Genome Biology

Coev2Net: a computational framework for boosting confidence in high-throughput protein-protein interaction datasets

Genome Biology 2012, **13**:R76 doi:10.1186/gb-2012-13-8-r76

Raghavendra Hosur (rhosur@mit.edu)
Jian Peng (pengjian@ttic.edu)
Arunachalam Vinayagam (vinu@genetics.med.harvard.edu)
Ulrich Stelzl (stelzl@molgen.mpg.de)
Jinbo Xu (j3xu@tti-c.org)
Norbert Perrimon (perrimon@receptor.med.harvard.edu)
Jadwiga Bienkowska (jbienkowska@gmail.com)
Bonnie Berger (bab@mit.edu)

ISSN 1465-6906

Article type Method

Submission date 3 May 2012

Acceptance date 14 August 2012

Publication date 31 August 2012

Article URL http://genomebiology.com/2012/13/8/R76

A computational framework for boosting confidence in high-throughput protein-protein interaction datasets

Raghavendra Hosur[1], Jian Peng[1,2], Arunachalam Vinayagam[3], Ulrich Stelzl[4], Jinbo Xu[2], Norbert Perrimon[3,5], Jadwiga Bienkowska[6,8] & Bonnie Berger[1,7,8]

1. Computer Science and Artificial Intelligence Laboratory, 32 Vassar Street, MIT, Cambridge-02139, MA.

2. Toyota Technological Institute, 6045 S. Kenwood Ave., Chicago-60637, IL.

3. Department of Genetics, 77 Avenue Louis Pasteur, Harvard Medical School, Boston-02115, MA.

4. Otto-Warburg Laboratory, Ihnestraβe 63-73, Max Planck Institute for Molecular Genetics, Berlin-D14195, Germany.

5. Howard Hughes Medical Institute, 20 Shattuck Street, Boston-02115, MA.

6. Computational Biology group, Biogen Idec, 14 Cambridge center, Cambridge-02142, MA.

7. Department of Mathematics, 77 Massachusetts Avenue, MIT, Cambridge-02139, MA.

8. Corresponding Authors, email: bab@mit.edu, jbienkowska@gmail.com.

Abstract

Improving the quality and coverage of the protein interactome is of tantamount importance for biomedical research, particularly given the various sources of uncertainty in high-throughput techniques. We introduce a structure-based framework, Coev2Net, for computing a single confidence score that addresses both false-positive and false-negative rates. Coev2Net is easily applied to thousands of binary protein interactions and has superior predictive performance over existing methods. We experimentally validate selected high-confidence predictions in the human MAPK network and show that predicted interfaces are enriched for cancer – related or damaging SNPs. Coev2Net can be downloaded at http://struct2net.csail.mit.edu.

Background

Protein-protein interactions (PPIs) play a critical role in all cellular processes, ranging from cellular division to apoptosis. Elucidating and analyzing PPIs is thus essential to understanding the underlying mechanisms in biology. Indeed, this has been a major focus of research in recent years, providing a wealth of experimental data about protein associations [1-9] Current PPI networks have been constructed using a number of techniques such as yeast-two-hybrid (Y2H), coimmuno or coaffinity purification followed by mass spectroscopy (coIP or coAP/MS) and curation of published low-throughput experiments [10-16]. Despite this tremendous push, the current coverage of PPIs is still rather poor (e.g. < 10% of interactions in humans) [17]. Additionally, despite considerable improvements in HTP techniques, they are still prone to spurious errors and systematic biases, yielding a significant number of false-positives and false-negatives [18,19,20,21]. This limits our ability to assess the true quality and coverage of the "interactome" [22,23,24].

Akin to sequencing of the human genome, complete high-confidence descriptions of protein-protein interactions is a fundamental step towards human interactome mapping [22,25]. Also present are the challenging issues of data quality and size estimation, as encountered in the human genome project [24,26,23]. However, unlike the challenges faced previously with sequencing, we still do not understand the rules of association of protein molecules, and are unable to distinguish between biophysical interactions, true biological interactions and false-positives [20]. Further unresolved questions as to the proportion of experimental artifacts in the current interactomes are coming to light as a consequence of the low degree of overlap between data curated from multiple high-throughput (as well as low-throughput) studies [27].

Several attempts have been made to characterize the quality of the interactions obtained from HTP experiments [23,28,7,24,29,30,31]. Experimental methods aim to limit false discovery by performing multiple iterations of the screen, which are time-consuming and expensive [29]. Secondary data, such as co-expression, co-localization, ontology correlation, topological features and orthology information are often used to further improve confidence in predicted interactions [32,33]. In addition to non-trivial correlations between these features (i.e. co-expression need not imply interaction), this data is not complete for all proteins. Furthermore, as more and more genomes are sequenced, only a fraction of proteins will have additional data to complement any experimental HTP study. Techniques

developed from integrating interactions observed in common across multiple secondary experimental assays of an initial network are laborious, expensive and time-consuming. Moreover, as suggested by Venkatesan et al. and Cusick et al. [27], the low overlaps achieved across different datasets highlight the differences in sampling and biases in experimental techniques rather than pinpoint the true interactions. Further, in many experimental methods, the confidence of observations is evaluated for that specific technique – they are seldom generalizable. Thus cost-effective and high-confident strategies are clearly required to complete the human interactome.

Recently, a number of algorithms have been developed to predict protein interactions by integrating complementary data such as sequence features and structural features [34-42,12]. Also recently, computational approaches to PPI prediction using structural information have been gaining much attention due to the rapid growth of the Protein Data Bank (PDB)[32,35,43-65]. An important advantage of structure-based approaches is their ability to identify the putative interface, thereby providing more information than any other high-throughput method. The common strategy of structure-based methods is to find a best-fit template complex structure for the two query sequences; the prediction is then based on the similarity of the two proteins to the template complex. Threading based approaches extend coverage further "into the twilight zone", making accurate predictions even when there is low sequence similarity (typically <40%) between the query proteins and the best-fit template complex [32,66,49]. However, to the best of our knowledge, there have been no studies that integrate HTP techniques with PPI prediction algorithms to quantitatively address both false-negative (FNR) and false-positive (FPR) issues.

In this paper, we introduce a general framework to predict, assess and boost confidence in individual interactions inferred from a HTP experiment. Our contribution is three-fold– 1) we develop a novel computational algorithm to quantitatively predict interactions, given just the protein sequences; 2) we show how the algorithm can be used in a general framework to quantify confidence in observed interactions; and 3) we demonstrate the utility of our structure-based framework in providing biologically significant additional information about binding sites, which is not provided by any other HTP method (either computational or experimental). We first validate our method on a high-confidence network in the recently investigated human Mitogen Activated Protein Kinase (MAPK) interactome [67,68]. We experimentally validate predicted high-confidence interactions for the MAPK

interactome using a complementary assay and show that the concordance between prediction and experimental validation is as good as the overlaps achieved in previous protocols involving multiple secondary assays [25]. Finally, we show that the interfaces predicted by our algorithm are enriched for functionally important sites in the context of signaling networks; and utilize this information to hypothesize a novel regulatory mechanism involving cross talk between the insulin and stress-response pathways via interactions between proteins MAPK6, YWHAZ and FOXO3 proteins.

Results

The Coev2Net framework for quantifying confidence in protein interactions

We developed Coev2Net (Fig 1), a framework for assessing confidence in protein interactions. To quantify confidence in an interactome, we incorporate high-confidence data sources, namely low throughput interactions and structural information. The framework gives a confidence score for each interaction, along with a predicted model of the binding interface for the proteins (Fig 1).

Inputs to the framework are a high-confidence network (usually much smaller than the HTP screen) and the interactions identified from the HTP experiment for which one wishes to quantify confidence. For every pair of interaction in the HTP screen, Coev2Net provides a score to assess their likelihood of being co-evolved from interacting homologous sequences (see Methods). To do this, Coev2Net first predicts a likely interface model for the two proteins, by threading [69] the sequences onto the best-fit template complex in our library. It then computes the likelihood of co-evolution of the two proteins (i.e. of the predicted interface) with respect to a probabilistic graphical model induced by the aligned interfaces of artificial homologous sequences (Fig 2, see Methods and Supplemental Material). By generating artificial sequences, we enrich the interfacial sequence/structure profiles for those protein-pairs with sparse interacting-sequence profiles and thus improve protein interface scoring accuracy. Note that this enrichment is carried out for all protein pairs, irrespective of the information content in their individual sequence profiles. These PGM scores are then input into a classifier trained on a small high-confidence network to compute a score between 0 and 1, representing the confidence of our method in that interaction (Fig 1). High-scoring interactions can then be investigated further using a

secondary experimental assay or taken as true positives for subsequent analyses. Additionally, since Coev2Net is a structure-based algorithm, it also produces as output a putative interface for the interacting pair (Fig 2). This information can be analyzed to design site-directed experiments to further characterize the specificity of the interaction.

Benchmarking Coev2Net

SCOPPI: We first benchmark Coev2Net on SCOPPI [70], a protein complex database. The database is divided into interacting family pairs for which multiple complexes have been solved. Rigorous cross-validation tests on the database indicate that Coev2Net achieves high accuracies, thereby validating our approach of modeling interface co-evolution as a high-dimensional sampling problem (Additional file 1, Fig S3). For the cross-validation tests, we considered only those family pairs in SCOPPI that have at least three non-redundant (sequence id < 50%) complexes. We randomly selected one as the test complex and used the other complexes within our Coev2Net protocol to simulate interacting homologs and construct the probabilistic graphical model (Fig 2). We additionally compared Coev2Net's performance on the SCOPPI dataset to another structure-based method, PRISM [45]. PRISM first identifies similar templates to two query structures by structural alignment. The final prediction is based upon the energy of complex formation calculated by docking these two predicted interfaces. We find that Coev2Net's performance, measured in terms of sensitivity and specificity, is much better than PRISM's on this dataset, (Additional file 1, Fig S3).

Furthermore, Coev2Net also performs well on SCOPPI family pairs not having more than two non-redundant complexes, indicating Coev2Net's ability to deal with limitations of both structural and sequence training data (Additional file 1, Fig S3).

MAPK interactome validation

To test the framework's ability to predict interactions for which there is often no structural data available and to assign confidence values to interactions, we re-trained Coev2Net on a high-quality human MAPK PPI network [67] and tested it on another high-quality MAPK network [68](Fig 3A,B,C). Oddly, these two MAPK networks are almost disjoint with only 6 overlapping interactions out of 4904 total interactions (Fig 3A). In the Bandyopadhyay set [67], we could make predictions for 461 interactions; in the Vinayagam set [68], 1025 interactions, and in the negatome (PDB-negative set), 330 non-interactors. To check for known complexes in the two MAPK networks, for each interaction, we ran BLAST against

the entire PDB to identify homologous complexes. We were able to find only 22 pairs for which a solved homologous complex exists in the PDB (we used a E-value cutoff of 1e-40). On the other hand, our threading-based approach can make predictions for ~1500 interactions in the MAPK networks, indicating that our method extends predictions to those pairs for which a clear homologous complex does not exist. The Bandyopadhyay set was further divided into a "core" set of interactions (640), of which we could make predictions for 173 pairs. The definitions for core set and non-core set were taken as in the original citation [67]. This core set of interactions contains high-confidence interactions that are conserved in yeast [67].

To test the accuracy of Coev2Net's predictions, we first validated our method via 5-fold cross-validation on the high-confidence core set of interactions in the Bandyopadhyay set (Fig 3C). In addition, to assess the contribution of co-evolutionary profiles for PPI predictions, we compared the performance of our method to Struct2Net and a "baseline" classifier that is trained on just the threading-based features (no inter-protein features). Note that all methods are evaluated on the same dataset (the core set). Fig 3C clearly shows that Coev2Net accurately predicts interactions even when only a distant homologous complex is available and thus fills the existing gap in structure-based methods for PPI prediction. In addition, Fig 3C also shows that including long-distance correlations as in Coev2Net aids in PPI prediction as compared to other threading-based methods.

We trained our final classifier on the entire Bandyopadhyay core data set, and predicted interactions in the Vinayagam dataset. For the predictions made for the latter dataset, we found that the experimentally validated coverage of our method (~55% with a confidence-score cutoff of 0.6) is significantly higher than that reported by other prediction methods based on conservation, genomic data, GO annotation and literature extractions (~14% to ~28%) [29], although each method was evaluated on a different network. Here, coverage is defined as the percentage of total predicted interactions for which we make a positive prediction and that were validated experimentally in the Y2H screen (571 predicted positive out of 1025 in the Vinayagam dataset). The cutoff of 0.6 was chosen since it corresponds to the maximum specificity and sensitivity of the logistic-regression classifier on the Bandyopadhyay core dataset.

Moreover, our predicted confidence scores are highly correlated with the experimental observation frequencies of Y2H screens on this network (Vinayagam dataset). To assess

significance, we divided our predictions into high confidence and low confidence based on the probability cutoff of 0.6. To categorize interactions as true positive (TP) or true negative (TN) in the Y2H screens, we assumed the cutoffs employed in Schwartz et al. (for a False Discovery Rate FDR < 5%, TP interactions should be observed at least twice when tested with <5 independent assays, and at least three times when tested with more assays)[29]. We then populated a 2x2 contingency table to test for association between our predicted label (interacting or non-interacting) and experimentally predicted label. We find that the predicted interactions correlate (P-value < 0.01, Fisher's test) with those deemed likely true positives from an experimental standpoint. Encouragingly, the percentage of our framework's predicted TP interactions that are confirmed positive (from an experimental standpoint) in the Vinayagam dataset is roughly 52% (294 TP, 571 predicted positive, a two-fold increase compared to previous methods on Y2H retesting of computational predictions [29]. Alternatively, training Coev2Net on the high confidence network in the Vinayagam dataset and testing it on the Bandyopadhyay core network yields similar results. By predicting only a fraction of interactions with high confidence, Coev2Net enables us to focus on only the most likely interactions, enabling a more accurate understanding of the biology (Fig 3B).

Experimental validation of predictions

The confidence scores given by our framework can be used to design additional experiments to enhance the quality of the initial interactome. We tested 19 randomly chosen high confidence interactions (confidence score > 0.6) using a complementary assay (LUMIER)[71]. Each pair, along with a control, was tested at least 3 times using the LUMIER assay. To confirm an interaction, the average result (i.e. fold change in luciferase intensity [RLU] as measured in a TECAN Infinite M200 luminescence plate reader) across the repeats had to be greater than 1.5 times the control. 14 interactions of the 19 interactions exhibited luciferase intensity greater than 1.5 times the control (Fig 3D). Additionally, if the repeat experiments were too variable to confidently assess the interaction (as measured using a z-score), the interaction pair was discarded. The z-score is calculated as:

$$z_{LUMIER} = \frac{\overline{RLU} - \overline{RLU_{control}}}{\sigma_{RLU}}$$

Eight out of the 19 interactions were discarded in this way as they registered a z-score of less than 1.5 and were deemed too variable. For additional experimental details we refer the readers to a more comprehensive interactome mapping analysis in [72]. Notably, 10 out of the remaining 11 were confirmed as true interactions i.e. registering average intensity above 1.5 times the control. Overlaps achieved by our method compare favorably with previous approaches, such as Braun et al. [25], in which an initial positive reference set (PRS) was re-tested experimentally using a LUMIER assay (Table 1). Furthermore, we evaluated the sequence identities between the interacting sequences and the templates used for predicting their interaction (see Table 1, Supp. Info). Interestingly, we find that all of them have a medium to low average sequence identity (15-30%), indicating that Coev2Net yields accurate predictions even in the "twilight zone" of sequence identities, where traditional homology methods usually fail. For example, IBIS [73], another homology/structure-based method can detect only 2 pairs from the 10 detected by Coev2Net and experimentally validated by the LUMIER assay.

Abundance of missense SNPs at predicted interfaces

In addition to the confidence scores, Coev2Net also provides a putative interface for the interaction. These interfaces can yield novel mechanistic insights into the protein-protein interaction and provide hypotheses about disease-associated mutations that occur at the interface. Missense SNPs occurring at the interface can potentially disrupt the interaction between the proteins, leading to abnormal functioning of the cell. We analyzed the predicted interfaces for existence of PolyPhen2 annotated missense mutations in dbSNP (build 131) [74]. PolyPhen2 classifies a SNP as "benign", "probably damaging", "possibly damaging" or "unknown" based on various features including conservation score, monomeric structure score (when available) and physicochemical properties [75,76]. It does not however account for SNPs occurring in potential interacting regions. Interestingly, SNPs annotated as damaging by PolyPhen2 are preferentially observed at the interface as compared to non-interfaces ($P = 0.0075$, Fisher's exact test, Fig 4A). Furthermore, if we take into account the number of interface and non-interface sites, we find that the predicted interfaces are enriched for damaging SNPs as compared to the rest of the protein ($P < 7e\text{-}8$, Fisher exact test). The same analysis with SNPs classified as benign by PolyPhen2 does not show up as highly significant ($P = 0.06$). We further analyzed the distribution of the SNPs in terms of their density at the interface and non-interface. Here again, we find that damaging

SNPs are preferentially located on the interface. We find that the average density of damaging SNPs at the predicted interfaces is significantly higher than their density at non-interface positions (Fig 4B; P < 1e-10, Mann-Whitney test); a bias also observed by Wang et al. recently [63]. For benign SNPs, the average density at the interface is lower than that at non-interfaces (Fig 4B; P < 1e-10, Mann-Whitney test). These analyses show that there is an evolutionary pressure to admit only benign SNPs at the interface, since any potentially damaging SNP will hinder the interaction.

To investigate the structural distribution of annotated mutations, we analyzed somatic mutations characterized in cancer to see if there is any preference for their location on the protein. We analyzed annotated mutations in the coding region deposited in the Cosmic database for their predicted location [77]. We only considered mutations that are annotated as either synonymous or missense. Interestingly, for these mutations we find that missense mutations are more prevalent on average at the PPI interface than synonymous mutations (P < 10e-20, Mann-Whitney test) (Fig 4C). This suggests that these mutations might be responsible for disruption of protein-protein interactions and the aberrant molecular signaling associated with cancer.

Finally, we looked at the predicted locations for some of the un-annotated mutations in kinases (from the MoKCa database [78]. As an example, we considered the BRAF protein as it contained the highest number of annotated mutations in the database. Coev2Net predicts an interaction between BRAF and PAK2, using the template structure 1G3N (chains E and F). Fig 5A shows the predicted interface for this interaction, with the annotated (magenta) and un-annotated (dark blue) mutations indicated. The presence of these mutations at the interface of the interacting proteins gives us an added insight into the investigation of such variations. Further study using this information can provide mechanistic details about how such mutations disrupt normal cellular signaling.

Novel potential cross-talk regulatory mechanism

Phosphorylation sites have been observed to be enriched at interfaces in solved structures [79]. This observation has mechanistic implications as the PPI can be used as an additional regulatory mechanism for phosphorylation, or the interaction could be a precursor to phosphorylation. An example for such a mechanism is found in the signaling protein YWHAZ [80]. Its phosphorylation is regulated by its dimerization, which buries the

phospho-sites on YWHAZ [81]. Our predictions revealed an interesting observation that suggests similar regulatory mechanisms in the MAPK interactome. Coev2Net predicts an interaction between MAPK6 and YWHAZ. Both are important signaling proteins, with much known about YWHAZ, including the experimental observation that MAPK8 regulates phosphorylation at S184 [82]. Relatively less is known about MAPK6's function and its substrates [83]. However, it is known that S189 is a phospho-site regulated by PAK1, PAK2 and PAK3 [84,85,86]. Interestingly, we found that the phosphorylation sites for both MAPK6 (S189) and YWHAZ (S184) lie within the predicted interface for the interaction (Fig 5B). This structural observation could imply that the interaction regulates downstream activities of MAPK6 and YWHAZ by controlling their phosphorylation. The most likely mechanism is that MAPK6 phosphorylates YWHAZ, thereby preventing its dimerization and regulating downstream activities of YWHAZ. Additionally, Coev2Net also predicts an interaction between MAPK6 and FOXO3. From a signaling context, these observations suggest a possible mechanism of cross talk between the MAPK and insulin pathways. Analysis and validation of such a hypothesis is however beyond the scope of the present study.

Discussion

We have proposed a novel structure-based computational approach to identify protein-protein interactions on a genome-wide scale. Using structural features, we have demonstrated that our method can not only identify true-interactions better than previous approaches, but also provide key biological insights that are absent from HTP experiments.
While it has been shown previously for some families that residues in and around the interface have correlated evolutionary histories, extracting such robust correlation signals for predictive purposes on a genome scale has remained difficult due to limited known interacting homologs. In the context of homology search for only monomers, enriching a multiple sequence alignment with artificial sequences has proven to be effective in the case of limited homologs [87,88]. Utilizing a statistical model for constructing evolutionarily correlated interacting homologs for a given interacting pair of proteins, we are able to simulate homologous sequences and predict PPIs from correlations at the interface of these homologs. The excellent performance of our method helps corroborate the hypothesis of residue-level correlations for a wide variety of protein-protein interactions and provides an efficient way of using these correlations for predictive purposes.

As more and more HTP data for mapping the interactome are gathered, there would be a necessary demand for automatic protocols to evaluate the data quality and estimate the confidence in individual interactions. In particular, transient interactions have been notoriously difficult to elucidate and validate. We have shown that confidence in PPIs investigated through high throughput techniques can be quantified and enhanced by our proposed complementary structure-based PPI prediction algorithm. Our PPI predictions on recent HTP human MAPK interactomes and further experimental validations have indicated the efficacy of our predicted confidence scores. Moreover, since our framework requires only the sequences of the two candidate proteins, it can be used as a complementary feature to other methods that rely on additional features [31,89].

Limited studies have been undertaken to link structural features to genome-wide interactomes to gain a mechanistic understanding of underlying biological processes. Our threading-based approach enables us to extend coverage of structure-based studies further than that possible by homology models (see section *MAPK interactome validation*). As a result, the predicted structures are more reliable and provide a sound basis for mechanistic hypotheses. We provide an anecdotal example by analyzing the distribution of annotated missense SNPs in our predicted models. In agreement with a recent study [63], we show that such mutations are enriched at the interfaces. Furthermore, detailed analysis of phosphorylation sites enables us to propose a cross-talk mechanism involving an atypical kinase, MAPK6. Predictions made by our model for the potential interactors of MAPK6 provide the basis for further exploration of the role of this relatively less-studied kinase.

Conventional homology-based methods such as interPrets [44], IBIS [73] and PRISM [45] perform well when a similar template is found in the PDB. Threading based-methods provide predictions even when such conventional methods cannot find a suitable template. Furthermore, as we show in this paper, accuracy achieved by our threading-based method is the best amongst current structure-based methods. Coev2Net acts as a complement to conventional homology methods whenever a clear template for prediction is not available and expands threading methods by incorporating coevolution of protein interfaces. However, performance of threading-based techniques has been shown to decline when the query sequences are distantly related to the template (sequence identities < 15-20%)

[49,65]. While we currently use RAPTOR for identifying the putative interface, we hope to further push this limit by integrating new threading programs like RAPTORX [90] and iWRAP [49], into Coev2Net. While we encode our interface profile as a spanning-tree based graphical model, we believe this is a simplistic approximation of the reality. More complicated graphs could potentially be required for particular families of interacting proteins. Finally, we note that transient interactions are notoriously difficult to predict using structure-based interactions. Our validation using a technique (LUMIER) that can detect even transient interactions provides some confidence in predictions of transient interactions by Coev2Net.

Methods

Coev2Net algorithm:

The Coev2Net algorithm can be roughly divided into three distinct stages, 1) identification of the putative binding interface, 2) evaluation of the compatibility of the interface with an interface co-evolution based model (see "Construction of the interface profile through simulated co-evolution" below), and 3) evaluation of the confidence score for the interaction.

Identification of the putative interface: The two query sequences are each threaded against a complex template library to search for the best template. We use a top-performing threader program "RAPTOR" [69,90] to look for the best template match. Given a set of potential template matches, the best match is selected based on the z-score of the alignment. In order to evaluate the putative interface implied by the alignment, we calculate its compatibility with respect to the co-evolutionary profile for that interface.

Evaluating the interface: The predicted interface is evaluated by computing the log-likelihood of the interface residues with respect to the interface profiles described below– a probabilistic graphical model (PGM) for interacting pairs ("positive") and another graphical model representing background correlations ("negative"). A high log-likelihood with respect to the "positive" PGM implies that the protein sequences show co-evolution at the interface, compatible with the model and are hence likely to interact.

Computing confidence score: Once we have the compatibility scores for the predicted interface, we use these as features to predict our confidence in the interaction. A logistic-regression classifier is trained on a high-confidence network, and is used to predict our confidence score for the interaction, which is the output of the classifier. Both alignment features (from stage 1: *Identification of interface*) and interface features (from stage 2: *Evaluating the interface*) are used as features in the classifier. If p is the probability of interaction (or our confidence score), then:

$$\log \frac{p}{1-p} = \alpha + \beta_1^T X_1 + \beta_2^T X_2 + \beta_i Y_i + \beta_+^T L_+ + \beta_-^T L_-$$

where, X_i are the alignment features for each protein in the interacting pair (these include sequence scores, secondary structure scores and protein lengths); Y_i is the size of the interface; L_+ is the log-likelihood score of the predicted interface with respect to the positive tree, and L_- the log-likelihood score of the predicted interface with respect to the negative tree.

Construction of the interface profile through simulated co-evolution

To construct an interface profile for a SCOPPI family, which consists of a family of protein complexes; we exploit the biological intuition that interacting proteins exhibit co-evolution at the interface. This co-evolution has been detected even in residues within 10-12 Angstroms at the interface [62,64,91-94]. In Coev2Net, the interface profile is a probabilistic graphical model (PGM), pre-computed for each SCOPPI family, and encodes the most significant pattern of interface correlations exhibited by the interacting members of the SCOPPI family. This model is computed by formulating interface co-evolution as a high-dimensional sampling problem (see Supplement for further details). The three main steps in this simulation are:

1) Seeding the co-evolution: We start the simulation from known complexes within a SCOPPI family. We first align the interfaces using a contact map alignment program, CMAPi [95], CMAPi employs a contact map representation to efficiently align multiple interfaces and thereby improves alignments as compared to other sequence and structure-based techniques. The simulation is performed on each aligned interface.

2) Simulating co-evolution for an interface: For each pair of aligned seed sequences (full proteins forming the complex), additional sequences are constructed via random mutations according to a probability distribution (Additional file 1, Fig S1) based on paired positions within interfaces of complexes. To perform a mutation at a contact, we first randomly fix one amino acid in the contact, and sample the contacting amino acid from a distribution conditioned on the fixed amino acid (see Additional file 1, Fig S2 for a schematic). The new contact thus has one amino acid as before, and the contacting amino acid mutated according to a conditional probability distribution. Each contact is treated independently, with 5% of the interface contacts mutated at each step. For non-contacting residues mutations are performed independently in the two proteins according to the BLOSUM62 matrix. Again, 5% of the non-contacting residues are mutated in one step. The percentage of mutations to carry out in one step (i.e. 5%) was chosen based on previous studies on simulated evolution for remote homolog detection [96].

   The new sequences are first aligned to the hidden Markov models (HMMs) representing the corresponding protein families, and the alignment scores computed. They are then accepted or rejected in a stochastic manner, based on their joint fitness score. The mutation and stochastic selection of interacting sequences can be viewed as a Markov Chain Monte Carlo (MCMC) algorithm [97] for a high-dimensional sampling problem – we rigorously prove this correspondence in the supplemental methods.

3) Learning the PGM: Once we have sufficient sequences (i.e. after the MCMC converges), we encode the pairwise correlations observed in these "interacting" sequences using a probabilistic graphical model (PGM). Our motivation for introducing a PGM are twofold: 1) analogous to a sequence profile, a PGM is a "profile" that can be used to score predicted interfaces, and 2) to explicitly capture long-distance correlations (non-contact-based) at or near the interface residues. We select 1000 interacting sequences per training complex as our interacting set (to avoid large sample-sample fluctuations, we select close to 2500 sequences for SCOPPI families having only one training complex). To model the correlations between residues of these interacting proteins, we use the Sanghavi-Tan-Willsky algorithm [98] to construct two trees - one for the simulated interacting proteins ("positive") and one for background correlations ("negative"). These two trees are

our interface profiles for the particular SCOPPI family and can be pre-computed before making any predictions. We restricted ourselves to spanning trees for ease of learning and inference. In fact, other inference methods, such as belief propagation, would work on a loopy graph (i.e. the loopy network of contacts at the interface) but their behavior is not easy to control and very sensitive to the initialization. Note that our profiles of the interface residues are different from the HMM ones since our interface profiles are purposely computed from only interacting sequences; the HMM is constructed from independent sequences that do not necessarily interact.

Evaluation of the classifiers:

The individual methods were evaluated based on their ability to correctly predict true-positives and true-negatives. To do this, we plot receiver operator characteristic (ROC) curves for each method. In our ROC curves, sensitivity is defined as *true-positives/(true-positives+false-negatives)* and specificity is defined as *true-negatives/(true-negatives+false-positives)*. For a high-confidence true-positive and true-negative dataset, we perform 5-fold cross-validation tests for each method (Coev2Net, Struct2Net and Baseline), and plot the average sensitivities (at particular specificities) for these 5 runs. For Coev2Net, we run the MCMC sampling 5 times, and average the performance across these 25 curves (5 MCMC x 5 CV). To compare against interPrets, we used a cutoff on the z-score computed by the algorithm to classify a prediction as positive or negative. Since there is no training required here, there was no need for a cross-validation. For the computationally intensive IBIS [73], we compared our predictions on the 10 pairs validated using the LUMIER assay.

Acknowledgements

We would like to thank George Tucker and Jason Trigg for discussion on methods. We would also like to thank Lenore Cowen and Noah Daniels for discussions on simulated evolution. Funding for the work was provided by NIH (grant number 1R01GM081871 to BB and grant number R01DK088718 to NP) and HHMI (to NP).

Competing Interests

None declared.

Author contributions

RH, JB, BB conceived and designed the study. RH designed and implemented the algorithm; JP and RH provided the proof. JP, JB and BB helped in designing the algorithm and interpretation of the results. JP, AV, JX and US provided tools, protocols and reagents. AV and US did the LUMIER experiments. AV, NP and US provided feedback to the manuscript and suggested applications for the algorithm. RH, JP, JB and BB wrote the manuscript. All authors read and approved the final manuscript.

Bibliography

[1] L Giot, J Bader, C Brouwer, A Chaudhari, B Kuang, Y L Hao, C E Ooi, B Godwin, E Vitols, G Vijayadamodar, P Pochart, H Machineni, M Welsh, Y Kong, B Zerhusen, R Malcolm, Z Varrone, A Collis, M Minto, S Burgess, L McDaniel, E Stimpson, F Spriggs, J Williams, K Neurath, N Ioime, M Agee, E Voss, K Furtak, R Renzulli *et al.*: A protein interaction Map of *Drosophila Melanogaster*. *Science* 2003, **302:**1727-1736.

[2] S Li, C Armstrong, N Bertin, H Ge, S Milstein, M Boxem, P O Vidalain, J D Han, A Chesneau, T Hao, D S Goldberg, N Li, M Martinez, J F Rual, P Lamesch, L Xu, M Tewari, S L Wong, L V Zhang, G F Berriz, L Jacotot, P Vaglio, J Reboul, T Hirozane-Kishikawa, Q Li, H W Gabel, A Elewa, B Baumgartner, D J Rose, H Yu *et al.*: A Map of the interactome Network of the metazoan *C. elegans*. *Science* 2004, **303:**540-543.

[3] J F Rual, K Venkatesan, T Hao, T Hirozane-Kishikawa, A Dricot, N Li, G F Berriz, F D Gibbons, M Dreze, N Ayivi-Guedehoussou, N Klitgord, C Simon, M Boxem, S Milstein, J Rosenberg, D S Goldberg, L V Zhang, S L Wong, G Franklin, S Li, J S Albala, J Lim, C Fraughton, E Llamosas, S Cevik, C Bex, P Lamesch, R S Sikorski, J Vandenhaute, H Y Zoghbi *et al.:* Towards a proteome-scale map of human protein-protein interaction network. *Nature* 2005, **437:**1173-1178.

[4] P Uetz, L Giot, G Cagney, T Mansfield, R Judson, R S Judson, J R Knight, D Lockshon, V Narayan, M Srinivasan, P Pochart, A Qureshi-Emili, Y Li, B Godwin, D Conover, T Kalbfleisch, G Vijayadamodar, M Yang, M Johnston, S Fields, and J M Rothberg: A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae. Nature* 2000, **403:**623-627.

[5] U Stelzl, U Worm, M Lalowski, C Haenig, F Brembeck, H Goehler, M Stroedicke, M Zenkner, A Schoenherr, S Koeppen, Timm, S Mintzlaff, C Abraham, N Bock, S Kietzmann, A Goedde, E Toksoz, A Droege, S Krobitsch, B Korn, W Birchmeier, H Lehrach, and E E Wanker: A human protein-protein interaction network: A resource for annotating the proteome. *Cell* 2005, **122:**957-968.

[6] H Yu, P Braun, M A Yildirim, I Lemmens, K Venkatesan, J Sahalie, T Hirozane-Kishikawa, F Gebreab, N Li, N Simonis, T Hao, J F Rual, A Dricot, A Vazquez, R R Murray, C

Simon, L Tardivo, S Tam, N Svrzikapa, C Fan, A S de Smet, A Motyl, M E Hudson, J Park, X Xin, M E Cusick, T Moore, C Boone, M Snyder, F P Roth *et al.*: High-quality binary protein interaction map of the yeast interactome network. *Science* 2008, **322:**104-10.

[7] N Simonis, J F Rual, A R Carvunis, M Tasan, I Lemmens, T Hirozane-Kishikawa, T Hao, J M Sahalie, K Venkatesan, F Gebreab, S Cevik, N Klitgord, C Fan, P Braun, N Li, N Ayivi-Guedehoussou, E Dann, N Bertin, D Szeto, A Dricot, M A Yildirim, C Lin, A S de Smet, H L Kao, C Simon, A Smolyar, J S Ahn, M Tewari, M Boxem, S Milstein *et al.*: Empirically controlled mapping of the *Caenorhabditis elegans* protein-protein interactome network. *Nat Methods* 2009, **6:**47-54.

[8] E Formstecher, S Aresta, V Collura, A Hamburger, A Meil, A Trehin, C Reverdy, V Betin, S Maire, C Brun, B Jacq, M Arpin, Y Bellaiche, S Bellusci, P Benaroch, M Bornens, R Chanet, P Chavrier, O Delattre, V Doye, R Fehon, G Faye, T Galli, J A Girault, B Goud, J de Gunzburg, L Johannes, M P Junier, V Mirouse, A Mukherjee *et al.*: Protein interaction mapping: a *Drosophila* case study. *Genome Res*, **15:**376-84.

[9] R M Ewing, P Chu, F Elisma, H Li, P Taylor, S Climie, L McBroom-Cerajewski, M D Robinson, L O'Connor, M Li, R Taylor, M Dharsee, Y Ho, A Heilbut, L Moore, S Zhang, O Ornatsky, Y V Bukhman, M Ethier, Y Sheng, J Vasilescu, M Abu-Farha, J P Lambert, H S Duewel, I I Stewart, B Kuehl, K Hogue, K Colwill, K Gladwish, B Muskat *et al.*: Large-scale mapping of human protein-protein interactions by mass spectrometry. *Mol Syst Biol* 2007, **3:**89.

[10] M E Sardiu and M P Washburn: Building protein-protein interaction networks with proteomics and informatics tools. *J Biol Chem*, **286:**23645-51.

[11] L Bonetta: Protein-protein interactions: Tools for the search. *Nature*, **468:**852.

[12] J G Lees, J K Heriche, I Morilla, J A Ranea, and C A Orengo: Systematic computational prediction of protein interaction networks. *Phys Biol* 2011, **8:**035008.

[13] T Kocher and G Superti-Furga: Mass spectrometry-based functional proteomics: from molecular machines to protein networks. *Nat Methods* 2007, **4**:807-15.

[14] K G Guruharsha, J F Rual, B Zhai, J Mintseris, P Vaidya, N Vaidya, C Beekman, C Wong, D Y Rhee, O Cenaj, E McKillip, S Shah, M Stapleton, K H Wan, C Yu, B Parsa, J W Carlson, X Chen, B Kapadia, K VijayRaghavan, S P Gygi, S E Celniker, R A Obar, and S Artavanis-Tsakonas: A protein complex network of *Drosophila melanogaster*. *Cell*, **147:**690-703.

[15] S R Collins, P Kemmeren, X C Zhao, J F Greenblatt, F Spencer, F C Holstege, J S Weissman, and N J Krogan: Toward a comprehensive atlas of the physical interactome of *Saccharomyces cerevisiae*. *Mol Cell Proteomics* 2007, **6:**439-50.

[16] A Elefsinioti, O S Sarac, A Hegele, C Plake, N C Hubner, I Poser, M Sarov, A Hyman, M Mann, M Schroeder, U Stelzl, and A Beyer: Large-scale de novo prediction of physical protein-protein association. *Mol Cell Proteomics* 2011, **10:**111010629.

[17] C Stark, B Breitkreutz, T Reguly, L Boucher, A Brietkreutz, and M Tyers: BIOGRID: A general repository for interaction datasets. *Nucleic Acids Research* 2006, **34:**535.

[18] D Sontag, R Singh, and B Berger: Probabilistic modeling of systematic errors in two-hybrid experiments. *Proceedings of the Pacific Symposium on Biocomputing* 2007, **12:**445-457.

[19] A Björkland, S Light, L Hedin, and A Elofsson: Quantitative assessment of the structural bias in protein-protein interaction assays. *Proteomics* 2008, **8:**4657-4667.

[20] A Vazquez, J Rual, and K Venkatesan: Quality control methodology for high-throughput protein-protein interaction screening. *Methods in Molecular Biology* 2011, **781:**279-294.

[21] C von Mering, R Krause, B Snel, M Cornell, S G Oliver, S Fields, and P Bork: Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* 2002, **417:**399-403.

[22] K Venkatesan, J.-F. Rual, A Vazquez, U Stelzl, I Lemmens, T Hirozane-Kishikawa, T Hao, M Zenkner, X Xin, K I Goh, M A Yildirim, N Simonis, K Heinzmann, F Gebreab, J M Sahalie, S Cevik, C Simon, A S de Smet, E Dann, A Smolyar, A Vinayagam, H Yu, D Szeto, H Borick, A Dricot, N Klitgord, R R Murray, C Lin, M Lalowski, J Timm *et al.*: An empirical framework for binary interactome mapping. *Nature Methods* 2008, **6:**83-90.

[23] J Yu and R L Finley: Combining multiple positive training sets to generate confidence scores for protein-protein interactions. *Bioinformatics* 2009, **25:**105-11.

[24] M Dreze, D Monachello, C Lurin, M E Cusick, D E Hill, M Vidal, and P Braun: High-quality binary interactome mapping. *Methods Enzymol* 2010, **470:**281-315.

[25] P Braun, M Tasan, M Dreze, M Barrios-Rodiles, I Lemmens, H Yu, J M Sahalie, R R Murray, L Roncari, A S de Smet, K Venkatesan, J F Rual, J Vandenhaute, M E Cusick, T Pawson, D E Hill, J Tavernier, J L Wrana, F P Roth, and M Vidal: An experimentally derived confidence score for binary protein-protein interactions. *Nature Methods* 2009, **6:**91-97.

[26] L Sambourg and N Thierry-Mieg: New insights into protein-protein interaction data lead to increased estimates of the S. cerevisiae interactome size. *BMC Bioinformatics* 2010, **11:**605.

[27] M Cusick, H Yu, A Smolyar, K Venkatesan, A Carvunis, N Simonis, J F Rual, H Borick, P Braun, M Dreze, J Vandenhaute, M Galli, J Yazaki, D E Hill, J R Ecker, F P Roth, and M Vidal: Literature-curated protein interaction datasets. *Nature Methods* 2009, **6:**39-46.

[28] S Suthram, T Shlomi, E Ruppin, R Sharan, and T Ideker: A direct comparison of protein interaction confidence assignment schemes. *BMC Bioinformatics* 2006, **7:**360.

[29] A Schwartz, J Yu, K Gardenour, R Finley Jr, and T Ideker: Cost-effective strategies for completing the interactome. *Nature Methods* 2009, **6:**55-61.

[30] H Choi, B Larsen, Z.-Y. Lin, A Breitkreutz, D Mellacheruvu, D Fermin, Z Qin, M Tyers, A.-C. Gingras, and A Nesvizhskii: SAINT: probabilistic scoring of affinity purification-mass spectrometry data. *Nature Methods* 2011, **8:**70-73.

[31] J Bader, A Chaudhuri, J Rothberg, and J Chant: Gaining confidence in high-throughput protein interaction networks. *Nature Biotech* 2003, **22:**78-85.

[32] R Singh, J Xu, and B Berger: Struct2Net: Integrating Structure Into Protein-Protein Interaction Prediction. *Proceedings of the Pacific Symposium on Biocomputing* 2006, **11:**403-414.

[33] R Jansen, H Yu, D Greenbaum, Y Kluger, N J Krogan, S Chung, A Emili, M Snyder, J F Greenblatt, and M Gerstein: A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* 2003, **302:**449-53.

[34] A Ben-Hur and W Noble: Kernel methods for predicting protein-protein interactions. *Bioinformatics* 2005, **21:**38.

[35] D Betel, K Breitkreuz, R Isserlin, D Dewar-Barch, M Tyers, and C Hogue: Structure-Templated Predictions of Novel Protein Interactions from Sequence Information. *PLOS Computational Biology* 2007, 3:e182.

[36] L Burger and E Nimwegen: Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method. *Molecular Systems Biology* 2008, **4:**165.

[37] M Deng, S Mehta, F Sun, and T Chen: Inferring domain-domain interactions from protein-protein interactions. *Genome Research* 2002, **12:**1540-1548.

[38] J A Encinar, G Fernandez-Ballester, I E Sanchez, E Hurtado-Gomez, F Stricher, P Beltrao, and L Serrano: ADAN: a database for prediction of protein-protein interaction of modular domains mediated by linear motifs. *Bioinformatics* 2009, **25:**2418-2424.

[39] J Shen, J Zhang, X Luo, W Zhu, K Yu, K Chen, Y Li, and H Jiang: Predicting protein-protein interactions based only on sequences information. *Proceedings Of The National Academy Of Sciences* 2007, **104:**4337-4341.

[40] A Valencia and F Pazos: Computational methods for the prediction of protein interactions. *Current Opinion in Structural Biology* 2002, **12:**368-373.

[41] A Valencia and F Pazos: *In silico* two-hybrid system for the selection of physically interacting protein pairs. *Proteins* 2002, **47:**219-227.

[42] S Gomez, W Noble, and A Rzhetsky: Learning to predict protein-protein interactions from protein sequences. *Bioinformatics* 2003, **19:**1875-1881.

[43] P Aloy and R Russell: Structural systems biology: modelling protein interactions. *Nature Reviews Molecular Cell Biology* 2006, **7:**188-197.

[44] P Aloy and R B Russell: Interrogating protein interactions networks through structural biology. *Proceedings of the National Academy of Sciences* 2002, **99:**5896-5901.

[45] A Aytuna, A Gursoy, and O Keskin: Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces. *Bioinformatics* 2005, **21:**2850-2855.

[46] A M Edwards, B Kus, R Jansen, D Greenbaum, J Greenblatt, and M Gerstein: Bridging structural biology and genomics: assessing protein interaction data with known complexes. *Trends Genet* 2002, **18:**529-36.

[47] N Fukuhara, N Go, and T Kawabata: Prediction of Interacting Proteins from Homology-modeled Complex Structure Using Sequence and Structure scores. *Biophysical Journal* 2007, **3:**13-26.

[48] N Fukuhara, N Go, and T Kawabata: HOMCOS: a server to predict interacting protein pairs and interacting sites by homology modeling of complex structures. *Nucleic Acids Research (Web Server Issue)* 2008, **36:**185.

[49] R Hosur, J Xu, J Bienkowska, and B Berger: iWRAP: an interface threading approach with application to cancer-related protein-protein interactions. *Journal of Molecular Biology* 2011, **405:**1295-1310.

[50] Y Huang, D Hang, L Lu, L Tong, M Gerstein, and G Montelione: Targeting the human cancer pathway protein interaction network by strutural genomics. *Molecular and Cellular Proteomics* 2008, **7:**2048-2060.

[51] P Kim, L Lu, Y Xia, and M Gerstein: Relating three-dimensional structures to protein networks provides evolutionary insights. *Science* 2006, **314:**1938-1941.

[52] W Kittichotirat, M Guerquin, R E Bumgarner, and R Samudrala: Protinfo PPC: a web server for atomic level prediction of protein complexes. *Nucleic Acids Res* 2009, **37:**519.

[53] P Kundrotas and I Vakser: Accuracy of protein-protein binding sites in high-throughput template-based modeling. *PLOS Computational Biology* 2010, **6:**1000727.

[54] H Lu, L Lu, and J Skolnick: Development of Unified Statistical Potentials Describing Protein-Protein Interactions. *Biophysical Journal* 2003, **84:**1895-1901.

[55] L Lu, H Lu, and J Skolnick: MULTIPROSPECTOR: An algorithm for the prediction of protein-protein interactions by multimeric threading. *Proteins* 2002, **49:**350-364.

[56] S Mukherjee and Y Zhang: Protein-protein complex structure predictions by multimeric threading and template recombination. *Structure* 2011, **19:**955-66.

[57] A Stein, R Russell, and P Aloy: 3did: interacting protein domains of known three-dimensional structure. *Nucleic Acids Research* 2005, **33:**D413-D417.

[58] A. Stein, R. Mosca, and P. Aloy: Three-dimensional modeling of protein interactions and complexes is going 'omics. *Curr Opin Struct Biol* 2011, **21:**200-8.

[59] N Tuncbag, A Gursoy, and O Keskin: Prediction of protein-protein interactions: unifying evolution and structure at protein interfaces. *Phys Biol* 2011, **8:**035006.

[60] N Tuncbag, A Gursoy, R Nussinov, and O Keskin: Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using PRISM. *Nat Protoc* 2011, **6:**1341-54.

[61] M Tyagi, K Hashimoto, B A Shoemaker, S Wuchty, and A R Panchenko: Large-scale mapping of human protein interactome using structural complexes. *EMBO Rep* 2012, **13:**266-271.

[62] M Tyagi, R R Thangudu, D Zhang, S H Bryant, T Madej, and A R Panchenko: Homology Inference of Protein-Protein Interactions via Conserved Binding Sites. *PLoS One* 2012, **7:**28896.

[63] X Wang, X Wei, B Thijssen, J Das, S Lipkin, and H Yu: Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nature Biotechnology* 2012, **30:**159-164.

[64] M. N. Wass, A. David, and M. J. Sternberg: Challenges for the prediction of macromolecular interactions. *Curr Opin Struct Biol* 2011, **21:**382-90.

[65] L Pulim, J Bienkowska, and B Berger: LTHREADER: Prediction of extracellular Ligand-Receptor interactions in cytokines using localized threading. *Protein Science* 2008, **17:**279-292.

[66] R Singh, D Park, J Xu, R Hosur, and B Berger: Struct2Net: a web service to predict protein-protein interactions using a structure-based approach. *Nucleic Acids Research (Web Server Issue)* 2010, **38:**W508-W515.

[67] S Bandyopadhyay, C Chiang, J Srivastava, M Gersten, S White, R Bell, C Kurschner, C Martin, M Smoot, S Sahasrabudhe, D Barber, S Chanda, and T Ideker: A human MAP kinase interactome. *Nature Methods* 2010, **7:**801-805.

[68] A Vinayagam, U Stelzl, R Foulle, S Plassmann, M Zenkner, J Timm, H Assmus, M Andrade-Navarro, and E Wanker: A directed protein interaction network for investigating intracellular signal transduction. *Science Signaling* 2011, **4:**rs8.

[69] J Xu, M Li, D Kim, and Y Xu: RAPTOR: Optimal Protein Threading by Linear Programming. *J Bioinform Comput Biol* 2003, **1:**95-117.

[70] C Winter, A Henschel, W K Kim, and M Schroeder: SCOPPI: a structural classification of protein-protein interfaces. *Nucleic Acids Research (Database issue)* 2006, **34:**310-314.

[71] M Barrios-Rodiles, K Brown, B Ozdamar, R Bose, Z Liu, R Donovan, F Shinjo, Y Liu, J Dembowy, I Taylor, V Luga, N Przulj, M Robinson, H Suzuki, Y Hayashizaki, I Jurisica, and J Wrana: High-Throughput mapping of a dyamic signaling network in mammalian cells. *Science* 2005, **307:**1621-1625.

[72] A Hegele, A Kamburov, A Grossmann, C Sourlis, S Wowro, M Weimann, C Will, V Pena, R LŸhrmann, and U Stelzl: Dynamic protein-protein interaction wiring of the human spliceosome. *Molecular Cell* 2012, **45:**567-580.

[73] A Shoemaker, D Zhang, M Tyagi, R Thangudu, J Fong, A Marchler-Bauer, S Bryant, T Madej, and A Panchenko: IBIS (Inferred Biomolecular Interaction Server) reports, predicts and integrates multiple types of conserved interactions for proteins. *Nucleic Acids Research (Database Issue)* 2012, **40:**D834-D840.

[74] S Sherry, M Ward, M Kholodov, J Baker, L Phan, E Smigielski, and K Sirotkin: dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research* 2001, **29:**308-311.

[75] V Ramensky, P Bork, and S Sunyaev: Human non-synonymous SNPs: server and survey. *Nucleic Acids Research* 2002, **30:**3894-3900.

[76] I Adzhubei, S Schmidt, L Peshkin, V Ramensky, A Gerasimova, P Bork, A Kondrashov, and S Sunyaev: A method and server for predicting damaging missense mutations. *Nature Methods* 2010, **7:**248-249.

[77] S Forbes, N Bindal, S Bamford, C Cole, C Kok, D Beare, M Jia, R Shepherd, K Leung, A Menzies, J Teague, P Campbell, M Stratton, and P Futreal: COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Research (Database Issue)* 2011, **39:**945-950.

[78] C Richardson, Q Gao, C Mitsopoulous, M Zvelebil, L Pearl, and F Pearl: MoKCa database – mutations of kinases in cancer. *Nucleic Acids Research (Database Issue)* 2009, **37:**824-831.

[79] H Nishi, K Hashimoto, and A Panchenko: Phosphorylation in protein-protein binding: effect on stability and function. *Structure* 2011, **19:**1807-1815.

[80] D Morrison: The 14-3-3 proteins: integrators of diverse signaling cues that impact cell fate and cancer development. *Trends in Cell Biology 2008*, **19:**16-23.

[81] J Woodcock, J Murphy, F Stomski, M Berndt, and A Lopez: The dimeric versus monomeric status of 14-3-3zeta is controlled by phosphoryltion of Ser58 at the dimeric interface. *J Biol Chem* 2003, **278:**36323-36327.

[82] K Yoshida, T Yamaguchi, T Natsume, D Kufe, and Y Miki: JNK phosporylation of 14-3-3 proteins regulates nuclear targeting of c-Abl in the apoptotic response to DNA damage. *Nature Cell Biology* 2005, **7:**278-285.

[83] C Julien, P Coulombe, and S Meloche: Nuclear export of ERK3 by a CRM1-dependent mechanism regulates its inhibitory action on cell-cycle progression. *J Biol Chem* 2003, **278:**42615-42624.

[84] F Opperman, F Gnad, J Olsen, R Hornberger, Z Greff, G Keri, M Mann, and H Daub: Large-scale proteomics analysis of the human kinome. *Mol Cell Proteomics* 2009, **8:**1751-1764.

[85] P Deleris, M Trost, I Topsirovic, P Tanguay, K Borden, P Thibault, and S Meloche: Activation loop phosphorylation of ERK3/ERK4 by group I p21-activated kinases (PAKs) defines a novel PAK-ERK3/4-MAPK-activated protein kinase 5 signaling pathway. *J Biol Chem* 2011, **286:**6470-6478.

[86] N Dephoure, C Zhou, J Villen, S Beausoleil, C Bakalarski, S Elledge, and S Gygi: A quantitative atlas of mitotic phosphorylation. *Proc Natl Acad Sci USA* 2008, **105:**10762-10767.

[87] A Kumar and L Cowen: Augmented training of Hidden Markov Models to recognize remote homologs via simulated evolution. *Bioinformatics* 2009, **25:**1602-1608.

[88] A Kumar and L Cowen: Recognition of beta-structural motifs using Hidden Markov Models trained with simulated evolution. *Bioinformatics* 2010, **26:**287.

[89] H Huang and J Bader: Precision and recall estimates for two-hybrid screens. *Bioinformatics* 2009, **25:**372-378.

[90] J Peng and J Xu: RaptorX: Exploiting structure information for protein alignment by statistical inference. *Proteins* 2011, **79:**161-71.

[91] M Kann, B Shoemaker, A Panchenko, and T Przytycka: Correlated evolution of interacting proteins: Looking behind the mirror tree. *J Mol Biol* 2009, **385:**91-98.

[92] A K Ramani and E M Marcotte: Exploiting the co-evolution of interacting proteins to discover interaction specificity. *J Mol Biol* 2003, **327:**273-84.

[93] F Pazos, D Juan, J M Izarzugaza, E Leon, and A Valencia: Prediction of protein interaction based on similarity of phylogenetic trees. *Methods Mol Biol* 2008, **484:**523-35.

[94] A Panjkovich and P Aloy: Predicting protein-protein interaction specificity through the integration of three-dimensional structural information and the evolutionary record of protein domains. *Mol Biosyst* 2010, **6:**741-9.

[95] V Pulim, J Bienkowska, and B Berger: Optimal contact map alignment of protein-protein interfaces. *Bioinformatics* 2008, **24:**2324-2328.

[96] N Daniels, R Hosur, B Berger, and L Cowen: SMURFLite: combining simplified markov random fields with simulated evolution improves remote homology detection for beta-structural proteins into the twilight zone. *Bioinformatics* 2012, **28:**1216-1222.

[97] Jun S Liu, Monte Carlo strategies in scientific computing. New York: Springer, 2001.

[98] S Sanghvi, V Tan, and A Willsky: Learning graphical models for hypothesis testing. Statistical Signal Processing Workshop (SSP), 2007.

Figure legends

Fig 1. Framework for assessing confidence in a HTP PPI screen. Coev2Net, trained on a high-quality PPI network, is able to assign structure-based confidence scores for HTP PPI networks. Each node represents a protein and each edge the putative interaction between the two proteins. The thickness of an edge describes structure-based confidences of putative PPIs.

Fig 2. Flowchart of Coev2Net. Left: MCMC sampling to generate synthetic homologous sequences for each complex template. Right: 1) For given query protein pairs, the best template (from the structural library) is identified by protein threading; 2) structural and sequence features are extracted from the interfacial alignment and residue correlations scored w.r.t. the profile PGM; and 3) a classifier gives the probability of interaction for the query protein pair.

Fig 3. A) Overlap of the Vinayagam (blue) and Bandyopadhyay (red) datasets (left)**.** The study by Bandyopadhyay et al. reveals 2269 interactions with 641 "core" interactions supported by multiple lines of evidence, whereas the Vinayagam dataset has 2626 interactions connecting 1126 proteins. Differences in the two experimental techniques are highlighted by the fact that only 170 nodes and 6 interactions overlap in the two sets. **B)** Coev2Net predicted high-confidence network is shown on the right. Edge colors correspond to the dataset they come from. MAPK6 has the highest degree, and its label is shown explicitly. **C)** Comparisons of performance on MAPK network for Coev2Net and Struct2Net (iWRAP+DBLRAP) [49,32,66] in terms of sensitivity and specificity. Coev2Net performs much better than previous methods on this dataset (core network of Bandyopadhyay et al.), and its performance is robust with respect to the randomness in MCMC sampling. The classifier (Fig 2) is trained and tested via 5-fold cross-validation on the core network. The MCMC procedure is repeated 5 times to assess robustness of the predictions and the corresponding error bars are indicated. 'Baseline' method represents a logistic regression classifier with just the alignment features and no PPI (inter-protein) features. **D)** Experimental validation of predicted high-confidence interactions using LUMIER assay. Typically a fold increase of 1.5 is considered as a true positive.

Fig 4. Predicted interfaces are enriched for SNPs in the Coev2Net predicted high-confidence MAPK network. **A**) Relative distribution of PolyPhen annotated mutations at the interface and non-interface. **B)** SNP (PolyPhen annotated) prevalence at the interface and non-interface. **C)** Somatic mutations characterized as "missense" preferentially fall on the interface (bottom). The white circles represent corresponding means. Error bars represent the 75%-25% data range.

Fig 5. A) Predicted interface for the interaction between BRAF (light blue) and PAK2 (red surface). Cancer associated mutations that are annotated are shown in magenta. In dark blue we indicate mutations that are predicted to be associated with cancer but with no current annotations. Rest of the template structure is shown in gray. Mutations were taken from MoKCa database [78]. **B)** Predicted interface for the interaction between MAPK6 (yellow) and YWHAZ (cyan). Phosphorylation sites on the proteins are indicated in red (S189 for MAPK6 and S184 for YWHAZ). The template used for the prediction was 1F5Q (chains A and B).

Yeast strains implementation	#validated (LUMIER)	Y2H PPIs	%overlap
Y strain 2m 1 reporter 1mM_3-AT (Braun et al.)	19	33	57
Y strain 2m 2 reporters 1mM_3-AT (Braun et al.)	13	22	59
Y strain CEN 1 reporter 1mM_3-AT (Braun et al.)	17	23	74
MaV CEN 2 reporters 20 mM_3-AT (Braun et al.)	9	14	64
Our prediction	**14**	**19**	**74**
Our prediction*	**10**	**11**	**91**

Table 1. Comparison of overlaps achieved by Braun et al. and our method when some of the initial Y2H interaction pairs are re-tested using LUMIER assay. * These pairs have LUMIER assay z-scores > 1.5.

Abbreviations

coAP, co-affinity purification; coIP, co-immunoprecipitation; FDR, false-discovery rate; FNR, false-negative rate; FPR, false-positive rate; GO, gene ontology; HMM, Hidden Markov Model; HTP, high-throughput; LUMIER, luminescence-based mammalian interactome mapping; MAPK, mitogen-activated protein kinase; MCMC, Markov chain Monte Carlo; PGM, probabilistic graphical model; PPI, protein-protein interaction; ROC, receiver operator characteristics; SNP, single nucleotide polymorphism; TN, true negative; TP, true positive; Y2H, yeast-2-hybrid.

Additional files

Additional file 1

Title: Supplementary methods

Description: Supplementary methods on the algorithm, results on benchmarking and comparison with other methods.

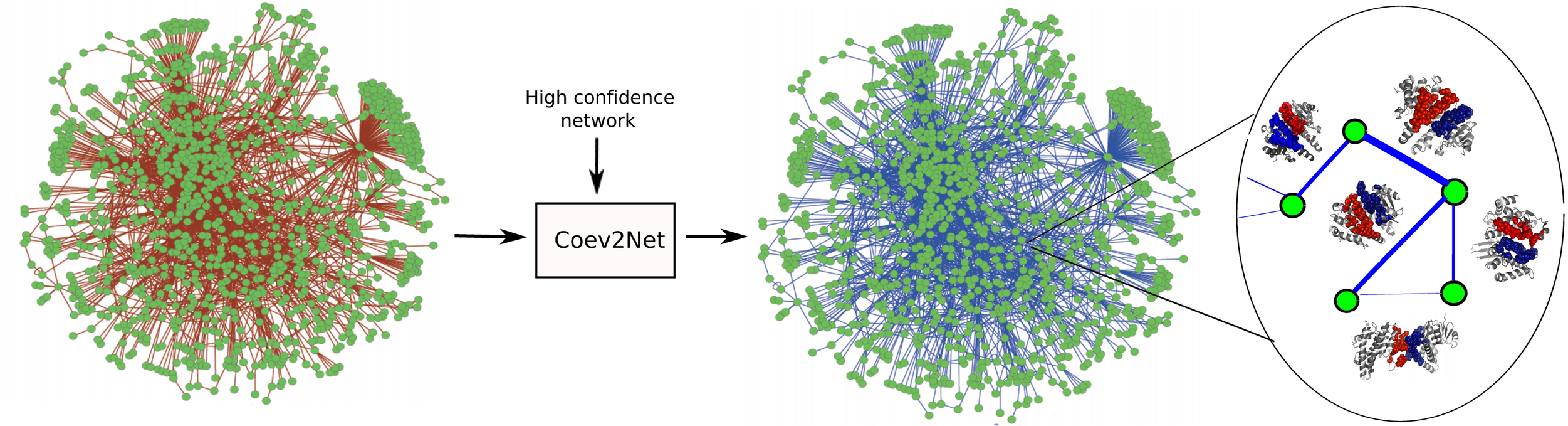


Figure 1

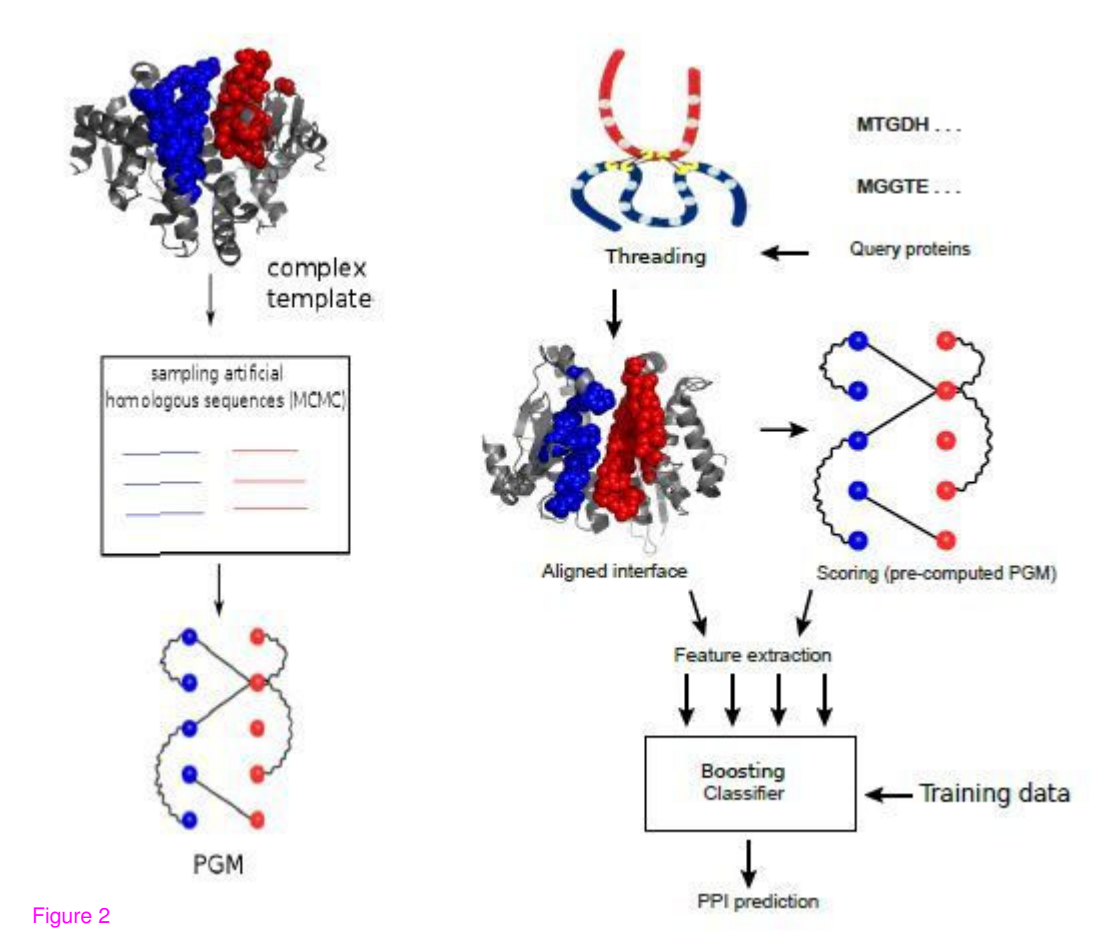

complex
template
sampling artificial
homologous sequences (MCMC)
PGM
MTGDH . . .
MGGTE . . .
Threading
Query proteins
Aligned interface
Scoring (pre-computed PGM)
Feature extraction
Boosting
Classifier
Training data
PPI prediction

Figure 2

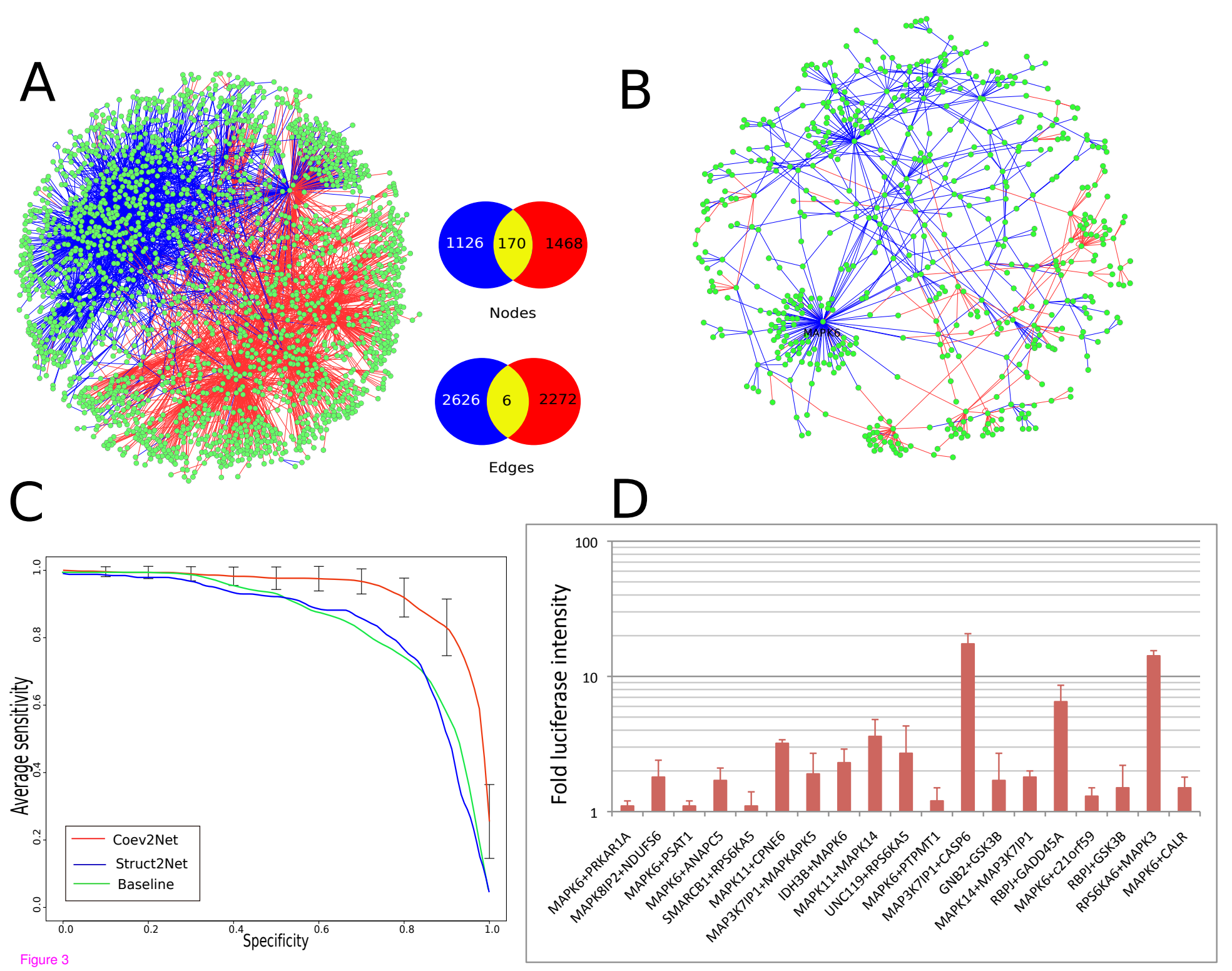
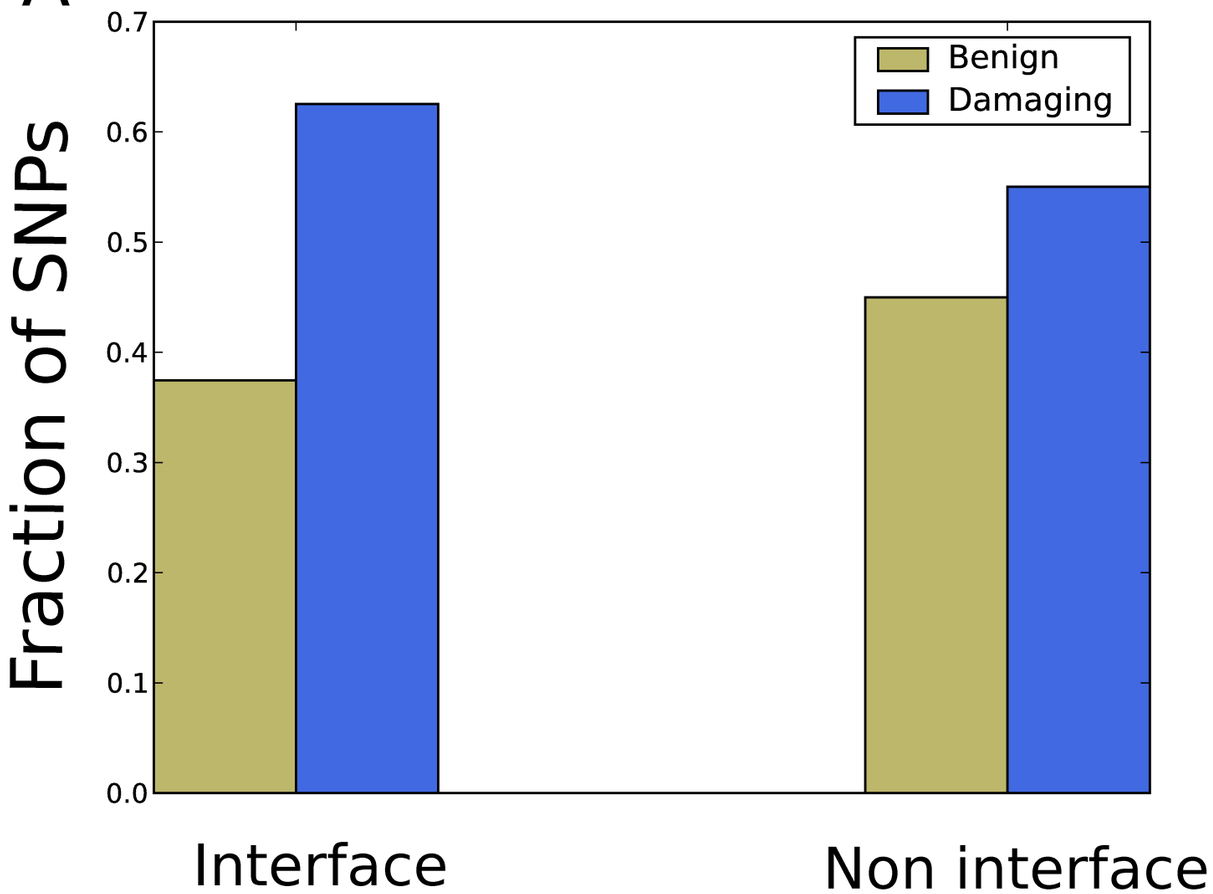
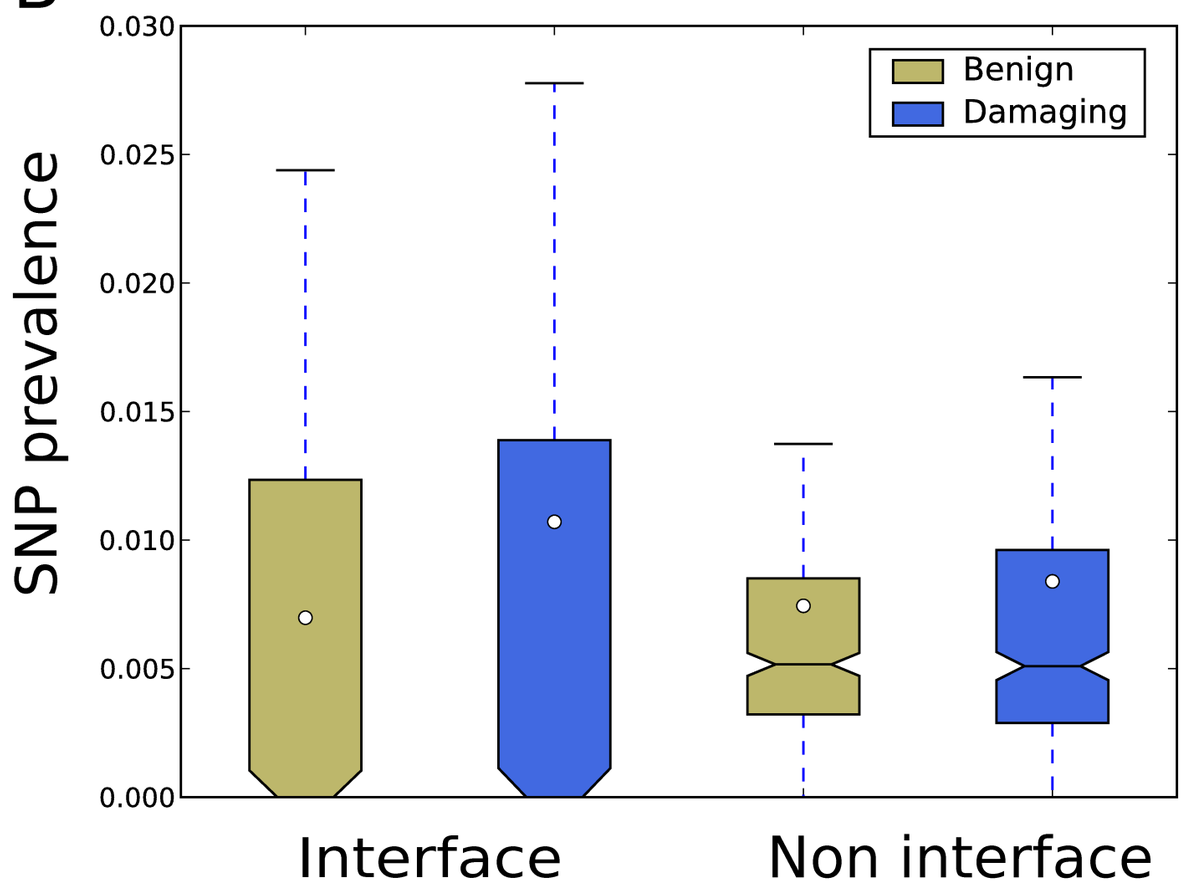

A
1126
170
1468
Nodes
2626
6
2272
Edges
B
MAPK6
C
Average sensitivity
Specificity
Coev2Net
Struct2Net
Baseline
D
Fold luciferase intensity
100
10
1
MAPK6+PRKAR1A
MAPK8IP2+NDUFS6
MAPK6+PSAT1
MAPK6+ANAPC5
SMARCB1+RPS6KA5
MAPK11+CPNE6
MAP3K7IP1+MAPKAPK5
IDH3B+MAPK6
MAPK11+MAPK14
UNC119+RPS6KA5
MAPK6+PTPMT1
MAP3K7IP1+CASP6
GNB2+GSK3B
MAPK14+MAP3K7IP1
RBPJ+GADD45A
MAPK6+c21orf59
RBPJ+GSK3B
RPS6KA6+MAPK3
MAPK6+CALR

Figure 3

A


Fraction of SNPs
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Benign
Damaging
Interface
Non interface

B


SNP prevalence
0.000
0.005
0.010
0.015
0.020
0.025
0.030
Benign
Damaging
Interface
Non interface

C

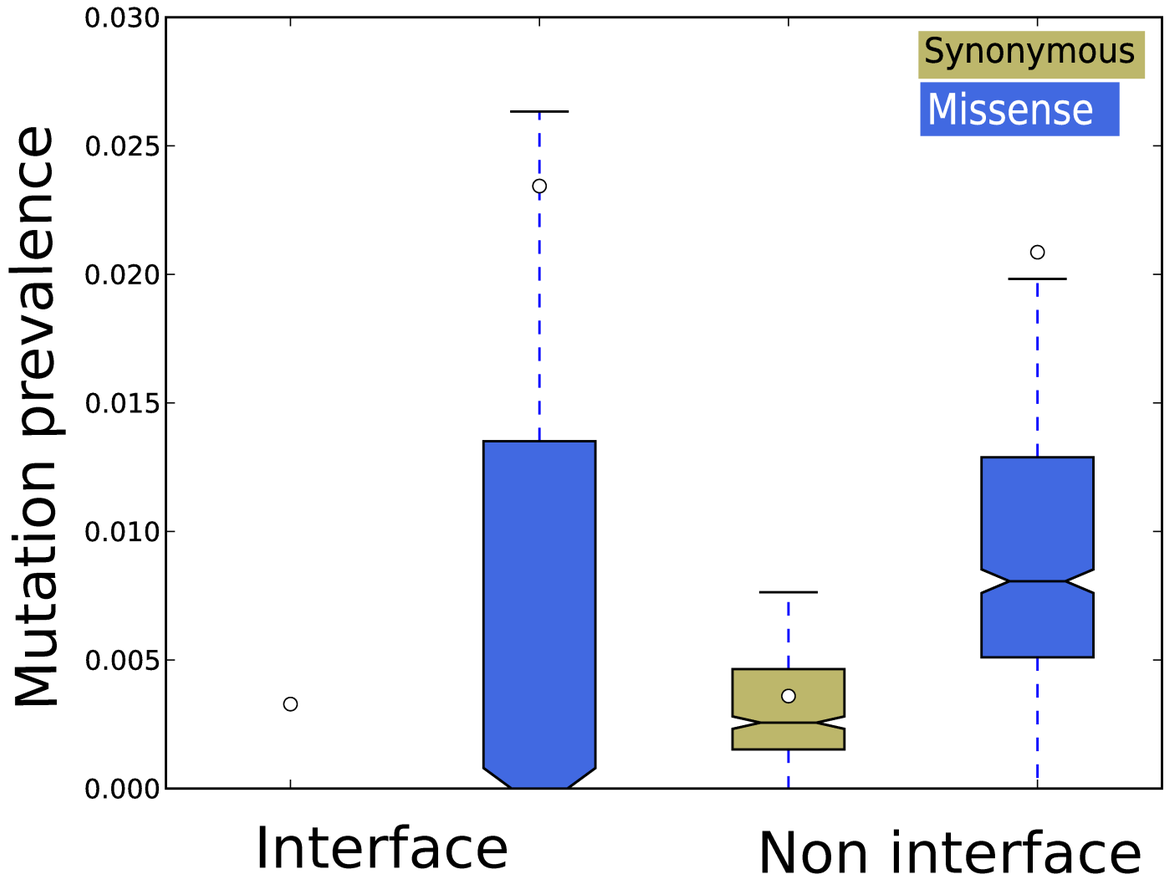
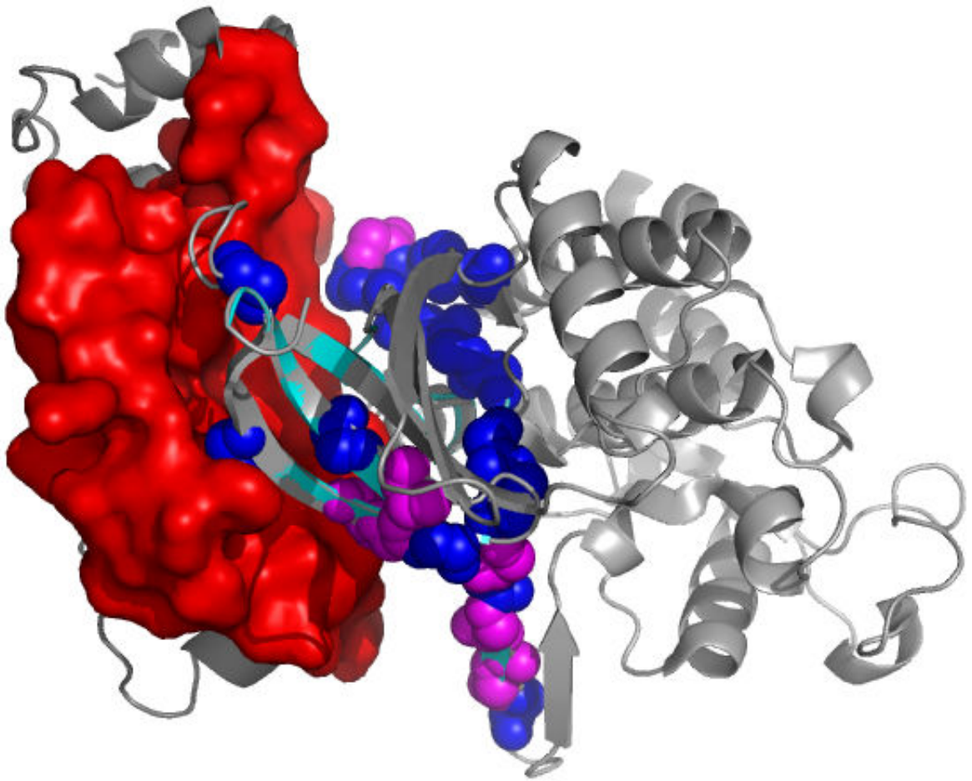

Mutation prevalence
0.000
0.005
0.010
0.015
0.020
0.025
0.030
Synonymous
Missense
Interface
Non interface

Figure 4

A



B

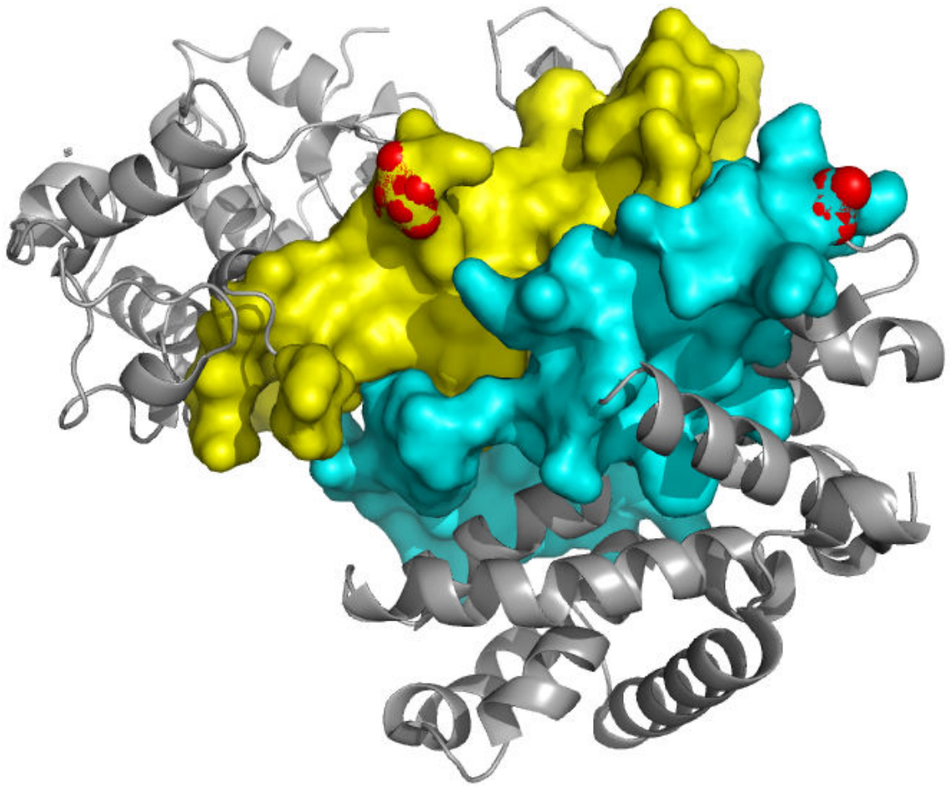


Figure 5

Additional files provided with this submission:

Additional file 1: Framework_Supplement.pdf, 692K
http://genomebiology.com/imedia/9729628617730508/supp1.pdf