Active Site Network, an approach for understanding the evolution of protein function

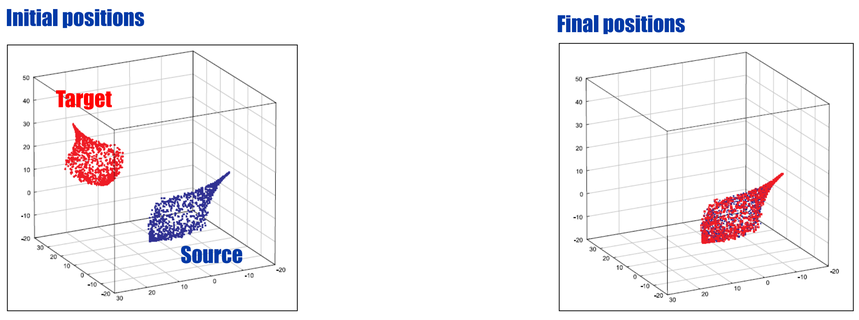
Pål Røynestad, Soumi Sengupta, Eivind Almaas
Norwegian University of Science and Technology

Summary

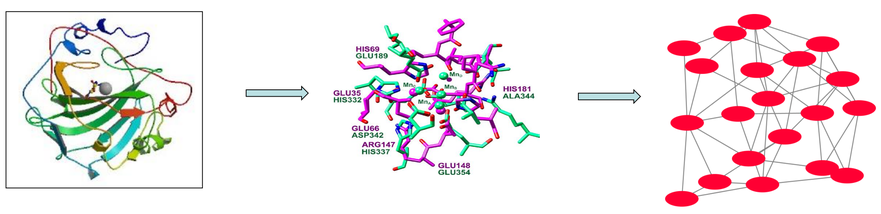
- The biological and chemical functions of enzymes, the class of proteins that carry out the majority of chemical processes in cellular metabolism, is heavily connected with the 3D structure of a particularly localized region called the active site
- We have developed a method for pair-wise comparisons of protein active sites that will form the basis for a protein similarity network, the active-site network (ASN).
- By focusing on the active sites of proteins, our approach is inherently substrate centric. Thus, it will allow for the systematic discovery of novel genes important for the Human Disease Network, identify potentially unintended pharmaceutical drug targets, as well as groups of proteins which may be of interest for enzyme engineering.
- Such a substrate-centric view of the proteome will allow the systematic discovery of unintended pharmaceutical drug targets, as well as groups of proteins, which may be of interest for enzyme engineering. As active sites are fairly easy to identify in metalloproteins, a preliminary network based on roughly 10,000 metalloproteins from the Protein Data Bank has been constructed, and other methods are currently being explored for other protein classes.

Methods

- To build the ASN, protein structures of metalloproteins were downloaded from the PDB database.
- In order to do pairwise comparisons between the downloaded protein structures, proteins were represented as point clouds with all atoms within 5Å of the metal ion being considered the active site. To do the comparison, the iterative closest point (ICP) algorithm was used.

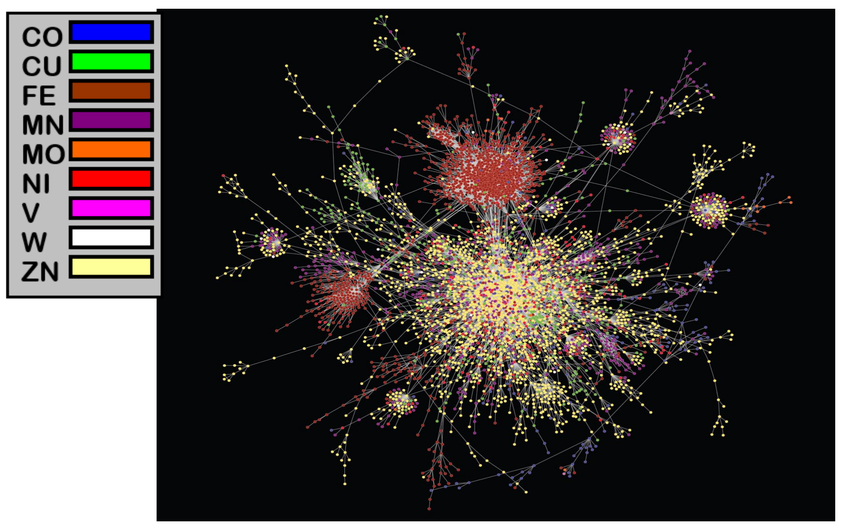


- To construct a network, we start with the pre-determined protein structures. After identifying the active site (for metalloproteins we assume that the protein is centered around the metal ion), we represent the protein structure as a set of geometric points. We then calculate the root mean square deviation (RMSD) score. Low RMSDs then suggest similar active site geometries. Finally, including only significant pair-wise scores, the final ASN can be constructed.

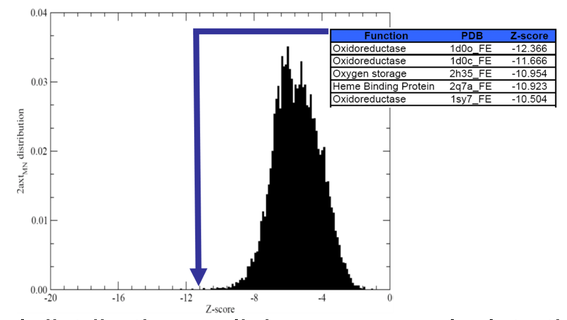


Results

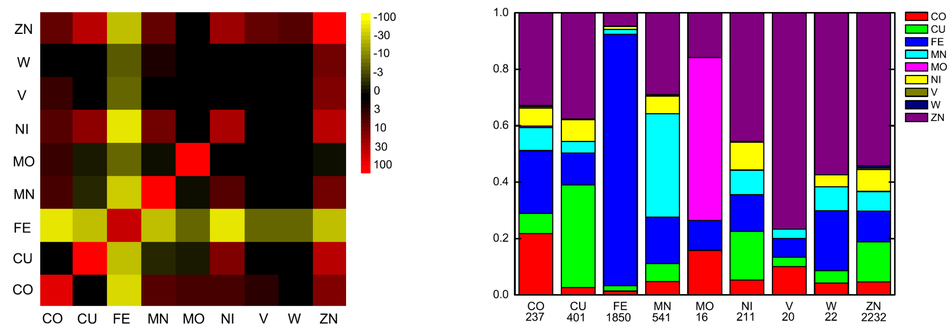
- Final ASN network of metalloproteins, consisting of 8361 protein, with an average degree of 4.1.



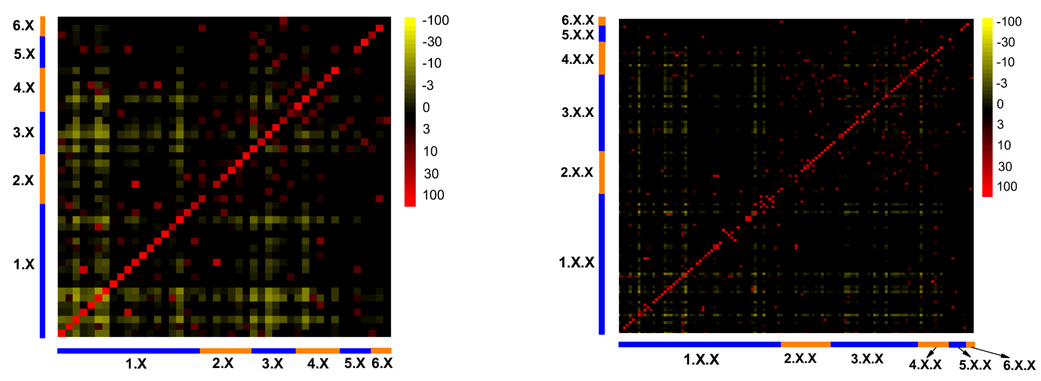
- Previous studies of the active site of the water-oxidizing complex of photosystem II found an ancient connection between it and a number of bi-manganese proteins of diverse function [1]. The ASN approach is able to extend this as well to other types of metalloproteins.



- Looking at the metal distribution on links, we can calculate the degree of overrepresentation of certain metal-metal links in the network.



- Using Enzyme Accession numbers, we can see if similar protein function tends to be found between connected proteins. Here we find that proteins of similar function are connected much more often than we would expect from random chance, suggesting that the method is producing reasonable results.



- Attempts have been made to examine proteins of unknown function. One example is in *Mycobacterium tuberculosis:* Rv2431c is a PE family protein with unknown function. Mn ion is bound to its active site by Glu and Asp. Most of its first neighbours are hydrolase and their active sites constitutes of Glu, Asp and His. Therefore the probable function of the protein may involve hydrolytic activity.

[1] J. Raymond et al, Coord. Chem. Rev. 252, 377 (2008).