Allosteric Pathway Formation in Globin Proteins Explored With Network Analysis Models

J. Pfeffer and K.N. Woods

Carnegie Mellon University
Pittsburgh, PA 15213, USA
jpfeffer@cs.cmu.edu, knwoods@cmu.edu

ABSTRACT

In the context of drug design and protein engineering, understanding the nature of allosteric networks is of particular interest. A pre-condition for successfully developing a rational design of proteins that incorporates allostery is the capability of understanding the structure and dynamics of protein internal communication pathways and networks. In this article, we compare computational approaches utilizing network analysis models and methods with results from THz time-scale spectroscopy experiments. We are able to show that community detection applied on network models of coevolution within a protein family can provide insight about the signal propagation pathways in protein subfamilies. Experimental results connect these communities with protein dynamics. We deduce that different routes of communication may be linked with the discrimination and dynamics of ligand binding.

Keywords: allostery, protein engineering, coevolution networks, community detection

1 Introduction

In contrast to orthosteric drugs that bind to active sites of proteins, allosteric drugs are intended to bind to non-active sites which allows for more specialized functionalities that may have less significant side effects. Proteins are inherently dynamic, and this property is critical for the allostery they exhibit. It is possible to use small molecules to both activate and inhibit protein function at a distance from the active site. However, for effective drug design based on protein structure it is necessary to unravel protein internal dynamics and signaling pathways that are triggered by different molecule ligand binding.

We use myoglobin (Mb) to study protein dynamics. Mb is responsible for storing and transporting oxygen (O_2) in mammalian muscle tissue. However, Mb also binds other small molecules, e.g. monoxide (CO), nitrous oxide (NO), and cyanide (CN^-). In this article, we focus on the network analysis aspects of our analysis. In particular, we discuss extraction of the network information utilizing multiple sequence alignment (MSA) and mutual information (MI) calculations, visualization of protein networks, and application of community detection algorithms. We also apply THz time-scale spectroscopy experiments in the frequency region of the protein spectra that is related to localized internal protein interactions. Details about the THz time-scale spectroscopy experiments as well as other aspects related to this work are discussed in [8].

2 Evolutionary Protein Networks

Our initial investigation involves studying the evolutionary signal propagation network of proteins within the globin family. We utilize experimental methods that are capable of detecting both the global protein fluctuations that are connected with conserved interactions across the subfamily and simultaneously, the more local-

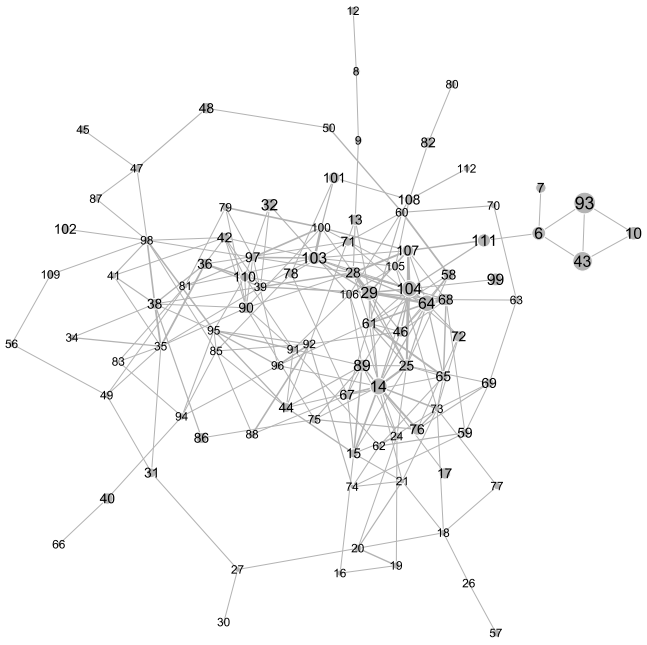


Figure 1: Network of the 96 resolved residues from the MSA of the globin protein family. Nodes represent amino acids (residues) with sperm whale myoglobin as reference structure (1mbn.pdb). Edges result from MI calculations.

ized intra-protein interactions that support specific signal propagation pathways formed by coevolving residues.

We use networks [5] to model coevolution within a protein family and social network analytical methods [2] to study the structure of these network models. We extract the coevolved information of the globin family through MSA of 4,630 sequences (max. 20% gaps) from the Pfam database and use sperm whale myoglobin as reference structure (1mbn.pdb). In terms of network models, nodes are amino acids (residues) that are conserved over multiple proteins within the protein family. To define the edges of the network, we apply MI calculations with a z-score threshold of 6.5 to reveal the extent of coevolutionary relationships between pairs of residues. Conservation and MI were determined with the MISTIC web server [7]. The resulting network can be seen in Figure 1; numbering of residues refers to the reference protein. The visualization position of nodes is calculated by applying distance scaling optimization [1], a network layout algorithm based on multi-dimensional scaling. The size of the nodes denotes conservation from the MSA. The line weights represent MI of pairs of residues, i.e. coevolution.

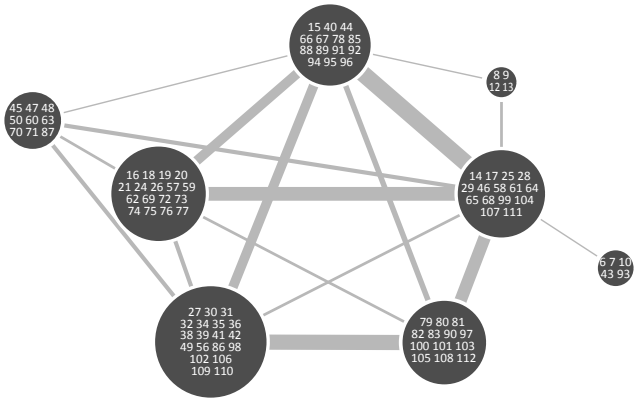


Figure 2: Aggregated visualization of communities showing the residues that are part of the 8 communities. Node size is number of residues per community. Edge width is number of connections between residues of different communities.

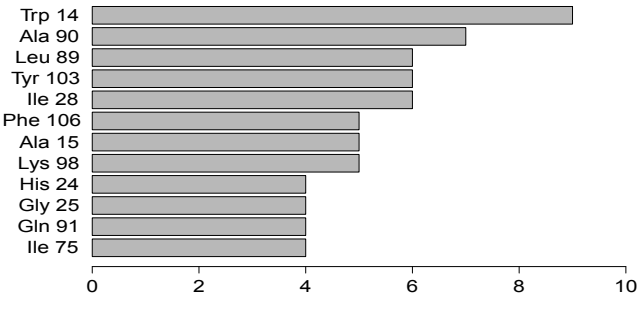


Figure 3: Top *Linker residues*. Number of connections to other communities.

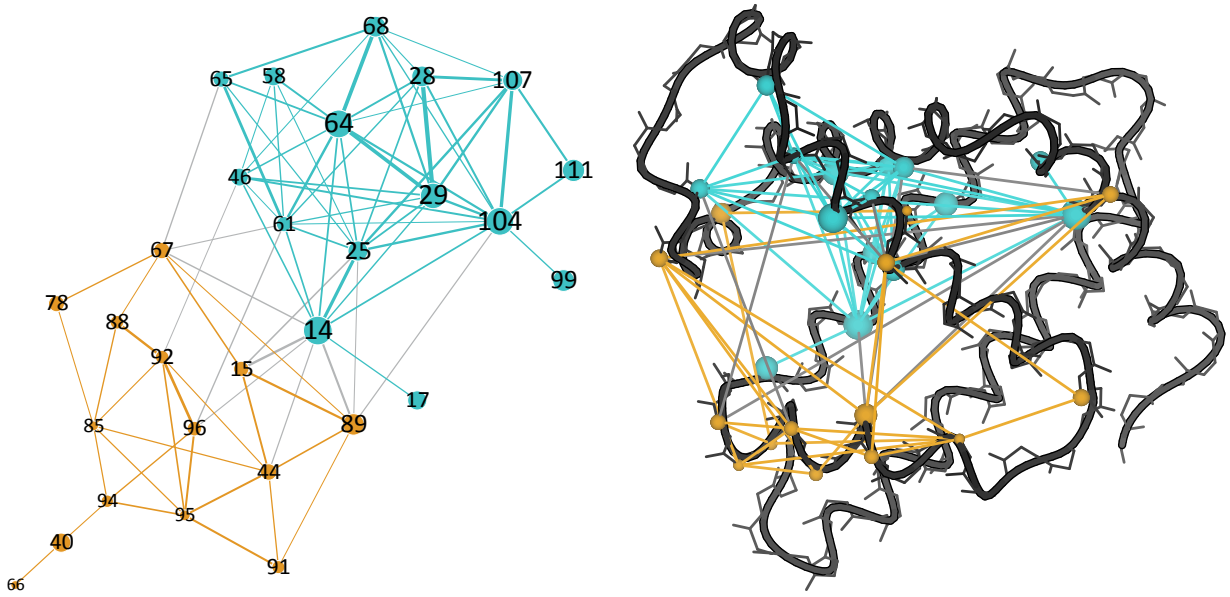
The visualization shows a dense area connecting several residues of the functional core of the protein and residues in the heme distal pocket, e.g. His 64, Val 68. This area of the network also include the pairs of residues with the strongest connections (MI). Note, that the two highest conserved residues, His 93, that is connected to the heme group, and Phe 43 do not share significant MI scores with the central residues and are structurally located on the periphery. In general, different areas of the network result from different sub-sets of the evolutionary-connected protein family and thus may be related to different functionalities. To reveal these areas of the network we perform community analysis.

3 Community Analysis

Communities in networks are groups of nodes that are denser connected within than among the groups. We apply Newman's clustering algorithms [6] to identify non-overlapping communities of residues. The clustering reveals 8 groups as optimal grouping with a modularity value of 0.51. The groups of residues as well as the connections among the communities are drawn in an aggregated network in Figure 2. Two of these groups are interesting and are visualized in Figure 4. The two dimensional network visualization (Figure 4(a)) of the evolutionary network shows that these two communities are structurally sparsely connected. In Figure 4(b) the nodes (residues) and edges of these communities are mapped onto a cartoon representation of the protein reference structure of Mb revealing their physical proximity. These communities of coevolved residues show high overlap with patterns identified with Force Distribution Analysis [8].

3.1 Communities as Signal Propagation Pathways

Community 1 consists mostly of residues on helix F that suggests that this set of residues create a pathway that connects areas proximal and distal to the heme. Community 2 consists of highly conserved residues (Trp 14, Leu 29, His64, Leu 104) as well as a couple of other functionally important residues. These two groups are very similar to the allosteric networks that are identified with force distribution analysis (FDA) in Mb bound with specific ligands[8]. We propose that the optimal route of communication (represented by community 2 of the network community analysis) is tied with O_2 ligand binding. A second group of residues (community 1) that form a less direct pathway of information flow in the protein network may be associated with binding to other small molecule ligands [8].



(a) 2-Dimensional network visualization

(b) Network mapped onto a cartoon representation

Figure 4: Two of eight network communities from the coevolution network in 2-dimensional network and in 3-dimensional protein visualization. Node sizes represent conservation, different node colors indicate groups. The edges within the distinct communities are colored with the group color; links between groups are drawn in gray. The position of the nodes in the 3D visualization is the position of the respective Cα atom.

3.2 Linker Residues

Looking at the intersection of these two communities, we find that Trp 14 plays a very crucial role in inter-community connections. Interestingly, it was shown in previous work that this residue is important for long range communication and energy transfer in the protein [3]. We assume that these *linker residues*, that connect different communities of coevolved residues, are essential for long-range communication within distal areas of the protein. Consequently, we calculate all linker residues of the globin family.

Starting with the result of the community analysis described above, we remove links from the coevolution network (Figure 1) that are within communities to identify all residues with connections to other communities. Figure 3 enumerates the residues with the most connections to other communities—we suggest that these are the linker residues of the globin family. We can see that Trp 14 is the top-ranked residue in this list.

4 TeraHertz Time-Scale Spectroscopy

To determine how protein dynamics are interconnected with signal propagation within this protein family, we triangulate our computational approaches with experiments that identify and characterize the allosteric networks that form in response to the binding of specific ligands. Explicitly, we perform THz time-scale spectroscopy measurements in the $20 - 170cm^{-1}$ spectral region on Mb, a prominent member of the globin family, to characterize both the global fluctuations that reflect protein intrinsic dynamics as well as the localized interactions that are tied with the formation of specific allosteric pathways (signal propagation pathways) in the protein three-dimensional structure[4].

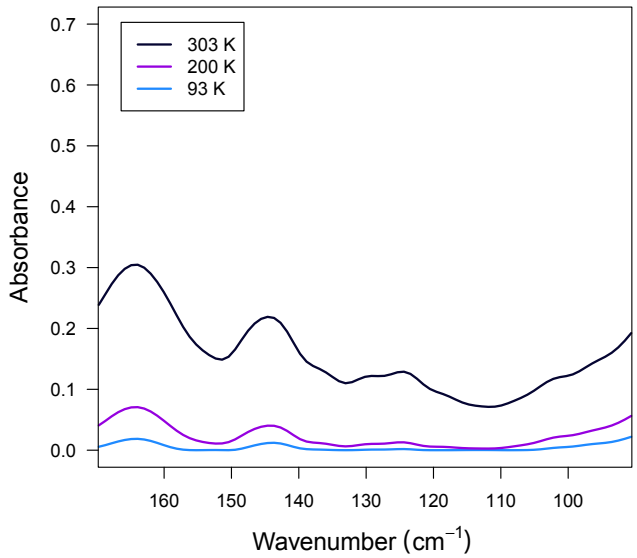
The THz time-scale spectroscopy experiments include carbon monoxide binding myoglobin $MbCO$ and oxygen binding myoglobin MbO_2. In Figure 5 the higher frequency region of the protein spectra are plotted and highlight the localized interactions among the coevolved residues. The very distinct spectra indicate that different residues are excited and different signal propagation pathways are formed in the protein three-dimensional structure. In addition, the intramolecular dynamics of $MbCO$ (Figure 5(a)) reflect harmonic motion that suggests a suboptimal pathway while the anharmonic motion in MbO_2 binding depicts an optimal pathway in the protein structure with respect to storing and transporting oxygen. For more details on the THz time-scale spectroscopy of the protein spectra see [8].

5 Conclusions

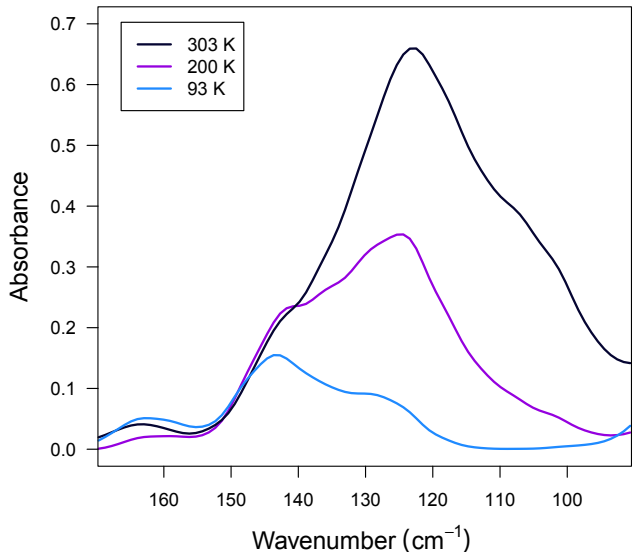
In this paper we have shown that network analysis models and methods, particularly community analysis, can be used to identify different areas in a network of coevolved resides. These communities seem to be cor-

related to signal propagation pathways in the protein. Linker residues connecting multiple communities may play an important role in long range communication and energy transfer.

Preliminary results indicate that allosteric communication within the globin family comprises multiple pathways of communication that are characteristically intertwined with functional associations and dynamics. Particularly, we find that ligand binding may be the source of the diverse, altered routes of communication within the protein family and within these routes there are particular, key residues that have evolved to mediate signaling by means of coherent thermal fluctuations along the alternate pathways. We propose that the most optimal route of communication in the Mb protein sequences is directly tied with O_2 ligand binding whereas binding of other small molecule ligands exploit less direct pathways of information flow in the protein interaction network.



(a) *MbCO*



(b) MbO_2

Figure 5: Experimental THz spectrum of a hydrated film sample of (a) *MbCO* and (b) MbO_2 in the $100 - 170cm^{-1}$ spectral region at 93 K, 200 K, and 303 K. The fluctuations detected in this region of the spectrum relate to the localized protein interactions that highlight specific allosteric pathways that form in response to ligand binding.

Future work will examine whether we can describe and predict the molecular mechanisms that are tied with the network of interactions that propagate an allosteric signal within a protein subfamily, not only for globins, but also for other medically relevant protein families. And if so, whether there is a potential to use this information in both assessing the formation of potential diseases within a protein sub-family, and conversely as an aid in engineering allosteric drug therapies to combat the disease.

References

[1] Brandes, U. and C. Pich (2007). Eigensolver methods for progressive multidimensional scaling of large data. *Proceedings of the 14th International Symposium on Graph Drawing (GD'06)*, 42–53.

[2] Hennig, M., U. Brandes, J. Pfeffer, and I. Mergel (2012). *Studying Social Networks. A Guide to Empirical Research*. Frankfurt: Campus Verlag.

[3] Hochstrasser, R. M. and D. K. Negus (1984). Picosecond fluorescence decay of tryptophans in myoglobin. *Proceedings of the National Academy of Science 81*(14), 4399–4403.

[4] Liu, Y. and I. Bahar (2012). Sequence evolution correlates with structural dynamics. *Molecular Biology and Evolution 29*(9), 2253–2263.

[5] Newman, M., A.-L. Barabasi, and D. J. Watts (2006). *The Structure and Dynamics of Networks*. Princeton, NJ: Princeton University Press.

[6] Newman, M. E. J. and M. Girvan (2004). Finding and evaluating community structure in networks. *Physical Review E 69*(2), 026113.

[7] Simonetti, F. L., E. Teppa, A. Chernomoretz, M. Nielsen, and C. Marino Buslje (2013). MISTIC: mutual information server to infer coevolution. *Nucleic acids research 41*(Web Server issue), W8–14.

[8] Woods, K. (2014). Using THz time-scale infrared spectroscopy to examine the role of collective, thermal fluctuations in the formation of myoglobin allosteric communication pathways and ligand specificity. *Soft Matter*, DOI: 10.1039/C3SM53229A.