Determination of Specificity Residues in Two Component Systems using Graphlets

Faruck Morcos[1], Charles Lamanna[2], Nitesh V. Chawla[3], and Jesús A. Izaguirre[4]

[1] Department of Computer Science and Engineering
Cushing Hall 214
University of Notre Dame, Notre Dame, IN 46556, USA
Email: amorcosg@nd.edu

[2] Department of Computer Science and Engineering
384 Fitzpatrick Hall
University of Notre Dame, Notre Dame, IN 46556, USA
Email: clamanna@nd.edu

[3] Department of Computer Science and Engineering
353 Fitzpatrick Hall
University of Notre Dame, Notre Dame, IN 46556, USA
Email: nchawla@nd.edu

[4] Department of Computer Science and Engineering
326C Cushing Hall
University of Notre Dame, Notre Dame, IN 46556, USA
Email: izaguirr@nd.edu

Correspondence:
amorcosg@nd.edu

Keywords:
domain-domain specificity, graph-based learning, two component systems, graphlets

BIOCOMP'09

Determination of Specificity Residues in Two Component Systems using Graphlets

F. Morcos[1,2], **C. Lamanna**[1], **N. V. Chawla**[1,2] **and J. A. Izaguirre** [1,2]
[1] Department of Computer Science and Engineering , University of Notre Dame, Notre Dame, IN, USA
[2]Center for Complex Networks Research , University of Notre Dame, Notre Dame, IN, USA

Abstract— *This work presents a novel method for the identification of specificity residues in two component systems based on the discovery of graphlet signatures. We use network representations of 3-D structures and sequence of proteins, experimental data and graph-based learning to detect graphlet signatures that potentially are responsible for phosphotranfer specificity between Histidine Kinase (HK) and Response Regulator (RR) domains. This approach is applied to the system of HK and RR in E. coli. Structural regions were found for Histidine Kinases RstB and Response Regulator RstA and confirmed using experimental data. In addition, some hypothetical regions of specificity were proposed to explain cross talk between the HKs that phosphorylate YhfA and the RRs that interact with UhpB. Such an approach offers the ability to identify domain specificity residues in two component systems in silico.*

Keywords: domain-domain specificity, graph-based learning, two component systems, graphlets

1. Introduction

At the present time, descriptions of interactions between domain families lack specificity at a domain level. Although data exists for more general domain families, interactions between individual domains lack specific dynamics and estimates for interaction. One of the most important mechanisms for signal transduction in a bacterial cell is termed the Two Component System (TCS). This mechanism serves as a phosphotransfer system for the transmission of information across membranes and within the cytoplasm. This process is illustrated in Figure 1.

The standard composition of a TCS includes an enzyme called a Histidine Kinase (HK) and a Response Regulator (RR) (red and blue, respectively, in Figure 1), which sense external stimuli and signal appropriate responses. A typical Two Component System contains a HK with a sensor domain, usually found in the membrane, which identifies external stimulus (step 1 in Figure 1). The HK catalyzes ATP and autophosphorylates its histidine residue (step 2), then the phosphoryl group of the HK covalently modifies the RR (step 3) that in turn, serves as a trigger of several cellular responses (step 4) and in some cases as transcription factors [12].

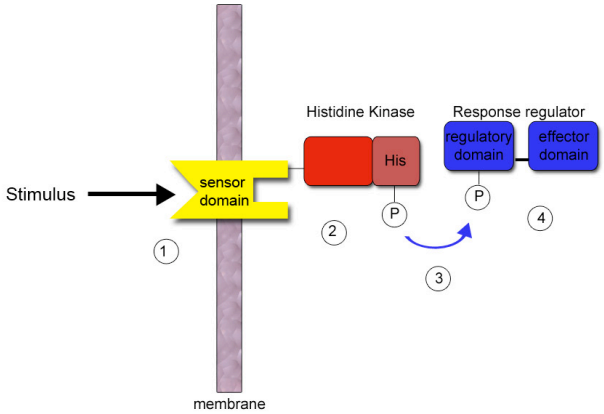

Fig. 1: **Phosphotransfer in TCS.** A HK responds to an external stimulus and transfers a phosphate group to the regulatory domain of the RR.

Phosphorylation in a RR induces conformational changes that activate its regulatory domain. The TCS help regulate processes like osmoregulation, transport, metabolism and chemotaxis. Two Component Systems are neither observed in humans nor most eukaryotes [9]. However, understanding their specificity could have medical implications in the development of antibacterial drugs as Two Component Systems are prevalent in pathogenic bacteria.

There is typically a one to one interaction mapping between an HK and an RR; however, in some cases, one Histidine Kinase activates several Response Regulators and vice versa. This system underscores a need to identify which Histidine Kinase is connected to which Response Regulator. Histidine Kinases and Response Regulators that belong to the same operon are called cognate. Nevertheless, it is common to find an *orphan* HK that has a locus distant from its corresponding interacting response regulators, a fact that complicates HK-RR matching. Kinases and Response Regulators share sequence and structural characteristics; consequently, the mechanism for pair specificity and cross-talk prevention remains to be solved.

We propose a novel computational technique to estimate domain-domain interaction specificity between Response Regulators and Histidine Kinases which not only considers sequence but also 3-D structure in an attempt to capture the

regions responsible for these particular domain interactions.

2. Identifying Specificity with Graphlets

This approach utilizes the 3-D structures and sequences of Histidine Kinases and Response Regulators to develop corresponding network representations. In these network representations, each node represents a residue (amino acid) and an edge between two nodes signifies that the nodes are sufficiently close in the 3-D structure (i.e. a distance less than 5 Å) or contiguous in the protein sequence. Such a representation yields a network with edges along the protein backbone and clustering among residues in close proximity.

Once the domains have been transformed into networks, graphlet signatures are identified to devise domain-domain specificity. This permits the quantification of specificity between Histidine Kinases and Response Regulator domains. Furthermore, domain interaction data from experimental phosphotransfer events were used to train and validate our models.

2.1 Methodology

We have developed a methodology to the problem of domain-domain interaction specificity in a Two Component System. We propose the following series of steps as a means of identifying key residues:

1) Obtain 3-D structural data of Histidine Kinases and Response Regulators (using homology modeling as necessary).
2) Align the sequences of the Histidine Kinases. Independently align the sequences of the Response Regulators.
3) Convert the 3-D structural data for all Histidine Kinases and Response Regulators into network representations.
    a) Create a node for each residue in the protein sequence.
    b) Draw an edge between nodes that are sufficiently close as per the 3-D structural data (e.g. <5 Å).
4) Divide the network representations of the Histidine Kinases and Response Regulators into positive and negative examples of specificity.
    a) Examine the case of RstB, CpxA, UhpB and YfhA; as well as their respective cross-talk.
5) Utilize SUBDUE in order to find substructures in the network representations that corresponds to interaction specificity.
    a) Maximization of the number of substructures belonging to positive examples but *not* negative examples of protein network representations.
6) Correlate key substructures in network representations to protein sequence.
7) Compare substructures with previously identified residues that influence specificity.

The following subsections describe these steps in greater detail.

2.2 3-D Structural Data

We compiled and organized a knowledge base of Protein Data Bank (PDB) [5] files containing the three dimensional atomic structure of proteins belonging to the Two Component System. The structural information of PDB files is obtained by crystallography and nuclear magnetic resonance methods and are stored in the Protein Data Bank (accessible on-line at http://www.pdb.org). For those proteins without readily available 3-D representations, homology modeling was employed. Lastly, the residue sequences composing these 3-D models were aligned in order to be compared.

2.3 Conversion of 3-D Structural Data into Network

The approach selects all the alpha carbon (α-carbon) atoms of the structural data, as these atoms are considered representative of each amino acid side chain. The network representation labels these α-carbon with the amino acid side chain it represents. The resulting protein network representation is a more sparse fashion than if we had used atomic bonds.

This algorithm creates a matrix that contains the distances between each α-carbon of the protein. It treats each residue in the molecule as a node in a graph and, based on the previously mentioned distance matrix, draws an edge between them if the distances between their corresponding α-carbons is beneath a threshold value of t. This threshold value t is a pliable parameter for our algorithm. A proximity topology between residues at a certain threshold can provide insight into critical residues in addition to their biochemical composition

After creating this matrix, our program identifies the nodes and edges that best represent the 3-D structural data. These network representations are then output in file formats that permit visualization and analysis of the protein network representations.

2.4 Network Representation of Proteins

Figure 2 shows an example of a network generated with our framework and visualized in Cytoscape [10]. In this figure, every node represents a residue in a protein and the edges are drawn between residues that are closer than $t = 10$ Å. This representation is particularly dense; consequently, our method uses a distance of 5 Å, as it yields a sparse network that maintains biological significance.

An example of the visualization of both protein and structural data for the Two Component System proteins Envz (pdb:1joy) and Spo0F (pdb:1f51) can be seen in Figures 3a and 3b, respectively. These representations were generated following steps (1-3) of our methodology. Although the graphs are relatively sparse, they encode key sequence and

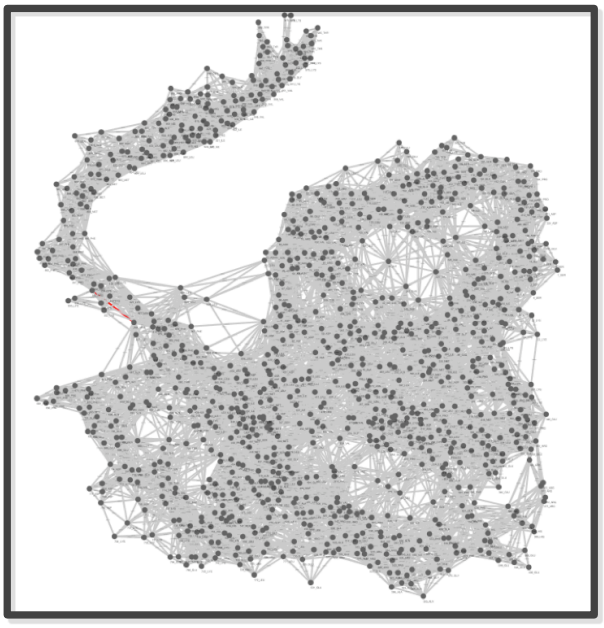

Fig. 2: Network visualization of a protein 3-D structure using Cytoscape. This network was created with the program described in Section 2.3.

structural information. Furthermore, we colored the residues red that are known to impact the specificity of proteins [11].

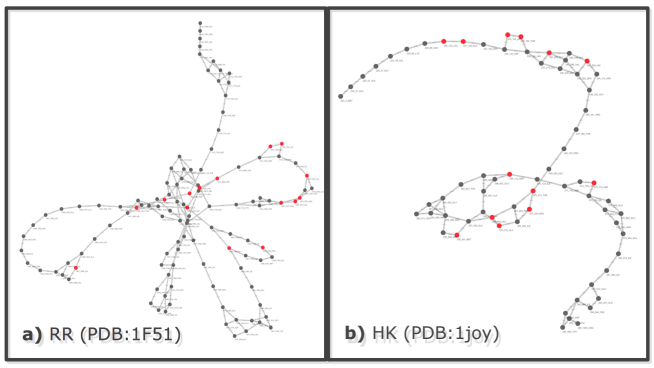

Fig. 3: **a)** Network transformation of protein Spo0F (pdb:1f51) . **b)** Network representation of Histidine Kinase EnvZ (pdb:1joy) .

2.5 Identification of Graphlets with SUBDUE

After generating the network representations of the 3-D protein structures, we utilized SUBDUE to isolate recurring motifs or substructures within these networks. SUBDUE offers the ability to identify recurring substructures within a set of networks as well as graph learning capabilities [3]. SUBDUE has been used for a variety of applications including the study of proteins [4].

Repeating substructures within the 3-D protein structure underlie key components of the specificity of a protein interaction. SUBDUE is able to identify those graphlets correlating to protein specificity through its supervised graph based learning. These recurring graphlets represent protein signatures that define interaction specificity and thereby predict specificity.

We use supervised graph based learning to identify key graphlets. The SUBDUE Set Cover algorithm identifies substructures that belong to a particular set of networks (the "positively labeled" network) that do not belong to another set of networks (the "negatively labeled" networks). The Set Cover algorithm attempts to find the graphlets or substructures of a particular network representation that maximize the following: the number of positively labeled examples containing that graphlet, plus the number of negative examples not containing that graphlet, divided by the number of positive examples plus the number of negative examples.

The result would then capture the graphlets that underscore particular specificity properties, allowing the differential analysis of Histidine Kinase and Response Regulator protein interactions.

3. Application to Two Component Systems in E. Coli

We used our methodology in a set of Two Component System pairs found in the model bacteria *E. coli*. For the case of two component systems, *E. coli* contains a number of HK-RR pairs that have been identified and tested experimentally for phosphotransfer. Yamamoto et al. [13], performed a series of experiments to investigate phosphotransfer in *E. coli* for a group of 55 Kinases and response regulators. They could identify which HK would interact with which RR as well as cross-talk events in these proteins. Some of these findings are summarized in Figure 4, where the left column illustrates the set of HKs and the right column a set of RRs.

Figure 4 also shows which pairs are cognate (black line) and which have non-cognate interactions or crosstalk (red lines). We obtain the protein sequences for these proteins using the Uniprot database [2] and their 3D structures from the Protein Data Bank [5]. For those cases where there was no crystal or NMR structure we obtained such structures from the Swiss Model Homology Database [7]. This repository constructs hypothetical 3-D structures of proteins based on their sequence and similarity with other proteins that have experimentally determined 3-D structures. We used these proteins in order to obtain positive and negative examples of HK-RR interactions. The combination of positive and negative samples provide a mechanism for pinpointing those components that do (or do not) impact domain-domain specificity.

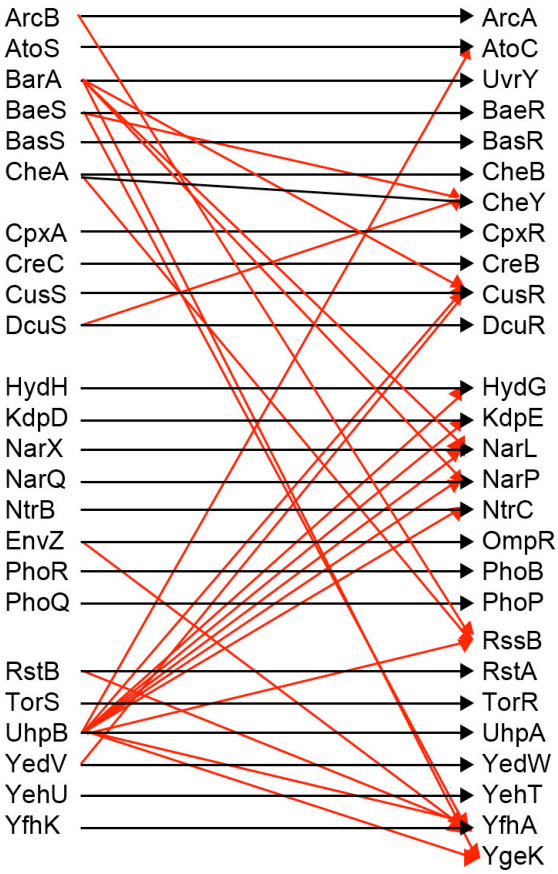

Fig. 4: Two Component System Pairs as found in *E. coli* [11]. Histidine Kinases are on the left and Response Regulators on the right. Cognate pairs are linked by black lines and non-cognate interactions by red lines.

3.1 Predictive Capabilities

We used the SUBDUE graph based learning to identify key structural distinctions between Response Regulators in the *E. Coli* two component system. We labeled *positive* examples of a particular protein interaction specificity along with *negative* examples. The substructures identified by the set cover algorithm would be closely tied to the specificity of that particular protein, allowing our methodology to predict future specificity in protein interactions. Identifying a recurring substructure in these proteins that yields a particular specificity would provide a mechanism for identifying proteins with similar properties. Furthermore, our approach would be able to adaptively identify substructures that are specific to individual cognate pairs in the *E. Coli* Two Component system and thereby predict interaction specificity. Additionally, since there is extensive experimental coverage of the *E. Coli* Two Component system, we were able to compare the accuracy of our computational approach.

As our primary example, we sought to identify those substructures that cause Histidine Kinases to interact/interfere with the Response Regulator CpxR. Such an approach required that we label the Histidine Kinases that interact with CpxR as per the results of Yamamoto et. al [13]. Consequently, we first labeled those Histidine Kinases that were "positive" examples of specificity as those proteins that interacted with CpxR. These proteins were: CheA, CpxA, DcuS, EnvZ, PhoR and RstB. Furthermore, all other Histidine Kinases did *not* have any form of interaction with CpxR, and would therefore be labeled as negative examples: ArcB, AtoS, BaeS, BasS, CitA, CreC, CusS, EvgS, HydH, KdpD, NarQ, NarX, NtrB, PhoQ, PhoQ, QseC, TorS, UhpB, YedV, YehU and YfhK. A visualization of those Histidine Kinases that interact with CpxR can be found in Figure 5 and can be compared against the larger HK-RR set in Figure 4.

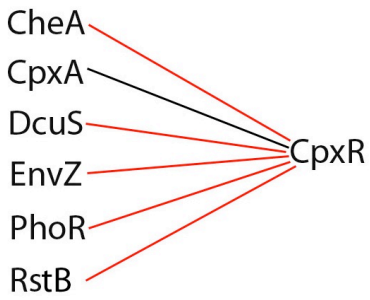

Fig. 5: The Response Regulator CpxR is on the right. Those Histidine Kinases with which it interacts are on the left. These Histidine Kinases comprise the positive set.

We were able to identify substructures that existed in four of the aforementioned six proteins labeled as "positive," and none of the graphs labeled as "negative." Although there were several such substructures, they were all centered on the same ten contiguous residues of the respective Histidine Kinases. The significance of this is described in greater detail in later sections. Nonetheless, these substructures would allow for the identification of Histidine Kinases specific (or non-specific) to CpxR at a rate of 91.3%. This identification represented a high level of precision considering the several stages involved in our current methodology.

Likewise, we utilized supervised graph based learning to find the graphlets that underscore specificity of Response Regulator interactions. The particular example we explored was those Response Regulators that showed specificity towards the Histidine Kinase RstB [13]. RstB interacts with the Response Regulator RstA as a cognate pair, in addition to engaging in cross talk with the Response Regulators HydG, CpxR, RssB, and YfhA (seen in Figure 6). Consequently, we labeled as positive examples the Response Regulators that interact with RstB and for which 3-D modeling was available: CpxA, RstB and YfhA. Alternatively, we then labeled eighteen other *E. Coli* Response Regulators as negative: ArcA, BasR, CC1181, CheB, CheY, CreB, CusR, EvgA, KdpE, NarP, NtrC, PhoB, PhoP, QseB, TorR, YedW, YehT and YpdB. These Response Regulators labeled as negative

had been experimentally shown to lack specificity towards RstB in the form of acting as a cognate pair or sustained cross-talk [13].

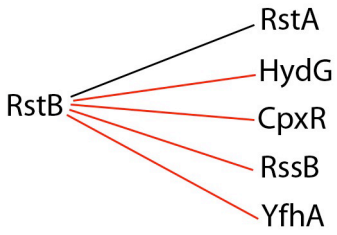

Fig. 6: The Histidine Kinase RstB is on the left. Those Response Regulators with which it is cognate or participates in crosstalk can be found on the right. These Response Regulators represent the positive set.

We then identified key structures that influence the specificity between Response Regulators and RstA. We found recurring substructures in the positively labeled graphs that did not also exist in the negatively labeled graphs. By using these substructures as the signature of Response Regulators that would interact with RstA, we detected specific properties that were shared by all three positive examples, and only one of the eighteen negative examples. This degree of accuracy signifies the ability to identify specificity (and non-specificity) for Response Regulators interacting with RstB at an accuracy of 95.2%.

3.2 Comparison with Known Specificity Regions

After having identified such substructures or graphlets that make phosphotranfer specific in a set of TCS of *E. coli*, it was important to determine if such substructures have a biological meaning and, furthermore, if such elements have been identified before. For the case of the RR RstA, the best substructure found by our method actually identified a region that coincides with the molecular interface in the only crystal structure close to a HK-RR bound structure. This graphlet is illustrated in Figure 7, which is shown to map a region (highlighted in orange) in the tertiary structure representation of the response regulator RstA. Figure 7 also shows the position of this region in the alignment produced by Laub et al. with other RRs. The region we identified is shown with some markers (asterisks) that represent the interface between HK-RR in the PDB structure 1F51.

In the case of the HK, we identified key substructures in those Kinases specific to the Response Regulator CpxR. These substructures would impact the specificity of the HK RstB, as both CpxR and RstB belong to the same positive set as described in Section 3.1. The resulting substructure found to be responsible for specificity in RstB were located in a loop between two alpha helices in the HK RstB. This

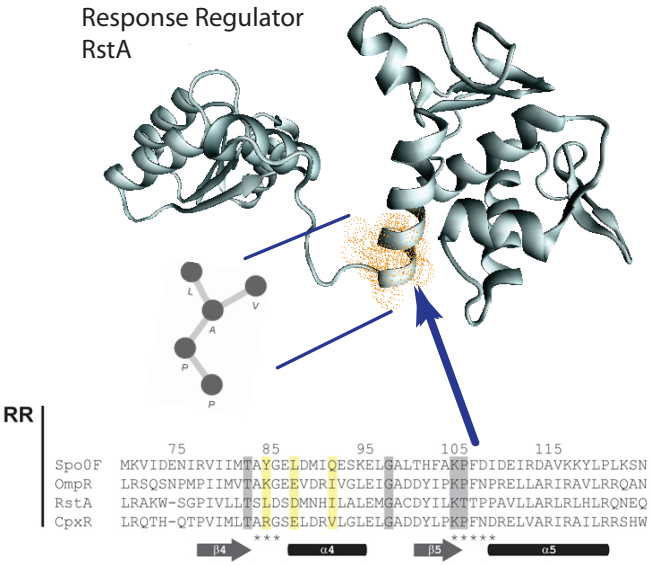

Fig. 7: 3D structure of Response Regulator RstA. The highlighted region identifies the graphlet signature found by our method. Alignment from [10].

region was shown experimentally to be the most important region for specificity in HKs EnvZ, CpxA, Spo0B and RstB. The sequence-based mutual information method proposed by Skerker et al. [11] could not identify this region because this region belongs to a loop section that is hard to align. The specificity graphlet signature as well as the 3D structure of HK RstB are shown in Figure 8. The loop region we identified is also highlighted in the loop of RstA (shown in orange).

This result provided experimental evidence of the significance of our findings and gave biological relevance to our method to find specificity regions in TCS.

3.3 Hypothetical Specificity Regions

We tested our methodology in a second set of TCS proteins in order to find specificity regions in the Response Regulators that interact with UhpB and the Histidine Kinases that interact with YfhA. These two are interesting cases where we observe cross talk. Hence, it is of relevance to try to determine which regions might be responsible for this phenomenon. We applied our method and looked for common substructures in response regulators NarL, NarP CusR, NtrC, HydG, KdpE and UhpA. All these RR are known to have phosphotransfer activity with the HK UhpB (see Figure 9a). We identified a substructure that was able to recover such interactions with an accuracy of 0.875. This substructure is shown in Figure 10a.

We also investigated the case of cross-talk between ki-

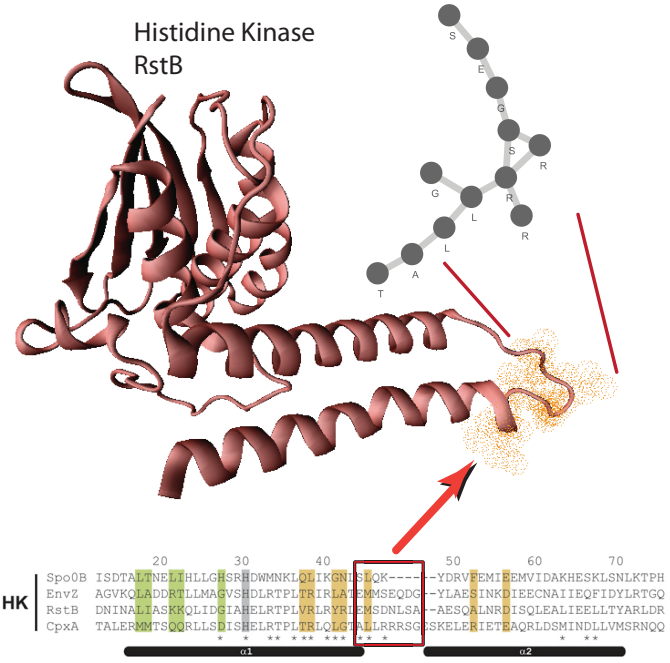

Fig. 8: 3D structure of Histidine Kinase RstB. The highlighted region identifies the graphlet signature found by our method. Alignment taken from [10].

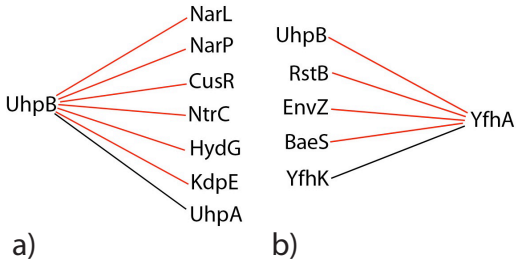

Fig. 9: **a)** The Histidine Kinase UhpB is on the left. Those Response Regulators with which it is cognate or participates in crosstalk can be found on the right. These Response Regulators represent the positive set. **b)** The Response Regulator YfhA is on the left. Those Histidine Kinases with which it interacts are on the left. These Histidine Kinases comprise the positive set.

nases UhpB, RstB, EnvZ, BaeS and YfhK with the response regulator YfhA. These HKs except Yfhk are non-cognate kinases that phosphorylate YfhA (see Figure 9b). Using our DDI specificity identification method, we found a substructure composed of five residues that was able to discern phosphotransfer interactions with an accuracy of 0.904762, whenever we combined the positive and negative sets. This region of specificity along with its graphlet is shown in Figure 10b.

According to our methodology, these regions and substructures could play an important role for specificity in-

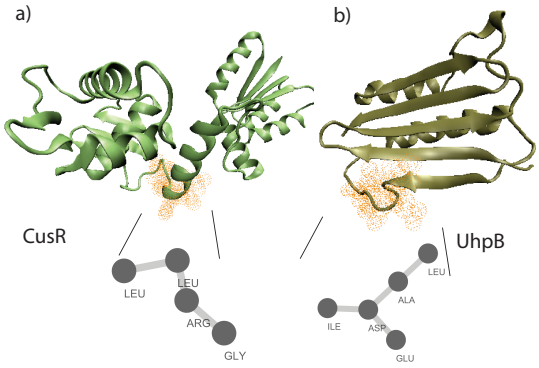

Fig. 10: **a)** Specificity graphlet in response regulator CusR. **b)** Specificity graphlet in histidine kinase UhpB.

volving proteins interacting with UhpB and YfhA. These results bring a hypothesis of why cross-talk happens in these two particular systems. If these findings are experimentally validated, then our method for specificity estimation could bring important specificity information to many more TCS systems.

4. Related Work

A novel experimental technique to test specificity of Histidine Kinases and Response Regulators in vitro was developed by Laub et al. [8], [1]. This biochemical approach was able to identify *in vivo* HK-RR pairings in an *in vitro* set up. This method called *phosphotransfer profiling* supports the hypothesis that Kinases have an *in vitro* kinetic preference towards their *in vivo* response regulators. Computational approaches to predict specificity in the Two Component System have focused on sequence. The premise is that HK-RR specificity is directly related to the particular constitution of the residues in each Histidine Kinase and Response Regulator.

One hypothesis suggests that residues close to the active site aspartate of Response Regulators contribute to the HK-RR pairing. These residues would interact directly with residues around the phosphohistidine in Kinases [6]. Fabret et al. did a comparison between the residues near the aspartate in several response regulators and noticed that these residues belong to loops connecting α-helices and β-sheets. Mutations on these loops produced a change in Kinase specificity.

5. Future Work

The approach proposed in this paper could be extended to identify specificity regions in a larger set of Histidine Kinase and Response Regulator pairs. Also, a more extensive investigation of the predictive power of this analysis could be performed, allowing for its application to new biological pathways with unknown interaction pairings. Furthermore,

a validation of this *in silico* approach by experimental biological methods would provide evidence for the accuracy of graphlet signatures in capturing domain-domain interaction specificity.

6. Conclusions

In this work, we have presented a network based approach to study molecular specificity between proteins. In particular, we focused our investigation on a very important pathway known as the two component system. This pathway is prevalent mainly in bacterial organisms and is the basis of more complex pathways in eukaryote cells. The networks approach proposed herein allows for the use of structural information to study specificity in domain-domain interactions, an examination that has been primarily investigated with sequence information or with experimental approaches.

A network representation of protein data, as opposed to atomic or sequence representations provides an extra insight by reducing the complexities of atomic considerations and adding extra relationships to the simpler use of only sequence. In this project we were able to combine many different sources of biological information ranging from protein databases, structural data, and homology modeling of proteins into rich network models. We also used graph learning techniques to identify those patterns that underpin domain-domain interaction specificity. We were able to achieve this using analysis tools like SUBDUE and by transforming protein structure and sequence into network representations. Additionally, we were able to study proteins from the organism *E. coli* that provided us with the data to build predictive models of domain-domain interaction pairing using subgraph identification.

We studied the specific case of the Histidine Kinase RstB and its corresponding response regulator interactors. Through our methodology, we were able to identify graph motifs in their protein structure that were shared only by those pairs participating in phosphotransfer within these sets of proteins. These graph motifs were able to identify domain-domain specificity in these cases with accuracies exceeding 90%. We compared these motifs with experimentally confirmed regions of domain-domain interaction specificity. These signatures turned out to be very important regions for HK-RR specificity, particularly in the case of Histidine Kinases wherein we found a region of specificity that was confirmed experimentally. Furthermore, this region could not be found by the sequence based method in [11]. This result provides support for the use of network methods that take into account both sequence and structure. Finally, we hypothesize about potential specificity substructures found by our methodology. These regions could give us an insight of why cross talk occurs within HK UhpB and its interacting response regulators as well as RR YhfA and its corresponding kinases. Further experimental validation could give additional support for the relevance of our methodology.

References

[1] E. G. Biondi, J. M. Skerker, M. Arif, M. S. Prasol, B. S. Perchuk, and M. T. Laub. A phosphorelay system controls stalk biogenesis during cell cycle progression in *Caulobacter crescentus*. *Mol Microbiol*, 59(2):386–401, Jan 2006.

[2] U. Consortium. The Universal Protein Resource (UniProt). *Nucleic Acids Res*, 35(Database issue):D193–D197, Jan 2007.

[3] D. Cook and L. Holder. *Mining Graph Data*. John Wiley and Sons, 2006.

[4] D. J. Cook, L. B. Holder, S. Su, R. Maglothin, and I. Jonyer. Structural mining of molecular biology data. 20(4):67–74, July–Aug. 2001.

[5] N. Deshpande, K. J. Addess, W. F. Bluhm, J. C. Merino-Ott, W. Townsend-Merino, Q. Zhang, C. Knezevich, L. Xie, L. Chen, Z. Feng, R. K. Green, J. L. Flippen-Anderson, J. Westbrook, H. M. Berman, and P. E. Bourne. The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema. *Nucleic Acids Res*, 33(Database issue):D233–D237, Jan 2005.

[6] C. Fabret, V. A. Feher, and J. A. Hoch. Two-component signal transduction in bacillus subtilis: how one organism sees its world. *J Bacteriol*, 181(7):1975–1983, Apr 1999.

[7] J. Kopp and T. Schwede. The swiss-model repository of annotated three-dimensional protein structure homology models. *Nucleic Acids Res*, 32(Database issue):D230–D234, Jan 2004.

[8] M. T. Laub, E. G. Biondi, and J. M. Skerker. Phosphotransfer profiling: systematic mapping of two-component signal transduction pathways and phosphorelays. *Methods Enzymol*, 423:531–548, 2007.

[9] G. A. Petsko and D. Ringe. *Protein Structure and Function*. New Science Press, 2004.

[10] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 13(11):2498–2504, Nov 2003.

[11] J. M. Skerker, B. S. Perchuk, A. Siryaporn, E. A. Lubin, O. Ashenberg, M. Goulian, and M. T. Laub. Rewiring the specificity of two-component signal transduction systems. *Cell*, 133(6):1043–1054, Jun 2008.

[12] A. M. Stock, V. L. Robinson, and P. N. Goudreau. Two-component signal transduction. *Annu Rev Biochem*, 69:183–215, 2000.

[13] K. Yamamoto, K. Hirao, T. Oshima, H. Aiba, R. Utsumi, and A. Ishihama. Functional characterization in vitro of all two-component signal transduction systems from escherichia coli. *J Biol Chem*, 280(2):1448–1456, Jan 2005.