Large-scale mapping of human protein interactions using structural complexes

– Suppl. Online Materials

Manoj Tyagi, Kosuke Hashimoto, Benjamin A. Shoemaker, Stefan Wuchty and Anna R. Panchenko

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

Supplementary Background:

Mapping of complete interactomes using structural complexes.

A few attempts have been made to map the interactome of yeast on a large scale using structural complexes. For example, Aloy and Russell introduced a scoring scheme to assess the fit of any interacting pair on a structural complex of homologous proteins (Aloy & Russell, 2002). They searched for Pfam domains in contact in the same structure, and inferred potential interactions from homologous structural complexes. The authors found 59 out of 2590 high-throughput interactions using the yeast interactome that could be mapped to structurally inferred complexes. The advantage of this method lies in the evaluation of the statistical significance and reliability of predicted interactions using empirical potentials. Later Kim et al. addressed the topic of finding mutually exclusive and interfaces by mapping the yeast interactome using sequence similarity between proteins from high-throughput interactions and known protein complexes (Kim et al, 2006). As a result, they composed a network containing 873 nodes and 1269 edges. Most recently a PRISM protocol was introduced for the prediction of protein-protein interactions on the proteome scale using known template protein-protein interfaces (Ogmen et al, 2005; Tuncbag et al, 2011). The idea of this approach relies on the similarity of interfaces (not necessarily folds) of interacting proteins. The authors looked for similarity between surface regions of target proteins to known template interfaces and considered both geometric complementarity and evolutionary conservation of binding hot spots.

The framework described in this study uses homology inference similar to the first two methods. The strength of our approach is that it ensures close evolutionary relationships between structural complexes and target proteins, and verifies the interactions and binding interfaces by several means. First, it examines the evolutionary conservation between homologous complexes under the assumption that if the binding site is conserved among non-redundant homologs, it is more likely to be biologically relevant (sites that are not conserved and lineage specific might be excluded unless they are very similar to the query based on the ranking score, see below). Second, our approach uses algorithms to infer correct biological units and, finally, it applies a

rigorous scheme to rank binding sites with respect to their relevance to the target protein. Such a ranking scheme includes sequence-PSSM score (where the Position Specific Scoring Matrix is constructed based on the alignment of binding sites from homologous complexes), overall sequence similarity between target protein and its homologs with known complexes, and the number of interface contacts. It should be mentioned that such a rigorous verification of interactions upon homology inference is essential since common descent does not necessarily imply similarity in function or interactions. Annotations transferred from one protein to a homolog may result in incorrect functional or interolog assignment at larger evolutionary distances, even for close homologs if they have different binding specificities. Since binding specificity is usually determined by the structural and sequence features of protein interaction interfaces, it is essential to detect and transfer binding sites correctly. To verify and guide predictions based on inference, one needs to ensure similarity between sites on the unknown target protein and on the conserved binding sites detected in homologs. Moreover, the PDB asymmetric unit which is usually used to infer interactions does not necessarily correspond to the biological state of a given protein. According to several studies (Jefferson et al, 2006; Xu et al, 2008), more than 20% of PDB complexes represent crystallographic packing errors and about 30% of PDB entries should be reconstructed by applying crystallographic symmetry operations.

Supplementary Methods:

Major steps of the IBIS method for predicting protein interaction partners and binding site locations

- ***Collecting homologs with observed interactions***

To infer interactions based on homology we first collected template proteins with known structures that are similar to a given query protein and have at least 80% sequence identity and more than 80% of the query sequence aligned using cBlast. For each template protein we retrieved all homologous (with more than 30% identity) and structurally-similar proteins with the known structural complexes from the Protein Data Bank. Template and homologous structural complexes were structurally aligned using the

VAST algorithm. Subsequently, homologous complexes were grouped based on their binding site similarity, assuming that a binding site is functionally important and is not lineage specific if it is evolutionarily conserved among non-redundant homologs.

- ***Measuring binding site similarity***

We cluster domain-binding sites into groups based on their sequence and structure similarity. The similarity score between two positions *i* and *j* of two binding sites is defined as (Thangudu et al, 2010):

$$S_{ij} = H(a_i, a_j)\Delta_{ij} + \boldsymbol{\theta}\Delta_{ij} + w(1 - \Delta_{ij})$$

where *H* is the corresponding element of the BLOSUM62 matrix; Δ_{ij} is equal to 1 if two positions are aligned and 0 otherwise; $\boldsymbol{\theta}$ is an additional weight of "+1" added to each structurally equivalent position, and *w* is a gap penalty of "-4". The overall similarity score between two binding sites is calculated by summing S_{ij} over all positions in the gapped alignment.

- ***Clustering of binding sites***

The binding sites of the homologous structure neighbors are clustered using a complete-linkage clustering algorithm, which calculates the distance between two clusters as the maximum distance between their members. A distance cutoff value to define the clusters is chosen using a pseudo-free energy function from a study which maximizes the mean similarity of members within a cluster and minimizes the complexity of the description provided by cluster membership (Slonim et al, 2005). At the end of this procedure sets of binding residues ("binding sites") from different homologs of the query protein are grouped together based on their similarity.

- ***Ranking of binding site clusters***

All binding site clusters are ranked in terms of biological relevance and similarity to the query. First, we check whether the same or similar binding sites reoccur in diverse protein complexes and assess their conservation within the cluster. Clusters that have

more than one non-redundant protein (at a sequence identity threshold of 90%) in the cluster are called “conserved binding site” clusters. Those clusters with only one non-redundant protein complex are considered “singletons” and usually correspond to either lineage specific binding modes or cases lacking enough conservation evidence.

Second, since the larger interfaces are more likely to be biologically relevant, the ranking score also includes a term corresponding to the number of interfacial contacts averaged over all homologous complexes ($Z_{contact}$). Another term in the ranking score accounts for the relevance of a given binding site cluster to the query. A position specific score matrix (PSSM) is constructed based on the binding site alignment using the implicit pseudo-count method. The aligned binding site region of the query protein is scored against the PSSM, and a sequence-PSSM score is calculated (Z_{PSSM}). A higher sequence-PSSM score implies a higher probability of this site being biologically relevant for annotating the given query. In addition, we calculate the average sequence identity between the query and all cluster members over the whole structure-structure alignment (not just binding sites) to estimate the evolutionary distance between the query protein and the group of homologous structure neighbors (Z_{pcnt}).
All components of the ranking score (*i.e.* PSSM, conservation, contact number, and percent identity of the alignment) are converted to Z-scores, and their weighted combination is used where weights were determined empirically.

- ***Validation of interactions using the PISA algorithm***

Interfaces present in PDB asymmetric units (ASU) are validated using the PISA (Protein Interfaces, Surfaces, and Assemblies server) algorithm (Krissinel & Henrick, 2007) which is considered to be one of the best methods for identifying biologically relevant interfaces present in crystal structures. PISA is an automated method for detecting macromolecular assemblies based on the analysis of interfaces and stability of assemblies reported in crystal structures. PISA uses chemical thermodynamics calculations to compute a set of macromolecular assemblies, which are expected to be stable in solution and presumed to represent the biological form of a protein in the cell.

- ***Assessing the accuracy of the method***

The accuracy of the IBIS method for predicting protein interaction partners and binding site locations, was found to have 88% sensitivity and a recall value of 71% (Tyagi et al, 2011). We address the question of predicting the biological interfaces that are not present in the PDB asymmetric unit and need to be reconstructed by applying crystallographic symmetry operations. We show that almost half of such interfaces can be reconstructed by IBIS without the prior knowledge of crystal parameters of the query protein.

References

Aloy P, Russell RB (2002) Interrogating protein interaction networks through structural biology. *Proc Natl Acad Sci U S A* **99:** 5896-5901

Jefferson ER, Walsh TP, Barton GJ (2006) Biological units and their effect upon the properties and prediction of protein-protein interactions. *J Mol Biol* **364:** 1118-1129

Kim PM, Lu LJ, Xia Y, Gerstein MB (2006) Relating three-dimensional structures to protein networks provides evolutionary insights. *Science* **314:** 1938-1941

Krissinel E, Henrick K (2007) Inference of macromolecular assemblies from crystalline state. *J Mol Biol* **372:** 774-797

Ogmen U, Keskin O, Aytuna AS, Nussinov R, Gursoy A (2005) PRISM: protein interactions by structural matching. *Nucleic Acids Res* **33:** W331-336

Slonim N, Atwal GS, Tkacik G, Bialek W (2005) Information-based clustering. *Proc Natl Acad Sci U S A* **102:** 18297-18302

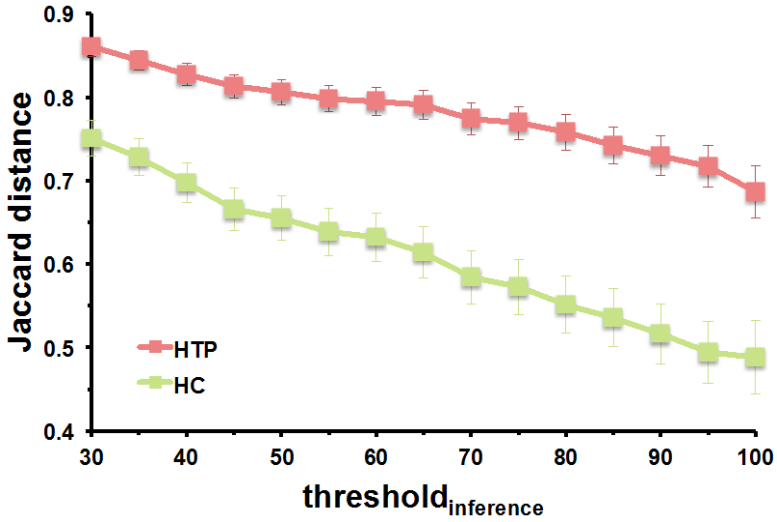
Thangudu RR, Tyagi M, Shoemaker BA, Bryant SH, Panchenko AR, Madej T (2010) Knowledge-based annotation of small molecule binding sites in proteins. *BMC Bioinformatics* **11:** 365

Tuncbag N, Gursoy A, Nussinov R, Keskin O (2011) Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using PRISM. *Nat Protoc* **6:** 1341-1354

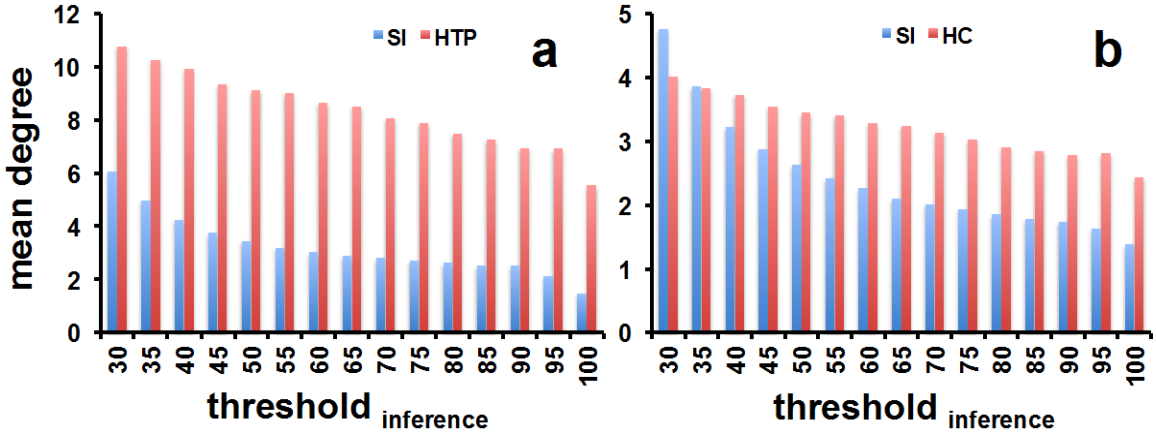
Tyagi M, Thangudu RR, Zhang D, Bryant SH, Madej T, Panchenko AR (2011) Homology inference of protein-protein interactions via conserved binding sites *PloS One*

Xu Q, Canutescu AA, Wang G, Shapovalov M, Obradovic Z, Dunbrack RL, Jr. (2008) Statistical analysis of interface similarity in crystals of homologous proteins. *J Mol Biol* **381:** 487-507

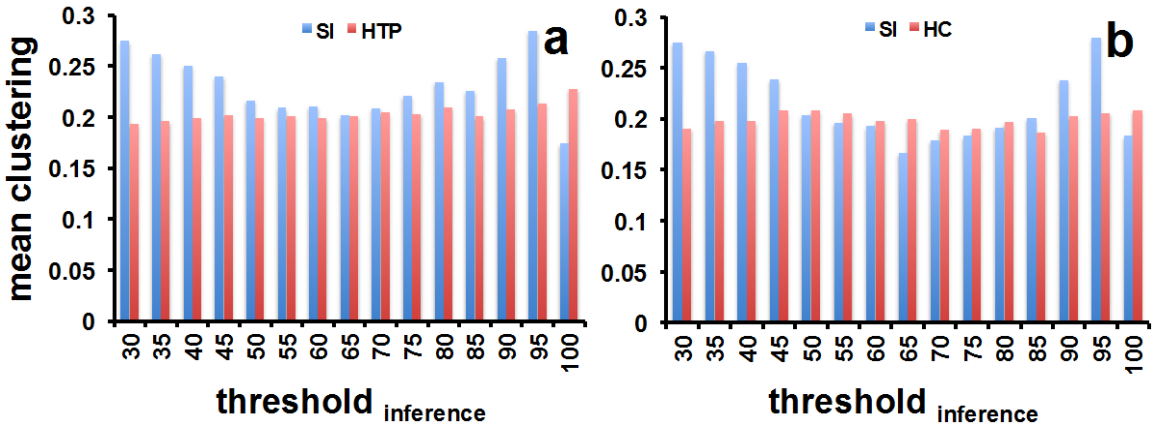
Supplementary Figures:



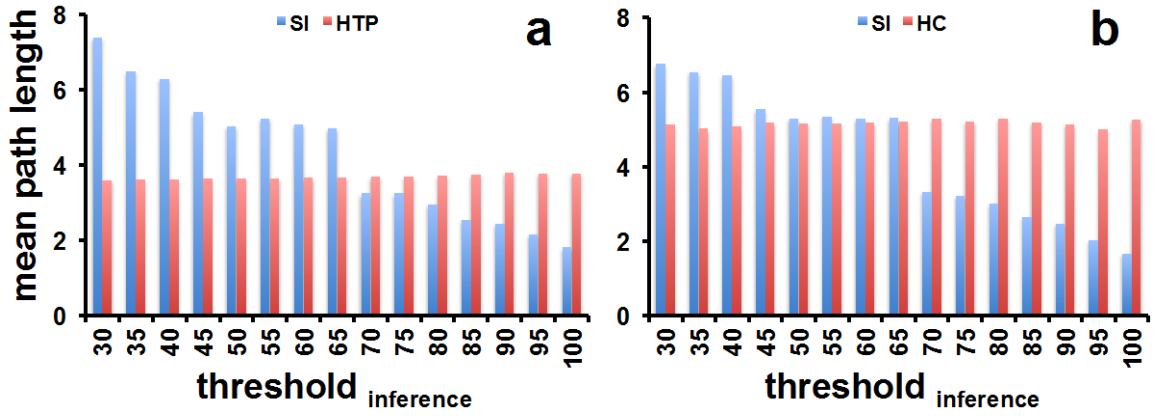
Suppl. Figure 1| Comparisons of structurally inferred (SI) to high-throughput (HTP) and high-confidence (HC) interactions of single proteins. We calculated Jaccard distances between interactions of a given protein that appeared in the HTP, HC and SI networks at different inference thresholds. We only considered proteins that appeared in both networks. Averaging Jaccard indexes over all proteins, error bars indicate 95% confidence intervals.



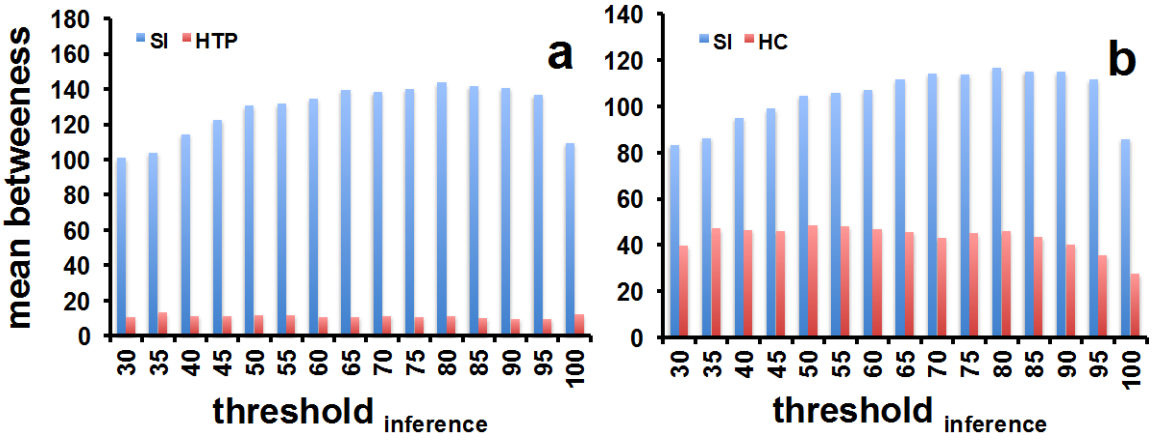
Suppl. Figure 2| Mean node degrees are plotted for **(a)** structurally inferred (SI) and high-throughput (HTP) networks and **(b)** structurally inferred (SI) and high confidence (HC) networks using different inference thresholds. In both cases, we only considered interactions between proteins that appear in both compared networks. Furthermore, structurally inferred interactions were either experimentally 'observed' or derived from conserved binding site clusters.



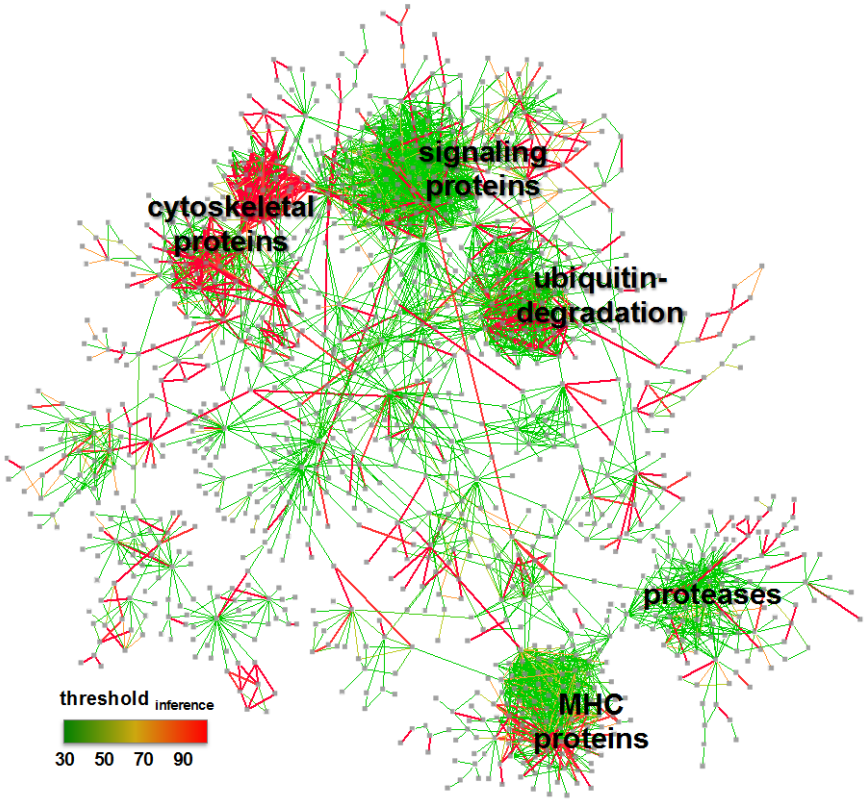
Suppl. Figure 3| Mean clustering coefficients are plotted for **(a)** structurally inferred (SI) and high-throughput (HTP) networks and **(b)** structurally inferred (SI) and high confidence (HC) networks using different inference thresholds. In both cases, we only considered interactions between proteins that appeared in both compared networks. Furthermore, structurally inferred interactions were either experimentally 'observed' or derived from conserved binding site clusters.



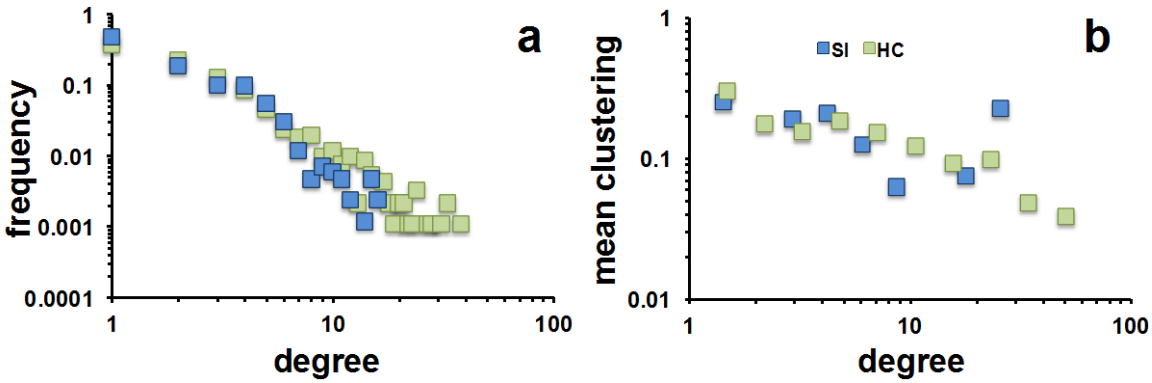
Suppl. Figure 4| Mean shortest path lengths are plotted for **(a)** structurally inferred (SI) and high-throughput (HTP) networks and **(b)** structurally inferred (SI) and high confidence (HC) networks using different inference thresholds. In both cases, we only considered interactions between proteins that appear in both compared networks. Furthermore, structurally inferred interactions were either experimentally 'observed' or derived from conserved binding site clusters.



Suppl. Figure 5| Mean betweenness centralities are plotted for **(a)** structurally inferred (SI) and high-throughput (HTP) networks and **(b)** structurally inferred (SI) and high confidence (HC) networks using different inference thresholds. In both cases, we only considered interactions between proteins that appear in both compared networks. Furthermore, structurally inferred interactions were either experimentally 'observed' or derived from conserved binding-site clusters.



Suppl. Figure 6| Utilizing all structurally inferred interactions, we found several modules that were functionally coherent.



Suppl. Figure 7| After we structurally inferred protein interactions with inference threshold of more than 50% (SI) we only considered experimentally determined high confidence (HC) and structural interactions between proteins that appeared in both networks. Furthermore, structurally inferred interactions were either experimentally 'observed' or derived from conserved binding site clusters. In **(a)** we found that distributions of interacting proteins in all networks have a power-law tail. **(b)** Indicating modularity in the given networks, we observed a power-law dependence between the number of interaction partners and the clustering coefficient in all networks.