RESEARCH ARTICLE

REVISED Protein sites with more coevolutionary connections tend to evolve slower, while more variable protein families acquire higher coevolutionary connections [version 2; referees: 2 approved]

Sapan Mandloi , Saikat Chakrabarti

Department of Structural Biology and Bioinformatics Division, Council of Scientific and Industrial Research, Indian Institute of Chemical Biology, Kolkata, West Bengal, 700032, India

v2 **First published:** 10 Apr 2017, **6**:453 (doi: 10.12688/f1000research.11251.1)
Latest published: 07 Jul 2017, **6**:453 (doi: 10.12688/f1000research.11251.2)

Abstract

Background: Amino acid exchanges within proteins sometimes compensate for one another and could therefore be co-evolved. It is essential to investigate the intricate relationship between the extent of coevolution and the evolutionary variability exerted at individual protein sites, as well as the whole protein.

Methods: In this study, we have used a reliable set of coevolutionary connections (sites within 10Å spatial distance) and investigated their correlation with the evolutionary diversity within the respective protein sites.

Results: Based on our observations, we propose an interesting hypothesis that higher numbers of coevolutionary connections are associated with lesser evolutionary variable protein sites, while higher numbers of the coevolutionary connections can be observed for a protein family that has higher evolutionary variability. Our findings also indicate that highly coevolved sites located in a solvent accessible state tend to be less evolutionary variable. This relationship reverts at the whole protein level where cytoplasmic and extracellular proteins show moderately higher anti-correlation between the number of coevolutionary connections and the average evolutionary conservation of the whole protein.

Conclusions: Observations and hypothesis presented in this study provide intriguing insights towards understanding the critical relationship between coevolutionary and evolutionary changes observed within proteins. Our observations encourage further investigation to find out the reasons behind subtle variations in the relationship between coevolutionary connectivity and evolutionary diversity for proteins located at various cellular localizations and/or involved in different molecular-biological functions.

Open Peer Review

Referee Status: ✓ ✓

	Invited Referees 1	 2
REVISED version 2 published 07 Jul 2017	✓ report	
version 1 published 10 Apr 2017	? report	✓ report

1 **Anna Panchenko** , National Institutes of Health, USA

2 **Ramanathan Sowdhamini** , Tata Institute of Fundamental Research, India

Discuss this article

Comments (0)

Corresponding author: Saikat Chakrabarti (saikat@iicb.res.in)

Author roles: Mandloi S: Data Curation, Writing – Review & Editing; **Chakrabarti S**: Conceptualization, Supervision, Writing – Original Draft Preparation

Competing interests: No competing interests were disclosed.

How to cite this article: Mandloi S and Chakrabarti S. **Protein sites with more coevolutionary connections tend to evolve slower, while more variable protein families acquire higher coevolutionary connections [version 2; referees: 2 approved]** *F1000Research* 2017, **6**:453 (doi: 10.12688/f1000research.11251.2)

Grant information: The work is supported by CSIR network project fund (BSC 0121).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

First published: 10 Apr 2017, **6**:453 (doi: 10.12688/f1000research.11251.1)

REVISED **Amendments from Version 1**

Version 2 of this manuscript includes the following changes:

1: Change of a line in the Abstract from “Correlated mutation or coevolution of positions in a protein is tightly linked with the protein’s respective evolutionary rate”, to “Amino acid exchanges within proteins sometimes compensate for one another and could therefore be co-evolved.”

2: Included detail of coevolutionary programs’ range of values.

3: Incorporated result of the MISTIC server to compare with our results.

4: We have included Figure S7 as a MISTIC server Circos representation for CD01291 and CD00164 families.

5: We have included Table S2 which contains the MISTIC server’s web link results for a few randomly selected CDD families.

6: Included mention of project GAP 362 in the Grant information.

See referee reports

Introduction

According to the neutral theory of evolution, the functionality of a protein with a disadvantageous mutation can be restored by another mutation that compensates for the first to sustain the function[1]. Such compensating mutations, together with other factors arising due to common functional, structural and folding constraints, lead to correlations between different positions in a protein or protein family. Coordinated changes of amino acid residues are typically acquired by examining covariation between two aligned positions. A large number of computational methods have been proposed[2–11] to quantify the covariation between two protein sites in a given multiple sequence alignment (MSA). Most methods are based on variation of mutual information[12–17], maximum likelihood approximations[18], Bayesian probabilities[19], and phylogenetic approaches[20,21]. Newer methods successfully implement direct coupling analysis[22], Protein Sparse Inverse COVariance: PSICOV[23] and Matrix Match Maker[24] algorithms to identify coevolving sites. These previous studies demonstrate that sequence covariation is powerful in detecting protein-protein interactions, ligand receptor binding, and the folding structure of a protein. In addition to direct physical interactions, distantly located coevolving amino acid residues are reported to be energetically coupled[25] or subject to similar functional constraints[26]. Compensated amino acid substitutions have been described in previous works in terms of their locations in structure and their physico-chemical properties[3,20,21]. Coevolutionary signals coming from residue charge compensating mutations have been found to be stronger compared to size compensating mutations[3,21,27]. Despite the fact that coevolution has been found to be rather weak in many cases, correlated mutations have had comparative success in predicting protein secondary and tertiary structures, and in some cases protein interaction partners[28–30].

Coevolution is difficult to detect due to various reasons, such as the variable nature of compensatory mutations, the strong dependence of covariations on evolutionary distances, and the number of sequences in the alignment. Hence, it is crucial to understand how coevolutionary processes are related to evolutionary diversity within protein families. Despite significant efforts in this field, the relationship between evolutionary conservation and the extent of coevolution is not well understood. For example, it is not clear whether families with higher evolutionary diversity would exhibit more coevolutionary connections or not. Similarly, at the residue level, this relationship needs to be thoroughly examined. An earlier study by Fodor and Aldrich[31] observed a lack of agreement between correlated mutation methods, and the resultant differences might have been caused by differing sensitivities to background conservation. In a previous study, it was also indicated that residues, which form many coevolutionary connections with other residues, are more evolutionary conserved and are involved in specific functionally important interactions and conformational changes[32].

A complete understanding of protein evolution and coevolution will require a large scale analysis of important factors that determine the selective forces acting on different residues of a protein to be coevolved. Here, we present a study that undertook a detailed analysis to investigate the relationship between evolutionary conservation and the extent of coevolution within a protein. This relationship could be dependent on the reliability of the predicted coevolved sites as there are no direct ways to validate the coevolutionary connectivity. Therefore, it is a good idea to use multiple coevolution extracting algorithms and filter out a reliable set within protein sites. Similarly, spatial proximity between the coevolved sites might provide additional reliability about the predicted coevolved sites[33]. We examined the evolutionary conservation using the popular AL2CO[34] program within 19,736, 35,514, 50,217, and 56,879 coevolved site pairs (located within 10Å spatial distance), which were identified by approaches, such as mutual information (MIp program[6]), McLachlan amino acid similarity matrix based techniques (McBASC program[27]), Direct Coupling Analysis (DCA program[22]), and Protein Sparse Inverse COVariance method (PSICOV program[23]), from 753 curated protein family alignments, available from the Conserved Domain Database (CDD[35]). Our study suggests the hypothesis that a higher number of coevolutionary connection is likely to be observed for a particular site that is less evolutionary variable, while a higher number of coevolutionary connections can be observed for a protein family that has higher evolutionary variability. We found that the sites with a higher number of coevolutionary connections have a much higher tendency to be conserved compared to the sites with a smaller number of connections. These sites might act as ‘hub points’, and therefore changes in these sites would affect many other connected sites. We further investigated the impact of important structural properties, like secondary structures, solvent accessibilities and hydrogen bonding of the coevolved sites, to understand the reasons behind the observed correlation between coevolution and evolutionary diversity. Our findings indicate that coevolved sites are generally preferred at a solvent accessible/hydrogen bonded/helical state compared to a solvent buried/non-hydrogen bonded/β strand state. However, discernable differences in evolutionary conservation between the higher and lesser coevolved sites were observed only for sites located at solvent accessible states compared to buried states. We also examined whether the observed negative (anti) correlation between coevolution and evolutionary conservation for a protein family can be under the influence of its cellular localization or the type of functions with which it is involved. Coevolution analysis for the whole protein suggests that the cytoplasmic and extracellular proteins possess moderately stronger negative (anti) correlation between the number of coevolutionary connections and their average evolutionary conservation.

Methods

Dataset

We collected 753 protein domain alignments from the Conserved Domain Database (CDD; https://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml[35]) version 2.13, for which at least one 3D structure entry and more than 50 protein sequences are available. An alignment length threshold (>=100) was also applied to exclude smaller proteins. A complete list of protein families is provided in Table S1.

Identification of coevolved sites

Mutual information[6] (Suppl. Mat. Ref.) is widely used measure to estimate the covariation between sites in protein families. In this analysis, we used a mutual information based method to estimate coevolutionary connection between two sites of a protein family. This method (MIp) is based on information theory that accurately estimates the expected levels of background coming from random and phylogenetic signals. Removal of the phylogenetic and random background allows identifying substantially more coevolving positions in protein families. Altogether we identified 19,736 (out of total 36,616) coevolved site pairs located within 10Å spatial distance from the 753 family alignments, with a MIp Z-score cutoff of 4.0 or higher.

McBASC[27] (http://fodorwebsite.appspot.com//covariance1_1.zip) was used to calculate the simple inter-position coevolution for the 753 protein family alignments. McBASC provides high score for non-conserved and co-varying positions from a multiple sequence alignment. The calculation of McBASC was performed as described in Fodor and Aldrich 2004, using the software provided by the authors (http://www.afodor.net/). McBASC does not use any structural or phylogenetic information in the calculation of coevolution. We identified 35,514 (out of total 95,866) coevolved site pairs located within 10Å spatial distance from the 753 family alignments with McBASC Z-score cutoff of 4.0 or higher.

DCA[22] (Direct Coupling Analysis) aims at predicting coevolving residues based on the maximum entropy principle. DCA is also used in predicting inter and intra domain contacts. This method is used in separating direct and indirect correlation between residues. DCA analysis was implemented with MATLAB code kindly provided to us by Domenico L. Gatti (Supplementary File 1). We identified 50,217 (out of total 1,61,332) coevolved site pairs located within 10Å spatial distance from the 753 family alignments with DCA Z-score cutoff of 4.0 or higher.

PSICOV[23] (Protein Sparse Inverse COVariance) method is developed with the specific goal of separating direct from indirect coupling between residues. PSICOV takes into account the global correlations between pairs. Modified MATLAB code (without the default minimum requirement of 500 sequences), which was kindly provided to us by Domenico L. Gatti (Supplementary File 1), was used in this study. We identified 56,879 (out of total 162,336) coevolved site pairs located within 10Å spatial distance from the 753 family alignments with PSICOV Z-score cutoff of 4.0 or higher.

Random selection of non-coevolved sites and pairs

Site pairs other than those involved in coevolutionary connections were considered as non-coevolutionary sites. We randomly selected non-coevolved sites from each protein family (Supplementary File 2). For each randomly selected non-coevolved site (i), neighboring non-coevolved sites were selected based on the structural distance (<10Å) and sequence distance filters (>i±6 positions). Similar numbers of non-coevolved site pairs were selected randomly 10 times. We performed similar correlation analysis between the numbers of spatial neighbors and evolutionary conservation of non-coevolved sites.

Calculation of amino acid conservation

Analysis of positional conservation in a sequence alignment can aid in the detection of functionally and/or structurally important residues. The AL2CO[34] program performs conservation analysis in a comprehensive and systematic way. It was used to calculate the conservation index for each position for a given multiple sequence alignment. Twelve different strategies of conservation index calculation have been implemented in the AL2CO program (http://prodata.swmed.edu/al2co/al2co.php). For this analysis, we used independent count (sequence weighting scheme) and matrix based sum-of-pair[36] (conservation calculation method) measure scoring scheme to calculate evolutionary conservation of each coevolved sites or column in the alignment. A higher AL2CO score indicates higher conservation index.

Calculation of spatial distances and structural properties

Representative three-dimensional (3D) structures were collected for each family from the Protein Data Bank (PDB; http://www.rcsb.org/pdb/home/home.do)[37]. Spatial distances were calculated using atom coordinates supplied in the individual PDB file. Structural properties, such as solvent accessibility, secondary structures, and hydrogen bonds, were computed from the protein structure using the JOY package[38] (http://mizuguchilab.org/joy/) Solvent accessibility was measured using the PSA program from the JOY package, and residues that had an accessible surface area <7% were treated as solvent buried or inaccessible. Similarly, secondary structures (helix, strand and coil) and hydrogen bonding patterns were estimated using the SSTRUC and HBOND programs from the JOY package[39], respectively.

Collection of Gene Ontology information

The Gene Ontology (http://www.geneontology.org/)[40] covers three classes/domains: cellular localization, molecular function and biological process. Functional information of each CDD family was collected from Gene Ontology database using the UNIPROT[39] ID of the representative protein structure as a query. We mapped 517, 720, and 634 protein domain families into cellular localization, molecular function and biological process, respectively.

Mapping conservation and coevolutionary connection onto 3D structure

Mapping of evolutionary conservation and coevolutionary information onto the 3D structure was done using in-house Perl scripts (Supplementary File 3). B-Factor column in PDB file was substituted with evolutionary conservation score and colored according to B-Factor ranging blue (low conservation) to red (high conservation). Lines connecting C-alpha atoms of residues represent coevolutionary connection between those residues.

Results and discussion

Coevolution versus evolutionary diversity at the site level

Coevolutionary connections between protein sites were identified from multiple sequence alignments of 753 protein domain families by algorithms employing differing approaches, such as mutual information (MIp program[6]), McLachlan amino acid similarity matrix based techniques (McBASC program[27]), Direct Coupling Analysis (DCA program[22]), and Protein Sparse Inverse COVariance method (PSICOV program[23]). Minimal overlaps were observed for coevolved sites predicted by these programs (Figure S1), supporting previous interpretations that differences in the preferred level of background conservation may exist within each program to identify coevolved residue pairs[6].

The pattern of evolutionary diversity within the coevolved sites was examined using evolutionary conservation scoring approaches (*e.g.*, AL2CO). Figure 1 plots the average conservation scores of sites having higher or lower coevolutionary connections (A: MIp; B: McBASC; C: DCA; D: PSICOV programs, respectively). Figure 1 suggests that highly coevolved sites possess higher average AL2CO scores, depicting higher evolutionary conservation. Coevolutionary connections, even though selected based on a strong statistically significant threshold (Z-score >4), might contain background noise resulting in an unreliable

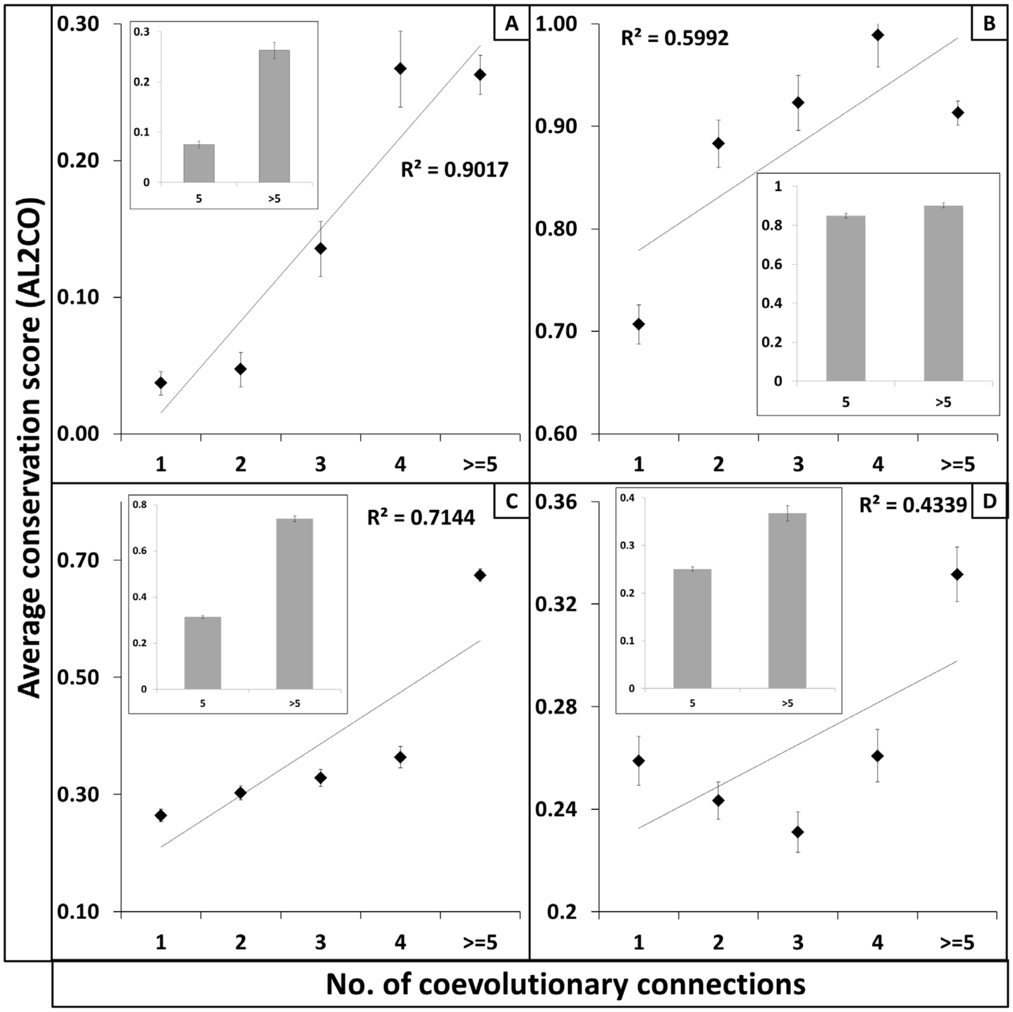


Figure 1. Relationship between coevolutionary connections and evolutionary conservation for protein sites. X-axes show the coevolutionary connection (represented in bins of 1 and 5) per site whereas Y-axes represent the average evolutionary conservation index (CI) estimated by the AL2CO program. Each vertical panel (panel **A**–**D**) represents results obtained from coevolution predicted by various programs (Panel **A**: MIp; **B**: McBASC; **C**: DCA; **D**: PSICOV). Panels provide correlation data for the coevolved sites that are located within or equal to 10Å. The coefficient of determination (R^2) indicates how well the data points fit to the linear regression model between coevolutionary connection and evolutionary conservation. The observed scale of coevolution values obtained from multiple covevolutionary programs varies a lot. The probable reason for such observation can be the algorithm used by individual programs for calculation of covariation/coevolution.

relationship between coevolution and evolutionary conservation. To disprove this, we performed similar analysis using non-coevolutionary random sites and found that there is a smaller correlation between non-coevolved sites having higher or lower structural distance based neighbors (<10Å) and their evolutionary conservation (Figure S2).

Observation of strong positive correlation between coevolutionary connections and evolutionary conservation within the coevolved sites selected based on structural proximity suggests that highly coevolved protein sites tend to evolve slower.

Influence of structural environment. The structural environment of a protein site is a critical factor that can influence its evolutionary diversity pattern[41,42]. To understand the reasons behind the observed phenomenon where higher coevolutionary connections are found for sites that are less diversified, we investigated the roles of structural environments, such as solvent accessibility state, and secondary structural content of the coevolved sites.

We have observed more coevolutionary connections for sites that are solvent accessible compared to that observed within buried sites (Figure 2A). Interestingly, solvent accessible sites that possess lower numbers (<3) of coevolutionary connections (LCC) are consistently less conserved compared to the sites that have relatively higher number (>3) of coevolutionary connections (HCC) (Figure 2A). Although the similar trend is also observed within the solvent buried sites, the differences of conversation indices between the HCC and LCC are more prominent within solvent accessible state compared to that observed at buried state (Figure 2B).

Higher abundance of coevolutionary connections is also observed for sites that are involved in hydrogen bonding compared to those are not involved in hydrogen bonding. However, no discernable differences in evolutionary conservation were observed between the higher and lesser coevolved sites involved in hydrogen bonding compared to those that do not have hydrogen bonding (Figure S3).

Slightly higher abundance of coevolutionary connections was observed for sites that were located in helix compared to those forming strands. No discernable differences in evolutionary conservation were observed between the higher and lesser coevolved

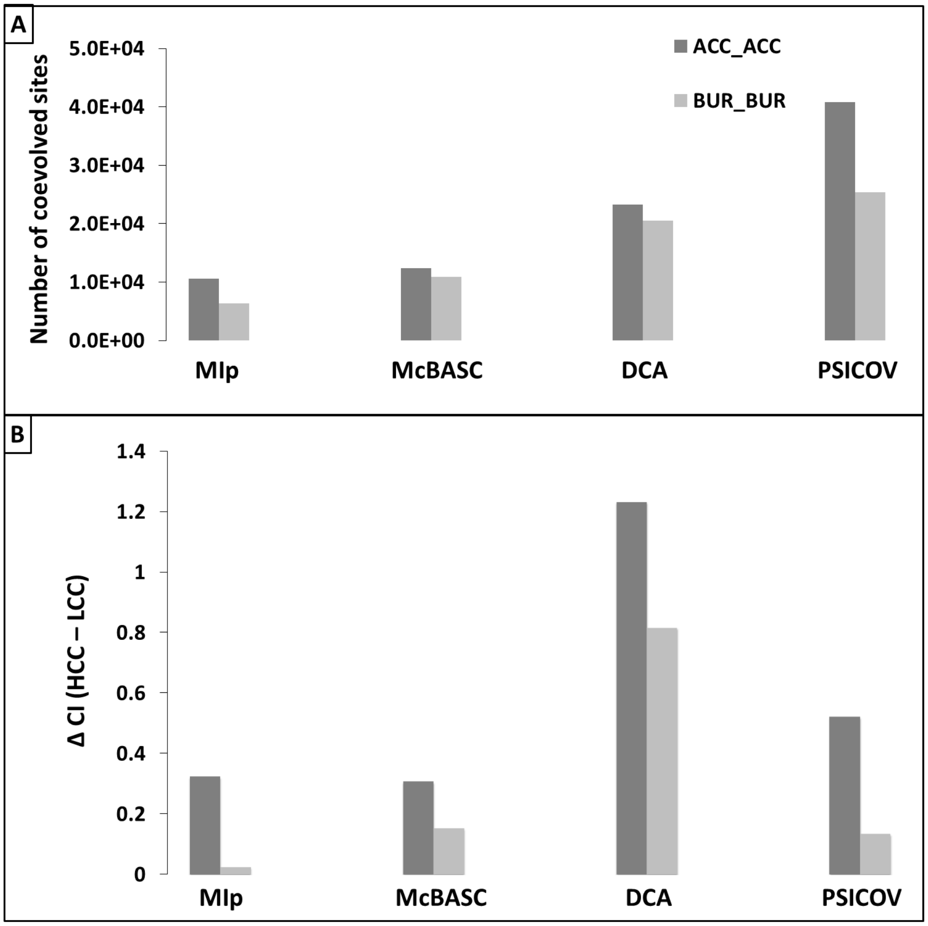


Figure 2. Analysis for sites involved in solvent accessible and buried environment. (**A**) Number of coevolved sites involved in forming coevolutionary pairs, where both sites are present in solvent accessible (ACC_ACC; dark grey) and buried (BUR_BUR; light grey) environments. (**B**) Difference of conversation indices (CI) between higher coevolutionary connection (HCC) and lower coevolutionary connection (LCC) sites involved in ACC_ACC and BUR_BUR environments. LCC: less than or equal to 3 coevolutionary connections; HCC: higher than 3 coevolutionary connections.

sites located at helical environments compared to those that form strands (Figure S4).

Influence of functional involvement. We also investigated the relationship between coevolutionary connection and evolutionary conservation for protein sites with respect to their functional involvement. However, functional sites (e.g., active sites, protein or ligand binding sites) do not show significantly higher positive correlation between coevolutionary connection and evolutionary conservation, and no discernable differences were observed among the correlation coefficients between coevolutionary connection and evolutionary conservation observed for various types of functional sites (data not visualised).

Coevolution versus evolutionary diversity at the protein/family level

It is important to know how the evolutionary conservation profile of the whole protein or family influences the coevolutionary connections within its sites. Figure 3 and Figure S5 plot the average conservation scores of protein families (considering all gapless columns of the family alignment) with respect to the total number of coevolved sites observed within those families. Our results suggest strong negative correlation between the number of coevolved sites found within a protein family and its average conservation score. This finding indicates that, in general, more conserved proteins/families tend to possess lower coevolutionary connections, whereas proteins/families with less stringent evolutionary pressure might engineer more intra-coevolutionary connections.

We further investigated the influence of cellular localization and biological-molecular functions of the proteins that displayed correlation between the coevolutionary connections and evolutionary conservation. We categorized the representative proteins from 517, 720, 634 families into cellular localization, molecular function and biological processes, respectively, using their Gene Ontology annotations. For example, 54%, 15% and 12% of the 517 families, having at least one pair of coevolved sites, reside within cytoplasm, nucleus and membrane, respectively (Figure S6). Similarly, 55%, 17% and 10% coevolved protein families are involved in catalysis (enzyme), nucleic acid and ion binding functions, respectively. Coevolved proteins were also found to be abundant in various metabolic functions (Figure S6). Table 1 provides the R^2 and slope (m) values between the coevolutionary

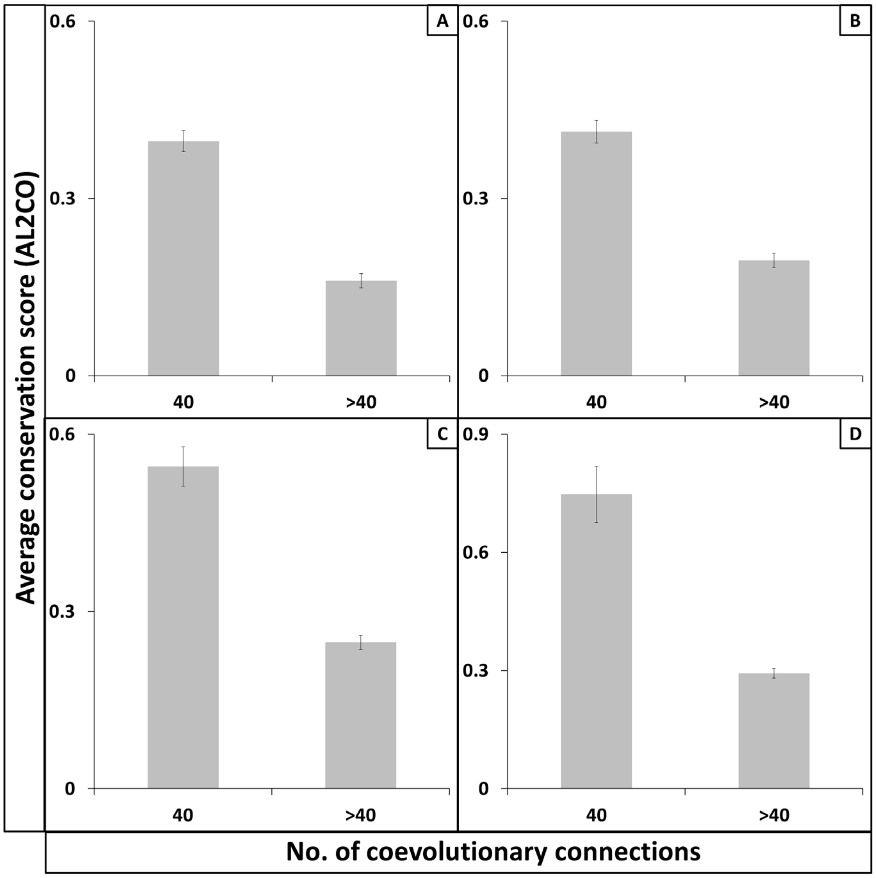


Figure 3. Relationship between coevolutionary connections and evolutionary conservation for the full-length protein. X-axes show the coevolutionary connections (represented in bins of 40) of protein families whereas Y-axes represent the average evolutionary conservation score of the same families estimated by the AL2CO program. Panels show the data extracted from all 753 CDD families. Each panel (**A**–**D**) represents results obtained from coevolution predicted by various programs (MIp, McBASC, DCA and PSICOV, respectively).

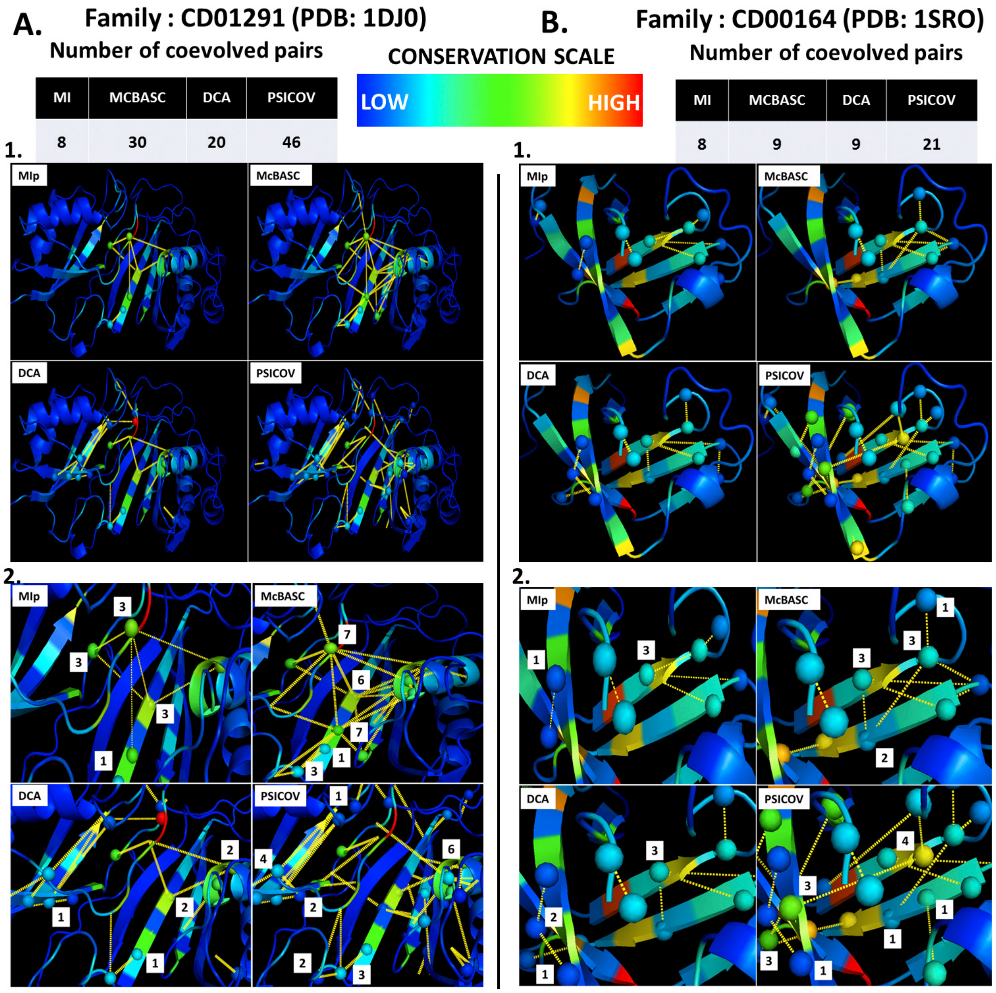
Table 1. Correlation between the coevolutionary connections and the evolutionary conservation of proteins with respect to their Gene Ontology classification. R^2: coefficient of determination; m: slope of line for relationship between the coevolutionary connections (predicted by MI, McBASC, DCA and PSICOV programs) and the evolutionary conservation of proteins with respect to their most frequently observed Gene Ontology based cellular localizations, molecular functions, and biological processes.

Program	Gene Ontology											
	Cellular localizations				Molecular functions				Biological processes			
	Cytoplasm	Nucleus	Membrane	Extracellular space	Catalysis	Nucleic acid binding	Ion binding	Ligand binding	Anabolic processes	Catabolic processes	Other metabolic processes	Transport
MIp	R^2 = 0.63	R^2 = 0.12	R^2 = 0.6	R^2 = 0.70	R^2 = 0.82	R^2 = 0.43	R^2 = 0.51	R^2 = 0.40	R^2 = 0.84	R^2 = 0.62	R^2 = 0.77	R^2 = 0.54
	m = -0.03	m = -0.03	m = -0.06	m = -0.06	m = -0.05	m = -0.05	m = -0.05	m = -0.02	m = -0.01	m = -0.02	m = -0.04	m = -0.06
MCBASC	R^2 = 0.68	R^2 = 0.62	R^2 = 0.2	R^2 = 0.53	R^2 = 0.55	R^2 = 0.53	R^2 = 0.55	R^2 = 0.25	R^2 = 0.15	R^2 = 0.40	R^2 = 0.51	R^2 = 0.24
	m = -0.04	m = -0.07	m = -0.03	m = -0.06	m = -0.08	m = -0.06	m = -0.05	m = -0.02	m = -0.01	m = -0.04	m = -0.04	m = -0.05
DCA	R^2 = 0.93	R^2 = 0.47	R^2 = 0.54	R^2 = 0.37	R^2 = 0.62	R^2 = 0.83	R^2 = 0.63	R^2 = 0.68	R^2 = 0.27	R^2 = 0.54	R^2 = 0.81	R^2 = 0.48
	m = -0.06	m = -0.09	m = -0.08	m = -0.05	m = -0.10	m = -0.07	m = -0.07	m = -0.04	m = -0.02	m = -0.14	m = -0.07	m = -0.09
PSICOV	R^2 = 0.41	R^2 = 0.01	R^2 = 0.42	R^2 = 0.61	R^2 = 0.74	R^2 = 0.67	R^2 = 0.4	R^2 = 0.35	R^2 = 0.16	R^2 = 0.43	R^2 = 0.89	R^2 = 0.0006
	m = -0.05	m = -0.01	m = -0.05	m = -0.10	m = -0.27	m = -0.10	m = -0.04	m = -0.04	m = 0.01	m = -0.24	m = -0.07	m = -0.01

connection and evolutionary conservation for proteins categorized in certain cellular localization. Cytoplasmic and extracellular proteins show slightly stronger anti-correlation between the number of their coevolutionary connections and evolutionary conservation. Similarly, proteins involved in catalysis and nucleic acid binding type of molecular functions show moderately stronger negative correlation, whereas proteins involved in miscellaneous metabolic processes, which mostly include generic carbohydrate and glutamine metabolisms and nitrogen fixation processes, exhibit stronger negative correlation between coevolutionary connections within the protein and its average conservation (Table 1).

Example cases

Figure 4A provides an example case where coevolutionary connections are overlaid with evolutionary conservation scores onto the 3D structure of a representative protein (PDB code: 1DJ0) from the pseudouridine synthase domain family (CDD code: CD01291). 8, 30, 20 and 46 coevolutionary connections were predicted by MIp[6], McBASC[27], DCA[22] and PSICOV23 methods, respectively. Interestingly, in this family, the average conservation score (AL2CO score: 0.65) for all sites are quite low (as shown by color coding), despite having higher coevolutionary connections. Hence, observations in this family support the hypothesis that a higher number of coevolutionary connections can be expected for a protein family that has higher evolutionary variability or lower evolutionary conservation. Similarly, Figure 4B provides a case where coevolutionary connections are projected onto the 3D structure of a representative protein (PDB code: 1SRO) from the ribosomal protein S1 domain (CDD code: CD00164). It is evident from Figure 4B that the number of coevolutionary connections is relatively low in this family, while the overall


A.
Family : CD01291 (PDB: 1DJ0)
Number of coevolved pairs
CONSERVATION SCALE
LOW
HIGH
B.
Family : CD00164 (PDB: 1SRO)
Number of coevolved pairs
1.
MIp
McBASC
DCA
PSICOV
2.
1
2
3
4
6
7

MI	MCBASC	DCA	PSICOV
8	30	20	46

MI	MCBASC	DCA	PSICOV
8	9	9	21

Figure 4. Coevolutionary connection and evolutionary conservation are projected onto the 3D structures of proteins from two different protein families. Panel **A1** provides an example of higher coevolutionary connections (average >20) with respect to an overall lower evolutionary conservation (average AL2CO score: 0.65) status projected on the 3D structure of a representative protein (PDB code: 1DJ0) from the pseudouridine synthase domain family (CDD code: CD01291). Panel **B1** represents a case [representative protein (PDB code: 1SRO; CDD code: CD00164) from the ribosomal protein S1 domain] where lower coevolutionary connections (average <10), with respect to overall higher evolutionary conservation (average AL2CO score: 1.63) status are observed. Lower panels (**A2** and **B2**) show examples from the same families (zoomed image) of higher coevolutionary connections for sites that have relatively higher evolutionary conservation.

evolutionary conservation (indicated by color coding) is higher (AL2CO score: 1.63). Hence, observations in this protein support the hypothesis that lower evolutionary connections can be expected for a less evolutionary variable protein. Interestingly, sites within the 1SRO protein show a similar trend as observed in the 1DJ0 protein (panels A2 and B2 of Figure 4), and higher numbers of coevolutionary connections are observed for protein sites that are less evolutionary variable.

Dataset 1. Predicted data for coevolution and conservation

http://dx.doi.org/10.5256/f1000research.11251.d157108

Files of coevolutionary sites predicted by four programs with conservation score predicted by AL2CO program with 10Å filter.

In order to compare our findings, we have performed a similar analysis using the MISTIC[43] (Mutual Information Server To Infer Coevolution) server, taking 20 randomly selected protein families from our dataset as case study. Interestingly, the observed coevolutionary network predicted for CD01291 and CD00164 families (discussed above) are similar to our study (Figure S7). MISTIC results also show that the CD01291 family has higher coevolutionary network connections for fewer variables sites whereas the CD00164 family has less coevolutionary connections and is overall less conserved. The MISTIC server's web link results for other protein families, are available in Table S2.

Conclusions

Over the years, it has become apparent that intra protein coevolution is an important evolutionary phenomenon to maintain proteins' functional flexibility. However, the signs of coevolution are subtle, and as a consequence, hard to detect. The majority of sites in a protein coevolve to some degree, in that they contribute more or less to structural integrity and, thus, function of the protein. However, some sites will more directly influence each other. By definition, coevolution is closely connected to the evolutionary variability of a protein. Hence, it is essential to investigate the intricate relationship between the extent of coevolution and the evolutionary variability exerted at individual protein sites, as well as the whole protein. However, it is also relevant to check the reliability of the predicted coevolved sites before deriving any hypothesis between coevolution and evolutionary conservation. Therefore, we employed multiple algorithms for the detection of coevolutionary connection and used a structural proximity based filtration system to validate the coevolutionary connections within protein sites.

In this study we have not checked/compared the difference between the two concepts of covariation and coevolution. We have used different programs (MIp, McBASC, DCA and PSICOV) which calculate covariation among protein sites in tree-independent manner. In this study, it was assumed that observed patterns of covariation are caused by molecular coevolution and they were treated synonymously. To the best of our knowledge, this is the first time where such a detailed analysis is performed to investigate any existing correlation between the coevolution and evolutionary conservation. Based on our observations, we propose an interesting hypothesis that a higher number of coevolutionary connection is associated for a protein site that is less evolutionary variable, while a higher number of the coevolutionary connections can be observed for a protein family that has higher evolutionary variability. The obvious question is why such apparently contrasting relationship exists. One probable explanation could be that these highly coevolved sites might act as 'coevolutionary hubs', and therefore changes at these sites would affect many other connected sites. On the contrary, the evolutionary selection pressure needs to be lower at the whole protein for more sites to be involved in covariation. Probably, sites that are critical to maintain structural integrity and functional flexibility are co-varying with many other sites, but the extent of variation is limited. Hence, the critical balance between covariation and evolutionary conservation is maintained via these 'coevolutionary hub' sites. However, to be rich in a coevolutionary connection, a protein requires evolutionary flexibility so that correlated or compensatory mutations can be arranged with response to an initial change. Hence, higher coevolutionary connection is observed for families that are more evolutionary variable than others.

Data availability

Dataset 1: Predicted data for coevolution and conservation. Files of coevolutionary sites predicted by four programs with conservation score predicted by AL2CO program with 10Å filter. doi, 10.5256/f1000research.11251.d157108[44]

Author contributions

SC designed the work, analyzed and interpreted data; SM conceived the experiments, analyzed and interpreted data. SC and SM wrote the manuscript.

Competing interests

No competing interests were disclosed.

Grant information

The work is supported by the CSIR network project fund (BSC 0121) and GAP 362.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

SC acknowledges CSIR-IICB for infrastructural support and Department of Biotechnology for Ramalingaswami fellowship. SM acknowledges Department of Science and Technology for INSPIRE fellowship.

Supplementary material

Table S1: List of CDD families. File contains CDD family ID, PDB ID and UNIPROT ID used in the study.

Click here to access the data.

Table S2: MISTIC server results for CDD protein families.

Click here to access the data.

Supplementary File 1: Program for coevolutionary connection prediction. MATLAB code for DCA and PSICOV coevolutionary connection prediction program. This code is provided by Domenico L. Gatti with permission.

Click here to access the data.

Supplementary File 2: Program to extract non-coevolved sites.

Click here to access the data.

Supplementary File 3: Program and data-files to map conservation and coevolutionary information onto PDB.

Click here to access the data.

Figure S1: Number of coevolved pairs predicted (in blue, pink, green and yellow) by different programs (MIp, MCBASC, DCA and PSICOV, respectively) and common pairs (in black) between them.

Click here to access the data.

Figure S2: Relationship between the number of structural neighbor and evolutionary conservation for non-coevolved protein sites. X-axes show the numbers of structural neighbor (represented in bins of 1) per site whereas Y-axes represent the average evolutionary conservation index estimated by the AL2CO program. Each panel (panel **A–D**) represents results obtained from the number of non-coevolved pairs similar to the number of coevolved pairs predicted by various coevolution programs (A: MIp; B: McBASC; C: DCA; D: PSICOV). Panels provide correlation data for the non-coevolved sites (i) that are located within or equal to 10Å (and at sequence position between i and >i±6). The coefficient of determination (R^2) indicates how well the data points fit to the linear regression model between coevolutionary connection and evolutionary conservation.

Click here to access the data.

Figure S3: Analysis for sites involved in H-bonding. (**A**) Number of coevolved sites involved in forming coevolutionary pairs, where both sites are either involved in H-bonding (HBY_HBY; dark grey) or not involved in H-bonding (HBX_HBX; light grey). (**B**) Difference of conversation indices (CI) between higher coevolutionary connection (HCC) and lower coevolutionary connection (LCC) sites involved in HBY_HBY or HBX_HBX. LCC: less than or equal to 3 coevolutionary connections; HCC: higher than 3 coevolutionary connections.

Click here to access the data.

Figure S4: Analysis for sites involved in different secondary structures. (**A**) Number of coevolved sites involved in forming coevolutionary pairs, where both sites are located in helix (H_H; dark grey) and strand (E_E; light grey). (**B**) Difference of conversation indices (CI) between higher coevolutionary connection (HCC) and lower coevolutionary connection (LCC) sites involved in H_H and E_E. LCC: less than or equal to 3 coevolutionary connections; HCC: higher than 3 coevolutionary connections.

Click here to access the data.

Figure S5: Relationship between coevolutionary connections and evolutionary conservation for the full-length protein. X-axes show the coevolutionary connection (represented in bins of 10) of protein families whereas Y-axes represent the average evolutionary conservation score of the same families estimated by the AL2CO program. Panels show the data extracted from all 753 CDD families. Each panel (**A–D**) represents results obtained from coevolution predicted by various programs (MIp, MCBASC, DCA and PSICOV, respectively).

Click here to access the data.

Figure S6: Gene Ontology distribution of the protein families used for this study. (**1**) Representative proteins from 517 CDD families were assigned to cellular localization, whereas the same from 720 and 624 families could be assigned to at least one (**2**) molecular function or (**3**) biological process (**3**), respectively. Details can be found in Methods.

Click here to access the data.

Figure S7: MISTIC server results as Circos representation for (A) CD01291 and (B) CD00164 families. MI Circos is a sequential circular representation of the MSA and the information it contains. Colored square boxes of the second circle indicate the MSA position conservation (highly conserverd positions are shown in red, while less conserved ones are shown in blue). Lines connect pairs of positions with MI greater than 6.5 (Marino Buslje *et al.*, 2009). **Red** edges represent the top 5%, **black** ones are between 70% and 95%, and **gray** edges account for the remaining 70%.

Click here to access the data.

References

1. Kimura M: **The Neutral Theory of Molecular Evolution.** Cambridge: Cambridge University Press, 1994.

2. Taylor WR, Hatrick K: **Compensating changes in protein multiple sequence alignments.** *Protein Eng.* 1994; **7**(3): 341–8.
PubMed Abstract | Publisher Full Text

3. Chelvanayagam G, Eggenschwiler A, Knecht L, *et al.*: **An analysis of simultaneous variation in protein structures.** *Protein Eng.* 1997; **10**(4): 307–16.
PubMed Abstract | Publisher Full Text

4. Pazos F, Helmer-Citterich M, Ausiello G, *et al.*: **Correlated mutations contain information about protein-protein interaction.** *J Mol Biol.* 1997; **271**(4): 511–23.
PubMed Abstract | Publisher Full Text

5. Oliveira L, Paiva AC, Vriend G: **Correlated mutation analyses on very large sequence families.** *Chembiochem.* 2002; **3**(10): 1010–7.
PubMed Abstract | Publisher Full Text

6. Dunn SD, Wahl LM, Gloor GB: **Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction.** *Bioinformatics.* 2008; **24**(3): 333–40.
PubMed Abstract | Publisher Full Text

7. Martin LC, Gloor GB, Dunn SD, *et al.*: **Using information theory to search for co-evolving residues in proteins.** *Bioinformatics.* 2005; **21**(22): 4116–24.
PubMed Abstract | Publisher Full Text

8. Goh CS, Bogan AA, Joachimiak M, *et al.*: **Co-evolution of proteins with their interaction partners.** *J Mol Biol.* 2000; **299**(2): 283–93.
PubMed Abstract | Publisher Full Text

9. Goh CS, Cohen FE: **Co-evolutionary analysis reveals insights into protein-protein interactions.** *J Mol Biol.* 2002; **324**(1): 177–92.
PubMed Abstract | Publisher Full Text

10. Fares MA, McNally D: **CAPS: coevolution analysis using protein sequences.** *Bioinformatics.* 2006; **22**(22): 2821–2.
PubMed Abstract | Publisher Full Text

11. Yip KY, Patel P, Kim PM, *et al.*: **An integrated system for studying residue coevolution in proteins.** *Bioinformatics.* 2008; **24**(2): 290–2.
PubMed Abstract | Publisher Full Text

12. Buslje CM, Santos J, Delfino JM, *et al.*: **Correction for phylogeny, small number of observations and data redundancy improves the identification of coevolving amino acid pairs using mutual information.** *Bioinformatics.* 2009; **25**(9): 1125–31.
PubMed Abstract | Publisher Full Text | Free Full Text

13. Gouveia-Oliveira R, Pedersen AG: **Finding coevolving amino acid residues using row and column weighting of mutual information and multi-dimensional amino acid representation.** *Algorithms Mol Biol.* 2007; **2**: 12.
PubMed Abstract | Publisher Full Text | Free Full Text

14. Korber BT, Farber RM, Wolpert DH, *et al.*: **Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis.** *Proc Natl Acad Sci U S A.* 1993; **90**(15): 7176–80.
PubMed Abstract | Publisher Full Text | Free Full Text

15. Little DY, Chen L: **Identification of coevolving residues and coevolution potentials emphasizing structure, bond formation and catalytic coordination in protein evolution.** *PLoS One.* 2009; **4**(3): e4762.
PubMed Abstract | Publisher Full Text | Free Full Text

16. Fatakia SN, Costanzi S, Chow CC: **Computing highly correlated positions using mutual information and graph theory for G protein-coupled receptors.** *PLoS One.* 2009; **4**(3): e4681.
PubMed Abstract | Publisher Full Text | Free Full Text

17. Gao H, Dou Y, Yang J, *et al.*: **New methods to measure residues coevolution in proteins.** *BMC Bioinformatics.* 2011; **12**: 206.
PubMed Abstract | Publisher Full Text | Free Full Text

18. Pollock DD, Taylor WR, Goldman N: **Coevolving protein residues: maximum likelihood identification and relationship to structure.** *J Mol Biol.* 1999; **287**(1): 187–98.
PubMed Abstract | Publisher Full Text

19. Dimmic MW, Hubisz MJ, Bustamante CD, *et al.*: **Detecting coevolving amino acid sites using Bayesian mutational mapping.** *Bioinformatics.* 2005; **21**(Suppl 1): i126–35.
PubMed Abstract | Publisher Full Text

20. Fukami-Kobayashi K, Schreiber DR, Benner SA: **Detecting compensatory covariation signals in protein evolution using reconstructed ancestral sequences.** *J Mol Biol.* 2002; **319**(3): 729–43.
PubMed Abstract | Publisher Full Text

21. Choi SS, Li W, Lahn BT: **Robust signals of coevolution of interacting residues in mammalian proteomes identified by phylogeny-aided structural analysis.** *Nat Genet.* 2005; **37**(12): 1367–71.
PubMed Abstract | Publisher Full Text

22. Morcos F, Pagnani A, Lunt B, *et al.*: **Direct-coupling analysis of residue coevolution captures native contacts across many protein families.** *Proc Natl Acad Sci U S A.* 2011; **108**(49): E1293–301.
PubMed Abstract | Publisher Full Text | Free Full Text

23. Jones DT, Buchan DW, Cozzetto D, *et al.*: **PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments.** *Bioinformatics.* 2012; **28**(2): 184–90.
PubMed Abstract | Publisher Full Text

24. Rodionov A, Bezginov A, Rose J, *et al.*: **A new, fast algorithm for detecting protein coevolution using maximum compatible cliques.** *Algorithms Mol Biol.* 2011; **6**: 17.
PubMed Abstract | Publisher Full Text | Free Full Text

25. Lockless SW, Ranganathan R: **Evolutionarily conserved pathways of energetic connectivity in protein families.** *Science.* 1999; **286**(5438): 295–9.
PubMed Abstract | Publisher Full Text

26. Fares MA, Travers SA: **A novel method for detecting intramolecular coevolution: adding a further dimension to selective constraints analyses.** *Genetics.* 2006; **173**(1): 9–23.
PubMed Abstract | Publisher Full Text | Free Full Text

27. Olmea O, Rost B, Valencia A: **Effective use of sequence correlation and conservation in fold recognition.** *J Mol Biol.* 1999; **293**(5): 1221–39.
PubMed Abstract | Publisher Full Text

28. Göbel U, Sander C, Schneider R, *et al.*: **Correlated mutations and residue contacts in proteins.** *Proteins.* 1994; **18**(4): 309–17.
PubMed Abstract | Publisher Full Text

29. Kann MG, Shoemaker BA, Panchenko AR, *et al.*: **Correlated evolution of interacting proteins: looking behind the mirrortree.** *J Mol Biol.* 2009; **385**(1): 91–8.
PubMed Abstract | Publisher Full Text | Free Full Text

30. de Juan D, Pazos F, Valencia A: **Emerging methods in protein co-evolution.** *Nat Rev Genet.* 2013; **14**(4): 249–61.
PubMed Abstract | Publisher Full Text

31. Fodor AA, Aldrich RW: **Influence of conservation on calculations of amino acid covariance in multiple sequence alignments.** *Proteins.* 2004; **56**(2): 211–21.
PubMed Abstract | Publisher Full Text

32. Chakrabarti S, Panchenko AR: **Coevolution in defining the functional specificity.** *Proteins.* 2009; **75**(1): 231–40.
PubMed Abstract | Publisher Full Text | Free Full Text

33. Horner DS, Pirovano W, Pesole G: **Correlated substitution analysis and the prediction of amino acid structural contacts.** *Brief Bioinform.* 2008; **9**(1): 46–56.
PubMed Abstract | Publisher Full Text

34. Pei J, Grishin NV: **AL2CO: calculation of positional conservation in a protein sequence alignment.** *Bioinformatics.* 2001; **17**(8): 700–12.
PubMed Abstract | Publisher Full Text

35. Marchler-Bauer A, Anderson JB, Cherukuri PF, *et al.*: **CDD: a Conserved Domain Database for protein classification.** *Nucleic Acids Res.* 2005; **33**(Database issue): D192–6.
PubMed Abstract | Publisher Full Text | Free Full Text

36. Wang G, Dunbrack RL Jr: **Scoring profile-to-profile sequence alignments.** *Protein Sci.* 2004; **13**(6): 1612–26.
PubMed Abstract | Publisher Full Text | Free Full Text

37. Berman HM, Westbrook J, Feng Z, *et al.*: **The Protein Data Bank.** *Nucleic Acids Res.* 2000; **28**(1): 235–42.
PubMed Abstract | Publisher Full Text | Free Full Text

38. Mizuguchi K, Deane CM, Blundell TL, *et al.*: **JOY: protein sequence-structure representation and analysis.** *Bioinformatics.* 1998; **14**(7): 617–23.
PubMed Abstract | Publisher Full Text

39. UniProt Consortium: **The Universal Protein Resource (UniProt) in 2010.** *Nucleic Acids Res.* 2010; **38**(Database issue): D142–8.
PubMed Abstract | Publisher Full Text | Free Full Text

40. Ashburner M, Ball CA, Blake JA, *et al.*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet.* 2000; **25**(1): 25–9.
PubMed Abstract | Publisher Full Text | Free Full Text

41. Overington J, Donnelly D, Johnson MS, *et al.*: **Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds.** *Protein Sci.* 1992; **1**(2): 216–26.
PubMed Abstract | Publisher Full Text | Free Full Text

42. Lee S, Blundell TL: **Ulla: a program for calculating environment-specific amino acid substitution tables.** *Bioinformatics.* 2009; **25**(15): 1976–7.
PubMed Abstract | Publisher Full Text | Free Full Text

43. Simonetti FL, Teppa E, Chernomoretz A, *et al.*: **MISTIC: Mutual information server to infer coevolution.** *Nucleic Acids Res.* 2013; **41**(Web Server issue): W8–14.
PubMed Abstract | Publisher Full Text | Free Full Text

44. Mandloi S, Chakrabarti S: **Dataset 1 in: Protein sites with more coevolutionary connections tend to evolve slower, while more variable protein families acquire higher coevolutionary connections.** *F1000Research.* 2017.
Data Source

Open Peer Review

Current Referee Status: ✓ ✓

Version 2

Referee Report 10 July 2017

doi:10.5256/f1000research.13045.r24094

✓ **Anna Panchenko**
Computational Biology Branch, National Institutes of Health, Bethesda, MD, USA

The authors have addressed my comments.

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Referee Report 03 May 2017

doi:10.5256/f1000research.12138.r21734

✓ **Ramanathan Sowdhamini**
National Centre for Biological Sciences, Tata Institute of Fundamental Research, Bangalore, Karnataka, India

The evolutionary conservation of a large number of predicted co-evolving residue pairs has been investigated for possible correlation between the extent of conservation and the strength of co-evolving residue networks. Co-evolution has been predicted by four different popular algorithms. Residues with a high networking of co-evolved residues are found to be more evolutionarily conserved. However, the same trend is not true at the entire protein domain family level and evolutionarily conserved protein families appear to exhibit less co-evolved network of residues. Likewise, solvent-accessible residues were predicted to retain more co-evolutionary connections in comparison to solvent-buried residues. These are interesting observations, but the connections between these individual observations and possible implications/applications will be good to include in the paper.

Queries:

The first sentence in the Abstract could be changed to "Amino acid exchanges within proteins sometimes compensate for one another and could therefore be co-evolved." since this fact of tight linking is not well-known and forms one of the questions in this study.

Page 4: It will be nice to explain how the conservation score (within the AL2CO program) is calculated.

Was there no check for consensus in predicting the co-evolved residues? For instance, to see whichever are predicted by three or more methods … It will be interesting to examine the results for subset of such highly predicted co-evolved sites.

Page 4: “Representative three-dimensional (3D) structures were collected
for each family from the Protein Data Bank” – to provide details as to how they were selected?

Page 5: This statement “Observation of strong positive correlation between coevolutionary connections and evolutionary conservation within the coevolved sites selected based on structural proximity suggests that highly coevolved protein sites tend to evolve slower.” seems to be apparently counter-intuitive. How can highly conserved sites be co-evolving also? Highly conserved sites usually imply high degree of identity (self-amino acid preservation). If so, how a co-evolutionary index can be set up for two spatially proximate residues which remain identical? Please explain for the benefit of the readers.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Referee Expertise: Structural bioinformatics, genomics, genome analysis, protein-protein interactions

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 14 Jun 2017

Saikat Chakrabarti, Indian Institute of Chemical Biology, India

The evolutionary conservation of a large number of predicted co-evolving residue pairs has been investigated for possible correlation between the extent of conservation and the strength of co-evolving residue networks. Co-evolution has been predicted by four

different popular algorithms. Residues with a high networking of co-evolved residues are found to be more evolutionarily conserved. However, the same trend is not true at the entire protein domain family level and evolutionarily conserved protein families appear to exhibit less co-evolved network of residues. Likewise, solvent-accessible residues were predicted to retain more co-evolutionary connections in comparison to solvent-buried residues. These are interesting observations, but the connections between these individual observations and possible implications/applications will be good to include in the paper.

We thank the reviewer for the positive comments. Current study does not provide practical application in its current form, but does offer insight into the underlying properties of covariation/coevoluution methods and the relationship of these methods with evolutionary rate. However, the knowledge regarding the intricate relationship between evolutionary variability and coevolutionary connection is very important to gain insight about the dynamics and pattern of evolutionary history of protein families. The variable nature of this intricate balance is perhaps crucial in determining the overall conservation and/or flexibility of functionally important sites within certain protein families.

The first sentence in the Abstract could be changed to "Amino acid exchanges within proteins sometimes compensate for one another and could therefore be co-evolved." since this fact of tight linking is not well-known and forms one of the questions in this study.

We thank the reviewer for her thoughtful comment. We agree to change the line in the abstract.

Page 4: It will be nice to explain how the conservation score (within the AL2CO program) is calculated.

Information on conservation score calculation is provided in manuscript subsection “Calculation of amino acid conservation”. The AL2CO (manuscript reference: 34) program performs conservation analysis in a comprehensive and systematic way. We used independent count (sequence weighting scheme) and matrix based sum-of-pair (conservation calculation method) measure scoring scheme of AL2CO program to calculate evolutionary conservation of each coevolved sites or column in the alignment. These scoring functions use the summation of the products of frequencies for the column for every combination of amino acid *a* and *b* and multiplies the products by the corresponding BLOSUM62 matrix amino acid substitution frequencies.

Was there no check for consensus in predicting the co-evolved residues? For instance, to see whichever are predicted by three or more methods … It will be interesting to examine the results for subset of such highly predicted co-evolved sites.

We thank the reviewer for her insightful opinion. Figure S1 in supplementary file 1 shows number of coevolved pairs predicted (in blue, pink, green and yellow) by different programs (MIp, MCBASC, DCA and PSICOV, respectively) and common pairs (in black) between them. We have performed correlation analysis on consensus data (for ex. 3335 coevolved pairs predicted by all four programs), we have observed similar trend, but as the number of consensus predicted coevolved sites are not so large, we have not provided in the results.

Page 4: "Representative three-dimensional (3D) structures were collected for each family from the Protein Data Bank" – to provide details as to how they were selected?

We have selected 3D structure of each family from the conserved domain database (CDD) alignment (represented as first sequence in alignment file).

Page 5: This statement "Observation of strong positive correlation between coevolutionary connections and evolutionary conservation within the coevolved sites selected based on structural proximity suggests that highly coevolved protein sites tend to evolve slower." seems to be apparently counter-intuitive. How can highly conserved sites be co-evolving also? Highly conserved sites usually imply high degree of identity (self-amino acid preservation). If so, how a co-evolutionary index can be set up for two spatially proximate residues which remain identical? Please explain for the benefit of the readers.

We thank the reviewer for these useful comments. We agree with the apparent counter intuitiveness of the sentence. However, it is the fact that we observed. The obvious question is why such apparently contrasting relationship exists. Perhaps, both coevolutionary and evolutionary changes are dynamic processes and for a given protein site at a certain point of evolutionary conservation status, highest coevolutionary connections are observed. This evolutionary conservation status of the site is perhaps selected and maintained. One probable explanation could be that these highly coevolved sites might act as 'coevolutionary hub' and therefore changes at these sites would affect many other connected sites. However, we must mention that in this study the higher conservation is with respect to other coevolving sites and not necessarily meant completely conserved sites.

Competing Interests: No competing interests were disclosed.

Referee Report 25 April 2017

doi:10.5256/f1000research.12138.r21733

? **Anna Panchenko**

Computational Biology Branch, National Institutes of Health, Bethesda, MD, USA

This paper expands on a previous study (Ref 32) and shows that the number of inter-residue coevolutionary relationships can be correlated with the evolutionary conservation of a protein site and protein family. The authors applied four different algorithms to calculate the coevolutionary relationships between sites and an overall trend observed in this study is confirmed by different methods. Interestingly, the absolute scale of site conservation for sites with the same number of coevolutionary relationships can differ drastically between methods (for example McBASC and MIp on Figure 1). I wonder if the sets of pairwise correlated sites overlap between different methods. I would also suggest using MISTIC server which can provide information on conservation, coevolution and structure mapping. The relationships between coevolution and diversity of protein families is interesting and intriguing, can it be related to the quality of alignments, one of the major factors defining the accuracy of coevolutionary detection algorithms? It is important also to discuss the difference between covariation and coevolution, the latter is not necessarily the cause, see some recent studies: PMID:25944916).

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 14 Jun 2017

Saikat Chakrabarti, Indian Institute of Chemical Biology, India

This paper expands on a previous study (Ref 32) and shows that the number of inter-residue coevolutionary relationships can be correlated with the evolutionary conservation of a protein site and protein family. The authors applied four different algorithms to calculate the coevolutionary relationships between sites and an overall trend observed in this study is confirmed by different methods. Interestingly, the absolute scale of site conservation for sites with the same number of coevolutionary relationships can differ drastically between methods (for example McBASC and MIp on Figure 1).

We thank the reviewer for the positive comments. We agree with reviewer's comment, the observed scale of values obtained from multiple coevolutionary programs varies a lot. The probable reason for such observation can be the algorithm used by individual programs for calculation of covariation/coevolution. We are providing our point-by-point response in the following.

I wonder if the sets of pairwise correlated sites overlap between different methods.

Figure S1 in supplementary file 1 shows number of coevolved pairs predicted (in blue, pink, green and yellow) by different programs (MIp, MCBASC, DCA and PSICOV, respectively) and common pairs (in black) between them.

I would also suggest using MISTIC server which can provide information on conservation, coevolution and structure mapping.

We thank the reviewer for the comment. MISTIC (mutual information server to infer coevolution) is an online server, hence running for large number of protein families is not feasible. However, as case studies, we have performed the analysis on MISTIC server for 20 protein families (including CD01291 and CD00164 families provided in the paper). Result of MISTIC server for CD01291 is available at http://mistic.leloir.org.ar/results.php?jobid=201705252335594338 and for CD00164 at http://mistic.leloir.org.ar/results.php?jobid=201705269510226.
Circos representation of result for both the families is as follows:

CD01291 (Pseudouridine synthases family):

Link:http://mistic.leloir.org.ar/Results/job201705252335594338/circos/circos201705252335594338.pn

CD00164 (Ribosomal protein S1-like RNA-binding domain):

Link: http://mistic.leloir.org.ar/Results/job201705269510226/circos/circos201705269510226.png

MI Circo is a sequential circular representation of the MSA and the information it contains. Coloured square boxes of the second circle indicate the MSA position conservation (highly conserved positions are in red, while less conserved ones are in blue).
Lines connect pairs of positions with MI greater than 6.5 (Marino Buslje *et al*, 2009). **Red** edges represent the top 5%, **black** ones are between 70% and 95%, and **gray** edges account for the remaining 70%.

Interestingly observed coevolutionary network predicted for both the families are similar to our study. Where CD01291 family has higher coevolutionary network connections for fewer variables sites whereas CD00164 family has less coevolutionary connections and overall less conserved.

Result for other families:
CD01424 (MGS_CPS_II):
http://mistic.leloir.org.ar/results.php?jobid=20170611337071057
CD01887 (Initiation Factor 2 (IF2):
http://mistic.leloir.org.ar/results.php?jobid=20170611409092676
CD03377 (Thiamine pyrophosphate (TPP family):
http://mistic.leloir.org.ar/results.php?jobid=20170611947015544
CD03278 (ATP-binding cassette domain of barmotin):
http://mistic.leloir.org.ar/results.php?jobid=20170611948108735
CD03481 (Transducer domain):
http://mistic.leloir.org.ar/results.php?jobid=2017061194942951
CD01357 (Aspartase):
http://mistic.leloir.org.ar/results.php?jobid=20170611950419609

CD00036 (Chitin/cellulose binding domains):
http://mistic.leloir.org.ar/results.php?jobid=20170614134138131
CD00089 (Protein kinase C-related kinase homology region 1 (HR1)):
http://mistic.leloir.org.ar/results.php?jobid=20170614138375373
CD04371 (DEP domain):
http://mistic.leloir.org.ar/results.php?jobid=20170614139192179
CD00052 (Eps15 homology domain):
http://mistic.leloir.org.ar/results.php?jobid=20170614140306460
CD00173 (Src homology 2 (SH2) domain):
http://mistic.leloir.org.ar/results.php?jobid=20170614142492212
CD01926 (cyclophilin_ABH_like domain):
http://mistic.leloir.org.ar/results.php?jobid=201706111058271452
CD04912 (ACT domains located C-terminal):
http://mistic.leloir.org.ar/results.php?jobid=20170611105511638
CD00164 (Ribosomal protein S1-like RNA-binding domain):
http://mistic.leloir.org.ar/results.php?jobid=20170614204299821
CD01714 (The electron transfer flavoprotein (ETF)):
http://mistic.leloir.org.ar/results.php?jobid=201706111053598165
CD00585 (Peptidase C1B subfamily):
http://mistic.leloir.org.ar/results.php?jobid=201706111052469189
CD04867 (TGS domain-containing YchF GTP-binding protein):
http://mistic.leloir.org.ar/results.php?jobid=201706111051354847
CD02014 (Thiamine pyrophosphate (TPP) family):
http://mistic.leloir.org.ar/results.php?jobid=20170611953291860

The relationships between coevolution and diversity of protein families is interesting and intriguing, can it be related to the quality of alignments, one of the major factors defining the accuracy of coevolutionary detection algorithms?

We thank the reviewer for bringing very important point of quality of alignment for coevolutionary analysis. Quality of alignment is major factor in the analysis and for this reason we have utilized manually curated CDD alignments.

It is important also to discuss the difference between covariation and coevolution, the latter is not necessarily the cause, see some recent studies: PMID: 25944916).

We thank the reviewer for providing the information. In this study we have not checked/compared the difference between the two concepts of covariation and coevolution. We have used different programs (MIp, McBASC, DCA and PSICOV) which calculate covariation among protein sites in tree-independent manner. In this study, it was assumed that observed patterns of covariation are caused by molecular coevolution and they were treated synonymously.

Competing Interests: No competing interests were disclosed.

Referee Response 15 Jun 2017

Anna Panchenko, National Library of Medicine & NIH, USA

I would like thank the authors for adequately responding to my comments.

Competing Interests: No competing interests were disclosed.