We obtained human protein sequences from the UCSC genome browser (http://genome.ucsc.edu). Interacting domains were retrieved from Pfam repository (http://pfam.sanger.ac.uk/) [55]. The hypothesis behind the Pfam data retrieving was that the interacting domains confer binding capability to protein regions. The cut-off values used for domain assignment are (1) e-value of alignment $e^{< 1.0 \times 10^{-4}}$; (2) domain length > 12; (3) matched sequence length > 80% of domain length [26]. In particular, single interface proteins were designated as having few interaction interfaces (two at most) and multi-interface proteins having more than two interacting interfaces [27]. The numbers of HKH_MI and HKH_SI proteins are respectively 303 and 895. The numbers of MI and SI proteins belonging to TSH PPIN are 1705 and 6062, respectively.

2.3. Estimation of evolutionary rate

The evolutionary rates of human genes were calculated by dividing non-synonymous substitution rate (dN) with synonymous substitution rate (dS). The dN and dS values were retrieved from BioMart interface of Ensembl Version 87 (http://www.ensembl.org/biomart/martview) [28] for *Homo sapiens* (GRCh37) using one to one Human-Mouse as well as Human-Chimpanzee orthologous pairs.

2.4. Prediction of miRNA targets sites and gene expression level assessment

The number of miRNA targets per gene were obtained from TargetScan (release 6.2) (http://www.targetscan.org) [29] for its more reliable data over other databases. Tissue-wise RNA-seq gene expression data was obtained from the human protein atlas [30]. Average gene expression level of HK genes was calculated by considering only those tissues where it shows higher than mean expression level calculated for all tissues. Expression level for TS genes represents only the tissue where the desired gene is expressed at its highest level.

2.5. Collection of conformational and functional annotation

Protein conformational diversity data was acquired from CoDNaS database [31]. The database utilizes a total of 70,467 PDB structures (Protein Data Bank, a repository of biological macromolecular structure) [32], representing a set of 9398 monomeric proteins of the protein data bank. Conformational diversity was measured as the maximum RMSD (root-mean-square deviation measuring the average distance between the superimposed atoms) between available conformers of a protein. RMSD values were normalized to RMSD100 for all proteins with > 40 residues [33]. This provided us with 1094 human proteins with corresponding conformational diversity values.