Protein-protein interactions can be predicted using coiled coil co-evolution patterns

Pablo Miera, Gregorio Alanis-Lobato, Miguel A. Andrade-Navarro

Faculty of Biology, Johannes Gutenberg University Mainz, Gresemundweg 2, 55128 Mainz, Germany
Institute of Molecular Biology, Ackermannweg 4, 55128 Mainz, Germany

ABSTRACT

Protein–protein interactions are sometimes mediated by coiled coil structures. The evolutionary conservation of interacting orthologs in different species, along with the presence or absence of coiled coils in them, may help in the prediction of interacting pairs. Here, we illustrate how the presence of coiled coils in a protein can be exploited as a potential indicator for its interaction with another protein with coiled coils. The prediction capability of our strategy improves when restricting our dataset to highly reliable, known protein–protein interactions. Our study of the co-evolution of coiled coils demonstrates that pairs of interacting proteins can be distinguished from not interacting pairs by means of their structural information. This hints at the potential of our strategy to predict new protein–protein interactions.

1. Introduction

Proteins interact with a wide variety of biomolecules within a cell, from small molecule substrates to other proteins. In particular, protein–protein interactions (PPIs) are involved in many biological functions, for example in regulating enzymatic activity, mediating the assembly of protein complexes, and in developmental control (Braun and Gingras, 2012). Thus, knowing whether two proteins interact or not can give us important information about their function. For these reasons, there is a growing variety of methods currently available to detect PPIs, either in vitro, in vivo or computationally (Srinivasa Rao et al., 2014; Alainis-Lobato, 2015), and PPIs with different levels of confidence (experimentally described, predicted or inferred from other species) are collected in databases like HIPPIE (Schaefer et al., 2012), STRING (Sklarczyk et al., 2011), BioGRID (Chatr-Aryamontri et al., 2015), MINT (Licata et al., 2012), HAPPI (Chen et al., 2009), and iRefWeb (Turner et al., 2010).

The physical interaction between proteins is sometimes mediated by matching coiled coil regions (Lupas and Gruber, 2005; Wang et al., 2012; Watkins et al., 2015). Here, we propose that this can be exploited to identify novel interactions between proteins.

Protein–protein interactions are expected to be conserved in evolution, in a similar manner as orthologous proteins are. Conserved interacting orthologs in two species, known as interologs (Wallout et al., 2000; Yu et al., 2004), can thus be used to transfer PPI information across species. Given a conserved PPI across multiple species, if it is mediated by coiled coils (CCs) in some species but not in others, then this can be reflected by the co-evolution of the interacting CCs.

CCs can be lost or gained during the evolution of protein families; if gained, they are added to proteins rather than converted from existing alpha helices (Rackham et al., 2010). Such gains and losses make CCs good targets for analysis of correlated evolution in CC interacting partners under the assumption that CC loss or gain in one interacting partner will correlate with the corresponding CC loss or gain in the other interacting partner. Detection of such correlated gains and losses of CC would then support the prediction of PPIs.

Computational prediction of PPIs has been previously proposed using either structural (Aloy and Russell, 2003; Gonzalez and Liao, 2010; Chang et al., 2010) or evolutionary information (Pazos and Valencia, 2001; Jothi et al., 2005). Here, we use both strategies at the same time, as a proof of concept that the evolutionary study of CC in interologous proteins can be used as an indicative of a PPI. Our results show indeed a differential behavior in the correlated evolution of CCs between protein pairs known to interact in humans and randomly sampled protein pairs.

2. Methods

2.1. Description of the dataset

The full set of 20,166 proteins from the human protein–protein interaction (PPI) database HIPPIE v1.8 (Schaefer et al., 2012) was selected as the query dataset. A Reciprocal-Best-Hits-BLAST (RBHB)
strategy with default parameters was locally used to look for orthologous proteins to the initial human dataset in the following 50 complete reference proteomes: Escherichia coli s. K12, Bacillus subtilis s.IS68, Helicobacter pylori s. ATCC, Deinococcus radiodurans s. ATCC, Synechocystis sp. s. PCC 6803/Kazusa (Bacteria); Halobacterium salinarum s. ATCC, Korarchaeum cryptofilum s. OPF8, Nanoarchaeum equitans, Sulfolobus solfataricus s. ATCC (Archaea); Trypanosoma cruzi, Monosiga brevicollis, Dictyostelium discoideum, Giardia lamblia, Paramecium tetraurelia, Plasmodium falciparum isolate 3D7, Reticulomyxa filosa, Thalassiosira oceanica, Volvox carteri, Chlamydomonas reinhardtii, Selaginella moellendorffii, Oryza sativa subsp. japonica, Arabidopsis thaliana, Encephalitozoon cuniculi s. GB-M1, Schizosaccharomyces pombe, Saccharomyces cerevisiae s. ATCC, Klyveromyces lactis s. ATCC, Neurospora crassa s. ATCC, Rhizophus delemar, Trichoplax adhaerens, Strongylocentrotus purpuratus, Nematostella vectensis, Schistosoma mansoni, Ciona intestinalis, Branchiostoma floridae, Anolis carolinensis, Takifugu rubripes, Danio rerio, Xenopus tropicalis, Taeniopygia guttata, Gallus gallus, Ornithorhynchus anatinus, Bos taurus, Mus musculus, Homo sapiens (Eukarya). Proteomes were downloaded from the UniProt Knowledgebase database on May 2015 (The UniProt Consortium, 2015).

2.2. Coiled coil prediction

A CC motif consists of two or three alpha-helices twisted around one another, in either a parallel or anti-parallel configuration, characterized by a hydrophobic core and its surrounding electrostatic periphery (Lupas et al., 1991; Lupas and Gruber, 2005; Wang et al., 2012; Shin et al., 2014). The prediction of CC regions is possible due to their characteristic physicochemical properties and a typical heptad repeat pattern. Several tools have been developed for the prediction of such regions, using the comparison to proteins with known CCs to refine their predictions (Lupas et al., 1991; Walshaw and Woolfson, 2001; McDonnell et al., 2006; Fariselli et al., 2007; Trigg et al., 2011). The orthologous proteins were analyzed with the standalone version of PCOILS v1.01 (Lupas et al., 1991) to predict whether each protein has at least one CC region or not, using as parameters: window_size=21, and a minimum local probability of 90%. For each protein in the query dataset, a vector was built with the obtained results; one value per proteome: no orthologous protein (-), orthologous protein with CC (1), or orthologous protein but no CC (0). This means that each protein x is represented by a 50-element vector with entries equal to −1, 0 or 1 (Fig. 1).

2.3. Scoring of the vector pairs and evaluation of the scoring methods

Given two proteins A and B, each one represented by the 50-element vector described above, Aij being the total number of cases in which an entry of A is i, for i=(-1,0), Mij specifying the total number of cases in which an entry of A is i and an entry of B is j and P(a,b) being the joint probability of the particular values a and b occurring together in vectors A and B; ten different scoring methods were used to evaluate the likelihood that A and B interact. Four of these methods are naive (Eqs. 1–4) and six correspond to well-established measures from the field of information theory (Eqs. 5–10). The scoring methods are defined as follows: a) Method 1, sum all the coincidences [0 0] and [1 1] (Eq. (1)); b) Method 2, sum all the coincidences [0 0], [1 1], and [− −] (Eq. (2)); c) Method 3, same as Method 2 but penalizing mismatches [0 1] (Eq. (3)); d) Method 4, same as Method 3 but penalizing mismatches [0 −] or [1 −] (Eq. (4)); e) Dice score using the [1] values as reference (Dice, 1945) (Eq. (5)); f) Jaccard Index using the [1] values as reference (Jaccard, 1912) (Eq. (6)); g) Joint Entropy (Shannon, 1948) (Eq. (7)); h) Mutual Information (Shannon, 1948) (Eq. (8)); i) Dice score using the [0] values as reference (Eq. (9)); and j) Jaccard Index using the [0] values as reference (Eq. (10)). Note that the Dice and Jaccard indices measure how much overlap there is between vectors A and B, using different normalization factors. On the other hand, Joint Entropy and Mutual Information are associated with the amount of information held by A and B. While Joint Entropy quantifies the average amount of novel information that we obtain from observing vectors A and B, Mutual Information measures the information that they share (Shannon, 1948).

\[
Naive_{0,1}(A,B) = M_{11} + M_{00} \\
Naive_{0,-1}(A,B) = M_{11} + M_{00} + M_{--} \\
wAnaive_{0,1}(A,B)=M_{11}+M_{00}+M_{--}−0.25(M_{01}+M_{10}) \\
wBNaive_{0,1}(A,B)=M_{11}+M_{00}+M_{--}−0.25(M_{01}+M_{10})−0.1(M_{1--}+M_{0--}) \\
Dice_{1}(A,B) = \frac{2M_{11}}{A_{1} + B_{1}} \\
Jaccard_{1}(A,B) = \frac{M_{11}}{M_{11} + M_{01} + M_{10}} \\
H(A,B) = −\sum_{a \in A, b \in B} P(a,b) \log P(a,b) \\
I(A,B) = H(A) + H(B) − H(A,B) \\
D_{ice_{0}}(A,B) = \frac{2M_{00}}{A_{0} + B_{0}}
\]

![Fig. 1. Information vector built for each protein query (A, B), with three different possible values for each proteome (n): no ortholog protein (-), ortholog protein with CC (1), and ortholog protein without CC (0).](image-url)
\[ Jaccard(A, B) = \frac{M_{00}}{M_{00} + M_{01} + M_{10}} \]  

Eqs. 1–10. Scoring methods used to evaluate the vector pairs. “–”: No orthologous protein; “1”: Orthologous protein with CC; “0”: Orthologous protein without CC.

Interacting pairs of human proteins were obtained from HIPPIE v1.8 (Schaefer et al., 2012). A ROC curve was calculated for each of the scoring methods, using the HIPPIE pairs present within our dataset as the true set scores (1122 pairs), and the rest of the pairs of human proteins as the false set scores (1,109,673 pairs). The area under the ROC curve (AUC) was computed to evaluate the methods. Homodimers were left out of possible protein pairs for obvious reasons.

2.4. Paralogy information

The information about the paralog pairs was extracted on 9 Dec 2015 from the Ensembl database (Flicek et al., 2014), using the GRCh37 human archive. Only 1357 pairs from the complete dataset (1,110,795 pairs) are annotated as paralogs.

3. Results

3.1. Generation and filtering of informative vectors

Coiled coil (CC) mediated protein-protein interactions (PPI) are thought to be evolutionarily conserved (Wang et al., 2012). Emergence of CC motifs can occur as addition of new elements to existing domain organizations. If a new CC participates in a PPI with another CC, it can be expected that both interacting CCs will co-occur in the same organisms. One such example can be described for the interaction between KMD1A and RCOR1 (Fig. 2). This evolutionary correlation of CCs in two proteins can be used as a hint that these proteins interact.

In order to capture such evolutionarily conserved correlated emergence of CC regions in pairs of proteins, we examined a wide range of organisms, from unicellular organisms like Escherichia coli to the human proteome, covering a list of 50 well-studied model organisms (see Methods for a detailed list). The orthologs to all human proteins covered in the PPI database HIPPIE (Schaefer et al., 2012), a total of 20,166 proteins, were located using the Reciprocal-Best-Hits-BLAST strategy (RBHB). For every ortholog found, the standalone version of PCOILS (Lupas et al., 1991) was used to predict whether it has any CC or not, independently of its position.

This particular PPI was not evaluated in our analysis because the number of species with detected CCs was too high according to the cutoff used (see Methods for details).

The results form vectors that can have three different possible values for each proteome: no ortholog protein found (−), ortholog protein with CC (1), and ortholog protein without CC (0) (Fig. 1). The distribution of the values per vector show that in many cases most of the orthologs in a vector do not have CC (Fig. 3). They are therefore not informative vectors. Vectors with a balanced combination of values are needed in order to predict CC-mediated PPIs. The initial set of vectors was filtered to leave out those with less than ten and more than thirty positions with orthologs and CC, and orthologs but no CC. This procedure reduced the number of vectors from 20,166 to a filtered subset of 1491. The new set of vectors shows an adjusted proportion of values (48.36% ortholog but no CC, 44.45% ortholog and CC, 7.19% no ortholog), in contrast to the initial unbalanced number of orthologs without CC (78.17%, 13.51% and 8.32%, respectively).

3.2. Evaluation of vector pairs

Given the filtered set of 1491 vectors, there are 1,110,795 unique pairs to score. Those pairs were scored using four naive and six information theory scoring methods (see Methods). Naive methods are based on counting pairwise identities between the vector values (weighted or not), while the information theory methods rely on standard measures for comparing sample sets (Jaccard Index, Dice score, Joint Entropy, Mutual Information, and a variation of both the Jaccard and Dice indices).

The HIPPIE database of human PPIs scored according to experimental evidence (Schaefer et al., 2012) was used to evaluate the scored vector pairs. The 1,110,795 scored pairs were compared to the complete HIPPIE dataset (88,397 interacting pairs), finding that they have 1122 pairs in common. Using these pairs as the true protein-protein interacting pairs, a ROC curve was constructed based on the scores computed by the ten scoring methods for each potential protein interaction (Fig. 4). The Jaccard Index scoring method yields the best predicting performance (AUC=0.6021).

The selected dataset introduced an unwanted bias in the true pairs subset; while trying to predict human PPIs using CC co-evolution patterns, the dataset is not restricted solely to CC mediated interacting pairs. To check whether the prediction can be improved by focusing on these specific type of interactions, the true dataset was filtered to keep only the pairs with both human proteins predicted to have CCs. The Jaccard Index method was used to score the new subset of 882 true pairs. There is an improvement in the predictions (AUC=0.6805), limited to human protein pairs with CC (CC_both, Fig. 4). The distribution of the scores (CC-scores) is consistently different between the pairs annotated as true and as false in the updated dataset (full dataset in Supplementary File 1; Fig. 5). In a 0–1 range, their quartile distribution shows an enrichment of higher CC-scores in the true pairs: 0.4062, 0.4815 and 0.5455 in the true pairs (first, second and third quartile scores, respectively) versus 0.3143, 0.4062, and 0.4848 in the false pairs.

The updated dataset of true protein-protein interaction pairs includes pairs which, in human, are expected to interact via their CC; and thereby they can be predicted with a higher confidence using evolutionary information. The current approach is not flawless, as we ignore protein pairs that may interact in some organisms mediated by CC but not in human. However, as previously described for these cases, the results still show that the evolutionary information about CC regions may be taken as an indicative of the interaction between two human proteins.

3.3. Study of the paralog bias in the results

Up to when a paralog pair arose in evolution, both proteins share their ancestor orthologs. This entails that the vectors of paralog pairs could be biased, as they might have identical entries due to their shared ancestry. Since paralogs also tend to interact with each other (Pereira-Leal et al., 2007), we need to study if a significant part of our success in detecting PPIs using CC vectors could stem from this bias. To study in detail how the paralog pairs influence the evaluation of the vectors, several combinations between the true interacting pairs and the paralog pairs within the full set of vector pairs were evaluated (Table 1). Condition “a” indicates the current prediction result, taking into account all possible pairs. When the set of PPI pairs is broken down using the paralogy information, results differ considerably; predictions for those interacting pairs that are also paralogs have a much higher AUC (condition “b”, AUC=0.9659) than for those that are not paralogs (condition “c”, AUC=0.6619). The latter result is confirmed when paralog pairs are taken out of the subset of false pairs (condition “d”, AUC=0.6625).

As expected, paralogs have significantly similar CC vectors, which allows a good prediction of paralog pairs using them (condition “e”, AUC=0.8802). On the other hand, the fact that PPIs for paralogs can be even better predicted than paralogs using the CC vectors (condition “b”, AUC=0.9659) stresses out that there is additional important information given by PPI pairs in the CC vectors that cannot be explained based solely on paralogy.
Fig. 2. Correlated CC evolution in RCOR1 and KDM1A protein families. (A) 3D structure of the complex between human RCOR1 and KDM1A/LSD1 (PDB:2IW5; Yang et al., 2006). These two proteins interact through CC regions. (B) Multiple sequence alignment of the interacting CC regions in sequences of these families in representatives from fungi (S. pombe), plant (A. thaliana), and animals (N. vectensis, C. intestinalis, C. milii, D. rerio, L. chalumnae, mouse and human). These CCs are clearly present in animals but absent from the plant and fungi proteins, which have either a gap at the corresponding position, or non-homologous sequence. RCOR1 features: CC (orange) and SANT domain (red); KDM1A: CC (cornflower blue) and SWIRM (cyan). Structure represented using the UCSF Chimera package. Multiple sequence alignments are provided as supplementary material, and were generated using MUSCLE (Edgar, 2004) with default parameters followed by manual editing.

Fig. 3. Frequencies of vectors of orthologs by numbers of proteins predicted to have CCs. Results are shown for the complete set of vectors (white) and for the filtered subset (black). CC: Coiled Coil.

Fig. 4. ROC curves generated to evaluate the vector pairs using different scoring methods (see Methods). The area under the curve (AUC) is shown for each curve. MUSCLE (Edgar, 2004) with default parameters followed by manual editing.
The current evaluation of the results is based on the election of HIPPIE as the human PPI reference database, both on the use of the initial protein dataset and for the subsequent true/false annotation. To assess the impact of these decisions in the outcome of the study, the CC-scores of the interacting protein pairs were evaluated based on the HIPPIE score. The HIPPIE score is calculated combining the number and reliability of evidence supporting an interaction (Schaefer et al., 2012).

Several HIPPIE score cutoffs were used to select increasingly restricted subsets of interacting proteins. The CC-scores associated to each subset were evaluated, and the results show that the predicting precision of the CC co-evolution method improves when the cutoff is more stringent (Table 2). A major downside of rising the cutoff is the reduction of the number of pairs to be considered as true. The results suggest that the quality of the true dataset has a great impact in the evaluation of the CC-scores, and therefore in the assessment of the predictive power of the proposed method. Pairs predicted to interact with a high confidence are accordingly identified with a higher CC-score. As the scores calculated in the HIPPIE database reflect the reliability of the experimental evidence for a given PPI, this indicates that if the information improves the predictive power of the CC-score will also improve.

4. Discussion

Coiled coils (CC) often mediate protein-protein interactions (PPIs) (Lupas and Gruber, 2005; Wang et al., 2012). CCs emerge in addition to existing protein architectures (Rackham et al., 2010). Thus, we hypothesized that the emergence of the interacting CCs of interaction partners should be correlated, and thus correlated CC presence or absence in diverse species can be used as a predictor of the interaction between interologs.

The described method is limited by the number of proteomes used to generate the unique orthology vector for each human query protein. One future improvement of the method would involve the use of many more additional proteomes to make the vector more specific to detect certain co-evolution patterns. But given the current genome sequencing hotspots it would entail a bias towards a particular taxonomic group, as the sequenced genomes are not equally distributed along the tree of life (Hug et al., 2016). Efforts should be made in this direction to cover currently under-represented parts of the tree of life, which would allow for the generation of a large, balanced and completely sequenced proteome dataset in order to build specific vectors to compare to. Conclusions drawn from much more complex vectors would be more meaningful and therefore the prediction capabilities of the method would improve significantly.

We have developed an approach thought to be valid to predict new protein-protein interactions. Although this is a first step, the method is expected to improve by using more proteomes to make the comparable vectors more specific, as well as by progressing in the knowledge of true PPIs that could be used as gold standards for better evaluations. Advances in both fronts will equally improve the prediction capability of the method.

Author’s contributions

PM and MAN conceived the project. PM collected the data and performed the analyses, GAL was in charge of vector pair scorings. MAN supervised the project. PM drafted the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The authors thank all members of the CBDM group for helpful and constructive discussions. This work was supported with funds from the Center for Computational Sciences Mainz (CSM, Johannes Gutenberg University of Mainz, Germany).

The authors have no conflict of interest to declare.
Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jtbi.2016.11.001.

References


