MPC Corporation

Annual Progress Report: 2008 Formula Grant

Reporting Period

July 1, 2009 – June 30, 2010

Formula Grant Overview

The MPC Corporation received $115,727 in formula funds for the grant award period January 1, 2009, through June 30, 2010. Accomplishments for the reporting period are described below.

Research Project 1: Project Title and Purpose

Computational and Experimental Approaches for Predicting miRNAs and their Targets - Micro ribonucleic acids (miRNAs) are small RNAs that have important roles in regulating gene expression by binding to other RNAs. Deleterious polymorphisms in either the miRNA or its target will have significant effects on development and cell differentiation and are, therefore, associated with disease and congenital birth defects. While datasets of known miRNAs are accumulating at a steady rate, very little is known of the targets, and hence function of these miRNAs. This project seeks to develop computational prediction algorithms to predict miRNAs and their targets in an experimentally tractable organism, the sea urchin. High throughput experimental technologies will be used to test these predictions and refine the algorithm. This computational tool can then be used to predict miRNAs and targets in other organisms, including humans.

Duration of Project

1/1/2009 - 6/30/2010

Project Overview

The objective of this project is to develop computational algorithms to predict miRNAs and their targets. These predictions will be tested using high throughput experimental approaches in our sea urchin model system. Data obtained from the sea urchin will be used to refine the models and generate new predictions in an iterative process. These algorithms can eventually be used to predict miRNAs and their targets in any other organism for which there is a sequenced genome.

The specific aims of this project are (1) to develop and test algorithms to identify miRNAs from genome sequences and (2) to develop and test a software tool for predicting miRNA targets. We have already developed a computation tool to predict miRNAs from genome sequences. This approach was very successful in identifying already known miRNAs in several model organisms and in humans and, additionally, added to the database of potential miRNAs in these organisms. We have used these tools to predict sea urchin miRNAs; and, in the first part of the project, we

will verify these predictions and refine our algorithm. We will use quantitative polymerase chain reaction (PCR) to verify each predicted miRNA since we can test up to five miRNAs in as little as two hours. A strength of the sea urchin as a model system is that we can readily isolate all developmental stages and adult tissue types in order to perform a broad survey of miRNA expression. The pool of positive miRNA identified from the quantitative PCR assays will be confirmed by northern blot analyses. This process will generate a dataset of positive and negative predictions that can be used to refine our predictor. Eventually these algorithms will be used to predict miRNAs in other model organisms and in humans.

In the second aim we will test current predictors of miRNA targets and develop our own algorithms. This area of research is critical as miRNA targets are poorly known, but this information is essential for understanding how they function. We will explore Bayesian inference methods, supervised learning methods (e.g., support vector machines) or a combination of existing prediction methods with more empirical criteria (e.g., based on the thermodynamic properties of the miRNA-target interactions).

Principal Investigator

Veronica F. Hinman, PhD
Assistant Professor
Carnegie Mellon University
4400 5th Ave
Pittsburgh. PA 15213

Other Participating Researchers

Panayiotis Benos, PhD; Sabah Kadri, BSc - employed by University of Pittsburgh

Expected Research Outcomes and Benefits

The outcome of this project will be the development of computational tools to predict miRNAs and their targets. MiRNAs regulate gene expression through binding to other RNAs, and they are associated with a variety of genetic diseases. While very little is known about the function of miRNAs or their binding targets, such information is essential for understanding their roles in disease. These new computational tools will provide a set of predictions for miRNAs and their targets in humans and any other organisms that have a sequenced genome. These prediction tools will have been refined using data obtained from a deuterostome (sea urchin) organism.

Summary of Research Completed

We previously designed a computational prediction algorithm, HHMMiR, for discovery of miRNA from genome sequence. This algorithm was used, in addition to homology searches with the miRBase registry, to make genomic predictions for miRNAs in the sea urchin, as noted previously. To verify the predictions made and discover new miRNAs, if any, we used Illumina sequencing of small RNA libraries in sea urchin embryos. Mixed RNA populations from multiple embryonic stages (1, 2, and 3 days post-fertilization) in sea urchins were sent to the

Wistar Institute, Philadelphia, for small RNA library preparation and sequencing. We developed a computational pipeline (Fig. 1) to analyze the sequenced data and search for miRNAs. The sea urchin library returned ~13 million reads, which were then filtered and preprocessed to remove low quality 3' ends and linker sequences. The pre-processed reads were collapsed into "tags" based on sequence identity, and reads mapping to other non-coding RNAs (such as tRNAs, rRNAs, etc.) were removed. We found 44 conserved miRNAs, of which 11 were predicted previously with 100% homology search in the genome. The precursor sequences of the five most abundant miRNAs in the sample are shown in Fig. 2. miRDeep was used to make novel miRNA predictions in these samples. We found 13 potentially novel miRNAs using this method.

The ultimate aim of this research is to be able to make better predictions of miRNA-target interactions involved in regulatory gene networks. A number of methods have been developed for predicting targets of miRNAs including TargetScan, RNAhybrid, miRanda, and PITA. While miRanda and TargetScan use sequence complementarity and free energies of the RNA:RNA duplex, RNAhybrid makes predictions entirely based on thermodynamics, while PITA uses site accessibility as the main prediction criteria. All of these methods suffer from high false positive rates (Fig. 3). The underlying criterion for our improved prediction tool is whether the given untranslated region (UTR) sequence of a gene is under regulation by miRNA(s). We wanted to combine the prediction power of these weak classifiers into a single strong classifier. We used the supervised learning method, boosting, to achieve this goal.

For testing purposes, we use immunoprecipitation (IP) data from *C.elegans*. Two proteins from the GW182 family, AIN1 and AIN2, were used for the IP studies. We used 721 mRNA transcripts (558 had available UTR sequence available) enriched in both IP datasets, as the positive dataset. Similarly, we used the 51 miRNAs overlapping in both datasets as the miRNAs to be considered for this study. 652 mRNAs that were not enriched in both samples (474 of which had available UTR sequence available) were used as the negative dataset. On 10-fold cross validation, the sensitivity and false positive rate of prediction were 92.4% and 10.6%, with a standard deviation of 16% and 22%, respectively (Fig. 4). To eliminate noise from the datasets, we filtered the positive dataset by removing all UTRs that were not predicted as positive by any algorithm (67 were removed). UTRs were removed from the negative dataset if three or more algorithms predicted the UTR as positive (118 were removed), thereby drastically reducing the standard deviation of the accuracy statistics. The sensitivity and false positive rate were 90.4 ± 4.8% and 7.7 ± 5% respectively (Fig. 4).

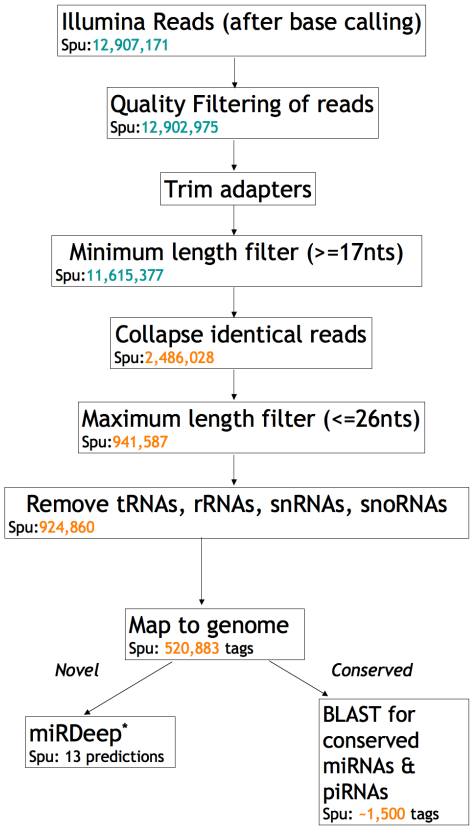

Figure 1: Computational pipeline to search for miRNAs in deep sequencing data.

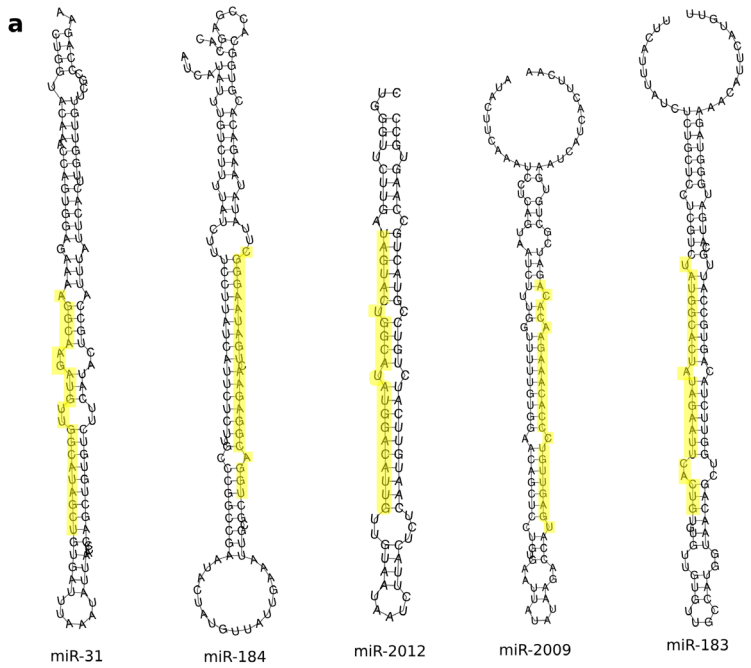

Figure 2: Stem loop structures of precursors of the most abundant sea urchin miRNAs (from the registry).

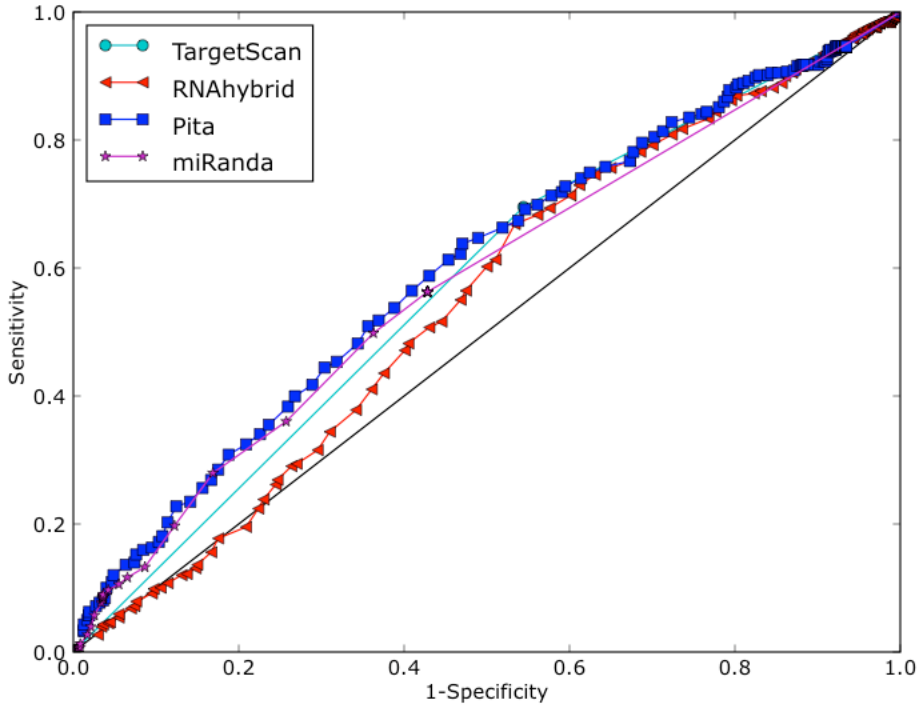

Figure 3: ROC curves for the four known miRNA target prediction algorithms using the *C.elegans* immunoprecipitation dataset.

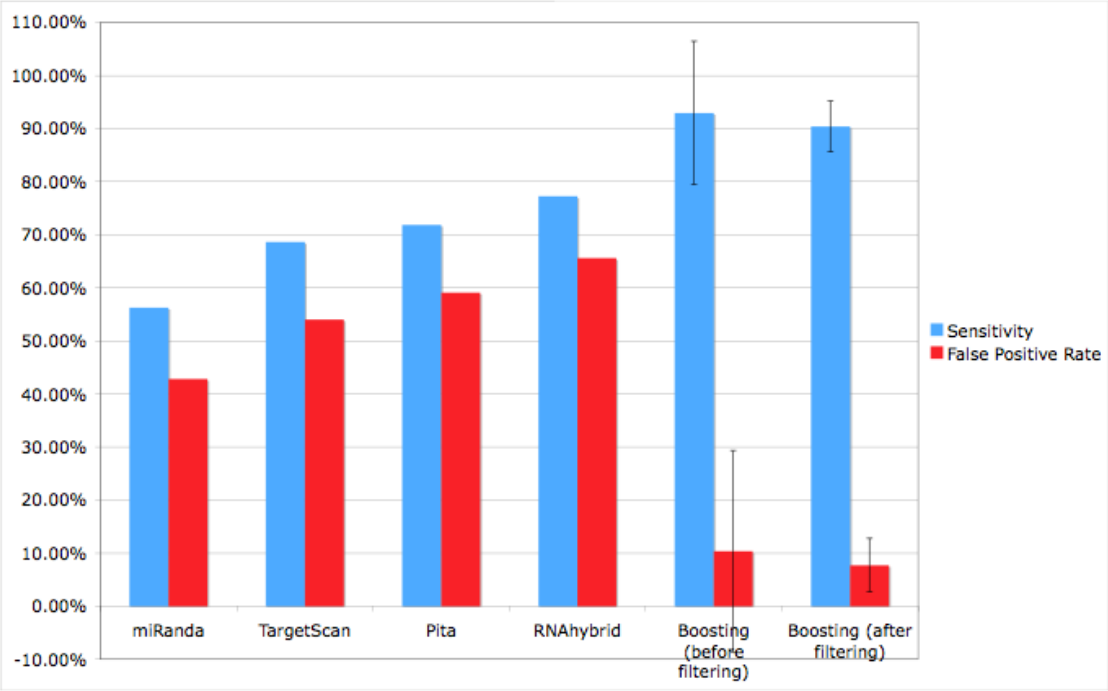

Figure 4: Comparison of performance between the four known prediction algorithms and our boosting tool, which combines the prediction power of all four methods.

Research Project 2: Project Title and Purpose

High-Throughput Estimation and Modeling of Protein Filament Distributions from Microscope Images - Proteins that form filamentous networks (e.g., microtubules, actin filaments) inside the cell play a crucial role in providing the cell its mechanical properties and, therefore, are related to many important biological processes, including wound healing and cancer metastases. Researchers who wish to characterize the effects of different drugs (or other alterations in the cell's environment) on these proteins are currently limited to performing qualitative studies because current imaging technology (including software) is insufficient for this task. Our goal in this project is to develop a software tool that can be used to extract the spatial distribution (structure) of these proteins directly from fluorescence microscopy images, thus facilitating future health sciences discoveries.

Duration of Project

1/1/2009 - 6/30/2010

Project Overview

The broad goal of our project is to develop software that will enable quantitative high-throughput studies of the distribution of filamentous protein networks (e.g., microtubules, actin filaments, and other related proteins) using currently available imaging technology. Filamentous distributions play a crucial role in providing the mechanical properties of cells and are also involved in important biological phenomena such as cancer metastases and wound healing. Thus, the availability of such a software tool would have a large impact in biology and health

sciences since it would allow researchers to perform quantitative high-throughput screening studies of the effects of drugs (or other experimental conditions) on these proteins.

Our specific aims in this study are to:

I) Develop methods for quantitative information extraction (number of filaments, their length distributions, etc.) directly from standard fluorescence microscopy images in a manner that is amenable to high-throughput studies and

II) Develop modeling and simulation approaches that will capture the cell-level essential characteristics for protein filament distributions based on information extracted from aim I

The principle behind the system is that very basic *a priori* knowledge can be used to formulate models of proteins from which artificial images are generated (according to initial estimates of the parameters of the model). Using input fluorescence image(s) of protein filaments of interest, the model parameters are iteratively modified until a specified similarity measure between the real input image(s) and the artificial ones is maximized. We will use an extensive validation strategy that includes experiments using both simulated data and data arising from real experiments.

Principal Investigator

Gustavo K. Rohde, PhD
Assistant Professor
HHC 122
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213

Other Participating Researchers

Robert F. Murphy, PhD; Aabid Shariff, BT - employed by Carnegie Mellon University

Expected Research Outcomes and Benefits

The methods developed in this study will enable quantitative measurement of physically meaningful parameters of filament distributions (i.e., filament number, length, and curvature) directly from standard fluorescence imaging. Currently, no technology exists for this task, and development of new software for this purpose would enable more definitive studies of the effects of drugs or other experimental conditions on these protein distributions. Together with auxiliary data (e.g., RNA or protein expression data), such measurement technology will also enable scientists to more easily determine which genes control certain parameters of protein distributions. Given that filamentous protein distributions have a crucial role in many health science related applications (e.g., cancer metastases), successful implementation of the methodology described below is likely to have a broad effect.

In addition to estimation of quantitative parameters as described above, we will also develop methods to create accurate network models to be used in large-scale, whole-cell models to capture and test predictions of cell behavior. In combination with existing methods that build object-based models, the methods developed will permit accurate information on the subcellular distributions of large numbers of proteins to be included in cell simulations. Altogether, these innovations are meant to facilitate the discovery of drugs and other experimental conditions controlling the behavior of these protein aggregations, leading to important discoveries in health sciences applications.

Summary of Research Completed

The active research in this project ended in the previous reporting period. During this final funding period, we applied for and received an NIH grant for $800,000 entitled "Automated High-Throughput Estimation and Modeling of Protein Network Distributions." We also submitted and published a paper in *Cytometry Part A*; the title of that publication is "A generative model of microtubule distributions, and indirect estimation of its parameters from fluorescence microscopy images," published in May 2010.

Research Project 3: Project Title and Purpose

Computational Modeling of miRNA Involvement in Pulmonary Gene Networks - MicroRNA (miRNA) genes are small, non-coding RNAs that act as gene expression regulators. They have been associated with various diseases (pulmonary fibrosis, heart failure, cancer) and important biological processes (cell differentiation, development). This project will develop computational methods to study miRNA's involvement in complex eukaryotic gene networks. New and existing computational methods will be evaluated on publicly available data. Experimental high-throughput data collected from lung epithelial cells—following stimulation by a factor associated with idiopathic pulmonary fibrosis (IPF)—will be analyzed as a proof-of-principle. The analysis will identify miRNAs involved in critical networks in cells related to IPF. Our results will also enhance the understanding of basic principles in cells' response to stimuli.

Duration of Project

1/1/2009 – 6/30/2010

Project Overview

It has been shown that miRNA genes are important regulators of gene expression and key players in many fundamental biological processes, like cell differentiation and development. More recently, their role in terminal diseases, like heart fibrosis, has been established. Finally, miRNAs are important markers for cancer. Despite their biological significance, the study of miRNAs' involvement in gene regulatory networks is limited. Part of the reason is that miRNA transcription is still poorly understood, thus making placement in gene networks difficult. We, and other laboratories, have recently analyzed high throughput data that provide insights for the characteristics of miRNA primary transcripts and promoter structures.

The research objective of this study is twofold: First, we will extend existing algorithms and develop new algorithms to model gene regulatory networks that involve miRNA genes. Currently, there are a number of algorithms that can model interactions between transcription factor (TF) genes with various degrees of success. We will extend these algorithms to accommodate miRNA genes, and we will evaluate them on publicly available datasets (microararrays and ChIP-chip). If needed, we will develop new algorithms based on the Bayesian network inference framework. In the second part of the project, we will use the best performing algorithm(s) to analyze high throughput data from lung epithelial cell lines related to IPF. The data from this study will provide a foundation for future follow-up studies to test the prediction efficiency of our algorithms in the involvement of miRNAs in IPF.

In summary, this project is expected to yield:
1. A robust method for modeling TF-miRNA interaction networks
2. A set of high-confidence predictions of miRNA-involving networks related to pulmonary fibrosis.

Principal Investigator

Panagiotis Benos, PhD
Associate Professor of Computational and Systems Biology
University of Pittsburgh, School of Medicine
3064 BST3
3501 Fifth Avenue
Pittsburgh, PA 15213

Other Participating Researchers

Tzu-Wei (Grace) Huang, BS- employed by University of Pittsburgh

Expected Research Outcomes and Benefits

Cells respond to extracellular stimuli by regulating the expression of their genes. An important part of cells' response is conducted through specific proteins, called transcription factors (TFs). miRNA genes are key regulators of many protein coding genes, including TFs; they are involved in many key biological processes and diseases. Unraveling the characteristics of TF-miRNA interactions at a systems level is important for understanding and treating these diseases.

This project focuses on the study of TF-miRNA interactions from a gene network perspective. The primary aim is to extend existing and develop new algorithms for modeling regulatory gene interactions, including miRNA genes. Evaluation of these algorithms will be performed on publicly available high-throughput datasets. By doing so, we will capitalize on our recent findings related to miRNA gene identification and transcription regulation. This part of the project will provide the computational framework for the analysis of gene networks and, as such, it will affect the study of many disease phenotypes.

In the second part of the project, we will perform a proof-of-principle analysis of existing data from epithelial lung cell lines related to pulmonary fibrosis. The outcome of this project portion will be a number of high-confidence predictions of miRNA-TF gene networks that are critical to the onset and progression of IPF, a devastating, terminal disease.

Summary of Research Completed

Summary: From June 2009 to March 2010, Grace Huang finalized Aim 1 (see below) by developing a novel method to assign weights on pairwise gene interactions, including microRNA genes, transcription factors, and other protein coding genes. This method combines sequence information (static) with gene expression data (dynamic). The novelty of the method lies in the way it combines different lines of evidence in a unified framework. It can also be expanded to include different types of information in the future, such as DNA methylation status, histone modifications, and other parameters. We are currently testing this method on datasets from idiopathic pulmonary fibrosis (IPF) and other lung diseases (Aim 2). We have also developed a web interface around this algorithm to make it accessible to researchers world-wide.
Below we present some of the results from the completion of Aim 1 and the ongoing research of Aim 2.

Results of Aim 1. Our graduate student, Grace Huang, developed a new method that combines sequence (static) data and gene expression (dynamic) data to infer relations among transcription factors (TFs), microRNA (miRNA), and other genes. Sequence data are used to score possible TF→gene interactions (here "gene" indicates miRNA gene, TF, or another protein coding gene) by examining various features of the corresponding gene's promoter. Currently, the features include the score for TF binding on the promoter as well as the accessibility of the promoter as evidenced by ChIP-chip or ChIP-seq experiments (H4 histone accetylation, CCCTC-binding factor [CTCF] binding, etc.). We plan to expand the feature set as more features become available. The miRNA→gene interactions (here "gene" indicates a TF or other protein coding gene) are inferred by combining predictions of five algorithms (including the popular TargetScan, PicTar, and PITA). Currently, the combined score is calculated empirically, but we plan to extend the method to combine the target prediction scores in a unified framework using a support vector machine (SVM) or another general machine learning algorithm.
Gene expression data for protein coding and miRNA genes are analyzed together, and pairwise expression correlations are identified between various genes (TFs, miRNAs, other genes) using standard statistical methods. The p-values for these associations are combined with the p-values from the static data to generate a confidence score. A new web server has been developed by our programmer (Harry Athanassiou; funded by other sources) to graphically present the output network of interactions. The pipeline of the algorithm is presented in Figure 1.

Results of Aim 2. In Aim 2 we proposed to use our algorithm to identify gene regulatory modules related to IPF. In particular, we were interested in identifying significant feed-forward loops (FFLs) (Figure 2). As a test case, we analyzed a dataset consisting of 17 IPF samples and seven control samples (matched mRNA + miRNA arrays), provided by our collaborator, Dr. Naftali Kaminski. Although our algorithm can only work with protein coding gene expression data, we wanted to test its efficiency on combined gene expression datasets. The mRNA and miRNA profiles were "matched," that is, they were collected from the same tissue from the same

patient. The preliminary analysis identified a number of potentially interesting FFLs (Figure 3). For example, we predict that miR-34a and miR-10b participate in an FFL by downregulating STAT3 and its target ASXL1.

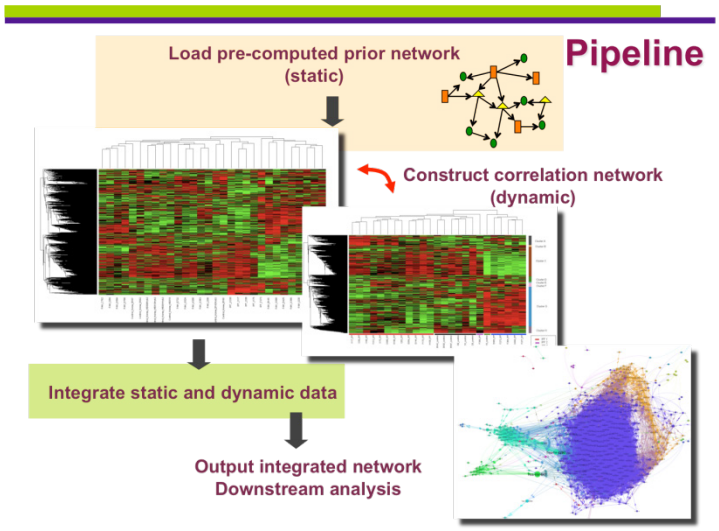

Figure 1. The pipeline of the algorithm that was developed through this project

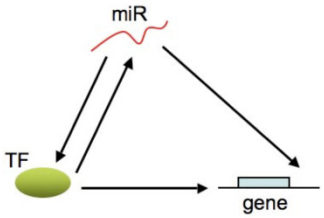

Figure 2. miRNA containing feed-forward-loops (FFLs)

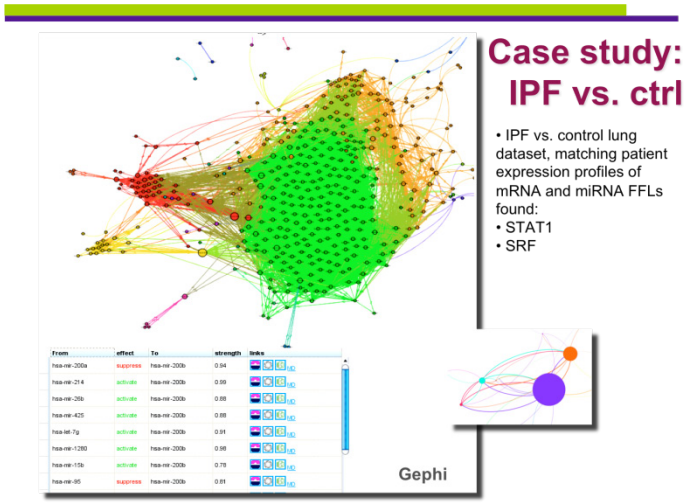

Figure 3. Test case of the algorithm: IPF vs. control.

Research Project 4: Project Title and Purpose

Modeling the Interactions of Dual Specificity Phosphatases with Inhibitors - Dual specificity phosphatases (DUSPs) have become important targets for drug discovery, with the establishment of their deregulatory role in cancer and Alzheimer's disease. However, deficiencies in scoring protein-ligand interactions and inadequate consideration of protein conformational flexibility have limited the utility of computational approaches. Our purpose is to develop efficient computational models and methods to study the interactions of DUSPs with small molecules toward developing selective and potent inhibitors. We will focus on the interactions of Cdc25B and MKP-1 with their inhibitors and substrates using elastic network models and various docking algorithms. Computational results will be analyzed in conjunction with experimental data from high throughput screening experiments conducted at the University of Pittsburgh Drug Discovery Institute.

Duration of Project

1/1/2009 - 6/30/2010

Project Overview

Protein tyrosine phosphatases (PTPs) are essential modulators of cellular processes. Dual-specificity phosphatases (DUSPs) constitute a subgroup of PTPs. Among members of DUSPs, mitogen-activated protein (MAP) kinase (MAPK) phosphatases (MKPs) and cell division cycle 25s (Cdc25s) currently serve as targets for drug discovery in many labs, including the University of Pittsburgh Drug Discovery Institute (UPDDI). In this project we will examine the interactions of Cdc25B and MKP-1 with their inhibitors and substrates using computational models and methods, known as protein-ligand docking simulations. Key mechanisms and interaction sites will be determined toward developing more selective and potent inhibitors.

Protein-ligand docking has become an integral part of drug discovery efforts, as the number of experimentally determined protein structures continues to increase. However, deficiencies in scoring protein-ligand interactions limit the use of docking simulations at various stages of the drug discovery process. This limitation essentially arises from inadequate consideration of the conformational flexibility, and associated entropy, which is not usually included in evaluating the binding free energies. Guidelines for addressing these deficiencies have been described in the literature. However, these have not been incorporated into computational models for protein-ligand docking simulations and virtual high-throughput screening applications.

Our goals are twofold: Our first aim is to work toward developing new tools for efficiently simulating protein-ligand interaction. Our second aim is to identify potential lead compounds for specific members of the therapeutically important family of DUSPs. Toward the first aim, we will perform, in addition to conventional docking simulations, newly developed methods based on elastic network models. These models were recently developed to rapidly assess the collective dynamics of proteins and the effects of ligand binding on their dynamics. A major advantage of these models is their ability to describe the conformational flexibility of proteins under equilibrium conditions using simple structural (topological) properties. We will use these models

for efficiently sampling protein conformations and bound ligand configurations. The generated ensemble of conformations will then be used in docking simulations. The distribution of conformations will provide a measure of their entropic characteristics.

Principal Investigator

Ivet Bahar, PhD
Professor and Chair of Computational Biology
University of Pittsburgh
3064 BST3
3501 Fifth Avenue
Pittsburgh, PA 15213

Other Participating Researchers

Ahmet Bakan, BS - employed by University of Pittsburgh

Expected Research Outcomes and Benefits

DUSPs are important, but poorly understood, cell signaling enzymes that remove phosphate groups from tyrosine and serine/threonine residues on their substrates. Deregulation of DUSPs has been implicated in cancer, obesity, diabetes, inflammation, and Alzheimer's disease. Due to their biological and biomedical significance, DUSPs have become the subject of drug discovery high-throughput screening (HTS) and focused compound library development efforts. However, progress in identifying selective and potent DUSP inhibitors has been limited by a lack of sufficient structural data on inhibitor-bound DUSPs. The shallow, nearly flat active sites, a shape required to achieve dual-specificity, has been another factor hampering the rational design of DUSP inhibitors. The present study aims at developing new computational models, methods, and tools that may potentially address these challenges and contribute toward the development of inhibitors against this important family of signaling enzymes.

In addition to designing DUSP potent inhibitors, this study aims at systematic development of a suite of docking programs that will efficiently take account of the intrinsic flexibility of the target proteins. Current docking simulations show that inadequate consideration of the conformational flexibility of the target protein and the entropic contribution to the selection and stabilization of particular bound conformers is a major drawback limiting the computer-aided design and selection of optimal inhibitors. This research will potentially alleviate this serious problem and contribute to development of more accurate protein-ligand modeling software for computer-aided drug design.

Summary of Research Completed

Specific Aim 1. To identify recurrent features and specific residues playing a key role in the inhibition of selected dual-specificity phosphatases (DUSPs), toward designing potent and selective inhibitors.

Summary We have further investigated the dynamics and druggability of DUSP6 using simulation methods and have identified a new potential site. Our results suggest that it is possible to identify allosteric inhibitors with better potency and selectivity properties. An RO1 grant application has been submitted to NIH to fund the proposed screening and testing research based on these results. This method had direct impact on our DUSP inhibitor discovery research.

All tyrosine phosphatases share a common mechanism of dephosphorylation, involving highly conserved residues: a catalytic cysteine temporarily accepts the phosphate group from the substrate, an arginine stabilizes the transition state, and a proton donated by an acidic residue in the highly flexible "general acid loop" (GAL) facilitates the hydrolysis of the intermediate phosphothioester. This mechanism is called the "catalytic triad," which in DUSP6 is formed by C293, R299, and D262 (Figure 1A). In substrate-inducible DUSPs such as DUSP6, D262 must be sufficiently close to the catalytic Cys for full activation (Rigas et al., 2001). This condition is presumably achieved by a suitable change in the GAL conformation. We recently identified a novel small molecule (BCI) in zebrafish screens that has no redox-active features and inhibits DUSP6 by a novel mechanism. Since the crystal structure for DUSP6 has been solved (PDB ID: 1mkp), we initially performed docking simulations to determine potential binding sites for BCI on MKP-3. These simulations indicated that BCI would preferentially bind the inactive form of DUSP6 at a site that would prevent the movement of the GAL to assume the functional closed state upon ERK binding (Figure 1A). The model also predicted that BCI might inhibit the inducible phosphatase activity of MKP-3 but would have no effect on its basal activity, which was confirmed experimentally. The discovery that the activity of DUSPs can be regulated through allosteric mechanisms has opened a new avenue to inhibitor discovery.

We further investigated the dynamics and druggability properties of the allosteric binding site by performing molecular dynamics (MD) simulations. Because there is no X-ray structure of DUSP6 in its activated state, we initially performed a comparative structural analysis of DUSP6 and DUSP10 as prototypical substrate-activatable and non-activatable DUSPs, respectively. In the low-activity form of DUSP6, (β3`) is part of a rigid beta-sheet that forces the GAL into an open position (Figure 1B, orange). In constitutively active DUSP10, β3` is deformed and the GAL is closed, positioning D262 close to the catalytic cavity (Figure 1B, green). Our MD simulations with a total time of 1 microsecond supported this model. In the high-activity form of DUSP6, the GAL was closed and β3` deformed (Figure 1C, blue), similar to DUSP10 (Figure 1B, green). Mechanistically, the model predicts that a rigid, structured beta-sheet containing β3-β3` strands joined by a short hairpin keeps the GAL in an open position through a proline residue (P241). In the high-activity conformation of DUSP6 (and in constitutively active DUSP10), β3` deforms, P241 moves away, and R299 and D262 reach their active state conformations.

Based on the above model and our results, we hypothesized that alternative sites that stabilize the conformation of the β3` strand could elicit the allosteric effect triggered by BCI binding. To identify such sites, we analyzed the druggability of the DUSP6 catalytic domain using a first principles-based method based on the work of Seco et al., 2009. In this method, the enrichment of isopropanol on particular sites on the protein surface is translated into a druggability index. We calculated the maximal achievable affinity for the BCI binding site to be 3 μM, which is consistent with the BCI's antiphosphatase activity (Molina et al., 2009). We also detected a second site with predicted nanomolar binding affinity located on the strand β 5 (Figure 2A).

DUSP6 makes extensive crystal contacts at this particular region (Figure 2B), in support of the propensity of this site to make intermolecular interactions. Four key residues on the neighboring DUSP6 molecule that make crystal contacts are shown in Figure 2C. Calculation of the buried surface area at this region yielded 139, 122, 97, and 75 $Å^2$ for W264, P334, M337, and L341, respectively. The respective interaction energies were calculated to be -3.0, -4.7, -3.1, and -2.5 kcal/mol, which cumulatively led to a 0.2 nM binding affinity. This value, comparable to our druggability predictions, provides independent support for targeting this particular pocket.

In summary, molecular modeling techniques have predicted a novel allosteric mechanism of inhibition for DUSP6, which was supported by experimental evidence. Also, the catalytic domains of DUSP6 have at least two allosteric binding sites for small molecules that could be exploited for small molecule inhibitor discovery.

Specific Aim 2. To assess the entropic contribution to binding free energy by sampling a representative ensemble of protein-ligand complex conformations using elastic network models.

Summary We previously showed the importance of protein entropic factors in binding of drug/inhibitor molecules. We investigated and improved a simulation-based method for binding site identification and druggability assessment.

Seco and Barril introduced an innovative, first principles-based technique to examine isopropanol accessibility of target protein surfaces in MD simulations, which is in principle applicable to different types of binding sites, including protein-protein interfaces, catalytic sites, or allosteric sites. Yet, the use of a hydrophobic/polar probe (isopropanol) precludes its applicability to binding sites containing charged residues such as the active site of protein tyrosine phosphatase 1B. The intermediate results were not easily interpretable, and the methodology contained redundant assumptions. We proposed: (1) a new methodology based on the same principles, but with simplified assumptions to facilitate the interpretation of the results, (2) a new fully automated algorithm that does not require user intervention, and (3) a new probe set (containing isopropanol, isobutane, acetamide, acetate, and isopropylamine) that expands the coverage of the methodology to binding sites with diverse physicochemical properties. We benchmarked the new methodology against a diverse set of binding sites from five proteins. We also investigated the effect of adopting different probe partial charge sets and probe compositions on the predictive ability of the method.

We developed tools for fully automated system preparation and druggability analysis. These tools will be released for public use by the end of 2010 with an accompanying journal article. The tools are expected to have an impact on guiding research in targeting proteins with feature-poor catalytic sites by helping identification of novel sites. This work was done in collaboration with scientists at GlaxoSmithKline.

With increasing structural data on metal-binding proteins, we have systematically analyzed the collective dynamics of a representative set of these proteins and tried to ascertain whether metal-binding sites have distinguishing features or positions that enable them to achieve their role in the activity of the proteins that bind them. We: (1) use elastic network models (Gaussian Network Model) to observe how metal-binding sites contribute to global dynamics, (2) analyze

the sites' solvent accessibility, and (3) use Markovian stochastic analysis to elucidate the sites' signal transduction properties and their involvement in allosteric communication.

Analyzing the global mobility profiles of 30 *apo-holo* pairs illustrated the insensitivity of global modes to the presence/absence of bound metal. Hence we chose 145 *holo* structures and examined whether metal-binding residues (Cysteines [C], Histidines [H], Glutamates [E] and Aspartates [D] or "CHED") occupy key positions (near hinges) and thus appear as minima in the global mode profile (Figure 3) The histogram in Panel A shows that metal-binding CHED residues, indicated by red bars, have a clear bias toward lower mobilities compared to all residues as well as their non-metal-binding counterparts. A closer analysis of the mobilities of each of the subsets showed that metal-coordinating histidine, glutamate, and aspartate (HED) prefer to be present in regions of lower mobility, while cysteine shows the opposite trend.

We analyzed solvent accessibility of the two subgroups of CHED residues. We saw a predisposition of metal-coordinating histidines, glutamates, and aspartates to occupy regions with limited, if any, exposure to solvent, in spite of being polar or charged. Thus, low solvent accessibility, along with the low mobility, can be used as "features" in facilitating the identification of potentially metal-coordinating HED residues in putative metal-binding proteins.

We introduced a Markov model for describing the stochastics of signal transmission in proteins modeled as networks of nodes and springs (Chennubhotla and Bahar 2007). The two main metrics of communication propensities are: (1) hitting time, $H(i,j)$, which represents an average number of steps for signals to be transmitted from residue j to residue i and (2) commute time, $C(i,j) = H(i,j) + H(j,i)$. These properties for a representative metal-binding protein have been shown in Figure 4 (A & B). Panel C shows the average hit time versus residue number, evaluated from the average of $H(i,j)$ over all starting points j. Red markers highlighting metal-binding residues occur at minima, pointing to efficient signal transmission. Panel D shows the average hit-time versus variance for each residue. The metal-binding residues identified by red markers show that they have fast and precise communication as shown by small standard deviation.

A comprehensive analysis of the average hit time versus standard deviation of residues participating in metal-binding in all 145 *holo* structures reiterated the above trend, where almost all metal-binding residues showed minimal hit times and minimal variance, pointing toward their fast and precise signal transmission properties. Thus, our results show that metal-binding sites might have an inherent preference to undergo minimal fluctuations in their positions, which might enable them to elicit cooperative responses; occupy central/buried positions despite being polar or charged; and possess unique signal transduction properties, which suggest that these sites may be selected to achieve optimum allosteric communication. These properties give us useful insights into the basic principles for designing metal-binding sites.

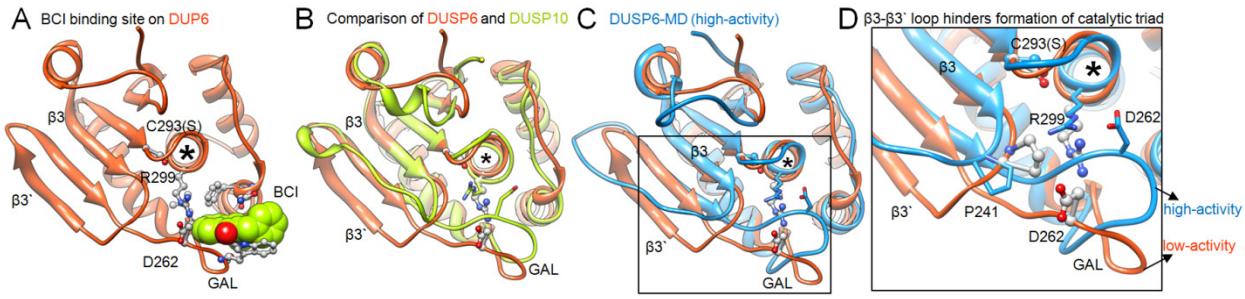

Figure 1. Molecular dynamics simulations of the BCI allosteric site. A. BCI (green) prevents closure of the general acid loop. Upon BCI binding, the GAL containing D262 is held in an open position, locking the enzyme in its low-activity form. The catalytic cavity is marked with a star. B. MKP-3/DUSP6 in its low activity form (orange) compared to constitutively active MKP-5/DUSP10 (green). In MKP-5/DUSP10 the GAL is closed and appears to be coupled to deformation of β3`. C. Comparison of molecular dynamics-derived DUSP6 in its high-activity form (DUSP6-MD, blue) shows β3`deformation similar to MKP-5/DUSP10. D. β3-β3` hairpin loop hinders formation of the catalytic triad. A rigid, structured beta-sheet containing β3-β3` strands joined by a short hairpin keep the GAL in an open position through a proline residue (P241). In the high-activity conformation of DUSP6 (and in constitutively active DUSP10), β3` deforms, P241 moves away, and R299 and D262 reach their active state conformations. This structural transition is presumed to be favored upon ERK binding. Data are from a 30 ns simulation; a snapshot from 20th ns is shown.

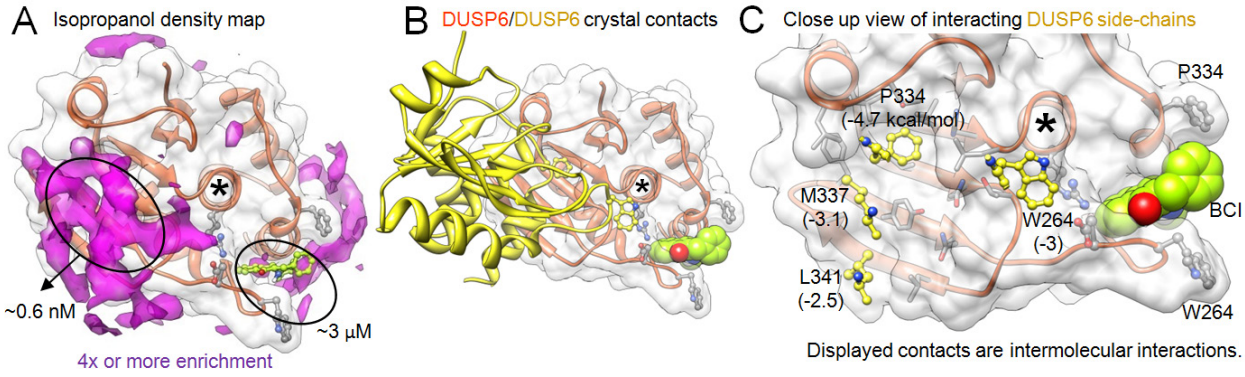

Figure 2. A novel allosteric small molecule binding site on DUSP6. A. Results from druggability simulations. Purple clouds correspond to regions enriched in isopropanol concentration, corresponding to energetically favorable sites for drug-like molecule binding. The BCI site was calculated as a μM site. A more distant site with nM druggability index was also observed. B. DUSP6 makes extensive crystal contacts at the distant site. Analysis of DUSP6/DUSP6 interactions (by reconstruction of the unit cell using crystal parameters) showed favorable interactions between monomers (second DUSP6 shown in yellow), suggesting the novel site can accommodate small molecules. C. Key interacting side-chains on neighboring DUSP6 monomers (shown in yellow) occupy binding pockets at the distant site and constitute anchors for biased docking and virtual screening studies for identification of new chemotypes. Interaction energies predicted by ANCHOR server are shown in parentheses.

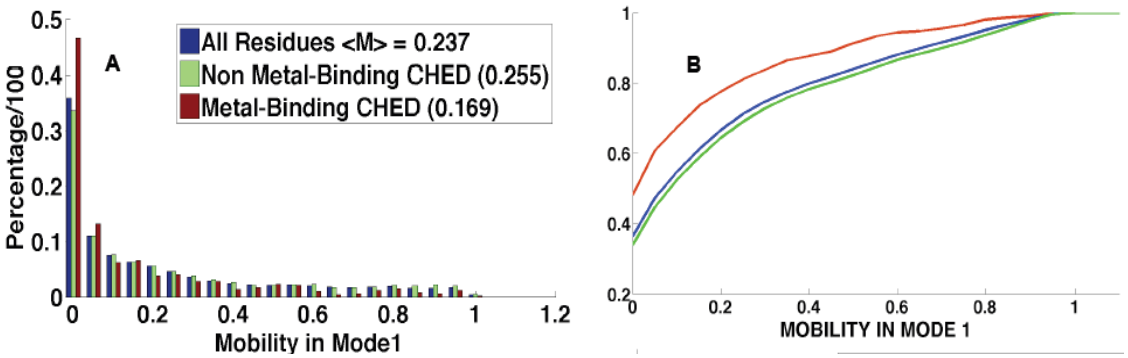

Figure 3 shows the mobility in Global Mode 1. A: displays the histograms of mobilities for the three subsets. For comparative purposes, the mobilities were normalized in the range [0, 1] for each protein. B: shows the cumulative distribution for three subsets.

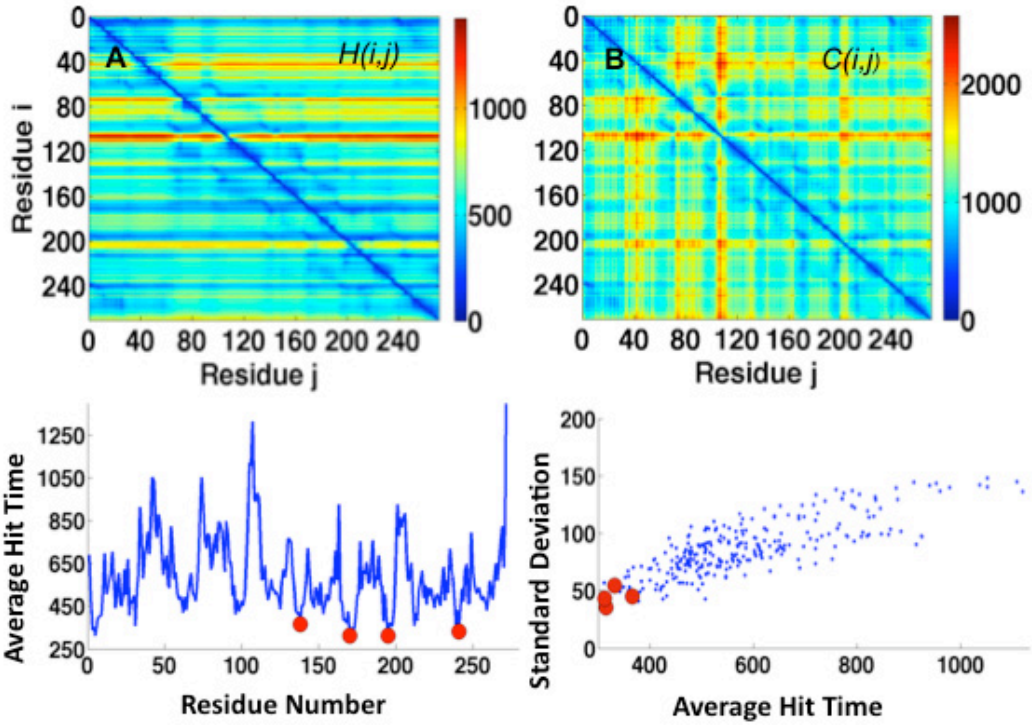

Figure 4. Signal propagation properties illustrated for a Zn2+ binding protein (1I6N). A: Hitting time B: Commute Time C: Residue number versus average hitting time, with red markers showing metal-binding sites that show minimal hit times. D: Average hit times versus standard deviation. Metal binding sites made prominent with red markers to highlight minimal variance.