

# Network Motifs Profile of Proteins Contact Map: a novel Pattern Recognition Tool for Fold Prediction

Ofer Rahat<sup>1\*</sup>, Merav Parter<sup>2</sup>, Uri Alon<sup>2</sup>, Gideon Schreiber<sup>1</sup>

1. Biochemistry Department, Weizmann Inst. of Science, Rehovot, Israel 76100

2. Molecular Cell Department, Weizmann Inst. of Science, Rehovot, Israel 76100

\*email: ofer.rahata@weizmann.ac.il

## Introduction

Proteins are the pixels of biological functions, the smallest functioning unit. Among the amazing properties of proteins is the ability to self-assemble. Indeed, in physiological conditions, proteins fold to the encoded structure, with an enormous variability. Since structure predictions is not available, the Structural Genomics Initiative (SGI) was launched recently, an initiative that enrich with new entries, the Protein Data Bank (PDB). This data bank is perhaps the database with the highest rate of price-per-bit of data. Each entry of the PDB relates to a protein (single chain or multi-chain) for which structural data is available, usually solved by either X-ray crystallography or NMR.

Yet, the number of known protein structures is a few orders of magnitude less than the number of sequences, and so structure prediction at different levels is of high interest. Abstraction of the structural data is a key step towards revealing of sequence-structure correlations, a first step in protein prediction algorithms. Furthermore, abstraction serves also to uncover structure-function relations. One successful example of such an abstraction is to build a library of small, continuous stretches of proteins and cluster it down to obtain fold families [1]. In this context, discovery of sequence-structure relations results in insights both for structure prediction of a given gene (which in turn may give an insight of the function), and also the opposite direction: finding new sequence for a given desirable fold (protein design).

Contact maps (or networks) have been studied as structure abstractions in order to accelerate design of decoy folds [2]. Vendruscolo ET. Al. [3] revealed the small-world architecture of these networks, and related hub vertices to energy hot-spots (hubs are residue with high connectivity). Dokholyan ET. Al. [4] correlated the graph connectivity with folding rate. It is interesting to note that folding time of proteins vary from a few milliseconds to a few minutes. Brinda et. Al. [5] further associated hubs with energy hotspots of protein-protein interactions.

Various definitions of these networks were suggested, usually based on a distance threshold between two non-covalent interacting atoms. These thresholds in turn affect tremendously network architecture, with a difference of as low as 0.2Å (way below the cited resolution in all structures) give rise to a different of as much as 30% of the edges. In this context, it is important to analyze contact maps using learning tools that are less

noise sensitive. We showed that cluster analysis is fruitful both in understanding mutation data [6] and in studying the evolution of protein-protein interactions [7], showing the network robustness with respect to threshold variation. We further applied Network Motif analysis, in order to reveal fold properties. Motifs are sub-graphs which appear in the network with a significantly higher frequency than random. The motifs are the building block of proteins structures, while the profile of motifs occurrence suggested by the analysis of over 2500 PDB structures, characterizes the architecture of fold.

### Network Motif Profiles

Secondary structures (i.e. helices, strands) are key descriptors of a protein fold. Prediction of secondary structure from sequence achieved a success rate of as high as 80%-85%, yet this rate has been stable for a long time. In this context, it is interesting to note, that expert assignment of secondary structure (that is, the definition of the helices and strands for a solved protein structure) is also ambiguous, at least at the level of 15%-20%. Previously [8] we showed that contact network analysis can be used to rediscover the secondary elements in an unsupervised manner. Using a novel random network model for proteins, we map the huge world of 6-nodes motifs. The most frequent motifs relate to helices and strands, whereas less frequent motifs (but still highly significant) relates to novel secondary structure elements. In addition, motifs in protein networks identify specific structure/function clefts.

### Evolution of Protein-Protein interactions

Protein-Protein interactions include enzyme-inhibitor, hormone-receptor, antibody-antigen etc., and occur in diverse affinities in the range of 15 orders of magnitude. Deriving the affinity from the contact map is very difficult, due to the non-additivity of its components<sup>6</sup>. We introduced a method to identify sub-networks of the atomic interaction networks, which are additive. These separation property turned to be evolutionary conserved among homologous interfaces. One example is the crystal structure of Hemoglobin, for which a wealth of complexes entries exist in PDB. We showed that the modularity is conserved throughout the species, even though the atomic details of the interactions vary. Another example is of the Human Growth-Hormone -receptor, in which different clusters are suggested to confer protein interfaces with different functional specificity.

### References

1. Bradley P, Malmström L, Qian B, Schonbrun J, Chivian D, Kim DE, Meiler J, Misura KM, Baker D. Free modeling with Rosetta in CASP6. *Proteins* 2005 **61 Suppl 7**:128-34.
2. Vendruscolo M, Najmanovich R, Domany E. Protein folding in contact map space, *Physical Review Letters* 1999; **82**:656-9.
3. Vendruscolo M, Dokholyan NV, Paci E, Karplus M. Small-world view of the amino acids that play a key role in protein folding. *Phys Rev E* 2002; **65**:061910.
4. Dokholyan NV, Li L, Ding F, Shakhnovich EI. Topological determinants of protein folding. *Proc Natl Acad Sci USA* 2002; **99(13)**:8637-8641.
5. Brinda KV, Vishveshwara S. Oligomeric protein structure networks: insights into protein-protein interactions. *BMC Bioinformatics* 2005; **6**:296-310.
6. Reichmann D, Rahat O, Albeck S, Meged R, Dym O, Schreiber G. The modular architecture of protein-protein binding interfaces. *Proc. Natl. Acad. Sci. US* 2005; **102**:57-62.

7. Rahat O, Yitzhaky A, Schreiber G., Cluster Conservation as a Novel Tool for Studying Protein-Protein Interactions Evolution. to appear in *Protein*.
8. Raveh B, Rahat O, Basri R, Schreiber G. Rediscovering secondary structures as network motifs - an unsupervised learning approach. *Bioinformatics* 2007; **23(2)**:e163-9.
9. Mintseris J, Zhiping W. Atomic Contact Vectors in Protein-Protein Recognition *Proteins* 2003; **53**:629-639.