2013

Discovering meaning from biological sequences: focus on predicting misannotated proteins, binding patterns, and G4-quadruplex secondary

Carson Michael Andorf
Iowa State University

Discovering meaning from biological sequences: focus on predicting misannotated proteins, binding patterns, and G4-quadruplex secondary structures

by

Carson Michael Andorf

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Computer Science

Program of Study Committee:
Vasant Honavar, Co-major Professor
Drena Dobbs, Co-major Professor
David Fernandez-Baca
Robert Jernigan
Taner Sen
Guang Song

Iowa State University
Ames, Iowa
2013

DEDICATION

TO MY FAMILY,
FOR THEIR OVERWHELMING SUPPORT

TABLE OF CONTENTS

Page

Page

ACKNOWLEDGEMENTS

I would like to thank my major professor Dr. Vasant Honavar. He has provided me with constant support, encouragement, mentoring, and inspiration. I have benefited greatly from his vast knowledge of machine-learning and bioinformatics. His help and insight has been a constant factor in my life and my research.

I would like to thank my co-major professor Dr. Drena Dobbs. She has provided me with support, mentoring, and guidance throughout my years at Iowa State. She was essential to my successful transition into bioinformatics. The foundation of my biological knowledge came from her teachings and from our conversations.

My sincere appreciation goes to Dr. Robert Jernigan, Dr. Taner Sen, and Dr. David Fernandez-Baca, and Dr. Guang Song. I would like to thank them for serving on my program of study committee, for reading my thesis and making constructive suggestions.

A special thanks goes to Dr. Gavin Naylor who allowed me to do a rotation in his lab. Also, I would like to thank Dr. Dan Voytas, Kathy Wiederin, and Trish Stauble. They were vital in getting me support through the IGERT program and keeping me involved in the Bioinformatics Program. I would like to thank Dr. Volker Brendel for his teachings and his mentoring outside of my academic work.

Many thanks go to past and present members in the Artificial Intelligence Research Lab and the Dobbs Lab for rewarding research discussions, and helpful comments on my thesis, research, and numerous manuscripts. I have developed many lifetime friendships within the laboratories, all of which I will cherish the rest of my life. I want to thank a close collaborator, Adrian Silvescu, whose knowledge and insight has been invaluable. I also want to thank Changhui Yan, Jun Zhang, Doina Caragea, Jae-Hyung Lee, Michael Terribilini, Jeff Sander, Pete Zaback, Dae-Ki Kang, Feihong Wu, Jyotishman Pathak, Cornelia Caragea, Jie Bao, Yasser El-

Manzalawy, Oksana Yakhnenko, Rasna Walia , Li Xue,and Dake Wang for their contributions to my research. I would like to express my special thanks to Jaime Reinoso-Castillo and Hector Leiva, whom I started my years at Iowa State with; their friendship and time spent studying was key for my early academic success.

Additionally, I would like to thank Dr. Carolyn Lawrence and the MaizeGDB team (Darwin Campbell, Dr. Lisa Harper, Dr. Jack Gardiner, Dr. Mary Schaeffer, Dr. Taner Sen, and Ethy Cannon) who were supportive and showed patience as I finished my Ph.D.

Thank you Carol Reese and other members of my family for the periodic reminders of the importance of my degree and giving me a goal to make you a proud aunt.

Finally, a very special thanks is due to my wife, Destri Andorf, for her love and support throughout my years at Iowa State and helping me bring my ideas to visual form. Also, to my daughters Carissa and Caelyn who have brought inspiration into my life. I hope the work done at Iowa State and in academia will provide a better future for them and give a clear path towards academics and higher learning. I want to give thanks to my parents (Michael and Beverly), my sister (Erica), and my two nephews (Blayke and Marcus) for their unconditional love, overwhelming support, and constant encouragement.

ABSTRACT

Proteins are the principal catalytic agents, structural elements, signal transmitters, transporters, and molecular machines in cells. Experimental determination of protein function is expensive in time and resources compared to computational methods. Hence, assigning proteins function, predicting protein binding patterns, and understanding protein regulation are important problems in functional genomics and key challenges in bioinformatics. This dissertation comprises of three studies. In the first two papers, we apply machine-learning methods to (1) identify misannotated sequences and (2) predict the binding patterns of proteins. The third paper is (3) a genome-wide analysis of G4-quadruplex sequences in the maize genome. The first two papers are based on two-stage classification methods. The first stage uses machine-learning approaches that combine composition-based and sequence-based features. We use either a decision trees (HDTree) or support vector machines (SVM) as second-stage classifiers and show that classification performance reaches or outperforms more computationally expensive approaches. For study (1) our method identified potential misannotated sequences within a well-characterized set of proteins in a popular bioinformatics database. We identified misannotated proteins and show the proteins have contradicting AmiGO and UniProt annotations. For study (2), we developed a three-phase approach: Phase I classifies whether a protein binds with another protein. Phase II determines whether a protein-binding protein is a hub. Phase III classifies hub proteins based on the number of binding sites and the number of concurrent binding partners. For study (3), we carried out a computational genome-wide screen to identify non-telomeric G4-quadruplex (G4Q)

elements in maize to explore their potential role in gene regulation for flowering plants. Analysis of G4Q-containing genes uncovered a striking tendency for their enrichment in genes of networks and pathways associated with electron transport, sugar degradation, and hypoxia responsiveness. The maize G4Q elements may play a previously unrecognized role in coordinating global regulation of gene expression in response to hypoxia to control carbohydrate metabolism for anaerobic metabolism. We demonstrated that our three studies have the ability to predict and provide new insights in classifying misannotated proteins, understanding protein binding patterns, and indentifying a potentially new model for gene regulation.

CHAPTER I

GENERAL INTRODUCTION

High throughput genome sequencing projects are producing larger and larger amounts of raw sequence data. The need to identify, label, and classify these sequences' biological functions and roles in a fast and efficient way has emerged, with increasing emphasis on the creation of high-quality, accurate functional predictions. Proteins are the principal catalytic agents, structural elements, signal transmitters, transporters, and molecular machines in cells. For this dissertation, three ways are explored to understand the function of proteins. First, by examining further the central dogma of biology, the relationship between sequence, structure, and function is considered. Second, our evolving understanding of a proteins' interaction partners is shown to help with functional annotation of proteins [1]. Protein-protein interactions are, therefore, critical to elucidating the role played by individual proteins in important biological pathways. Third, the transcription and translational control of genes that encode proteins can define whether and how proteins are expressed. Chromosomes carry genetic information and numerous DNA-sequence elements regulate the expression of that information through control of chromatin structure and transcription. An example of one of these elements is the G4-quadruplex, a non-B-form DNA secondary structure (DNA conformations other than the canonical right-handed double helical structure) that has received recent attention for its role in gene regulation.

Experimental determination of protein function, interaction relationships, and regulation lags significantly behind the rate of growth of sequence databases. This situation is likely to continue for the foreseeable future. Hence, assigning proteins a biological

knowledge label in the absence of functional information, i.e., from sequence alone, alone remains one of the most challenging problems in functional genomics [1]. It is also desirable to make this process quick, accurate, versatile, and updateable. For this reason, much work on producing such algorithms has been attempted. This dissertation takes a computational approach to addressing three bioinformatics problems: (1) identifying proteins that are likely to be functionally misannotated proteins; (2) labeling binding patterns of hub proteins; and (3) exploring the roles of G4-quadruplexes in regulating gene expression.

Data Mining Approaches for Discovery of Protein Sequence-Structure-Function Relationships

Recent advances in both the fields of machine-learning [2] and data mining [3] offer a promising approach to data-driven discovery of complex relationships from large data sets in general, and discovery of macro-molecular sequence, structure, evolution, expression, interaction, and function in particular [4, 5]. To summarize, the data mining approach uses a representative training dataset that encodes information about proteins with known functions to build a classifier for assigning proteins to one of the functional families represented in the training set. The resulting classifier can be used to assign novel protein sequences to one of the protein families represented in the training set after it has been evaluated using an independent test set. Several recent studies have explored data mining approaches for the automated construction of classifiers for assigning proteins to functional families. We describe recent work, the results, and the methods used for prediction (See **Tables 1 - 4**). Here we focus on machine-learning algorithms that are often used and have shown success in bioinformatics. These algorithms include the following:

Decision trees [6], Naïve Bayes, Markov models, and Support vector machines (SVM) [7]. These methods are described in greater details in the methods sections in each chapter.

Overview of Prediction of Misannotated Proteins

As more genomic sequences become available, methods for functional annotation of genes become increasingly important. Because experimental determination of protein structure and function is expensive and time-consuming, there is an increasing reliance on automated approaches for assignment of functional categories (e.g., Gene Ontology (GO) [8-10]) to protein sequences. An advantage of such automated methods is that they can be used to annotate hundreds to thousands of proteins in a matter of minutes, which makes their use especially attractive in large-scale genome-wide annotation efforts.

Most automated approaches to protein function annotation rely on transfer of annotations from previously annotated proteins, based on sequence similarity. Such annotations are susceptible to several sources of error, including errors in the original annotations from which new annotations are inferred, errors in the algorithms, bugs in the programs or scripts used to process the data, clerical errors on the part of human curators, etc. The effect of such errors can be magnified because they can propagate from one set of annotated sequences to another through widespread use of automated techniques for genome-wide functional annotation of proteins. Once introduced, such errors can go undetected. Because of the increasing reliance of biologists on reliable annotations for formulation of hypotheses, design of experiments, and interpretation of results, incorrect annotations can lead to wasted effort and erroneous conclusions. Computational approaches to checking automatically inferred annotations against independent sources of evidence and detecting potential annotation errors offer a potential solution to this problem

[11-21]. See **Table 5** for details about these methods. Currently, there are few high-throughput methods that exist or that have been widely applied to this problem. This is partly due to the possibility that the same pitfalls for annotating sequences may exist for checking annotations. Hence, there is a large and growing need for methods to predict misannotations and methods for dealing with these misannotations.

Overview of Protein-Protein Interaction Networks and the Relationship of Binding Patterns in Hub Proteins

Protein-protein interaction networks are usually represented as graphs with nodes corresponding to the proteins and edges denoting pairwise interactions. This is a simplistic representation that is not rich enough to encode types of interactions including interactions that involve more than two proteins. A single target protein can interact with a large number of partners: sometimes the interactions are mutually exclusive because of competition for the interaction sites on the target protein, other times interactions may be simultaneously possible, and even mutually cooperative [22, 23]. Distinguishing between these types of interactions is essential for uncovering the physical basis of interactions of a protein with other proteins, engineering the protein surfaces to manipulate synthetic pathways, or for designing drugs that bind specific targets [24-26]. Of particular interest are hub proteins, proteins that interact with a large number of other proteins in an interaction network. Hub proteins have been reported to play essential roles in cellular control and tend to be highly conserved across species [27]. Mutations in hub proteins can potentially disrupt interactions with its many interaction partners, which can turn out to be lethal [28-30]. Hence, it is especially important to understand the physical and structural basis of interactions of hub proteins with their partners. Hub proteins can be distinguished

based on structural and kinetics properties. Based on structural information, Kim et al. [31] defined two types of hub protein structures: singlish interface hubs (SIH) and multiple-interface hubs (MIH). SIH interacts with multiple partners at one or two binding sites and MIH interacts with multiple interaction partners through more than two binding sites. Recent studies [30-39] have explored the roles of SIH and MIH in protein-protein interactions and hence protein function. Hub proteins can also be classified based on the kinetic mode of interaction. Han et al. [40] described a classification model based on a bimodal distribution of co-expression of hub proteins with their interaction partners. Date hubs tend to display expression levels that have low correlation with those of their interaction partners (and tend to bind different partners at different time points or locations). Conversely, party hubs tend to exhibit expression levels that have high degree of correlation with those of their interaction partners (and tend to interact simultaneously with the partners). The analysis of party and date hubs provides additional insights into the structure of the underlying proteome and interactome.

Experimental characterization of hub proteins in terms of their structural and kinetic characteristics requires knowledge of the structures of complexes formed by such proteins in interaction with their binding partners [41, 42]. Because of the prohibitive cost and effort needed to determine the structures of complexes formed by hub proteins with their binding partners and the interfaces that mediate such interactions, there is an urgent need for reliable methods that predict the structural and kinetic characteristics of hubs from sequence information alone. In particular, there is a growing interest in purely sequence-based computational methods for discriminating between simultaneously possible versus

mutually exclusive interactions [36, 40, 43, 44] and predicting the number of binding sites available on the surface of a protein.

Our current understanding of protein-protein interaction networks is quite limited for a variety of reasons including the high rate of false positives (predicted interactions between proteins) associated with high throughput experiments, the low coverage of solved co-crystal structures in the Protein Data Bank (PDB), and the difficulty of obtaining reliable negative evidence that a protein does not interact with one or more other proteins. Hence, there is a growing interest in computational tools that provide useful insights into various structural aspects of protein interactions from protein sequence alone. **Table 6 - 8** shows recent work on SIH/MIH, date/party hubs, and distinguishing hub proteins from non-hub proteins [45-47].

Genome-Wide Analysis of G4-Quaduplex Motifs

The G4Qs and the ability to form them are hallmarks of most eukaryotic telomere repeat sequences, but they also occur at sites throughout the chromosome, where their role in gene regulation has received major attention in the last decade [48-50]. These non-telomeric G4Qs have been characterized in *E coli*, human, yeast, and other eukaryotes (See **Table 8**) [51-57]. These G4Q elements likely function as reversible structural motifs that have evolved a variety of functions. In particular, recognition of their role in cell cycle control and cancer-associated genes and pathways has fueled a rapid and recent expansion of research on G4Qs and is reflected in the numerous reviews in the last few years on their occurrence and functions (See **Table 9**) [58-64]. Significant discoveries of G4Q elements have come from bioinformatics analyses and data mining of whole genomes (See **Table 10**) [65-71]. G4Q motifs are demonstrated to occur frequently in the human genome [54,

70, 71], appearing in more than 40% of human gene promoters [72]. Together with their occurrence in human, mouse, and rat promoters [73] led to the hypothesis of a specific and conserved role for G4Qs in transcriptional regulation. A well characterized example is the c-MYC proto-oncogene, in which a G4Q motif within the promoter was shown to affect transcription *in vitro* [74] and was proposed to function as a molecular switch. Notably absent from these analyses are representatives of plant kingdom, despite the role of plant genetics in basic discoveries of nucleic acid functions, such as telomere functions, basic mechanisms of crossover and inheritance, and the discovery of transposable elements. To date, G4Q investigations in plants have been limited to one genome-wide analysis of Arabidopsis with a brief comparison to other plant species (see **Table 11**) [75, 76]. Though this study gave a broad overview of the frequency of G4Qs in some plants species, the role of G4Qs in plants remains unknown. Although plants do not succumb to cancer, they do share much of the genome maintenance and basic eukaryotic biology with species from other kingdoms. We propose that maize (*Zea mays* ssp. *mays)* would be a good model organism to explore the roles of G4Qs in plants, because maize is a well-described plant model species for genetics and, more recently, genomics research [77].

A Survey of Current Methods

In our literature review section we show previous work that relates to the three individual studies (Chapter 2, 3, and 4). Most of the literature fits the following criteria: a **recent** study showing new or state-of-the-art results, a **classic** study that was either the first of its kind or the original source of a result, a **review** paper that summarizes a field of study, or a **representative** paper that provides a good or representative example of similar papers in that field. The literature review is broken up in the following sections:

1. Predicting functionally misannotated proteins

    A. Approaches for functional annotation of proteins

        i. Similarity-based of sequence (**Table 1**)

        ii. Sequence motif-based (**Table 2**)

        iii. Composition and reduced alphabet (**Table 3**)

        iv. Protein interaction networks, protein structure, literature and text mining, gene expression, or the integration of other relevant data (**Table 4**)

    B. Misannotation of proteins (**Table 5**)

2. Labelling binding patterns of hub proteins

    A. Hubs and non-hubs in PPI networks (**Table 6**)

    B. Structural and kinetic annotation of protein hubs (**Table 7**)

3. Exploring the roles of G4-quadruplexes in the maize genome

    A. G4-quadruplex: Genome-wide surveys (**Table 8**)

    B. G4-quadruplex: Review papers (**Table 9**)

    C. G4-quadruplex: Bioinformatics (**Table 10**)

    D. G4-quadruplex: Plants (**Table 11**)

The literature review for each subject is organized into tables. Computational papers have four columns: the authors and year(s), a description of the paper, a summary of the results, and the method or type of method used. Review papers or biology-centered papers have three columns: the authors and year(s), a description of the approach and results, and keywords that best summarized key elements in the paper.

Table 1: Similarity based approaches functional annotation of proteins.

Author Year	Description	Results	Methods
Lopez and Pazos 2013 [78, 79]	Developed a method for predicting function to individual domains based on position specific scoring matrices (PSSM) build on sequences assigned to gene ontology terms. In addition this method predicts structural folds and identifies individual residues that may play functionally important roles.	On a set of 1017 protein chains, they presented ROC curves showing their method outperformed a PSI-BLAST based method. They also developed COPRED, a web-server that predicts function, functional sites, and folds at the domain level.	PSSM Gene Ontology Domain
Sleator et al. 2010 [80]	Reviewed various recent studies in protein function prediction from sequence.	This is a review paper so no results were reported, but they did address the strength and weaknesses of each approach, and what challenges remain in this area.	Automated Function Prediction Review
Xiong et al. 2011 [81]	Developed a multiple step method to assign function to sequence. The first step is to identify motifs from aligned sequences. The second step is to score the motifs based on the frequency and quality of the motifs. The final step is using the scored motifs as inputs to a random forest and SVM classifiers.	They tested their method with nucleosome occupancy and protein solubility datasets. They reported AUC values of 0.80 and 0.63 respectively for each dataset.	SVM Random Forest
Meng et al. 2009 [82]	Approach combined similarity-based with literature-based annotation for assigning GO labels to protein sequences. For the similarity-based annotation they used reciprocal best hits between rice and multiple other species. They validated the results literature with reviewed data, domains, or wet-lab data.	Using their sequence-based approach they functional annotated 6,286 proteins (including 2,870 hypothetical proteins). The literature-based approach annotated 2,810 proteins (1,673 received new GO terms). Overall 7,412 proteins were annotated.	Reciprocal BLAST Literature-mining
Finn et al. 2006 [83, 84] Sonnhammer et al. 1997 [85]	Developed PFam (Protein Family) database. Each protein database is represented by a hidden Markov model (HMM) based on the sequence alignment of the proteins in that family.	PFam is a widely used both in classifying proteins as well as a location to find well-characterized protein domain families. Their HMMR method to construct PFam families is also widely used.	PFam HMM
Vinayagam et al. 2006 [86]	Developed GOPET a Gene Ontology term Prediction and Evaluation Tool that uses an SVM based on a sequence kernel.	This web service gives experimental researchers as well as the bioinformatics community a valuable sequence annotation device. Additionally, GOPET also provides less significant annotation data which may serve as an extended discovery platform for the user.	SVM

- Table 1 continued -

Jones et al. 2005 [87]	Developed a simple method that given an unannotated sequence they use BLAST, to find similar sequences that have already been assigned GO terms by curators. They make their predictions based on the best five matching sequences	The precision and recall of estimates increases rapidly as the amount of distance permitted between a predicted term and a correct term assignment increases.	BLAST
Lanckriet et al. 2004 [88]	Used a SVM to predict ribosomal and membrane proteins. As a kernel to the SVM they used a wide variety of genome-wide measurements including BLAST, PFam HMM, protein interaction, gene expression, multiple sequence alignment, hydropathy profile, and random numbers.	Their combined method for using multiple sources of genome-wide information performed better than using any individual method alone. On the ribosomal proteins their method was able to predict with very high ROC scores (>.999) The membrane proteins predictions had a .922 ROC score. Showing this is a very effective method.	SVM, HMM, BLAST
Murvai et al. 2001 [89]	Developed an artificial neural network (ANN) to recognize domains in protein sequences. A query sequence is first compared to a database using BLAST, the output is encoded an input to the artificial neural network.	On a wide variety of domain data and functional data they were able to get correlation coefficients ranging from .85 to .98.	ANN, BLAST
Altschul et al. 1990,1997 [90, 91]	Developed BLAST (Basic Local Alignment and Search Tool), a homology based search tool to that finds local regions of similarity.	BLAST is the most widely use bioinformatics tool. BLAST is used for many tasks including inferring protein function based on BLAST hits.	BLAST
Gribskov et al. 1987 [92]	Developed a profile analysis method to detecting relationships between distantly related proteins by using position-specific scoring tables.	This has become a common technique in building relationships among aligned sequences.	Position-specific scoring matrix

Table 2: Motif based approaches for functional annotation of proteins.

Author Year	Description	Results	Methods
Saidi et al. 2010 [93]	Tested a wide variety of machine-learning algorithms based on the following features: motifs, n-grams, amino acid composition, and functional domains.	They reported accuracies on 5 protein datasets. Their best results reached 87.7% by using a fusion network of all their methods combined.	Decision trees SVM Naive Bayes ANN
Sarc et al. 2010 [94]	Created a method (GOPred) that combines the results from BLAST nearest neighbors, subsequence profile maps, and peptide statistics (from PEPSTATS-SVM) with several different voting schemes to predict GO functional annotations.	For a given protein, GOPred provides predicted GO terms along with probability scores for each method. It also shows a weighted mean score.	BLAST Profile maps SVM
Pagni et al. 2007, 2004 [95, 96]	Developed a web-server (MyHits) to functionally annotate proteins based on domains and motifs.	Their site offer a variety of tools including PSI-BLAST, ClustalW, T-Coffee, Jalview; a set of sequences; a set of motifs; and a set of matches between the sequences and motifs.	PSI-BLAST Pfsearch
Kunik et al. 2005 [97]	Reviewed current motifs along with a new unsupervised method (MEX) for finding motifs in sequences. They combined the motif information with a SVM to produce a functional classifier.	The classification results from MEX performed better than SVMProt and a SVM using the Smith-Waterman distance matrix.	Motif finding SVM
Wang et al. 2003, 2002 [98, 99]	Developed a decision tree approached based on the presence or absence of PROSITE motifs.	Their method outperformed a query based approached based on individual PROSITE motifs on a set of peptidase proteins.	Decision Tree PROSITE
Benhur and Brutlag 2003 [100] Huang et al, 2001 [101]	Developed e-motif and e-matrix, methods for detecting remote homology that is based on the presence of discrete sequence motifs. The motif content of a pair of sequences is used to define a similarity that is used as a kernel for a Support Vector Machine (SVM) classifier.	They tested the methods on two remote homology detection tasks: prediction of a previously unseen SCOP family and prediction of an enzyme class given other enzymes that have a similar function on other substrates. They find their methods perform significantly better than other sequence homology based tools.	SVM
Attwood et al. 2000 [102]	Developed the PRINTS database, a collection of protein family fingerprints base on sequence alignment.	The group of PRINTS motifs together is diagnostically more potent than single motifs by virtue of the biological context afforded by matching motif neighbors.	PRINTS
Mulder et al. 2005, 2003, 2002 [103-105]	Developed InterPro, a new integrated documentation resource for protein families, domains and functional sites, combining the information of the PROSITE, PRINTS, PFam and ProDom database projects.	They have merged and gathered locally the annotations from PRINTS, PROSITE and PFam.	PRINTS PROSITE PFam InterPro

- Table 2 continued -

Henikoff et al. 1999 [106]	Developed the Blocks Database, which contains ungapped multiple alignments for families documented in PROSITE. Blocks can be searched to classify new sequences.	Blocks is a commonly used database to search for protein function.	Blocks
Falquet et al. 2002 [107] Bairoch and Bucher 1994 [108] Hofmann et al. 1999 [109]	Developed the PROSITE database. PROSITE is a compilation of sites and patterns found in protein sequences; it can be used as a method of determining the function of uncharacterized proteins translated from genomic or cDNA sequences.	PROSITE is a common site to locate possible protein motifs that can be used to determine a proteins possible function.	PROSITE

Table 3: Composition, pseudo-composition, and reduced alphabet approaches for functional annotation of proteins.

Author Year	Description	Results	Methods
Feng et al. 2013 [110]	Developed a method and web server, iHSP-PseRAAAC, to predict heat shock protein families. The method used a SVM trained and tested on the composition of reduced alphabet representations of proteins.	They were able to correctly classify six types of heat shock proteins at a rate of 87%. This was done on a dataset with less than 40 % sequence identity.	SVM Composition Reduced alphabets
Huang et al. 2013, 2013 [111, 112]	Used a radial basis function neural networks trained on pseudo amino acid composition and PSSM representations of proteins to predict subcellular localization and membrane proteins.	They presented classification performance and computation time on six localization datasets and a membrane protein dataset with various flavors of their method.	SVM PSSM Pseudo composition
Hoze et al. 2013 [113]	Developed a SVM approach based on the amino acid composition representation of proteins to classify peptide cleavage prediction.	This method was able to predict peptide cleavage with the following top scoring performance measures: accuracy (0.75), sensitivity (0.81), and specificity (0.99).	SVM Composition
Chou et al. 2012, 2010,2010, 2010,2010, 2009,2007, 2007 [114-123]	Developed a series of algorithms including: mPLoc series and iLoc series. They used different algorithms such as fuzzy k-NN and SVM based on a series of features. These features include pseudo-amino acid composition, quasi-sequence-order effect, physicochemical composition, and other distance functions based on protein interactions.	Their methods were able to improve the performance of classification or offer unique insightful information on a variety of datasets.	Fuzzy k-Nearest Neighbor SVM

- Table 3 continued -

Huang et al. 2007 [124]	Developed ProLoc, an evolutionary support vector machine (ESVM) based classifier with automatic selection from a large set of physicochemical composition (PCC) features to design an accurate system for predicting protein subnuclear localization.	ProLoc utilizing the selected 33 and 28 physicochemical composition features has accuracies of 56.37% for SNL6 and 72.82% for SNL9, which are better than previous methods.	SVM
Han et al. 2006 [125]	Developed a system for predicting disordered regions in proteins based on decision trees and reduced amino acid composition. Concise rules based on biochemical properties of amino acid side chains are generated for prediction.	In cross-validation tests, with four groups of reduced amino acid composition, this system can achieve a recall of 80% at a 13% false positive rate for predicting disordered regions, and the overall accuracy can reach 83.4%. This prediction accuracy is comparable to most, and better than some, existing predictors. Advantages of this approach are high prediction accuracy for long disordered regions and efficiency for large-scale sequence analysis	Decision Trees Reduced Alphabets
Guo and Lin. 2006 [126, 127]	Developed TSSub for predicting localization in eukaryotic proteins. This system extracts features from both profiles and amino acid sequences. Different features are extracted from profiles by multiple probabilistic neural network (PNN) classifiers (the amino acid composition from whole profiles and N-terminus of profiles and the dipeptide composition) Also, a SVM classifier is added to implement the residue-couple feature extracted from amino acid sequences.	The overall accuracies of TSSub reach 93.0 and 77.4% on the Reinhardt and Hubbard's eukaryotic protein dataset and Huang and Li's eukaryotic protein dataset, respectively.	SVM Eukaryotes Composition
Huttenhower et al. 2006 [128, 129]	Developed Bayesian networks to predict functional relationships between proteins under a variety of conditions. This study considers the effect of network structure and compares expert estimated conditional probabilities with those learned using a generative method and a discriminative method.	They find that it is critical to consider variation across biological functions; even when global performance is strong, some categories are consistently predicted well, and others are difficult to analyze. All learned models outperform the equivalent expert estimated models, although this effect diminishes as the amount of available data decreases.	Expectation Maximization Extended Logistic Regression

- Table 3 continued -

Bhasin et al. 2005, 2004 [130-134]	Developed (ESLpred and PSLpred) hybrid methods using information that is based on amino acid composition, properties of amino acids, dipeptide composition, and PSI-BLAST results. Also, produced a method to predict nuclear receptors.	Using the hybrid method they were able to predict the localization of eukaryotic proteins better then by each method individually. They reported accuracies ranging from 68.2 to 95.3%.	SVM
Sarda et al. 2005 [135]	Developed pSLIP which uses SVMs in conjunction with multiple physicochemical properties of amino acids to predict protein subcellular localization in eukaryotes across six different locations. Unlike other algorithms, contextual information is preserved by dividing the protein sequences into clusters.	The algorithm was applied to the dataset provided by Park and Kanehisa and it obtained prediction accuracies for the different classes ranging from 87.7-97.0% with an overall accuracy of 93.1%	SVM Clustering
Yang and Chou 2004 [136]	Developed a bio-support vector machine that replaces the kernel function of a SVM with a similarity matrix for amino acids.	They used their method to predict HIV protease cleavage sites. Their method had a mean accuracy of 92%. Their method had a large advantage in reducing the model complexity and enhancing the model robustness.	SVM
Hua et al. 2001[137]	Developed a Support Vector Machine algorithm to predict the subcellular localization of proteins based on their amino acid composition.	The total prediction accuracies reach 91.4% for prokaryotic organisms (DS8) and 79.4% for eukaryotic organisms (DS7). This method had better results versus other methods using amino acid composition for prediction.	SVM
Reinhardt and Hubbard 1998 [138]	Trained neural networks to predict the subcellular location of proteins in prokaryotic or eukaryotic cells from their amino acid composition. Their dataset is one of the standards for testing subcellular localization methods.	For three possible subcellular locations in prokaryotic organisms this method had a prediction accuracy of 81%. For eukaryotic proteins an overall prediction accuracy of 66%was achieved for four localization sites.	ANN

Table 4: Protein interaction networks, protein structure, literature and text mining, gene expression, or the integration of other relevant data approaches for functional annotation of proteins.

Author Year	Description	Results	Methods
Hoehndorf et al. 2013 [139]	Described a method to assign functional annotation using phenotype information and GO annotations. They applied the approach to yeast, worm, zebrafish, fruitfly, and mouse.	They presented ROC curves with AUC values to show that predicting genetic and protein interactions improve if you also use gene function data. They were also able to classify a large number of novel genes in the datasets.	GO Phenotypes Genetic interactions PPI
Cozzetto et al. 2013 [140]	Developed a classifier that functional annotates (GO terms) proteins using many different types of data. The data include: sequence, gene expression, PPI, UniProt annotations. They also develop a new scoring metric to measure functional annotation.	They tested the method using 595 proteins from the 2011 Automated Function Prediction meeting and showed they were able to get recall values about 0.5 on molecular function annotations and 0.3 on biological process annotations.	GO PPI UniProt Gene expression
Sasidharen et al. 2012 [141]	Developed a software package, *GFam*, that does functional annotations of gene/protein families. This approach combine sequence and domain data.	Compared to data in InterProt for Arabidopsis, *GFam* increased sequence coverage by 7.2% and residue coverage by 14.6%.	Domain Greedy method InterProt
Wilikins et al. 2012 [142]	Reviewed pattern-based methods for functionally annotating proteins.	The types of patterns this paper reviews include: sequence-based, structure-based, phylogenomic, and synthesis through evolution. A evolutionary patterns and functional redesign case study is also performed.	Pattern-based
Frech et al. 2010 [143]	Developed a method to classify gene families across genomes by using existing curated gene families of reference species to classify genes of related genomes.	They tested their method on *C. elegans* and four related *Caenorhabditis* species. They compared their results with other curated data on chemosensory genes and found good agreement.	*C. elegans* Gene Families
Pandey et al. 2009 [144]	Developed a method to improve functional classification by including interrelationships between functional classes . They used a k-nearest neighbor approach using a measure that evaluates semantic similarity betweens nodes in an ontology.	They tested their method on several datasets to functional classify GO Biological process labels. They results show their method was able to produce more accurate predictions. They also were able to uncover previously unknown and novel annotations.	GO k-Nearest Neighbor

- Table 4 continued -

Malmstrom et al. 2007 [145]	Yeast proteins were parsed into 14,934 domains, and proteins with low sequence similarity to proteins of known structure were folded using the Rosetta de novo structure prediction method. This structural data was integrated with process, component, and function annotations from the Saccharomyces Genome Database to assign yeast protein domains to SCOP super-families using a simple Bayesian approach	They predicted the structure of 3,338 putative domains and assigned SCOP super-family annotations to 581 of them. They assigned structural annotations to 7,094 predicted domains based on fold recognition and homology modeling methods.	Structural information
Schuemie et al. 2007 [146]	Developed a method based on a combination of cross-species sequence homology searches and the corresponding literature. This approach facilitated the direct association between sequence data and information from biological texts describing function.	Clustering of the DEAD-box protein family of RNA helicases confirmed that the proteins shared similar morphology although functional subfamilies were accurately identified by this approach. They were able to predict novel nuclear proteins.	Homology search Clustering
Zhu et al. 2007 [147]	Used a previously published method GESTs to use information based on gene expression similarity, taxonomy similarity of the functional classes, and protein-protein interaction network data to predict functional classes in the Gene Ontology.	Based on the yeast protein-protein interaction data from MIPS and a dataset of gene expression profiles, they show that this method is powerful for predicting protein function to very specific terms.	PPI networks
Hayete et al. 2005 [148]	Developed GoTree, a method that uses GO domains and textual information based on similar proteins to predict a proteins GO label.	This method is more sensitive when compared to the InterPro2GO performance and suffers only some precision decrease. In comparison to the InterPro2GO this method improved the sensitivity by 22%, 27% and 50% for Molecular Function, Biological Process and Cellular GO terms respectively.	Decision Tree Text mining

- Table 4 continued -

Martin et al. 2004 [149]	Developed GOtcha, a method for predicting gene product function by annotation with GO terms. GOtcha uses associations with term-specific probability (P-score) measures of confidence. Term-specific probabilities are a novel feature that allows for the identification of conflicts or uncertainty in annotation.	GOtcha provided 60% better recovery of annotation terms and 20% higher selectivity than annotation with TOPBLAST.	Decision Tree expectation maximization
Lanckriet et al. 2004 [88]	Used a SVM to predict ribosomal and membrane proteins. As a kernel to the SVM they used a wide variety of genome-wide measurements including BLAST, PFam HMM, protein interaction, gene expression, multiple sequence alignment, hydropathy profile, and random numbers.	Their combined method for using multiple sources of genome-wide information performed better than using any individual method alone. On the ribosomal proteins their method was able to predict with very high ROC scores (>.999) The membrane proteins predictions had a .922 ROC score. Showing this is a very effective method.	SVM, HMM, BLAST
Jensen et al. 2003 [150]	Developed a neural network approach to predict human protein function according to Gene Ontology labels.	On a dataset of 11 different GO labels, their method was able to produce sensitivities of at least 50% with false positive rates of below 10% on all data and on the best categorizes 70% sensitivity with 5% false positive rates.	Neural networks
Deng et al. 2003 [151]	Developed an integrated probabilistic model to protein physical interactions, genetic interactions, highly correlated gene expressions, protein complexes, and domain structures to predict protein function. This model is an extension of Markovian random fields.	On a database based on MIPS protein function classification they were able to combine their methods and get a sensitivity and selectivity tradeoff of about 76% apiece. This method works well for predicting multiple functions for a single protein.	Markov random fields (MRF)
Claros and Vincens 1996 [152]	Developed a method using discriminant analysis using 47 parameters from a set of mitochondrial proteins found in SwissProt. A computational method that facilitates the analysis and objective prediction of mitochondrially imported proteins has been developed.	Based on amino acid sequence alone, 75-97% of the mitochondrial proteins studied have been predicted to be imported into mitochondria. The existence of mitochondrial-targeting sequences is predicted in 76-94% of the analyzed mitochondrial precursor proteins.	PCA

Table 5: Computational methods for prediction of misannotated proteins

Author Year	Description	Results
Bell et al. 2012 [153]	Performed a case study using the UniProtKB database to define a measure for the quality of large-scale annotation.	Power-law distributions were applied to the UniProtKB database to create a quality metric for functional annotation. They showed that the quality of the database has matured over time and that manual annotations were superior to automated annotations.
Faria et al. 2012 [14]	Presented a new method for literature-based GO term annotation by combining a literature-based approach with computational predictions from SGD.	On a subset of proteins from SGD, the proposed method was able to update 24% of the unknown proteins with a GO annotation.
Skunca et al. 2012 [154]	Introduced a method that evaluates electronic GO annotations by comparing successive releases of the UniProt Gene Ontology Annotation database.	They analyzed over 190,000 gene products and showed a lot of figures and graphs displaying the quality of the annotations. They found that the accuracy of electronic annotations is very comparable to expert-curated annotations.
Costanzo et al. 2011 [12]	Analyzed the GO annotations of UniProt proteins and discuss the quality of the annotations. They also developed an algorithm to learn relationships between GO terms.	The show that 64% of the UniPotKB proteins are misannotated. Their algorithm predicted 501relationships (with 94% precision) between molecular function terms.
Park et al. 2011, 2005 [18, 19]	Developed GOChase, a set of web-based utilities to detect and correct the errors in GO-based annotations.	They determine an error-conscious system for GO can help GO-based high-level analysis tools that use GO annotations. Functionalities like showing the evolution history and redirecting to the correct target term may benefit current GO browsers
Klimke et al. 2011 [16]	Discusses the current state and major issues of automated annotation of sequences.	The present the need for standards and guidelines for submission, retrieval, exchange, and analysis of data. They provide examples and advantages of using such standards.
Poptsova et al. 2010 [20]	Reviewed misannotations in public databases and possible reasons they occur. They explored both errors in gene calling and errors in functional annotation.	They found errors can occur at multiple stages of annotation. Errors also were more likely in earlier annotated genomes and that second generation tools had less misannotations.
Schnoes et al. 2009 [155]	Reported the misannotation rates for function in UniProtKB/Swiss-Prot, GenBank NR, UniProtKB/TrEMBL, and KEGG for 37 enzyme families.	For the four databases, levels of misannotations ranged from 0 - 63%, they also showed the levels of misannotation has been increasing from 1993 to 2005.
Jones et al. 2007 [156]	Present a new method based on maximal precision for estimating the error rate in the GO database.	They estimated the error rate to be between 28% - 30%. Annotations with the evidence code ISS (Inferred by Sequence Similarity) has estimated error rates at 49%, non-ISS annotations has error rates at 13% - 18%.
Gilks et al. 2005 [21]	Explore the possibility of chains of misannotation occurring in public databases. They developed a dynamical probabilistic model for these misannotation chains.	By exploring the consequences of the model for annotation quality they conclude that it is evident that this iterative approach leads to a systematic deterioration of database quality.

- Table 5 continued -

Green et al. 2005 [15]	Report on a new type of systematic annotation error in genome and pathway databases that results from the misinterpretation of partial Enzyme Commission (EC) numbers.	They observe this type of error in multiple databases, including KEGG, VIMSS and IMG, all of which assign genes to KEGG pathways. The Escherichia coli subset of the KEGG database exhibits this error for 6.8% of its gene-reaction assignments.
Dolan et al. 2005 [13]	Developed a methodology for assessing the consistency of GO annotations provided by different annotation groups. The method is completely general and can be applied to compare any two sets of GO annotations.	They present the results obtained by comparing GO annotations for mouse and human gene sets.
Naumoff et al. 2004 [17]	Developed a method to bridge gaps between genes that have been annotated and ones that have been experimentally studied but are not correlated with sequence data in current databases. When possible they collect a body of facts about experimental data, homology, unnoticed sequence data, and accurate information about gene context.	They show that a set of closely related sequences which have been annotated as ornithine carbamoyltransferases are actually putrescine carbamoyltransferases.

Table 6: Hubs and non-hubs in PPI networks.

Author Year	Description	Keywords
Song et al. 2013 [157]	Focused on the relationship between the number of interactions a protein has and its functional importance in a cellular network. The types of interactions are also explored and how they relate to protein essentiality.	Cellular networks Protein essentiality BioGRID GO Yeast
Banky et al. 2013 [158]	Developed a PageRank method for protein networks that combines information about connected nodes along with a nodes degree. This method identifies important nodes regardless if the node is a hub or not. This method was applied to s protein metabolic network.	PageRank Hub / non-Hubs Metabolic networks
Latha et al. 2011 [47]	Used 20 sequence-based features as input to a Bayesian model to predict hubs and hub behavior. These features include physiochemical, thermodynamic and conformational properties of amino acid residues. They were able to predict proteins from *E. coli* and human with accuracies ranging from 87.8 to 99.5%	Bayesian Human *E. coli*
Choura et al. 2011 [159]	Used a comparative modeling and docking-based methods to predict protein-protein complexes of hubs. They use the human NR-RTK network. Their results show that some interactions are mutually exclusive while others can occur at the same time.	Docking PPI Comparative Models Human
Zhang et al. 2010 [160]	Developed a method, CoEWC, to discovery new essential proteins. This method is based on properties of PPI networks and co-expression of interacting networks. This method outperform other classical measures when evaluated on the Yeast PPI network.	Essential proteins PPI Co-expression Yeast
Cho et al. 2010 [161]	Developed a method to provide hierarchical ordering of proteins based on a interactome network. This method also identifies hubs and provides a confidence score. They tested this method on the Yeast protein interactome network.	Hierarchical ordering Hub prediction Yeast
Patil et al. 2010 [39]	Reviewed features and properties that have been shown to be particular to hub proteins. These include: binding ability, disorder, surface charge, domain distribution, and functional domains.	Hub properties Disorder
Hsing et al. 2009, 2008 [45, 162]	Presented two studies in identifying hub-proteins. The first is based on GO annotations and interaction data. They used data from yeast, E. coli, fly, and human. The second method uses over 1300 protein features based on QSAR parameters, sequence features, interaction data, and functional annotations. They applied boosting trees to these features for the above mentioned species.	Boosting trees Hub prediction GO Yeast Human Fly E. coli
Manna et al. 2009 [163]	Discussed the properties of hub and non-hubs in protein interaction networks. They focused on the evolutionary rate between complex-forming and non-complex forming hubs and the structural disorder of hubs.	Human Disorder Evolutionary rate
Higurashi et al. 2008 [164]	Developed a method to identify hub proteins and further classifies the hubs based on the types of interactions (transient vs. permanent). This method is based on statistical analysis of PDB data. They also looked at properties of the different types of hubs including: disorder regions, structural flexibility, charge, and polarity.	Hub prediction PPI Transient interactions disorder

- Table 6 continued -

Lin et al. 2008 [165]	Developed a web-based method, *Hubba*, to identify and scores hub proteins. This method is based on using graph theory (Maximum Neighborhood Component (MNC and Density of MNC) on PPI data.	Hub prediction Graph theory PPI

Table 7: Structural and kinetic annotation of protein hubs.

Author Year	Description	Keywords
Chang et al. 2013 [166]	Explored two yeast interactive datasets to find behavior of date and party hub proteins. Properties that they analyzes include: expression patterns, topological roles, and functional classifications.	Yeast Date / Party hubs PPI
Goel et al. 2012, 2011 [167, 168]	Two papers. The first is a review of the roles and properties of Date hubs and analyses the interactions in the Yeast cell cycle. They classify Date proteins into two categories: dynamic hubs with static partners and static hubs with dynamic partners. The second paper presents a four-dimensional protein interaction network viewer and analyzes singlish date hub interactions in yeast.	Date hubs Yeast
Keskin et al. 2012, 2012, 2010, 2010, 2009, 2008, 2008, 2008 [27, 34, 169-174]	Presented a series of papers on predicting protein interfaces, hot spots, and hot regions. Hot regions are interacting residues that contribute more to binding free energy. Hot regions are clusters of hot spots. The hot spot predictor, HotPoint, is based on solvent accessibility and energy contribution rules. This method reaches 70% accuracy. The hot region prediction method, HotSpot, is based on HotPoint.	Protein interface prediction Hot spot Hot regions
Wang et al. 2012 [175]	Explored the properties of date and party hubs using three-dimensional structures, genomic essentiality, gene coexpression, and functional semantic similarity. They also concluded singlish proteins are most date hubs, but multi-interface are just as likely to be date or party hubs.	Date hubs Party hubs Singlish Multi-interface
Zhao et al. 2012 [176]	Developed a method to find associations between interfaces and function based on graph theory of multiple interface domains. They found that 40% of proteins have a multi-interface domain.	Multi-interface domains
Bhardwaj et al. 2011 [177]	Mapped three dimensional structure data onto PPI networks to classify singlish and multi-interface hubs. They also determine if individual interactions are permanent or transient. This work uses protein structures from PDB.	Singlish hubs Multi-interface hubs Permanent interactions Transient interactions
Dasgupta et al. 2011 [178]	Explored the structures of 16 date hub proteins to better understand how date hub proteins can bind different partners at overlapping interfaces. They looked at several properties of proteins including: surface area, compositions of amino acid residues and secondary structures, and side-chain orientations.	Date hub Overlapping interfaces

- Table 7 continued -

Mirzarezaee et al. 2010 [179]	Developed a few machine-learning algorithms (k-nearest neighbor and Bayes Classifier) to predict date and party hubs using feature reduction on a large set of sequenced-based features. These features include: amino acid sequences, domain contents, repeated domains, functional categories, biological processes, cellular compartments, disordered regions, and position specific scoring matrices. They tested their method on a set of Yeast proteins.	Date hubs Party hubs K-nearest neighbor Feature reduction
Agarwal et al. 2010 [180]	Critically analyzed the distinctions of hub proteins belonging to two classes (Date and Party). They show evidence that some of the original co-expression distinctions may be credited to a small subsets of the classes. They conclude the classifying proteins as date/party is not meaningful.	Date hubs Party hubs
Kahali et al. 2009 [36]	Described the evolutionary rate differences between date and party hubs in the Yeast PPI network. They conclude that party hubs have lower evolutionary rates and that it can be contributed to the order regions of those proteins.	Date hubs Party hubs Order/disorder
Batada et al. 2007, 2006 [181, 182]	Critically analyzed the conclusion that hubs can be classified as date and party. They concluded that the original datasets were too small and through appropriate statistical analysis that they could not support the date/party distinctions.	Date hubs Party hubs
Kim et al. 2006 [31]	Science paper. Combined structural modeling with network analysis on the yeast PPI network to define two types of hub proteins. Singlish hubs have mutually exclusive interactions where multi-interface hubs can interact with multiple partners at the same time. They also describe several properties that can be determined based on these two classes. The properties include: GO enrichment, co-expression, essentiality, evolutionary rate, and paralog percentage.	Singlish hubs Date hubs PPI Yeast
Ekman et al. 2006 [183]	Originally definined and described the different properties of date and party hubs. Demonstrated that Party hubs were less likely to have long disordered regions, more likely to interact with each other, and are more evolutionary conserved as compared to Date hubs.	Date hubs Party hubs

Table 8: G4-quadruplex: review papers.

Author Year	Description	Keywords
Maizels 2013, 2008, 2006 [48, 184, 185]	Provided an overview of G4 DNA and their potential to form in the genome. Discusses the potential role that G4-quadruplexes could have in the regulation of gene expression and how the structural mechanism could work (including G-Loops).	Gene regulation Gene expression G-loops
Chen and Yang 2012 [186]	Looked into the relationship between the sequence, structure, and stability of G4-quadruplexes. Gives examples of different types of structures that a G4-quadruplex can fold into (including parallel and antiparallel) and how it relates to the DNA sequence. Compares telomeric and promoter G-quadruplexes. Discusses how G4-quadruplexes interact with small molecules.	G4 structure Promoters
Bochman et al. 2012 [50]	Reviewed the characteristics of G4-quadruplexs structures, and their roles in genomic stability and cellular processes. Shows examples of several non-B-form DNA secondary structures. Looks at the role of G4-quadruplexes in telomeres, replication, transcription, and recombination.	non-B-form DNA Telomeres
Yang and Okamoto 2010 [58]	Discussed the recent advances in G4-quadruplex research in telomeres and the promoter regions of human oncogenes. The paper shows several example of oncogenes with G4-quadruplex structures including c-MYC, BCL-2, c-KIT, VEGF, and HIF1. Concludes with G4-quadruplexes as having potential for anticancer drug targets.	Human Promoters Oncogenes Drug targets
Brooks and Hurley 2010, 2009 [59, 187]	Reviewed the gene MYC, how it is overexpressed in various cancers, and models how non-duplex structures (including G4-quadruplexes) controls transcription. Discuses G4-quadruplexes in other oncogenes.	G4 structure Oncogenes i-motif MYC
Qin and Hurley 2008 [60]	Focused on G4-quadruplexes in the promoter regions of eukaryotes. Compares and give examples of two-tetrad, three-tetrad, and four-tetrad G4-quadruplexes. Discusses the role of quadruplexes in the promoters of c-MYC, VEGF, HIF-1, RET, KRAS, c-KIT, Bcl-2, PDGF-A, and c-Myb.	Promoters Two-tetrad Three-tetrad Four-tetrad HIF-1
Burge et al. 2006 [49]	Gave details on the sequence, structure, and topology of G4-quadruplexes. Provides examples of different telomere repeat sequences for different species. Proposes different topologies that a G4-quaruplex can form. Also, shows a multiple crystal structures formed by telomeric sequences and compares them to non-telomeric quadruplexes.	G4 structure Telomeric structure
Simonsson 2001 [188]	Gave in depth review of the structures of the G4-quadruplexes. Shows the G4-quadruplex motif and the role and structure of individual base pairs.	G4 motif G4 structure

Table 9: G4-quadruplex: genome-wide surveys.

Author Year	Description	Keywords
Eddy et al. 2011 [56]	Explored the relationship of RNA Pol II transcription complex pausing and G4Q motifs genome-wide in human. This paper compares genome-wide analysis of the pausing of the Pol II transcription complex downstream of the transcriptional start site, GC-richness, and the position of potential G4Qs. The study finds promoters of paused genes are enriched with G4Q.	Human Transcription pausing GC-richness
Todd et al. 2011 [54]	Examined the sequence similarity among G4Q across the human genome. This paper clusters and categorizes G4Q in the human genome based on sequence similarity. Explored individual clusters that are associated to particular functions including zinc fingers, leucocyte immunoglobulin-like receptor genes, and immunoglobulin genes.	Human Clustering Sequence similarity Zinc fingers Immunoglobulin
Capra et al. 2010 [57]	Demonstrated that G4Q are conserved and linked to genomic features in *Saccharomyces cerevisiae*. This paper found that G4Q were associated with promoter regions, rDNA, mitotic and meiotic break sites. G4Qs were also 10-fold more likely to be in the mitochondrial DNA than nuclear DNA.	Saccharomyces cerevisiae Promoter rDNA Double-strand break points
Halder et al. 2009, 2008 [52, 73]	Predicted G4-quadruplex motifs as nucleosome occupancy signals in *Saccharomyces cerevisiae* and human. Determined that G4Qs exclude nuclesomes in promoters and modulate transcription efficiency of promoters.	Saccharomyces cerevisiae Human Nucleosome occupancy Promoters
Du et al. 2009, 2008 [53, 189]	Investigated the regulatory role of G4Q in gene transcription in human. This paper looked at the potential of G4Q forming in the putative transcriptional regulatory regions in human genes. They found that G4Q found downstream of the transcriptional start site were associated with gene expression. A model is proposed on how G4Q leaves the DNA in an open state allowing for continued transcription and increased gene expression.	Human Transcription Gene Expression
Hupert et al. 2010, 2008, 2008, 2008, 2007, 2007, 2005 [71, 190-195]	Several papers that discussed the locations and potential roles that G4Q may play in the human genome. They show the distribution and enrichment of G4Q around the human TSS and UTRs.	Human TSS
Rawal et al. 2006 [51]	Described G4Q as a regulatory motif in *Escherichia coli*. A genome-wide analysis on over 61,000 open reading frames from 18 prokaryotes species and found that G4Q were found in promoter regions of genes associated with transcription, secondary metabolite biosynthesis, and signal transduction. The paper focused on mapping G4Q motifs to the regulatory network of *E.coli.*	Escherichia coli Cross-species Bacterial genomes

Table 10: G4-quadruplex: bioinformatics.

Author Year	Description	Keywords
Menendez et al. 2012 [65] **Kikin et al. 2006 [69]**	Developed a web-based server, QGRS-H Predictor, for finding and analyzing G4Qs in nucleotide sequences. The server also assigns a score to each G4Q based on the likelihood to form a stable G4-quadruplex. Other features include a homology score, homology map, and sequence viewer.	QGRS-H Predictor G-score
Wong et al. 2010 [66]	Created a database, QuadDB, of predicted G4Q in ten genomes (including human, but no plants). They also provide a G4Q prediction server called QuadPredict, which predicts G4Q along with thermal stability. They also allow users to download the pre-compiled version of *quadparser*.	QuadDB *quadparser* Thermal stability
Stegle et al. 2009 [67]	From sequence information alone, created a Bayesian prediction method for predicting the stability of G4Q. They also present a genome-wide analysis of the human genome using their tool.	Bayesian Human Stability
Mani et al. 2009 [68]	Combined a genome-wide G4Q analysis of the human genome with HapMap single nucleotide polymorphism (SNP) data to study the G4Q density in SNP hotspots. They presented evidence and a model how G4Q could be involved as determinants of recombination.	HapMap Human Hotspots Recombination
Yadav et al. 2008 [70]	Created a two-part database of G4Q motifs, QuadBase. The first part is EuQuad which is based on information of G4Q in the 10kb regions upstream of the transcriptional start site of human, chimp, rat, and mouse genes. The second part is ProQuad which has G4Q data for 146 prokaryotes. This online database also has the Pattern Search and Pattern Finder tools for discovering and searching for G4Q motifs.	QuadBase Promoter Human Prokaryotes
Huppert et al. 2005 [71]	Developed an algorithm to identify G4Q in DNA sequence and implemented a software tool called *quadparser*. They systematically studied loop lengths of potential G4Qs. They presented evidence that G4Q patterns were less likely to form in RNA forming sequences.	*quadparser* Loop lengths

Table 11: G4-quadruplex: Plants

Author Year	Description	Keywords
Takahashi et al. 2012 [75]	Characterized eight species' genomic G4Q representation (including 4 plant species: Arabidopsis, grape, rice, and poplar). They found increased presence of G4Q on the template strands near the TSS of genes and did GO enrichment on those genes.	Arabidopsis Grape Rice Poplar GO enrichment
Mullen et al. 2010 [76]	Reported the prevalence and significance of G4Q in the Arabidopsis genome. They compared these results with analysis of potential G4Q in 11 other species including human, maize, and rice. They show that three-tetrad G4Q motif are less prevalent in Arabidopsis as compared to other species and is actually less likely to appear in genic regions as compared to genic regions which is the opposite observation as compared to most other genome-wide studies on other species. They found that two-tetrad G4Q are more abundant and are more likely found in genes and in promoter regions.	Arabidopsis Maize Human Rice Comparative study Three-tetrad Two-tetrad

Summary

There have been many recent advances in using machine-learning techniques to predict sequence-function-structure relationships of proteins. Given the success or failures, advantages or disadvantages, and breakthroughs or pitfalls of these recent advances, we want to be able to improve the performance of these algorithms. We focus on improving the following performance criteria: classification performance, computational time, updateability, and flexibility. We approach these problems by building unique data representations, new or modified machine-learning algorithms, new knowledge representations, and/or new methods to interrupt the results from these algorithms. With these new methods and different representations of proteins, we offer an alternative, and sometime complementary, insight to the sequence-function-structure relationship of proteins.

Additionally, research on non-Watson-Crick base-pairing structures have received growing attention. Specifically, the four-stranded DNA structure G4-quadruplex is an

attractive target of research, especially its role in the promoters of oncogenes. Though this research field is relatively new and largely unexplored, progress already has been made in G4Q structures and their role in gene regulation [51, 56], gene expression [196], DNA-damage repair [197], transcriptional pausing [198], alternative polyadenylation, and mRNA shortening [199] . A G4Q sequence motif has been identified and has been used in several genome-wide studies. An area of research that has yet to be fully explored is the roles of G4Qs in plant species. We perform a genome-wide analysis of G4Qs in the model organism *Zea mays*, maize, and present potentially new forms of gene regulation in several metabolic pathways.

Dissertation Organization

In this dissertation, I aim to develop machine-learning methods for identification of biological labels for proteins and to use bioinformatics approaches to explore the role of G4-quaduplexes in gene regulation. We focus on machine-learning methods using sequence information alone. We developed a method to detect possible misannotations found in the Gene Ontology Database and a three-phase approach to identify binding patterns in hub proteins based on the yeast protein-protein interaction network. The final work analyzes the maize genome for the G4Q structural motif and explores the role of the G4Q in several metabolic pathways. This dissertation uses an alternative approach: chapters 2-4 are papers that have been published or are submitted to a journal for peer review.

Chapter 1: This chapter introduces the problems we studied, the relevant background material for this problem, a survey of current studies, and the outline of the dissertation.

Chapter 2: We report the development of a two-stage method, *HDTree*, consisting of a decision tree classifier based on combining the outputs of several composition-based and sequence homology based methods. We applied the *HDTree* to the problem of identifying potential misannotated proteins in the AmiGO database. We were able to locate hundreds of protein sequences that were potentially misannotated. The results have been published in the journal *BMC-Bioinformatics.* Carson Andorf contributed to experimental design, carried out computational experiments, and drafted the manuscript; Drena Dobbs and Vasant Honavar contributed to experimental design, discussions, and manuscript preparation.

Chapter 3: We report a three-phase approach for predicting the binding patterns of hub proteins. Phase I classifies a protein by whether or not it is likely to bind with another protein. Phase II determines if a protein-binding protein is a hub. Phase III predicts whether a protein-binding proteins as singlish-interface versus multiple-interface hubs and date versus party hubs. At each stage, we use sequence-based predictors trained using several standard machine-learning techniques. The results have been published in the journal *PLOS ONE.* Carson Andorf contributed to experimental design, carried out computational experiments and drafted the manuscript; Vasant Honavar and Taner Sen contributed to experimental design, discussions, and manuscript preparation.

Chapter 4: We report a bioinformatics approach to predict G4-quadruplexes in the maize genome and present evidence that maize G4Q elements may play a previously unrecognized role in coordinating global genomic responses to hypoxia and related energy crisis states. The results have been submitted to the journal *G3:Genes|Genomes|Genetics.* Carson Andorf contributed to experimental design, carried out the computational

experiments, and co-drafted the manuscript; Mykhailo Nibiletsky and Elizabeth Stroupe carried out the wet-lab experiments and reviewed the manuscript; Drena Dobbs, Karen Koch and Carolyn Lawrence contributed to experimental design, discussions, and manuscript review; Hank Bass led the design of computational and wet-lab experiments and co-drafted the manuscript.

Chapter 5: This chapter summarizes the reported work, describes individual contributions, and lays out potential future work.

References

1. Eisenberg D, Marcotte EM, Xenarios I, Yeates TO: **Protein function in the post-genomic era**. *Nature* 2000, **405**(6788):823-826.

2. Mitchell T: **Machine learning**. New York, NY, USA: McGraw Hill; 1997.

3. Witten IH, Frank E: **Data Mining: Practical Machine Learning Tools and Techniques with Java Implements**. San Mateo, CA: Morgan Kaufmann; 1999.

4. Baldi P, Brunak S: **Bioinformatics: The Machine Learning Approach**. Cambridge, MA: MIT Press; 1998.

5. Luscombe NM, Greenbaum D, Gerstein M: **What is bioinformatics? A proposed definition and overview of the field**. *Methods Inform Med* 2001, **40**(4):346-358.

6. Quinlan R: **Induction of decision tree**. *Machine Learning* 1986, **1**:81-106.

7. Vapnik V: **Statistical learning theory**. New York: Springer-Verlag; 1998.

8. Gene Ontology Consortium T: **Creating the gene ontology resource: design and implementation**. *Genome Res* 2001, **11**(8):1425-1433.

9. Gene Ontology Consortium T: **The Gene Ontology in 2010: extensions and refinements**. *Nucleic Acids Res* 2010, **38**(Database issue):D331-335.

10. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C *et al*: **The Gene Ontology (GO) database and informatics resource**. *Nucleic Acids Res* 2004, **32**(Database issue):D258-261.

11. Adams LG, Khare S, Lawhon SD, Rossetti CA, Lewin HA, Lipton MS, Turse JE, Wylie DC, Bai Y, Drake KL: **Enhancing the role of veterinary vaccines**

reducing zoonotic diseases of humans: linking systems biology with vaccine development. *Vaccine* 2011, **29**(41):7197-7206.

12. Costanzo MC, Park J, Balakrishnan R, Cherry JM, Hong EL: **Using computational predictions to improve literature-based Gene Ontology annotations: a feasibility study**. *Database (Oxford)* 2011, **2011**:bar004.

13. Dolan ME, Ni L, Camon E, Blake JA: **A procedure for assessing GO annotation consistency**. *Bioinformatics* 2005, **21**(1):i136-143.

14. Faria D, Schlicker A, Pesquita C, Bastos H, Ferreira AE, Albrecht M, Falcao AO: **Mining GO annotations for improving annotation consistency**. *PLoS One* 2012, **7**(7):e40519.

15. Green ML, Karp PD: **Genome annotation errors in pathway databases due to semantic ambiguity in partial EC numbers**. *Nucleic Acids Res* 2005, **33**(13):4035-4039.

16. Klimke W, O'Donovan C, White O, Brister JR, Clark K, Fedorov B, Mizrachi I, Pruitt KD, Tatusova T: **Solving the Problem: Genome Annotation Standards before the Data Deluge**. *Stand Genomic Sci* 2011, **5**(1):168-193.

17. Naumoff DG, Xu Y, Glansdorff N, Labedan B: **Retrieving sequences of enzymes experimentally characterized but erroneously annotated : the case of the putrescine carbamoyltransferase**. *BMC Genomics* 2004, **5**(1):52.

18. Park YR, Kim J, Lee HW, Yoon YJ, Kim JH: **GOChase-II: correcting semantic inconsistencies from Gene Ontology-based annotations for gene products**. *BMC Bioinformatics* 2011, **12 Suppl 1**:S40.

19. Park YR, Park CH, Kim JH: **GOChase: correcting errors from Gene Ontology-based annotations for gene products**. *Bioinformatics* 2005, **21**(6):829-831.

20. Poptsova MS, Gogarten JP: **Using comparative genome analysis to identify problems in annotated microbial genomes**. *Microbiology* 2010, **156**(Pt 7):1909-1917.

21. Gilks WR, Audit B, de Angelis D, Tsoka S, Ouzounis CA: **Percolation of annotation errors through hierarchically structured protein sequence databases**. *Math Biosci* 2005, **193**(2):223-234.

22. Gursoy A, Keskin O, Nussinov R: **Topological properties of protein interaction networks from a structural perspective**. *Biochem Soc Trans* 2008, **36**(Pt 6):1398-1403.

23. Kuzu G, Keskin O, Gursoy A, Nussinov R: **Constructing structural networks of signaling pathways on the proteome scale**. *Curr Opin Struct Biol* 2012.

24. Liu S, Zhu X, Liang H, Cao A, Chang Z, Lai L: **Nonnatural protein-protein interaction-pair design by key residues grafting**. *Proc Natl Acad Sci U S A* 2007, **104**(13):5330-5335.

25. Grigoryan G, Reinke AW, Keating AE: **Design of protein-interaction specificity gives selective bZIP-binding peptides**. *Nature* 2009, **458**(7240):859-864.

26. Fleishman SJ, Whitehead TA, Ekiert DC, Dreyfus C, Corn JE, Strauch EM, Wilson IA, Baker D: **Computational design of proteins targeting the conserved stem region of influenza hemagglutinin**. *Science* 2011, **332**(6031):816-821.

27. Keskin O, Gursoy A, Ma B, Nussinov R: **Principles of protein-protein interactions: what are the preferred ways for proteins to interact?** *Chem Rev* 2008, **108**(4):1225-1244.

28. Jeong H, Mason SP, Barabasi AL, Oltvai ZN: **Lethality and centrality in protein networks**. *Nature* 2001, **411**(6833):41-42.

29. Missiuro PV, Liu K, Zou L, Ross BC, Zhao G, Liu JS, Ge H: **Information flow analysis of interactome networks**. *PLoS Comput Biol* 2009, **5**(4):e1000350.

30. Zotenko E, Mestre J, O'Leary DP, Przytycka TM: **Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality**. *PLoS Comput Biol* 2008, **4**(8):e1000140.

31. Kim PM, Lu LJ, Xia Y, Gerstein MB: **Relating three-dimensional structures to protein networks provides evolutionary insights**. *Science* 2006, **314**(5807):1938-1941.

32. Aragues R, Sali A, Bonet J, Marti-Renom MA, Oliva B: **Characterization of protein hubs by inferring interacting motifs from protein interactions**. *PLoS Comput Biol* 2007, **3**(9):1761-1771.

33. Bellay J, Han S, Michaut M, Kim T, Costanzo M, Andrews BJ, Boone C, Bader GD, Myers CL, Kim PM: **Bringing order to protein disorder through comparative genomics and genetic interactions**. *Genome Biol* 2011, **12**(2):R14.

34. Cukuroglu E, Gursoy A, Keskin O: **Analysis of hot region organization in hub proteins**. *Ann Biomed Eng* 2010, **38**(6):2068-2078.

35. Fong JH, Panchenko AR: **Intrinsic disorder and protein multibinding in domain, terminal, and linker regions**. *Mol Biosyst* 2010, **6**(10):1821-1828.

36. Kahali B, Ahmad S, Ghosh TC: **Exploring the evolutionary rate differences of party hub and date hub proteins in Saccharomyces cerevisiae protein-protein interaction network**. *Gene* 2009, **429**(1-2):18-22.

37. Kim PM, Sboner A, Xia Y, Gerstein M: **The role of disorder in interaction networks: a structural analysis**. *Mol Syst Biol* 2008, **4**:179.

38. Pang K, Cheng C, Xuan Z, Sheng H, Ma X: **Understanding protein evolutionary rate by integrating gene co-expression with protein interactions**. *BMC Syst Biol* 2010, **4**:179.

39. Patil A, Kinoshita K, Nakamura H: **Hub promiscuity in protein-protein interaction networks**. *Int J Mol Sci* 2010, **11**(4):1930-1943.

40. Han JD, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, Dupuy D, Walhout AJ, Cusick ME, Roth FP *et al*: **Evidence for dynamically organized modularity in the yeast protein-protein interaction network**. *Nature* 2004, **430**(6995):88-93.

41. Kar G, Gursoy A, Keskin O: **Human cancer protein-protein interaction network: a structural perspective**. *PLoS Comput Biol* 2009, **5**(12):e1000601.

42. Kar G, Kuzu G, Keskin O, Gursoy A: **Protein-protein interfaces integrated into interaction networks: implications on drug design**. *Curr Pharm Des* 2012.

43. Jin G, Zhang S, Zhang XS, Chen L: **Hubs with network motifs organize modularity dynamically in the protein-protein interaction network of yeast**. *PLoS One* 2007, **2**(11):e1207.

44. Taylor IW, Linding R, Warde-Farley D, Liu Y, Pesquita C, Faria D, Bull S, Pawson T, Morris Q, Wrana JL: **Dynamic modularity in protein interaction networks predicts breast cancer outcome**. *Nat Biotechnol* 2009, **27**(2):199-204.

45. Hsing M, Byler K, Cherkasov A: **Predicting highly-connected hubs in protein interaction networks by QSAR and biological data descriptors**. *Bioinformation* 2009, **4**(4):164-168.

46. Nair MTaAS: **Prediction and disorderliness of hub proteins**. *International Journal of Bioinformatics Research* 2009, **1**(2):70-80.

47. Latha AB, Nair AS, Sivasankaran A, Dhar PK: **Identification of hub proteins from sequence**. *Bioinformation* 2011, **7**(4):163-168.

48. Maizels N: **Dynamic roles for G4 DNA in the biology of eukaryotic cells**. *Nat Struct Mol Biol* 2006, **13**(12):1055-1059.

49. Burge S, Parkinson GN, Hazel P, Todd AK, Neidle S: **Quadruplex DNA: sequence, topology and structure**. *Nucleic Acids Res* 2006, **34**(19):5402-5415.

50. Bochman ML, Paeschke K, Zakian VA: **DNA secondary structures: stability and function of G-quadruplex structures**. *Nat Rev Genet* 2012, **13**(11):770-780.

51. Rawal P, Kummarasetti VB, Ravindran J, Kumar N, Halder K, Sharma R, Mukerji M, Das SK, Chowdhury S: **Genome-wide prediction of G4 DNA as regulatory motifs: role in Escherichia coli global regulation**. *Genome Res* 2006, **16**(5):644-655.

52. Halder K, Halder R, Chowdhury S: **Genome-wide analysis predicts DNA structural motifs as nucleosome exclusion signals**. *Mol Biosyst* 2009, **5**(12):1703-1712.

53. Du Z, Zhao Y, Li N: **Genome-wide analysis reveals regulatory role of G4 DNA in gene transcription**. *Genome Res* 2008, **18**(2):233-241.

54. Todd AK, Johnston M, Neidle S: **Highly prevalent putative quadruplex sequence motifs in human DNA**. *Nucleic Acids Res* 2005, **33**(9):2901-2907.

55. Todd AK, Neidle S: **Mapping the sequences of potential guanine quadruplex motifs**. *Nucleic Acids Res* 2011, **39**(12):4917-4927.

56. Eddy J, Vallur AC, Varma S, Liu H, Reinhold WC, Pommier Y, Maizels N: **G4 motifs correlate with promoter-proximal transcriptional pausing in human genes**. *Nucleic Acids Res* 2011, **39**(12):4975-4983.

57. Capra JA, Paeschke K, Singh M, Zakian VA: **G-quadruplex DNA sequences are evolutionarily conserved and associated with distinct genomic features in Saccharomyces cerevisiae**. *PLoS Comput Biol* 2010, **6**(7):e1000861.

58. Yang D, Okamoto K: **Structural insights into G-quadruplexes: towards new anticancer drugs**. *Future Med Chem* 2010, **2**(4):619-646.

59. Brooks TA, Hurley LH: **The role of supercoiling in transcriptional control of MYC and its importance in molecular therapeutics**. *Nat Rev Cancer* 2009, **9**(12):849-861.

60. Qin Y, Hurley LH: **Structures, folding patterns, and functions of intramolecular DNA G-quadruplexes found in eukaryotic promoter regions**. *Biochimie* 2008, **90**(8):1149-1171.

61. Cahoon LA, Seifert HS: **Transcription of a cis-acting, noncoding, small RNA is required for pilin antigenic variation in Neisseria gonorrhoeae**. *PLoS Pathog* 2013, **9**(1):e1003074.

62. Beckett J, Burns J, Broxson C, Tornaletti S: **Spontaneous DNA lesions modulate DNA structural transitions occurring at nuclease hypersensitive element III(1) of the human c-myc proto-oncogene**. *Biochemistry* 2012, **51**(26):5257-5268.

63. Juranek SA, Paeschke K: **Cell cycle regulation of G-quadruplex DNA structures at telomeres**. *Curr Pharm Des* 2012, **18**(14):1867-1872.

64. Weng HY, Huang HL, Zhao PP, Zhou H, Qu LH: **Translational repression of cyclin D3 by a stable G-quadruplex in its 5' UTR: implications for cell cycle regulation**. *RNA Biol* 2012, **9**(8):1099-1109.

65. Menendez C, Frees S, Bagga PS: **QGRS-H Predictor: a web server for predicting homologous quadruplex forming G-rich sequence motifs in nucleotide sequences**. *Nucleic Acids Res* 2012, **40**(Web Server issue):W96-W103.

66. Wong HM, Stegle O, Rodgers S, Huppert JL: **A toolbox for predicting g-quadruplex formation and stability**. *J Nucleic Acids* 2010, **2010**.

67. Stegle O, Payet L, Mergny JL, MacKay DJ, Leon JH: **Predicting and understanding the stability of G-quadruplexes**. *Bioinformatics* 2009, **25**(12):i374-382.

68. Mani P, Yadav VK, Das SK, Chowdhury S: **Genome-wide analyses of recombination prone regions predict role of DNA structural motif in recombination**. *PLoS One* 2009, **4**(2):e4399.

69. Kikin O, D'Antonio L, Bagga PS: **QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences**. *Nucleic Acids Res* 2006, **34**(Web Server issue):W676-682.

70. Yadav VK, Abraham JK, Mani P, Kulshrestha R, Chowdhury S: **QuadBase: genome-wide database of G4 DNA--occurrence and conservation in human, chimpanzee, mouse and rat promoters and 146 microbes**. *Nucleic Acids Res* 2008, **36**(Database issue):D381-385.

71. Huppert JL, Balasubramanian S: **Prevalence of quadruplexes in the human genome**. *Nucleic Acids Res* 2005, **33**(9):2908-2916.

72. Clark DW, Phang T, Edwards MG, Geraci MW, Gillespie MN: **Promoter G-quadruplex sequences are targets for base oxidation and strand cleavage during hypoxia-induced transcription**. *Free Radic Biol Med* 2012, **53**(1):51-59.

73. Verma A, Halder K, Halder R, Yadav VK, Rawal P, Thakur RK, Mohd F, Sharma A, Chowdhury S: **Genome-wide computational and expression analyses reveal G-quadruplex DNA motifs as conserved cis-regulatory elements in human and related species**. *J Med Chem* 2008, **51**(18):5641-5649.

74. Siddiqui-Jain A, Grand CL, Bearss DJ, Hurley LH: **Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription**. *Proc Natl Acad Sci U S A* 2002, **99**(18):11593-11598.

75. Takahashi H, Nakagawa A, Kojima S, Takahashi A, Cha BY, Woo JT, Nagai K, Machida Y, Machida C: **Discovery of novel rules for G-quadruplex-forming**

sequences in plants by using bioinformatics methods. *J Biosci Bioeng* 2012, **114**(5):570-575.

76. Mullen MA, Olson KJ, Dallaire P, Major F, Assmann SM, Bevilacqua PC: **RNA G-Quadruplexes in the model plant species Arabidopsis thaliana: prevalence and possible functional roles**. *Nucleic Acids Res* 2010, **38**(22):8149-8163.

77. Bennetzen JL, Hake SC: **Handbook of Maize: Genetics and Genomics**. New York: Springer; 2009.

78. Lopez D, Pazos F: **COPRED: prediction of fold, GO molecular function and functional residues at the domain level**. *Bioinformatics* 2013.

79. Lopez D, Pazos F: **Concomitant prediction of function and fold at the domain level with GO-based profiles**. *BMC Bioinformatics* 2013, **14**(3):1471-2105.

80. Sleator RD, Walsh P: **An overview of in silico protein function prediction**. *Arch Microbiol* 2010, **192**(3):151-155.

81. Xiong H, Capurso D, Sen S, Segal MR: **Sequence-based classification using discriminatory motif feature selection**. *PLoS One* 2011, **6**(11):e27382.

82. Meng S, Brown DE, Ebbole DJ, Torto-Alalibo T, Oh YY, Deng J, Mitchell TK, Dean RA: **Gene Ontology annotation of the rice blast fungus, Magnaporthe oryzae**. *BMC Microbiol* 2009, **9 Suppl 1**:S8.

83. Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R *et al*: **Pfam: clans, web tools and services**. *Nucleic Acids Res* 2006, **34**(Database issue):D247-251.

84. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K *et al*: **The Pfam protein families database**. *Nucleic Acids Res* 2010, **38**(Database issue):D211-222.

85. Sonnhammer EL, Eddy SR, Durbin R: **Pfam: a comprehensive database of protein domain families based on seed alignments**. *Proteins* 1997, **28**(3):405-420.

86. Vinayagam A, del Val C, Schubert F, Eils R, Glatting KH, Suhai S, Konig R: **GOPET: a tool for automated predictions of Gene Ontology terms**. *BMC Bioinformatics* 2006, **7**:161.

87. Jones CE, Baumann U, Brown AL: **Automated methods of predicting the function of biological sequences using GO and BLAST**. *BMC Bioinformatics* 2005, **6**:272.

88. Lanckriet GR, Deng M, Cristianini N, Jordan MI, Noble WS: **Kernel-based data fusion and its application to protein function prediction in yeast**. *Pac Symp Biocomput* 2004:300-311.

89. Murvai J, Vlahovicek K, Szepesvari C, Pongor S: **Prediction of protein functional domains from sequences using artificial neural networks**. *Genome Res* 2001, **11**(8):1410-1417.

90. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, **215**(3):403-410.

91. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **25**(17):3389-3402.

92. Gribskov M, McLachlan AD, Eisenberg D: **Profile analysis: detection of distantly related proteins**. *Proc Natl Acad Sci U S A* 1987, **84**(13):4355-4358.

93. Saidi R, Maddouri M, Mephu Nguifo E: **Protein sequences classification by means of feature extraction with substitution matrices**. *BMC Bioinformatics* 2010, **11**:175.

94. Sarac OS, Atalay V, Cetin-Atalay R: **GOPred: GO molecular function prediction by combined classifiers**. *PLoS One* 2010, **5**(8):e12382.

95. Pagni M, Ioannidis V, Cerutti L, Zahn-Zabal M, Jongeneel CV, Falquet L: **MyHits: a new interactive resource for protein annotation and domain identification**. *Nucleic Acids Res* 2004, **32**(Web Server issue):W332-335.

96. Pagni M, Ioannidis V, Cerutti L, Zahn-Zabal M, Jongeneel CV, Hau J, Martin O, Kuznetsov D, Falquet L: **MyHits: improvements to an interactive resource for analyzing protein sequences**. *Nucleic Acids Res* 2007, **35**(Web Server issue):W433-437.

97. Kunik V, Solan Z, Edelman S, Ruppin E, Horn D: **Motif extraction and protein classification**. *Proc IEEE Comput Syst Bioinform Conf* 2005:80-85.

98. Wang X, Schroeder D, Dobbs D, Honavar V: **Data-Driven Discovery of Protein Function Classifiers: Decision Trees Based on Motifs Outperform Those Based on PROSITE Patterns and Profiles on Peptidase Families.** . In*: 2002*.

99. Wang X, Schroeder D, Dobbs D, Honavar V: **Automated data-driven discovery of protein function classifiers**. *Information Sciences* 2003, **155**:1-18.

100. Ben-Hur A, Brutlag D: **Remote homology detection: a motif based approach**. *Bioinformatics* 2003, **19**(1):i26-33.

101. Huang JY, Brutlag DL: **The EMOTIF database**. *Nucleic Acids Res* 2001, **29**(1):202-204.

102. Attwood TK, Croning MD, Flower DR, Lewis AP, Mabey JE, Scordis P, Selley JN, Wright W: **PRINTS-S: the database formerly known as PRINTS**. *Nucleic Acids Res* 2000, **28**(1):225-227.

103. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Barrell D, Bateman A, Binns D, Biswas M, Bradley P, Bork P *et al*: **The InterPro Database, 2003 brings increased coverage and new features**. *Nucleic Acids Res* 2003, **31**(1):315-318.

104. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Biswas M, Bradley P, Bork P, Bucher P *et al*: **InterPro: an integrated documentation resource for protein families, domains and functional sites**. *Brief Bioinform* 2002, **3**(3):225-235.

105. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bradley P, Bork P, Bucher P, Cerutti L *et al*: **InterPro, progress and status in 2005**. *Nucleic Acids Res* 2005, **33**(Database issue):D201-205.

106. Henikoff S, Henikoff JG, Pietrokovski S: **Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations**. *Bioinformatics* 1999, **15**(6):471-479.

107. Falquet L, Pagni M, Bucher P, Hulo N, Sigrist CJ, Hofmann K, Bairoch A: **The PROSITE database, its status in 2002**. *Nucleic Acids Res* 2002, **30**(1):235-238.

108. Bairoch A, Bucher P: **PROSITE: recent developments**. *Nucleic Acids Res* 1994, **22**(17):3583-3589.

109. Hofmann K, Bucher P, Falquet L, Bairoch A: **The PROSITE database, its status in 1999**. *Nucleic Acids Res* 1999, **27**(1):215-219.

110. Feng PM, Chen W, Lin H, Chou KC: **iHSP-PseRAAAC: Identifying the heat shock protein families using pseudo reduced amino acid alphabet composition**. *Anal Biochem* 2013.

111. Huang C, Yuan J: **Using radial basis function on the general form of Chou's pseudo amino acid composition and PSSM to predict subcellular locations of proteins with both single and multiple sites**. *Biosystems* 2013, **113**(1):50-57.

112. Huang C, Yuan JQ: **A multilabel model based on Chou's pseudo-amino acid composition for identifying membrane proteins with both single and multiple functional types**. *J Membr Biol* 2013, **246**(4):327-334.

113. Hoze E, Tsaban L, Maman Y, Louzoun Y: **Predictor for the effect of amino acid composition on CD4+ T cell epitopes preprocessing**. *J Immunol Methods* 2013, **391**(1-2):163-173.

114. Chou KC: **Prediction of protein cellular attributes using pseudo-amino acid composition**. *Proteins* 2001, **43**(3):246-255.

115. Chou KC, Cai YD: **Predicting protein quaternary structure by pseudo amino acid composition**. *Proteins* 2003, **53**(2):282-289.

116. Chou KC, Shen HB: **Plant-mPLoc: a top-down strategy to augment the power for predicting plant protein subcellular localization**. *PLoS One* 2010, **5**(6):e11335.

117. Chou KC, Shen HB: **Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites**. *J Proteome Res* 2007, **6**(5):1728-1734.

118. Chou WC, Yin Y, Xu Y: **GolgiP: prediction of Golgi-resident proteins in plants**. *Bioinformatics* 2010, **26**(19):2464-2465.

119. Shen HB, Chou KC: **Virus-mPLoc: a fusion classifier for viral protein subcellular location prediction by incorporating multiple sites**. *J Biomol Struct Dyn* 2010, **28**(2):175-186.

120. Shen HB, Chou KC: **Gneg-mPLoc: a top-down strategy to enhance the quality of predicting subcellular localization of Gram-negative bacterial proteins**. *J Theor Biol* 2010, **264**(2):326-333.

121. Shen HB, Chou KC: **Gpos-mPLoc: a top-down approach to improve the quality of predicting subcellular localization of Gram-positive bacterial proteins**. *Protein Pept Lett* 2009, **16**(12):1478-1484.

122. Shen HB, Chou KC: **Hum-mPLoc: an ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites**. *Biochem Biophys Res Commun* 2007, **355**(4):1006-1011.

123. Wu ZC, Xiao X, Chou KC: **iLoc-Gpos: a multi-layer classifier for predicting the subcellular localization of singleplex and multiplex Gram-positive bacterial proteins**. *Protein Pept Lett* 2012, **19**(1):4-14.

124. Huang H, Jedynak BM, Bader JS: **Where have all the interactions gone? Estimating the coverage of two-hybrid protein interaction maps**. *PLoS Comput Biol* 2007, **3**(11):e214.

125. Han P, Zhang X, Norton RS, Feng ZP: **Predicting disordered regions in proteins based on decision trees of reduced amino acid composition**. *J Comput Biol* 2006, **13**(10):1723-1734.

126. Guo J, Lin Y: **TSSub: eukaryotic protein subcellular localization by extracting features from profiles**. *Bioinformatics* 2006, **22**(14):1784-1785.

127. Guo J, Pu X, Lin Y, Leung H: **Protein subcellular localization based on PSI-BLAST and machine learning**. *J Bioinform Comput Biol* 2006, **4**(6):1181-1195.

128. Huttenhower C, Hibbs M, Myers C, Troyanskaya OG: **A scalable method for integration and functional analysis of multiple microarray datasets**. *Bioinformatics* 2006, **22**(23):2890-2897.

129. Huttenhower C, Troyanskaya OG: **Bayesian data integration: a functional perspective**. *Comput Syst Bioinformatics Conf* 2006:341-351.

130. Bhasin M, Garg A, Raghava GP: **PSLpred: prediction of subcellular localization of bacterial proteins**. *Bioinformatics* 2005, **21**(10):2522-2524.

131. Bhasin M, Raghava GP: **ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST**. *Nucleic Acids Res* 2004, **32**(Web Server issue):W414-419.

132. Bhasin M, Raghava GP: **GPCRpred: an SVM-based method for prediction of families and subfamilies of G-protein coupled receptors**. *Nucleic Acids Res* 2004, **32**(Web Server issue):W383-389.

133. Bhasin M, Raghava GP: **Classification of nuclear receptors based on amino acid composition and dipeptide composition**. *J Biol Chem* 2004, **279**(22):23262-23266.

134. Garg A, Bhasin M, Raghava GP: **Support vector machine-based method for subcellular localization of human proteins using amino acid compositions, their order, and similarity search**. *J Biol Chem* 2005, **280**(15):14427-14432.

135. Sarda D, Chua GH, Li KB, Krishnan A: **pSLIP: SVM based protein subcellular localization prediction using multiple physicochemical properties**. *BMC Bioinformatics* 2005, **6**:152.

136. Yang ZR, Chou KC: **Bio-support vector machines for computational proteomics**. *Bioinformatics* 2004, **20**(5):735-741.

137. Hua S, Sun Z: **Support vector machine approach for protein subcellular localization prediction**. *Bioinformatics* 2001, **17**(8):721-728.

138. Reinhardt A, Hubbard T: **Using neural networks for prediction of the subcellular location of proteins**. *Nucleic Acids Res* 1998, **26**(9):2230-2236.

139. Hoehndorf R, Hardy NW, Osumi-Sutherland D, Tweedie S, Schofield PN, Gkoutos GV: **Systematic analysis of experimental phenotype data reveals gene functions**. *PLoS One* 2013, **8**(4):e60847.

140. Cozzetto D, Buchan DW, Bryson K, Jones DT: **Protein function prediction by massive integration of evolutionary analyses and multiple data sources**. *BMC Bioinformatics* 2013, **14 Suppl 3**:S1.

141. Sasidharan R, Nepusz T, Swarbreck D, Huala E, Paccanaro A: **GFam: a platform for automatic annotation of gene families**. *Nucleic Acids Res* 2012, **40**(19):e152.

142. Wilkins AD, Bachman BJ, Erdin S, Lichtarge O: **The use of evolutionary patterns in protein annotation**. *Curr Opin Struct Biol* 2012, **22**(3):316-325.

143. Frech C, Chen N: **Genome-wide comparative gene family classification**. *PLoS One* 2010, **5**(10):e13409.

144. Pandey G, Myers CL, Kumar V: **Incorporating functional inter-relationships into protein function prediction algorithms**. *BMC Bioinformatics* 2009, **10**:142.

145. Malmstrom L, Riffle M, Strauss CE, Chivian D, Davis TN, Bonneau R, Baker D: **Superfamily assignments for the yeast proteome through integration of structure prediction with the gene ontology**. *PLoS Biol* 2007, **5**(4):e76.

146. Schuemie M, Chichester C, Lisacek F, Coute Y, Roes PJ, Sanchez JC, Kors J, Mons B: **Assignment of protein function and discovery of novel nucleolar proteins based on automatic analysis of MEDLINE**. *Proteomics* 2007, **7**(6):921-931.

147. Zhu M, Gao L, Guo Z, Li Y, Wang D, Wang J, Wang C: **Globally predicting protein functions based on co-expressed protein-protein interaction networks and ontology taxonomy similarities**. *Gene* 2007, **391**(1-2):113-119.

148. Hayete B, Bienkowska JR: **Gotrees: predicting go associations from protein domain composition using decision trees**. *Pac Symp Biocomput* 2005:127-138.

149. Martin DM, Berriman M, Barton GJ: **GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes**. *BMC Bioinformatics* 2004, **5**:178.

150. Jensen LJ, Gupta R, Staerfeldt HH, Brunak S: **Prediction of human protein function according to Gene Ontology categories**. *Bioinformatics* 2003, **19**(5):635-642.

151. Deng M, Zhang K, Mehta S, Chen T, Sun F: **Prediction of protein function using protein-protein interaction data**. *Proc IEEE Comput Soc Bioinform Conf* 2002, **1**:197-206.

152. Claros MG, Vincens P: **Computational method to predict mitochondrially imported proteins and their targeting sequences**. *Eur J Biochem* 1996, **241**(3):779-786.

153. Bell MJ, Gillespie CS, Swan D, Lord P: **An approach to describing and analysing bulk biological annotation quality: a case study using UniProtKB**. *Bioinformatics* 2012, **28**(18):i562-i568.

154. Skunca N, Altenhoff A, Dessimoz C: **Quality of computationally inferred gene ontology annotations**. *PLoS Comput Biol* 2012, **8**(5):e1002533.

155. Schnoes AM, Brown SD, Dodevski I, Babbitt PC: **Annotation error in public databases: misannotation of molecular function in enzyme superfamilies**. *PLoS Comput Biol* 2009, **5**(12):e1000605.

156. Jones CE, Brown AL, Baumann U: **Estimating the annotation error rate of curated GO database sequence annotations**. *BMC Bioinformatics* 2007, **8**:170.

157. Song J, Singh M: **From hub proteins to hub modules: the relationship between essentiality and centrality in the yeast interactome at different scales of organization**. *PLoS Comput Biol* 2013, **9**(2):e1002910.

158. Banky D, Ivan G, Grolmusz V: **Equal opportunity for low-degree network nodes: a PageRank-based method for protein target identification in metabolic graphs**. *PLoS One* 2013, **8**(1):e54204.

159. Choura M, Rebai A: **Structural analysis of hubs in human NR-RTK network**. *Biol Direct* 2011, **6**:49.

160. Zhang X, Xu J, Xiao WX: **A new method for the discovery of essential proteins**. *PLoS One* 2013, **8**(3):e58763.

161. Cho YR, Zhang A: **Identification of functional hubs and modules by converting interactome networks into hierarchical ordering of proteins**. *BMC Bioinformatics* 2010, **11 Suppl 3**:S3.

162. Hsing M, Byler KG, Cherkasov A: **The use of Gene Ontology terms for predicting highly-connected 'hub' nodes in protein-protein interaction networks**. *BMC Syst Biol* 2008, **2**:80.

163. Manna B, Bhattacharya T, Kahali B, Ghosh TC: **Evolutionary constraints on hub and non-hub proteins in human protein interaction network: insight from protein connectivity and intrinsic disorder**. *Gene* 2009, **434**(1-2):50-55.

164. Higurashi M, Ishida T, Kinoshita K: **Identification of transient hub proteins and the possible structural basis for their multiple interactions**. *Protein Sci* 2008, **17**(1):72-78.

165. Lin CY, Chin CH, Wu HH, Chen SH, Ho CW, Ko MT: **Hubba: hub objects analyzer--a framework of interactome hubs identification for network biology**. *Nucleic Acids Res* 2008, **36**(Web Server issue):W438-443.

166. Chang X, Xu T, Li Y, Wang K: **Dynamic modular architecture of protein-protein interaction networks beyond the dichotomy of 'date' and 'party' hubs**. *Sci Rep* 2013, **3**:1691.

167. Goel A, Wilkins MR: **Dynamic hubs show competitive and static hubs non-competitive regulation of their interaction partners**. *PLoS One* 2012, **7**(10):e48209.

168. Goel A, Li SS, Wilkins MR: **Four-dimensional visualisation and analysis of protein-protein interaction networks**. *Proteomics* 2011, **11**(13):2672-2682.

169. Cukuroglu E, Gursoy A, Keskin O: **HotRegion: a database of predicted hot spot clusters**. *Nucleic Acids Res* 2012, **40**(Database issue):D829-833.

170. Guney E, Tuncbag N, Keskin O, Gursoy A: **HotSprint: database of computational hot spots in protein interfaces**. *Nucleic Acids Res* 2008, **36**(Database issue):D662-666.

171. Keskin O, Tuncbag N, Gursoy A: **Characterization and prediction of protein interfaces to infer protein-protein interaction networks**. *Curr Pharm Biotechnol* 2008, **9**(2):67-76.

172. Tuncbag N, Gursoy A, Keskin O: **Identification of computational hot spots in protein interfaces: combining solvent accessibility and inter-residue potentials improves the accuracy**. *Bioinformatics* 2009, **25**(12):1513-1520.

173. Tuncbag N, Keskin O, Gursoy A: **HotPoint: hot spot prediction server for protein interfaces**. *Nucleic Acids Res* 2010, **38**(Web Server issue):W402-406.

174. Tuncbag N, Keskin O, Nussinov R, Gursoy A: **Fast and accurate modeling of protein-protein interactions by combining template-interface-based docking with flexible refinement**. *Proteins* 2012, **80**(4):1239-1249.

175. Wang H, Zheng H: **Correlation of genomic features with dynamic modularity in the yeast interactome: a view from the structural perspective**. *IEEE Trans Nanobioscience* 2012, **11**(3):244-250.

176. Zhao L, Hoi SC, Wong L, Hamp T, Li J: **Structural and functional analysis of multi-interface domains**. *PLoS One* 2012, **7**(12):e50821.

177. Bhardwaj N, Abyzov A, Clarke D, Shou C, Gerstein MB: **Integration of protein motions with molecular networks reveals different mechanisms for permanent and transient interactions**. *Protein Sci* 2011, **20**(10):1745-1754.

178. Dasgupta B, Nakamura H, Kinjo AR: **Distinct roles of overlapping and non-overlapping regions of hub protein interfaces in recognition of multiple partners**. *J Mol Biol* 2011, **411**(3):713-727.

179. Mirzarezaee M, Araabi BN, Sadeghi M: **Features analysis for identification of date and party hubs in protein interaction network of Saccharomyces Cerevisiae**. *BMC Syst Biol* 2010, **4**:172.

180. Agarwal S, Deane CM, Porter MA, Jones NS: **Revisiting date and party hubs: novel approaches to role assignment in protein interaction networks**. *PLoS Comput Biol* 2010, **6**(6):e1000817.

181. Batada NN, Reguly T, Breitkreutz A, Boucher L, Breitkreutz BJ, Hurst LD, Tyers M: **Still stratus not altocumulus: further evidence against the date/party hub distinction**. *PLoS Biol* 2007, **5**(6):e154.

182. Batada NN, Reguly T, Breitkreutz A, Boucher L, Breitkreutz BJ, Hurst LD, Tyers M: **Stratus not altocumulus: a new view of the yeast protein interaction network**. *PLoS Biol* 2006, **4**(10):e317.

183. Ekman D, Light S, Bjorklund AK, Elofsson A: **What properties characterize the hub proteins of the protein-protein interaction network of Saccharomyces cerevisiae?** *Genome Biol* 2006, **7**(6):R45.

184. Maizels N, Gray LT: **The G4 genome**. *PLoS Genet* 2013, **9**(4):18.

185. Eddy J, Maizels N: **Conserved elements with potential to form polymorphic G-quadruplex structures in the first intron of human genes**. *Nucleic Acids Res* 2008, **36**(4):1321-1333.

186. Chen Y, Yang D: **Sequence, stability, and structure of G-quadruplexes and their interactions with drugs**. *Curr Protoc Nucleic Acid Chem* 2012, **Chapter 17**:Unit17 15.

187. Brooks TA, Hurley LH: **Targeting MYC Expression through G-Quadruplexes**. *Genes Cancer* 2010, **1**(6):641-649.

188. Simonsson T: **G-quadruplex DNA structures--variations on a theme**. *Biol Chem* 2001, **382**(4):621-628.

189. Du Z, Zhao Y, Li N: **Genome-wide colonization of gene regulatory elements by G4 DNA motifs**. *Nucleic Acids Res* 2009, **37**(20):6784-6798.

190. Huppert JL: **Four-stranded DNA: cancer, gene regulation and drug development**. *Philos Trans A Math Phys Eng Sci* 2007, **365**(1861):2969-2984.

191. Huppert JL: **Hunting G-quadruplexes**. *Biochimie* 2008, **90**(8):1140-1148.

192. Huppert JL: **Four-stranded nucleic acids: structure, function and targeting of G-quadruplexes**. *Chem Soc Rev* 2008, **37**(7):1375-1384.

193. Huppert JL: **Structure, location and interactions of G-quadruplexes**. *FEBS J* 2010, **277**(17):3452-3458.

194. Huppert JL, Balasubramanian S: **G-quadruplexes in promoters throughout the human genome**. *Nucleic Acids Res* 2007, **35**(2):406-413.

195. Huppert JL, Bugaut A, Kumari S, Balasubramanian S: **G-quadruplexes: the beginning and end of UTRs**. *Nucleic Acids Res* 2008, **36**(19):6260-6268.

196. Endoh T, Kawasaki Y, Sugimoto N: **Translational halt during elongation caused by G-quadruplex formed by mRNA**. *Methods* 2013.

197. Douarre C, Mergui X, Sidibe A, Gomez D, Alberti P, Mailliet P, Trentesaux C, Riou JF: **DNA damage signaling induced by the G-quadruplex ligand 12459 is modulated by PPM1D/WIP1 phosphatase**. *Nucleic Acids Res* 2013, **41**(6):3588-3599.

198. Reinhold WC, Mergny JL, Liu H, Ryan M, Pfister TD, Kinders R, Parchment R, Doroshow J, Weinstein JN, Pommier Y: **Exon array analyses across the NCI-60 reveal potential regulation of TOP1 by transcription pausing at guanosine quartets in the first intron**. *Cancer Res* 2010, **70**(6):2191-2203.

199. Beaudoin JD, Perreault JP: **Exploring mRNA 3'-UTR G-quadruplexes: evidence of roles in both alternative polyadenylation and mRNA shortening**. *Nucleic Acids Res* 2013, **41**(11):5898-5911.

200. Yu H, Braun P, Yildirim MA, Lemmens I, Venkatesan K, Sahalie J, Hirozane-Kishikawa T, Gebreab F, Li N, Simonis N *et al*: **High-quality binary protein interaction map of the yeast interactome network**. *Science* 2008, **322**(5898):104-110.

201. Dreze M, Monachello D, Lurin C, Cusick ME, Hill DE, Vidal M, Braun P: **High-quality binary interactome mapping**. *Methods Enzymol* 2010, **470**:281-315.

202. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P *et al*: **A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae**. *Nature* 2000, **403**(6770):623-627.

203. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y: **A comprehensive two-hybrid analysis to explore the yeast protein interactome**. *Proc Natl Acad Sci U S A* 2001, **98**(8):4569-4574.

204. Walhout AJ, Vidal M: **High-throughput yeast two-hybrid assays for large-scale protein interaction mapping**. *Methods* 2001, **24**(3):297-306.

205. Venkatesan K, Rual JF, Vazquez A, Stelzl U, Lemmens I, Hirozane-Kishikawa T, Hao T, Zenkner M, Xin X, Goh KI *et al*: **An empirical framework for binary interactome mapping**. *Nat Methods* 2009, **6**(1):83-90.

206. Das J, Yu H: **HINT: High-quality protein interactomes and their applications in understanding human disease**. *BMC Syst Biol* 2012, **6**(1):92.

207. Bertin N, Simonis N, Dupuy D, Cusick ME, Han JD, Fraser HB, Roth FP, Vidal M: **Confirmation of organized modularity in the yeast interactome**. *PLoS Biol* 2007, **5**(6):e153.

208. Afridi TH, Khan A, Lee YS: **Mito-GSAAC: mitochondria prediction using genetic ensemble classifier and split amino acid composition**. *Amino Acids* 2011.

209. Tsai CJ, Ma B, Nussinov R: **Protein-protein interaction networks: how can a hub protein bind so many different partners?** *Trends Biochem Sci* 2009, **34**(12):594-600.

CHAPTER 2

EXPLORING INCONSISTENCIES IN GENOME-WIDE PROTEIN FUNCTION ANNOTATIONS: A MACHINE LEARNING APPROACH

Modified from a paper published in *BMC Bioinformatics*
8:284 (3 August 2007)

Carson Andorf, Drena Dobbs, Vasant Honavar

Abstract

Background

Incorrectly annotated sequence data are becoming more commonplace as databases increasingly rely on automated techniques for annotation. Hence, there is an urgent need for computational methods for checking consistency of such annotations against independent sources of evidence and detecting potential annotation errors. We show how a machine learning approach designed to automatically predict a protein's Gene Ontology (GO) functional class can be employed to identify potential gene annotation errors.

Results

In a set of 211 previously annotated mouse protein kinases, we found that 201 of the GO annotations returned by AmiGO appear to be *inconsistent* with the UniProt functions assigned to their human counterparts. In contrast, 97% of the predicted annotations generated using a machine learning approach were *consistent* with the UniProt annotations of the human counterparts, as well as with available annotations for these mouse protein kinases in the Mouse Kinome database.

Conclusions

We conjecture that most of our predicted annotations are, therefore, correct and suggest that the machine learning approach developed here could be routinely used to

detect potential errors in GO annotations generated by high-throughput gene annotation projects.

Authors from the original publication (Okazaki et al.: Nature 2002, 420:563-73) have provided their response to Andorf et al, directly following the correspondence.

Background

As more genomic sequences become available, functional annotation of genes presents one of the most important challenges in bioinformatics. Because experimental determination of protein structure and function is expensive and time-consuming, there is an increasing reliance on automated approaches to assignment of Gene Ontology (GO) [1] functional categories to protein sequences. An advantage of such automated methods is that they can be used to annotate hundreds or thousands of proteins in a matter of minutes, which makes their use especially attractive - if not unavoidable - in large-scale genome-wide annotation efforts.

Most automated approaches to protein function annotation rely on transfer of annotations from previously annotated proteins, based on sequence or structural similarity. Such annotations are susceptible to several sources of error, including errors in the original annotations from which new annotations are inferred, errors in the algorithms, bugs in the programs or scripts used to process the data, clerical errors on the part of human curators, among others. The effect of such errors can be magnified because they can propagate from one set of annotated sequences to another through widespread use of automated techniques for genome-wide functional annotation of proteins [2-5]. Once introduced, such errors can go undetected for a long time. Because of the increasing reliance of biologists and computational biologists on reliable functional annotations for formulation of hypotheses,

design of experiments, and interpretation of results, incorrect annotations can lead to wasted effort and erroneous conclusions. Computational approaches to checking automatically inferred annotations against independent sources of evidence and detecting potential annotation errors offer a potential solution to this problem [6-11].

Previous work of several groups, including our own [12-19] has demonstrated the usefulness of machine learning approaches to assigning putative functions to proteins based on the amino acid sequence of the proteins. On the specific problem of predicting the catalytic activity of proteins from amino acid sequence, we showed that machine learning approaches outperform methods based on sequence homology [13]. This is especially true when sequence identity among proteins with a specified function is below 10%; the accuracy of predictions by our HDTree classifier was 8%-16% better than that of PSI-BLAST [13]. The discriminatory power of machine learning approaches thus suggests they should be valuable for detecting potential annotation errors in functional genomics databases.

Here we demonstrate that a machine learning approach, designed to predict GO functional classifications for proteins, can be used to identify and correct potential annotation errors. In this study, we focused on a small but clinically important subset of protein kinases, for which we "stumbled upon" potential annotation errors while evaluating the performance of protein function classification algorithms. We chose a set of protein kinases categorized under the GO class GO0004672, Protein Kinase Activity, which includes proteins with serine/threonine (Ser/Thr) kinase activity (GO0004674) and tyrosine (Tyr) kinase activity (GO0004713). Post-translational modification of proteins by phosphorylation plays an important regulatory role in virtually every signaling pathway in

eukaryotic cells, modulating key biological processes associated with development and diseases including cancer, diabetes, hyperlipidemia and inflammation [20,21]. It is natural to expect that such well studied and functionally significant families of protein kinases are correctly annotated by genome-wide annotation efforts.

Results

The initial aim of our experiments was to evaluate the effectiveness of machine learning approaches to automate sequence-based classification of protein kinases into subfamilies. Because both the Ser/Thr and Tyr subfamilies contain highly divergent members, some of which share less than 10% sequence identity with other members, they offer a rigorous test case for evaluating the potential general utility of this approach. Previously, we developed HDTree [13], a two-stage approach that combines a classifier based on amino acid *k*-gram composition of a protein sequence, with a classifier that relies on transfer of annotation from PSI-BLAST hits (see **Methods** for details). A protein kinase classifier was trained on a set of 330 human protein kinases from the Ser/Thr protein kinase (GO0004674) and Tyr protein kinase (GO0004713) functional classes based on direct and indirect annotations assigned by AmiGO [22], a valuable and widely used tool for retrieving GO functional annotations of proteins. Performance of the classifier was evaluated, using 10-fold cross-validation, on two datasets: i) the dataset of 330 *human* protein kinases, and ii) a dataset of 244 *mouse* protein kinases drawn from the same GO functional classes. The initial datasets were not filtered based on evidence codes or sequence identity cutoffs.

Using the AmiGO annotations as reference, the resulting HDTree classifier correctly distinguished between Ser/Thr kinases and Tyr kinases in the human kinase

dataset with an overall accuracy of 89.1% and a kappa coefficient of 0.76. In striking contrast, the accuracy of the classifier on the mouse kinase dataset was only 15.1%; the correlation between the GO functional categories predicted by the classifier and the AmiGO reference labels was an alarming -0.40: 72 of the 244 mouse kinases were classified as Ser/Thr kinases, 105 as Tyr kinases, and 67 as "dual specificity" kinases (belonging to both GO0004674 and GO0004713 classes) (see **Table 1).**

Assuming the AmiGO annotations were correct, these results suggested that either this particular machine learning approach is extremely ineffective for classifying mouse protein labels, or that human and mouse protein kinases have so little in common that a classifier trained on the human proteins is doomed to fail miserably on the mouse proteins. In light of the demonstrated effectiveness of machine learning approaches on a broad range of classification tasks that arise in bioinformatics [23], and well-documented high degree of homology between human and mouse proteins [24], neither of these conclusions seemed warranted. Could this discrepancy be explained by the AmiGO annotations for mouse protein kinases? We proceeded to investigate this possibility.

A comparison of the distribution of Ser/Thr, Tyr, and dual specificity kinases in mouse versus human **(Figure 1a)** reveals a striking discordance: based on AmiGO annotations, mouse has many more Tyr and dual specificity kinases than human and only 40% as many Ser/Thr protein kinases. In contrast, as explained below, the fractions of Ser/Thr, Tyr, and dual specificity kinases based on UniProt annotations are very similar in mouse and human **(Figure 1b).** Furthermore, the predictions of our two-stage machine learning algorithm are in good agreement with the UniProt annotations for both human and mouse protein kinases **(Figures 1b and 1c).**

Examination of the GO evidence codes for the mouse protein kinases revealed that 211 of 244 mouse protein kinases included the evidence code “RCA,” "inferred from reviewed computational analysis" [see **Appendix C**], indicating that these annotations had been assigned using computational tools and reviewed by a human curator before being deposited in the database used by AmiGO. Notably, 28 of 33 (85%) mouse protein kinases with an evidence code other than RCA (e.g., "inferred from direct assay") were assigned "correct" labels, relative to the AmiGO reference, by the classifier trained on the human protein kinase data. Each of the 211 proteins with the RCA evidence code had at least one annotation that could be traced to the FANTOM Consortium and RIKEN Genome Exploration Research Group [25], a source of protein function annotations in the Mouse Genome Database (MGD) [24]. To further examine each of these 211 mouse protein kinases, we used the gene IDs obtained from AmiGO to extract information about each protein from UniProt [26]. We searched the UniProt records for mention of "Serine/Threonine" or "Tyrosine" (or their synonyms) in fields for protein name, synonyms, references, similarity, keywords, or function, and created a dataset in which each protein kinase had one of the corresponding UniProt labels: “Ser/Thr kinase,” "Tyr kinase,” or "dual specificity kinase" if both keywords were found. Results of our comparison of UniProt labels with AmiGO annotations for each class in this dataset of 211 mouse protein kinases are shown in **Figure 2a:** for 201 of the 211 cases with an RCA annotation code, the UniProt and AmiGO labels were inconsistent. Results of our comparison are shown in **Table 2** [see **Appendix D and E**].

This result led us to test the ability of the HDTree classifier trained on the human kinase dataset to correctly predict the family classifications for proteins in the mouse

kinase dataset, this time using UniProt instead of AmiGO annotations as the "correct" reference labels. Strikingly, the classifier (trained on the human kinase dataset) achieved a classification accuracy of 97.2%, with a kappa coefficient of 0.93, on the mouse kinase dataset. As illustrated in **Figure 2b,** the classifier correctly classified 205 out of the 211 mouse kinases into Ser/Thr, Tyr or dual specificity classes compared with 10 out of 211 for AmiGO. A direct comparison of classifiers based on UniProt annotations and AmiGO annotations can be seen in **Table 3**. This performance actually exceeded that of the same classifier tested on the human kinase dataset, for which an overall classification accuracy of 89.1%, with a kappa coefficient of 0.76, was obtained [see **Table 1** and see **Appendix A**

The HDTree method uses a decision tree built from the output from eight individual classifiers. A decision tree is built by selecting, in a greedy fashion, the individual classifier that provides the maximum information about the class label at each step, [27]. By examining the decision tree, it is easy to identify the individual classifiers that have the greatest influence on the classification. In the case of the kinase datasets used in this study, the classifiers constructed by the NB(k) algorithms using trimers and quadmers, NB(3) and NB(4), were found to provide the most information regarding class labels. This suggests that the biological "signals" detected by these classifiers are groups of 3-4 residues, not necessarily contiguous in the primary amino acid sequence, but often in close proximity or interacting within three-dimensional structures to form functional sites (e.g., catalytic sites, binding sites), an idea supported by the results of our previous work [13]. Notably, the NB(3) and NB(4) classifiers appear to contribute more to the ability to distinguish proteins with very closely related enzymatic activities than PSI-

BLAST. The PSI-BLAST results influenced the final classification, however, when the NB(3) and NB(4) classifiers disagreed on the classification.

Discussion

Examination of the Mouse Kinome Database [28] reveals that the majority of annotated mouse kinases have a human ortholog with sequence identity >90% [see **Appendix F and G**]. Results summarized in **Figures 1 and 2**, together with the assumption that the relative proportions of Ser/Thr, Tyr and dual specificity kinases should not be significant different in human and mouse, led us to conclude that UniProt derived annotations are more likely to be correct than those returned by AmiGO for this group of mouse protein kinases with the RCA evidence code**.** We have shared our findings with the Mouse Genome Database [24], which is in the process of identifying and rectifying the source of potential problems with these annotations.

Identifying potential annotation errors in a specific dataset such as the mouse kinase dataset solves only a part of a larger problem. Because annotation errors can propagate across multiple databases through the widespread - and often necessary - use of information derived from available annotations, it is important to track and correct errors in other databases that rely on the erroneous source. For example, using AmiGO, we retrieved 136 rat protein kinases for which annotations had been transferred from mouse protein kinases based on homology (indicated by the evidence code "ISS," 'inferred from sequence or structural similarity') with one of the 201 erroneously annotated mouse protein kinases. Examination of the UniProt records for these 136 rat protein kinases revealed that 94 of those labeled as "Ser/Thr" kinases by UniProt had AmiGO annotations of "Tyr" or "dual

specificity" kinase, and 42 of those labeled as "Tyr" kinases by UniProt had AmiGO annotations of "Ser/Thr" or "dual specificity" kinase [see **Appendix B and H**].

A recent study found that the GO annotations with ISS (inferred from sequence or structural similarity) evidence code could have error rates as high as 49% [29]. This argues for the development and large-scale application of a suite of computational tools for identifying and flagging potentially erroneous annotations in functional genomics databases. Our results suggest the utility of including machine learning methods among such a suite of tools. Large-scale application of machine learning tools to protein annotation has to overcome several challenges. Because many proteins are multi-functional, classifiers should be able to assign a sequence to multiple, not mutually exclusive, classes (the *multi label* classification problem), or more generally, to a subset of nodes in a directed-acyclic graph, e.g., the GO hierarchy, (the *structured label* classification problem). Fortunately, a number of research groups have developed machine learning algorithms for multi-label and structured label classification and demonstrated their application in large-scale protein function classification [30-33]. We can draw on recent advances in machine learning methods for hierarchical multi-label classification of large sequence datasets to adapt our method to work in such a setting. For example, a binary classifier can be trained to determine membership of a given sequence in the class represented by each node of the GO hierarchy, starting with the root node (to which trivially the entire dataset is assigned). Binary classifiers at each node in the hierarchy can then be trained recursively, focusing on the dataset passed to that node from its parent(s) in the GO hierarchy.

In this study, we have limited our attention to *sequence-based* machine learning methods for annotation of protein sequences. With the increasing availability of other types of data (protein structure, gene expression profiles, etc.), there is a growing interest in machine learning and other computational methods for genome-wide prediction of protein function using diverse types of information [34-39]. Such techniques can be applied in a manner similar to our use of sequence-based machine learning to identify potentially erroneous annotations in existing databases.

Conclusion

The increasing reliance on automated tools in genome-wide functional annotation of proteins has led to a corresponding increase in the risk of propagation of annotation errors across genome databases. Short of direct experimental validation of every annotation, it is impossible to ensure that the annotations are accurate. The results presented here and in recent related studies [6-11] underscore the need for checking the consistency of annotations against multiple sources of information and carefully exploring the sources of any detected inconsistencies. Addressing this problem requires the use of machine readable metadata that capture precise descriptions of all data sources, data provenance, background assumptions, and algorithms used to infer the derived information. There is also a need for computational tools that can detect annotation inconsistencies and alert data sources and their users regarding potential errors. Expertly curated databases such as the Mouse Genome Database are indispensable for research in functional genomics and systems biology, and it is important to emphasize that several measures for finding and correcting inconsistent annotations are already in place at MGD [24]. The present study suggests that additional measures, especially in the case of protein

annotations with RCA evidence code, can further increase the reliability of these valuable resources.

Methods

Classification Strategy

We constructed an HDTree binary classifier, described below, for each of the three kinase families. The first two kinase families correspond to the GO labels GO0004674 (Ser/Thr kinases) or GO0004713 (Tyr kinases) but not both; the third family corresponds to dual-specificity kinases that belong to both GO0004674 and GO0004713. Classifier #1 distinguishes between Ser/Thr kinases and the rest (Tyr and dual-specificity kinases). Similarly, classifier #2 distinguishes between Tyr kinases and the rest (Ser/Thr and dual specificity kinases). Classifier #3 distinguishes dual-specificity kinases from the rest (those with only Ser/Thr or Tyr activity), based on the predictions generated by classifier #1 and classifier #2 as follows: If only classifier #1 generates a positive prediction, the corresponding sequence is classified as (exclusively) a Ser/Thr kinase. If only classifier #2 generates a positive prediction, the corresponding sequence is classified as (exclusively) Tyr kinase. If both classifiers generate a positive prediction or if both classifiers generate a negative prediction, the corresponding sequence is classified as a dual-specificity kinase. We interpret the disagreement between the classifiers as indicative of signaling evidence that the protein is neither exclusively Ser/Thr nor Tyr, and hence, likely to have dual specificity. More sophisticated evidence combination methods could be used instead. However, this simple technique worked sufficiently well in the case of this dataset (see **Table 4**).

HDTree Method

As noted above, an HDTree binary classifier [13] is constructed for each of the three kinase families. Each HDTree binary classifier is a decision tree classifier that assigns a class label to a target sequence based on the binary class labels output by the Naïve Bayes, NB k-gram, NB(k), and PSI-BLAST classifiers for the corresponding kinase families. Because there are eight classifiers Naïve Bayes, NB 2-gram , NB 3-gram , NB 4-gram , NB(2), NB(3), NB(4), and PSI-BLAST, the input to a HDTree binary classifier for each kinase family consists of an 8-tuple of class labels assigned to the sequence by the corresponding 8 classifiers. The output of the HDTree classifier for kinase family *c* is a binary class label (1 if the predicted class is *c*; 0 otherwise). Thus, each HDTree classifier is a decision tree classifier that is trained to predict the binary class label of a query sequence based on the 8-tuple of class labels predicted by the eight individual classifiers. Because HDTree is a decision tree, it is easy to determine which individual classifier(s) provided the most information in regards to the predicted class label. In the resulting tree, nodes near the top of the tree provided the most information about the class label. Thus, HDTree can also facilitate identification of the determinative biological sequence signals. We used the Weka version 3.4.4 implementation [40] (J4.8) of the C4.5 decision tree learning algorithm [27].

We describe below, a class of probabilistic models for sequence classification.

Classification Using a Probabilistic Model

We start by introducing the general procedure for building a classifier from a probabilistic generative model.

Suppose we can specify a probabilistic model α for sequences defined over some alphabet $\sum$ (which in our case is the 20-letter amino acid alphabet). The model α

specifies for any sequence $\overline{S} = s_1,...,s_n$, the probability $P_\alpha(\overline{S} = s_1,...,s_n)$ of generating the sequence $\overline{S}$. Suppose we assume that sequences belonging to class c_j are generated by the probabilistic generative model $\alpha(c_j)$.

Then, $P_\alpha(\overline{S} = s_1,...,s_n \mid c_j) = P_{\alpha(c_j)}(\overline{S} = s_1,...,s_n)$ is the probability of $\overline{S}$ given that the class is c_j. Therefore, given the probabilistic generative model for each of the classes in *C* (the set of possible mutually exclusive class labels) for sequences over the alphabet $\sum$, we can compute the most likely class label $c(\overline{S})$ for any given sequence $\overline{S} = s_1,...,s_n$ as follows: $c(\overline{S}) = \arg\max_{c_j \in C} P_\alpha(\overline{S} = s_1,...,s_n \mid c_j)P(c_j)$. Hence, the goal of a machine learning algorithm for sequence classification is to estimate the parameters that describe the corresponding probabilistic models from data. Different classifiers differ with regard to their ability to capture the dependencies among the elements of a sequence.

In what follows, we use the following notations.

$n = |\overline{S}|$ = the length of the sequence $|\overline{S}|$

k = the size of the k-gram (k-mer) used in the model

s_i = the i^{th} element in the sequence $\overline{S}$

c_j = the j^{th} class in the class set C

Naïve Bayes Classifier

The Naïve Bayes classifier assumes that each element of the sequence is independent of the other elements given the class label. Consequently,

$$c(\overline{S}) = \arg\max_{c_j \in C} P_\alpha \prod_{i=1}^{n} P_\alpha(s_1 \mid c_j) \cdots P_\alpha(s_n \mid c_j) P(c_j)$$

Note that the Naive Bayes classifier for sequences treats each sequence as though it were simply a *bag* of letters. We now consider two Naive Bayes-like models based on *k* -grams.

Naïve Bayes *k*-grams Classifier

The Naive Bayes *k*-grams (NB *k*-grams) [12,13,41] method uses a sliding a window of size *k* along each sequence to generate a *bag* of *k*-grams representation of the sequence. Much like in the case of the Naive Bayes classifier described above treats each *k*-gram in the bag to be independent of the others given the class label for the sequence. Given this probabilistic model, the standard method for classification using a probabilistic model can be applied. The probability model associated with Naïve Bayes *k*-grams:

$$P_\alpha(\overline{S} = [S_1 = s_1, \ldots, S_n = s_n]) = \arg\max_{c_j \in C} P_\alpha \prod_{i=1}^{n-k+1} P_\alpha(S_i = s_i, \ldots, S_{i+k-1} = s_{i+k-1} \mid c_j) P(c_j)$$

A problem with the NB *k*-grams approach is that successive *k*-grams extracted from a sequence share *k*-1 elements in common. This grossly and systematically violates the independence assumption of Naive Bayes.

Naïve Bayes (*k*)

We introduce the Naive Bayes (*k*) or the NB(*k*) model [12,13,41] to explicitly model the dependencies that arise as a consequence of the overlap between successive *k*-grams in a sequence. We represent the dependencies in a graphical form by drawing edges between the elements that are directly dependent on each other.

Using the Junction Tree Theorem for graphical models [42], it can be proved [41] that the correct probability model α that captures the dependencies among overlapping *k*-grams is given by:

$$P_\alpha(\overline{S} = [S_1 = s_1, ..., S_n = s_n]) = \frac{\prod_{i=1}^{n-k+1} P_\alpha(S_i = s_i, ..., S_{i+k-1} = s_{i+k-1})}{\prod_{i=2}^{n-k+1} P_\alpha(S_i = s_i, ..., S_{i+k-2} = s_{i+k-2})}$$

Now, given this probabilistic model, we can use the standard approach to classification given a probabilistic model. It is easily seen that when $k = 1$, Naive Bayes 1-grams as well as Naive Bayes (1) reduce to the Naive Bayes model.

The relevant probabilities required for specifying the above models can be estimated using standard techniques for estimation of probabilities using Laplace estimators [43].

PSI-Blast

We used PSI-BLAST (from the latest release of BLAST) [44] to construct a binary classifier for each class. We used the binary class label predicted by the PSI-BLAST based classifier as an additional input to our HD-Tree classifier. Given a query sequence to be classified, we use PSI-BLAST to compare the query sequence against a reference protein sequence database, i.e., the training set used in the cross-validation process. We run PSI-BLAST with the query sequence against the reference database. We assign to the query sequence the functional class of the top scoring hit (the sequence with the lowest e-value) from the PSI-BLAST results. The resulting binary prediction of the PSI-BLAST classifier for class *c* is 1 if the class label for the top scoring hit is *c*. Otherwise, it is 0. An e-value cut-off of 0.0001 was used for PSI-BLAST, with all other parameters set to their default values.

Performance Evaluation

The performance measures [45] used to evaluate each of the different classifiers trained using machine learning algorithms are summarized in **Tables 5 and 6**.

Authors' contributions

CA conceived of and designed the study, carried out the data analysis and visualization, developed the Java computer code, and drafted the manuscript. DD and VH contributed to the design of the study, analysis and interpretation of results, and writing of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The authors wish to thank Masaaki Furuno, David Hill, Judith Blake, Richard Baldarelli, Piero Carninci, Yoshihide Hayashizaki and the other members of Mouse Genome Informatics, the FANTOM2 project, and AmiGO. Their work has provided invaluable resources, data, and tools to the public. We appreciate their prompt attention to the potential errors identified in this work (among thousands of correctly annotated proteins). We also would like to thank Shankar Subramaniam of the University of California, San Diego and Pierre Baldi of the University of California, Irvine for helpful comments on an earlier draft of this paper. This research was supported in part by grants from the National Science Foundation (0219699) and the National Institutes of Health (GM066387) to Vasant Honavar and Drena Dobbs. Carson Andorf has been supported in part by a fellowship funded by an Integrative Graduate Education and Research Training (IGERT) award (9972653) from the National Science Foundation. The authors are grateful to members of their research groups for helpful comments throughout the progress of this research.

References

1. The Gene Ontology Consortium: **Gene ontology: tool for the unification of biology.** *Nature Genet* 2000, **25:** 25–29.

2. Doerks T, Bairoch A, Bork P. **Protein annotation: detective work for function prediction.** *Trends Genet* 1998. **14**: 248–250

3. Bork P, Koonin EV. **Predicting functions from protein sequences--where are the bottlenecks?** *Nat Genet* 1998, 18(4):313-318.

4. Gilks WR, Audit B, de Angelis D, Tsoka S, Ouzounis CA. **Percolation of annotation errors through hierarchically structured protein sequence databases.** *Math Biosci* 2005, **193(2)**:223-234.

5. Gilks WR, Audit B, De Angelis D, Tsoka S, Ouzounis CA. **Modeling the percolation of annotation errors in a database of protein sequences**. *Bioinformatics* 2002 **18**: 1641-1649.

6. Naumoff DG, Xu Y, Glansdorff N, Labedan B: **Retrieving sequences of enzymes experimentally characterized but erroneously annotated : the case of the putrescine carbamoyltransferase.** *BMC Genomics* 2004, **5**:52.

7. Green ML, Karp PD: **Genome annotation errors in pathway databases due to semantic ambiguity in partial EC numbers.** *Nucleic Acids Res* 2005, **33**: 4035-4039.

8. Dolan ME, Ni L, Camon E, Blake JA: **A procedure for assessing GO annotation consistency.** *Bioinformatics 2005*, **21:** 136-143.

9. Park YR, Park CH, Kim JH: **GOChase: correcting errors from gene ontology-based annotations for gene products.** *Bioinformatics* 2005, **21**: 829-831.

10. Devos D, Valencia A. **Practical limits of function prediction.** *Proteins* 2000, **41(1)**: 98-107.

11. Levy ED, Ouzounis CA, Gilks WR, Audit B. **Probabilistic annotation of protein sequences based on functional classifications**. *BMC Bioinformatics* 2005, **6**: 302.

12. Andorf C, Silvescu A, Dobbs D, Honavar V: **Learning classifiers for assigning protein sequences to gene ontology functional families.** *In: Fifth Int Conf Knowledge Based Computer Systems 2004,* India: 256-265. [http://www.cs.iastate.edu/~honavar/Papers/nbk.pdf].

13. Andorf C, Silvescu A, Dobbs D, Honavar V. **Learning classifiers for assigning protein sequences to Gene Ontology functional families: combining of function annotation using sequence homology with that based on amino acid k-gram composition yields more accurate classifiers than either of the**

individual approaches. Department of Computer Science, Iowa State University, 2004. [http://www.cs.iastate.edu/~andorfc/hdtree/HDtree2006.pdf].

14. Ben-Hur A and Brutlag D. **Remote homology detection: a motif based approach.** *Bioinformatics* 2003, **19:** i26-i33.

15. Hayete B, Bienkowska JR. **Gotrees: predicting go associations from protein domain composition using decision trees**. *Pac Symp Biocomput*. 2005:127-138.

16. Martin DM, Berriman M, Barton GJ. **GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes**. *BMC Bioinformatics.* 2004, **5**:178.

17. Murvai J, Vlahovicek K, Szepesvari C, Pongor S. **Prediction of protein functional domains from sequences using artificial neural networks.** *Genome Research*. 2001, **11**:1410-1417.

18. Vinayagam A, del Val C, Schubert F, Eils R, Glatting KH, Suhai S, Konig R. **GOPET: a tool for automated predictions of Gene Ontology terms.** *BMC Bioinformatics*. 2006, **7**:161.

19. Zhu M, Gao L, Guo Z, Li Y, Wang D, Wang J, Wang C. **Globally predicting protein functions based on co-expressed protein-protein interaction networks and ontology taxonomy similarities.** *Gene.* 2007, **391(1-2)**:113-119.

20. Gallego M, Virshup DM: **Protein serine/threonine phosphatases: life, death, and sleeping.** *Curr Opin Cell Biol* 2005, **17**: 197-202.

21. Bourdeau A, Dube N, Tremblay ML: **Cytoplasmic protein tyrosine phosphatases, regulation and function: the roles of PTP1B and TC-PTP.** *Curr Opin Cell Biol 2005,* **17**: 203-209.

22. Gene Ontology Consortium: **The Gene Ontology (GO) project in 2006.** *Nucleic Acids Res 2006,* **34**: 322-326.

23. Larranaga P, Calvo B, Santana R: **Machine learning in bioinformatics.** *Brief Bioinform 2006*, **7**: 86-112.

24. Eppig JT, Bult CJ, Kadin JA: **The Mouse Genome Database (MGD): from genes to mice--a community resource for mouse biology.** *Nucleic Acids Res* 2005, **33**: 471-475.

25. Okazaki Y, Furuno M: **Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs.** *Nature 2002*, **420**: 563-573.

26. Bairoch A, Apweiler R, Wu CH: **The Universal Protein Resource (UniProt).** *Nucleic Acids Res 2005*, **33**: 154-159.

27. Quinlan JR: *C4.5: Programs for Machine Learning*. Morgan Kauffman; 1993.

28. Caenepeel S, Charydczak G, Sudarsanam S, Hunter T, Manning G: **The mouse kinome: discovery and comparative genomics of all mouse protein kinases.** *PNAS 2004*, **101:** 11707-11712.

29. Jones CE, Brown AL, Baumann U. **Estimating the annotation error rate of curated GO database sequence annotations.** *BMC Bioinformatics* 2007, **8(1)**: 170.

30. Tsoumakas G and Katakis I. **Multi-label classification: An overview.** *Int J Data Warehousing and Mining* 2007, **3(3)**:1-13.

31. Barutcuoglu Z, Schapire RE, Troyanskaya OG. **Hierarchical multi-label prediction of gene function.** *Bioinformatics* 2006, **22(7)**:830-836

32. Rousu J, Saunders C, Szedmak S, Shawe-Taylor J. **Kernel-Based Learning of Hierarchical Multilabel Classification Models.** *J Mach Learn Res* 2006, **7:** 1601–1626

33. Blockeel H, Schietgat L, Struyf J, Dzeroski S, Clare A. **Decision Trees for Hierarchical Multilabel Classification: A Case Study in Functional Genomics.** In: *Proceedings of 10th European Conference on Principles and Practice of Knowledge Discovery in Databases*, Berlin: Springer, Lecture Notes in Computer Science 2006. Vol. 4213 pp.18-29.

34. Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D. **A combined algorithm for genome-wide prediction of protein function.** *Nature* 1999, **402**: 83–86.

35. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. **Assigning protein functions by comparative genome analysis: protein phylogenetic profiles**. *Proc Natl Acad Sci U S A* 1999, **96**: 4285–4288.

36. Eisen MB, Spellman PT, Brown PO, Botstein D. **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci U S A* 1998, **95**:14863–14868.

37. Karaoz U, Murali TM, Letovsky S, Zheng Y, Ding C, Cantor C, Kasif S. **Whole-genome annotation by using evidence integration in functional-linkage networks**. *Proc Natl Acad Sci U S A* 2004, **101**: 2888–2893.

38. Nariai N, Kolaczyk ED, Kasif S. **Probabilistic protein function prediction from heterogeneous genome-wide data.** *PLoS ONE* 2007, **2**: e337.

39. Xiong J, Rayner S, Luo K, Li Y, Chen S. **Genome wide prediction of protein function via a generic knowledge discovery approach based on evidence integration.** *BMC Bioinformatics* 2006, **7**: 268.

40. Witten I, Frank E: **Data mining in bioinformatics using Weka.** *Data Mining: Practical machine learning tools and techniques*. 2nd Edition, San Francisco: Morgan Kaufmann; 2005.

41. Silvescu A, Andorf C, Dobbs D, Honavar V: **Inter-element dependency models for sequence classification Technical report**. Department of Computer Science, Iowa State University, 2004. [http://www.cs.iastate.edu/silvescu/papers/nbktr/nbktr.ps].

42. Cowell R, Dawid A, Lauritzen S, Spiegelhalter D: *Probabilistic Networks and Expert Systems*. Springer; 1999.

43. Mitchell, T. *Machine learning*. New York, USA: McGraw Hill; 1997.

44. Altschul S, Madden T, Schaffer A, Zhang J, Miller W, Lipman D: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acid Res 1997,* **2**: 3389 – 3402.

45. Baldi P, and Brunak S. *Bioinformatics: The Machine Learning Approach*. Cambridge, MA: MIT Press; 1998.

Figures

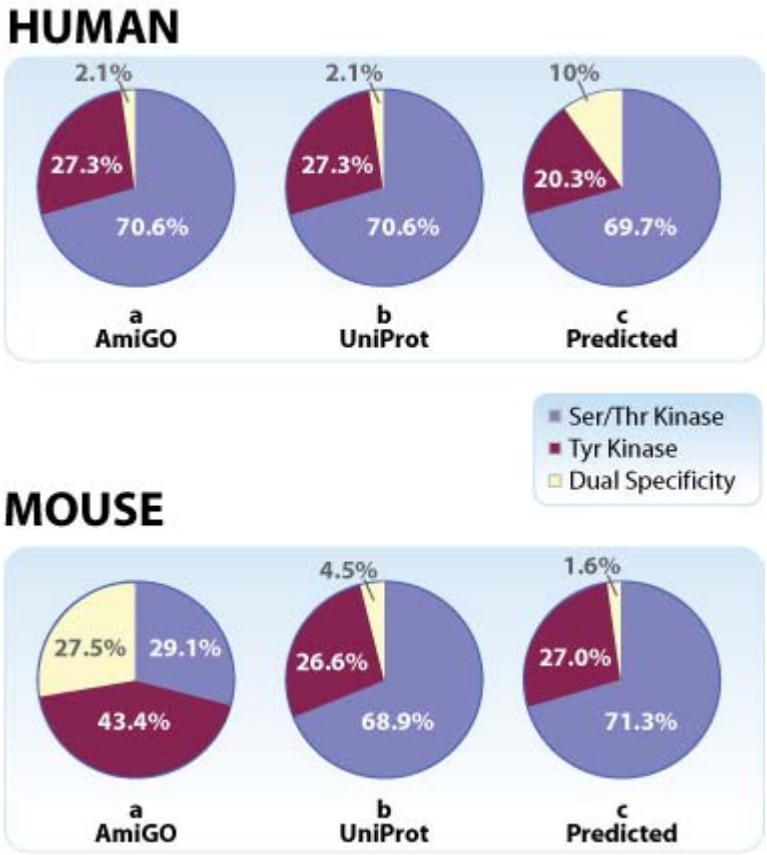


Figure 1: Distribution of Ser/Thr, Tyr, and dual specificity kinases among annotated protein kinases in human versus mouse genomes (See **Appendix I** for details). Pie charts illustrate the functional family distribution of protein kinases in human (top) versus mouse (bottom), based on: **a. AmiGO functional classifications:** Ser/Thr (GO0004674) [Blue]; Tyr (GO0004713) [Red] or "dual specificity" (proteins with both GO classifications) [Yellow]. **b. UniProt annotations:** classification based on UniProt records containing the key words Ser/Thr [Blue], Tyr [Red], or dual specificity [Yellow] (see **Appendix D**). **c. Predicted annotations by the HDTree classifier:** The classifier was built on human proteins with functional labels Ser/Thr (GO0004674)[Blue], Tyr (GO0004713) [Red] or "dual specificity" [Yellow] derived from AmiGO and verified by UniProt **(Appendix A)**.

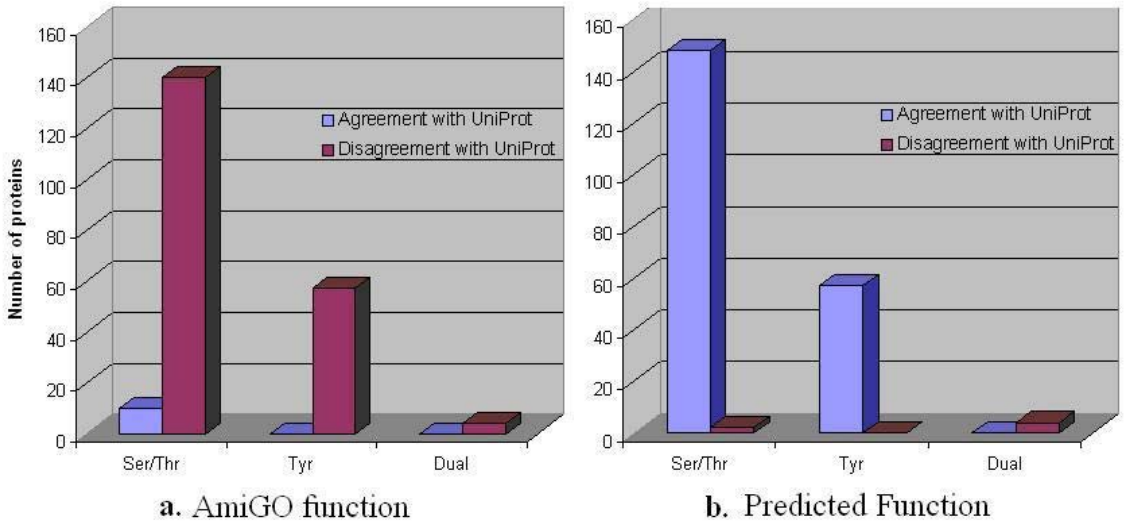


Figure 2: Comparison of UniProt annotations of mouse protein kinases sequences with annotations from AmiGO or predicted by HDTree. The bar charts illustrate the number of proteins that were in agreement (blue)/disagreement (red) with the annotations found in UniProt. Proteins that belong to each of the three functional classes found in the UniProt records are represented by two bars. The blue bar represents the number of proteins in which UniProt and the given method share the same annotation (*agreement*) for that function. The red bar represents the number of proteins in which UniProt and the given method have different annotations (*disagreement*) for that function. **a.** AmiGO vs. UniProt annotations **b.** HDTree predictions vs. UniProt annotations. (Additional details are provided in **Appendix A and E**).

Tables

Table 1. Comparison of performance of classifiers to predict AmiGO annotations- The performance measures accuracy, kappa coefficient correlation coefficient, precision, and recall are reported for two of the HDTree classifiers. The first classifier is trained on 330 human proteins. The performance is based on 10-fold cross-validation. The second classifier is trained on the 330 human proteins and tested on 244 mouse proteins. The annotations for the mouse and human proteins were obtained from AmiGO.

	Accuracy	Kappa Coefficient	Correlation Coefficient			Precision			Recall		
			Ser/Thr	Tyr	Dual	Ser/Thr	Tyr	Dual	Ser/Thr	Tyr	Dual
Human	89.1	0.76	0.82	0.86	0.30	0.97	1.00	0.15	0.95	0.74	0.71
Mouse	15.1	-0.40	-0.40	-0.43	-0.01	0.17	0.11	0.25	0.41	0.07	0.01

Table 2 . Comparison of AmiGO and UniProt annotations for 211 mouse protein kinases with RCA Evidence code - Each of the 211 mouse kinase proteins with an RCA evidence code used in this study has an AmiGO and a UniProt annotation. This table shows the number of proteins that have the nine possible combinations of the AmiGO and UniProt annotations. Each row of the table represents one of the three possible UniProt labels and each column represents each of the three AmiGO annotations. Each entry of the table shows the number of proteins with the corresponding annotation. Please note, all entries along the diagonal (in bold) show the number of proteins where the AmiGO and UniProt annotations were in agreement. All other entries show the number of proteins where AmiGO and UniProt were in disagreement. (Additional details provided in **Appendix D and E**).

KINASE FAMILY	AmiGO Ser/Thr	AmiGO Tyr	AmiGO Dual specificity
UniProt Ser/Thr	**10**	105	35
UniProt Tyr	54	**0**	3
UniProt Dual specificity	0	4	**0**

Table 3. Comparison of performance of classifiers based on AmiGO annotations and UniProt annotations- The performance measures accuracy, kappa coefficient correlation coefficient, precision, and recall are reported for two of the HDTree classifiers. Both classifiers were trained on 330 human proteins and tested on 211 mouse proteins with RCA evidence codes in AmiGO. The first classifier was trained and tested with annotations provided by UniProt and the second classifier used annotations obtained from AmiGO.

	Accuracy	Kappa Coefficient	Correlation Coefficient			Precision			Recall		
			Ser/Thr	Tyr	Dual	Ser/Thr	Tyr	Dual	Ser/Thr	Tyr	Dual
UniProt	97.1	0.93	0.98	0.94	0.00	0.97	0.97	0.00	0.99	1.00	0.00
AmiGO	4.2	-0.37	-0.64	-0.85	0.00	0.06	0.00	0.00	0.14	0.00	0.00

Table 4. Classification Schema for Classier #3 (Method to predict Dual Specificity)- HDTree Classifier #3 uses the outputs form HDTree Classifier #1 and HDTree Classifier #2 to distinguish between dual-specificity kinases, Ser/Thr kinases, and Tyr kinases. See Methods section for details on each classifier.

Prediction of classifier #1 (Ser/Thr)	Prediction of classifier #2 (Tyr)	**New Prediction of classifier #3 (Dual, Ser/Thr,Tyr)**
Yes	Yes	**Dual**
Yes	No	**exclusively Ser/Thr**
No	Yes	**exclusively Tyr**
No	No	**Dual**

Table 5. Performance measures definitions for binary classification- The performance measures *accuracy, precision, recall, correlation coefficient, and kappa coefficient* are used to evaluate the performance of our machine learning approaches [45]. *Accuracy* is the fraction of overall predictions that are correct. *Precision* is the ratio of predicted true positive examples to the total number of actual positive examples. *Recall* is the ratio of predicted true positives to the total number of examples predicted as positive. *Correlation coefficient* measures the correlation between predictions and actual class labels. *Kappa coefficient* is used as a measure of agreement between two random variables (predictions and actual class labels).

The table summarizes the definitions of performance measures in the 2-class setting (binary classification), where M = the total number of classes and N = the total number of examples. *TP, TN, FP, FN* are the true positives, true negatives, false positives, and false negatives for the given confusion matrix.

Performance Measure	Binary Classification
Accuracy	$\frac{TP + TN}{TP + FP + TN + FN}$
Precision	$\frac{TP}{TP + FN}$
Recall	$\frac{TP}{TP + FP}$
Correlation Coefficient	$\frac{TP * TN - FP * FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}}$
Kappa Coefficient	$\frac{(TP * + TN) - ((TP + FN) * (TP + FP) + (TN + FN) * (TN + FP))}{N - ((TP + FN) * (TP + FP) + (TN + FN) * (TN + FP))}$

Table 6. Performance measure definitions for multi-class classification- The performance measures *accuracy, precision, recall, correlation coefficient, and kappa coefficient* are used to evaluate the performance of our machine learning approaches [45]. *Accuracy* is the fraction of overall predictions that are correct. *Precision* is the ratio of predicted true positive examples to the total number of actual positive examples. *Recall* is the ratio of predicted true positives to the total number of examples predicted as positive. *Correlation coefficient* measures the correlation between predictions and actual class labels. *Kappa coefficient* is used as a measure of agreement between two random variables (predictions and actual class labels).

The table displays the general definition of each measure, where M = the total number of classes and N = the total number of examples, x_{ik} represents the number of examples in row i and column k of the given confusion matrix.

Performance Measure	General Classification
Accuracy (class i)	$\frac{\sum_{i=1}^{M} x_{ii}}{N}$
Precision (class i)	$\frac{x_{ii}}{\sum_{k=1}^{M} x_{ki}}$
Recall (class i)	$\frac{x_{ii}}{\sum_{k=1}^{M} x_{ik}}$
Correlation Coefficient (class i)	$\frac{(x_{ii} * \sum_{h \neq i} x_{hh}) - (\sum_{k=1}^{M} x_{ki} * \sum_{j=1}^{M} x_{ij})}{\sqrt{(x_{ii} + \sum_{k=1}^{M} x_{ki})(x_{ii} + \sum_{j=1}^{M} x_{ij})(\sum_{h \neq i} x_{hh} + \sum_{k=1}^{M} x_{ki})(\sum_{h \neq i} x_{hh} + \sum_{j=1}^{M} x_{ij})}}$
Kappa Coefficient	$\frac{\sum_{i=1}^{M} x_{ii} - \sum_{h=1}^{M} (\sum_{k=1}^{M} x_{kh} * \sum_{j=1}^{M} x_{hj})}{N - \sum_{h=1}^{M} (\sum_{k=1}^{M} x_{kh} * \sum_{j=1}^{M} x_{hj})}$

Supplementary Information found in the Appendices Section

Appendix A: Supplementary Data

Supplementary Data: Machine Learning approaches to predict Gene Ontology and/or UniProt Functional labels.
Description: The data provided represent the results and performance of all the machine learning approaches used in this study.

Appendix B: – Supplementary Note

Supplementary Note: Because there is only a non-curated reference to the work done on "Rat ISS GO annotations from MGI's mouse gene data," we provide the abstract and a link to the original reference report.
Description: Because there is only a non-curated reference to the work done on "Rat ISS GO annotations from MGI's mouse gene data," we provide the abstract and a link to the original reference report in this file.

Appendix C: – Supplementary Table 1

Supplementary Table 1: Evidence Codes for AmiGO annotations
Description: A table displaying the Evidence Codes for AmiGO annotations of the mouse protein kinases used in this study.

Appendix D: – Supplementary Table 2

Supplementary Table 2: AmiGO annotations versus UniProt annotations (with UniProt Evidence)
Description: A table comparing the annotations found in the AmiGO server with the annotations found in UniProt.

Appendix E: – Supplementary Table 3

Supplementary Table 3: AmiGO labels, UniProt labels, and Predicted Labels for each mouse kinase protein
Description: A table comparing the predicted annotations from our three machine learning classifiers with the annotations of AmiGO and UniProt.

Appendix F: – Supplementary Table 4

Supplementary Table 4: Mouse Kinases having a Human Ortholog
Description: A table displaying the human orthologs for the mouse kinases used in this study. The table also displays the identity between these orthologs.

Appendix G: – Supplementary Table 5

Supplementary Table 5: Number of Mouse kinases having a specified level of sequence identity with their human orthologs. (Summary statistics for Supplementary Table 4)
Description: A table displaying the summary statistics of Supplementary Table 4.
Appendix H: – Supplementary Table 6

Supplementary Table 6: The UniProt and AmiGO annotations for the Rat kinase proteins with Mouse orthologs
Description: This table displays the UniProt and AmiGO annotations for rat kinase proteins that were annotated based on a mouse ortholog.

Appendix I: – Supplementary Table 7

Supplementary Table 7: Distribution of protein classes for Human and Mouse proteins annotated by AmiGO, UniProt, and HDTree
Description: This table is a representation of the data used in Figure 1 which is a pie chart showing the distribution of human and mouse protein classes based on annotations found in AmiGO, UniProt, and predicted by HDTree.

CHAPTER 3.

PREDICTING THE BINDING PATTERNS OF HUB PROTEINS: A STUDY USING YEAST PROTEIN INTERACTION NETWORKS

Modified from a paper published in *PLOS ONE*
8(2) (19 February 2013)

Carson Andorf, Vasant Honavar, Taner Sen

Abstract

Background

Protein-protein interactions are critical to elucidating the role played by individual proteins in important biological pathways. Of particular interest are hub proteins that can interact with large numbers of partners and often play essential roles in cellular control. Depending on the number of binding sites, protein hubs can be classified at a structural level as singlish-interface hubs (SIH) with one or two binding sites, or multiple-interface hubs (MIH) with three or more binding sites. In terms of kinetics, hub proteins can be classified as date hubs (i.e., interact with different partners at different times or locations) or party hubs (i.e., simultaneously interact with multiple partners).

Methodology

Our approach works in 3 phases: Phase I classifies if a protein is likely to bind with another protein. Phase II determines if a protein-binding (PB) protein is a hub. Phase III classifies PB proteins as singlish-interface versus multiple-interface hubs and date versus party hubs. At each stage, we use sequence-based predictors trained using several standard machine learning techniques.

Conclusions

Our method is able to predict whether a protein is a protein-binding protein with an accuracy of 94% and a correlation coefficient of 0.87; identify hubs from non-hubs with 100% accuracy for 30% of the data; distinguish date hubs/party hubs with 69% accuracy and area under ROC curve of 0.68; and SIH/MIH with 89% accuracy and area under ROC curve of 0.84. Because our method is based on sequence information alone, it can be used even in settings where reliable protein-protein interaction data or structures of protein-protein complexes are unavailable to obtain useful insights into the functional and evolutionary characteristics of proteins and their interactions.

Availability: We provide a web server for our three-phase approach: http://hybsvm.gdcb.iastate.edu.

Introduction

Proteins are the principal catalytic agents, structural elements, signal transmitters, transporters and molecular machines in cells. Functional annotation of proteins remains one of the most challenging problems in functional genomics, however, our evolving understanding of a proteins' interaction partners helps in functional annotation of proteins [1]. Protein-protein interactions are therefore critical to elucidating the role played by individual proteins in important biological pathways. Such networks are typically constructed using high throughput techniques (e.g., yeast two-hybrid (Y2H) experiments).

Our current understanding of protein-protein interaction networks is quite limited for a variety of reasons. The challenge of reliable and complete determination of the interactome is far from being fully addressed due to the high rate of false positives. These false positives are associated with high throughput experiments, the low coverage of solved co-crystal structures in the Protein Data Bank (PDB), and the difficulty of obtaining reliable negative evidence that a protein does not interact with one or more other proteins. For example, Y2H experiments focus on pair-wise interactions between proteins and provide, at best, rather indirect evidence for higher order interactions e.g., those that require three proteins to come together to form a complex. Even in the case of pairwise interactions, Y2H experiments only provide evidence that a pair of proteins is likely to interact *in vitro*, without offering any insights into the physical basis of such interactions, or whether such interactions may actually occur *in vivo* [2-6]. It is well known that data from high-throughput Y2H experiments are notoriously noisy and suffer from a high false positive rate [7]. The high-quality datasets tend to have low-coverage e.g., it is estimated up to 95% of the human interactome is unmapped [8]. Furthermore, whether a particular

protein-protein interaction is experimentally observed depends on the specific experimental conditions. It is therefore critical to validate the putative interactions between proteins suggested by Y2H experiments using additional experimental or computational studies. As a result, considerable amount of recent work has focused on creating high-quality interaction datasets by systematically removing errors and low-quality interactions or by combining multiple sources of evidence [2,3,8,9]. Hence, there is considerable interest in reliable prediction of protein-protein interactions.

Protein-protein interaction networks are usually represented and visualized as graphs in which the nodes correspond to the proteins and edges denote their possible pairwise interactions. Such a representation is simply not rich enough to encode interactions that involve more than two proteins, nor do they distinguish between them. Furthermore, a single target protein can interact with a large number of partners: some of these interactions may be mutually exclusive because of competition between potential binding partners for the same interaction sites on the target protein. Other interactions may be simultaneously possible, and in many instances, even mutually cooperative, i.e., binding of one partner to the target protein may prepare the target for binding to a second partner [10,11]. Distinguishing between these various types of interactions is essential for uncovering the physical basis of interactions of a protein with other proteins, engineering the protein surfaces to manipulate synthetic pathways, or for designing drugs that bind specific targets [12-14]. However, answering such questions is extremely difficult in the absence of direct experimental evidence, e.g., structures of complexes formed by a protein when it interacts with one or more other proteins or results of site-specific mutation experiments that identify the protein surface residues that play essential roles in such

interactions. Unfortunately, experimental determination of protein-protein complexes or of binding sites is notoriously time-consuming and expensive. Hence, there is a growing interest in computational tools that provide useful insights into various structural aspects of protein interactions from protein sequence alone.

Of particular interest in this study are hub proteins, i.e., proteins that interact with large numbers of partners [15]. It is worth noting that "large numbers of partners" is a relative term and is arbitrarily defined. In several studies, hub proteins are defined as those with 5 or more interaction partners [15-18]. The choice of five (as opposed to some other number) or more interacting partners as the defining characteristic of hub proteins is somewhat arbitrary. The quality of protein-protein interaction data (false positives, incomplete coverage) presents additional challenges in categorizing proteins into hubs and non-hubs. These difficulties notwithstanding, hub proteins have been reported to play essential roles in cellular control and tend to be highly conserved across species [19]. Mutations in hub proteins can potentially disrupt its interactions with its many interaction partners, which can turn out to be lethal for the cell's survival [20-22]. Hence, it is especially important to understand physical and structural basis of interactions of hub proteins with their partners. Recent studies suggest that hubs are more diverse than previously thought and show striking differences in number of binding sites and kinetics of binding. Kim et al. [16] combined three-dimensional structure information, known domain-domain interaction data, and protein-interaction data to define two types of hub protein structures. The first type of hub proteins, called singlish interface hubs (SIH), interacts with multiple partners at one or two binding sites. Because the interactions rely on binding events at one or two binding sites, interactions with the different partners tend

to be mutually exclusive. The second type of hub proteins, called multiple-interface hubs (MIH), interacts with multiple interaction partners through more than two binding sites (See Figure 1). Recent studies [16,22-30] have explored the roles of SIH and MIH in protein-protein interactions and hence protein function. Kim et al. [16], who were among the first to analyze the properties of SIH and MIH proteins, found that MIH were twice as likely (compared to SIH) to be *essential* for survival and perhaps as a consequence, more conserved across species with implications for determining the evolutionary rates for protein hubs. They also found that MIH proteins are more likely to be members of large stable structural complexes. SIH and MIH also differ in terms of network expansion during evolution: SIH appear to follow the canonical preferential gene duplication model whereas MIH do not [16]. A recent study showed SIH tend to display higher degrees of disorder than MIH [28]. Table 1 summarizes the results of previous studies [16,28] that have compared the properties of SIH and MIH.

Hub proteins can also be classified based on the kinetic mode of interaction. Han et al. [31] recently described an expression-based classification model for hub proteins. This classification is based on a bimodal distribution of co-expression of hub proteins with their interaction partners [31]. Date hubs tend to display expression levels that have low correlation with those of their interaction partners (and tend to bind different partners at different time points or locations). Conversely, party hubs tend to exhibit expression levels that have high degree of correlation with those of their interaction partners (and tend to interact simultaneously with the partners). See Figure 1 for an illustration of date hubs versus party hubs. The analysis of party and date hubs provides additional insights into the structure of the underlying proteome and interactome. For example, date hubs contribute

to global network stability and connectivity by acting as a inter-module linkers [10] that serve as regulators, mediators, or adapters. In contrast, party hubs act as intra-module linkers that coordinate a specific process or assist the formation of a specific protein complex [31,32]. In these intermolecular interactions, the residues that contribute the most to binding (hot spots) for date hubs tend to be spatially near each other (forming hot regions) [25]. Date hubs are likely to evolve faster than party hubs [33]. Table 2 summarizes the conclusions of previous studies that have compared the properties of date hubs and party hubs [25,31,33,34]. The differences between the two types of hub proteins strongly suggest that they might play different functional roles. SIH tend to be date hubs whereas MIH tend to be party hubs [16]; but there are exceptions. It should be no surprise that SIH tend be date hubs: the number of binding sites that a hub protein has limits the number of partners with which it can interact at the same time. However, the converse does not necessarily hold, i.e., not every date hub is a SIH. A date hub may only have one or two concurrent interactions at any given time, but each of these interactions may involve different binding sites. Hence a date hub can in general be a SIH or a MIH. Similarly a party hub tends to be a MIH, since many concurrent interactions require many interaction sites, but a MIH can be a party hub or a date hub based on the interaction kinetics. Recent studies have focused on the role of hubs in interaction networks and in particular, the differences in SIH versus MIH and date hubs versus party hubs [22-24,26,27,29-31,35,36].

Experimental characterization of hub proteins in terms of their structural and kinetic characteristics requires knowledge of the structures of complexes formed by such proteins in interaction with their binding partners [35,37]. Because of the prohibitive cost and effort needed to determine the structures of complexes formed by hub proteins with

their binding partners and the interfaces that mediate such interactions, there is an urgent need for reliable methods for predicting the structural and kinetic characteristics of hubs from sequence information alone. In particular, there is a growing interest in purely sequence-based computational methods for discriminating between simultaneously possible versus mutually exclusive interactions[27,31,38,39] and predicting the number of binding sites available on the surface of a protein.

There has been considerable work on machine learning approaches for distinguishing hub proteins from non-hub proteins [40-42]. Mirzarezaee et al. have recently proposed methods for distinguishing between date hubs (that interact with one partner at a time) and party hubs (that simultaneously interact with many partners) [15] using 17 features including 4 composition measurements, grouping of 48 physicochemical properties, six GO term features, domain occurrence, disordered regions, and position specific scoring matrices (PSSM). They reported correlation coefficients of 0.74 for both date and party hubs. In light of these results, a natural question to ask is whether similar or better performance can be achieved from information based solely on the sequence of the hub protein.

Against this background, we introduce a three-phase machine learning approach (See Figure 2). Phase I predicts if a protein physically binds with other proteins (protein-binding (PB) versus non-protein-binding (NPB)). If a protein is predicted to be a PB protein, that protein goes through the second and third phase of predictions. Phase II uses sequence similarity to determine the potential number of interaction sites for the input sequence based on a weighted-average of the number of interactors of the top scoring BLAST hits. Phase III applies methods for predicting both structure (singlish vs. multiple)

and kinetics (date vs. party) classifications of protein-binding proteins using information derived from only the sequence of the protein (See Figure 3). Our experiments show that our method is able to predict whether a protein is a protein-binding protein with an accuracy of 94%, 0.93 area under a ROC curve (AUC) and a correlation coefficient of 0.87; identify hubs from non-hubs with 100% accuracy for 30% of the data (with the rest being flagged as putative hubs or putative non-hubs depending on the sequence similarity to known hubs/non-hubs in our dataset); distinguish date hubs/party hubs with 69% accuracy and AUC of 0.68; and SIH/MIH with 89% accuracy, 0.85 AUC. The method can be used even in settings where reliable protein-protein interaction data, or structures of protein-protein complexes are unavailable, to obtain useful insights into the functional and evolutionary characteristics of proteins and their interactions. In addition, our method does not rely on computationally expensive multiple sequence alignments, the presence of functional or structural domains, or additional functional annotations (e.g. GO terms), allowing for fast and updateable predictions.

It should be noted that categorizing hub proteins into structural and kinetic classes presents many challenges. SIH and date proteins are defined by the absence of concurrent interaction partners or interaction sites. However, it is difficult to reliably determine the absence of interaction between a protein and one or more putative interaction partners because of the lack of experimental data under a broad range of conditions. It is thus possible that some proteins labelled as SIH in our dataset are in fact MIH where not all interaction partners have been identified. Conversely, because of the high false positive rates associated with high-throughput experiments, some proteins labelled as MIH or party hubs are in fact SIH. These sources for errors in the protein-protein interaction data need

to be kept in mind in interpreting the results of our study as well as other similar analyses of protein-interaction data.

A web server for the three-phase approach for automated PB/NPB, SIH/MIH, and date/party prediction is available at http://hybsvm.gdcb.iastate.edu.

Results and Discussion

Our approach to classifying proteins based on binding patterns is a 3-phase approach: Phase I predicts if a protein is likely to bind with another protein, i.e., protein-binding (PB). Phase II determines if a protein-binding protein is a hub. Phase III classifies PB proteins as singlish-interface versus multiple-interface hubs and date versus party hubs, based on sequence information alone. We present results of experiments for each of the three phases.

In this study, we use a simple encoding of protein sequences using the probability distribution short (*k*-letter) subsequences (*k*-grams) of amino acids. In our experiments, we used values of k ranging from *k*=1 (amino acid composition) through k=4 (dimers, trimers, and tetramers). Larger values of *k* were not considered, because we run out of data to reliably estimate the model parameters. We use a range of standard machine learning methods implemented in Weka version 3.6.0: J4.8 version [43] of the C4.5 decision tree learning algorithm (Decision Tree) [44], SMO version [45] of the support vector machine (SVM) [46] with a polynomial kernel, Multilayer Perception neural network (ANN) [43], and Naïve Bayes algorithm [43]. In addition, in Phase I and III, we use a two-stage ensemble classifier, *HybSVM*, which uses an SVM to combine the outputs of a set of predictors. We compare the results of predictors trained using machine learning methods with two baseline methods: the first baseline method classifies proteins based on the

number of SCOP [47,48] and PFAM [49] domains (domain-based method) present in the sequence. The second baseline method classifies each protein based on the class-label of its nearest PSI-BLAST hit. To evaluate predictors constructed using machine learning we used 10-fold cross-validation. Because any single measure e.g., accuracy, provides at best partial information about the performance of a predictor, we use a set of measures including accuracy, precision, recall, correlation coefficient, F-measure, and area under the Receiver Operating Characteristic (ROC) curve. Additional details can be found in the Methods section of the paper.

Predicting protein binding proteins (Phase I)

To evaluate our method to discriminate proteins that bind to other proteins from those that bind to other substrates (e.g., small ligands), we assembled Dataset 1, which consists of 5,010 proteins including 3,418 proteins that bind to one or more proteins and 1,592 that bind to small ligands, but are not known to bind to other proteins. As mentioned in the introduction, creating a set of proteins that do not bind to any other protein is a difficult challenge due to low-coverage and high false-positive rates in available protein-protein interaction data. Here we use the information coming from ligand-binding experiments to obtain "negative data", i.e., non-protein-binding proteins: considering the inaccuracies in the protein-protein interaction data, if a protein has no experimental evidence of binding with another protein, but with a ligand, then we assume that the protein is non-protein binding. Our hypothesis here is that if a protein interacts with a ligand and no experimental data are available for its interaction with another protein, then the lack of evidence of protein-protein interaction is less likely due to the incompleteness in the data and more likely due to the lack of protein binding activity. Thus, we assembled

a set of ligand-binding proteins and filtered out those that had high sequence similarity to proteins known to bind with other proteins to obtain a set of non-protein binding proteins. The methodology (described in detail in the Methods section) is not without its drawbacks: it disregards ligand-interacting proteins that are also involved in protein-protein interactions *in vivo* but lacking the confirmation of *in vitro* experimental data.

As shown in Tables S1 and S2, the ability to distinguish protein-binding proteins from non-protein-binding proteins varies as a function of the machine learning method used and the size of the *k*-gram used. The accuracies ranged from 74.4% (Decision Tree, k=2) to 87.2% (SVM, k=2). Simply predicting each protein as belonging to the majority class yields an accuracy of 68.2% (see Domain-based method). Most of the methods were able to achieve accuracies well above 68.2%. The precision values ranged from 0% to 81%, recall from 0% to 93%, and correlation coefficient from 0.00 to 0.69. Figure 4 shows ROC curves for each of the methods. These curves show no single method outperforms all others over the entire range of tradeoffs between precision and recall. This suggests the possibility of using an ensemble of classifiers that takes advantage of the complementary information provided by the individual classifiers.

To examine this possibility, we built *HybSVM* for Phase I, which constructs a support vector machine (SVM) classifier that takes as input, for each protein sequence to be classified, the outputs of seven classifiers as well as the PSI-BLAST method and produces as output, a class label for the protein. The 7 classifiers used are: NB(1), NB(2), NB(3), NB(4), NB 2-gram, NB 3-gram, NB 4-gram. PSI-BLAST performs well on sequences with high sequence similarity whereas the NB(k) and NB k-gram methods perform well on sequences with high *k*-gram composition similarity. Logistic regression

models are applied to the *HybSVM* classifier to get a probability score for each prediction. These scores are then used to evaluate the quality of each prediction.

Table 3 compares the performance of the *HybSVM* classifiers for Phase I against other standard machine learning approaches. *HybSVM* had an accuracy of 94.2% (an improvement of 6% in absolute terms over NB 4-gram) and a correlation coefficient of 0.87 (an improvement of 0.15 over NB 4-gram). For each performance measure the *HybSVM* method had the highest value for Dataset 1. *HybSVM* for Phase I also outperforms the other methods over the entire range of tradeoffs between precision and recall on a ROC curve (Figure 4).

Predicting hub proteins (Phase II)

Since our overall goal is to predict structural and kinetic classes for hub proteins and these classifiers need to be trained on hub-only proteins, we need a method to (1) identify hub proteins, (2) filter out non-hub proteins, and/or (3) flag proteins that have potential of being non-hubs. To evaluate this type of method, we assembled Dataset 2, consisting of 4,036 proteins including 1,741 hub proteins and 2,295 non-hub proteins. The dataset was derived from high confidence protein-protein interaction data from BioGrid [50] by labelling proteins with more than 5 interaction partners as hubs and proteins with fewer than 3 interaction partners as non-hubs. Proteins with 3, 4, or 5 interaction partners were not used in the dataset because, given the incompleteness of experimentally determined interactions, their categorization into hubs versus non-hubs is likely to be less reliable than the rest of the proteins in the dataset.

We used a simple homology-based method to classify proteins into hubs and non-hubs. A protein is classified as a hub if each of the top 4 hits returned by PSI-BLAST [51]

search correspond to hub proteins (See Methods for details). Similarly, a protein is classified as a non-hub if all of the top hits are non-hub proteins. A protein is flagged as being likely a hub or non-hub based on the majority of the class-labels of the four top hits. If no hits are reported, the protein is flagged as having no known label. In addition to our predictions, in our web server, we report the number of interaction partners belonging to the top hit, the range of interaction partners of the top four hits, and a predicted number of interaction partners (based on the number of interaction partners of the top four BLAST hits weighted by the BLAST score of each hit). This simple sequence-based method correctly classified 536 hub proteins and 630 non-hub proteins (approximately 30% of the data). No proteins were incorrectly classified as hubs or non-hubs.

Predicting structural and kinetic classes for hub proteins (Phase III)

Structural prediction: discriminating SIH from MIH hub proteins. To evaluate structural predictions on hub proteins, we created Dataset 3. The dataset consists of 155 hub proteins including 35 SIH and 120 MIH. The dataset is a subset of data originally compiled by Kim et al. [16], but has been filtered to remove highly homologous sequences (50% or more sequence identity within at least 80% of the length of the sequence).

Tables S3 and S4 show the ability to distinguish SIH and MIH (Dataset 3) based on several standard machine learning approaches with varying sizes *k*-grams. The accuracies ranged from 67.7% (Decision Tree, k=2) to 81.2% (Naive Bayes, k=3). Several classifiers actually had accuracies below 77.4% (e.g., SVM, k=1). The precision values ranged from 0% to 86%, recall from 0% to 63%, and correlation coefficient from 0.00 to 0.41. Figure 4 shows ROC curves for each of the methods. Again, these curves show no single method outperforms all others. On Dataset 3, each of the machine learning methods used here

outperformed the simple domain-based method (note that the simple domain-based method had both 0.00 precision and recall because it was unable to predict any SIH proteins correctly).

To validate how well interaction sites of SIH and MIH can be predicted on Dataset 3, we ran a subset of the data through the interaction site predictor ISIS [52] and the target specific interaction site predictor NPS-HomPPI [53] with default settings. Both methods generally under-predicted the number of interaction sites and in many cases the methods predicted few or no interaction sites on the hub proteins.

Table 4 shows that the individual methods perform well on assigning hubs to classes based on structural characteristics. No single *k*-value is optimal for all methods; optimal values of *k* vary with the size and complexity of the dataset. Variables such as number of proteins, size of proteins, and homology between proteins all play an important role in developing an appropriate model for our classifiers. Therefore, it is difficult to design a single model or choose a single optimal value of *k* for any dataset without prior knowledge of the data. We also observe that proteins within a class are assigned different labels by classifiers that correspond to different choices of *k*. Our results show that a single classifier does not classify all the proteins correctly, yet a vast majority of the proteins (over 93%) have at least one classifier that correctly predicts its class. Again, we used the *HybSVM* method, this time for Phase III classifications, to take advantage of the complementary information provided by the individual classifiers.

From the results shown in Table 4, we can see that *HybSVM* outperforms all other individual methods on 5 of the 6 performance measures. For Dataset 3, *HybSVM* improved accuracy by 5.2% (89.0%) and correlation coefficient by 0.22 (0.69) over the

previous best classifier, NB 4-gram. This method also had the highest AUC with a value of 0.85 (an improvement of 0.14 over SVM, the next highest performing method). NB(k) had the highest recall value at 0.84 (it was able to correctly label more SIH proteins), but it came at the cost of a low precision (0.31) and lower correlation coefficient (0.44). The threshold can be adjusted for *HybSVM* to achieve a better recall based on *HybSVM* having very balanced precision and recall scores (values of 0.75 and 0.77) and the highest f-measure (76.0). Figure 4 shows ROC curves for each of the methods on Dataset 3. Although *HybSVM* did not always outperform the other methods over the entire range of tradeoffs between precision and recall, it did outperform the other methods for a specific range of false positive rates (from 0.0 to 0.4 for SIH and from 0.25 to 0.5 MIH). No single method significantly outperformed *HybSVM.* It is worth noting that *HybSVM* method is especially attractive if there is little tolerance for false positives. In contrast, each of the other methods (with the exception of domain-based method) works relatively well, in settings where there is greater tolerance for higher false positive rates.

A closer examination of the results for Dataset 3 shows that many of the misclassified hub proteins are close to the arbitrary boundary between SIH and MIH. This raises the question as to whether the labels could be more reliably predicted if the arbitrary cut-off on the number of interfaces is altered (See Figure S1). For example, hubs with 4 or fewer interaction sites were labelled with an accuracy of 72%. However, the accuracy of classification of hubs with 3 or fewer interfaces, the cut-off value for distinguishing SIH and MIH, was considerably lower. The sensitivity of predictions for MIH improves as the number of interfaces of the hub protein increases (See Figure S2). The sensitivity of predicting a protein hub with four or more interfaces is 96% (119/124) and 97% (108/111)

for protein hubs with 5 or more interfaces. The majority (11/18) of our misclassifications were caused by a strong homology between proteins with differing numbers of interaction partners that were 2 or fewer (e.g. a protein with two interaction partners had a strong homology with a protein that had four interaction partners). Details of the misclassifications can be found in Table 5.

Kinetic prediction: discriminating Date from Party hub proteins. To assess kinetic predictions on hub proteins, we created Dataset 4. Dataset 4 contains 199 hub proteins including 91 date hubs and 108 party hubs. Dataset 4 was originally created by Han et al. [31]; this dataset had relatively low sequence homology so no sequences were removed. Tables S5 and S6 show the results of using standard machine learning approaches on this dataset. The accuracies for Dataset 4 ranged from 51.0% to 66.2% and the correlation coefficients from 0.01 to 0.30; precision from 50% to 70% and recall from 42% to 62%. The results of HybSVM on Dataset 4 (see Table 6) provided an accuracy of 69.2% and a correlation coefficient of 0.37, which are comparatively better than the best individual method (NB 3-gram). HybSVM had a marginally lower AUC value (0.68) as compared to the Naive Bayes (0.70). HybSVM also had the best F-Score (62.6) and precision (0.71). Figure 4 shows ROC curves for each of the methods on Dataset 4. The ROC curves were similar to the curves generated by building classifiers on the SIH/MIH dataset. The results show that *HybSVM* method can be used in settings where a low false positive rate is desirable.

A study by Mirzarezaee et al. has recently proposed methods for distinguishing date hub proteins from party hub proteins [15] using a variety of features including 4 composition measurements, 48 physicochemical properties, six GO term features, domain

occurrence presence, disordered regions, and position specific scoring matrices (PSSM). They reported accuracies of up to 77% with correlation coefficients of 0.74 for both date and party hubs. Their dataset also used yeast proteins, but it was a different set of proteins and consisted of over 5,000 non-hub proteins. Their methodology consisted of classifying proteins into the following four classes: non-hub, intermediately connected, date, and party. The *HybSVM* approach we report here focused instead on the binary classification task of distinguishing date hubs from party hubs. Our method does not need functional annotations (GO terms) of proteins, their domain composition, or their sequence alignments with their homologs. Our method also provides probability scores for each prediction. These scores allow an investigator to trade-off the reliability of predictions against the coverage of the predictions. *HybSVM* runs quickly and is easy to implement and update, which are ideal characteristics to serve the method through a web server. A web server implementation of *HybSVM* can be found here: http://hybsvm.gdcb.iastate.edu.

Validating the three-phase approach

To validate our three-phase approach, we tested each phase on additional datasets. Since Dataset 1 (PB versus NPB) was created independently of Dataset 3 (SIH versus MIH) and Dataset 4 (Date versus Party), we used these two datasets along with the data used in the Mirzarezaee paper (Date versus Party) as a test set for the *HybSVM* classifier for predicting protein-binding proteins (Phase I). The union of these three datasets included 900 yeast hub proteins. The Phase I *HybSVM* classifier predicted 99.7% of the proteins as protein-binding proteins. Only three multi-interface proteins were misclassified.

We also used the data from the Mirzarezaee study as an additional test set to predict hub proteins (Phase II) and for the *HybSVM* classifier to discriminate date hubs from party hubs (Phase III). The Phase II classifier correctly predicted 147 proteins as hub proteins and 116 as likely hub-proteins. The classifier misclassified 45 proteins as non-hub proteins and 23 as likely non-hubs proteins (12% error rate). All other proteins were labeled as being of unknown category. The Phase III classifier used to discriminate date and party hub proteins predicted 67.9% of the 546 proteins correctly with a correlation coefficient of 0.36. One of the advantages of our approach over the Mirzarezaee study is that a probability score is assigned to the predictions. In this example, a majority of the misclassifications had a probability score under 0.70. Predictions with higher scores are more reliable. For example, in the case of predictions with score greater than 0.70 (337 proteins), accuracy improves to 74.2% (0.46 correlation coefficient). The predictions with score greater than 0.90 (78 proteins) yield even more reliable results: 84.6% accuracy, and 0.54 correlation coefficient. These results show that investigators can benefit from our method, which needs only sequence information, to control the quality of the predictions by sacrificing the coverage of the classifier. SIH and MIH class labels were not readily available for the Mirzarezaee dataset, so the structural classifier of Phase III was not evaluated on this dataset.

Conclusion

We have demonstrated that it is possible to fairly reliably classify proteins in a three-phase approach: the first phase distinguishes protein-binding (PB) versus non-protein-binding (NPB) proteins; the second phase predicts if the protein is likely to be a hub; the third phase classifies protein-binding proteins into SIH versus MIH and date

versus party hubs. Our approach uses only sequence information and therefore will be highly useful for the analysis of proteins lacking structural information. These classifications provide insights into the structural and kinetic characteristics of the corresponding proteins in the absence of interaction networks, expression data, three-dimensional structure, sequence alignment, functional annotations, domains, or motifs. We note that the performance of our classifier for predicting structural characteristics of hubs (i.e., classifying hubs into SIH versus MIH) is better than that of the classifier for predicting kinetic or expression related characteristics of hubs (i.e., classifying hubs into date versus party hubs).

Materials and Methods

Here we used four datasets for training and testing classifiers for different phases of prediction. Because protein interaction datasets tend to have high false positive rates, when building these datasets, our main goal was to use high-quality data. Our second goal was to remove sequence bias in the datasets. The first dataset consists of proteins that are involved in binding with other proteins (PB) and proteins that are not involved in PB (NPB). This dataset was used in the first phase of our prediction. The second dataset consists of hub and non-hub proteins. This set of proteins was used in the second phase of predictions. Datasets 3 and 4 were used by the third phase to distinguish singlish/multiple and date/party hub proteins (Figure 2).

Dataset 1 – Protein-binding (PB) versus non-protein-binding (NPB) proteins

The first dataset consists of two subsets of proteins. The first subset is generated using high-quality sets of proteins that are known to interact with other proteins. These proteins form a protein-binding (PB) subset. The second subset consists of proteins that

are unlikely to bind with other proteins (NPB). To create the PB subset, 3,640 yeast proteins were downloaded from HINT [9] (High-quality protein interactomes) (http://hint.yulab.org/)--HINT is a database of high-quality protein-protein interactions for different organisms, which was obtained by integrating protein-protein interaction data from various sources and filtered to remove low-quality interactions.

The NPB subset consists of proteins that bind with small molecules, but not with proteins. Identifying such a subset is a challenging task, because the available protein-protein interaction data are incomplete at best. It has been estimated that the fraction of identified interactions of the full human interactome is between 5% and 13% [8,54,55] and up to 30% for the yeast interactome [54]. The efforts to increase the coverage will most likely increase the false positive rate as well [56]. Therefore, it is inevitable that any NPB dataset will be subject to these inherent limitations of incompleteness and incorrectness in experimental protein-protein interaction sets. Considering these limitations, we used the following methodology to create the NPB subset: a set of 8,443 proteins were downloaded from BindingDB [57] (http://www.bindingdb.org/bind/index.jsp). This includes the entire set of protein targets that bind to small-molecules. In order to filter proteins that are interacting with other proteins, these 8,443 BindingDB proteins were BLASTed [58] against the PB set and any protein that had a positive hit was removed. Additional filtering was performed with the remaining BindingDB proteins against the 5,000 yeast proteins that have an experimental protein-protein interaction evidence in BioGrid [50] (http://thebiogrid.org/). The remaining set of non-interacting proteins was 4,567 proteins. To minimize sequence bias, we clustered the proteins in both subsets where at least 80% of the sequence shared 50% or more sequence identity. A representative sequence was

randomly chosen for each cluster to obtain the final dataset. The resulting dataset, Dataset 1, consists of a total of 5,010 proteins including 3,418 proteins in the PB subset and 1,592 proteins in the NPB subset.

Dataset 2 – Hub proteins versus non-hub proteins

Manna et al. [18] had previously created a dataset of hubs and non-hubs. This dataset was originally assembled by downloading human protein-protein interaction data from BioGRID [50]. Any protein that had more than five interactions was labeled as a hub, proteins with fewer than three interactions were labeled as non-hub. Proteins with three, four, or five interactions were not considered as they were near the arbitrary cut-off value for defining a hub and had high potential for being mislabeled. Their resulting dataset included 2,221 hub proteins and 2,889 non-hub proteins. The data ranged from proteins with a single interaction partner to 170 interaction partners. To minimize sequence bias in this dataset, we applied the same methodology we used to obtain Dataset 1: we clustered the protein where at least 80% of each sequence shared 50% or more sequence identity and randomly chose a representative sequence from each cluster. The resulting dataset, Dataset 2, consists of 4,036 proteins including 1,741 hub proteins and 2,295 non-hub proteins.

Dataset 3 - Singlish interface hubs (SIH) versus multiple interface hubs (MIH)

Previously Kim et al. [16] created SIH and MIH datasets by combining yeast interaction data from various sources [4,16,59-63] and associating these proteins with Pfam domains [49,64], which were then subsequently mapped onto known PDB structures using iPfam [65]. They filtered out interactions that were not consistent with protein complexes as defined by iPfam to obtain a yeast protein interaction network. Kim et al.

used this set to analyze evolutionary patterns in hub proteins and it uses a robust structure-based definition of hubs, which was useful in our study. Here, we follow their definition of hubs such that a protein is defined as a hub protein if it has five or more interaction partners. A hub protein with 1 or 2 mutually exclusive interactions is defined as singlish-interface hub (SIH); a hub protein with 3 or more mutually exclusive interactions is defined as multi-interface hub (MIH). The original dataset consists of 1,269 interactions involving 873 proteins with 167 hub proteins including 36 SIH and 131 MIH proteins. We downloaded the original dataset from http://sin.gersteinlab.org. We filtered non-hub proteins out and applied the same sequence filtering that we used for Dataset 1 and Dataset 2. The resulting dataset, Dataset 3, consists of 155 hub proteins with 35 SIH and 120 MIH proteins.

Dataset 4 - Date hubs versus Party hubs

Han et al. [31] created a protein set of date and party hubs by merging the results of multiple methods [4,31,32,59-62,66-71]. Similar to Kim et al. [16], they defined a protein as a hub protein if it has five or more interaction partners. They based their definition of date and party hubs on co-expression patterns: hubs that have low degree of co-expression with their interaction partners (Pearson correlation coefficient of 0.5 or lower) are assumed to bind different partners at different time points or locations and are classified as date hubs. In the same vein, hubs that exhibit high degree of co-expression with their interaction partners (Pearson correlation coefficient greater than 0.5) are assumed to interact simultaneously with their interaction partners. Their resulting yeast interaction dataset consists of 1,379 proteins and 2,493 interactions, which contains both hub and non-hub proteins. We filtered non-hub proteins out and applied the same sequence filtering

that we used on the previous datasets. The sequence bias was already removed in the original dataset; so no additional sequences were removed. The resulting set, Dataset 4 consists of 199 hub proteins including 91 date hubs and 108 party hubs.

Overlap among Dataset 3 and Dataset 4

It is worth noting that the SIH-MIH and Date-Party classes show some overlap. Figure 5 is a Venn diagram showing the distribution of the 41 proteins in their respective classes. For example, 6 singlish-interface proteins are also date hub proteins. Similarly, there are 2 singlish-party hubs, 6 multi-date hubs, and 27 multi-party hubs. Figure 6 show examples of singlish-interface date hub and multi-party hub proteins respectively.

Classification framework

For each class within our dataset we built a binary classifier that predicts if that protein belonged to that class or not. The reported accuracy estimates are based on stratified 10-fold cross validation. Each of the individual classifiers is described below:

Machine learning methods

Support Vector Machines. A support vector machine (SVM), given a training set, that is linearly separable in a kernel-induced feature space, implements a linear decision boundary that maximizes the margin of separation between the classes [46]. If the dataset is not perfectly separable, slack variables are used to minimize the number of misclassified training examples. Logistic regression models were applied to the outputs of the SVM to get a probability score. These scores can be used to evaluate the quality of a prediction. Even if the overall accuracy of the prediction model does not meet high standards, the quality of individual predictions may be suitable based on the probability score. The score also allows an investigator to determine and set the trade-off between the sensitivity and

selectivity of the classified (as shown in the ROC-curves). The scores range from 0.5 (50% or equal probability of belonging to either class) to 1 (100% probability of belonging to the specified class). We used the Weka version 3.6.0 SMO implementation [45] of the support vector machine algorithm [46] with a polynomial kernel.

Naive Bayes Multinomial Classifier. The Naive Bayes multinomial classifier models each sequence by a bag of letters sampled from a fixed alphabet. In our case, the bag of letters is the amino acid composition of a protein sequence. Thus, each element (amino acid) of the sequence is assumed to be independent of the other elements in the sequence given the class label. Based on this assumption, a multinomial Naive Bayes classifier can be built over all of the sequences for a given class. This is similar to the bag of words approach previously used for text classification [72].

Naive Bayes k-grams Classifier. The Naive Bayes *k*-grams (NB *k*-grams) method [73] uses a sliding a window of size *k* along each sequence to generate a *bag* of *k*-grams representation of the sequence. Much like in the case of the Naive Bayes classifier described above, the Naïve Bayes *k*-grams classifier treats each *k*-gram in the bag to be independent of the others given the class label for the sequence. Given this probabilistic model, the previously outlined method for classification using a probabilistic model can be applied. A problem with the NB *k*-grams approach is that successive *k*-grams extracted from a sequence share *k*-1 elements in common. This grossly and systematically violates the independence assumption of Naive Bayes.

NB (k). NB(k) [73] constructs, for each class, a Markov model of order *k* -1. It modifies the Naïve Bayes model to explicitly model the dependencies (of order *k-1*) between the letters of a sequence. It is easily seen that when *k* = 1, Naive Bayes k-grams as

well as Naive Bayes (1) reduce to the Naive Bayes model. The relevant probabilities required for specifying the above models can be estimated using standard techniques for estimation of probabilities using Laplace estimators [74].

Naïve Bayes (NB) *k*-grams and NB(*k*) models were constructed and evaluated on the dataset with *k* ranging from 1 to 4. Values of *k* larger than 4 were not considered because at higher values of *k*, the available data are insufficient to obtain reliable probability estimates.

PSI-Blast. The homology-based tool PSI-BLAST [51] version 2.2.9 was used to construct a binary classifier for each class. We used the binary class label predicted by the PSI-BLAST-based classifier as an additional input to our *HybSVM* classifier. Given a query sequence to be classified, we used PSI-BLAST to compare the query sequence against the training set. In 10-fold cross-validation, we ran PSI-BLAST with the query sequences in a given fold against the reference database comprised of the remaining nine folds. We repeated this process for each of the ten folds. For *HybSVM,* a class was assigned to the query sequence based on the top-scoring hit (i.e., the sequence with the lowest e-value) from the PSI-BLAST results. The resulting binary prediction of the PSI-BLAST classifier for class *c* is 1 if the class label for the top scoring hit is *c*. Otherwise, it is 0. An e-value cut-off of $1x10^{-4}$ was used for PSI-BLAST, with all other parameters set to their default values. For predicting hub proteins, the four top-scoring hit were used. If there was a consensus among the top four hits then the class label of the four hits is assigned to the query sequence. If three of the four top-scoring hits had the same class, the query sequence is labeled as 'likely' belonging to that class. In addition to the prediction, we report the number of interaction partners belonging to the top hit, the range of

interaction partners of the top four hits, and a weighted average (based on the BLAST score) of number of interaction partners of the top four hits.

Domain-based Method. The domain-based method builds a classifier by using a class-conditional probability distribution based on the frequency of SCOP [47,48] and PFAM [49] domains in the following manner. For each protein, the count for each type of domain was determined by the number of domains listed at the Saccharomyces Genome Database (SGD) [75]. This method was used to rule out a simple direct correlation between the number of domains and the number of interaction sites on a hub protein.

Decision Tree. A Decision Tree builds a predictive model by recursively partitioning the dataset based on choosing features that provide the most information gain. In our example, the feature set is the observed k-gram composition of amino acids given a class. For binary classes (e.g., SIH versus MIH), the decision tree algorithm chooses a *k*-gram feature that partitions the data to maximize the information gain between classes. The process is recursively repeated on the new partitions until no more information gain can be achieved. Additional techniques are performed (e.g., pruning) to help prevent overtraining. For these experiments we used the commonly used decision tree algorithm C4.5 [44] implemented as the J4.8 algorithm [43] in Weka version 3.6.0.

Multi-layer perceptron. A multi-layer perceptron, often referred to as a multilayer artificial neural network [76,77] (ANN) implements a non-linear decision function by using a weighted linear combination of non-linear (typically sigmoid) transformations of linear functions of input features. The ANN is typically trained using error back-propagation or generalized gradient descent algorithm that minimizes a function of the

error between the desired and actual outputs of the ANN. We used the Multilayer Perceptron artificial neural network (ANN) implementation [43] in Weka version 3.6.0.

HybSVM Method. We introduce *HybSVM* classifier that is a support vector machine (SVM) classifier that assigns the class label to a target sequence based on the class labels output by the Naive Bayes (NB), NB k-gram, NB(k) classifiers, and an additional attribute, the output from the PSI-BLAST classifiers. Since there are eight classifiers Naive Bayes, NB 2-gram, NB 3-gram, NB 4-gram, NB(2), NB(3), NB(4), and PSI-BLAST, the input to the *HybSVM* classifier consists of a 8-tuple vector of class labels assigned to the sequence by the 8 classifiers. A SVM is trained to predict the class label for each sequence based on the 8-tuple of class labels predicted by the eight individual classifiers.

Performance evaluation

The performance measures *accuracy, precision, recall, f-measure, and correlation coefficient* are used to evaluate the performance of our machine learning approaches [78]. *Accuracy* is the fraction of overall predictions that are correct. *Precision* is the ratio of predicted true positive examples to the total number of actual positive examples. *Recall* is the ratio of predicted true positives to the total number of examples predicted as positive. The *F-measure* (F1 score) is the harmonic mean of precision and recall. The F-measure has a range between 0 (worst value) and 1 (best value). *Correlation coefficient* measures the correlation between predictions and actual class labels. The correlation coefficient has a range of -1 (worst value) and 1 (best value).

Table S7 summarizes the definitions of performance measures in the two-class setting (binary classification), where M represents the total number of classes and N

represents the total number of examples. *TP, TN, FP,* and *FN* are the true positives, true negatives, false positives, and false negatives for each of our classification problems. For example, when predicting SIH proteins: TP refers to a SIH correctly predicted, FP to MIH predicted as SIH, FN as SIH predicted as a MIH, and TN to MIH correctly predicted.

Where possible, we used the area under the receiver operating characteristic (AUC) curve. The ROC curve plots the true positive rate versus false positive rate for a binary classifier system (a protein belongs to a given class or not) as its discrimination threshold is varied. The area ranges from 0 (worst) to 1 (best); the value of 0.5 refers to the expected value of a random method.

Acknowledgements

We would like to thank Rasna Walia (Iowa State University) and Li Xue (Iowa State University) for their critical reading of the manuscript. The work of Vasant Honavar while working at the National Science Foundation was supported by the National Science Foundation. Any opinion, finding, and conclusions contained in this article are those of the authors and do not necessarily represent the views of the National Science Foundation.

References

1. Eisenberg D, Marcotte EM, Xenarios I, Yeates TO (2000) **Protein function in the post-genomic era**. *Nature* 405: 823-826.

2. Yu H, Braun P, Yildirim MA, Lemmens I, Venkatesan K, et al. (2008) **High-quality binary protein interaction map of the yeast interactome network**. *Science* 322: 104-110.

3. Dreze M, Monachello D, Lurin C, Cusick ME, Hill DE, et al. (2010) **High-quality binary interactome mapping.** *Methods Enzymol* 470: 281-315.

4. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, et al. (2000) **A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae.** *Nature* 403: 623-627.

5. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, et al. (2001) **A comprehensive two-hybrid analysis to explore the yeast protein interactome.** *Proc Natl Acad Sci* U S A 98: 4569-4574.

6. Walhout AJ, Vidal M (2001) **High-throughput yeast two-hybrid assays for large-scale protein interaction mapping.** *Methods* 24: 297-306.

7. Huang H, Jedynak BM, Bader JS (2007) **Where have all the interactions gone? Estimating the coverage of two-hybrid protein interaction maps.** *PLoS Comput Biol* 3: e214.

8. Venkatesan K, Rual JF, Vazquez A, Stelzl U, Lemmens I, et al. (2009) **An empirical framework for binary interactome mapping.** *Nat Methods* 6: 83-90.

9. Das J, Yu H (2012) **HINT: High-quality protein interactomes and their applications in understanding human disease.** *BMC Syst Biol* 6: 92.

10. Gursoy A, Keskin O, Nussinov R (2008) **Topological properties of protein interaction networks from a structural perspective.** *Biochem Soc Trans* 36: 1398-1403.

11. Kuzu G, Keskin O, Gursoy A, Nussinov R (2012) **Constructing structural networks of signaling pathways on the proteome scale.** *Curr Opin Struct Biol*.

12. Liu S, Zhu X, Liang H, Cao A, Chang Z, et al. (2007) **Nonnatural protein-protein interaction-pair design by key residues grafting.** *Proc Natl Acad Sci* U S A 104: 5330-5335.

13. Grigoryan G, Reinke AW, Keating AE (2009) **Design of protein-interaction specificity gives selective bZIP-binding peptides**. *Nature* 458: 859-864.

14. Fleishman SJ, Whitehead TA, Ekiert DC, Dreyfus C, Corn JE, et al. (2011) **Computational design of proteins targeting the conserved stem region of influenza hemagglutinin.** *Science* 332: 816-821.

15. Mirzarezaee M, Araabi BN, Sadeghi M (2010) **Features analysis for identification of date and party hubs in protein interaction network of Saccharomyces Cerevisiae**. *BMC Syst Biol* 4: 172.

16. Kim PM, Lu LJ, Xia Y, Gerstein MB (2006) **Relating three-dimensional structures to protein networks provides evolutionary insights.** *Science* 314: 1938-1941.

17. Ekman D, Light S, Bjorklund AK, Elofsson A (2006) W**hat properties characterize the hub proteins of the protein-protein interaction network of Saccharomyces cerevisiae?** *Genome Biol* 7: R45.

18. Manna B, Bhattacharya T, Kahali B, Ghosh TC (2009) **Evolutionary constraints on hub and non-hub proteins in human protein interaction network: insight from protein connectivity and intrinsic disorder.** *Gene* 434: 50-55.

19. Keskin O, Gursoy A, Ma B, Nussinov R (2008) **Principles of protein-protein interactions: what are the preferred ways for proteins to interact?** *Chem Rev* 108: 1225-1244.

20. Jeong H, Mason SP, Barabasi AL, Oltvai ZN (2001) **Lethality and centrality in protein networks**. *Nature* 411: 41-42.

21. Missiuro PV, Liu K, Zou L, Ross BC, Zhao G, et al. (2009) **Information flow analysis of interactome networks.** *PLoS Comput Biol* 5: e1000350.

22. Zotenko E, Mestre J, O'Leary DP, Przytycka TM (2008) W**hy do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality.** *PLoS Comput Biol* 4: e1000140.

23. Aragues R, Sali A, Bonet J, Marti-Renom MA, Oliva B (2007) **Characterization of protein hubs by inferring interacting motifs from protein interactions.** *PLoS Comput Biol* 3: 1761-1771.

24. Bellay J, Han S, Michaut M, Kim T, Costanzo M, et al. (2011) **Bringing order to protein disorder through comparative genomics and genetic interactions.** *Genome Biol* 12: R14.

25. Cukuroglu E, Gursoy A, Keskin O (2010) **Analysis of hot region organization in hub proteins.** *Ann Biomed Eng* 38: 2068-2078.

26. Fong JH, Panchenko AR (2010) Intrinsic disorder and protein multibinding in domain, terminal, and linker regions. Mol Biosyst 6: 1821-1828.

27. Kahali B, Ahmad S, Ghosh TC (2009) **Exploring the evolutionary rate differences of party hub and date hub proteins in Saccharomyces cerevisiae protein-protein interaction network.** *Gene* 429: 18-22.

28. Kim PM, Sboner A, Xia Y, Gerstein M (2008) **The role of disorder in interaction networks: a structural analysis**. *Mol Syst Biol* 4: 179.

29. Pang K, Cheng C, Xuan Z, Sheng H, Ma X (2010) **Understanding protein evolutionary rate by integrating gene co-expression with protein interactions.** *BMC Syst Biol* 4: 179.

30. Patil A, Kinoshita K, Nakamura H (2010) **Hub promiscuity in protein-protein interaction networks**. *Int J Mol Sci* 11: 1930-1943.

31. Han JD, Bertin N, Hao T, Goldberg DS, Berriz GF, et al. (2004) **Evidence for dynamically organized modularity in the yeast protein-protein interaction network**. *Nature* 430: 88-93.

32. Agarwal S, Deane CM, Porter MA, Jones NS (2010) **Revisiting date and party hubs: novel approaches to role assignment in protein interaction networks**. *PLoS Comput Biol* 6: e1000817.

33. Bertin N, Simonis N, Dupuy D, Cusick ME, Han JD, et al. (2007) **Confirmation of organized modularity in the yeast interactome.** *PLoS Biol* 5: e153.

34. Afridi TH, Khan A, Lee YS (2011) **Mito-GSAAC: mitochondria prediction using genetic ensemble classifier and split amino acid composition**. *Amino Acids*.

35. Kar G, Gursoy A, Keskin O (2009) **Human cancer protein-protein interaction network: a structural perspective.** *PLoS Comput Biol* 5: e1000601.

36. Tsai CJ, Ma B, Nussinov R (2009) **Protein-protein interaction networks: how can a hub protein bind so many different partners?** *Trends Biochem Sci* 34: 594-600.

37. Kar G, Kuzu G, Keskin O, Gursoy A (2012) **Protein-protein interfaces integrated into interaction networks: implications on drug design**. *Curr Pharm Des*.

38. Jin G, Zhang S, Zhang XS, Chen L (2007) **Hubs with network motifs organize modularity dynamically in the protein-protein interaction network of yeast**. *PLoS One* 2: e1207.

39. Taylor IW, Linding R, Warde-Farley D, Liu Y, Pesquita C, et al. (2009) **Dynamic modularity in protein interaction networks predicts breast cancer outcome**. *Nat Biotechnol* 27: 199-204.

40. Hsing M, Byler K, Cherkasov A (2009) **Predicting highly-connected hubs in protein interaction networks by QSAR and biological data descriptors**. *Bioinformation* 4: 164-168.

41. Nair MTaAS (2009) **Prediction and disorderliness of hub proteins**. International *Journal of Bioinformatics Research* 1: 70-80.

42. Latha AB, Nair AS, Sivasankaran A, Dhar PK (2011) **Identification of hub proteins from sequence**. *Bioinformation* 7: 163-168.

43. Frank E, Hall M, Trigg L, Holmes G, Witten IH (2004) **Data mining in bioinformatics using Weka**. *Bioinformatics* 20: 2479-2481.

44. Quinlan JR (1993) **C4.5: Programs for Machine Learning.** San Francisco, CA, USA: Morgan Kaufmann Publishers.

45. Platt J (1998) **Fast training of support vector machines using sequential minimal optimization.** In: Scholkopf B, Burges CJC, Smola AJ, editors. Advances in Kernel Methods - Support Vector Learning. Cambridge, MA, USA: MIT Press.

46. Vapnik V (1998) **Statistical learning theory**. New York, NY, USA: Wiley.

47. Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C, et al. (2004) **SCOP database in 2004: refinements integrate structure and sequence family data.** *Nucleic Acids Res* 32: D226-229.

48. Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJ, et al. (2008) **Data growth and its impact on the SCOP database: new developments**. *Nucleic Acids Res* 36: D419-425.

49. Finn RD, Mistry J, Tate J, Coggill P, Heger A, et al. (2010) **The Pfam protein families database**. *Nucleic Acids Res* 38: D211-222.

50. Stark C, Breitkreutz BJ, Chatr-Aryamontri A, Boucher L, Oughtred R, et al. (2011) **The BioGRID Interaction Database: 2011 update**. *Nucleic Acids Res* 39: D698-704.

51. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 25: 3389-3402.

52. Ofran Y, Rost B (2007) **ISIS: interaction sites identified from sequence**. *Bioinformatics* 23: e13-16.

53. Xue LC, Dobbs D, Honavar V (2011) **HomPPI: a class of sequence homology based protein-protein interface prediction methods**. *BMC Bioinformatics* 12: 244.

54. Baker M (2012) **Proteomics: The interaction map**. Nature 484: 271-275.

55. Liu CH, Li KC, Yuan S (2012) **Human Protein-Protein Interaction Prediction by A Novel Sequence-Based Coevolution Method: Coevolutionary Divergence.** *Bioinformatics*.

56. De Las Rivas J, Fontanillo C (2010) **Protein-protein interactions essentials: key concepts to building and analyzing interactome networks**. *PLoS Comput Biol* 6: e1000807.

57. Liu T, Lin Y, Wen X, Jorissen RN, Gilson MK (2007) **BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities**. *Nucleic Acids Res* 35: D198-201.

58. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) **Basic local alignment search tool**. *J Mol Biol* 215: 403-410.

59. Mewes HW, Frishman D, Guldener U, Mannhaupt G, Mayer K, et al. (2002) **MIPS: a database for genomes and protein sequences**. *Nucleic Acids Res* 30: 31-34.

60. Ito T, Tashiro K, Muta S, Ozawa R, Chiba T, et al. (2000) **Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins**. *Proc Natl Acad Sci* U S A 97: 1143-1147.

61. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, et al. (2002) **Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry**. *Nature* 415: 180-183.

62. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, et al. (2002) **Functional organization of the yeast proteome by systematic analysis of protein complexes.** *Nature* 415: 141-147.

63. Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM, et al. (2002) **DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions.** *Nucleic Acids Res* 30: 303-305.

64. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, et al. (2002) **The Pfam protein families database.** *Nucleic Acids Res* 30: 276-280.

65. Finn RD, Marshall M, Bateman A (2005) **iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions**. *Bioinformatics* 21: 410-412.

66. Bader GD, Betel D, Hogue CW (2003) **BIND: the Biomolecular Interaction Network Database**. *Nucleic Acids Res* 31: 248-250.

67. Dandekar T, Snel B, Huynen M, Bork P (1998) **Conservation of gene order: a fingerprint of proteins that physically interact.** *Trends Biochem Sci* 23: 324-328.

68. Fromont-Racine M, Mayes AE, Brunet-Simon A, Rain JC, Colley A, et al. (2000) **Genome-wide protein interaction screens reveal functional networks involving Sm-like proteins.** *Yeast* 17: 95-110.

69. Fromont-Racine M, Rain JC, Legrain P (1997) **Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens**. *Nat Genet* 16: 277-282.

70. Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D (1999) **A combined algorithm for genome-wide prediction of protein function**. *Nature* 402: 83-86.

71. von Mering C, Krause R, Snel B, Cornell M, Oliver SG, et al. (2002) **Comparative assessment of large-scale data sets of protein-protein interactions**. *Nature* 417: 399-403.

72. Andrew Mccallum KN (1998) **A Comparison of Event Models for Naive Bayes Text Classification**. AAAI-98 Workshop on 'Learning for Text Categorization'. pp. 41-48.

73. Andorf C, Dobbs D, Honavar V (2007) **Exploring inconsistencies in genome-wide protein function annotations: a machine learning approach**. *BMC Bioinformatics* 8: 284.

74. Mitchell T (1997) **Machine learning**. New York, NY, USA: McGraw Hill.

75. Nash R, Weng S, Hitz B, Balakrishnan R, Christie KR, et al. (2007) **Expanded protein information at SGD: new pages and proteome browser**. *Nucleic Acids Res* 35: D468-471.

76. McCulloch WS, Pitts W (1990) **A logical calculus of the ideas immanent in nervous activity**. 1943. *Bull Math Biol* 52: 99-115; discussion 173-197.

77. Rosenblatt R (1962) **Principles of Neurodynamics**. New York: Spartan Books.

78. Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H (2000) **Assessing the accuracy of prediction algorithms for classification: an overview**. *Bioinformatics* 16: 412-424.

79. Ignatev A, Kravchenko S, Rak A, Goody RS, Pylypenko O (2008) **A structural model of the GDP dissociation inhibitor rab membrane extraction mechanism**. *J Biol Chem* 283: 18377-18384.

80. Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, et al. (2002) **The Protein Data Bank.** *Acta Crystallogr D Biol Crystallogr* 58: 899-907.

81. Rak A, Pylypenko O, Durek T, Watzke A, Kushnir S, et al. (2003) **Structure of Rab GDP-dissociation inhibitor in complex with prenylated YPT1** GTPase. *Science* 302: 646-650.

82. Groll M, Schellenberg B, Bachmann AS, Archer CR, Huber R, et al. (2008) A **plant pathogen virulence factor inhibits the eukaryotic proteasome by a novel mechanism**. *Nature* 452: 755-758.

Figures

Figure 1. Descriptions of the singlish-date, singlish-party, multi-date, and multi-party classes - Descriptions for each type of hub are described below. The rows of the table represent the singlish and multi-interface hub proteins. The columns represent the date and party hubs. The intersection of the column and row displays a picture showing examples of the type and number of interfaces involved for that class.

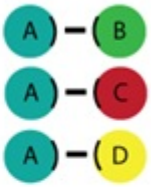
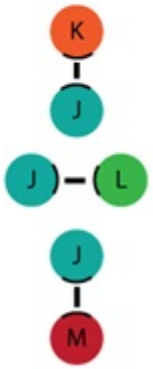
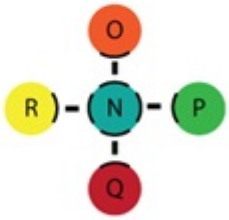
	Date Hub that binds with partners at different times and locations	**Party** Hub that binds with most of its partners simultaneously
Singlish Hub with one or two binding sites	 A interacts with B, C, D at one interaction site at three different times and locations.	rarely occurs
Multi Hub with three or more binding sites	 J interacts with K, L, M at three different interaction sites and locations.	 N simultaneously interacts with O, P, Q, R at four different interaction sites.

Figure 2. Three-phase method to predict protein-binding proteins, hub proteins, singlish interface/multiple interface (SIH/MIH), and Date/Party hubs - Phase I predicts if a protein physically binds with other proteins (protein-binding (PB) versus non-protein-binding (NPB)). If a protein is predicted to be a PB protein in Phase I, that protein is further classified in Phase II and Phase III. Phase II uses sequence similarity to determine the potential number of interaction sites for the input sequence and if that protein is likely to be a hub protein. Phase III applies methods for predicting both structural (singlish vs. multiple) and kinetic (date vs. party) classifications of protein hub proteins. All methods for each of the three phases make predictions from sequence alone.

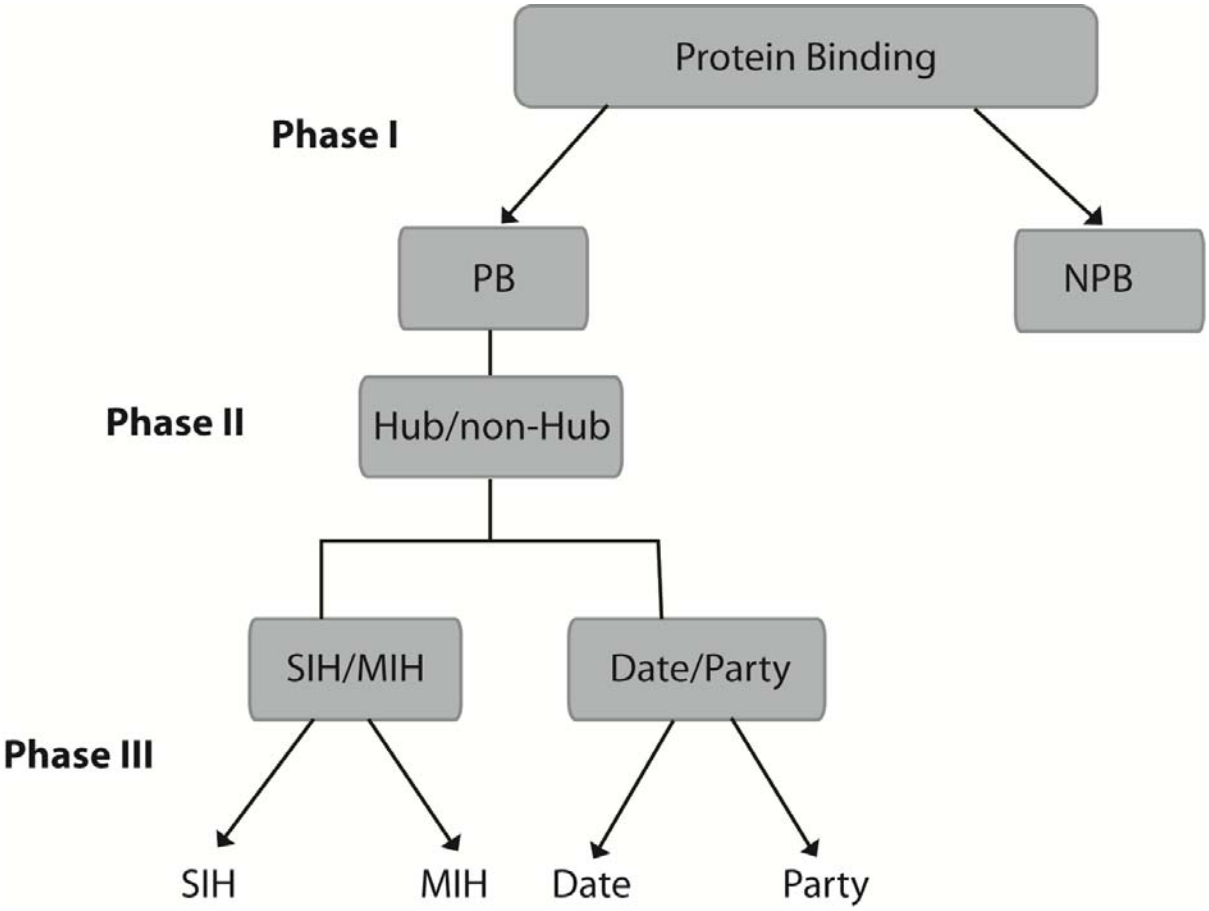


Figure 3. HybSVM method - *HybSVM* is a two-stage machine learning method. The first step of the algorithm is to convert sequence data into a composition-based data representation (monomer, dimer, trimer, and tetramer). These four new data representations are used as inputs to 7 machine learning algorithms based on the NB(k) and NB k-gram approaches (Stage 1). An eighth method based on PSI-BLAST is applied to the original sequence data. The outputs of each of the eight outputs are converted into a binary vector of length 8. The resulting vector is used as input to a SVM to produce the final output (Stage 2).

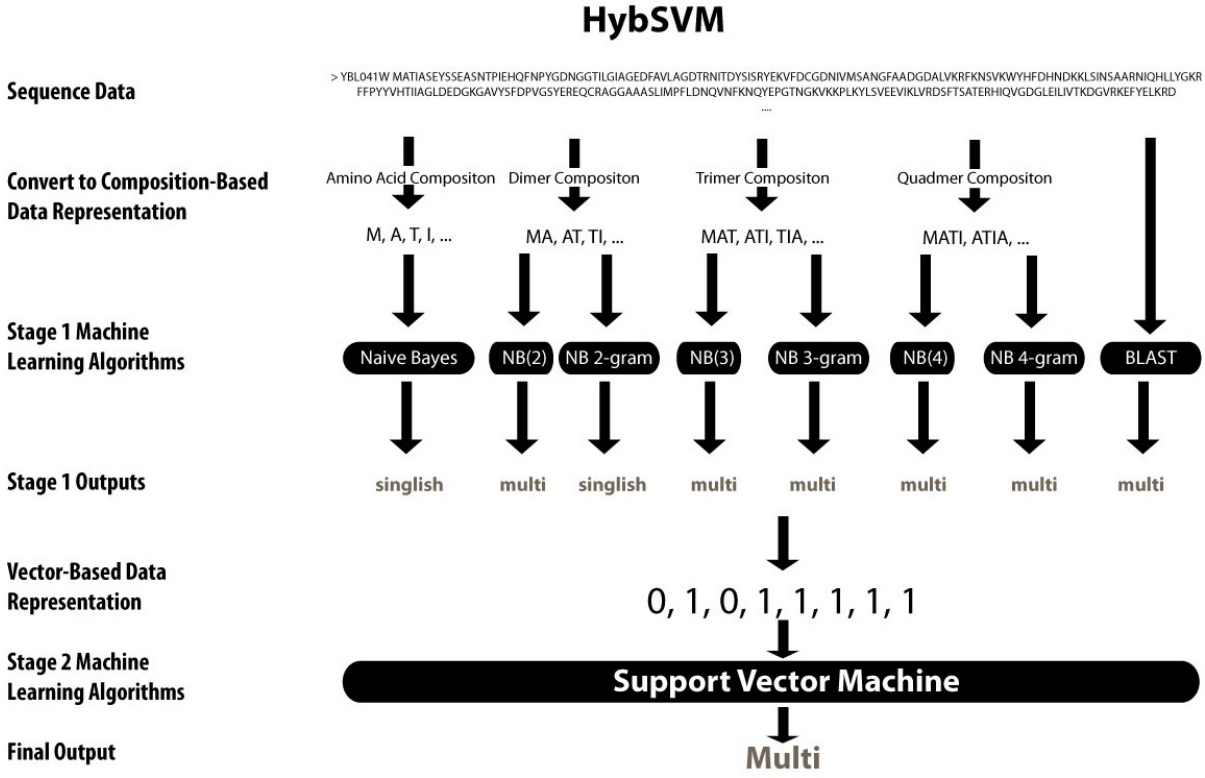


Figure 4. Receiver-operator characteristics (ROC) curve for Datasets 1, 3, and 4 - The curve describes the tradeoff between sensitivity and specificity at different thresholds for various predictors. A simple domain-based method is included as a baseline for comparison. The figure includes ROC curves for protein-binding (PB) versus non-protein-binding (NPB), singlish-interface versus multi-interface hub proteins, and date versus party hub proteins.

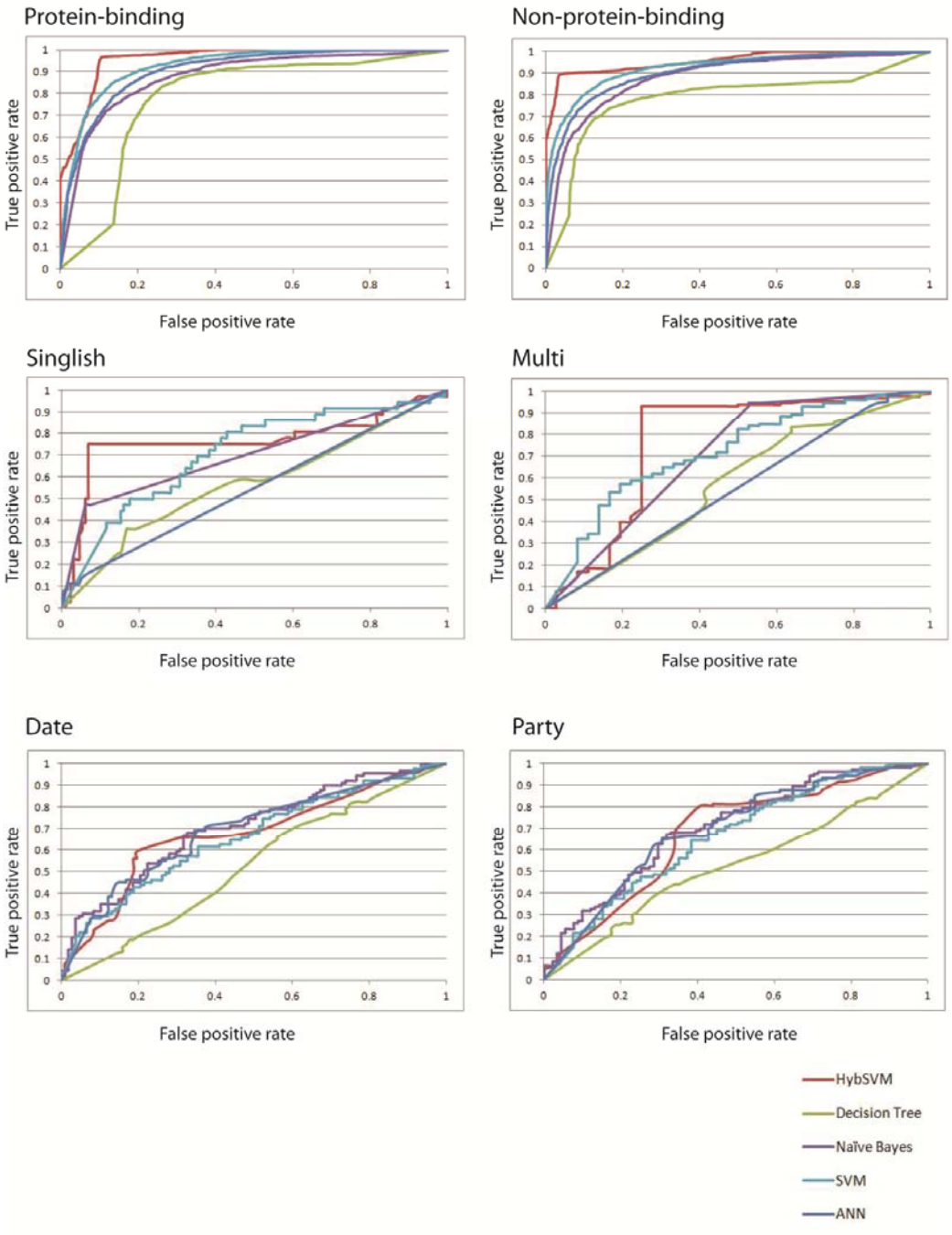


Figure 5. Venn-diagram for Dataset 3 and Dataset 4 - Each of the 272 hub proteins belong to one or more of the following classes: singlish, multi, date, party. Dataset 3 consists of 35 singlish hub proteins and 120 multi hub proteins (Yellow circles). Dataset 4 consists of 91 date hub proteins and 108 party hub proteins (Blue circles). Please see text for more details about the datasets.

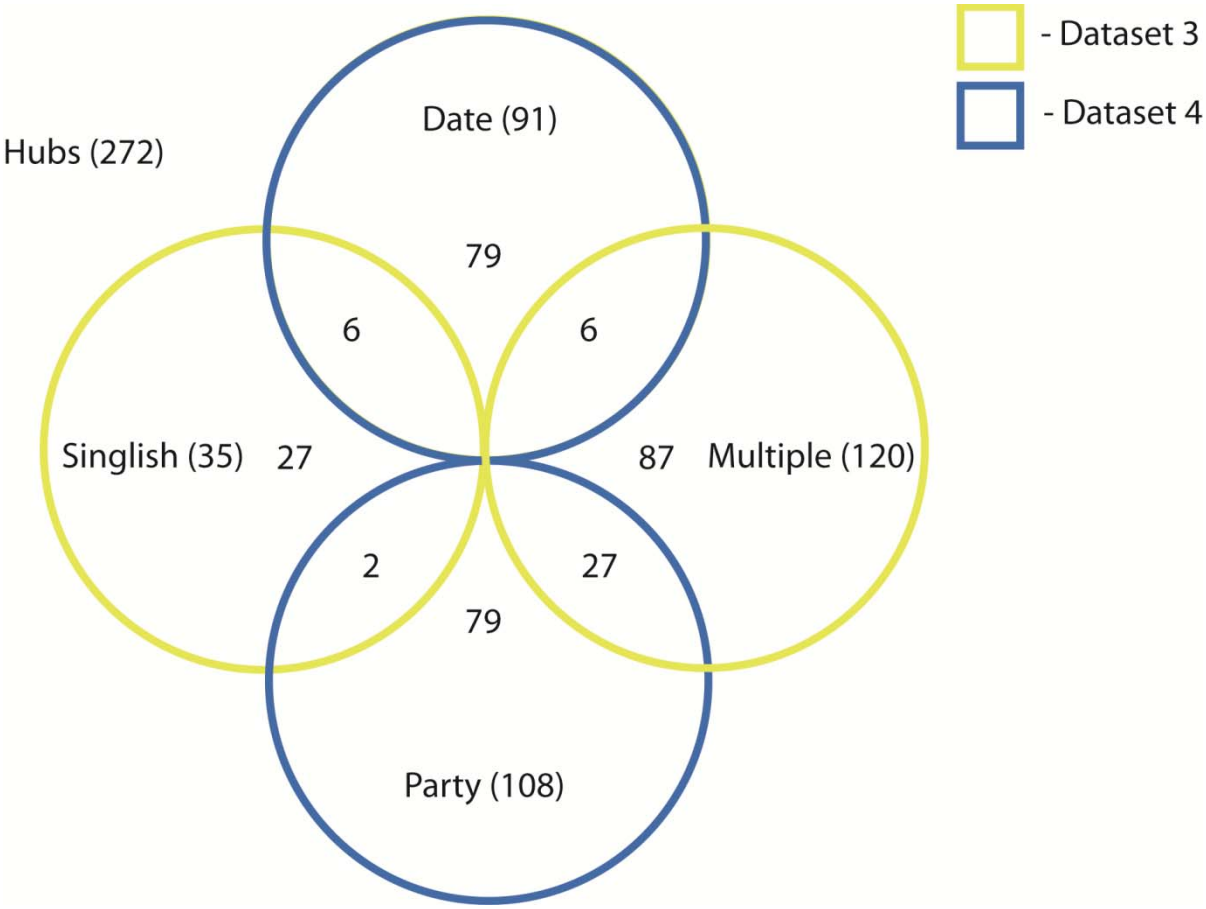
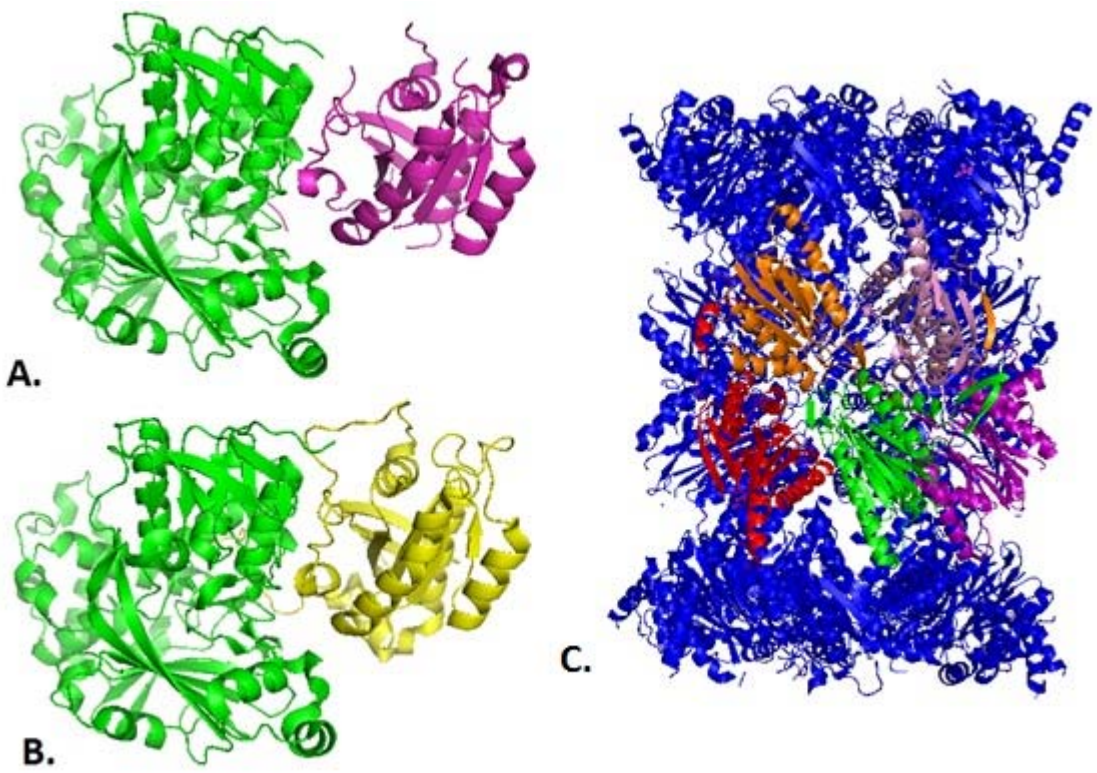


Figure 6. Example of a singlish-interface date and multi-party hub proteins - Images A and B show the quaternary structure for the singlish-date protein Rab GDP dissociation inhibitor alpha (GDI1, YER136W) binding with two different proteins. Image C shows the quaternary structure for the yeast protein beta 6 subunit of the 20S proteasome (PRE7, YBL041W) binding with multiple proteins at the same time. **A:** GDI1 (green) binding with GTP-binding protein YPT31/YPT8 (purple). PDB ID of the complex: 3cpj [79,80]. **B:** GDI1 (green) binding with GTP-binding proteinYPT1 (yellow). PDB ID of the complex: 1ukv [80,81]. The protein binds at one location (singlish-interface) with one partner at a time (date). **C:** PRE7 (green) binds with PUP1 (orange), PUP3 (red), C5 (pink), PRE4 (purple). PDB ID of the complex: 3bdm [80,82]. The protein binds at multiple locations (multi-interface) with many partners at same time (party).



Tables

Table 1. Properties of singlish and multiple-interface yeast protein hubs.

Properties	**Singlish-interface**	**Multiple-interface**
Essential	No	Yes
Conserved	No	Yes
Co-expression	Limited	High
3D Structure	Smaller, less stable	Larger, stable
Canonical preferential gene duplication	Yes	No
Disorder	High	Low

The properties for each type of interface are listed based on observed tendencies seen in Dataset 3 [16,28].

Table 2. Properties of date and party yeast protein hubs.

Properties	**Date**	**Party**
Evolutionary rate	Faster	Slower
Interactome connectivity	Intermodule	Intramodule
Structural interaction	Few interaction sites	Many interaction sites
Hot spots	More organized in hot regions	Less organized in hot regions
Hot regions	Covers a larger fraction of the interface region, larger number of distinct hot regions	Covers a smaller fraction of the interface region, smaller numbe distinct hot regions

The properties for each type of interface are listed based on observed tendencies seen in Dataset 4 [25,31,33].

Table 3. Dataset 1 (protein-binding vs. non-protein-binding, i.e. PB vs. NPB) prediction results from classifiers trained using machine learning methods.

Approach	Best k	Accuracy	F1 Score	Precision	Recall	C.C.	AUC
NB k-gram	4	88.2	80.1	.75	.86	.72	.85
NB(k)	3	86.4	78.5	.79	.78	.69	.83
Decision Tree	1	81.6	69.9	.72	.68	.57	.78
SVM	2	87.2	78.9	.82	.76	.70	.84
ANN	2	86.9	77.6	.83	.73	.71	.84
Naive Bayes	2	82.1	72.9	.70	.76	.60	.88
Domain-based	N/A	68.2	0.0	.00	.00	.00	.50
Homology-based	N/A	52.7	49.1	.37	.73	.15	N/A
HybSVM	N/A	**94.2**	**90.5**	**.92**	**.89**	**.87**	**.93**

Accuracy, F-measure (F1 Score), precision, recall, correlation coefficient (C.C.), and area under the receiver operating characteristic curve (AUC) of classification for the multi-interface versus singlish-interface dataset are presented. Accuracy and F-measure are reported in percentage. For each machine learning approach, values of k ranged from 1 to 4. Only the classifier with the best performing k-value (as defined by highest correlation coefficient) is shown. Our methods were estimated by cross-validation. The highest performing value(s) for each performance measure is highlighted in bold.

Table 4. Dataset 3 (SIH vs. MIH) prediction results from classifiers trained using machine learning methods.

Approach	Best k	Accuracy	F1 Score	Precision	Recall	C.C.	AUC
NB k-gram	4	83.8	53.8	.42	.75	.47	.71
NB(k)	3	83.2	45.3	.31	**.84**	.44	.69
Decision Tree	3	71.0	40.3	.38	.43	.21	.57
SVM	2	76.1	41.0	.46	.37	.27	.62
ANN	2	79.0	14.1	.60	.08	.10	.55
Naive Bayes	3	81.2	52.8	.62	.46	.41	.70
Domain-based	N/A	76.4	0.0	.00	.00	-.01	.42
Homology-based	N/A	66.4	46.6	.74	.34	.32	N/A
HybSVM	N/A	**89.0**	**76.0**	**.75**	.77	**.69**	**.85**

Accuracy, F-measure (F1 Score), precision, recall, correlation coefficient (C.C.), and area under the receiver operating characteristic curve (AUC) of classification for the multi-interface versus singlish-interface dataset are presented. Accuracy and F-measure are reported in percentage. For each machine learning approach, values of k ranged from 1 to 4. Only the classifier with the best performing k-value (as defined by highest correlation coefficient) is shown. Our methods were estimated by cross-validation. The highest performing value(s) for each performance measure is highlighted in bold.

Table 5. Details for misclassified proteins in Dataset 3 using HybSVM.

	Misclassified proteins		Protein with highest homology			
Class	**Gene**	**Interfaces**	**Gene**	**Interfaces**	**e-value**	**Difference**
singlish	RHO1	3	CDC42	2	1E-52	1
singlish	STE11	4	CLA4	2	2E-37	2
singlish	ARP2	3	ACT1	2	1E-102	1
singlish	MAK5	9	DRS1	2	6E-42	7
singlish	PRP28	8	DRS1	2	2E-49	6
singlish	PUB1	3	SGN1	1	1E-05	2
singlish	CMD1	4	MLC1	2	1E-15	2
singlish	SEC22	7	YKT6	2	9E-10	5
singlish	SNP1	3	SGN1	1	8E-05	2
multi	YKT6	2	SEC22	7	9E-05	-5
multi	CDC42	2	RHO1	3	1E-05	-1
multi	ACT1	2	ARP2	3	1E-05	-1
multi	SGN1	1	PUB1	3	1E-05	-2
multi	YTA7	1	RPT4	6	8E-05	-5
multi	MLC1	2	CMD1	4	1E-05	-2
multi	MTR3	1	(No hit)	N/A	N/A	N/A
multi	BOI2	2	(No hit)	N/A	N/A	N/A
multi	CLA4	2	STE11	4	2E-05	-2

Details for the misclassified proteins in Dataset 3 based on using the *HybSVM* method including: actual class (multi, singlish), gene name, and number of interfaces as predicted by Kim et al. [16] are shown. For each misclassified protein, information about the protein with the highest homology based on the nearest BLAST hit is also shown. This information includes: gene name**,** number of interfaces as predicted by Kim et al. [16], e-value of the BLAST results between the two proteins, and the difference between the number of predicted interfaces for the misclassified protein and its nearest BLAST hit.

Table 6. Dataset 4 (Date vs. Party hubs) predictions from classifiers trained using machine learning methods.

Approach	Best k	Accuracy	F1 Score	Precision	Recall	C.C.	AUC
NB k-gram	3	67.1	59.8	.54	**.67**	.33	**.71**
NB(k)	3	65.1	58.0	.53	.64	.29	.65
Decision Tree	1	53.5	55.4	.50	.62	.08	.53
SVM	3	62.1	59.0	.59	.59	.24	.66
ANN	2	66.2	55.5	.70	.46	.30	.69
Naive Bayes	1	65.2	57.5	.66	.51	.29	.70
Domain-based	N/A	59.1	30.2	.62	.20	.14	.57
Homology-based	N/A	29.8	22.0	.22	.22	-.43	N/A
HybSVM	N/A	**69.2**	**62.6**	**.71**	.56	**.37**	.68

Accuracy, F-measure (F1 Score), precision, recall, correlation coefficient (C.C.), and area under the receiver operating characteristic curve (AUC) of classification for the multi-interface versus singlish-interface dataset are presented. Accuracy and F-measure are reported in percentage. For each machine learning approach, values of k ranged from 1 to 4. Only the classifier with the best performing k-value (as defined by highest correlation coefficient) is shown. Our methods were estimated by cross-validation. The highest performing value(s) for each performance measure is highlighted in bold.

Supporting Information

Supporting Figures and Tables can be found in Appendices J -R.

Appendix J: – Figure S1

The accuracy curve of predicting singlish-interface and multiple-interface hub proteins as a function of the number of interaction sites. The curve shows the prediction accuracy for proteins with number of interactions sites less than the given maximum threshold. For example, the value of 5 on the x-axis refers to all hub proteins with 5 or fewer interfaces and the value on the curve (83%) at x=5, represents the accuracy of this set.

Appendix K: – Figure S2

The sensitivity curve of predicting singlish-interface and multiple-interface hub proteins as a function of the number of interaction sites. The curve shows the prediction accuracy for proteins with number of interactions sites more than the given minimum threshold. For example, the value of 5 on the x-axis refers to all hub proteins with 5 or more interfaces and the value on the curve (97%) at x=5, represents the sensitivity of this set.

Appendix L: – Table S1

Accuracy, precision, recall, and correlation coefficient (CC) of classification for the protein-binding versus non-protein-binding dataset are presented for internal machine learning methods. For each machine learning approach, values of k ranged from 1 to 4. The performance of the results were estimated using cross-validation. The highest performing value(s) for each performance measure is highlighted in bold.

Appendix M: – Table S2

Accuracy, precision, recall, and correlation coefficient (CC) of classification for the protein-binding versus non-protein-binding dataset are presented for standard machine learning methods. For each machine learning approach, values of k ranged from 1 to 2. The performances of the results were estimated using cross-validation. The highest performing value(s) for each performance measure is highlighted in bold.

Appendix N: – Table S3

Accuracy, precision, recall, and correlation coefficient (CC) of classification for the multi-interface versus singlish-interface dataset are presented for internal machine learning methods. For each machine learning approach, values of k ranged from 1 to 4. The performances of the results were estimated using cross-validation. The highest performing value(s) for each performance measure is highlighted in bold.

Appendix O: – Table S4

Accuracy, precision, recall, and correlation coefficient (CC) of classification for the multi-interface versus singlish-interface dataset are presented for standard machine learning methods. For each machine learning approach, values of k ranged from 1 to 3. The performances of the results were estimated using cross-validation. The highest performing value(s) for each performance measure is highlighted in bold.

Appendix P: – Table S5

Accuracy, precision, recall, and correlation coefficient (CC) of classification for the date versus party dataset are presented for internal machine learning methods. For each machine learning approach, values of k ranged from 1 to 4. The performances of the results were estimated using cross-validation. The highest performing value(s) for each performance measure is highlighted in bold.

Appendix Q: – Table S6

Accuracy, precision, recall, and correlation coefficient (CC) of classification for the date versus party dataset are presented for standard machine learning methods. For each machine learning approach, values of k ranged from 1 to 3. The performances of the results were estimated using cross-validation. The highest performing value(s) for each performance measure is highlighted in bold.

Appendix R: – Table S7

The formula for binary classification for each of our five performance measures is provided. *TP, TN, FP, FN* are the true positives, true negatives, false positives, and false negative predictions.

CHAPTER 4.

G-QUADRUPLEX SEQUENCES ARE FOUND IN GENES REGULATED BY HYPOXIA, LOW SUGAR, AND NUTRIENT DEPRIVATION IN MAIZE (ZEA MAYS SSP. MAYS L.)

Modified from a paper to be submitted to G3: Genes|Genomes|Genetics in December 2013.

Carson M. Andorf, Mykhailo Kopylov, Drena Dobbs, Karen E. Koch, M. Elizabeth Stroupe, Carolyn J. Lawrence, and Hank W. Bass

Abstract

The 4-stranded G-quadruplex (G4Q) elements are abundant, cis-acting elements in DNA and RNA that function in maintenance and expression of genes in prokaryotes and eukaryotes. To investigate their roles in the plant kingdom, we computationally identified potential G4Qs in the maize (*Zea mays* L.) genome. We found 149,988 non-telomeric G4Qs, with 43,174 remaining after repeat sequence masking. Nearly one quarter of the filtered gene set transcripts contained one or more G4Q elements, with positional hot spots occurring in the antisense/template strands of the 5' UTR and of the 5' end of the first intron. Representative genic G4Q oligonucleotide sequences showed quadruplex formation *in vitro*. G4Q-containing genes were over-represented in genes with roles in metabolic pathways related to hypoxia, glycolysis, sugar degradation, inositol metabolism, and base-excision repair. In addition, G4Qs were prevalent in genes for signaling pathways, including the hypoxia response, AMPK/SnRK, and DJ-1/GATase1. From these results, we propose that maize G4Q elements comprise a large class of cis-acting elements ideally positioned to aid expression of genes involved in adaptive metabolism of low-oxygen or low-sugar conditions. The G4Qs are likely to have, therefore, widespread and

previously unrecognized significance in linking energy crisis perception to genomic response in plants.

Introduction

G-quadruplex (G4Q) elements are four-stranded nucleic acid structures that can form in DNA or RNA and switch back and forth between G4Q and non-G4Q conformations (1-7). G4Q intramolecular elements are made of 4-guanine planar stacks, similar to small stem loop structures, but forming a four-stranded helix with three loops. These G4Q cis-acting elements are typical of the species-specific G-rich strands of most telomeric repeats (8-10). However, the vital roles of these G4Qs in transcription, translation, replication, and recombination were not uncovered until relatively recently (7,11). The genomic distribution of G4Q elements in and around genes has prompted numerous investigations into their roles as cis-acting regulatory elements (2,7,12-15). Finding G4Qs in the promoters of mammalian genes for cell-cycle control and cancer has further fueled a dramatic expansion and acceleration of G4Q research studies (3,6,16-22). Aside from their role in cell-cycle control, G4Q elements are implicated in several other key biological processes that include hypoxia, signaling for DNA repair, and helicase-associated genome maintenance (23,24).

Genome-wide computational screens for G4Qs have established their widespread occurrence in prokaryotic and eukaryotic species including *E. coli*, budding yeast, and humans (12,25-30). Such computational analyses provide valuable information and starting points for species-specific and comparative genomic analyses of G4Q biology (14,31-36).

To date, the non-telomeric G4Qs in the human genome are among the most well characterized (12,14,36). For instance, over 40% of human gene promoters contain G4Qs, with similar prevalence and conservation in other mammalian species (24,37). Direct evidence for their role in transcription is available in a few specific cases, including the *c-MYC* proto-oncogene, where the hypersensitive site $NHEIII_1$ was shown to require G4Q formation for functional regulation both *in vitro* and *in vivo* (38). Subsequent studies implicated G4Qs in transcriptional regulation of other cancer genes, such as *KRAS* (39), c-myb (40), c-kit (41), VEGF (42), PDGFR- β (43), HIF-1α (23), bcl-2 (44), Rb (45), RET (46) and hTERT (47). Ongoing pharmacological (48), mutational (49), and evolutionary research (50) into G4Q functions are providing new opportunities for development of G4Q-targeted drugs to treat certain human diseases (6,16,51)

Among the eukaryotes, the majority of G4Q research is limited to members of a single kingdom, the opisthokonts (animals and fungi). Investigation of G4Qs elsewhere on the eukaryotic tree of life, such as the plant kingdom, will provide a more comprehensive understanding of both the conserved and diverged use of G4Qs as functional elements. Investigations in plant species offer wide eukaryotic comparisons as well as the potential for identifying regulatory roles of value to world food and biomass production. Initial research into plant G4Qs showed these elements to be prevalent in Arabidopsis transcriptomes (52) and genomes of other plant species (53). A comparative survey of G4Qs in four plant species established that plant G4Qs are often located on the template strand near the transcription start site of genes, and that especially in rice, roles described by gene ontology were enriched for the terms, chloroplast, nucleus, and histone (52,53).

Among the model plant species, maize has a rich history of genetic discoveries regarding transmission genetics, mobile DNA elements, epigenetics, and genetic diversity (reviewed in 54). The maize genome is comprised of ten chromosomes, with a current genome sequence assembly of ~2.1 Gbp, similar in size and complexity to that of humans (55). These attributes, together with the agronomic value of maize, make it a valuable species for experimental investigation of non-telomeric G4Q elements in plants. Here we describe the identification of more than 43,000 potential G4Q elements in the non-repetitive portion of the maize genome. Their non-random distribution in genes and the functional classification of those genes is considered in light of possible biological roles for plant G4Qs.

Material and Methods

The Maize Genome, Gene Models, and Syntenic Grass Gene Lists.

Maize Reference Genome: The maize inbred line B73 genome sequence is available via GenBank Accessions: CM000777 - CM000786 and GK000031 - GK000034 (56). The current annotated assembly (B73 RefGen_v2: Release 5b.60) is accessible online via the MaizeGDB, http://www.maizegdb.org (57) and MaizeSequence, http://www.maizesequence.org. This version of the reference assembly consists of 10 pseudomolecule chromosomes (Chromosome1 - Chromosome10), 2 plasmid genomes (chloroplast and mitochondria), and a set of unmapped contigs (UNMAPPED). The 10 pseudomolecule chromosomes and unmapped contigs comprise approximately 2.07 billion base pairs (including 25.8 million unsequenced gap spaces, designated as “N”s, which constitute 1.2% of the genome sequence build.

Masked Genome: The non-repetitive portion of the maize genome can be analyzed as a "masked" subset of the genomic data available from http://gramene.org/Zea_mays/Info/Index (55,58), and is masked for the Munich Information Centre for Protein Sequences Repeat Catalog (MIPS/REcat) repeats (58) (http://mips.helmholtz-muenchen.de/plant/) and the Transposable Elements (TE) Consortium repeats, including the long-terminal repeat (LTR) exemplars (59). Collectively, these repetitive regions account for approximately 83% of the maize genome. The remaining regions are referred to as the "masked genome", and they comprise 358 Mbp of primarily unique or low copy sequence.

Gene Models: The B73 RefGen_v2 maize reference genome includes two sets of predicted gene models. The working gene set (WGS) consists of 109,704 non-overlapping candidate gene elements, produced by the union of genes from two different gene finding approaches – evidence-based gene predictions using GeneBuilder (55,60), and *ab initio* gene prediction models using FGENESH (61,62). The filtered gene set (FGS) is a strict subset of the WGS that excludes transposons, pseudogenes, contaminants, and other low-confidence annotations. The FGS consists of 39,570 gene models. Gene models in both sets may have multiple transcripts per model. In this study, only the transcript models designated as 'canonical' were used because each gene locus has one canonical model, and it is the one with the greatest supporting evidence.

Maize Syntenic Orthologs: We analyzed syntenic orthologs (also called syntelogs) of maize genes from other species of the grass family, *Poaceae*. A set of pangrass syntenic orthologs was compiled (56) for *Sorghum bicolor* (sorghum), *Setaria italica* (foxtail millet), *Oryza sativa* (rice), and *Brachypodium distachyon* (purple false

brome or brachy). The transcript sequences for the gene models from these species were downloaded from Phytozome at http://www.phytozome.net; *Sorghum bicolor* release 79: 25,507 gene models; *Setaria italica* release 164: 23,216 gene models; *Oryza sativa* release 193: 66,338 gene models; *Brachypodium distachyon* release 192: 31,229 gene models (56). The number of syntenic orthologs between maize and individual grass species were: 25,507 for sorghum, 23,216 for foxtail millet, 21,528 for rice, and 20,644 for brachy. Here we define a "conserved gene" as a maize gene model that has at least one syntenic ortholog with any of these four other grasses, and a "highly-conserved gene" as a maize gene model with syntenic orthologs in all four other grasses. Using these criteria, we analyzed 27,009 conserved and 15,949 highly conserved maize genes.

Quadruplex Prediction:

In this study, we define G4Q elements in maize as contiguous single-stranded sequences that match the following default quadruplex predicting formula, $G_{3+}L_{1-7}G_{3+}L_{1-7}G_{3+}L_{1-7}G_{3+}$ or it's complementary sequence, according to the *quadparser* software (63). The resulting G4Q elements, named G4v2_1 through G4v2_149988 (Supplemental Table 1), are each assigned a chromosome, start and stop coordinates, and strand (+ or -).

Analysis of Maize G4Q Distribution Relative to Gene Structure

We took two approaches to determining whether G4Qs occur in locations that correlate with the structure of genes. The first approach divided the gene region into six arbitrary bins based on their distance from the Transcription Start Site (TSS), and calculated the frequency of G4Q occurrence within each bin. The six bins were as follows: 1,000 to 301 bp upstream, 300 to 101 bp upstream, 100 to 1 bp upstream, 1 to 100 bp downstream, 101 to 300 bp downstream, and 301 to 1000 bp downstream. To calculate the

frequency with which a G4Q element belonged to a specific bin for all gene models in the FGS, the total number of G4Qs observed in that bin was divided by the total gene space occupied by the bin. We refer to this value as the "G4Q density" or number of G4Q elements per Mbp (see Fig. 1B). The second approach calculated the frequency with which G4Q elements overlap gene structural features (i.e., the TSS, coding start site (AUG), and the first exon/intron boundary (EX-IN) (see Fig. 2A). This approach focused on a 10 kb region centered on a given gene feature (e.g., TSS) for each gene model in the FGS. We refer to the observed number of G4Q elements that overlap (any base pair) of a gene feature as the "overlap count". If the10 kb region extended into or beyond the BAC boundary of the assembly, the region was ignored because the absolute distance between coordinates in different BACs is not guaranteed due to the fact that the order and orientation between BACs is not fully resolved in the current maize assembly. To express the frequency of overlap between G4Q elements and gene structural features as a percentage, we divided the overlap count by the total count of G4Q elements within the 10 kb region and multiplied by 100.

Metabolic Pathway Analyses

For visualizing metabolic pathway relationships among genes with similar types of G4Qs (e.g., the A5U list), we used the online interactive databases MaizeCyc and MapMan. (64) (http://mapman.gabipd.org). The online MaizeCyc web site and pathway tools software are jointly hosted by MaizeGDB (http://maizecyc.maizegdb.org/) and Gramene (http://pathway.gramene.org/MAIZE/class-tree?object=Pathways) as described by Monaco et al. (64). The MapMan site is available at http://mapman.gabipd.org (65). We used the cellular omics tool of version 2.0.2 of MaizeCyc (built on RefGen_v2) for

visualizing pathway data. Supplementary figures were built using version 3.5.0 of the standalone MapMan program based on the Zm_GENOME_RELEASE_09 mapping dataset for B73 RefGen_v2. Both tools were designed for visualizing microarray data. In MaizeCyc, reaction lines representing enzymatic steps connecting substrates and products are color-coded based on the expression level of genes. The line color schemes we applied are summarized in Figure 4 and Supplemental Table 3. In these MaizeCyc output html files, we used "red" or "orange" for genes with G4Qs, "blue" for reactions associated with maize genes that lack G4Qs, or "grey" for reactions lacking any assigned maize gene model.

Maize Gene Nomenclature

In maize, the locus (gene) is represented by as a unique italicized lower case word or phrase. Known dominant alleles are represented with an upper case first letter. For example, the *shrunken1* locus is represented as lower case for both the locus itself and for known recessive alleles, whereas a known dominant allele would be referred to as *Shrunken1* (this is in keeping with standard maize gene nomenclature; see http://maizegdb.org/maize_nomenclature.php). For the *shrunken1* locus, it is known that the allele in B73 is dominant, so the B73 variant should be listed as *Shrunken1*. However, because the dominant versus recessive character is not known for the B73 allele of most genes, we refer to the B73 sequence-based genes/gene models here using the lower case and italicized locus names without implying dominance values for any of the genes described (see Table 1). New gene names that were assigned in this study were checked for appropriate convention (courtesy M. Schaeffer, MaizeGDB curator) primarily, or

secondarily using considerations of named homologs in other species or the presence of characteristic conserved domains (66).

Oligonucleotide Folding Assays

Oligonucleotides and G4Q folding conditions: All oligonucleotides were received as salt-free purified and hydrated to 300 μM (Eurofins-MWG-Operon, Huntsville, Alabama, USA). Oligonucleotide sequences for the G-rich strand of Q4Qs were either wild-type (wt) sequences or corresponding mutant (mut) control sequences (G4Q-incompatible), with variant bases underlined. The oligonucleotide sequences (5' to 3') correspond to: human telomere repeats: HsTelo4xG_wt, GGATACTTAGGGTTAGGGTTAGGGTTAGGGCGAGTC; HsTelo4xG_mut, GGATACTTAGAGTTAGCGTTAGCGTTAGCGCGAGTC; maize telomere repeats: ZmTelo4xG_wt, GGATACTTTAGGGTTTAGGGTTTAGGGTTTAGGGCGAGTC; ZmTelo4xG_mut, GGATACTTTAGCGTTTAGAGTTTAGCGTTTAGAGCGAGTC; *shrunken1* A5U: sh1_A5U_wt, GGGAGGGAGGGTTTCTCTGGGACGGGAGAGGG, sh1_A5U_mut, GGGAGTGAGGGTTTCTCTGTGACGGGAGAGTG; *hexokinase4* A5U: hex4_A5U_wt, CGGGGGTGTTGAAGGGAGGAGGAGGGAGGGG; hex4_A5U_mut CGACGGTGTTGAAGCGAGGAGGAGCGAGCGG; *hexokinase4* A5I1: hex4_A5I1_wt, TGGGGTGGGGGGGGAGCGGG; hex4_A5I1_mut, TGGAGTCGGAGGAGAGCGCG; hex4_ATG_wt, CGGGGGGATGGGGCGGGTCGGG; hex4_ATG_mut, CGAGGCGATGAGGCGAGTCGAG. Oligonucleotides were diluted to 10 μM in 10 mM tetrabutyl-ammonium phosphate buffer pH 7.5 (TBA) supplemented with or without 100 mM KCl, heated in a 1.5 mL polypropylene microfuge tube to 95°C for 15 minutes on a

heat block, followed by slow cooling via ~8 hour drift down to room temperature and used for spectra or stored at 4°C.

Thermal difference absorption spectroscopy: Oligonucleotides were subjected to folding conditions as described above, diluted to 2.5 μM, and placed in screw-cap quartz cuvettes with 10 mm optical path length for spectroscopy using a Cary 300 Bio UV-Vis spectrophotometer equipped with Varian Cary Peltier cooler. Each measurement was determined from the average of three scans (230 - 330 nm) at 30 nm/min and 1 nm data intervals. The first spectra were taken at 25°C; then samples were heated to 90°C for 20 minutes and the second spectra were collected. Thermal difference spectra (values at 90°C minus values at 25°C) were calculated and normalized by setting the value at 330 nm to zero and the value at the highest positive peak to one.

Circular dichroism (CD) spectroscopy: CD spectra (Supplemental Figure 1) were also collected on the 10 □M folded oligonucleotide samples without dilution at 25°C on an AVIV 202 CD spectrometer using a quartz cuvette with 1 mm optical path. Data were collected at a 200-330 nm range using 3 scans at 15 nm/min, 0.33 s settling time and 1 nm bandwidth. Buffer baseline was recorded using the same parameters and subtracted from the sample spectra.

Results

Genome-wide Survey of Predicted G4Qs in Maize

We analyzed the entire maize genome using the default G4Q settings for the *quadparser* software. From this we identified and named nearly 150,000 G4Q elements, averaging 27 bases in size, and summarized in Figure 1. Each G4Q is given a unique name (G4v2-quad1 – G4v2-quad149988) and is defined by the chromosome number, plus the

beginning- and ending-coordinate, followed by the strand designation (Supplemental Table 1: "+" for top strand read from short arm telomere to long arm telomere, and "–" for bottom strand). The G4Qs showed a global relative density of 74 G4Q/Mbp, but nearly twice that amount in the masked, non-repetitive portion of the maize genome (Fig. 1A). The RefGen_v2 whole maize genome consists of 2.07 billion base pairs of which 1.2% is gap sequence (a fixed number of N's that defines the type of gap between sequence contigs) and 83% is repetitive elements. A large number of G4Qs were detected in repeated regions, and many were found in a few repetitive elements, such as *Xilon*. These and all other G4Qs not located in low-copy regions of the maize genome were largely excluded from this study, which focuses on the G4Qs in the masked genome.

Over 43,000 G4Qs remained after repetitive sequences were masked, and nearly 12,000 G4Qs were associated with gene models. Among the 39,626 canonical transcripts from the high quality filtered gene set (see methods), 9,572 have one or more G4Q elements. We examined the location of the G4Qs in genes that we segmented into various structural components and boundary areas (Fig. 1B). Using TSS-aligned, gene-average plots of G4Qs, normalized per bp of each segment, we found that the highest G4Q density mapped to the first 100 bp downstream of the TSS (segment A in Fig. 1B). In this region, G4Q density peaked at about 500 G4Qs per Mbp, nearly 10 X more abundant than that of the whole genome, and 4 X more abundant than that of the masked genome. The first 100 bp just upstream of the TSS was also enriched for G4Qs, which reached densities over 350 G4Qs per Mbp of sequence. The segments were binned relative to TSS alone (Fig. 1B), leading to various numbers of introns, ORFs, and even downstream sequences in some of the plotted gene segments (e.g, segment "C" in Fig. 1B). The third highest segment for

G4Q density is the 100 bp just upstream of the TSS (segment "z", Fig. 2B), in the promoter region. The G4Q density in this region is ~4 X greater than that in the overall masked genome space, suggestive of a role for these elements in regulating some aspect of gene function.

G4Q Element Hot Spots Map to Regulatory Regions for Gene Expression

Having established an enrichment of G4Qs in maize genes relative to genomic DNA that is not accounted for simply by GC content, we next looked more closely at the location of these elements in relation gene structure. Takahashi et al. (53) showed that the TSS region of the template strand was a hot spot for G4Q elements in four plant species, Arabidopsis, grape, rice, and poplar. If that location is conserved in plants, then we would expect maize should also show a similar hot spot. We examined this possibility along with several other gene structure boundaries as summarized in Figure 2. The segments of an illustrative intron-containing gene were used to define the boundaries around which averaged trend plots are diagrammed (Fig. 2A). Boundaries that emphasize important positional information include the TSS (Fig. 2A, using the canonical transcript, and orienting all genes in the same direction), the Exon-Intron boundary ("Ex-In", Fig. 2A, limited in these graphs to the 1st intron), and the start of the open reading frame ("AUG", Fig. 2A, defined by the start codon of the canonical transcript model).

Based on analyses of gene boundary trend plots, we detect several instances of non-random peaks of G4Q locations. At low resolution (viewing a 10-kb window) we observed a clear association between G4Qs and the TSS (Fig. 2B), with the peak region falling just downstream of the TSS. Looking closer at this region and separating the signals by strand, we found that nearly all of the TSS-associated G4Qs (blue lines, Figure

2) are located on the antisense or template strand (orange lines, Figure 2). For these plots, each base pair is incremented if it overlaps with any portion of a contiguous predicted G4Q. Similarly, we found another dramatic intronic peak limited to the antisense/template strand at a precise position, within the first 50 bp of the 5' end of intron 1 (Figure 2D). These antisense, Intron-1 G4Qs occur at variable locations relative to the TSS, and contribute to some of the signal downstream of the TSS (Fig. 2B and 2C). These two types of G4Q hot spots are respectively denoted "A5U" for antisense, 5' UTR, and "A5I1" for antisense, 5' end of intron-1. Together, they account for the vast majority of genic G4Qs, and at least one of these two types of antisense G4Qs appears in over 4,000 maize genes (10% of the maize genome) having at least one these two antisense G4Q elements.

A smaller peak in positions of G4Qs occurred on the sense/coding strand near the start codon (AUG boundary plots, Fig. 2E, 2F). Considering only the sense/coding strand (Fig. 2F), positional peaks of G4Qs are just upstream or downstream of the start codon, with many ORFs (about 0.5% of those in the genome) having a start codon within a G4Q element (green line at intersection with 0 distance from the boundary, Fig. 2F). These elements were denoted "AUG" for proximity to the start codon and presence in the mRNA. We observed 481 maize genes with AUG G4Q elements by this criteria, and while smaller in number, these elements are interesting because they could function at the transcriptional and/or translational level.

Genes with G4Qs in similar positions could theoretically also share some aspects of regulation or functions. If the presence of these elements is non-functional or coincidental, then no evolutionary conservation would be expected. To test for the most parsimonious explanation, i.e., that they are conserved for functional reasons, we examined conserved

orthologs of maize G4Q-containing genes and plotted the results as shown in Figure 3. The positional frequency of antisense G4Qs for maize genes were classified according to the quality of gene model evidence within maize and according to the degree of conservation between these genes and their syntenic counterparts in four relatives of maize. These species are, in order of increasing evolutionary divergence, sorghum, foxtail millet, brachy, and rice. This analysis shows that the "lowest quality" genes (i.e., those in the within the working gene set (WGS), but rejected from the higher quality filtered gene set (FGS); Fig. 3B, orange segment) show negligible likelihood of carrying antisense/template G4Qs. These rejected genes include pseudogenes, gene fragments, or gene-like sequences for which mRNAs evidence is lacking or minimal. Within the FGS, we observe an increasing prevalence of antisense/template G4Q elements as we go from genes found only in maize (Fig. 3B, blue segment), to those found in maize and any two of the four pan-grass relatives (Fig. 3B, gold segment). The most G4Qs occurred in the most highly-conserved genes; those with known syntenic orthologs (syntelogs) in all four of the other pan-grass species (Figure 3B, blue segment). Taken together, this analysis supports the idea that the maize genic G4Qs represent a large group of broadly-conserved, functional, cis-acting elements, frequently retained in highly-conserved grass genes.

Metabolic-Pathway Analysis Associates Maize G4Qs with Genes for Responses to Hypoxia

We next examined another prediction of the computational screen for maize G4Q-contianing genes: that potential for co-regulated roles could be revealed by analysis of genes in which G4Qs reside. If so, then functional classification of genes would reveal not only this trend, but also point to a possible function for G4Qs. We explored the biological function of maize G4Q-containing genes using metabolic pathway databases, MaizeCyc

(67) and MapMan (64,65), looking specifically for non-random distribution of these genes among the many pathways depicted.

We used the MaizeCyc gene expression omics tool to color-code the presence or absence of at G4Q-containing genes as shown in Figure 4. According to the positional hotspot peaks of G4Q locations, we developed three sets of gene lists, A5U, A5I1, and AUG (defined in Figure 2, Lists 1-5). We show one of these groups, A5U, using orange and red lines indicate enzymatic steps to which one of our listed genes has been mapped (Fig. 4A). Several pathways were identified by this method and are enlarged for each of the three categories A5U, A5I1, and AUG (Fig. 4B). These include aerobic respiration, Glycolysis I, parts of the TCA cycle, and two inositol phosphate-related pathways (Fig. 4B, pathways 5 and 6). Omics overview pathway html files, like that shown for the A5U (A5U-List1, A5U-List2, Fig. 4A), were produced for these and the other two gene lists (A5I1-List3, AUG-List4, and AUG-List5) available as web-browser files for mouse-over identification of the pathways with G4Q-containing genes (Supplemental Table 3 and links therein).

Among the notable G4Q-containing genes observed using the MaizeCyc omics viewer were those reported to modulate responses to hypoxia in diverse plant species (68). Select examples from specific biological processes or pathways are listed in Table 1, along with the gene model ID, chromosome location, and class of G4Q contained therein. For instance, in maize, the sucrose synthase gene encoded by *shrunken1* contains an A5U G4Q element and this gene is also induced by hypoxia (69) and expressed in the low-oxygen region of the endosperm of developing maize seed (70).

An unexpected and recurring set of genes observed in our G4Q screen, but not previously highlighted in plant responses to hypoxia are those involving *myo*-inositol and phytic acid biosynthesis (Fig. 4B, pathways 5 and 6, and Table 1). Detection of these pathways raises questions of possible relationships between inositol metabolism and plant hypoxia. We also note that G4Q elements need not all be associated with a shared function. In fact, given the myriad processes templated by nucleic acids, it is possible that we have detected some pathways that merely coincidentally linked by shared G4Qs.

To further explore the possible link between hypoxia-responsiveness and G4Q-containing genes, we cross-referenced our gene lists with those from transcriptomic analysis of hypoxic responses (68) and those from a recent study showing that hypoxia triggers acquisition of male germ-cell fate, and found that genes associated with hypoxia were more likely to have G4Qs than other genes (1.5 X enrichment for genes in "A5U List 1"). In addition to hypoxia-associated genes (71), we found an intriguing abundance of G4Qs (especially the A5U type) in signaling genes associated with energy homeostasis. Due to the overlapping nature of low-oxygen and low-sugar signaling, as well as their response pathways, we included many of them in Table 1. For instance, many genes associated with the TOR, AMP kinase (AMPK)/Snf-related kinase (SnRK), and oxidative stress signaling (DJ-1/PARK7) pathways have one or more G4Q elements (Table 1), possibly signifying a global, coordinate role for G4Qs that corresponds to the regulatory, signaling, and metabolic responses modulating adjustments to energy stress (72-74). In addition to these, we included examples of the base-excision, repair-pathway genes (Table 1), because of their role in redox-associated transcriptional regulation at G4Q sites in specific human genes (24,75).

Maize G4Q elements can form quadruplex structures in vitro.

We demonstrated that selected G4Q sequences from this study can indeed form G4Q structures *in vitro* using G4Q folding assays as summarized in Figure 5. Single-stranded synthetic oligonucleotides were incubated under G4Q-forming conditions and thermal difference spectra showed a diagnostic G4Q-specific increase in absorption at 295 nm. This A_{295} signature (arrow, Figure 5F, *H.s.* Telomere panel) was observed in a positive control sample of human telomere repeat DNA, $(TTAGGG)_n$, and also in a plant telomere oligonucleotide sample $(TTTAGGG)_n$, under the same conditions (Fig. 5F). The locations of several genic G4Qs are diagramed (Fig. 5B-E) and their capacity to adopt G4Q structures is shown for four different G4Q elements in the *shrunken1* and *hexokinas4* genes. Using oligonucleotides with mutations altered the G-tracts ("mut" in Fig 5.), or in the absence of potassium, we observed a failure form quadruplex structures in these assays. These results were corroborated using circular dichroism spectroscopy (Supplemental Figure 1), confirming that computationally predicted G4Q elements can adopt G4Q structures *in vitro*.

Discussion

Computational prediction and analysis of G4Q elements from the genome sequence of a major crop species reveals that these elements in maize are wide-spread, preferentially located in genes, and present in those with shared functions in hypoxic responses, energy metabolism, and inositol phosphate metabolism. Using the *quadparser* algorithm to define G4Q elements, we identified nearly 150,000 G4Qs, collectively covering over 2.22 Mbp of the maize genome. The ~43,000 G4Qs found in the masked region of the maize genome is similar in numerical scale to that, ~40,000, found in rice (53). To the extent that these are

functional DNA elements, G4Qs may represent one of the largest groups of cis-acting genetic elements known, not only for plants, but for eukaryotic species in general. Other major genic cis-acting elements include general transcriptional initiation motifs (TATA boxes) and RNA processing signals (splicing elements, poly-A addition signals). This study, together with that of Takahashi et al. (53), establishes G4Q elements as a major, and largely uncharacterized entry in the encyclopedia of DNA elements in plants. The relative paucity of G4Q elements in Arabidopsis (53) is an intriguing observation, possibly reflecting a reduced capacity for Arabidopsis to resolve non-telomeric G4Qs in comparison to maize.

Systematic examination of the location of the maize G4Q elements revealed two major hot-spots in maize genes, A5U and A5I1, two locations that together account for more than 90% of all the gene-associated G4Qs. While these positions suggest roles in transcription initiation or elongation, the G4Q elements may also function as nucleation or recruitment sites for DNA replication proteins, strand-unwinding activities, DNA-repair proteins, or chromatin-remodeling functions.

Given the locations of the A5U and A5I1 elements, we speculate a role for these maize G4Q elements in transcription, with a more specific prediction that they impact the processivity of RNA polymerase II. In non-plant species, the functionality of G4Qs has been documented for a growing list of genes, even though many of them contain G4Q elements upstream of the TSS, in non-transcribed promoter areas (76). One of the most well characterized examples of G4Q in gene regulation is the nuclease hypersensitive element, NHE III, located upstream of the human *c-Myc* gene. This element was first identified as a DNAse hypersensitive site, and later shown by mutational, chemical foot-

printing, and reporter gene assays to require G4Q formation for functionality (38). Currently, *c-Myc* gene regulation is being targeted using compounds that bind G4Q elements *in vivo* as potential cancer therapeutic (51).

A primary expectation at the outset of this study was that identification of maize G4Q elements could reveal genes with shared regulation and function. Prompted by the detection of a G4Q element in the promoter of a maize sucrose synthase gene, *shrunken1*, we extended analyses to the entire genome. Metabolic pathway analysis, summarized in Figure 4, showed a striking pattern of G4Qs common to key enzymes in energy metabolism. Plant cells under hypoxic conditions must adjust to a life-threatening energy crisis in which carbohydrate metabolism and redox reactions must be modified. Hallmarks of metabolic adjustments for hypoxic-survival , summarized in recent studies and reviews, include substrate-level ATP production, increased glycolytic flux, tight regulation of mobile reductants such as malate, oxoglutarate, and oxaloacetate, and sensitive integration of sugar signaling pathways (68,72,73,77-80). Remarkably, many of these same pathways were highlighted when metabolic impacts of G4Q-containing genes were visualized using the omics viewer tools of the MaizeCyc database (Fig. 4, Table 1). Associations persisted when we examined a more specific subset of genes including transcription factors involved in signaling of hypoxic and low sugar status (Table 1).

In considering the two bodies of information discussed here - the G4Q genic location hot spots, and the functional classification of genes that contain them, we propose a generalized model in Figure 6 to integrate our findings and suggest mechanistic hypotheses. We have shown that G4Q elements represent a common feature of genes responsive to energy crises. The hypothesis that the G4Q elements in maize represent a

common link in genetic responses to multiple overlapping energy crisis signals (Fig. 6A) is supported by the prevalence of G4Q elements in genes encoding members of the AMPK/SnRK, NrF2-realated, and hypoxia-responsive transcription factors (Table 1). Target genes also included a large number of metabolic enzymes that produce ATP under low-oxygen or low-sugar conditions. A simple mechanistic model for the possible function of G4Qs in gene expression is summarized (Fig. 6B) for both transcription and translation. In this model, the G4Q acts as a physical impediment, or "kink" in the template to reduce or block RNA or protein polymerization. Resolution of the "kink" occurs through the activity of trans-acting factors in response to energy crisis input signals. Transcriptional blockage by quadruplex structures and their stabilization by G4Q-binding molecules, or their resolution by helicases or other proteins, has been described for several, animal-gene systems, including *c-Myc*, *KRAS*, and *VEGF* (39,81,82). These and related G4Q polymerase stop-assays for quadruplex function were taken into consideration in developing this model (83). There is little doubt that the mode of action may be more complex and multifactorial than a simple, G4Q-kink-based block that is resolved for an increase gene expression. For instance, mechanistic models from research on animal gene promoter G4Q elements invoke interactions of multiple trans-acting factors in and around the quadruplex (reviewed in 4). For these reasons, we include the concept of "licensing" G4Q-containing genes under stress conditions (Fig. 6B), with direct resolution being but one of several molecular mechanisms in play. It is recognized that not all genes are activated by hypoxia, and not all hypoxia-induced maize genes (e.g., *alcohol dehydrogenase1*) contain G4Q elements. It will be important, therefore, to select a subset of these genes for detailed analysis in order to more clearly determine the functional role or

roles for G4Q elements in plant species. Clearly, the abundance of G4Qs provides a way for them to aid in concerted regulation of genes responding to disparate energy stress signals.

One unexpected finding was the occurrence of A5U G4Q elements in maize homologs of hypoxia-responsive transcription factors belonging to the HRE and RAP2 group VII ethylene response factors (84). These plant-specific, oxygen-sensing transcription factors were recently found to be stabilized under low oxygen conditions via a Cys-dependent, N-end turnover pathway (71,85). Once stabilized, they are proposed to enter the nucleus and mediate global transcriptional responses. Our findings point to a possible cis-acting element with which the HRE and HRAP2 transcription factors may interact to modulate expression of specific sets of target genes. This situation is analogous to the regulation of HIF-1α, an animal hypoxia-response transcription factor, which is also is stabilized under low oxygen conditions, but by a prolyl-hydroxylase instead of the N-end degradation pathway. Interestingly, the human HIF-1α gene also has an A5U G4Q element, possibly indicative of conserved use of G4Qs in response to hypoxia.

Finally, we consider the implications of a major and unexpected enrichment of G4Q elements in genes for biosynthesis of *myo*-inositol and phytic acid metabolism and in genes for SnRK signaling pathways (Fig. 4, pathways number 5 and 6). We propose that conservation and positioning of these G4Qs are not coincidental, but rather represent functional links connecting these pathways. Two important ways in which G4Qs in these genes might aid survival are discussed. The first is direct metabolic effects, and the second is on signaling systems that ensure low-energy metabolic programs are operative.

First, recent evidence indicates an increasingly close relationship between myo-inositol metabolism and stresses involving cellular energy status(68,77). Levels of this sugar-alcohol rise during hypoxia, nutrient deprivation, and perturbations of related sensing systems (72,73,86). The same stresses also increase levels of the myo-inositol metabolites, galactinol, raffinose, and phytate (73,87,88), and likely also involve the phosphatidyl inositols (73,86,87). Suggested stress roles of raffinose, galactinol, and phytate include a possible parallel to their well-known, storage functions in seeds (86). However, other roles may be more important to survival of starvation or hypoxia. Raffinose and galactinol, for example, can also arise during remobilization of starch or cell-wall polysaccharides (86), and further, can scavenge reactive oxygen species *in vitro* and *in vivo (86)*. In addition, sugar-alcohol forming reactions can serve as sinks for excess reductant (89,90) and biosynthesis of myo-inositol can compete with starch formation in a potentially advantageous way under stress (91-93). The non-storage, direct metabolic roles of phytate also affect the centrally-important balance of Pi, PPi, and adenylates (73,87), which would be especially important during low-oxygen and starvation conditions (68,86). Examples include the switch from ATP- to PPi-driven glycolysis (PDK to PPDK) under conditions of starvation (77,85,94-96), hypoxia (95,96) and the importance of PPi in hypoxic sieve tube elements of phloem (77,85,94). Recent work has also indicated that functions of the PPi-dependent tonoplast extend well beyond contributions to vacuolar pH, and are sugar-responsive (96).

Second, new research is revealing an even greater importance of the rapid, indirect action of myo-inositol metabolism in signaling systems. Previous work demonstrated diverse roles for myo-inositol-derived signaling molecules that include inositol tri-

phosphate (Ins(1,4,5)P_3) and phosphatidyl-inositol cascades (97). In addition, raffinose synthase can facilitate sucrose metabolism and myo-inositol cycling without production of hexoses (86), an advantageous pathway under stress, because hexose-based signals of carbohydrate abundance could otherwise counter adjustments to their limited supply (73). Recently, myo-inositol metabolism has moved into a potentially central position as a hub of interaction between two major sensing systems for nutrient and energy status of cells (96,98-101). The TOR and SnRK complexes, respectively, sense nutrient abundance and deprivation (72,73), thus mediating feast and famine responses from cellular to whole organism levels (72,73,102,103). The responsiveness of both systems is vital to the adjustment of metabolism for survival of cellular energy crises (96,98-101,104). Both are also integrally involved in responses to hypoxia and nutrient deprivation (73). In this regard, sugar-modifying enzymes are not viewed as merely metabolic substrates, but also as key signaling molecules, possibly requiring stringent control.

Taken together, these functions for inositol and sugar signaling may inspire the development of more elaborate models of energy crisis adaptation in plants. The current study has identified exciting new directions for investigating plant genetic responses to cellular energy and redox predicaments. Future experiments will be needed to test the ideas proposed in our model for the mechanism of G4Q action in maize. Further interrogation of these abundant cis-acting elements holds great potential for understanding and ultimately manipulating plant growth and development, adaptation, and overall food or biomass productivity.

Acknowledgement

We are grateful to M. Schaeffer for assistance with maize gene nomenclature, and to B.P. Chadwick, H. Cui, and J.H. Dennis for helpful discussions.

Funding

This work was supported by USDA-ARS and grants from the National Science Foundation (PGRP IOS-1025954 to HWB, PGRP IOS-1116561 to KEK and coworkers) and the USDA (NRI-Plant Biochemistry 07-03580 to KEK and coworkers).

References

1. Simonsson, T. (2001) G-quadruplex DNA structures--variations on a theme. *Biol Chem*, **382**, 621-628.

2. Burge, S., Parkinson, G.N., Hazel, P., Todd, A.K. and Neidle, S. (2006) Quadruplex DNA: sequence, topology and structure. *Nucleic Acids Res*, **34**, 5402-5415.

3. Qin, Y. and Hurley, L.H. (2008) Structures, folding patterns, and functions of intramolecular DNA G-quadruplexes found in eukaryotic promoter regions. *Biochimie*, **90**, 1149-1171.

4. Brooks, T.A., Kendrick, S. and Hurley, L. (2010) Making sense of G-quadruplex and i-motif functions in oncogene promoters. *Febs J*, **277**, 3459-3469.

5. Huppert, J.L. (2010) Structure, location and interactions of G-quadruplexes. *Febs J*, **277**, 3452-3458.

6. Yang, D. and Okamoto, K. (2010) Structural insights into G-quadruplexes: towards new anticancer drugs. *Future Med Chem*, **2**, 619-646.

7. Bochman, M.L., Paeschke, K. and Zakian, V.A. (2012) DNA secondary structures: stability and function of G-quadruplex structures. *Nat Rev Genet*, **13**, 770-780.

8. Sen, D. and Gilbert, W. (1988) Formation of parallel four-stranded complexes by guanine-rich motifs in DNA and its implications for meiosis. *Nature*, **334**, 364-366.

9. Sundquist, W.I. and Klug, A. (1989) Telomeric DNA dimerizes by formation of guanine tetrads between hairpin loops. *Nature*, **342**, 825-829.

10. Williamson, J.R., Raghuraman, M.K. and Cech, T.R. (1989) Monovalent cation-induced structure of telomeric DNA: the G-quartet model. *Cell*, **59**, 871-880.

11. Maizels, N. and Gray, L.T. (2013) The G4 genome. *PLoS Genet*, **9**, e1003468.

12. Todd, A.K., Johnston, M. and Neidle, S. (2005) Highly prevalent putative quadruplex sequence motifs in human DNA. *Nucleic Acids Res*, **33**, 2901-2907.

13. Hershman, S.G., Chen, Q., Lee, J.Y., Kozak, M.L., Yue, P., Wang, L.S. and Johnson, F.B. (2008) Genomic distribution and functional analyses of potential G-quadruplex-forming sequences in Saccharomyces cerevisiae. *Nucleic Acids Res*, **36**, 144-156.

14. Huppert, J.L. and Balasubramanian, S. (2005) Prevalence of quadruplexes in the human genome. *Nucleic Acids Res*, **33**, 2908-2916.

15. Maizels, N. (2006) Dynamic roles for G4 DNA in the biology of eukaryotic cells. *Nat Struct Mol Biol*, **13**, 1055-1059.

16. Brooks, T.A. and Hurley, L.H. (2009) The role of supercoiling in transcriptional control of MYC and its importance in molecular therapeutics. *Nat Rev Cancer*, **9**, 849-861.

17. Huppert, J.L. (2007) Four-stranded DNA: cancer, gene regulation and drug development. *Philos Trans A Math Phys Eng Sci*, **365**, 2969-2984.

18. Cahoon, L.A. and Seifert, H.S. (2013) Transcription of a cis-acting, noncoding, small RNA is required for pilin antigenic variation in Neisseria gonorrhoeae. *PLoS Pathog*, **9**, e1003074.

19. Beckett, J., Burns, J., Broxson, C. and Tornaletti, S. (2012) Spontaneous DNA lesions modulate DNA structural transitions occurring at nuclease hypersensitive element III(1) of the human c-myc proto-oncogene. *Biochemistry*, **51**, 5257-5268.

20. Juranek, S.A. and Paeschke, K. (2012) Cell cycle regulation of G-quadruplex DNA structures at telomeres. *Curr Pharm Des*, **18**, 1867-1872.

21. Weng, H.Y., Huang, H.L., Zhao, P.P., Zhou, H. and Qu, L.H. (2012) Translational repression of cyclin D3 by a stable G-quadruplex in its 5' UTR: implications for cell cycle regulation. *RNA Biol*, **9**, 1099-1109.

22. Yuan, L., Tian, T., Chen, Y., Yan, S., Xing, X., Zhang, Z., Zhai, Q., Xu, L., Wang, S., Weng, X. *et al.* (2013) Existence of G-quadruplex structures in promoter region of oncogenes confirmed by G-quadruplex DNA cross-linking strategy. *Sci Rep*, **3**, 1811.

23. De Armond, R., Wood, S., Sun, D., Hurley, L.H. and Ebbinghaus, S.W. (2005) Evidence for the presence of a guanine quadruplex forming region within a polypurine tract of the hypoxia inducible factor 1alpha promoter. *Biochemistry*, **44**, 16341-16350.

24. Clark, D.W., Phang, T., Edwards, M.G., Geraci, M.W. and Gillespie, M.N. (2012) Promoter G-quadruplex sequences are targets for base oxidation and strand cleavage during hypoxia-induced transcription. *Free Radic Biol Med*, **53**, 51-59.

25. Rawal, P., Kummarasetti, V.B., Ravindran, J., Kumar, N., Halder, K., Sharma, R., Mukerji, M., Das, S.K. and Chowdhury, S. (2006) Genome-wide prediction of G4 DNA as regulatory motifs: role in Escherichia coli global regulation. *Genome Res*, **16**, 644-655.

26. Halder, K., Halder, R. and Chowdhury, S. (2009) Genome-wide analysis predicts DNA structural motifs as nucleosome exclusion signals. *Mol Biosyst*, **5**, 1703-1712.

27. Du, Z., Zhao, Y. and Li, N. (2008) Genome-wide analysis reveals regulatory role of G4 DNA in gene transcription. *Genome Res*, **18**, 233-241.

28. Todd, A.K. and Neidle, S. (2011) Mapping the sequences of potential guanine quadruplex motifs. *Nucleic Acids Res*, **39**, 4917-4927.

29. Eddy, J., Vallur, A.C., Varma, S., Liu, H., Reinhold, W.C., Pommier, Y. and Maizels, N. (2011) G4 motifs correlate with promoter-proximal transcriptional pausing in human genes. *Nucleic Acids Res*, **39**, 4975-4983.

30. Capra, J.A., Paeschke, K., Singh, M. and Zakian, V.A. (2010) G-quadruplex DNA sequences are evolutionarily conserved and associated with distinct genomic features in Saccharomyces cerevisiae. *PLoS Comput Biol*, **6**, e1000861.

31. Menendez, C., Frees, S. and Bagga, P.S. (2012) QGRS-H Predictor: a web server for predicting homologous quadruplex forming G-rich sequence motifs in nucleotide sequences. *Nucleic Acids Res*, **40**, W96-W103.

32. Wong, H.M., Stegle, O., Rodgers, S. and Huppert, J.L. (2010) A toolbox for predicting g-quadruplex formation and stability. *J Nucleic Acids*, **2010**.

33. Stegle, O., Payet, L., Mergny, J.L., MacKay, D.J. and Leon, J.H. (2009) Predicting and understanding the stability of G-quadruplexes. *Bioinformatics*, **25**, i374-382.

34. Mani, P., Yadav, V.K., Das, S.K. and Chowdhury, S. (2009) Genome-wide analyses of recombination prone regions predict role of DNA structural motif in recombination. *PLoS One*, **4**, e4399.

35. Kikin, O., D'Antonio, L. and Bagga, P.S. (2006) QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences. *Nucleic Acids Res*, **34**, W676-682.

36. Yadav, V.K., Abraham, J.K., Mani, P., Kulshrestha, R. and Chowdhury, S. (2008) QuadBase: genome-wide database of G4 DNA--occurrence and conservation in human, chimpanzee, mouse and rat promoters and 146 microbes. *Nucleic Acids Res*, **36**, D381-385.

37. Verma, A., Halder, K., Halder, R., Yadav, V.K., Rawal, P., Thakur, R.K., Mohd, F., Sharma, A. and Chowdhury, S. (2008) Genome-wide computational and expression

analyses reveal G-quadruplex DNA motifs as conserved cis-regulatory elements in human and related species. *J Med Chem*, **51**, 5641-5649.

38. Siddiqui-Jain, A., Grand, C.L., Bearss, D.J. and Hurley, L.H. (2002) Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. *Proc Natl Acad Sci U S A*, **99**, 11593-11598.

39. Cogoi, S. and Xodo, L.E. (2006) G-quadruplex formation within the promoter of the KRAS proto-oncogene and its effect on transcription. *Nucleic Acids Res*, **34**, 2536-2549.

40. Palumbo, S.L., Memmott, R.M., Uribe, D.J., Krotova-Khan, Y., Hurley, L.H. and Ebbinghaus, S.W. (2008) A novel G-quadruplex-forming GGA repeat region in the c-myb promoter is a critical regulator of promoter activity. *Nucleic Acids Res*, **36**, 1755-1769.

41. Fernando, H., Reszka, A.P., Huppert, J., Ladame, S., Rankin, S., Venkitaraman, A.R., Neidle, S. and Balasubramanian, S. (2006) A conserved quadruplex motif located in a transcription activation site of the human c-kit oncogene. *Biochemistry*, **45**, 7854-7860.

42. Guo, K., Gokhale, V., Hurley, L.H. and Sun, D. (2008) Intramolecularly folded G-quadruplex and i-motif structures in the proximal promoter of the vascular endothelial growth factor gene. *Nucleic Acids Res*, **36**, 4598-4608.

43. Qin, Y., Fortin, J.S., Tye, D., Gleason-Guzman, M., Brooks, T.A. and Hurley, L.H. (2010) Molecular cloning of the human platelet-derived growth factor receptor beta (PDGFR-beta) promoter and drug targeting of the G-quadruplex-forming region to repress PDGFR-beta expression. *Biochemistry*, **49**, 4208-4219.

44. Dexheimer, T.S., Sun, D. and Hurley, L.H. (2006) Deconvoluting the structural and drug-recognition complexity of the G-quadruplex-forming region upstream of the bcl-2 P1 promoter. *J Am Chem Soc*, **128**, 5404-5415.

45. Xu, Y. and Sugiyama, H. (2006) Formation of the G-quadruplex and i-motif structures in retinoblastoma susceptibility genes (Rb). *Nucleic Acids Res*, **34**, 949-954.

46. Guo, K., Pourpak, A., Beetz-Rogers, K., Gokhale, V., Sun, D. and Hurley, L.H. (2007) Formation of pseudosymmetrical G-quadruplex and i-motif structures in the proximal promoter region of the RET oncogene. *J Am Chem Soc*, **129**, 10220-10228.

47. Palumbo, S.L., Ebbinghaus, S.W. and Hurley, L.H. (2009) Formation of a unique end-to-end stacked pair of G-quadruplexes in the hTERT core promoter with implications for inhibition of telomerase by G-quadruplex-interactive ligands. *J Am Chem Soc*, **131**, 10878-10891.

48. Chen, Y. and Yang, D. (2012) Sequence, stability, and structure of G-quadruplexes and their interactions with drugs. *Curr Protoc Nucleic Acid Chem*, **Chapter 17**, Unit17 15.

49. Pontier, D.B., Kruisselbrink, E., Guryev, V. and Tijsterman, M. (2009) Isolation of deletion alleles by G4 DNA-induced mutagenesis. *Nat Methods*, **6**, 655-657.

50. Verma, A., Yadav, V.K., Basundra, R., Kumar, A. and Chowdhury, S. (2009) Evidence of genome-wide G4 DNA-mediated gene expression in human cancer cells. *Nucleic Acids Res*, **37**, 4194-4204.

51. Brooks, T.A. and Hurley, L.H. (2010) Targeting MYC Expression through G-Quadruplexes. *Genes Cancer*, **1**, 641-649.

52. Mullen, M.A., Olson, K.J., Dallaire, P., Major, F., Assmann, S.M. and Bevilacqua, P.C. (2010) RNA G-Quadruplexes in the model plant species Arabidopsis thaliana: prevalence and possible functional roles. *Nucleic Acids Res*, **38**, 8149-8163.

53. Takahashi, H., Nakagawa, A., Kojima, S., Takahashi, A., Cha, B.Y., Woo, J.T., Nagai, K., Machida, Y. and Machida, C. (2012) Discovery of novel rules for G-quadruplex-forming sequences in plants by using bioinformatics methods. *J Biosci Bioeng*, **114**, 570-575.

54. Bennetzen, J.L. and Hake, S.C. (2009) *Handbook of Maize: Genetics and Genomics*. Springer, New York.

55. Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T.A. *et al.* (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science*, **326**, 1112-1115.

56. Schnable, J.C., Freeling, M. and Lyons, E. (2012) Genome-wide analysis of syntenic gene deletion in the grasses. *Genome Biol Evol*, **4**, 265-277.

57. Sen, T.Z., Andorf, C.M., Schaeffer, M.L., Harper, L.C., Sparks, M.E., Duvick, J., Brendel, V.P., Cannon, E., Campbell, D.A. and Lawrence, C.J. (2009) MaizeGDB becomes 'sequence-centric'. *Database (Oxford)*, **2009**, bap020.

58. Youens-Clark, K., Buckler, E., Casstevens, T., Chen, C., Declerck, G., Derwent, P., Dharmawardhana, P., Jaiswal, P., Kersey, P., Karthikeyan, A.S. *et al.* (2011) Gramene database in 2010: updates and extensions. *Nucleic Acids Res*, **39**, D1085-1094.

59. Nussbaumer, T., Martis, M.M., Roessner, S.K., Pfeifer, M., Bader, K.C., Sharma, S., Gundlach, H. and Spannagl, M. (2013) MIPS PlantsDB: a database framework for comparative plant genome research. *Nucleic Acids Res*, **41**, D1144-1151.

60. SanMiguel, P., Gaut, B.S., Tikhonov, A., Nakajima, Y. and Bennetzen, J.L. (1998) The paleontology of intergene retrotransposons of maize. *Nat Genet*, **20**, 43-45.

61. Milanesi, L., D'Angelo, D. and Rogozin, I.B. (1999) GeneBuilder: interactive in silico prediction of gene structure. *Bioinformatics*, **15**, 612-621.

62. Salamov, A.A. and Solovyev, V.V. (2000) Ab initio gene finding in Drosophila genomic DNA. *Genome Res*, **10**, 516-522.

63. Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N. *et al.* (2012) Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res*, **40**, D1178-1186.

64. Monaco, M.K., Sen, T.Z., Dharmawardhana, P.D., Ren, L., Schaeffer, M., Naithani, S., Amarasinghe, V., Thomason, J., Harper, L., Gardiner, J. *et al.* (2013) Maize Metabolic Network Construction and Transcriptome Analysis. *Plant Gen.*, **6**.

65. Thimm, O., Blasing, O., Gibon, Y., Nagel, A., Meyer, S., Kruger, P., Selbig, J., Muller, L.A., Rhee, S.Y. and Stitt, M. (2004) MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J*, **37**, 914-939.

66. Marchler-Bauer, A., Zheng, C., Chitsaz, F., Derbyshire, M.K., Geer, L.Y., Geer, R.C., Gonzales, N.R., Gwadz, M., Hurwitz, D.I., Lanczycki, C.J. *et al.* (2013) CDD: conserved domains and protein three-dimensional structure. *Nucleic Acids Res*, **41**, D348-352.

67. Karp, P.D., Paley, S.M., Krummenacker, M., Latendresse, M., Dale, J.M., Lee, T.J., Kaipa, P., Gilham, F., Spaulding, A., Popescu, L. *et al.* (2010) Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. *Brief Bioinform*, **11**, 40-79.

68. Bailey-Serres, J., Fukao, T., Gibbs, D.J., Holdsworth, M.J., Lee, S.C., Licausi, F., Perata, P., Voesenek, L.A. and van Dongen, J.T. (2012) Making sense of low oxygen sensing. *Trends Plant Sci*, **17**, 129-138.

69. Sachs, M.M., Freeling, M. and Okimoto, R. (1980) The anaerobic proteins of maize. *Cell*, **20**, 761-767.

70. Zeng, Y., Wu, Y., Avigne, W.T. and Koch, K.E. (1998) Differential regulation of sugar-sensitive sucrose synthases by hypoxia and anoxia indicate complementary transcriptional and posttranscriptional responses. *Plant Physiol*, **116**, 1573-1583.

71. Licausi, F., Kosmacz, M., Weits, D.A., Giuntoli, B., Giorgi, F.M., Voesenek, L.A., Perata, P. and van Dongen, J.T. (2011) Oxygen sensing in plants is mediated by an N-end rule pathway for protein destabilization. *Nature*, **479**, 419-422.

72. Robaglia, C., Thomas, M. and Meyer, C. (2012) Sensing nutrient and energy status by SnRK1 and TOR kinases. *Curr Opin Plant Biol*, **15**, 301-307.

73. Dobrenel, T., Marchive, C., Azzopardi, M., Clement, G., Moreau, M., Sormani, R., Robaglia, C. and Meyer, C. (2013) Sugar metabolism and the plant target of rapamycin kinase: a sweet operaTOR? *Front Plant Sci*, **4**, 93.

74. Xu, X.M., Lin, H., Maple, J., Bjorkblom, B., Alves, G., Larsen, J.P. and Moller, S.G. (2010) The Arabidopsis DJ-1a protein confers stress protection through cytosolic SOD activation. *J Cell Sci*, **123**, 1644-1651.

75. Ruchko, M.V., Gorodnya, O.M., Pastukh, V.M., Swiger, B.M., Middleton, N.S., Wilson, G.L. and Gillespie, M.N. (2009) Hypoxia-induced oxidative base modifications in the VEGF hypoxia-response element are associated with transcriptionally active nucleosomes. *Free Radic Biol Med*, **46**, 352-359.

76. Maizels, N. and Gray, L.T. (2013) The G4 genome. *PLoS Genet*, **9**, 18.

77. Bailey-Serres, J. and Voesenek, L.A. (2008) Flooding stress: acclimations and genetic diversity. *Annu Rev Plant Biol*, **59**, 313-339.

78. Bouche, N. and Fromm, H. (2004) GABA in plants: just a metabolite? *Trends Plant Sci*, **9**, 110-115.

79. Foyer, C.H. and Noctor, G. (2005) Redox homeostasis and antioxidant signaling: a metabolic interface between stress perception and physiological responses. *Plant Cell*, **17**, 1866-1875.

80. Foyer, C.H. and Noctor, G. (2011) Ascorbate and glutathione: the heart of the redox hub. *Plant Physiol*, **155**, 2-18.

81. Brown, R.V., Danford, F.L., Gokhale, V., Hurley, L.H. and Brooks, T.A. (2011) Demonstration that drug-targeted down-regulation of MYC in non-Hodgkins lymphoma is directly mediated through the promoter G-quadruplex. *J Biol Chem*, **286**, 41018-41027.

82. Breit, J.F., Ault-Ziel, K., Al-Mehdi, A.B. and Gillespie, M.N. (2008) Nuclear protein-induced bending and flexing of the hypoxic response element of the rat vascular endothelial growth factor promoter. *Faseb J*, **22**, 19-29.

83. Sun, D. and Hurley, L.H. (2010) Biochemical techniques for the characterization of G-quadruplex structures: EMSA, DMS footprinting, and DNA polymerase stop assay. *Methods Mol Biol*, **608**, 65-79.

84. Nakano, T., Suzuki, K., Fujimura, T. and Shinshi, H. (2006) Genome-wide analysis of the ERF gene family in Arabidopsis and rice. *Plant Physiol*, **140**, 411-432.

85. Gibbs, D.J., Lee, S.C., Isa, N.M., Gramuglia, S., Fukao, T., Bassel, G.W., Correia, C.S., Corbineau, F., Theodoulou, F.L., Bailey-Serres, J. *et al.* (2011) Homeostatic response to hypoxia is regulated by the N-end rule pathway in plants. *Nature*, **479**, 415-418.

86. Valluru, R. and Van den Ende, W. (2011) Myo-inositol and beyond--emerging networks under stress. *Plant Sci*, **181**, 387-400.

87. Sulpice, R., Trenkamp, S., Steinfath, M., Usadel, B., Gibon, Y., Witucka-Wall, H., Pyl, E.T., Tschoep, H., Steinhauser, M.C., Guenther, M. *et al.* (2010) Network analysis of enzyme activities and metabolite levels and their relationship to biomass in a large panel of Arabidopsis accessions. *Plant Cell*, **22**, 2872-2893.

88. Wouters, A., Boeckx, C., Vermorken, J.B., Van den Weyngaert, D., Peeters, M. and Lardon, F. (2013) The intriguing interplay between therapies targeting the epidermal growth factor receptor, the hypoxic microenvironment and hypoxia-inducible factors. *Curr Pharm Des*, **19**, 907-917.

89. Keunen, E., Peshev, D., Vangronsveld, J., W, V.D.E. and Cuypers, A. (2013) Plant sugars are crucial players in the oxidative challenge during abiotic stress: extending the traditional concept. *Plant Cell Environ.*

90. Nishizawa, A., Yabuta, Y. and Shigeoka, S. (2008) Galactinol and raffinose constitute a novel function to protect plants from oxidative damage. *Plant Physiol*, **147**, 1251-1263.

91. Loescher, W.H. (1987) Physiology and metabolism of sugar alcohols in higher plants. *Physiologia Plantarum*, **70**, 553-557.

92. Shen, B., Hohmann, S., Jensen, R.G. and Bohnert a, H. (1999) Roles of sugar alcohols in osmotic stress adaptation. Replacement of glycerol by mannitol and sorbitol in yeast. *Plant Physiol*, **121**, 45-52.

93. Sickler, C.M., Edwards, G.E., Kiirats, O., Gao, Z. and Loescher, W. (2007) Response of mannitol-producing Arabidopsis thaliana to abiotic stress. *Functional Plant Biology*, **34**, 382-391.

94. Fukao, T., Yeung, E. and Bailey-Serres, J. (2011) The submergence tolerance regulator SUB1A mediates crosstalk between submergence and drought tolerance in rice. *Plant Cell*, **23**, 412-427.

95. Huber, S.C. and Akazawa, T. (1986) A novel sucrose synthase pathway for sucrose degradation in cultured sycamore cells. *Plant Physiol*, **81**, 1008-1013.

96. Koch, K. (2004) Sucrose metabolism: regulatory mechanisms and pivotal roles in sugar sensing and plant development. *Curr Opin Plant Biol*, **7**, 235-246.

97. Ferjani, A., Segami, S., Horiguchi, G., Muto, Y., Maeshima, M. and Tsukaya, H. (2011) Keep an eye on PPi: the vacuolar-type H+-pyrophosphatase regulates postgerminative development in Arabidopsis. *Plant Cell*, **23**, 2895-2908.

98. Koch, K.E. (1996) Carbohydrate-modulated gene expression in plants. *Annu Rev Plant Physiol Plant Mol Biol*, **47**, 509-540.

99. Adams, L.G., Khare, S., Lawhon, S.D., Rossetti, C.A., Lewin, H.A., Lipton, M.S., Turse, J.E., Wylie, D.C., Bai, Y. and Drake, K.L. (2011) Enhancing the role of

veterinary vaccines reducing zoonotic diseases of humans: linking systems biology with vaccine development. *Vaccine*, **29**, 7197-7206.

100. Ruan, Y.L., Jin, Y., Yang, Y.J., Li, G.J. and Boyer, J.S. (2010) Sugar input, metabolism, and signaling mediated by invertase: roles in development, yield potential, and response to drought and heat. *Mol Plant*, **3**, 942-955.

101. Tiessen, A. and Padilla-Chacon, D. (2012) Subcellular compartmentation of sugar signaling: links among carbon cellular status, route of sucrolysis, sink-source allocation, and metabolic partitioning. *Front Plant Sci*, **3**, 306.

102. Caldana, C., Li, Y., Leisse, A., Zhang, Y., Bartholomaeus, L., Fernie, A.R., Willmitzer, L. and Giavalisco, P. (2013) Systemic analysis of inducible target of rapamycin mutants reveal a general metabolic switch controlling growth in Arabidopsis thaliana. *Plant J*, **73**, 897-909.

103. Xiong, Y., McCormack, M., Li, L., Hall, Q., Xiang, C. and Sheen, J. (2013) Glucose-TOR signalling reprograms the transcriptome and activates meristems. *Nature*, **496**, 181-186.

104. Bihmidine, S., Lin, J., Stone, J.M., Awada, T., Specht, J.E. and Clemente, T.E. (2013) Activity of the Arabidopsis RD29A and RD29B promoter elements in soybean under water stress. *Planta*, **237**, 55-64.

Figures

Figure 1. **G4Q density within specific genomic regions-** G4Q density within specific genomic regions. **(A)** Histogram represents average number of G4Q per Mb for the B73 RefGen_v2 (Whole Genome; gray), the B73 RefGen_v2 assembly masked to remove repetitive elements (Masked Genome; black) for entire maize genome, including both gene space, and non-gene space. **(B)** Schematic of gene-associated segments analyzed. Segments X - Z represent binned regions upstream of the transcription start site (TSS) and segments A-C represents regions downstream from the TSS. For example, segment X represents sequences in the region between 1,000 and 300 bp upstream of the TSS, and segment A represents sequences in the region between 1 and 100 bp downstream of the TSS. **(C)** Histogram illustrating the G4Q density in each gene segment defined in Fig. 1B. Values for each segment represent the average density computed over all genes in the Masked Genome (aligned relative to the TSS).

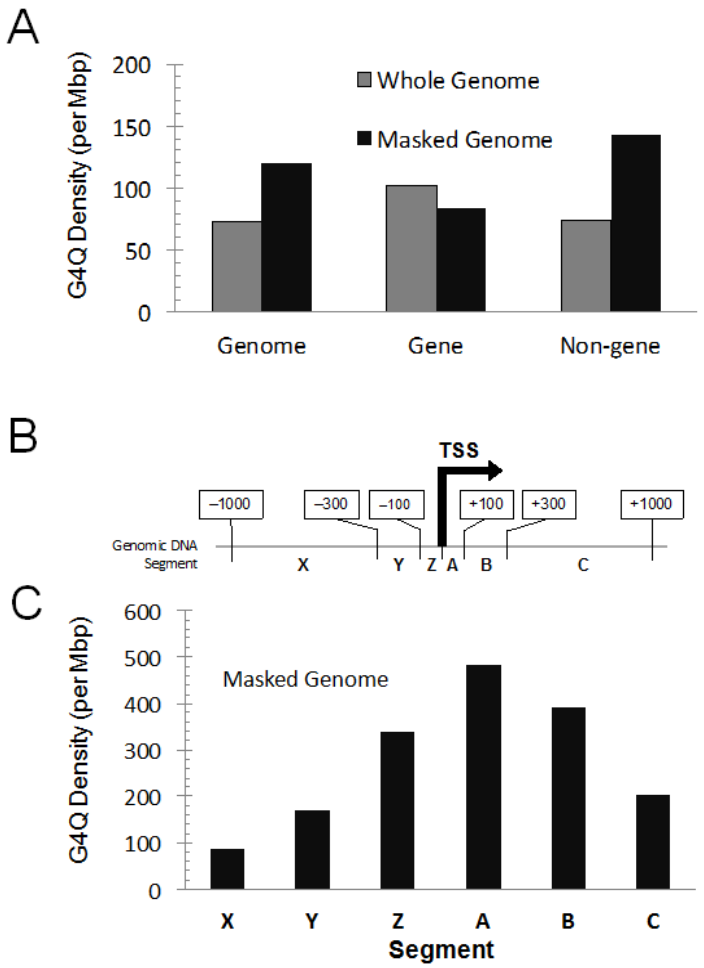


Figure 2. **Positional enrichment of G4Qs around specific gene structure boundaries- (A)** Schematic of gene structure boundary definitions for the TSS, AUG (start codon), and EX-IN (5' end of Exon1 – 3' end of Intron 1) boundary. The arrow extending from the TSS in Fig. 2A denotes the orientation of the canonical transcript for each gene model. The abundance of G4Qs at specific positions relative to each feature is expressed as the ratio of observed G4Qs over expected, where expected is defined as the number of G4Qs in the masked genome divided by the masked genome size, and is are plotted for the following features: **(B)** TSS, **(C)** TSS (zoomed), **(D)** EX-IN, **(E)** AUG, and **(F)** AUG (zoomed). For each panel, the boundary of interest is labeled at the top of the panel and is designated graphically as a vertical dashed line. The panels show the computed fold increase in G4Q elements (over the expected number) that overlap a given position (X axis) relative to the boundary. Values for elements found in either strand of the duplex DNA sequence (blue); in the sense/coding stand only (green), in the antisense/template strand only (orange). Peaks selected to make gene lists [L1-L5] are shown in brackets in panels **(C)** [L1, L2], **(D)** [L3], and **(F)** [L4, L5]. The number of genes in each list is indicated parenthetically; corresponding designations and complete gene lists are provided in Supplemental Table 1.

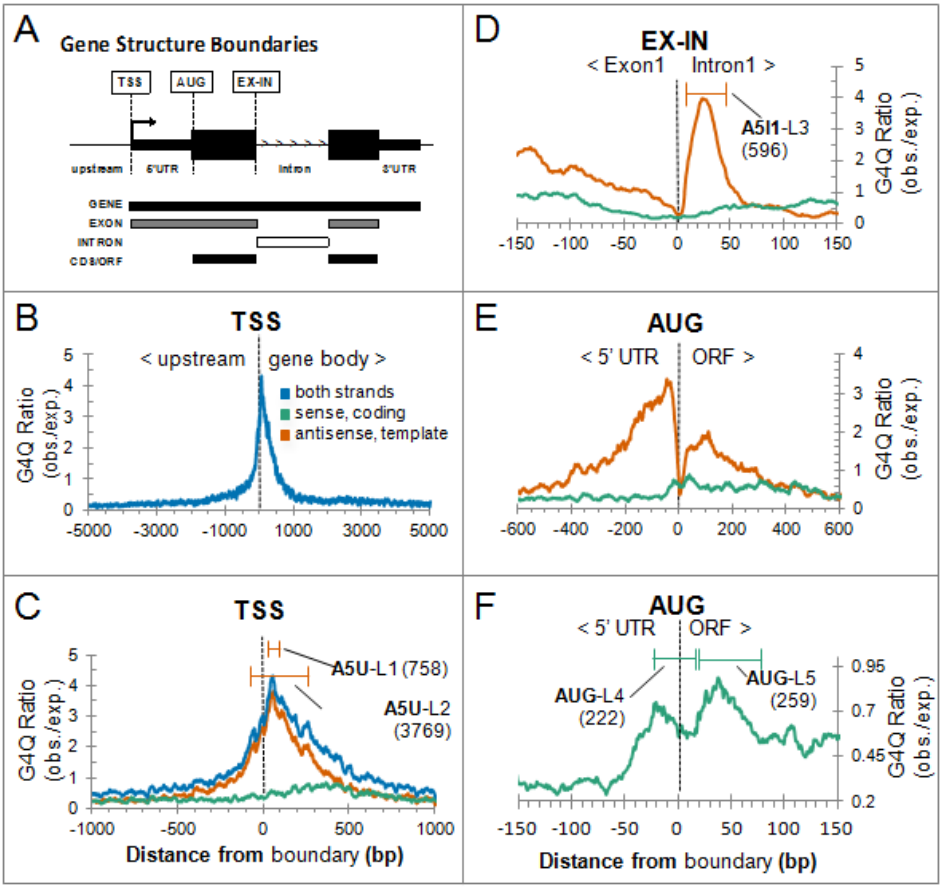


Figure 3. **Conservation of TSS-associated antisense (template strand) G4Qs in maize and maize relatives- (A)** Four horizontal, segmented bars represent the number of genes (X axis) in each of four nested subsets of maize annotated gene models. From the top, the maize WGS is made up of 109,461 genes. The second bar from the top, FGS, is defined as a high-confidence subset of the WGS. The third and fourth bars represent subsets of the FGS that have syntenic orthologs (syntelogs) in other grass species. These latter "syntenic subsets" comprise the conserved (i.e., one or more species containing syntelogs to the maize gene; third horizontal bar) and highly conserved (i.e., 5 grass species contain syntelogs to the maize gene; fourth horizontal bar) gene sets. Each of the top three bars is segmented into two colors where the gray segment contains only genes that meet criteria to be included in the subsequent subset. The orange segment of WGS represents genes that were filtered out to designate a high-quality FGS. The green segment contains FGS-specific genes that have no syntelogs. Within the syntenic subsets, the gold segment contains FGS-specific maize genes that are represented by a syntelog in one to four of the five grass species interrogated. The blue segment, which encompasses the fourth bar, represents maize genes with syntelogs in all five grass species compared. **(B)** Following the color coding convention defined in A, template/antisense G4Q ratios (Y axis) for the four annotated gene subsets are plotted against distance from TSS (X axis). Note that gene sets with the most supporting evidence (blue), including expression evidence and degree of evolutionary conservation, contain the highest signal for G4Q occurrence near the TSS.

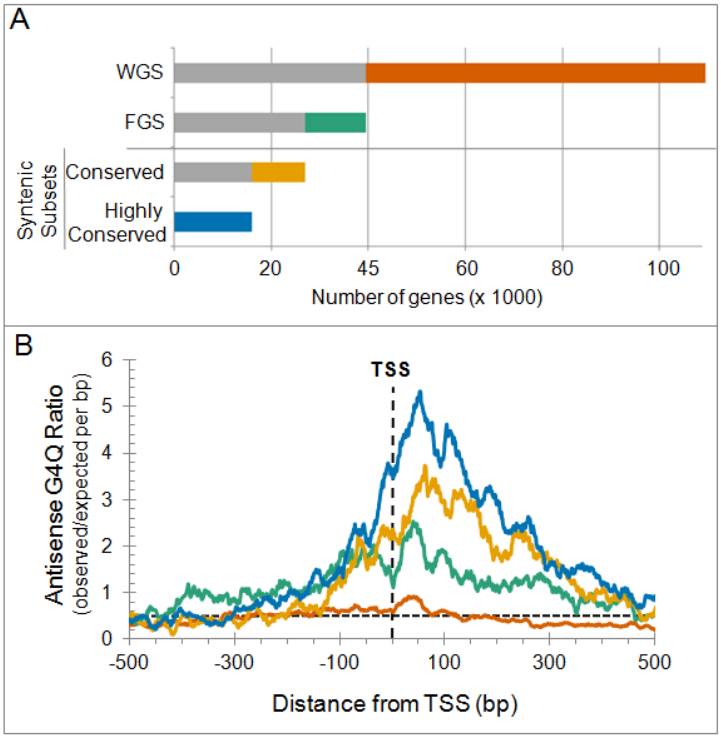
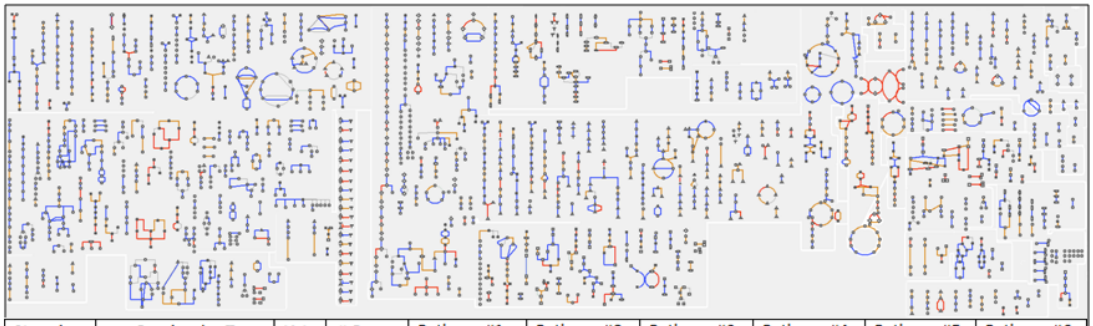


Figure 4. Metabolic Pathways highlighted with G4Q gene lists- (A) A diagram providing a schematic view of all pathways of *Zea mays* ssp. *mays* metabolism (as found in version 2.02 of MaizeCyc). Nodes represent metabolites and lines represent reactions, highlighted as follows: Red: At least one metabolite has been assigned to a gene model in G4Q List 1 (see Fig. 2C and Supplemental Table 2); Orange: At least one metabolite has been assigned to a gene model in G4Q List 2; Blue: A gene model(s) has been assigned to a metabolite, but none of the gene models are in List 1 or 2; Gray: No gene model has been assigned to the metabolite. (See Supplemental Table 3 for an interactive version of this image). (**B**) Table showing the quadruplex types associated with the five gene lists [L1 - L5], along with highlighted diagrams for six selected pathways (see panel C) and color coded as described above. (**C**) Names of the 6 selected pathways used in Panel B, as defined in MaizeCyc.

A



B

Strand	Quadruplex Type	List	# Genes	Pathway #1	Pathway #2	Pathway #3	Pathway #4	Pathway #5	Pathway #6
Template	A5U (39 to 62)	L1	745						
	A5U (-81 to 273)	L2	3769						
	A5I1 (9 to 43)	L3	596						
Coding	AUG (-20 to 20)	L4	222						
	AUG (20 to 80)	L5	259						

C

6 selected pathways

1. Aerobic respiration -- electron donor II
2. Glycolysis I
3. Cytosolic glycolysis, pyruvate dehydrogenase & TCA cycle
4. Sucrose degradation III
5. 1D-myo-inositol hexakisphosphate biosynthesis V (from Ins(1,3,4)P3)
6. Lipid-dependent phytate biosynthesis II

Figure 5. Schematics showing G4Qs on selected maize gene models- Each schematic shows 750 bases after the TSS and the last 250 bases of the canonical transcript for each given gene model. The arrow on top strand denotes the TSS, arrow on bottom strand denotes AUG, location of G4Qs are depicted as a three sheet stack on the appropriate strand, coding regions are wide black boxes, UTR regions are narrow black boxes, and introns are arrowed lines. **(A)** Maize telomere (TTTAGGG repeat) **(B)** *shruken1*, sucrose synthase with an A5U overlapping the TSS **(C)** maize *hexokinase3*, a hexokinase domain protein with three quadruplexes: A5U near the TSS, AUG overlapping the start codon, and a A5I1 in the first intron **(D)** maize *AP2-EREBP-transcription factor20,* RAP2 homologous protein with a A5U and a G4Q on the template strand immediately after an alternative transcript coding stop site **(E)** maize *AP2-EREBP-transcription factor16*, RAP2 homologous protein with three tandem A5Us (quadruplex patch) between the TSS and start codon. (**F**) Normalized UV absorbance thermal difference spectra for selected synthetic oligonucleotides in human telomeric repeat, maize telomeric repeat, and *shruken1* A5U. Human telomeric repeat was used in this experiment as a positive control. G4Q-characteristic TDS profile and prominent negative peak at A295 were obtained only for WT sequences annealed in the presence of 100 mM potassium (filled triangles) but not for those annealed in TBA phosphate buffer alone (open triangles). Mutant sequences did not show G4Q-characteristic signature independently of presence of potassium (filled and open circles respectively). (**G**) Normalized UV absorbance thermal difference spectra for three types of G4Q in maize *hexokinase4*.

- Figure 5 continued -

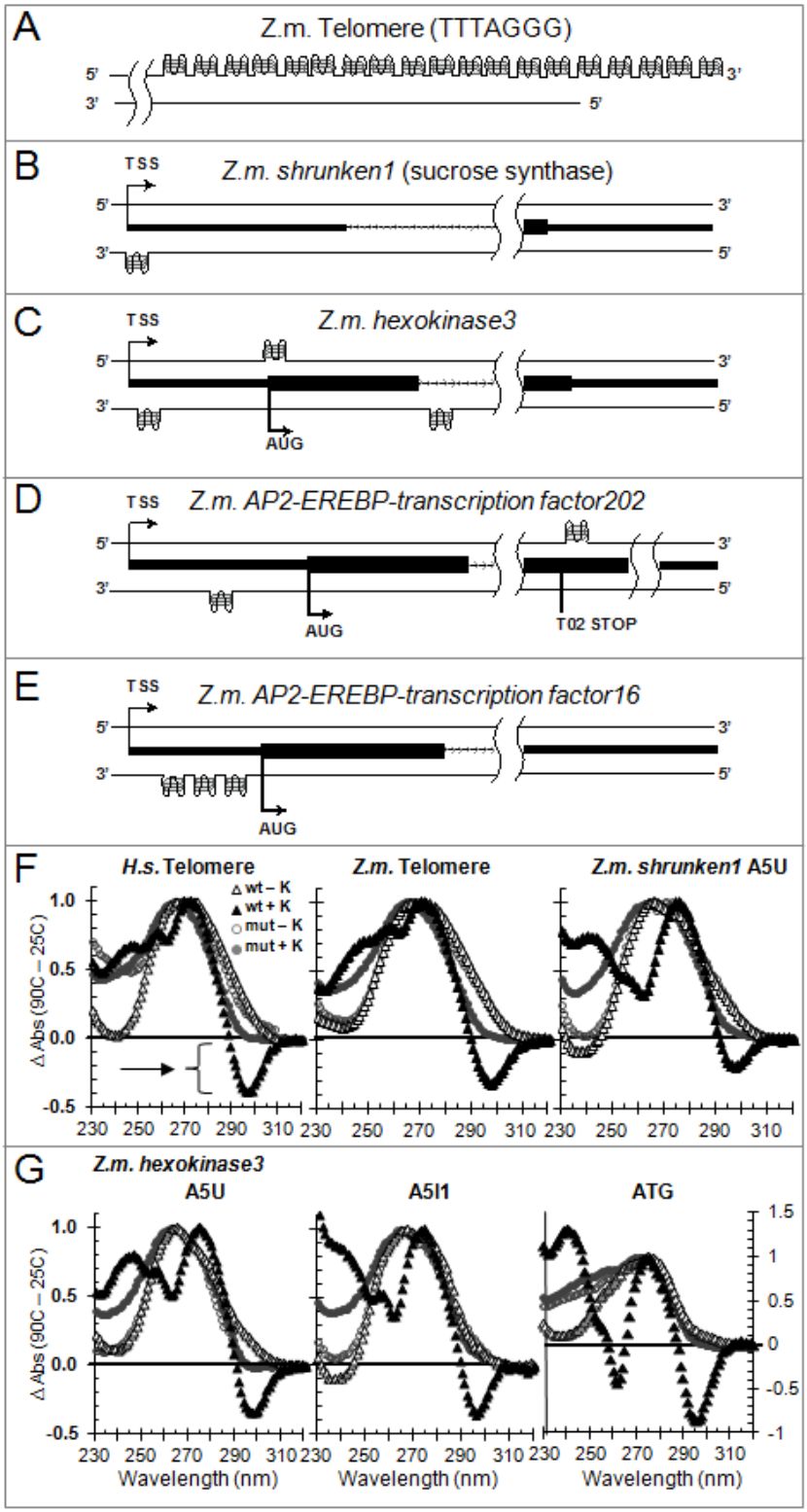
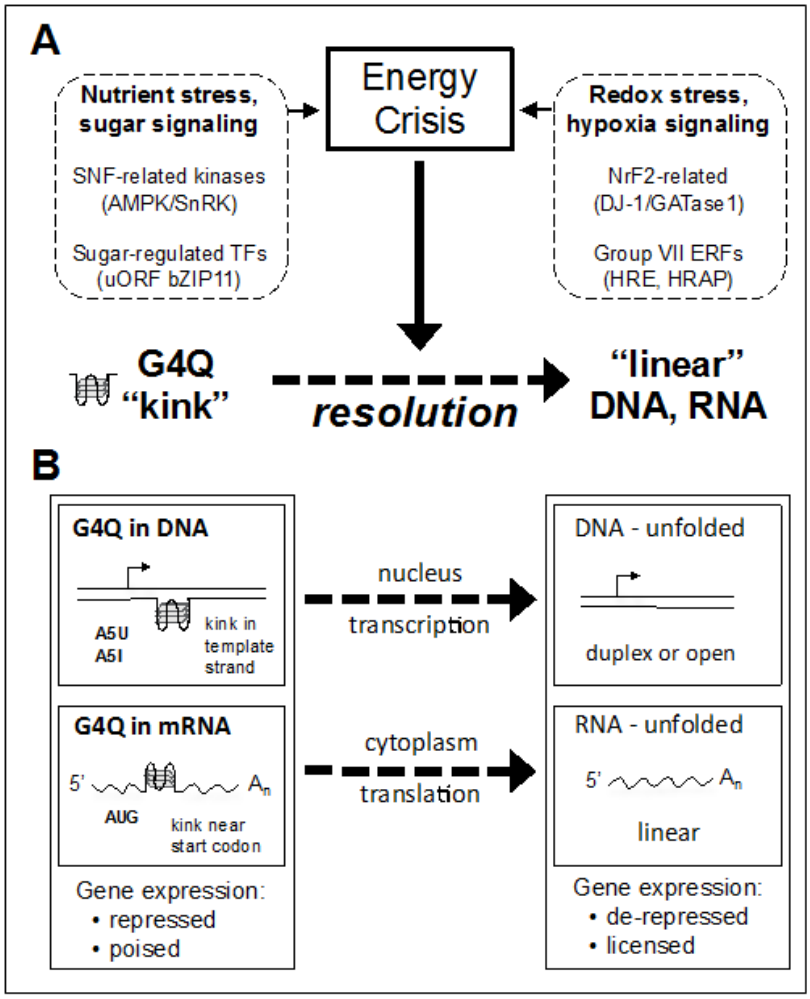


Figure 6. Model for role of G4Qs in expression of energy crisis-responsive genes. In this model, the G4Q elements classified in this study as A5U, A5I, or AUG type quadruplexes are proposed to represent negative regulatory elements in genes capable of induction by various energy stress signals, including low oxygen (hypoxia), low sugars (starvation/famine), or other signals (redox and nutrient balance). **(A)** The generalized model with several different causes of energy crisis (dashed boxes) and their signaling pathway components (parentheses) imply a broad, regulatory role for G4Qs as cis-acting elements available for coordinate genetic responses to these related metabolic and signaling pathways. "*Resolution*" denotes an unknown factor or process that functions to resolve, or "unkink" the quadruplex elements into a more favorable structure to permit expression or positive regulation of the G4Q-containing genes. **(B)** The effects of a specific G4Q element on transcriptional or translational regulation occur in separate cellular compartments, but are proposed function in both cases as a physical impediment to nuclear RNA polymerase or cytoplasmic ribosomal protein synthesis. For the regulation depicted in these panels (transcription versus translation), the "resolution" factors or processes are presumed to be different.



Tables

Table 1. Maize G4Q-containing Nuclear Genes with Key Roles in Energy Metabolism, Hypoxia, and Nutrient Signaling Pathways.

Functional Class	**Maize Locus Name**[a]	**Protein Product**[b] **(PFAM Domain)**	**G4Q**[c]	**Gene Model ID**[d]	**Genetic Bin**[e]
sugar metabolism	*sh1, shrunken1*	sucrose synthase (PF00862, PF00534)	A5U	GRMZM2G089713	9.01
	amyb1, beta amylase1	beta-amylase (PF01373)	S3U	GRMZM2G450125	1.01
	**amyb2, beta amylase2*	beta-amylase (PF01373)	2-S3U[#]	GRMZM2G025833	5.03
sugar transport	*sut1, sucrose transporter1*	sucrose transporter (cd06174[f])	A5U	GRMZM2G034302	1.04
inositol modification	*ipp2k, inositol polyphospate 2-kinase*	Inositol-pentakisphosphate 2-kinase (PF06090)	A5U	GRMZM2G067299	10.07
	**lpa3, low phytic acid3*	inositol kinase, ribokinase (PF00294)	2-A5U, AUG	GRMZM2G361593	1.1
energy, glycolysis	**hex4, hexokinase4*	hexokinase (PF00349, PF03727)	A5U, A5I1, AUG	GRMZM2G068913	3.05
	eno1, enolase1	enolase (PF03952, PF00113, PF07476)	A5U, A5I1, AUG	GRMZM2G064302	9.02
	pdlk1, pyruvate dehydrogenase (lipoamide) kinase1	mitochondrial pyruvate dehydrogenase kinase (PF10436, PF02518)	A5U	GRMZM2G107196	1.04
	**pyk3, pyruvate kinase3*	cytosolic pyruvate kinase (PF00224, PF02887)	A5U	GRMZM2G150098	2.01

- Table 1 continued -

	pdh2, *pyruvate dehydrogensase2*	pyruvate dehydrogenase (PF02779, PF02780)	A5U, AI	GRMZM2G043198	1.06
energy, NADPH production	*pdg2,* *6-phosphogluconate dehydrogenase2*	NADPH producing dehydrogenase of the oxidative PPP (PF03446, PF00393)	A5U, A5I1, SI1	GRMZM2G145715	3.05
energy, oxo-glutarate metabolism	*omt1,* *oxoglutarate malate transporter1*	plastidic 2-oxoglutarate/malate transporter (PF00939)	2-AORF	GRMZM2G383088	10.03
	gln1, *glutamine synthetase1*	glutamine synthetase; glutamate-ammonia ligase, chloroplast (PF03951, PF00120)	A5U	GRMZM2G098290	10.07
energy, respiration	**sudh1,* *succinate dehydrogenase1*	mitochondrial succinate dehydrogenase (PF00890, PF02910)	2-A5U	GRMZM2G064799	7
	**nad3,* *NADH-ubiquinone oxidorectase3*	NADH-ubiquinone oxidoreductase 10.5 kDa subunit (PF05047)	AORF-Exon2	GRMZM2G008464 (T02, P02)	8.03
	**ccr2,* *ubiquinol-cytochrome c reductase2*	ubiquinol-cytochrome c reductase complex 8.0 kDa subunit (PF05365)	A5I1	GRMZM2G064896	2.05
	**cox6b,* *cytochrome-c oxidase subunit VIb*	cytochrome oxidase c subunit VIb (PF02297)	A5U	GRMZM5G815839 (T02, P02)	9.04

- Table 1 continued -

energy signaling, TOR pathway	**raptor1,* *RAPTOR protein homolog1*	RAPTOR, TOR complex subunit (PF02985, PF00400)	2-A5U	GRMZM2G048067	10
energy signaling, AMPK/SnRK pathway	**snrk1a1, snf1-related kinase 1-like1*	Snf1-related AMPK, SnRK subunit (PF00069, PF07714, PF00627, PF02149)	A5U, SI1	GRMZM2G077278	6.06
	**snrk1a2, snf1-related kinase 1-like2*	Snf1-related AMPK, SnRK subunit (PF00069, PF07714, PF02149)	SI10	GRMZM2G180704	2.06
	**snrk1a3, snf1-related kinase 1-like3*	Snf1-related AMPK, SnRK subunit (PF04739)	A5U	GRMZM2G138814	1.03
	**snrk1bc1, snf1-related kinase -like1*	AMPK, SnRK1 subunits, with CMB48 and two CBS domains (PF00571)	2-A5U, A5I3	GRMZM2G047774	1.12
	**snrk1bc2, snf1-related kinase -like2*	AMPK, SnRK1 subunits, with CMB48 and two CBS domains (PF00571)	4-A5U	GRMZM2G014170	5
	**snrkc2, snf4-related kinase 2*	SNF4-related AMPK, SnRK subunit 2 with two CBS domains (PF00571)	A5U	GRMZM2G051764 (T03, P03)	3.09
energy signaling, sucrose-regulated TF, bZIP11 family	**bzip11a, basic leucine zipper transcription factor 11a (ZmbZIP84*)	bZIP11 family TF with sucrose-regulated group 1[g] uORF transcript (PF07716, PF00170)	AUG	GRMZM2G361611	4.1
	**bzip11b, basic leucine zipper transcription factor 11b (ZmbZIP60*)	bZIP11 family TF with sucrose-regulated group 1 uORF transcript (PF07716, PF00170)	2-A5U	GRMZM2G444748	5.03

- Table 1 continued -

	**bzip11c, basic leucine zipper transcription factor 11c (ZmbZIP12, lip15*)	bZIP11 family TF with sucrose-regulated group 1 uORF transcript (PF07716, PF00170)	A5U, A5I	GRMZM2G448607	6.01
oxidative stress response	*cat1, catalase1*	catalase (PF00199, PF06628)	A5U	GRMZM2G088212	5.03
oxidative stress signaling	**dj1a1, DJ-1A GATase1-domains1*	DJ-1a/PARK7-like GATase1-like (PF01965)	A5U	GRMZM2G102927	3.01
	**dj1a2, DJ-1A GATase1-domains2*	DJ-1a/PARK7-like GATase1-like (PF01965)	A5U	GRMZM2G024959	4.05
	**dj1a3, DJ-1A GATase1-domains3*	DJ-1a/PARK7-like GATase1-like (PF01965)	A5U	GRMZM2G127812	4.05
	**dj1c, DJ-1C GATase1-domains*	DJ-1c/PARK7-like GATase1-like (PF01965)	A5U	GRMZM2G117189 (T02, P02)	1.07
hypoxia transcription factor, group VII^{h} ERFs	**hre1, hypoxia responsive ERF homologous1 (ZmEREB67)*	hypoxia responsive ERF with AP2, CRIB domain, N-terminal MVLSAEI (PF00786, PF00847)	A5U	GRMZM2G114820	9.04
	**hre2, hypoxia responsive ERF homologous2 (ZmEREB102)*	hypoxia responsive ERF with AP2 domain, N-terminal MCGGAIL (PF00847)	A5U	GRMZM2G052667	7.02
	**hre3, hypoxia responsive ERF homologous3 (ZmEREB202)*	hypoxia responsive ERF with AP2 domain, N-terminal MCGGAIL (PF00847)	A5U	GRMZM2G148333	2.06

- Table 1 continued -

	**hrap1,* *hypoxia responsive RAP2 homologous1 (ereb14)*	hypoxia responsive ERF with AP2 domain and N-terminal MCGGAIL (PF00847)	A5U/TSS	GRMZM2G018398	4.07
	**hrap2,* *hypoxia responsive RAP2 homologous2 (ZmEREB160)*	hypoxia responsive ERF2 with AP2 domain protein and N-terminal MCGGAIL (PF00847)	3-A5U	GRMZM2G171179	9.01
transcription factor	*arf25,* *ARF-transcription factor 25 (ZmARF25)*	auxin Response Factor (PF02362, PF06507, PF02309)	2-A5U	GRMZM2G116557 (T02, P02)	8.06
	ca2p7, *CCAAT-HAP2-transcription factor 27 (ZmCA2P7)*	CCAAT box binding factor, (PF02045, PF01406)	3-A5U	GRMZM2G126957 (T02, P02)	10.06
Chromatin	*htr105,* *histone H3 105*	histone H3.2 (PF00125)	A5U	GRMZM2G130079	9.02
	htr103, *histone H3 103*	histone H3.3 (PF00125)	A5U	GRMZM2G078314	3.06
Chromatin	*chr112a,* *chromatin remodeling 112a*	SNF2 superfamily, with HIRAN, DEAD box helicase, RING and RAD5 domains (PF08797, PF00176, PF00097, PF00271)	2-A5U	GRMZM2G030768	4.05
DNA repair, base excision repair (BER) pathway[j]	**ogg1, 8-oxoguanine DNA glycolyase*	N-glycosylase/DNA lyase, OGG1-like (PF07934, PF00730)	A5U	GRMZM2G139031	5.04

- Table 1 continued -

	**endo3, endonuclease III*	endonuclease III (PF00730, PF00633)	SORF-exon7	GRMZM2G113228	10.02
	**dmag2a, DNA-2-methyladenine glycosylase II A*	DNA-3-methyladenine glycosylase II (PF00730)	A5U/TSS	GRMZM2G117574	6.07
	**dmag2b, DNA-2-methyladenine glycosylase II B*	DNA-3-methyladenine glycosylase II (PF00730)	A5ORF	GRMZM2G114592	6.07
	**dmag1, methyladenine glycolsylase I*	DNA-3-methyladenine glycosylase I (PF03352)	A5U	GRMZM2G171317	6.02
	**udg, uracil-DNA glycosylase*	uracil-DNA glycosylase (PF03167)	A5U, A5ORF	GRMZM2G040627	2.01
	hmg1, high mobility group protein1	high mobility group protein B1 (PF00505)	A5U	GRMZM5G834758 (T03, P03)	5.03
	pcna1, proliferating cell nuclear antigen1	proliferating cell nuclear antigen (PF00705, PF02747)	A5U, AP-140	GRMZM2G030523	5.08
	dpole2, DNA polymerase epsilon subunit 2	DNA polymerase epsilon subunit 2 (PF12213, PF04042)	A5U/TSS	GRMZM2G154267	1.06
	**dpold3a, DNA polymerase delta subunit 3 locus a*	DNA polymerase delta subunit 3 (PF00281, PF00673, PF09507)	AORF-exon6	GRMZM2G435338	3.02
	**dpold3b, DNA polymerase delta subunit 3 locus b*	DNA polymerase delta subunit 3 (PF09507)	A5U	GRMZM2G005536	8.03
	**dpold3c, DNA polymerase delta subunit 3 locus c*	DNA polymerase delta subunit 3 (PF09507)	A5U	GRMZM2G148626 (T02, P02)	7.01

Footnotes:

a Maize Locus Name; current gene name from MaizeGDB (MaizeGDB.org) annotation, listed as the short name followed by the full name. Genes named anew in this study are indicated with an asterisk preceding the short name. Synonyms are given for some loci/gene names in parentheses following the full name. Some genes are designated

using suggested nomenclature from specialized databases, including those curated by "ChromDB" (the chromatin database, chromdb.org) for chromatin remodeling gene (prefix "chr") and histone H3 (prefix "htr") and those for plant transcription factors, curated by "Grassius" (grassius.org, Gray et al., 2009) for ethylene response element binding protein (prefix "ZmEREB"), basic region leucine zipper (prefix "ZmbZIP"), auxin response factor (prefix "ZmARF" or "arf") and CCAAT-HAP2 family (prefix "ZmCA2P" or "ca2p"). The bzip11c locus is also known to encode the gene known as *lip15*, *low temperature induced transcription factor15*.

b The column named "Protein Product" refers to the name of the encoded gene product or enzyme known or predicted, along with the PFAM domain identifier in parenthesis, as reported from MaizeGDB. Predicted proteins are all from the only or 1st transcript model (T01, P01), unless otherwise designated. Abbreviations used in this column in order of appearance: PF/PFAM , protein family database (http://pfam.sanger.ac.uk/, Punta et al., 2011,); NADPH, the reduced form of nicotinamide adenine dinucleotide phosphate; PPP, pentose phosphate pathway; NADH, the reduced form of nicotinamide adenine dinucleotide; RAPTOR, regulatory-associated protein of mTOR (mammalian target of rapamycin); TOR, target of rapamycin; AMPK, 5' adenosine monophosphate-activated protein kinase; SnRK, SNF-(sucrose non-fermenting)-related serine/threonine-protein kinase; CBM48, carbohydrate binding module 48, a domain that is a member of the N-terminal Early set domain, a glycogen binding domain associated with the catalytic domain of AMP-activated protein kinase beta subunit; CBS, cystathionine beta synthase domain (also known as Bateman domain) that regulates the activity of associated enzymatic and transporter domains in response to binding molecules with adenosyl groups, AMP and ATP, or s-adenosylmethionine; bZIP11, basic leucine zipper transcription factor of the bZIP11 family of the group S bZIP TFs (Jakoby et al., 2002); DJ-1/PARK7, oncogene DJ1/Parkinson Disease 7; GATase1; type 1 glutamine amidotransferase; ERF; ethylene response factor; AP2, APETALA2; CRIB, Cdc42/Rac interactive binding; SNF2, sucrose non-fermenting 2; HIRAN, HIP116, Rad5p N-terminal; RING, really interesting new gene; RAD5, radiation sensitive 5; OGG1, 8-oxoguanine DNA glycolyase 1.

c G4Q elements are listed as defined in this study (Fig. 2). Multiple elements in the same class are by leading number and dash. For example 2-A5U indicates two G4Q elements within the Antisense 5' UTR. Classifier abbreviations: A5U, antisense strand 5' UTR; A5U/TSS, A5U/overlapping with the transcription start site; A5I1, antisense 5' end of Intron 1; AUG , sense strand near the "AUG" start codon; S3U, sense strand 3' UTR; AI, antisense strand non-first Intron followed by the intron number; AORF, antisense ORF followed by the exon number for intron-containing genes; SI, sense strand intron followed by intron number; AP, antisense promoter followed by a dash and the number of bases upstream of the TSS.

d Gene Model ID: From maizesequence.org, from the filtered gene set of B73 line of maize. For genes with multiple transcript models, Predicted proteins are all from the only or 1st transcript model (T01, P01), unless otherwise designated.

e Genetic BIN refers to the chromosome number followed by the genetic linkage bin as defined by Davis et al., (1999) for the map UMC 98. Current markers delineating the genetic bins of maize are at MaizeGDB ().

f The protein encoded by the sut1 locus lacks a PFAM domain but does contain the conserved domain cd06174; MFS, The Major Facilitator Superfamily (from Conserved Domain Database, CDD, at NCBI, http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml, Marchler-Bauer et al., 2013)

g Group 1 refers to the uORF-containing "homology group 1" bZIP genes (Hayden and Jorgensen, 2007).

h The group VII ERFs (from Nakano et al., 2006) include hypoxia responsive ERF TFs (loci/gene name "hre") and hypoxia responsive related to AP2 (RAP2.2/RAP2.12) TFs (loci/gene name "hrap") The HRE and HRAP genes listed here fall in the AP2-EREBP (APETALA2 – ethylene response element binding protein) group as described by Gray et al., (2009) and curated by the grass TF database, grassius.org. The recommended grassius gene names are listed here as synonyms in parentheses.

i The locus/gene *bzip11c* is also known as *lip15*, *low temperature induced protein15*.

j The base excision repair pathway from the KEGG pathway zma03410 (KEGG, http://www.genome.jp/dbget-bin/www_bget?zma03410) are included here because of its role G4Q-mediated hypoxia-induced transcription (Clark et al., 2012).

Supporting Information

Supporting Figures and Tables can be found in Appendices S -V.

Appendix S: – Supplementary Figure 1

CD spectra of the same oligonucleotide samples as in shown in Figure 6F. Here, the diagnostic G-quadruplex spectra are shown using CD. Oligonucleotides were annealed in either 10mM TBA-phosphate buffer (gray) or in 10mM TBA-phosphate buffer supplemented with 100mM KCl (black). G4Q-specific spectra were obtained for every oligonucleotide tested only in presence of potassium. (**A**) Human telomere repeat. Negative peak around 240 and two positive peaks at 265 and 287 are characteristic for human telomere repeat and suggest a formation of mixed type parallel/antiparallel quadruplex (**B**) Maize telomere repeat. Strong positive peak at 287 taken together with the absence of negative peak at 260 and presence of a valley between 235 and 245 also suggests mixed parallel/anitparallel quadruplex. (**C,D,E,F**) Oligonucleotides with genomic sequences showing characteristic CD-signatures of parallel quadruplex structures with negative peak at 240 and strong positive peak at 260 nm.

Appendix T: – Supplemental Table 1.

List of all maize G4Q elements.

Appendix U: – Supplemental Table 2.

Lists of gene-associated maize G4Q elements.

Appendix V: – Supplemental Table 3.

MaizeCyc Links. Links to HTML files of MaizeCyc omics viewer output for G4Q genes from the A5U, A5I1, and AUG gene lists.

CHAPTER 5.

CONCLUSIONS

Summary and discussion

Recent completion of many genome projects has generated large amounts of protein sequence data. An urgent and challenging task is to identify, label, and classify these proteins into related groups. These groups include functional classes, structural classes, interaction partners, binding patterns, regulation, and expression. However, identifying these classifications using traditional biological experiments lags far behind the increasing speed of sequencing technologies. Computational methods that can identify and classify related proteins in large scale are in urgent need. In this dissertation, three studies are presented. First, we developed a two-stage classifier, *HDTree*, to predict potential misannotations within the AmiGO web server database. Second, we developed a three-phase method to predict the binding patterns of proteins. Finally, we carried out a genome-wide analysis of the maize genome for G4Q motif occurrence in regulatory elements of genes involved in specific types of metabolism that suggest a regulatory role in coordinating global genomic response to hypoxia and related energy crisis states.

The experimental results presented in this dissertation have shown that:

- Our machine-learning method *HDTree* was able to identify 201 mouse protein kinases returned from AmiGO that were functionally misannotated. Our classifier was able to predict annotations that were consistent with the human counterparts with 97% accuracy.
- Our three-phase machine-learning based approach was able to predict whether a protein is a protein-binding protein with an accuracy of 94% (0.87 correlation

coefficient); distinguish hubs from non-hubs with 100% accuracy for 30% of the data; distinguish between date hubs/party hubs with 69% accuracy and AUC of 0.68; and distinguish SIH/MIH with 89% accuracy and AUC of 0.84.

- A genome-wide scan of the maize genome for the G4Q motif identified 149,988 G4Qs in maize, with 43,174 occurring in the non-repetitive genomic fraction. These elements showed a pronounced tendency to occupy non-random locations within genes. On the template strand of the genomic DNA, G4Qs were markedly enriched in 5' UTR region just downstream of the transcription start site, and at the 5' end of the first intron. These two G4Q hot spots were limited to the antisense/template strand, suggesting a broad and previously unrecognized role in transcription for up to thousands of maize genes. We highlighted several metabolic pathways that show an overabundance of genes with G4Qs in hot spot regions and postulate a potential role G4Qs have in hypoxia response.

The first two studies (Chapter 2 and Chapter 3) show examples (H*DTree* and *HybSVM*) of a general framework for learning classifiers by combining sequence-based and homology-based representations of proteins using a two-layer classifier. Although we focused our work on learning classifiers to predict misannotations of Gene Ontology labels and binding patterns of hub proteins, this framework can be easily extended to a variety of other biologically significant problems.

A third study (Chapter 4) focuses on a genome-wide analysis of the maize genome of how non-Watson-Crick structural conformations known as G4-quadruplexes have potential in gene regulation. This type of analysis can be extended to any sequenced genome.

Beyond the work in this dissertation, other published or presented results can be found in the publication section in **Appendix W** (Biographical Sketch).

Contributions to the field

The main contributions of this dissertation include the following:

Fully updatable sequence-based methods that need one pass through the data.

Many modern techniques cannot make this claim. The methods NB(k) and NB k-gram are based on a Naïve Bayes and k-order Markov models, respectively. To build these models we need to estimate the probability of each element in the model and store it in a probability table. The probability table is constructed by taking the counts of each individual k-gram seen in training data and dividing by the overall counts of all k-grams seen in the training data. These counts can easily be obtained by passing through each sequence and updating each count as they occur. Actually, multiple k-values can actually be counted in parallel. In our examples with protein sequences, we compute the counts of 1-grams, 2-grams, 3-grams, and 4-grams at the same time as we pass through the data. So with a single pass of the data we are able to build four separate probability tables. In the case of NB(k), we also need the estimates of the (k-1)-grams. These can be easily obtained my marginalizing over k-grams. This can be done in linear time in order of the number of possible (k-1)-grams which is less (and usually much less) than the total number of k-grams that we counted in the training data. Therefore, the models for the NB(k) and NB k-gram algorithms can be constructed in linear time in order of the size of the training set. As new data become available, the classifier can be updated by taking the k-gram counts from the new data and adding to the counts of the previous data. The new probabilities will be based on the updated counts. This also can be done in linear time.

Many modern techniques are not linear algorithms and are not fully updatable. An example of such a technique involves SVMs. SVMs are one of the most popular techniques for predicting protein function. SVMs can also be used on sequence data and on k-gram representation of sequences. Where our NB methods are based on generative models (since they are built on the underlying distributions of the data), SVMs are discriminative classifiers that focus on simultaneously minimizing the empirical classification error and maximizing the geometric margin between the classes. Since SVMs use regression and not strict counts of the data, multiple passes through the data are needed to minimize the classification error and maximize the margin. Therefore, SVMs are much slower than our NB methods for large data and/or k-values and some implementations for SVM are impractical for relatively large values of k. This problem is amplified when using data that are consistently being updated. Many biological databases fall into this category. An example is the GenBank database [1]. The GenBank release 194, (February 2013), had over 150 billion nucleotide bases from approxiametly 162 million sequences. Another example is the Protein Data Bank (PDB) [2, 3]. PDB has been growing at an exponential rate (http://www.rcsb.org/pdb/statistics/contentGrowthChart.do?content=total&seqid=100). In May 2013, there are over 83,000 protein and 2,500 nucleic acid structures in PDB. To maintain an updated classifier for PDB you would have to retrain you classifier as new data becomes present. In our models this is just updating dozens of sequences worth of k-gram counts a day. These updates would take seconds to do on a standard machine. In the case of an SVM classifier this would be updating the counts plus retraining the model with tens of thousands of protein sequences. A lot of time and energy (possibly days of

computational time and most likely expensive hardware would be needed) would be focused on building and maintaining the classifier rather than focused on using the classifier to get meaningful results. It is very feasible that more data would become present before the new classifier could be built. Our method partially addresses this problem by extending a new possibility on how to learn from a growing data source.

A high-performance classifier that predicts protein function based on sequence alone.

Many current techniques use a "kitchen sink" approach to classifying proteins into functional classes. They use as much information as possible in regards to a protein. This information could include secondary or tertiary structure, interactions with other proteins, DNA, RNA, or other small molecules, gene expression data, microarray data, sequence similarity among other proteins, textual information in the proteins annotation, literature mining on this protein (e.g., PubMed abstracts), along with the proteins sequence and composition. This list is far from comprehensive, but it shows the vast amount of information available for different proteins. Among all this information the easiest and most common information for a protein is its sequence. Once a protein is isolated it only takes days to hours to sequence the protein. Also, it is very easy to infer a protein sequence from the DNA of a given gene. Usually additional information is not present for a protein of interest and to obtain this information could take weeks, months, or even years in extreme cases. Other times it makes little sense to use a machine- learning approach when additional data exists (e.g., the function is already in the PubMed abstract, or a homolog with high sequence similarity exists in a common public database). Potentially the most interesting problem arises when no related sequences are available or low sequence identity exists within a class. We have shown our method performs quite well

with sequences that share the same function, but have low sequence identity (below 30 percent). In fact, we have shown our method still performs well when the sequence identity is below 10 percent and the sequence similarity is so low that a self-comparison using PSI-BLAST would return no significant hits. For these reasons, our method has improved the way protein function can be predicted from sequence.

Sequence-based method can determine "functional motifs."

Another major advantage of our approach is that our prediction results are easy to interrupt. This is unlike other "black box" approaches that only give you a prediction and little information behind the predication. Our method is based on using k-gram compositions of amino acids. To build a Naïve Bayes classifier, a probability table estimating the probabilities of the given model is needed. In our case, we estimate the probability of a given k-gram occurring given its class. Once a classifier is built, this probability table can be used to give insight into the protein's predicted class. In the case of predicting a protein kinase gene ontology label, we used the log-likihood ratio of the k-gram probabilities for the given classes and a simple independence test to provide valuable insight into the protein's sequence-structure-function relationship. We were able to determine four tetramers with potential functional/structural significance within the kinase class. With further analysis, we found one of these locations was the active site of the protein, two locations were involved in interactions with other proteins, and the forth site was not found in literature, but was in contact with the three previous sites in the tertiary structure. We hypothesized that this tetramer region may be important in maintaining the structure of the protein. This method can contribute to the field by being extended beyond

kinases to predict other potentially significant regions important in the sequence-structure-function relationship of other type of proteins.

A machine-learning approaches based on protein sequence to predict potential misannotations.

Our method can build classifiers that predict GO labels based on a protein's sequence. These classifiers are trained on proteins with pre-existing GO labels. Unfortunately, not all the annotations are correct. These proteins are annotated in a wide variety of ways, each with a level of reliability assigned to it. For example, a protein can be assigned a GO label based on having a high sequence similarity with another protein that already has an assigned GO label. This method is called ISS, inferred from sequence similarity. The problem with this method is that there is no guarantee that the highly similar sequence shares the same function with the sequence in question and worse it is possible that this original sequence does not have the correct annotation. Many scenarios exist where current methods (many being computational) could potentially incorrectly annotate a sequence. Yet, these public data are being used pervasively.

In our work, we addressed whether a machine-learning algorithm could be used to help predict and identify potentially misannotated proteins with a frequently accessed database. Our HDTree method was able to identify and offer potential correct annotations to hundreds of protein kinases from the AmiGO web server. These corrections help improve the field of machine-learning to build better classifiers, and allow the biological fields to do better science. This is the first machine-learning approach in recent literature focused on detecting misannotations in a biological context.

A machine-learning approach to predict binding patterns of proteins.

Proteins can bind with molecules including DNA, RNA, other proteins, ligands, or other small molecules. These interactions are usually essential for most biological processes. In many cases, proteins need to interact with other proteins to be functional. Most proteins interact with a few proteins, but some proteins interact with a large number of other proteins. In protein-protein interaction networks, proteins that interact with many other proteins are called hub proteins. Hubs proteins can be further distinguished as having mutually exclusive interactions (date) or multiple simultaneous interactions (party). Hubs can also have many interactions sites (multi-interface) or a small number of interaction sites (singlish-interface). Distinguishing these types of binding patterns in proteins can give further insight into other characteristics including: essentiality, disorderness, evolutionary rate, co-expression, 3D structure, and structural interactions. We developed a three phase method to classify proteins at multiple levels of binding patterns: binding/non-binding, hub/non-hub, date/party, singlish/multiple-interface. Other methods have been developed to predict hub proteins and date/party hubs, but our work is the first to explore using machine-learning approaches to predicting binding-proteins and to combine different levels of prediction in a multi-phase process.

An accessible binding pattern prediction server and a graphical user interface driven standalone version of HDSVM method.

We made our binding pattern prediction method accessible by building an online server. The server can be used to predict the following binding patterns: (1) classifies whether a protein is likely to bind with another protein; (2) predicts the number of interaction partners and if a protein-binding (PB) protein is a hub; and (3) classifies PB proteins as singlish-interface versus multiple-interface hubs and date versus party hubs. In

addition, this server includes the original data sets, instructions, and a web-based form to use our three-phase approach on novel data provided by the user.

Also, we have developed an easy to use stand alone graphical user interface driven version of *NB(k)*, *NB k-gram*, and *HybSVM*. The software package was implemented in Java and freely available by request. We want other groups to be able to easily reproduce the results presented in this dissertation and test and evaluate the software on other biological problems.

The online server of our program can be located at:

http://hybsvm.gdcb.iastate.edu/HybSVM/

Genome-wide study of the G4-quadruplex on the maize genome

Research on the structure, location, and functional role of the non-Watson-Crick, four-stranded G4-quadruplex structure has received a lot of recent attention. Much of this attention focussed on the overabundance of the G4Q in the promoter of human genes and in particular, in the promoters of oncogenes. Genome-wide surveys have been performed on human and other eukaryotic genome, but little attention has been paid to plant species. Currently, only two publications have explored the presence of G4Q in plants and both were comparative analyses of the G4Q distributions in Arabidopsis and a few other plant genomes. We carried out an in-depth analysis of the presence of G4Qs in maize and related grass genomes finding enrichment of G4Qs in genes in metabolic pathways associated with electron transport, sugar degradation, and response to hypoxia. We conclude that G4Qs may have a widespread and previously unrecognized capacity for coordinating global genomic responses to hypoxia and related stresses.

Future work on protein classification

Identification and classification of proteins into meaningful categories is a driver for biological research. Development of computational methods to predict such categories in a quick and reliable fashion is still in its initial stage. This problem provides challenging targets for machine-learning research to learn models for classification. The performance of the sequence-based classifiers from this study and methods from other studies can still be improved and better understood. The following are some potential directions for future studies:

1. Extending our previous methods in the context of specific biological problems. Specific example include identifying and/or predicting CIS and TRANS regulatory states, structural domains, DNA binding patterns, RNA binding patterns, and structural classes.
2. Extending our previous methods to predict motifs in proteins without using multiple sequence alignment. This can be performed by using the estimated probabilities of small conserved regions among a related set of proteins. By focusing on high probabilities, with large log likelihood ratios, and small variance, it is possible to have a narrow enough search space to experimentally search for overlapping regions. The overlapping regions become specified motifs. This method will be used to predict signal sites, localization sites, interaction sites, target sites, catalytic sites, and any other possible active sights in proteins.
3. Extending our previous methods to predict the involvement of individual residues with various biological processes by focusing on a window of amino acids around a residue instead on the whole sequence. Previous methods [4-8] have focused on

predicting interactions sites (e.g., protein-protein interactions, protein-DNA interactions, and protein-RNA interactions). These methods could be extended by using our proposed methods and focusing on other functional relevant processes such as active sites and localization sites.

4. Extending our previous methods by using reduced alphabets. It would be potentially insightful to perform a systematic study of the relationship among alphabet size, sequence similarity, the size of the dataset, and the prediction performance in protein classification problems. As the alphabet size increases the classifier gains sensitivity at the cost of lost selectivity. This cost is amplified as the sequence similarity becomes smaller. When the amount of data available is less there is less desire to have a highly sensitive classifier. The goal is to find the ideal trade-off between sensitivity and selectivity given the available data. Information theory could be used to determine the minimal information necessary to build a reliable classifier for a given dataset. Reduced alphabets that would be of interest to explore include: alphabets based on physicochemical properties of amino acids, substitution matrices, structural matrices (e.g. Miyazawa–Jernigan (MJ) matrix [9-14]), and random alphabets.
5. Extending our previous computational methods to predict protein classifications based on taxonomies, directed acyclic graphs (DAG), or other structural hierarchies. Currently there are growing amounts of classification schemas based on these approaches. Examples include the Gene Ontology project (DAG), the Structural Classification of Proteins (SCOP) [15-17] (taxonomy), and EC classification [18] (taxonomy). Current methods that perform predictions on these

classifications must take a subset of the data that corresponds to a specific level in the hierarchy. Our methods would provide a quick and flexible way to carry out predictions at any given level of the hierarchy or prediction of any path in the hierarchy.

6. Developing additional web servers for prediction on various classification tasks. These tasks will include Gene Ontology functional labels, protein secondary structure, and subcellular localization data sets. Each of these web servers would provide an easy to use web interface for prediction, a stand-alone version of the software, the data used to train the classifier, and user information for the effective use of a web server.
7. Extending our binding pattern method to predict binding locations. Instead of classifying proteins as a whole, subsequences (centered around each amino acid) could be used as input to a machine-learning approach. For a given protein, the method would classify each amino acid based on whether it is a binding location. Based on all the predictions, local clusters on the primary sequence or on the 3D-structure (if available) can be identified as potential binding sites or patches.
8. Extending our binding pattern prediction server to identify potential binding partners in yeast or other species. Given a protein sequence, PFAM domains can be found on that sequence. These domain(s) may be known to interact with other PFAM domains (data found in iPFAM). Any known protein that contains these second sets of domains are now potential interaction partners to the original sequence. The set of known interacting PFAM domains in iPFAM can be further filtered to create a more stringent set. These filters could include: only domains

that also interact in known orthologs (or syntelogs) and domains that interact in proteins with the same subcellular localization.

Future work toward understanding the biology of G4-quadruplexes

Identification and analysis of G4Q in the maize genome gave new insights into the roles of such structural elements in gene regulation and pathway coordination. Research in G4Q elements is an emerging field and has only just begun on plants. The following are some potential directions for future studies:

1. Exploring the differences in G4Q tendencies in plants and animals. G4Q elements are enriched in the promoters regions of human and plant genes. In humans the enrichment shows a bimodal distribution on the coding strand with a large peak just upstream of the TSS and a smaller peak slightly downstream of the TSS. In most plants there is a single peak on the template strand just downstream of the TSS. An exception is in Arabidopsis where G4Q elements are actually underrepresented in gene regions. [19] Future work could focus on the difference between the enrichments on different strands for plants versus animals.
2. Exploring whether G4Qs in the 3' UTR in plant genes have a role in alternative polyadenylation and mRNA shortening. This analysis has been conducted on humans [20], but has yet to be carried out on plants.
3. Exploring recent discoveries of RNA aptamers binding to ligand that allows for G4Q formation (also R-loops) [21, 22]. Our work focused on a genome-wide scan of the G4Q motif. Recent work has shown that other structural features near the G4Q region may interact, allow, or disallow the formation of the G4Q structure.

An additional scan for other known structural element motifs and their proximity to G4Q locations could easily be performed.

4. Explore co-occurrence of other genome features (e.g., Robertson's *Mutator* transposable elements [23]) beyond genes and gene features. Through preliminary work we found there may be a high co-occurrence of *Mu* elements near G4Q locations. Other studies in humans [24] have looked at SNP data in regards to G4Q locations, especially SNPs that could potentially disrupt the structural formation of a G4Q.
5. Explore different types of G4Qs including non-canonical G4Q motifs, smaller G-repeats, and larger loop sizes. Several databases [25-29] currently exist that contain known and/or potential G4Q forming sequences including non-canonical G4Qs. Recent research [30, 31] has also showed the effect of loop size and G4Q stability. Our current work uses the standard canonical definition for G4Q motif, but a more systematic approach can be taken varying the size of G-repeats and loops.
6. Cluster G4Qs based on sequence similarity and find any relationships in the clusters. Our current work treats each G4Q the same regardless of length, loop-size, G-repeats, or sequence similarity. Our work could be extended by taking the G4Qs in maize, clustering them based on sequence similarity, and exploring each cluster to see if there is correlation to strand, gene localization, types of gene regulation, or genome location.

References

1. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL: **GenBank**. *Nucleic Acids Res* 2008, **36**(Database issue):D25-30.

2. Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, Feng Z, Gilliland GL, Iype L, Jain S *et al*: **The Protein Data Bank**. *Acta Crystallogr D Biol Crystallogr* 2002, **58**(Pt 6 No 1):899-907.

3. Berman HM, Coimbatore Narayanan B, Di Costanzo L, Dutta S, Ghosh S, Hudson BP, Lawson CL, Peisach E, Prlic A, Rose PW *et al*: **Trendspotting in the Protein Data Bank**. *FEBS Lett* 2013, **587**(8):1036-1045.

4. Andorf C, Dobbs D, Honavar V: **Exploring inconsistencies in genome-wide protein function annotations: a machine learning approach**. *BMC Bioinformatics* 2007, **8**:284.

5. Sen TZ, Kloczkowski A, Jernigan RL, Yan C, Honavar V, Ho KM, Wang CZ, Ihm Y, Cao H, Gu X *et al*: **Predicting binding sites of hydrolase-inhibitor complexes by combining several methods**. *BMC Bioinformatics* 2004, **5**:205.

6. Terribilini M, Lee JH, Yan C, Jernigan RL, Carpenter S, Honavar V, Dobbs D: **Identifying interaction sites in "recalcitrant" proteins: predicted protein and RNA binding sites in rev proteins of HIV-1 and EIAV agree with experimental data**. *Pac Symp Biocomput* 2006:415-426.

7. Terribilini M, Lee JH, Yan C, Jernigan RL, Honavar V, Dobbs D: **Prediction of RNA binding sites in proteins from amino acid sequence**. *RNA* 2006, **12**(8):1450-1462.

8. Yan C, Terribilini M, Wu F, Jernigan RL, Dobbs D, Honavar V: **Predicting DNA-binding sites of proteins from amino acid sequence**. *BMC Bioinformatics* 2006, **7**:262.

9. Esteve JG, Falceto F: **A general clustering approach with application to the Miyazawa-Jernigan potentials for amino acids**. *Proteins* 2004, **55**(4):999-1004.

10. Li H, Tang C, Wingreen NS: **Designability of protein structures: a lattice-model study using the Miyazawa-Jernigan matrix**. *Proteins* 2002, **49**(3):403-412.

11. Miyazawa S, Jernigan RL: **How effective for fold recognition is a potential of mean force that includes relative orientations between contacting residues in proteins?** *J Chem Phys* 2005, **122**(2):024901.

12. Miyazawa S, Jernigan RL: **Long- and short-range interactions in native protein structures are consistent/minimally frustrated in sequence space**. *Proteins* 2003, **50**(1):35-43.

13. Miyazawa S, Jernigan RL: **A new substitution matrix for protein sequence searches based on contact frequencies in protein structures**. *Protein Eng* 1993, **6**(3):267-278.

14. Miyazawa S, Kinjo AR: **Properties of contact matrices induced by pairwise interactions in proteins**. *Phys Rev E Stat Nonlin Soft Matter Phys* 2008, **77**(5 Pt 1):051910.

15. Hardin CC, Henderson E, Watson T, Prosser JK: **Monovalent cation induced structural transitions in telomeric DNAs: G-DNA folding intermediates**. *Biochemistry* 1991, **30**(18):4460-4472.

16. Murzin AG, Brenner SE, Hubbard T, Chothia C: **SCOP: a structural classification of proteins database for the investigation of sequences and structures**. *J Mol Biol* 1995, **247**(4):536-540.

17. Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJ, Chothia C, Murzin AG: **Data growth and its impact on the SCOP database: new developments**. *Nucleic Acids Res* 2008, **36**(Database issue):D419-425.

18. Bairoch A: **The ENZYME database in 2000**. *Nucleic Acids Res* 2000, **28**(1):304-305.

19. Mullen MA, Olson KJ, Dallaire P, Major F, Assmann SM, Bevilacqua PC: **RNA G-Quadruplexes in the model plant species Arabidopsis thaliana: prevalence and possible functional roles**. *Nucleic Acids Res* 2010, **38**(22):8149-8163.

20. Huppert JL, Bugaut A, Kumari S, Balasubramanian S: **G-quadruplexes: the beginning and end of UTRs**. *Nucleic Acids Res* 2008, **36**(19):6260-6268.

21. Bing T, Chang T, Yang X, Mei H, Liu X, Shangguan D: **G-quadruplex DNA aptamers generated for systemin**. *Bioorg Med Chem* 2011, **19**(14):4211-4219.

22. Tucker WO, Shum KT, Tanner JA: **G-quadruplex DNA aptamers and their ligands: structure, function and application**. *Curr Pharm Des* 2012, **18**(14):2014-2026.

23. McCarty DR, Settles AM, Suzuki M, Tan BC, Latshaw S, Porch T, Robin K, Baier J, Avigne W, Lai J *et al*: **Steady-state transposon mutagenesis in inbred maize**. *Plant J* 2005, **44**(1):52-61.

24. Baral A, Kumar P, Halder R, Mani P, Yadav VK, Singh A, Das SK, Chowdhury S: **Quadruplex-single nucleotide polymorphisms (Quad-SNP) influence gene expression difference among individuals**. *Nucleic Acids Res* 2012, **40**(9):3800-3811.

25. Cer RZ, Donohue DE, Mudunuri US, Temiz NA, Loss MA, Starner NJ, Halusa GN, Volfovsky N, Yi M, Luke BT *et al*: **Non-B DB v2.0: a database of predicted**

non-B DNA-forming motifs and its associated tools. *Nucleic Acids Res* 2013, **41**(Database issue):D94-D100.

26. Kikin O, D'Antonio L, Bagga PS: **QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences**. *Nucleic Acids Res* 2006, **34**(Web Server issue):W676-682.

27. Kikin O, Zappala Z, D'Antonio L, Bagga PS: **GRSDB2 and GRS_UTRdb: databases of quadruplex forming G-rich sequences in pre-mRNAs and mRNAs**. *Nucleic Acids Res* 2008, **36**(Database issue):D141-148.

28. Yadav VK, Abraham JK, Mani P, Kulshrestha R, Chowdhury S: **QuadBase: genome-wide database of G4 DNA--occurrence and conservation in human, chimpanzee, mouse and rat promoters and 146 microbes**. *Nucleic Acids Res* 2008, **36**(Database issue):D381-385.

29. Zhang R, Lin Y, Zhang CT: **Greglist: a database listing potential G-quadruplex regulated genes**. *Nucleic Acids Res* 2008, **36**(Database issue):D372-376.

30. Guedin A, Gros J, Alberti P, Mergny JL: **How long is too long? Effects of loop size on G-quadruplex stability**. *Nucleic Acids Res* 2010, **38**(21):7858-7868.

31. Pandey S, Agarwala P, Maiti S: **Effect of loops and g-quartets on the stability of RNA g-quadruplexes**. *J Phys Chem B* 2013, **117**(23):6896-6905.

APPENDICES

APPENDIX A

Supplementary Data: Machine-Learning Results

Machine-Learning approaches to predict Gene Ontology and/or UniProt Functional labels:

Experiments #1 - #7 show the results of several machine-learning approaches used in this study. After the summary of each experiment, five sections summarize the performance obtained in each experiment. The first section displays the number of correctly and incorrectly classified instances along with the percent accuracy and percent error for each of the three classifiers (See **Methods** for details on each of the classifiers). The section summarizes the correlation coefficients and kappa coefficients for each of the methods. The third, fourth, and fifth sections show the individual performance of each of the three classifiers. Each row represents the functional class provided by the data label source (either AmiGO or UniProt). Each column represents the functional class *predicted* by the given classifier. The far right column shows the Recall for each of the classes and the last row shows the Precision for each of the classes (in **Methods**). The accuracy of the classifier can be found where the Recall column and Precision row intersect.

Experiment #1: Human cross-validation Summary

Training Method: 10-fold cross-validation
Species: Human
Evidence Code: All
Data Label Source: AmiGO
Total Number of Instances: 330

	Classifier #1		Classifier #2		**Classifier #3**	
	Instances	Percent	Instances	Percent	**Instances**	**Percent**
Correctly Classified	302	91.5%	316	95.8%	**294**	**89.1%**
Incorrectly Classified	28	8.5%	14	4.2%	**36**	**10.9%**

	Correlation Coefficient			Kappa Coefficient
	Ser/Thr	Tyr	Dual	
Classifier #1	0.79	--	--	0.79
Classifier #2	--	0.89	--	0.89
Classifier #3	0.82	0.86	0.30	0.76

=== Classifier #1 Overall Performance Evaluation ===

classified as →	GO0004674 (Ser/Thr)	GO0004713 (Tyr) and Dual	Recall
GO0004674(Ser/Thr)	227	6	0.97
GO0004713(Tyr) and Dual	22	75	0.77
Precision →	0.91	0.93	Accuracy = 0.92

=== Classifier #2 Overall Performance Evaluation ===

classified as →	GO0004713 (Tyr)	GO0004674 (Ser/Thr) and Dual	Recall
GO0004713(Tyr)	81	9	0.90
GO0004674(Ser/Thr) and Dual	5	235	0.98
Precision →	0.94	0.96	Accuracy = 0.96

=== Classifier #3 Overall Performance Evaluation ===

classified as →	GO0004713 (Tyr)	GO0004674 (Ser/Thr)	Dual	Recall
GO0004713 (Tyr)	67	6	17	0.74
GO0004674 (Ser/Thr)	0	222	11	0.95
Dual	0	2	5	0.71
Precision →	1.00	0.97	0.15	Accuracy = 0.89

Experiment #2: Human Training Set / Mouse Test Set Summary

Training Method: Trained on Human / Tested on Mouse
Species: Human (Training) / Mouse (Testing)
Evidence Code: All
Data Label Source: AmiGO
Total Number of Instances: 330 (Training) / 244 (Testing)

	Classifier #1		Classifier #2		Classifier #3	
	Instances	Percent	Instances	Percent	Instances	Percent
Correctly Classified	57	23.4%	86	35.2%	37	15.1%
Incorrectly Classified	187	76.6%	158	64.8%	207	84.9%

	Correlation Coefficient			Kappa Coefficient
	Ser/Thr	Tyr	Dual	
Classifier #1	-0.43	--	--	-0.30
Classifier #2	--	-0.40	--	-0.38
Classifier #3	-0.40	-0.43	-0.01	-0.40

=== Classifier #1 Overall Performance Evaluation ===

classified as →	GO0004674 (Ser/Thr)	GO0004713 (Tyr) and Dual	Recall
GO0004674(Ser/Thr)	29	42	0.41
GO0004713(Tyr) and Dual	145	28	0.16
Precision →	0.17	0.40	Accuracy = 0.23

=== Classifier #2 Overall Performance Evaluation ===

classified as →	GO0004713 (Tyr)	GO0004674 (Ser/Thr) and Dual	Recall
GO0004713(Tyr)	7	99	0.07
GO0004674(Ser/Thr) and Dual	59	79	0.57
Precision →	0.11	0.44	Accuracy = 0.35

=== Classifier #3 Overall Performance Evaluation ===

classified as →	GO0004713 (Tyr)	GO0004674 (Ser/Thr)	Dual	Recall
GO0004713 (Tyr)	7	96	3	0.07
GO0004674 (Ser/Thr)	42	29	0	0.41
Dual	17	49	1	0.01
Precision →	0.11	0.17	0.25	Accuracy = 0.15

Experiment #3: Human Training Set/ Mouse Test Set (RCA only) Summary

Training Method: Trained on Human / Tested on Mouse
Species: Human (Training) / Mouse (Testing)
Evidence Code: RCA (inferred from Reviewed Computational Analysis)
Data Label Source: AmiGO
Total Number of Instances: 330 (Training) / 211 (Testing)

	Classifier #1		Classifier #2		Classifier #3	
	Instances	Percent	Instances	Percent	Instances	Percent
Correctly Classified	13	6.2%	43	20.4%	9	4.2%
Incorrectly Classified	198	93.8%	168	79.6%	202	95.8%

	Correlation Coefficient			Kappa Coefficient
	Ser/Thr	Tyr	Dual	
Classifier #1	-0.85	--	--	-0.60
Classifier #2	--	-0.64	--	-0.57
Classifier #3	-0.64	-0.85	0.00	0.50

=== Classifier #1 Overall Performance Evaluation ===

classified as →	GO0004674 (Ser/Thr)	GO0004713 (Tyr) and Dual	Recall
GO0004674(Ser/Thr)	9	55	0.14
GO0004713(Tyr) and Dual	143	4	0.03
Precision →	0.06	0.07	Accuracy = 0.06

=== Classifier #2 Overall Performance Evaluation ===

classified as →	GO0004713 (Tyr)	GO0004674 (Ser/Thr) and Dual	Recall
GO0004713(Tyr)	0	109	0.00
GO0004674(Ser/Thr) and Dual	59	43	0.42
Precision →	0.00	0.28	Accuracy = 0.20

=== Classifier #3 Overall Performance Evaluation ===

classified as →	GO0004713 (Tyr)	GO0004674 (Ser/Thr)	Dual	Recall
GO0004713 (Tyr)	0	109	0	0.00
GO0004674 (Ser/Thr)	55	9	0	0.14
Dual	4	34	0	0.00
Precision →	0.00	0.06	0.00	Accuracy = 0.04

Experiment #4: Human Training Set / Mouse Test Set (at least one RCA code) Summary

Training Method: Trained on Human / Tested on Mouse
Species: Human (Training) / Mouse (Testing)
Evidence Code: at least one RCA evidence code (but also including other evidence codes in the annotation)
Data Label Source: AmiGO
Total Number of Instances: 330 (Training) / 211 (Testing)

	Classifier #1		Classifier #2		Classifier #3	
	Instances	Percent	Instances	Percent	Instances	Percent
Correctly Classified	26	12.3%	58	27.5%	9	4.2%
Incorrectly Classified	185	87.7%	153	72.5%	202	95.8%

	Correlation Coefficient			Kappa Coefficient
	Ser/Thr	Tyr	Dual	
Classifier #1	-0.68	--	--	-0.42
Classifier #2	--	-0.56	--	-0.52
Classifier #3	-0.56	-0.68	0.00	-0.37

=== Classifier #1 Overall Performance Evaluation ===

classified as →	GO0004674 (Ser/Thr)	GO0004713 (Tyr) and Dual	Recall
GO0004674(Ser/Thr)	9	42	0.18
GO0004713(Tyr) and Dual	143	17	0.11
Precision →	0.06	0.29	Accuracy = 0.12

=== Classifier #2 Overall Performance Evaluation ===

classified as →	GO0004713 (Tyr)	GO0004674 (Ser/Thr) and Dual	Recall
GO0004713(Tyr)	0	94	0.00
GO0004674(Ser/Thr) and Dual	59	58	0.50
Precision →	0.00	0.38	Accuracy = 0.27

=== Classifier #3 Overall Performance Evaluation ===

classified as →	GO0004713 (Tyr)	GO0004674 (Ser/Thr)	Dual	Recall
GO0004713 (Tyr)	0	94	0	0.00
GO0004674 (Ser/Thr)	42	9	0	0.18
Dual	17	49	0	0.00
Precision →	0.00	0.06	0.00	Accuracy = 0.04

Experiment #5: Human Training Set / Mouse Test Set (UniProt label) Summary

Training Method: Trained on Human / Tested on Mouse
Species: Human (Training) / Mouse (Testing)
Evidence Code: All
Data Label Source: UniProt
Total Number of Instances: 330 (Training) / 244 (Testing)

	Classifier #1		Classifier #2		Classifier #3	
	Instances	Percent	Instances	Percent	Instances	Percent
Correctly Classified	234	95.9%	241	98.8%	233	95.4%
Incorrectly Classified	10	4.1%	3	1.2%	11	4.6%

	Correlation Coefficient			Kappa Coefficient
	Ser/Thr	Tyr	Dual	
Classifier #1	0.90	--	--	0.90
Classifier #2	--	0.97	--	0.97
Classifier #3	0.96	0.90	0.43	0.90

=== Classifier #1 Overall Performance Evaluation ===

classified as →	GO0004674 (Ser/Thr)	GO0004713 (Tyr) and Dual	Recall
GO0004674(Ser/Thr)	166	2	0.99
GO0004713(Tyr) and Dual	8	68	0.89
Precision →	0.95	0.97	Accuracy = 0.96

=== Classifier #2 Overall Performance Evaluation ===

classified as →	GO0004713 (Tyr)	GO0004674 (Ser/Thr) and Dual	Recall
GO0004713(Tyr)	64	1	0.98
GO0004674(Ser/Thr) and Dual	2	177	0.99
Precision →	0.97	0.99	Accuracy = 0.99

=== Classifier #3 Overall Performance Evaluation ===

classified as →	GO0004713 (Tyr)	GO0004674 (Ser/Thr)	Dual	Recall
GO0004713 (Tyr)	64	0	1	0.98
GO0004674 (Ser/Thr)	2	166	0	0.99
Dual	0	8	3	0.27
Precision →	0.97	0.95	0.75	Accuracy = 0.95

Experiment #6: Human Training Set / Mouse Test Set (RCA only/UniProt label) Summary

Training Method: Trained on Human / Tested on Mouse
Species: Human (Training) / Mouse (Testing)
Evidence Code: All
Data Label Source: UniProt
Total Number of Instances: 330 (Training) / 211 (Testing)

	Classifier #1		Classifier #2		Classifier #3	
	Instances	Percent	Instances	Percent	Instances	Percent
Correctly Classified	205	97.1%	209	99.1%	205	97.1%
Incorrectly Classified	6	2.9%	2	0.9%	6	2.9%

	Correlation Coefficient			Kappa Coefficient
	Ser/Thr	Tyr	Dual	
Classifier #1	0.93	--	--	0.93
Classifier #2	--	0.98	--	0.98
Classifier #3	0.98	0.94	0.00	0.93

=== Classifier #1 Overall Performance Evaluation ===

classified as →	GO0004674 (Ser/Thr)	GO0004713 (Tyr) and Dual	Recall
GO0004674(Ser/Thr)	148	2	0.99
GO0004713(Tyr) and Dual	4	57	0.93
Precision →	0.97	0.97	Accuracy = 0.97

=== Classifier #2 Overall Performance Evaluation ===

classified as →	GO0004713 (Tyr)	GO0004674 (Ser/Thr) and Dual	Recall
GO0004713(Tyr)	57	0	1.00
GO0004674(Ser/Thr) and Dual	2	152	0.99
Precision →	0.97	1.00	Accuracy = 0.99

=== Classifier #3 Overall Performance Evaluation ===

classified as →	GO0004713 (Tyr)	GO0004674 (Ser/Thr)	Dual	Recall
GO0004713 (Tyr)	57	0	0	1.00
GO0004674 (Ser/Thr)	2	148	0	0.99
Dual	0	4	0	0.00
Precision →	0.97	0.97	0.00	Accuracy = 0.97

Experiment #7: Human Training Set / Mouse Test Set (at least one RCA code/UniProt label)*:

* Please note that these results were identical to Experiment #6 results since the evaluation is based on UniProt labels and not the labels given by AmiGO.

Comparing AmiGO and UniProt Labels:

Experiments #8 - #10 show tables corresponding to the number of proteins with the given functional labels given by AmiGO and by UniProt. These data were taken directly from each database; no machine learning approaches were used. Columns represent the number of proteins retrieved by AmiGO that had the corresponding Column header as a functional label given by AmiGO. Rows represent the number of proteins retrieved by AmiGO that had the corresponding Row header as functional evidence within UniProt.

Experiment #8: Mouse Data (All): UniProt versus AmiGO

UniProt Labels	GO0004713 (Tyr)	GO0004674 (Ser/Thr)	Dual	UniProt Total
GO0004713 (Tyr)	8	41	16	65
GO0004674 (Ser/Thr)	91	28	49	168
Dual	7	2	2	11
AmiGO Total →	106	71	67	Total # proteins =244

Instances in Agreement:	38	15.6%
Instances in Disagreement:	206	84.4%

Experiment #9: Mouse Data (RCA evidence code only): UniProt versus AmiGO

UniProt Labels	GO0004713 (Tyr)	GO0004674 (Ser/Thr)	Dual	UniProt Total
GO0004713 (Tyr)	0	54	3	57
GO0004674 (Ser/Thr)	105	10	35	150
Dual	4	0	0	4
AmiGO Total →	109	64	38	Total # proteins =211

Instances in Agreement:	10	4.7%
Instances in Disagreement:	201	95.3%

Experiment #10: Mouse Data (at least one RCA evidence code): UniProt versus AmiGO

UniProt Labels	GO0004713 (Tyr)	GO0004674 (Ser/Thr)	Dual	UniProt Total
GO0004713 (Tyr)	0	41	16	57
GO0004674 (Ser/Thr)	91	10	49	150
Dual	3	0	1	4
AmiGO Total →	94	51	66	Total # proteins =211

Instances in Agreement:	11	5.2%
Instances in Disagreement:	200	94.8%

APPENDIX B

Supplementary Note:

Because there is only a non-curated reference to the work done on “Rat ISS GO annotations from MGI's mouse gene data”, we provide the abstract and a link to the original reference report. The Rat Genome Database ID for this report is 1578720. The original report was created on April, 6, 2006.

Abstract

All the annotations assigned to mouse genes by the Mouse Genome Database (http://www.informatics.jax.org/) were uploaded from the Gene Ontology Consortium website (http://www.geneontology.org/GO.current.annotations.shtml). Annotations with "IEA" or "ND" Evidence Codes were removed and the remainder loaded onto the corresponding orthologous rat genes with an evidence code for the annotation to the rat gene of ISS. The decision to use ISS for the evidence code was made following a personal communication between Dr. Susan Bromberg (RGD) and Dr. Judith Blake (MGI/GO Consortium).

Original link:

http://rgd.mcw.edu/tools/references/references_view.cgi?id=1578720

APPENDIX C.

Supplementary Table 1:

Evidence Codes for AmiGO annotations of mouse protein kinases (See Table legend below).

Gene ID	AmiGO label	4674 Evidence Code	4713 Evidence Code
2610018G03Rik	4713		RCA
Acvr1b	4713		RCA
Acvr2a	4713		RCA
Acvr2b	4713		RCA
Acvrl1	4713		RCA
Adrbk1	4713		RCA
Akt1	4713		RCA
Alk	4674	RCA	
Araf	4713		RCA
Atm	4674	TAS	
Aurka	4713		RCA
Aurkb	4713		RCA
Axl	4674	RCA	
Blk	4674	RCA	
Bmpr1a	4713		RCA
Bmpr1b	4713		RCA
Bmpr2	4713		RCA
Bmx	4674	RCA	
Btk	4674 / 4713	RCA	IDA
Camk1	4713		RCA
Camk1d	4674	ISS	
Camk1g	4674	RCA	
Camk2a	4674	IMP	
Camk2b	4674 / 4713	IMP	RCA
Camk2g	4674 / 4713	IMP/RCA	RCA
Camk4	4674	TAS	
Camkk1	4674 / 4713	ISS	RCA
Ccrk	4674 / 4713	RCA	RCA
Cdc2a	4713		RCA
Cdc2l5	4674 / 4713	RCA	RCA
Cdk5	4674	IDA	
Cdk7	4674 / 4713	ISS	RCA
Cdk9	4713		RCA
Cdkl1	4674 / 4713	RCA	RCA
Cdkl3	4674 / 4713	RCA	RCA
Cdkl4	4674 / 4713	RCA	RCA
Chek1	4713		RCA
Chek2	4713		RCA
Chuk	4713		RCA
Cit	4674	IDA	
Clk1	4674 / 4713	IDA	IDA
Clk2	4713		IDA

- Supplementary Table 1 continued -

Clk3	4713		IDA
Clk4	4713		IDA
Cpne3	4674	RCA	
Csf1r	4674	RCA	IEA
Csk	4674	RCA	
Csnk1d	4713		RCA
Csnk1e	4713		RCA
Csnk1g2	4713		RCA
Csnk2a2	4674 / 4713	RCA	RCA
Dapk2	4713		RCA
Dapk3	4713		RCA
Dcamkl2	4674 / 4713	RCA	RCA
Ddr1	4674	RCA	
Dmpk	4713		RCA
Dyrk1a	4713		ISS
Egfr	4713		IDA
Eif2ak1	4713		RCA
Eif2ak3	4713		RCA
Eif2ak4	4674 / 4713	IDA	RCA
Epha1	4674 / 4713	RCA	RCA
Epha2	4674	RCA	
Epha3	4674 / 4713	RCA	IDA
Epha4	4674	RCA	
Epha5	4674	RCA	
Epha6	4674	RCA	
Epha7	4674	RCA	
Epha8	4674	RCA	
Ephb2	4674 / 4713	RCA	IDA
Ephb3	4674 / 4713	RCA	TAS
Ephb4	4674	RCA	
Ephb6	4674	RCA	
Erbb2	4674 / 4713	RCA	RCA
Ern2	4713		RCA
Fgfr1	4674 / 4713	RCA	TAS
Fgfr2	4674	RCA	
Fgfr3	4713		IDA
Fgfr4	4674	RCA	
Fgr	4674	RCA	
Flt1	4674	RCA	
Flt3	4674	RCA	
Flt4	4674	RCA	
Fyn	4713		IDA
Gprk2l	4674 / 4713	RCA	RCA
Gprk5	4674 / 4713	RCA	RCA
Gprk6	4713		RCA
Grk1	4713		RCA
Gsg2	4674	IDA	
Gsk3b	4674	IDA	

- Supplementary Table 1 continued -

Hck	4674	RCA	
Hipk2	4674	ISS	
Hipk3	4713		RCA
Hunk	4713		RCA
Ick	4674	IDA	
Igf1r	4674	RCA	
Ikbkb	4713		RCA
Ikbke	4674 / 4713	IDA	RCA
Ilk	4674	ISS	
Insrr	4674	RCA	
Irak3	4674 / 4713	RCA	RCA
Itk	4674	RCA	
Jak1	4713		IDA
Jak2	4674 / 4713	RCA	IDA
Jak3	4713		IDA
Kdr	4674	RCA	
Kit	4674 / 4713	RCA	IDA
Ksr1	4713		RCA
Lats1	4713		RCA
Lck	4674	RCA	
Limk1	4713		RCA
Lrrk1	4674 / 4713	RCA	RCA
Ltk	4674	RCA	
Lyn	4713		IDA
Map2k3	4713		RCA
Map2k5	4713		RCA
Map3k12	4713		RCA
Map3k14	4713		RCA
Map3k3	4713		RCA
Map3k4	4713		RCA
Map3k7	4713		RCA
Map3k8	4713		RCA
Map4k1	4674 / 4713	RCA	RCA
Map4k2	4713		RCA
Mapk1	4674 / 4713	ISS	RCA
Mapk10	4713		RCA
Mapk11	4713		RCA
Mapk12	4713		RCA
Mapk13	4713		RCA
Mapk14	4713		RCA
Mapk3	4713		RCA
Mapk7	4713		RCA
Mapk8	4713		RCA
Mapk9	4713		RCA
Mapkapk2	4713		RCA
Mapkapk5	4713		RCA
Mark1	4674 / 4713	RCA	RCA
Mark2	4713		RCA

- Supplementary Table 1 continued -

Mast1	4713		RCA
Mast2	4674 / 4713	IDA	RCA
Mastl	4674	RCA	
Matk	4674	RCA	
Melk	4674 / 4713	ISS	RCA
Mertk	4674	RCA	
Met	4674 / 4713	RCA	IDA
Mknk1	4713		RCA
Mos	4713		RCA
Musk	4713		TAS
Mylk2	4674	RCA	
Nek11	4674 / 4713	RCA	RCA
Nek2	4713		RCA
Nek4	4713		RCA
Nek6	4674 / 4713	RCA	RCA
Nek7	4674	RCA	
Nlk	4674 / 4713	RCA	RCA
Npr1	4674 / 4713	RCA	RCA
Oxsr1	4674 / 4713	RCA	RCA
Pak1	4674 / 4713	ISS	RCA
Pak2	4674	ISS	
Pak3	4713		RCA
Pak4	4674 / 4713	RCA	RCA
Pak7	4674 / 4713	IDA / RCA	RCA
Pask	4674 / 4713	RCA	RCA
Pbk	4674 / 4713	ISS / RCA	RCA
Pctk1	4713		RCA
Pctk3	4713		RCA
Pdgfra	4674	RCA	
Pdgfrb	4674	RCA	
Pdpk1	4674	IDA	
Pftk1	4674 / 4713	ISS	RCA
Phkg1	4713		RCA
Pim1	4713		RCA
Pim2	4674 / 4713	IDA	RCA
Pink1	4713		RCA
Pkmyt1	4713		RCA
Pkn2	4674 / 4713	RCA	RCA
Plk1	4713		RCA
Plk2	4713		RCA
Plk4	4713		RCA
Pnck	4713		RCA
Prkaca	4674 / 4713	IDA	RCA
Prkca	4674 / 4713	IDA	RCA
Prkcb1	4674 / 4713	RCA	RCA
Prkcc	4674 / 4713	RCA	RCA
Prkch	4713		RCA
Prkci	4713		RCA

- Supplementary Table 1 continued -

Prkcm	4713		RCA
Prkcz	4713		RCA
Prkg2	4713		RCA
Prkx	4674 / 4713	ISS	RCA
Prpf4b	4713		RCA
Ret	4674 / 4713	RCA	TAS
Ripk1	4713		RCA
Ripk5	4674 / 4713	RCA	RCA
Rock1	4713		RCA
Ror1	4674 / 4713	RCA	TAS
Ror2	4674 / 4713	RCA	TAS
Rps6ka1	4713		RCA
Rps6ka3	4674	IDA	
Rps6ka5	4674 / 4713	RCA	RCA
Rps6kb2	4713		RCA
Rps6kl1	4674	RCA	
Sbk1	4674 / 4713	RCA / ISS	RCA
Sgk2	4713		RCA
Sgk3	4674	IDA	
Slk	4713		RCA
Snrk	4674 / 4713	RCA	RCA
Src	4674 / 4713	RCA	IMP
Srpk1	4713		RCA
Srpk2	4713		RCA
Stk10	4674 / 4713	TAS	RCA
Stk16	4713		RCA
Stk17b	4674 / 4713	RCA	RCA
Stk23	4674	IDA	
Stk32b	4713		RCA
Stk36	4674 / 4713	RCA	RCA
Stk38l	4674	ISS / RCA	
Syk	4674 / 4713	RCA	IDA
Tbk1	4674 / 4713	RCA	RCA
Tec	4674	RCA	
Tek	4674	RCA	
Tgfbr1	4713		RCA
Tgfbr2	4713		RCA
Tie1	4674	RCA	
Tlk1	4674 / 4713	RCA	RCA
Tlk2	4713		RCA
Tnk1	4674	RCA	
Tnk2	4674 / 4713	RCA	RCA
Tssk1	4713		RCA
Tssk2	4713		RCA
Tssk6	4674 / 4713	RCA	RCA
Vrk3	4674	RCA	
Yes1	4674	RCA	
Zap70	4713		IDA

Legend for Supplementary Table 1:

Evidence Codes for AmiGO annotations

The evidence codes for all 244 mouse protein kinases used in this study are displayed in this table. All ***Mouse Gene ID*** numbers were obtained from the AmiGO protein record. The ***Mouse AmiGO Label*** field is "4713" (Tyr) if a query in AmiGO for the GO label GO0004713 returns the corresponding protein for mouse proteins, "4674" (Ser/Thr) if a query in AmiGO for the GO label GO0004674 returns the corresponding protein for mouse proteins, or "4674 / 4713" if a query in AmiGO for both GO labels GO0004674 and GO0004713 returns the corresponding protein. The ***4674 Evidence Code*** and ***4713 Evidence Code*** fields contain the evidence code(s)* provided by AmiGO for a given protein belonging to either the Gene Ontology family GO0004674 or GO0004713. If this field is empty, then the given protein was not included in the list of proteins returned by AmiGO for that Gene Ontology family.

*Examples of evidence codes provided by AmiGO:

- IC: Inferred by Curator
- IDA: Inferred from Direct Assay
- IEA: Inferred from Electronic Annotation
- IEP: Inferred from Expression Pattern
- IGI: Inferred from Genetic Interaction
- IMP: Inferred from Mutant Phenotype
- IPI: Inferred from Physical Interaction
- ISS: Inferred from Sequence or Structural Similarity
- NAS: Non-traceable Author Statement
- ND: No biological Data available
- RCA: inferred from Reviewed Computational Analysis
- TAS: Traceable Author Statement
- NR: Not Recorded

More details on evidence code: http://www.geneontology.org/GO.evidence.shtml

APPENDIX D

Supplementary Table 2:

AmiGO annotations versus UniProt annotations [with UniProt Evidence] (See Table legend below)

Gene ID	AmiGO label	UniProt label	UniProt Evidence
2610018G03Rik	4713	4674	Serine/threonine-protein kinase MST4
Acvr1b	4713	4674	Serine/threonine-protein kinase
Acvr2a	4713	4674	Serine/threonine-protein kinase
Acvr2b	4713	4674	Serine/threonine-protein kinase
Acvrl1	4713	4674	Serine/threonine-protein kinase
Adrbk1	4713	4674	Serine/threonine-protein kinase
Akt1	4713	4674	RAC-alpha serine/threonine-protein kinase
Alk	4674	4713	ALK tyrosine kinase receptor precursor
Araf	4713	4674	A-Raf proto-oncogene serine/threonine-protein kinase
Atm	4674	4674	Serine/threonine-protein kinase
Aurka	4713	4674	Serine/threonine-protein kinase 6
Aurkb	4713	4674	Serine/threonine-protein kinase
Axl	4674	4713	Proto-oncogene tyrosine-protein kinase MER precursor
Blk	4674	4713	Tyrosine-protein kinase BLK
Bmpr1a	4713	4674	Serine/threonine-protein kinase
Bmpr1b	4713	4674	Serine/threonine-protein kinase
Bmpr2	4713	4674	Belongs to the Ser/Thr protein kinase family
Bmx	4674	4713	Cytoplasmic tyrosine-protein kinase BMX
Btk	4674 / 4713	4713	Tyrosine-protein kinase BTK
Camk1	4713	4674	Serine/threonine-protein kinase
Camk1d	4674	4674	Belongs to the Ser/Thr protein kinase family
Camk1g	4674	4674	Serine/threonine-protein kinase
Camk2a	4674	4674	Serine/threonine-protein kinase
Camk2b	4674 / 4713	4674	Serine/threonine-protein kinase
Camk2g	4674 / 4713	4674	Serine/threonine-protein kinase
Camk4	4674	4674	Belongs to the Ser/Thr protein kinase family
Camkk1	4674 / 4713	4674	Serine/threonine-protein kinase
Ccrk	4674 / 4713	4674	Serine/threonine-protein kinase
Cdc2a	4713	4674	Serine/threonine-protein kinase
Cdc2l5	4674 / 4713	4674	Serine/threonine-protein kinase
Cdk5	4674	4674	Belongs to the Ser/Thr protein kinase family
Cdk7	4674 / 4713	4674 / 4713	Protein-tyrosine kinase; Belongs to the Ser/Thr protein kinase
Cdk9	4713	4674 / 4713	Protein-tyrosine kinase; Belongs to the Ser/Thr protein kinase
Cdkl1	4674 / 4713	4674	Serine/threonine-protein kinase
Cdkl3	4674 / 4713	4674	Serine/threonine-protein kinase
Cdkl4	4674 / 4713	4674	Serine/threonine-protein kinase
Chek1	4713	4674	Serine/threonine-protein kinase
Chek2	4713	4674	Serine/threonine-protein kinase
Chuk	4713	4674	Serine/threonine-protein kinase
Cit	4674	4674 / 4713	Dual specificity protein kinase activity
Clk1	4674 / 4713	4674 / 4713	Phosphorylates serines, threonines and tyrosines
Clk2	4713	4674 / 4713	Tyrosine-protein kinase, Serine/threonine-protein kinase

- Supplementary Table 2 continued -

Clk3	4713	4674 / 4713	Tyrosine-protein kinase, Serine/threonine-protein kinase
Clk4	4713	4674 / 4713	Tyrosine-protein kinase, Serine/threonine-protein kinase
Cpne3	4674	4674	Serine/threonine-protein phosphatase
Csf1r	4674	4713	protein tyrosine-kinase transmembrane receptor
Csk	4674	4713	Tyrosine-protein kinase CSK
Csnk1d	4713	4674	Serine/threonine-protein kinase
Csnk1e	4713	4674	Serine/threonine-protein kinase
Csnk1g2	4713	4674	Serine/threonine-protein kinase
Csnk2a2	4674 / 4713	4674	Serine/threonine-protein kinase
Dapk2	4713	4674	Belongs to the Ser/Thr protein kinase family
Dapk3	4713	4674	Belongs to the Ser/Thr protein kinase family
Dcamkl2	4674 / 4713	4674	Serine/threonine-protein kinase
Ddr1	4674	4713	Tyrosine kinase DDR
Dmpk	4713	4674	Belongs to the Ser/Thr protein kinase family
Dyrk1a	4713	4674 / 4713	Serine/threonine-protein kinase; Tyrosine-protein kinase
Egfr	4713	4713	Tyrosine-protein kinase
Eif2ak1	4713	4674	Serine/threonine-protein kinase
Eif2ak3	4713	4674	Serine/threonine-protein kinase
Eif2ak4	4674 / 4713	4674	Serine/threonine-protein kinase
Epha1	4674 / 4713	4713	Tyrosine-protein kinase receptor
Epha2	4674	4713	Tyrosine-protein kinase receptor
Epha3	4674 / 4713	4713	Tyrosine-protein kinase receptor
Epha4	4674	4713	Tyrosine-protein kinase receptor
Epha5	4674	4713	Tyrosine-protein kinase receptor
Epha6	4674	4713	Tyrosine-protein kinase receptor
Epha7	4674	4713	Tyrosine-protein kinase receptor
Epha8	4674	4713	Tyrosine-protein kinase receptor
Ephb2	4674 / 4713	4713	Tyrosine-protein kinase receptor
Ephb3	4674 / 4713	4713	Tyrosine-protein kinase receptor
Ephb4	4674	4713	Tyrosine-protein kinase receptor
Ephb6	4674	4713	Tyrosine-protein kinase receptor
Erbb2	4674 / 4713	4713	Receptor tyrosine-protein kinase erbB-2 precursor
Ern2	4713	4674	Serine/threonine-protein
Fgfr1	4674 / 4713	4713	Belongs to the Tyr protein kinase family
Fgfr2	4674	4713	Belongs to the Tyr protein kinase family
Fgfr3	4713	4713	Belongs to the Tyr protein kinase family
Fgfr4	4674	4713	Belongs to the Tyr protein kinase family
Fgr	4674	4713	Proto-oncogene tyrosine-protein kinase FGR
Flt1	4674	4713	Belongs to the Tyr protein kinase family
Flt3	4674	4713	Belongs to the Tyr protein kinase family
Flt4	4674	4713	Belongs to the Tyr protein kinase family
Fyn	4713	4713	Proto-oncogene tyrosine-protein kinase Fyn
Gprk2l	4674 / 4713	4674	Serine/threonine-protein kinase
Gprk5	4674 / 4713	4674	Serine/threonine-protein kinase
Gprk6	4713	4674	Serine/threonine-protein kinase
Grk1	4713	4674	Serine/threonine-protein kinase
Gsg2	4674	4674	Serine/threonine-protein kinase
Gsk3b	4674	4674	Belongs to the Ser/Thr protein kinase family
Hck	4674	4713	Tyrosine-protein kinase HCK

- Supplementary Table 2 continued -

Hipk2	4674	4674	serine/threonine-protein kinase
Hipk3	4713	4674	Belongs to the Ser/Thr protein kinase family
Hunk	4713	4674	Serine/threonine-protein kinase MAK-V
Ick	4674	4674	Serine/threonine-protein kinase ICK
Igf1r	4674	4713	Belongs to the Tyr protein kinase family
Ikbkb	4713	4674	Serine/threonine-protein kinase
Ikbke	4674 / 4713	4674	Serine/threonine-protein kinase
Ilk	4674	4674	Serine/threonine-protein kinase
Insrr	4674	4713	Tyrosine-protein kinase
Irak3	4674 / 4713	4674	Serine/threonine-protein kinase
Itk	4674	4713	Tyrosine-protein kinase ITK/TSK
Jak1	4713	4713	Tyrosine-protein kinase
Jak2	4674 / 4713	4713	Tyrosine-protein kinase JAK2
Jak3	4713	4713	Tyrosine-protein kinase
Kdr	4674	4713	Has a tyrosine-protein kinase activity
Kit	4674 / 4713	4713	Tyrosine-protein kinase
Ksr1	4713	4674 / 4713	Ser_thr_pkin;Tyr_pkinase
Lats1	4713	4674	Serine/threonine-protein kinase LATS1
Lck	4674	4713	Proto-oncogene tyrosine-protein kinase LCK
Limk1	4713	4674	Serine/threonine-protein kinase
Lrrk1	4674 / 4713	4674	Leucine-rich repeat serine/threonine-protein kinase 1
Ltk	4674	4713	Leukocyte tyrosine kinase receptor precursor
Lyn	4713	4713	Tyrosine-protein kinase Lyn
Map2k3	4713	4674	serine/threonine-protein kinase
Map2k5	4713	4674	serine/threonine-protein kinase
Map3k12	4713	4674	serine/threonine-protein kinase
Map3k14	4713	4674	serine/threonine-protein kinase
Map3k3	4713	4674	serine/threonine-protein kinase
Map3k4	4713	4674	serine/threonine-protein kinase
Map3k7	4713	4674	serine/threonine-protein kinase
Map3k8	4713	4674	serine/threonine-protein kinase
Map4k1	4674 / 4713	4674	serine/threonine-protein kinase
Map4k2	4713	4674	serine/threonine-protein kinase
Mapk1	4674 / 4713	4674	serine/threonine-protein kinase
Mapk10	4713	4674	serine/threonine-protein kinase
Mapk11	4713	4674	serine/threonine-protein kinase
Mapk12	4713	4674	serine/threonine-protein kinase
Mapk13	4713	4674	serine/threonine-protein kinase
Mapk14	4713	4674	serine/threonine-protein kinase
Mapk3	4713	4674	serine/threonine-protein kinase
Mapk7	4713	4674	serine/threonine-protein kinase
Mapk8	4713	4674	serine/threonine-protein kinase
Mapk9	4713	4674	Serine/threonine-protein kinase
Mapkapk2	4713	4674	Belongs to the Ser/Thr protein kinase family
Mapkapk5	4713	4674	Belongs to the Ser/Thr protein kinase family
Mark1	4674 / 4713	4674	Serine/threonine-protein kinase
Mark2	4713	4674	Serine/threonine-protein kinase MARK2
Mast1	4713	4674	Microtubule-associated serine/threonine-protein kinase 1
Mast2	4674 / 4713	4674	serine/threonine-protein kinase

- Supplementary Table 2 continued -

Mastl	4674	4674	serine/threonine-protein kinase
Matk	4674	4713	Megakaryocyte-associated tyrosine-protein kinase
Melk	4674 / 4713	4674	Serine/threonine-protein kinase
Mertk	4674	4713	Proto-oncogene tyrosine-protein kinase MER precursor
Met	4674 / 4713	4713	Tyrosine-protein kinase
Mknk1	4713	4674	MAP kinase-interacting serine/threonine-protein kinase 2
Mos	4713	4674	Proto-oncogene serine/threonine-protein kinase mos
Musk	4713	4713	Muscle, skeletal receptor tyrosine protein kinase precursor
Mylk2	4674	4674	Serine/threonine-protein kinase
Nek11	4674 / 4713	4674	Serine/threonine-protein kinase
Nek2	4713	4674	Serine/threonine-protein kinase Nek2
Nek4	4713	4674	Serine/threonine-protein kinase Nek3
Nek6	4674 / 4713	4674	Serine/threonine-protein kinase
Nek7	4674	4674	Serine/threonine-protein kinase
Nlk	4674 / 4713	4674	Serine/threonine kinase NLK
Npr1	4674 / 4713	4674	Serine/threonine-protein kinase
Oxsr1	4674 / 4713	4674	Serine/threonine-protein kinase
Pak1	4674 / 4713	4674	Serine/threonine-protein kinase
Pak2	4674	4674	Serine/threonine-protein kinase
Pak3	4713	4674	Serine/threonine-protein kinase PAK 3
Pak4	4674 / 4713	4674	Serine/threonine-protein kinase
Pak7	4674 / 4713	4674	Serine/threonine-protein kinase
Pask	4674 / 4713	4674	serine/threonine-protein kinase
Pbk	4674 / 4713	4674	Serine/threonine-protein kinase
Pctk1	4713	4674	Serine/threonine-protein kinase PCTAIRE-1
Pctk3	4713	4674	Serine/threonine-protein kinase PCTAIRE-3
Pdgfra	4674	4713	Tyrosine-protein kinase
Pdgfrb	4674	4713	Tyrosine-protein kinase
Pdpk1	4674	4674 / 4713	Phosphorylated on tyrosine and serine/threonine
Pftk1	4674 / 4713	4674	Serine/threonine-protein kinase
Phkg1	4713	4674	Serine/threonine-protein kinase
Pim1	4713	4674	Proto-oncogene serine/threonine-protein kinase Pim-1
Pim2	4674 / 4713	4674	Serine/threonine-protein kinase Pim-2
Pink1	4713	4674	Serine/threonine-protein kinase PINK1, mitochondrial precursor
Pkmyt1	4713	4674 / 4713	Membrane-associated tyrosine- and threonine-specific cdc2-inhibitory kinase
Pkn2	4674 / 4713	4674	Serine/threonine-protein kinase N2
Plk1	4713	4674	Serine/threonine-protein kinase PLK1
Plk2	4713	4674	Serine/threonine-protein kinase PLK2
Plk4	4713	4674	Serine/threonine-protein kinase PLK4
Pnck	4713	4674	Belongs to the Ser/Thr protein kinase family
Prkaca	4674 / 4713	4674	Belongs to the Ser/Thr protein kinase family
Prkca	4674 / 4713	4674	Serine/threonine-protein kinase
Prkcb1	4674 / 4713	4674	Serine/threonine-protein kinase
Prkcc	4674 / 4713	4674	Serine/threonine-protein kinase
Prkch	4713	4674	Serine/threonine-protein kinase
Prkci	4713	4674	Belongs to the Ser/Thr protein kinase family
Prkcm	4713	4674	Serine/threonine-protein kinase D1
Prkcz	4713	4674	Serine/threonine-protein kinase

- Supplementary Table 2 continued -

Prkg2	4713	4674	Belongs to the Ser/Thr protein kinase family
Prkx	4674 / 4713	4674	Serine/threonine-protein kinase
Prpf4b	4713	4674	Serine/threonine-protein kinase
Ptk2	4674	4674	Serine/threonine-protein kinase PTK2/STK2
Ptk6	4674	4713	Tyrosine-protein kinase 6
Pxk	4674	4674	Ser_thr_pkinase
Ret	4674 / 4713	4713	Proto-oncogene tyrosine-protein kinase
Ripk1	4713	4674	Receptor-interacting serine/threonine-protein kinase 3
Ripk5	4674 / 4713	4674	Receptor-interacting serine/threonine-protein kinase 5
Rock1	4713	4674	Serine/threonine-protein kinase
Ror1	4674 / 4713	4713	Tyrosine-protein kinase transmembrane receptor
Ror2	4674 / 4713	4713	Tyrosine-protein kinase transmembrane receptor
Rps6ka1	4713	4674	Serine/threonine-protein kinase
Rps6ka3	4674	4674	Serine/threonine kinase
Rps6ka5	4674 / 4713	4674	Serine/threonine kinase
Rps6kb2	4713	4674	Serine/threonine-protein kinase
Rps6kl1	4674	4674	Belongs to the Ser/Thr protein kinase family
Sbk1	4674 / 4713	4674	Serine/threonine-protein kinase SBK1
Sgk2	4713	4674	Serine/threonine-protein kinase Sgk2
Sgk3	4674	4674	Serine/threonine-protein kinase Sgk3
Slk	4713	4674	STE20-like serine/threonine-protein kinase
Snf1lk2	4674	4674	Serine/threonine-protein kinase SNF1-like kinase 2
Snrk	4674 / 4713	4674	SNF-related serine/threonine-protein kinase
Src	4674 / 4713	4713	tyrosine-protein kinase
Srpk1	4713	4674	Serine/threonine-protein kinase
Srpk2	4713	4674	Serine/threonine-protein kinase
Stk10	4674 / 4713	4674	Serine/threonine-protein
Stk16	4713	4674	Serine/threonine-protein kinase 16
Stk17b	4674 / 4713	4674	Serine/threonine-protein
Stk23	4674	4674	Serine/threonine-protein
Stk32b	4713	4674	Serine/threonine-protein kinase 32B
Stk36	4674 / 4713	4674	Serine/threonine-protein kinase 36
Stk38l	4674	4674	Serine/threonine-protein
Syk	4674 / 4713	4713	Tyrosine-protein kinase SYK
Tbk1	4674 / 4713	4674	Serine/threonine-protein kinase TBK1
Tec	4674	4713	Tyrosine-protein kinase Tec
Tek	4674	4713	TEK receptor tyrosine kinase genes
Tgfbr1	4713	4674	Serine/threonine-protein kinase
Tgfbr2	4713	4674	Serine/threonine-protein kinase
Tie1	4674	4713	Tyrosine-protein kinase receptor Tie-1 precursor
Tlk1	4674 / 4713	4674	Serine/threonine-protein kinase tousled-like 1
Tlk2	4713	4674	Serine/threonine-protein kinase tousled-like 2
Tnk1	4674	4713	Non-receptor tyrosine-protein kinase TNK1
Tnk2	4674 / 4713	4713	Tyrosine kinase non-receptor protein 2
Tssk1	4713	4674	Testis-specific serine/threonine-protein kinase
Tssk2	4713	4674	Testis-specific serine/threonine-protein kinase
Tssk6	4674 / 4713	4674	Testis-specific serine/threonine-protein kinase 6
Ttbk2	4674	4674	Serine/threonine kinase
Txk	4674	4713	Tyrosine-protein kinase TXK

- Supplementary Table 2 continued -

Tyk2	4674	4713	Non-receptor tyrosine-protein kinase TYK2
Tyro3	4674	4713	Tyrosine-protein kinase receptor TYRO3 precursor
Vrk1	4674	4674	Serine/threonine-protein kinase
Vrk2	4674	4674	Serine/threonine-protein kinase
Vrk3	4674	4674	Serine/threonine-protein kinase
Yes1	4674	4713	Proto-oncogene tyrosine-protein kinase Yes
Zap70	4713	4713	Tyrosine-protein kinase ZAP-70

Legend for Supplementary Table 2:

AmiGO annotations versus UniProt annotations (with UniProt Evidence)

This table displays the AmiGO annotation and UniProt annotations for each of the 244 mouse protein kinases used in this study. The *Mouse Gene ID* numbers were obtained from each of the AmiGO protein records. The *AmiGO Label* field is "4713" (Tyr) if a query in AmiGO for the GO label GO0004713 returns the corresponding protein for mouse proteins, "4674" (Ser/Thr) if a query in AmiGO for the GO label GO0004674 returns the corresponding protein for mouse proteins, or "4674 / 4713" if a query in AmiGO for both GO labels GO0004674 and GO0004713 returns the corresponding protein. The *UniProt Label* field is " 4713" if a search in UniProt with the AmiGO Gene ID returns a mouse protein that contains a reference to the functional class protein-tyrosine kinase activity, "4674" if a search in UniProt with the AmiGO Gene ID returns a mouse protein that contains a reference to the functional class serine/threonine kinase activity, or "4674 / 4713" if a search in UniProt returns a mouse protein that contains a reference to the functional class serine/threonine kinase activity and protein-tyrosine kinase activity or any evidence that would suggest dual specificity. The *UniProt Evidence* field contains at least one example of the evidence found in the UniProt record (within *protein name, synonyms, references, similarity, keywords*, or *function)* to support the label found in the *UniProt Label* field.

APPENDIX E.

Supplementary Table 3:

Comparison of AmiGO labels, UniProt labels, and Predicted Labels for each mouse kinase protein (See Table legend below)

Gene ID	AmiGO label	UniProt label	Prediction of Classifier #1	Prediction of Classifier #2	Prediction of Classifier #3
2610018G03Rik	4713	4674	4674	4674	4674
Acvr1b	4713	4674	4674	4674	4674
Acvr2a	4713	4674	4674	4674	4674
Acvr2b	4713	4674	4674	4674	4674
Acvrl1	4713	4674	4674	4674	4674
Adrbk1	4713	4674	4674	4674	4674
Akt1	4713	4674	4674	4674	4674
Alk	4674	4713	4713	4713	4713
Araf	4713	4674	4674	4674	4674
Atm	4674	4674	4674	4674	4674
Aurka	4713	4674	4674	4674	4674
Aurkb	4713	4674	4674	4674	4674
Axl	4674	4713	4713	4713	4713
Blk	4674	4713	4713	4713	4713
Bmpr1a	4713	4674	4674	4674	4674
Bmpr1b	4713	4674	4674	4674	4674
Bmpr2	4713	4674	4674	4674	4674
Bmx	4674	4713	4713	4713	4713
Btk	4674 / 4713	4713	4713	4713	4713
Camk1	4713	4674	4674	4674	4674
Camk1d	4674	4674	4674	4674	4674
Camk1g	4674	4674	4674	4674	4674
Camk2a	4674	4674	4674	4674	4674
Camk2b	4674 / 4713	4674	4674	4674	4674
Camk2g	4674 / 4713	4674	4674	4674	4674
Camk4	4674	4674	4674	4674	4674
Camkk1	4674 / 4713	4674	4674	4674	4674
Ccrk	4674 / 4713	4674	4674	4674	4674
Cdc2a	4713	4674	4674	4674	4674
Cdc2l5	4674 / 4713	4674	4674	4674	4674
Cdk5	4674	4674	4674	4674	4674
Cdk7	4674 / 4713	4674 / 4713	4674	4674	4674
Cdk9	4713	4674 / 4713	4674	4674	4674
Cdkl1	4674 / 4713	4674	4674	4674	4674
Cdkl3	4674 / 4713	4674	4674	4674	4674
Cdkl4	4674 / 4713	4674	4674	4674	4674
Chek1	4713	4674	4674	4674	4674
Chek2	4713	4674	4674	4674	4674
Chuk	4713	4674	4674	4674	4674
Cit	4674	4674 / 4713	4674	4674	4674
Clk1	4674 / 4713	4674 / 4713	4674	4713	4674/4713

- Supplementary Table 3 continued -

Clk2	4713	4674 / 4713	4674	4674	4674
Clk3	4713	4674 / 4713	4674	4674	4674
Clk4	4713	4674 / 4713	4674	4713	4674/4713
Cpne3	4674	4674	4674	4674	4674
Csf1r	4674	4713	4713	4713	4713
Csk	4674	4713	4713	4713	4713
Csnk1d	4713	4674	4674	4674	4674
Csnk1e	4713	4674	4674	4674	4674
Csnk1g2	4713	4674	4674	4674	4674
Csnk2a2	4674 / 4713	4674	4674	4674	4674
Dapk2	4713	4674	4674	4674	4674
Dapk3	4713	4674	4674	4674	4674
Dcamkl2	4674 / 4713	4674	4674	4674	4674
Ddr1	4674	4713	4713	4713	4713
Dmpk	4713	4674	4674	4674	4674
Dyrk1a	4713	4674 / 4713	4674	4713	4674/4713
Egfr	4713	4713	4674	4713	4674/4713
Eif2ak1	4713	4674	4674	4674	4674
Eif2ak3	4713	4674	4674	4674	4674
Eif2ak4	4674 / 4713	4674	4674	4674	4674
Epha1	4674 / 4713	4713	4713	4713	4713
Epha2	4674	4713	4713	4713	4713
Epha3	4674 / 4713	4713	4713	4713	4713
Epha4	4674	4713	4713	4713	4713
Epha5	4674	4713	4713	4713	4713
Epha6	4674	4713	4713	4713	4713
Epha7	4674	4713	4713	4713	4713
Epha8	4674	4713	4713	4713	4713
Ephb2	4674 / 4713	4713	4713	4713	4713
Ephb3	4674 / 4713	4713	4713	4713	4713
Ephb4	4674	4713	4713	4713	4713
Ephb6	4674	4713	4713	4713	4713
Erbb2	4674 / 4713	4713	4713	4713	4713
Ern2	4713	4674	4674	4674	4674
Fgfr1	4674 / 4713	4713	4713	4713	4713
Fgfr2	4674	4713	4713	4713	4713
Fgfr3	4713	4713	4713	4713	4713
Fgfr4	4674	4713	4713	4713	4713
Fgr	4674	4713	4713	4713	4713
Flt1	4674	4713	4713	4713	4713
Flt3	4674	4713	4713	4713	4713
Flt4	4674	4713	4713	4713	4713
Fyn	4713	4713	4713	4713	4713
Gprk2l	4674 / 4713	4674	4674	4674	4674
Gprk5	4674 / 4713	4674	4674	4674	4674
Gprk6	4713	4674	4674	4674	4674
Grk1	4713	4674	4674	4674	4674
Gsg2	4674	4674	4674	4674	4674

- Supplementary Table 3 continued -

Gsk3b	4674	4674	4674	4674	4674
Hck	4674	4713	4713	4713	4713
Hipk2	4674	4674	4674	4674	4674
Hipk3	4713	4674	4674	4674	4674
Hunk	4713	4674	4674	4674	4674
Ick	4674	4674	4674	4674	4674
Igf1r	4674	4713	4713	4713	4713
Ikbkb	4713	4674	4674	4674	4674
Ikbke	4674 / 4713	4674	4674	4674	4674
Ilk	4674	4674	4674	4674	4674
Insrr	4674	4713	4713	4713	4713
Irak3	4674 / 4713	4674	4674	4674	4674
Itk	4674	4713	4713	4713	4713
Jak1	4713	4713	4713	4713	4713
Jak2	4674 / 4713	4713	4713	4713	4713
Jak3	4713	4713	4713	4713	4713
Kdr	4674	4713	4713	4713	4713
Kit	4674 / 4713	4713	4713	4713	4713
Ksr1	4713	4674 / 4713	4674	4674	4674
Lats1	4713	4674	4674	4674	4674
Lck	4674	4713	4713	4713	4713
Limk1	4713	4674	4674	4674	4674
Lrrk1	4674 / 4713	4674	4674	4674	4674
Ltk	4674	4713	4713	4713	4713
Lyn	4713	4713	4713	4713	4713
Map2k3	4713	4674	4674	4674	4674
Map2k5	4713	4674	4674	4674	4674
Map3k12	4713	4674	4674	4674	4674
Map3k14	4713	4674	4674	4674	4674
Map3k3	4713	4674	4674	4674	4674
Map3k4	4713	4674	4674	4674	4674
Map3k7	4713	4674	4674	4674	4674
Map3k8	4713	4674	4674	4674	4674
Map4k1	4674 / 4713	4674	4674	4674	4674
Map4k2	4713	4674	4674	4674	4674
Mapk1	4674 / 4713	4674	4674	4674	4674
Mapk10	4713	4674	4674	4674	4674
Mapk11	4713	4674	4674	4674	4674
Mapk12	4713	4674	4674	4674	4674
Mapk13	4713	4674	4674	4674	4674
Mapk14	4713	4674	4674	4674	4674
Mapk3	4713	4674	4674	4674	4674
Mapk7	4713	4674	4674	4674	4674
Mapk8	4713	4674	4674	4674	4674
Mapk9	4713	4674	4674	4674	4674
Mapkapk2	4713	4674	4674	4674	4674
Mapkapk5	4713	4674	4674	4674	4674
Mark1	4674 / 4713	4674	4674	4674	4674

- Supplementary Table 3 continued -

Mark2	4713	4674	4674	4674	4674
Mast1	4713	4674	4674	4674	4674
Mast2	4674 / 4713	4674	4674	4674	4674
Mastl	4674	4674	4674	4674	4674
Matk	4674	4713	4713	4713	4713
Melk	4674 / 4713	4674	4674	4674	4674
Mertk	4674	4713	4713	4713	4713
Met	4674 / 4713	4713	4713	4713	4713
Mknk1	4713	4674	4674	4674	4674
Mos	4713	4674	4674	4674	4674
Musk	4713	4713	4713	4713	4713
Mylk2	4674	4674	4674	4674	4674
Nek11	4674 / 4713	4674	4674	4674	4674
Nek2	4713	4674	4674	4674	4674
Nek4	4713	4674	4674	4674	4674
Nek6	4674 / 4713	4674	4674	4674	4674
Nek7	4674	4674	4674	4674	4674
Nlk	4674 / 4713	4674	4674	4674	4674
Npr1	4674 / 4713	4674	4713	4713	4713
Oxsr1	4674 / 4713	4674	4674	4674	4674
Pak1	4674 / 4713	4674	4674	4674	4674
Pak2	4674	4674	4674	4674	4674
Pak3	4713	4674	4674	4674	4674
Pak4	4674 / 4713	4674	4674	4674	4674
Pak7	4674 / 4713	4674	4674	4674	4674
Pask	4674 / 4713	4674	4674	4674	4674
Pbk	4674 / 4713	4674	4674	4674	4674
Pctk1	4713	4674	4674	4674	4674
Pctk3	4713	4674	4674	4674	4674
Pdgfra	4674	4713	4713	4713	4713
Pdgfrb	4674	4713	4713	4713	4713
Pdpk1	4674	4674 / 4713	4674	4674	4674
Pftk1	4674 / 4713	4674	4674	4674	4674
Phkg1	4713	4674	4674	4674	4674
Pim1	4713	4674	4674	4674	4674
Pim2	4674 / 4713	4674	4674	4674	4674
Pink1	4713	4674	4674	4674	4674
Pkmyt1	4713	4674 / 4713	4674	4674	4674
Pkn2	4674 / 4713	4674	4674	4674	4674
Plk1	4713	4674	4674	4674	4674
Plk2	4713	4674	4674	4674	4674
Plk4	4713	4674	4674	4674	4674
Pnck	4713	4674	4674	4674	4674
Prkaca	4674 / 4713	4674	4674	4674	4674
Prkca	4674 / 4713	4674	4674	4674	4674
Prkcb1	4674 / 4713	4674	4674	4674	4674
Prkcc	4674 / 4713	4674	4674	4674	4674
Prkch	4713	4674	4674	4674	4674

- Supplementary Table 3 continued -

Prkci	4713	4674	4674	4674	4674
Prkcm	4713	4674	4674	4674	4674
Prkcz	4713	4674	4674	4674	4674
Prkg2	4713	4674	4674	4674	4674
Prkx	4674 / 4713	4674	4674	4674	4674
Prpf4b	4713	4674	4674	4674	4674
Ptk2	4674	4674	4713	4713	4713
Ptk6	4674	4713	4713	4713	4713
Pxk	4674	4674	4674	4674	4674
Ret	4674 / 4713	4713	4713	4713	4713
Ripk1	4713	4674	4674	4674	4674
Ripk5	4674 / 4713	4674	4674	4674	4674
Rock1	4713	4674	4674	4674	4674
Ror1	4674 / 4713	4713	4713	4713	4713
Ror2	4674 / 4713	4713	4713	4713	4713
Rps6ka1	4713	4674	4674	4674	4674
Rps6ka3	4674	4674	4674	4674	4674
Rps6ka5	4674 / 4713	4674	4674	4674	4674
Rps6kb2	4713	4674	4674	4674	4674
Rps6kl1	4674	4674	4674	4674	4674
Sbk1	4674 / 4713	4674	4674	4674	4674
Sgk2	4713	4674	4674	4674	4674
Sgk3	4674	4674	4674	4674	4674
Slk	4713	4674	4674	4674	4674
Snf1lk2	4674	4674	4674	4674	4674
Snrk	4674 / 4713	4674	4674	4674	4674
Src	4674 / 4713	4713	4713	4713	4713
Srpk1	4713	4674	4674	4674	4674
Srpk2	4713	4674	4674	4674	4674
Stk10	4674 / 4713	4674	4674	4674	4674
Stk16	4713	4674	4674	4674	4674
Stk17b	4674 / 4713	4674	4674	4674	4674
Stk23	4674	4674	4674	4674	4674
Stk32b	4713	4674	4674	4674	4674
Stk36	4674 / 4713	4674	4674	4674	4674
Stk38l	4674	4674	4674	4674	4674
Syk	4674 / 4713	4713	4713	4713	4713
Tbk1	4674 / 4713	4674	4674	4674	4674
Tec	4674	4713	4713	4713	4713
Tek	4674	4713	4713	4713	4713
Tgfbr1	4713	4674	4674	4674	4674
Tgfbr2	4713	4674	4674	4674	4674
Tie1	4674	4713	4713	4713	4713
Tlk1	4674 / 4713	4674	4674	4674	4674
Tlk2	4713	4674	4674	4674	4674
Tnk1	4674	4713	4713	4713	4713
Tnk2	4674 / 4713	4713	4713	4713	4713
Tssk1	4713	4674	4674	4674	4674

- Supplementary Table 3 continued -

Tssk2	4713	4674	4674	4674	4674
Tssk6	4674 / 4713	4674	4674	4674	4674
Ttbk2	4674	4674	4674	4674	4674
Txk	4674	4713	4713	4713	4713
Tyk2	4674	4713	4713	4713	4713
Tyro3	4674	4713	4713	4713	4713
Vrk1	4674	4674	4674	4674	4674
Vrk2	4674	4674	4674	4674	4674
Vrk3	4674	4674	4674	4674	4674
Yes1	4674	4713	4713	4713	4713
Zap70	4713	4713	4713	4713	4713

Legend for Supplementary Table 3:

AmiGO labels, UniProt labels, and Predicted Labels for each mouse kinase protein

For the 244 mouse protein kinase used in this study, each row in the table contains the AmiGO annotation, UniProt annotation and predictions made by each of the three machine-learning classifiers tested. These predictions are based on using a classifier built on 330 human proteins with protein gene ontology functional labels from AmiGO and verified by UniProt. Because the training data are fixed, note that the predictions will be the same regardless of what labels we use on the test set for evaluation purposes.

The ***Mouse Gene ID*** was obtained from each of the AmiGO protein records. The *AmiGO Label* field is "4713" (Tyr) if a query in AmiGO for the GO label GO0004713 returns the corresponding protein for mouse proteins, "4674" (Ser/Thr) if a query in AmiGO for the GO label GO0004674 returns the corresponding protein for mouse proteins, or "4674 / 4713" if a query in AmiGO for both GO labels GO0004674 and GO0004713 returns the corresponding protein. The ***UniProt Label*** field is " 4713" if a search in UniProt with the AmiGO Gene ID returns a mouse protein that contains a reference to the functional class protein-tyrosine kinase activity, "4674" if a search in UniProt with the AmiGO Gene ID returns a mouse protein that contains a reference to the functional class serine/threonine kinase activity, or "4674 / 4713" if a search in UniProt returns a mouse protein that contains a reference to the functional class serine/threonine kinase activity and protein-tyrosine kinase activity or any evidence that would suggest dual specificity. The ***Prediction of classifier #1*** field contains the prediction of the first HDTree classifier that distinguishes between GO0004674 and not GO0004674 (GO0004713 and Dual). This classifier was built on human proteins. The ***Prediction of classifier #2*** field contains the prediction of the second HDTree classifier that distinguishes between GO0004713 and not GO0004713 (GO0004674 and Dual). This classifier was built on human proteins. The ***Prediction of classifier #3*** field contains the prediction of the third classifier that distinguishes between GO0004674, GO0004713 and Dual. The third classifier combines the outputs of the first two classifiers to distinguish between GO0004674 and GO0004713 and Dual. (See **Supplementary Data** for details)

APPENDIX F

Supplementary Table 4:

Mouse Kinases having a Human Ortholog (See Table legend below)

Gene ID (AmiGO)	Jackson Lab Symbol (Mouse Kinome)	Gene ID (Mouse Kinome)	Human Ortholog (Human Kinome)	Identity between orthologs
2610018G03Rik				
Acvr1b	Acvr1b	ALK4	ALK4	99.31%
Acvr2a				
Acvr2b	Acvr2b	ACTR2B	ACTR2B	100.00%
Acvrl1				
Adrbk1	Adrbk1	BARK1	BARK1	99.62%
Akt1	Akt1	AKT1	AKT1	99.61%
Alk	Alk	ALK	ALK	98.15%
Araf	Araf	ARAF	ARAF	98.45%
Atm	Atm	ATM	ATM	N/A
Aurka				
Aurkb				
Axl	Axl	AXL	AXL	97.40%
Blk	Blk	BLK	BLK	93.63%
Bmpr1a	Bmpr1a	BMPR1A	BMPR1A	99.65%
Bmpr1b	Bmpr1b	BMPR1B	BMPR1B	98.96%
Bmpr2	Bmpr2	BMPR2	BMPR2	98.70%
Bmx	Bmx	BMX	BMX	94.44%
Btk	Btk	BTK	BTK	98.81%
Camk1				
Camk1d	E030025C11Rik	CAMK1d	CAMK1d	100.00%
Camk1g	Camk1g	CAMK1g	CAMK1g	98.43%
Camk2a	Camk2a	CaMK2a	CaMK2a	100.00%
Camk2b	Camk2b	CaMK2b	CaMK2b	100.00%
Camk2g	Camk2g	CaMK2g	CaMK2g	100.00%
Camk4	Camk4	CaMK4	CaMK4	99.61%
Camkk1	Camkk1	CAMKK1	CAMKK1	93.45%
Ccrk	4932702G04Rik	CCRK	CCRK	94.04%
Cdc2a	Cdc2a	CDC2	CDC2	97.18%
Cdc2l5				
Cdk5	Cdk5	CDK5	CDK5	99.65%
Cdk7	Cdk7	CDK7	CDK7	97.18%
Cdk9	Cdk9	CDK9	CDK9	98.32%
Cdkl1	Cdkl1	CDKL1	CDKL1	95.42%
Cdkl3	Cdkl3	CDKL3	CDKL3	93.29%

- Supplementary Table 4 continued -

Cdkl4	AU067824	CDKL4	CDKL4	91.87%
Chek1	Chek1	CHK1	CHK1	96.89%
Chek2	Chek2	CHK2	CHK2	92.13%
Chuk	Chuk	IKKa	IKKa	97.08%
Cit				
Clk1	Clk	CLK1	CLK1	94.32%
Clk2	Clk2	CLK2	CLK2	98.42%
Clk3		CLK3	CLK3	100.00%
Clk4	Clk4	CLK4	CLK4	98.73%
Cpne3				
Csf1r				
Csk	Csk	CSK	CSK	99.19%
Csnk1d	Csnk1d	CK1d	CK1d	100.00%
Csnk1e	Csnk1e	CK1e	CK1e	100.00%
Csnk1g2				
Csnk2a2	Csnk2a2	CK2a2	CK2a2	99.30%
Dapk2	Dapk2	DAPK2	DAPK2	98.86%
Dapk3	Dapk3	DAPK3	DAPK3	95.44%
Dcamkl2	6330415M09Rik	DCAMKL2	DCAMKL2	97.29%
Ddr1	Ddr1	DDR1	DDR1	92.91%
Dmpk	Dm15	DMPK1	DMPK1	93.31%
Dyrk1a	Dyrk1a	DYRK1A	DYRK1A	99.69%
Egfr	Egfr	EGFR	EGFR	98.84%
Eif2ak1				
Eif2ak3				
Eif2ak4				
Epha1	Epha1	EphA1	EphA1	93.85%
Epha2	Epha2	EphA2	EphA2	96.12%
Epha3	Mark3	EphA3	EphA3	100.00%
Epha4	Epha4	EphA4	EphA4	100.00%
Epha5	Epha5	EphA5	EphA5	98.83%
Epha6	Epha6	EPHA6	EPHA6	98.67%
Epha7	Epha7	EphA7	EphA7	99.61%
Epha8	Epha8	EphA8	EphA8	93.41%
Ephb2	Ephb2	EphB2	EphB2	100.00%
Ephb3	Ephb3	EphB3	EphB3	99.62%
Ephb4	Ephb4	EphB4	EphB4	99.65%
Ephb6	Ephb6	EphB6	EphB6	93.90%
Erbb2	Erbb2	ErbB2	ErbB2	98.45%
Ern2				
Fgfr1	Fgfr1	FGFR1	FGFR1	100.00%
Fgfr2	Fgfr2	FGFR2	FGFR2	99.64%

- Supplementary Table 4 continued -

Fgfr3	Fgfr4	FGFR4	FGFR4	97.47%
Fgfr4	Fgfr4	FGFR4	FGFR4	97.47%
Fgr	Fgr	FGR	FGR	92.00%
Flt1	Flt1	FLT1	FLT1	91.46%
Flt3	Flt3	FLT3	FLT3	93.11%
Flt4	Flt4	FLT4	FLT4	91.69%
Fyn	Fyn	FYN	FYN	99.20%
Gprk2l	Gprk2l	GPRK4	GPRK4	85.55%
Gprk5	Gprk5	GPRK5	GPRK5	98.86%
Gprk6	Gprk6	GPRK6	GPRK6	97.34%
Grk1				
Gsg2				
Gsk3b	Gsk3b	GSK3B	GSK3B	100.00%
Hck	Hck	HCK	HCK	94.42%
Hipk2	Hipk2	HIPK2	HIPK2	99.39%
Hipk3	Hipk3	HIPK3	HIPK3	97.87%
Hunk	Hunk	HUNK	HUNK	98.08%
Ick	Ick	ICK	ICK	97.15%
Igf1r	Igf1r	IGF1R	IGF1R	97.41%
Ikbkb	Ikbkb	IKKb	IKKb	96.73%
Ikbke	Ikbke	IKKe	IKKe	91.51%
Ilk	Taf10	ILK	ILK	99.22%
Insrr	Insrr	IRR	IRR	93.70%
Irak3	Irak3	IRAK3	IRAK3	83.15%
Itk	Itk	ITK	ITK	96.80%
Jak1	Jak1	JAK1	JAK1	97.83%
Jak2	Jak2	JAK2	JAK2	97.12%
Jak3	Jak3	JAK3	JAK3	82.72%
Kdr	Kdr	KDR	KDR	96.33%
Kit	Kit	KIT	KIT	94.94%
Ksr1	Ksr	KSR1	KSR1	95.83%
Lats1	Lats1	LATS1	LATS1	98.69%
Lck	Lck	LCK	LCK	97.21%
Limk1	Limk1	LIMK1	LIMK1	97.37%
Lrrk1	D130026O16Rik	LRRK1	LRRK1	94.37%
Ltk	Ltk	LTK	LTK	88.89%
Lyn	Lyn	LYN	LYN	97.62%
Map2k3	Map2k3	MAP2K3	MAP2K3	98.09%
Map2k5	Map2k5	MAP2K5	MAP2K5	98.36%
Map3k12				
Map3k14				
Map3k3	Map3k3	MAP3K3	MAP3K3	99.23%

- Supplementary Table 4 continued -

Map3k4	Map3k4	MAP3K4	MAP3K4	98.07%
Map3k7		MAP3K7	MAP3K7	96.47%
Map3k8		MAP3K8	MAP3K8	94.70%
Map4k1				
Map4k2				
Mapk1				
Mapk10				
Mapk11				
Mapk12				
Mapk13				
Mapk14				
Mapk3				
Mapk7				
Mapk8				
Mapk9				
Mapkapk2	Mapkapk2	MAPKAPK2	MAPKAPK2	98.47%
Mapkapk5	Mapkapk5	MAPKAPK5	MAPKAPK5	98.59%
Mark1	B930025N23Rik	MARK1	MARK1	99.21%
Mark2	Mark2	MARK2	MARK2	100.00%
Mast1	Mast1	MAST1	MAST1	99.64%
Mast2	Mtssk	MAST2	MAST2	98.91%
Mastl	2700091H24Rik	MASTL	MASTL	87.32%
Matk				
Melk	Melk	MELK	MELK	95.26%
Mertk	Mertk	MER	MER	93.61%
Met	Met	MET	MET	97.71%
Mknk1				
Mos		MOS	MOS	75.09%
Musk	Musk	MUSK	MUSK	97.15%
Mylk2				
Nek11				
Nek2	Nek2	NEK2	NEK2	93.56%
Nek4	Nek4	NEK4	NEK4	96.48%
Nek6	Nek6	NEK6	NEK6	98.41%
Nek7	Nek7	NEK7	NEK7	98.41%
Nlk	Nlk	NLK	NLK	100.00%
Npr1				
Oxsr1				
Pak1	Pak1	PAK1	PAK1	99.60%
Pak2	Pak2	PAK2	PAK2	99.60%
Pak3	Pak3	PAK3	PAK3	98.81%
Pak4	Pak4	PAK4	PAK4	98.41%
Pak7				
Pask	Pask	PASK	PASK	88.93%

- Supplementary Table 4 continued -

Pbk	Topk-pending	PBK	PBK	89.55%
Pctk1				
Pctk3				
Pdgfra	Pdgfra	PDGFRa	PDGFRa	98.32%
Pdgfrb	Pdgfrb	PDGFRb	PDGFRb	94.71%
Pdpk1				
Pftk1				
Phkg1	Phkg	PHKg1	PHKg1	93.31%
Pim1	Pim1	PIM1	PIM1	95.26%
Pim2	Pim2	PIM2	PIM2	92.16%
Pink1	1190006F07Rik	PINK1	PINK1	81.25%
Pkmyt1				
Pkn2				
Plk1	Plk	PLK1	PLK1	96.84%
Plk2	Snk	PLK2	PLK2	99.60%
Plk4	Stk18	PLK4	PLK4	95.67%
Pnck				
Prkaca				
Prkca	Prkca	PKCa	PKCa	100.00%
Prkcb1				
Prkcc				
Prkch				
Prkci				
Prkcm				
Prkcz				
Prkg2				
Prkx				
Prpf4b				
Ptk2				
Ptk6				
Pxk				
Ret	Ret	RET	RET	95.42%
Ripk1	Ripk1	RIPK1	RIPK1	76.58%
Ripk5				
Rock1	Rock1	ROCK1	ROCK1	100.00%
Ror1	Ror1	ROR1	ROR1	98.70%
Ror2	Ror2	ROR2	ROR2	96.09%
Rps6ka1				
Rps6ka3				
Rps6ka5				
Rps6kb2				
Rps6kl1				
Sbk1				
Sgk2	Sgk2	SGK2	SGK2	94.57%
Sgk3	Sgk3	SGK3	SGK3	98.45%

- Supplementary Table 4 continued -

Slk	Stk2	SLK	SLK	98.46%
Snf1lk2				
Snrk	Snrk	SNRK	SNRK	95.67%
Src	Src	SRC	SRC	99.20%
Srpk1	Srpk1	SRPK1	SRPK1	90.59%
Srpk2	Srpk2	SRPK2	SRPK2	93.73%
Stk10				
Stk16				
Stk17b				
Stk23				
Stk32b				
Stk36				
Stk38l				
Syk	Syk	SYK	SYK	99.20%
Tbk1	Tbk1	TBK1	TBK1	97.97%
Tec	Tec	TEC	TEC	96.00%
Tek	Tek	TIE2	TIE2	99.63%
Tgfbr1	Tgfbr1	TGFBR1	TGFBR1	100.00%
Tgfbr2	Tgfbr2	TGFBR2	TGFBR2	98.31%
Tie1	Tie1	TIE1	TIE1	99.63%
Tlk1	Tlk1	TLK1	TLK1	100.00%
Tlk2	Tlk2	TLK2	TLK2	99.64%
Tnk1	Tnk1	TNK1	TNK1	93.51%
Tnk2				
Tssk1	Stk22a	TSSK1	TSSK1	92.34%
Tssk2	Stk22b	TSSK2	TSSK2	94.64%
Tssk6				
Ttbk2	B930008N24Rik	TTBK2	TTBK2	98.44%
Txk	Txk	TXK	TXK	87.70%
Tyk2	Tyk2	TYK2	TYK2	91.73%
Tyro3	Tyro3	TYRO3	TYRO3	98.85%
Vrk1	Vrk1	VRK1	VRK1	93.63%
Vrk2	Vrk2	VRK2	VRK2	84.85%
Vrk3	AI428238	VRK3	VRK3	81.25%
Yes1				
Zap70	Zap70	ZAP70	ZAP70	95.60%

Legend for Supplementary Table 4:

Mouse Kinases having a Human Ortholog

All 244 mouse protein kinases used in this study were compared to the mouse proteins found in the Mouse Kinome. If a match in the Kinome database [1-3] was found the corresponding human ortholog and its sequence identity to the mouse protein are displayed in the table (167 proteins in total). If a match was not found, the row in the table is left blank (77 proteins in total). To match up the AmiGO proteins with the proteins found in Mouse Kinome [1] we used the ***Mouse Gene ID*** obtained from the AmiGO record and the ***Jackson Lab Symbol*** found in the second table of the Mouse Kinome [1-3]. The Mouse Kinome database directly provided the ***Human Ortholog ID*** and the percent sequence identity between the Mouse and Human orthologs (***Percent Identity between Orthologs*** field above). We did not compute this identity directly for this study. A brief summary of this table can be found in **Supplementary Table 5**.

References for Supplementary Table 4:

1. Caenepeel, S., Charydczak, G., Sudarsanam, S., Hunter, T. & Manning, G. *PNAS* **101**, 11707-11712. (2004). http://kinase.com/mouse/

2. Manning, G., Whyte, D.B., Martinez, R., Hunter, T., & Sudarsanam, S. *Science* **298**, 1912-1934 (2002). http://kinase.com/human/kinome/

3. http://kinase.com/mouse/tables/Table2.xls

APPENDIX G.

Supplementary Table 5:

Sequence Identity cutoff with Human Ortholog	Number of mouse proteins	Percent of 167 mouse kinases with Human Ortholog	Percent of all 244 mouse kinases
100%	19	11.4%	7.8%
99%	46	27.5%	18.9%
95%	118	70.7%	48.4%
90%	154	92.2%	63.1%
85%	160	95.8%	65.6%
80%	165	98.8%	67.6%
75%	167	100%	68.4%

Legend for Supplementary Table 5:

Number of Mouse kinases having a specified level of sequence identity with their human orthologs. (Summary statistics for Supplementary Table 4).

We compared the sequence identities between each of the 244 mouse protein kinases used in this study with their human orthologs found in the Mouse Kinome (See **Supplementary Table 4)**. This table summarizes the number of proteins that had sequence identities greater than a fixed cutoff value.

APPENDIX H

Supplementary Table 6:

The UniProt and AmiGO annotations for the Rat kinase proteins with Mouse orthologs (See Table legend below)

Mouse / Rat Gene ID	Mouse AmiGO label	Rat AmiGO label by ISS Annotation	UniProt label for Mouse
Acvr2a	4713	4713	4674
Acvr2b	4713	4713	4674
Acvrl1	4713	4713	4674
Adrbk1	4713	4713	4674
Akt1	4713	4713	4674
Alk	4674	4674	4713
Araf	4713	4713	4674
Aurkb	4713	4713	4674
Axl	4674	4674	4713
Blk	4674	4674 / 4713	4713
Bmpr1a	4713	4713	4674
Bmpr2	4713	4713	4674
Btk	4674 / 4713	4674 / 4713	4713
Camk1	4713	4713	4674
Camk1g	4674	4674	4674
Camk2a	4674	4674	4674
Camk2b	4674 / 4713	4674 / 4713	4674
Camk2g	4674 / 4713	4674 / 4713	4674
Camkk1	4674 / 4713	4713	4674
Ccrk	4674 / 4713	4674 / 4713	4674
Cdc2a	4713	4713	4674
Cdc2l5	4674 / 4713	4713	4674
Cdk5	4674	4674	4674
Cdk7	4674 / 4713	4713	4674 / 4713
Cdk9	4713	4713	4674 / 4713
Cdkl1	4674 / 4713	4674 / 4713	4674
Cdkl3	4674 / 4713	4674 / 4713	4674
Chek1	4713	4713	4674
Chek2	4713	4713	4674
Chuk	4713	4713	4674
Clk1	4674 / 4713	4713	4674 / 4713
Clk3	4713	4713	4674 / 4713
Clk4	4713	4713	4674 / 4713
Csf1r	4674	4674	4713
Csk	4674	4674	4713
Csnk1d	4713	4713	4674
Csnk1e	4713	4713	4674
Csnk1g2	4713	4713	4674
Csnk2a2	4674 / 4713	4713	4674
Dapk2	4713	4713	4674
Dapk3	4713	4713	4674

- Supplementary Table 6 continued -

Ddr1	4674	4713	4713
Dmpk	4713	4713	4674
Dyrk1a	4713	4713	4674 / 4713
Eif2ak1	4713	4713	4674
Eif2ak3	4713	4713	4674
Eif2ak4	4674 / 4713	4674 / 4713	4674
Epha1	4674 / 4713	4674	4713
Epha2	4674	4674	4713
Epha3	4674 / 4713	4674	4713
Epha5	4674	4674	4713
Epha6	4674	4674	4713
Epha7	4674	4674	4713
Epha8	4674	4674	4713
Ephb3	4674 / 4713	4674	4713
Ephb6	4674	4674	4713
Erbb2	4674 / 4713	4674	4713
Ern2	4713	4713	4674
Fgfr1	4674 / 4713	4674	4713
Fgfr2	4674	4674	4713
Fgfr4	4674	4674	4713
Fgr	4674	4674	4713
Flt1	4674	4674	4713
Flt3	4674	4674	4713
Flt4	4674	4674	4713
Fyn	4713	4713	4713
Gprk2l	4674 / 4713	4713	4674
Gprk5	4674 / 4713	4674 / 4713	4674
Gprk6	4713	4713	4674
Grk1	4713	4713	4674
Hck	4674	4674	4713
Hipk2	4674	4674	4674
Hipk3	4713	4713	4674
Ick	4674	4674	4674
Igf1r	4674	4674	4713
Ikbkb	4713	4713	4674
Ikbke	4674 / 4713	4713	4674
Ilk	4674	4674	4674
Irak3	4674 / 4713	4674 / 4713	4674
Itk	4674	4674 / 4713	4713
Jak1	4713	4713	4713
Jak2	4674 / 4713	4674 / 4713	4713
Jak3	4713	4713	4713
Kit	4674 / 4713	4674	4713
Ksr1	4713	4713	4674 / 4713
Lck	4674	4674	4713
Limk1	4713	4713	4674
Lyn	4713	4713	4713
Map2k3	4713	4713	4674

- Supplementary Table 6 continued -

Map2k5	4713	4713	4674
Map3k12	4713	4713	4674
Map3k14	4713	4713	4674
Map3k3	4713	4713	4674
Map3k4	4713	4713	4674
Map3k7	4713	4713	4674
Map3k8	4713	4713	4674
Map4k1	4674 / 4713	BOTH	4674
Map4k2	4713	4713	4674
Mapk1	4674 / 4713	BOTH	4674
Mapk10	4713	4713	4674
Mapk11	4713	4713	4674
Mapk12	4713	4713	4674
Mapk13	4713	4713	4674
Mapk14	4713	4713	4674
Mapk3	4713	4713	4674
Mapk7	4713	4713	4674
Mapk8	4713	4713	4674
Mark1	4674 / 4713	4674 / 4713	4674
Mast2	4674 / 4713	4674 / 4713	4674
Mastl	4674	4674	4674
Matk	4674	4674	4713
Melk	4674 / 4713	4674 / 4713	4674
Met	4674 / 4713	4674	4713
Mylk2	4674	4674	4674
Nek11	4674 / 4713	4674 / 4713	4674
Nek6	4674 / 4713	4674 / 4713	4674
Nek7	4674	4674	4674
Npr1	4674 / 4713	4674	4674
Oxsr1	4674 / 4713	4674 / 4713	4674
Pak1	4674 / 4713	4674 / 4713	4674
Pak2	4674	4674	4674
Pak4	4674 / 4713	4674 / 4713	4674
Pak7	4674 / 4713	4713	4674
Pask	4674 / 4713	4674 / 4713	4674
Pbk	4674 / 4713	4674 / 4713	4674
Pdgfra	4674	4674	4713
Pdgfrb	4674	4674	4713
Pftk1	4674 / 4713	4674 / 4713	4674
Pkn2	4674 / 4713	4674 / 4713	4674
Prkca	4674 / 4713	4674 / 4713	4674
Prkcb1	4674 / 4713	4674 / 4713	4674
Prkcc	4674 / 4713	4674 / 4713	4674
Ptk2	4674	4674 / 4713	4674
Ptk6	4674	4674	4713
Pxk	4674	4674	4674
Rock1	4713	4713	4674
Ror2	4674 / 4713	4674	4713

- Supplementary Table 6 continued -

Rps6ka1	4713	4713	4674
Rps6ka5	4674 / 4713	4674 / 4713	4674
Rps6kb2	4713	4713	4674
Sgk2	4713	4713	4674
Snrk	4674 / 4713	4713	4674
Src	4674 / 4713	4713	4713
Srpk1	4713	4713	4674
Srpk2	4713	4713	4674
Stk10	4674 / 4713	4674 / 4713	4674
Stk16	4713	4713	4674
Stk17b	4674 / 4713	4713	4674
Syk	4674 / 4713	4674 / 4713	4713
Tbk1	4674 / 4713	4674 / 4713	4674
Tec	4674	4674	4713
Tek	4674	4674	4713
Tgfbr1	4713	4713	4674
Tgfbr2	4713	4713	4674
Tie1	4674	4674	4713
Tlk1	4674 / 4713	4713	4674
Tlk2	4713	4713	4674
Tnk1	4674	4674	4713
Tnk2	4674 / 4713	4674 / 4713	4713
Tssk1	4713	4713	4674
Tssk2	4713	4713	4674
Yes1	4674	4713	4713
Zap70	4713	4713	4713

Legend for Supplementary Table 6:

The UniProt and AmiGO annotations for the Rat kinase proteins with Mouse orthologs

In this study 136 rat proteins had a mouse ortholog with a "potentially incorrect" AmiGO annotation. This table displays the Gene ID for the mouse and rat (they are the same because they are orthologs), the mouse GO label provided by AmiGO, the rat GO label by ISS (inferred from sequence similarity) with the mouse protein, and the corresponding UniProt annotation.

The ***Mouse/Rat Gene ID*** was obtained from each of the AmiGO protein records. The ***Mouse AmiGO Label*** field is "4713" (Tyr) if a query in AmiGO for the GO label GO0004713 returns the corresponding protein for mouse proteins, "4674" (Ser/Thr) if a query in AmiGO for the GO label GO0004674 returns the corresponding protein for mouse proteins, or "4674 / 4713" if a query in AmiGO for both GO labels GO0004674 and GO0004713 returns the corresponding protein. The ***Rat GO label by ISS Annotation*** field is "4713" if a query in AmiGO for the GO label GO0004713 returns the corresponding protein for rat proteins with an evidence code of ISS by association, "4674" if a query in

AmiGO for the GO label GO0004674 returns the corresponding protein for rat proteins with an evidence code of ISS by association with a mouse protein or “4674 / 4713” if a query in AmiGO for the GO label GO0004674 and GO0004713 returns the corresponding protein for rat proteins with an evidence code of ISS by association with a mouse protein. The ***UniProt Label for Mouse*** field is “ 4713” if a search in UniProt with the AmiGO Gene ID returns a mouse protein that contains a reference to the functional class “protein-tyrosine kinase activity”, ”4674” if a search in UniProt with the AmiGO Gene ID returns a mouse protein that contains a reference to the functional class “serine/threonine kinase activity”, or “4674 / 4713” if a search in UniProt returns a mouse protein that contains a reference to the functional class “serine/threonine kinase activity” and “protein-tyrosine kinase activity” or any evidence that would suggest dual specificity.

APPENDIX I

Supplementary Table 7:

Species (Data Source)	0004674 Ser/Thr Kinase	0004713 Tyr Kinase	0004674 and 0004713 Dual Specific
Human (AmiGO)	233 (70.6%)	90 (27.3%)	7 (2.1%)
Human (UniProt)	233 (70.6%)	90 (27.3%)	7 (2.1%)
Human (HDTree)	230 (69.7%)	67 (20.3%)	33 (10.0%)
Mouse (AmiGO)	71 (29.1%)	106 (43.4%)	67 (27.5%)
Mouse (UniProt)	168 (68.9%)	65 (26.6%)	11 (4.5%)
Mouse (HDTree)	174 (71.3%)	66 (27.0%)	4 (1.6%)

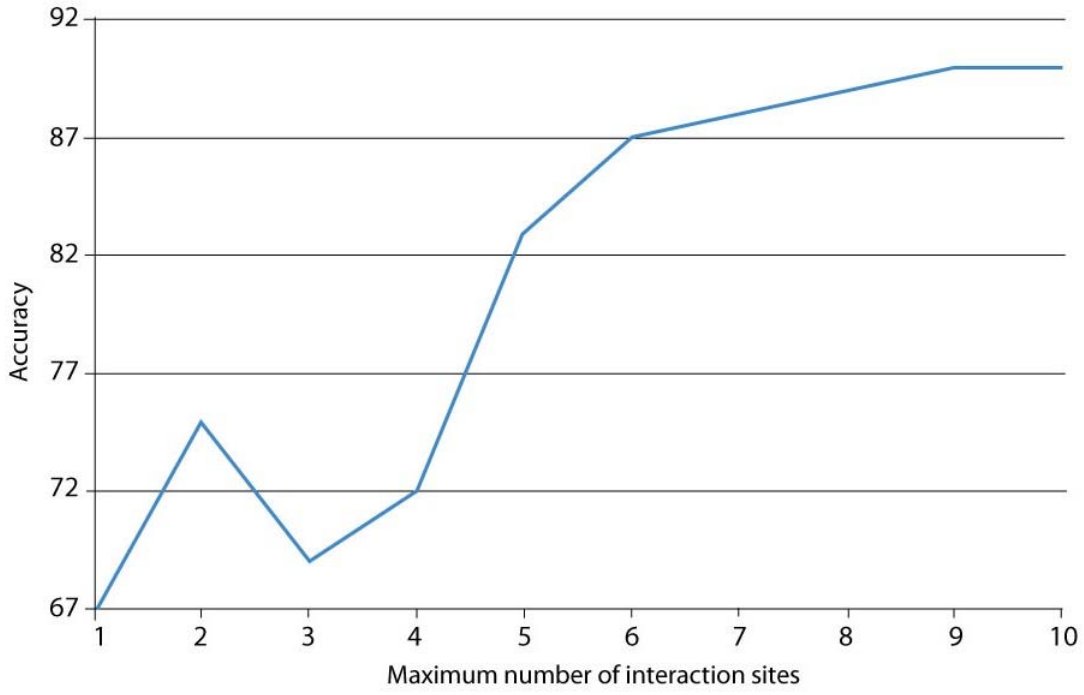
Legend for Supplementary Table 7:

Distribution of protein classes for Human and Mouse proteins annotated by AmiGO, UniProt, and HDTree.

The table shows the distributions of each class of kinases (GO0004674, GO0004713, and Dual Specificity) for proteins retrieved from AmiGO, verified by UniProt, and predicted by the HDTree method. Each entry contains the number of proteins that belongs to the given class and its percentage compared to all the kinases for the given source. The Human dataset contains 330 proteins and the Mouse dataset contains 244 proteins. A pie chart of these distributions is shown in **Figure 1.**

APPENDIX J

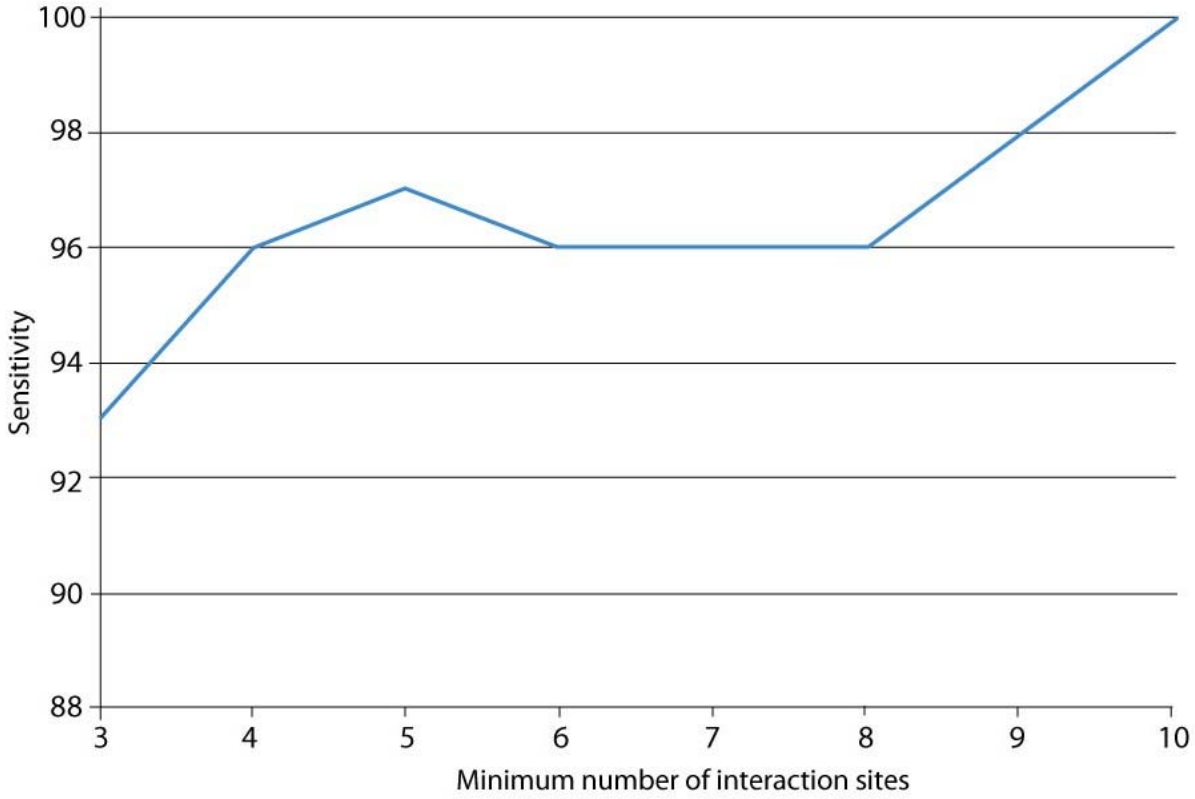
Figure S1. The accuracy curve of predicting singlish-interface and multiple-interface hub proteins as a function of the number of interaction sites. The curve shows the prediction accuracy for proteins with number of interactions sites less than the given maximum threshold. For example, the value of 5 on the x-axis refers to all hub proteins with 5 or fewer interfaces and the value on the curve (83%) at x=5, represents the accuracy of this set.



The accuracy curve of predicting singlish-interface and multiple-interface hub proteins as a function of the number of interaction sites. The curve shows the prediction accuracy for proteins with number of interactions sites less than the given maximum threshold. For example, the value of 5 on the x-axis refers to all hub proteins with 5 or fewer interfaces and the value on the curve (83%) at x=5, represents the accuracy of this set.

APPENDIX K

Figure S2. **The sensitivity curve of predicting singlish-interface and multiple-interface hub proteins as a function of the number of interaction sites.** The curve shows the prediction accuracy for proteins with number of interactions sites more than the given minimum threshold. For example, the value of 5 on the x-axis refers to all hub proteins with 5 or more interfaces and the value on the curve (97%) at x=5, represents the sensitivity of this set.



The sensitivity curve of predicting singlish-interface and multiple-interface hub proteins as a function of the number of interaction sites. The curve shows the prediction accuracy for proteins with number of interactions sites more than the given minimum threshold. For example, the value of 5 on the x-axis refers to all hub proteins with 5 or more interfaces and the value on the curve (97%) at x=5, represents the sensitivity of this set.

APPENDIX L

Table S1. Dataset 1 results on our internal machine-learning methods

Approach	k	Accuracy	Precision	Recall	C.C
NB k-gram	1	81.4	.78	.68	.59
	2	82.5	.80	.69	.61
	3	84.5	.81	.72	.65
	4	88.2	.75	.86	.72
NB(k)	2	83.9	.80	.72	.64
	3	86.4	.79	.78	.69
	4	85.8	.59	**.93**	.66
Domain-based	N/A	68.2	.00	.00	.00
Homology-based	N/A	52.7	.37	.73	.15
HybSVM	**N/A**	**94.2**	**.92**	.89	**.87**

Accuracy, precision, recall, and correlation coefficient (CC) of classification for the protein-binding versus non-protein-binding dataset are presented for internal machine-learning methods. For each machine-learning approach, values of k ranged from 1 to 4. The performance of the results were estimated using cross-validation. The highest performing value(s) for each performance measure is highlighted in bold.

APPENDIX M

Table S2. Dataset 1 results on standard machine-learning methods

Approach	k	Accuracy	Precision	Recall	C.C
Decision Tree	1	81.6	.72	.68	.57
	2	74.4	.60	.60	.41
SVM	1	85.2	.83	.67	.64
	2	87.2	.82	.76	.70
ANN	1	85.4	.81	.70	.65
	2	86.9	.83	.73	.71
Naive Bayes	1	81.8	.72	.70	.58
	2	82.1	.70	.76	.60
HybSVM	**N/A**	**94.2**	**.92**	**.89**	**.87**

Accuracy, precision, recall, and correlation coefficient (CC) of classification for the protein-binding versus non-protein-binding dataset are presented for standard machine-learning methods. For each machine-learning approach, values of k ranged from 1 to 2. The performances of the results were estimated using cross-validation. The highest performing value(s) for each performance measure is highlighted in bold.

APPENDIX N

Table S3. Dataset 3 results on our internal machine-learning methods

Approach	k	Accuracy	Precision	Recall	C.C
NB k-gram	1	72.9	.37	.39	.20
	2	78.7	.48	.53	.36
	3	82.5	.25	**.90**	.42
	4	83.8	.42	.75	.47
NB(k)	2	81.9	.51	.62	.44
	3	83.2	.31	.84	.44
	4	69.6	.74	.40	.35
Domain-based	N/A	76.4	.00	.00	-.01
Homology-based	N/A	66.4	.74	.34	.32
HybSVM	**N/A**	**89.0**	**.75**	.77	**.69**

Accuracy, precision, recall, and correlation coefficient (CC) of classification for the multi-interface versus singlish-interface dataset are presented for internal machine-learning methods. For each machine-learning approach, values of k ranged from 1 to 4. The performances of the results were estimated using cross-validation. The highest performing value(s) for each performance measure is highlighted in bold.

APPENDIX O

Table S4. Dataset 3 results on standard machine-learning methods

Approach	k	Accuracy	Precision	Recall	C.C
Decision Tree	1	71.6	.30	.24	.07
	2	67.7	.32	.37	.13
	3	71.0	.38	.43	.21
SVM	1	77.4	.00	.00	.00
	2	76.1	.46	.37	.27
	3	80.6	**.86**	.17	.23
ANN	1	76.7	.00	.00	.00
	2	78.0	.56	.09	.08
	3	76.7	.38	.05	.03
Naive Bayes	1	70.3	.40	.63	.29
	2	73.5	.43	.49	.28
	3	81.2	.62	.46	.41
HybSVM	**N/A**	**89.0**	.75	**.77**	**.69**

Accuracy, precision, recall, and correlation coefficient (CC) of classification for the multi-interface versus singlish-interface dataset are presented for standard machine-learning methods. For each machine-learning approach, values of k ranged from 1 to 3. The performances of the results were estimated using cross-validation. The highest performing value(s) for each performance measure is highlighted in bold.

APPENDIX P

Table S5. Dataset 4 results on our internal machine-learning methods

Approach	k	Accuracy	Precision	Recall	C.C
NB k-gram	1	64.1	.48	.64	.27
	2	66.1	.49	**.68**	.31
	3	67.1	.54	.67	.33
	4	60.6	.53	.57	.20
NB(k)	2	64.6	.45	.67	.28
	3	65.1	.53	.64	.29
	4	57.5	.58	.53	.15
Domain-based	N/A	59.1	.62	.30	.14
Homology-based	N/A	29.8	.22	.22	-.43
HybSVM	**N/A**	**69.2**	**.71**	.56	**.37**

Accuracy, precision, recall, and correlation coefficient (CC) of classification for the date versus party dataset are presented for internal machine-learning methods. For each machine-learning approach, values of k ranged from 1 to 4. The performances of the results were estimated using cross-validation. The highest performing value(s) for each performance measure is highlighted in bold.

APPENDIX Q

Table S6. Dataset 4 results on standard machine-learning methods

Approach	k	Accuracy	Precision	Recall	C.C
Decision Tree	1	53.5	.50	**.62**	.08
	2	51.0	.46	.43	.01
	3	52.0	.48	.48	.03
SVM	1	62.1	.61	.50	.23
	2	58.1	.55	.52	.15
	3	62.1	.59	.59	.24
ANN	1	64.6	.69	.42	.27
	2	66.2	.70	.46	.30
	3	65.2	.67	.47	.28
Naive Bayes	1	65.2	.66	.51	.29
	2	64.1	.65	.48	.26
	3	62.6	.61	.52	.24
HybSVM	**N/A**	**69.2**	**.71**	.56	**.37**

Accuracy, precision, recall, and correlation coefficient (CC) of classification for the date versus party dataset are presented for standard machine-learning methods. For each machine-learning approach, values of k ranged from 1 to 3. The performances of the results were estimated using cross-validation. The highest performing value(s) for each performance measure is highlighted in bold.

APPENDIX R

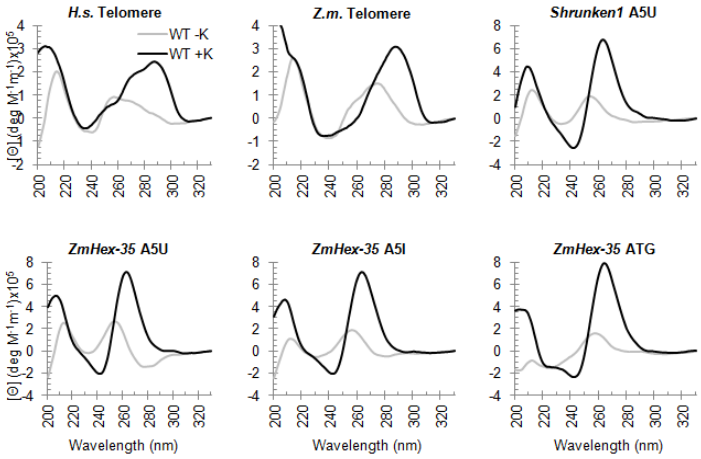
Table S7. Performance measures.

Performance Measure	Formula for Binary Classification
Accuracy	$\frac{TP + TN}{TP + FP + TN + FN}$
Precision	$\frac{TP}{TP + FP}$
Recall	$\frac{TP}{TP + FN}$
Correlation Coefficient	$\frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}}$
F-measure	$2\left(\frac{precision \times recall}{precision + recall}\right)$

The formula for binary classification for each of our five performance measures is provided. *TP, TN, FP, FN* are the true positives, true negatives, false positives, and false negative predictions.

APPENDIX S

Supplementary Figure 1: **CD spectra for maize G4Q oligonucleotides in vitro.** CD spectra of the same oligonucleotide samples as in shown in Figure 6F. Here, the diagnostic G-quadruplex spectra are shown using CD. Oligonucleotides were annealed in either 10mM TBA-phosphate buffer (gray) or in 10mM TBA-phosphate buffer supplemented with 100mM KCl (black). G4Q-specific spectra were obtained for every oligonucleotide tested only in presence of potassium. (**A**) Human telomere repeat. Negative peak around 240 and two positive peaks at 265 and 287 are characteristic for human telomere repeat and suggest a formation of mixed type parallel/antiparallel quadruplex (**B**) Maize telomere repeat. Strong positive peak at 287 taken together with the absence of negative peak at 260 and presence of a valley between 235 and 245 also suggests mixed parallel/anitparallel quadruplex. (**C,D,E,F**) Oligonucleotides with genomic sequences showing characteristic CD-signatures of parallel quadruplex structures with negative peak at 240 and strong positive peak at 260 nm.



Hs4G_bw2
WT: GGATACTTAGGGTTAGGGTTAGGGTTAGGGCGAGTC
Mut: GGATACTTAGAGTTAGCGTTAGCGTTAGCGCGAGTC

HxK_A5I_v2
WT: TGGGGTGGGGGGGGAGCGGG
Mut: TGGAGTCGGAGGAGAGCGCG

sh1_full_v4
WT: GGGAGGGAGGGTTTCTCTGGGACGGGAGAGGG
Mut:GGGAGTGAGGGTTTCTCTGTGACGGGAGAGTG

Zm4xG_bw2
WT: GGATACTTTAGGGTTTAGGGTTTAGGGTTTAGGGCGAGTC
Mut: GGATACTTTAGCGTTTAGAGTTTAGCGTTTAGAGCGAGTC

HxK_A5U_v2
WT: CGGGGGTGTTGAAGGGAGGAGGAGGGAGGGG
Mut: CGACGGTGTTGAAGCGAGGAGGAGCGAGCGG

HxK_ATG_v2
WT: CGGGGGTGTTGAAGGGAGGAGGAGGGAGGGG
Mut: CGACGGTGTTGAAGCGAGGAGGAGCGAGCGG

APPENDIX T

Supplemental Table 1. List of all maize G4Q elements.

Supplemental Table 1 is is too large to be included. It is avaiable by request from the author or will be avaialable at the journal website.

APPENDIX U

Supplemental Table 2. Lists of gene-associated maize G4Q elements.

Supplemental Table 2 is too large to be included. It is avaiable by request from the author or will be avaialable at the journal website.

APPENDIX V

Supplemental Table 3. MaizeCyc Links. Links to HTML files of MaizeCyc omics viewer output for G4Q genes from the A5U, A5I1, and AUG gene lists.

Link to html file displaying MaizeCyc Hits, September, 2013.	Quadruplex Type	List from Figure 2	Distance range	# Genes	Color
http://ftp.maizegdb.org/g4/pathways/ZmG4Qs_A5U_List1_List2_MaizeCYC/	Antisense 5' UTR (A5U)	List 1	(39 to 62)	745	Red
		List 2	(-81 to 273)	3769	Orange
http://ftp.maizegdb.org/g4/pathways/ZmG4Qs_A5I1_List3_MaizeCYC/	Antisense 5' end of Intron 1 (A5I1)	List 3	(9 to 43)	596	Red
http://ftp.maizegdb.org/g4/pathways/ZmG4Qs_AUG_List4_List5_MaizeCYC/	AUG translational start (AUG)	List 4	(-20 to 20)	222	Red
		List 5	(20 to 80)	259	Orange

APPENDIX W

BIOGRAPHICAL VITA

A. Education

INSTITUTION AND LOCATION	DEGREE	YEAR(s)	FIELD OF STUDY
Wartburg College, Waverly, Iowa	B.A.	1996 - 2000	Computer Science and Mathematics
Iowa State University, Ames, Iowa	Ph.D.	2000 - 2013	Computer Science

B. Positions and Honors.

Positions and Employment

1996 - 2000 Undergraduate, Dept of Computer Science, Wartburg College, Waverly, IA
1998 - 1999 Computer programmer. John Deere Waterloo Works, Waterloo, IA
2000 - 2002 Research Fellow, Iowa State University funded by IGERT: NSF, Ames, IA
2002 - 2002 Bioinformatician and Computational Biologist. Pioneer Hi-Bred, IA
2002 - 2004 Research Fellow, Iowa State University funded by Pioneer Hi-Bred, IA
2006 - 2008 Bioinformatician/Computational Biologist. NewLink Genetics, Ames, IA
2008 - NOW Bioinformatics Engineer. USDA-ARS, Ames, IA

Honors

1996 Mathematics Field Day Scholarship
1996 State of Iowa Scholar, Wartburg College
1996 - 2000 Wartburg College Regents Scholar
1997 - 2000 Kappa Mu Epsilon, Mathematics Honor Society
1998 Chellevold Mathematics Scholarship, Wartburg College
2002 - 2004 Baker Center / Pioneer Hi-Bred Bioinformatics Fellowship
2002 - Arthur A. Collins Computer Science Scholarship
2002 - Iowa State University Computer Science Honor Society

C. Publications

Refereed (journal)

1. **Andorf CM**, Honavar V, Sen TZ. (2013) Predicting the binding patterns of hub proteins: a study using yeast protein interaction networks. PLoS One. 2013;8(2):e56833. doi: 10.1371/journal.pone.0056833. Epub 2013 Feb 19.

2. Cannon EK, Birkett SM, Braun BL, Kodavali S, Jennewein DM, Yilmaz A, Antonescu V, Antonescu C, Harper LC, Gardiner JM, Schaeffer ML, Campbell DA, **Andorf CM**, Andorf DE, Lisch D, Koch KE, McCarty DR, Quackenbush J, Grotewold E, Lushbough CM, Sen TZ, Lawrence CJ. (2011) POPcorn: An Online Resource Providing Access to Distributed and Diverse Maize Project Data. Int J Plant Genomics. 2011;2011:923035. Epub 2011 Dec 27.

3. Schaeffer ML, Harper LC, Gardiner JM, **Andorf CM**, Campbell DA, Cannon EK, Sen TZ, Lawrence CJ. (2011) MaizeGDB: curation and outreach go hand-in-hand. Database (Oxford). 2011 May 29;2011:bar022. Print 2011.

4. Harper LC, Schaeffer ML, Thistle J, Gardiner JM, **Andorf CM**, Campbell DA, Cannon EK, Braun BL, Birkett SM, Lawrence CJ, Sen TZ. (2011) The MaizeGDB Genome Browser tutorial: one example of database outreach to biologists via video. Database (Oxford). 2011 May 9;2011:bar016. doi: 10.1093/database/bar016. Print 2011. PMID: 21565781

5. **Andorf, C.M**., Lawrence, C.J., Harper, L.C., Schaeffer, M.L., Campbell, D.A., Sen, T.Z. (2010). The Locus Lookup tool at MaizeGDB: identification of genomic regions in maize by integrating sequence information with physical and genetic maps.. Bioinformatics. 2010 Feb 1;26(3):434-6.

6. Sen, T. Z., Harper, L. C., Schaeffer, M. L., **Andorf, C. M**., Seigfried, T. E., Campbell, D. A., Lawrence, C. J. (2010) Choosing a genome browser for a Model Organism Database: surveying the Maize community. Database Vol. 2010:baq007; doi:10.1093/database/baq007.

7. Sen, T.Z., **Andorf, C.M**., Schaeffer, M.L., Harper, L.C., Sparks, M.E., Duvick, J., Brendel, V.P., Cannon, E., Campbell, D.A., Lawrence, C.J. (2009) MaizeGDB becomes 'sequence-centric' Database. 2009:Vol. 2009:bap020.

8. **Andorf, C**.**M** , Dobbs, D., and Honavar, V. (2007). Exploring inconsistencies in genome-wide protein function annotations: a machine-learning approach. *BMC Bioinformatics*. 2007 Aug 3;8:284.

Journal publication in preparation:

9. **Andorf, C.M.**, Mibiletsky, M., Dobbs, D.,Koch, K., Stroupe, M.E., Lawrence, C., and Bass, H. (2013) Genome-wide screen for G-quadruplex elements in maize (Zea mays L.) reveals networks of genes induced by hypoxia and associated with energy crisis metabolism. To be submitted: *Genes|Genomes|Genetics*.

Refereed (conference, and workshop)

1. **Andorf, C. M**., Cannon, E. K., Portwood, J., Braun, B., Birkett, S., Harper, L.C., Schaeffer, M. L., Gardiner, J. M., Jayasingam, D., Campbell, D. A., Richter, J., Sen, T. Z., Lawrence, C. J. (2013) MaizeGDB: everything old is new again! In: 54th Annual Maize Genetics Conference (MGC 2013), Short Talk, St. Charles, Illinois, USA.

2. **Andorf, C. M.** (2012) The New Maize Genome Database. In: 48th Annual Illinois Corn Breeders' School, Invited Talk, Champaign, Illinois, USA.

3. **Andorf, C. M.,** Cannon, E. K., Braun, B., Birkett, S., Harper, L.C., Schaeffer, M. L., Gardiner, J. M., Jayasingam, D., Campbell, D. A., Sen, T. Z., Lawrence, C. J. (2011) Reinventing MaizeGDB. In: 53rd Annual Maize Genetics Conference (MGC 2011), Short Talk, St. Charles, Illinois, USA.

4. Lawrence, C. J., Cannon, E., **Andorf, C. M**., Campbell, D. A., Harper, L.C., Schaeffer, M. L., Sen, T. Z. (2010) Sequence resources at MaizeGDB with emphasis on POPcorn: a PrOject Portal for corn In: 52nd Annual Maize Genetics Conference (MGC 2010), Short Talk, Riva del Garda(Trento), Italy.

5. Sen, T. Z., **Andorf, C. M**., Campbell, D. A., Schaeffer, M. L., Harper, L. C., Lawrence, C. J. (2009) The MaizeGDB Genome Browser In: 51st Annual Maize Genetics Conference (MGC 2009), Short Talk, St. Charles, Illinois.

6. **Andorf, C**., Silvescu, A., Dobbs, D. and Honavar, V. Learning Classifiers for Assigning Protein Sequences to Gene Ontology Functional Families. In: Proceedings of the Fifth International Conference on Knowledge Based Computer Systems (KBCS 2004), India.

7. **Andorf, C**., Dobbs, D., and Honavar, V. (2002). Discovering Protein Function Classification Rules from Reduced Alphabet Representations of Protein Sequences. In: Proceedings of the Conference on Computational Biology and Genome Informatics. Durham, North Carolina.

8. Honavar, V., **Andorf, C**., Caragea, D., Silvescu, A., Reinoso-Castillo, J., and Dobbs, D. (2001). Ontology-Driven Information Extraction and Knowledge Acquisition from Heterogeneous, Distributed Biological Data Sources. In: Proceedings of the IJCAI-2001 Workshop on Knowledge Discovery from Heterogeneous, Distributed, Autonomous, Dynamic Data and Knowledge Sources.

9. Caragea, D., Silvescu, A., Pathak, J., Bao, J., **Andorf, C**., Dobbs, D., and Honavar, V. (2005) Information Integration and Knowledge Acquisition from Semantically Heterogeneous Biological Data Sources. Data Integration in Life Sciences (DILS 2005), San Diego, Berlin: Springer-Verlag. Lecture Notes in Computer Science. In press, 2005.

10. Caragea, D., Bao, J., Pathak, J., **Andorf, C**, Dobbs, D., and Honavar, V. (2005). Information Integration from Semantically Heterogeneous Biological Data Sources. Proceedings of the Third International Workshop on Biological Data Management (BIDM 2005), Copenhagen, IEEE Computer Society, In press, 2005.

11. Silvescu, A., Reinoso-Castillo, J., **Andorf, C**., Honavar, V., and Dobbs, D. (2001). Ontology-Driven Information Extraction and Knowledge Acquisition from Heterogeneous, Distributed Biological Data Sources. In: Proceedings of the IJCAI-2001 Workshop on Knowledge Discovery from Heterogeneous, Distributed, Autonomous, Dynamic Data and Knowledge Sources.

Refereed Extended Abstracts in Conferences

1. Wimalanathan, K., **Andorf, C. M**., et al. (2013) Functional annotation of B73 gene models: A machine learning approach. In: 55th Annual Maize Genetics Conference (MGC 2013), Poster Program, St. Charles, Illinois, USA.

2. Mauch, E. **Andorf, C. M**., et al. (2013) Compare Identity By State Relationships of the Ames Diversity Panel using TYPSimSelector. In: 55th Annual Maize Genetics Conference (MGC 2013), Poster Program, St. Charles, Illinois, USA.

3. Harper, L.C., **Andorf, C. M**., et al. (2013) How to Access and Use the New MaizeGDB Website. In: 55th Annual Maize Genetics Conference (MGC 2013), Poster Program, St. Charles, Illinois, USA.

4. Portwood, J.L., Cannon, E. K., **Andorf, C. M**., et al. (2013) MaizeGDB has evolved! In: 55th Annual Maize Genetics Conference (MGC 2013), Poster Program, St. Charles, Illinois, USA.

5. Gardiner, J. M., Cannon, E. K., Wimalanathan, K., **Andorf, C. M**., et al. (2013) Gene Expression Analysis Tools at MaizeGDB. In: 55th Annual Maize Genetics Conference (MGC 2013), Poster Program, St. Charles, Illinois, USA.

6. Richter, J.D., Gardiner, J. M., Harper, L.C., Schaeffer, M. L., Cannon, E. K., **Andorf, C. M**., et al. (2013) MaizeGDB Genome Browser. In: 55th Annual Maize Genetics Conference (MGC 2013), Poster Program, St. Charles, Illinois, USA.

7. Walsh, J., Sen, T. Z., Schaeffer, M. L., Gardiner, J. M., Harper, L.C., Cannon, E. K., **Andorf, C. M**., et al. (2013) CycTools: An Interface for Exploring and Updating BioCyc Databases. In: 55th Annual Maize Genetics Conference (MGC 2013), Poster Program, St. Charles, Illinois, USA.

8. Sen, T. Z., et al. (2013) Metabolic Pathway Resources at MaizeGDB. In: 55th Annual Maize Genetics Conference (MGC 2013), Poster Program, St. Charles, Illinois, USA.

9. **Andorf, C. M.,** Rao, B.S., Andorf, D.E., Cannon, E. K., Braun, B. L., Birkett, S. M., Campbell, D. A., Gardiner, J. M., Harper, L.C., Schaeffer, M. L., Sen, T. Z., Lawrence, C. J. (2012) Alpha.MaizeGDB.org: MaizeGDB's interface redesign. In: 54th Annual Maize Genetics Conference (MGC 2012), Poster Program, Portland Oregon, USA.

10. Campbell, D. A., Coe, E., Sachs, M., **Andorf, C. M**., Birkett, S. M., Braun, B. L., Cannon, E. K., Gardiner, J. M., Harper, L.C., Schaeffer, M. L., Sen, T. Z., Lawrence, C. J. (2012) How well do you know YOUR maize research community? In: 54th Annual Maize Genetics Conference (MGC 2012), Poster Program, Portland Oregon, USA.

11. Harper, L.C., **Andorf, C. M.,** Birkett, S. M., Braun, B. L., Campbell, D. A., Cannon, E. K., Gardiner, J. M., Schaeffer, M. L., Sen, T. Z., Lawrence, C. J. (2012) Learn How to Use MaizeGDB. In: 54th Annual Maize Genetics Conference (MGC 2012), Poster Program, Portland Oregon, USA.

12. Schaeffer, M. L., **Andorf, C. M**., Birkett, S. M., Braun, B. L., Campbell, D. A., Cannon, E. K., Gardiner, J. M., Harper, L.C., Sen, T. Z., Lawrence, C. J. (2012) MaizeGDB - use cases that leverage new data and tools. In: 54th Annual Maize Genetics Conference (MGC 2012), Poster Program, Portland Oregon, USA.

13. Wimalanathan, K., Gardiner, J. M., Nahal, H., Cannon, E. K., Patel, R., Braun, B. L., Schaeffer, M. L., Harper, L.C., **Andorf, C. M**., Campbell, D. A., Birkett, S. M., Sen, T. Z., Provart, N., Lawrence, C. J. (2012) Gene Expression Resources Available from MaizeGDB. In: 54th Annual Maize Genetics Conference (MGC 2012), Poster Program, Portland Oregon, USA.

14. Cannon, E. K., **Andorf, C. M**., Braun, B. L., Campbell, D. A., Schaeffer, M. L., Yeh, C., Schnable, P., Lawrence, C. J. (2012) Diversity Data at MaizeGDB. In: 54th Annual Maize Genetics Conference (MGC 2012), Poster Program, Portland Oregon, USA.

15. **Andorf, C. M**., Cannon, E. K., Braun, B., Birkett, S., Harper, L.C., Schaeffer, M. L., Gardiner, J. M., Jayasingam, D., Campbell, D. A., Sen, T. Z., Lawrence, C. J. (2011) Reinventing MaizeGDB. In: 53rd Annual Maize Genetics Conference (MGC 2011), Poster Program, St. Charles, Illinois, USA.

16. Schaeffer, M. L., Gardiner, J. M., Campbell, D. A., Birkett, S., Braun, B., **Andorf, C. M.,** Cannon, E. K., Harper, L. C., Sen, T. Z., Lawrence, C. J. (2011) MaizeGDB Curation Activities -- Diverse Genomes, Gene Expression, Community GO Annotation. In: 53rd Annual Maize Genetics Conference (MGC 2011), Poster Program, St. Charles, Illinois, USA.

17. Campbell, D. A., **Andorf, C. M**., Cannon, E. K., Braun, B., Birkett, S., Gardiner, J. M., Harper, L.C., Schaeffer, M. L., Sen, T. Z., Lawrence, C. J. (2011) MaizeGDB virtualized infrastructure. In: 53rd Annual Maize Genetics Conference (MGC 2011), Poster Program, St. Charles, Illinois, USA.

18. Braun, B., **Andorf, C. M**., Cannon, E. K., Birkett, S., Gardiner, J. M., Harper, L.C., Schaeffer, M. L., Campbell, D. A., Lawrence, C. J., Sen, T. Z. (2011) The MaizeGDB Genome Browser: Tools and Resources. In: 53rd Annual Maize Genetics Conference (MGC 2011), Poster Program, St. Charles, Illinois, USA.

19. Harper, L.C., Schaeffer, M. L., Thistle, J., Gardiner, J. M., Campbell, D. A., **Andorf, C. M**., Cannon, E. K., Braun, B., Birkett, S., Sen, T. Z., Lawrence, C. J. (2011) The new MaizeGDB video tutorial page. In: 53rd Annual Maize Genetics Conference (MGC 2011), Poster Program, St. Charles, Illinois, USA.

20. Bodnar, A., **Andorf, C. M**., Lawrence, C. J. (2011) MaizeResearch.org: Outreach for the maize genetics community. In: 53rd Annual Maize Genetics Conference (MGC 2011), Poster Program, St. Charles, Illinois, USA.

21. **Andorf, C. M**., Sen, T. Z., Harper, L.C., Schaeffer, M. L., Campbell, D. A., Lawrence, C. J. (2010) MaizeGDB Tools and Resources In: 52nd Annual Maize Genetics Conference (MGC 2010), Poster Program, Riva del Garda(Trento), Italy.

22. Schaeffer, M. L., Harper, L.C., Campbell, D. A., **Andorf, C. M**., Cannon, E., Sen, T. Z., Lawrence, C. J. (2010) MaizeGDB: The Data In: 52nd Annual Maize Genetics Conference (MGC 2010), Poster Program, Riva del Garda(Trento), Italy

23. Harper, L.C., Schaeffer, M. L., Campbell, D. A., **Andorf, C. M**., Cannon, E., Sen, T. Z., Lawrence, C. J. (2010) How to use the new sequence-based functionalities at MaizeGDB In: 52nd Annual Maize Genetics Conference (MGC 2010), Poster Program, Riva del Garda(Trento), Italy.

24. Duvik, J., Liu, J., Schlueter, S., Sen, T. Z., **Andorf, C. M**., Wilkerson, M., Lawrence, C. J., Brendel, V. (2010) Maize Community Annotation Project to Improve Gene Structures In: 52nd Annual Maize Genetics Conference (MGC 2010), Poster Program, Riva del Garda(Trento), Italy.

25. **Andorf, C. M**., Sen, T. Z., Harper, L.C., Schaeffer, M. L., Campbell, D. A., Lawrence, C. J. (2009) MaizeGDB: Web Interface and New Features In: 51st Annual Maize Genetics Conference (MGC 2009), Poster Program, St. Charles, Illinois.

26. Harper, L.C. Sen, T. Z., **Andorf, C. M**., Schaeffer, M. L., Campbell, D. A., Lawrence, C. J. (2009) How to use MaizeGDB In: 51st Annual Maize Genetics Conference (MGC 2009), Poster Program, St. Charles, Illinois.

27. Harper, L. C., Campbell, D. A., Schaeffer, M. L., **Andorf, C. M**., Sen, T. Z., Zimmerman, S.A., Sachs M., Lawrence, C. J. (2009) New and Improved

Phenotypes in MaizeGDB In: 51st Annual Maize Genetics Conference (MGC 2009), Poster Program, St. Charles, Illinois.

28. Sen, T. Z., **Andorf, C. M**., Campbell, D. A., Schaeffer, M. L., Harper, L. C., Lawrence, C. J. (2009) The MaizeGDB Genome Browser In: 51st Annual Maize Genetics Conference (MGC 2009), Poster Program, St. Charles, Illinois. \

29. **Andorf, C. M**., Sen, T. Z., Cannon E., Campbell, D. A., Schaeffer, M. L., Harper, L. C., Lawrence, C. J. (2010) MaizeGDB: Tools And Resources In: Plant & Animal Genomes XVIII Conference (PAG 2010), Computer Demonstration, San Diego, California.

30. **Andorf, C. M.,** Sen, T. Z., Cannon E., Campbell, D. A., Schaeffer, M. L., Harper, L. C., Lawrence, C. J. (2010) MaizeGDB: Tools And Resources In: Plant & Animal Genomes XVIII Conference (PAG 2010), Poster Program, San Diego, California.

31. Harper, L. C., Schaeffer, M. L., Sen, T. Z., **Andorf, C. M**., Cannon E., Campbell, D. A., Lawrence, C. J. (2010) How To Use The New Functionalities At MaizeGDB: Now Sequence-Centric! In: Plant & Animal Genomes XVIII Conference (PAG 2010), Poster Program, San Diego, California.

32. Sen, T. Z., **Andorf, C. M**., Campbell, D. A., Schaeffer, M. L., Harper, L. C., Lawrence, C. J. (2009) The MaizeGDB Genome Browser. In: Plant & Animal Genomes XVII Conference (PAG 2009), Poster Program, San Diego, California.

33. Schaeffer, M. L., **Andorf, C. M**., Campbell, D. A., Harper, L. C., Sen, T. Z., Lawrence, C. J. (2009) MaizeGDB -- New Data, Access And Community Curation. In: Plant & Animal Genomes XVII Conference (PAG 2009), Poster Program, San Diego, California.

34. **Andorf, C**., Dobbs, D., Honavar, V. (2006) Learning Classifiers for Assigning Protein Sequences to Subcellular Localization Families. In: Proceedings of the Annual Meeting of the International Society for Computational Biology (ISMB 2006), Poster Program, Fortaleza, Brazil.

35. **Andorf, C**., Dobbs, D., Honavar, V. (2005) Learning Classifiers for Assigning Protein Sequences to Gene Ontology Functional Families. In: Proceedings of the Annual Meeting of the International Society for Computational Biology (ISMB 2005), Poster Program, Detroit, Michigan.

36. Caragea, D., Silvescu, A., Pathak, J., Bao, J., **Andorf, C**., Yan, C., Dobbs, D. and Honavar, V. (2005). Knowledge Acquisition from Autonomous, Distributed, Semantically Heterogeneous Data Sources. In: Proceedings of the Annual Meeting of the International Society for Computational Biology (ISMB 2005), Poster Program, Detroit, Michigan.

37. Pathak, J., Bao, J., Caragea, D., Silvescu, A., **Andorf, C**., Yan, C., Dobbs, D. and Honavar, V. (2005). INDUS: A System for Information Integration and Knowledge Acquisition from Autonomous, Distributed, and Semantically Heterogeneous Data Sources. In: The Program of the Annual Meeting of the International Society for Computational Biology (ISMB 2005), Demo Program, Detroit, Michigan.

38. Caragea, D., Pathak, J., Bao, J., Silvescu, A., **Andorf, C**., Dobbs, D., and Honavar, V. (2005). Information Integration from Semantically Heterogeneous Biological Data Sources. In: Proceedings of the 3rd International Workshop on Biological Data Management (BIDM 2005), DEXA Workshops 2005, Copenhagen, Denmark. Pp. 580-584. IEEE Computer Society.

Abstract Coordinator

1. 55th Annual Maize Genetics Conference (MGC 2013), Abstract Coordinator, St. Charles, Illinois, USA.

2. 54th Annual Maize Genetics Conference (MGC 2012), Abstract Coordinator, Portland, Oregon, USA.

3. 53rd Annual Maize Genetics Conference (MGC 2011), Abstract Coordinator, St. Charles, Illinois, USA.

4. 52nd Annual Maize Genetics Conference (MGC 2010), Abstract Coordinator, Riva del Garda(Trento), Italy.

5. 51st Annual Maize Genetics Conference (MGC 2009), Abstract Coordinator, St. Charles, Illinois, USA.

D. Research Support

1R43HG004021-01 C. Link (PI) 2/01/07 – 7/30/07
NIH SBIR Phase I
VmatchNL – a User-friendly Graphical Interface for Large-scale Genome Analysis.
The goal of this project is to develop a front end to the Vmatch bioinformatics software that will facilitate use of this powerful string matching program for biologists and other non-programming-savvy researchers.
Role: Programmer / Research Scientist (2007)

1R43HG004180-01 W. Young (PI) 3/01/07 – 8/31/07
NIH SBIR Phase I
Exon Boundary Tags (EBTs) for Human Functional Genome Annotation.
This project will develop a novel technology to quantitatively sequence exon boundary tags, with applications in human genome annotation and comparison of transcription levels of alternative splice forms in different cell lines.
Role: Programmer / Research Scientist (2007)

021969 Honavar and Dobbs (co-PI) 2002-2006
National Science Foundation

ITR: Algorithms and Software for Knowledge Acquisition from Heterogeneous Distributed Data
The goal of my project was to develop algorithms and software for a new generalized version of the Naïve Bayes algorithm to use alternative representations of proteins for prediction of novel proteins. These prediction problems include function, structure, subcellular localization, and protein-protein interactions.
Role: Graduate Stipend (2004 – 2006)

GM066387 Honavar and Dobbs (co-PI) 2003-2007
National Institutes of Health
Discovering Protein Sequence-Structure-Function Relationships, Biological Information Science and Technology Initiative
The goal of my project was to test and experimentally validate algorithms and software for a new generalized version of the Naïve Bayes algorithm to use alternative representations of proteins for prediction of novel proteins. These prediction problems include function, structure, subcellular localization, and protein-protein interactions.
Role: Graduate Stipend (2004 – 2006)

Honavar (PI) 2002 - 2004
Pioneer Hi-Bred
Pioneer Hi-Bred Research and Training Grant
The goal of my project was to rewrite the Protein Family Database (PFam) by rebuilding the hidden Markov models using alternative representations of the protein sequences. This work, in terms of the database's predictability of previously unknown proteins, was able to increase selectivity without sacrificing sensitivity.
Role: Graduate Stipend (2002 – 2004)

9972653 Honavar, Voytas, Carpenter, Schnable, Wendel (co-PI) 1999 - 2005
National Science Foundation
Integrative Graduate Education and Research Traineeship (IGERT) program
The goal of my project was to develop a Decision Tree algorithm to use alternative representations of proteins to predict novel proteins' function
Role: Graduate Stipend (2000 – 2002)