

# UC San Diego

## Research Theses and Dissertations

### **Title**

Biomolecular Interactions Using Machine Learning

### **Permalink**

<https://escholarship.org/uc/item/00f756qs>

### **Author**

Bock, Joel R.

### **Publication Date**

2003

Peer reviewed

UNIVERSITY OF CALIFORNIA, SAN DIEGO

**Biomolecular Interactions Using Machine Learning**

A dissertation submitted in partial satisfaction of the  
requirements for the degree  
Doctor of Philosophy

in

Bioengineering

by

Joel Robert Bock

Committee in charge:

Professor David A. Gough, Chair  
Professor Philip E. Bourne  
Professor Charles Elkan  
Professor Julian I. Schroeder  
Professor Shankar Subramaniam

2003

UMI Number: 3091349

**UMI**<sup>®</sup>

---

UMI Microform 3091349

Copyright 2003 by ProQuest Information and Learning Company.

All rights reserved. This microform edition is protected against  
unauthorized copying under Title 17, United States Code.

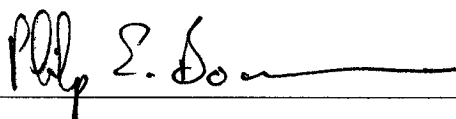
ProQuest Information and Learning Company  
300 North Zeeb Road  
P.O. Box 1346  
Ann Arbor, MI 48106-1346

Copyright  
Joel Robert Bock, 2003  
All rights reserved.

The dissertation of Joel Robert Bock is approved, and  
it is acceptable in quality and form for publication on  
microfilm:

  
\_\_\_\_\_

*Strombrun subcommittee*  
\_\_\_\_\_

  
\_\_\_\_\_

  
\_\_\_\_\_

  
\_\_\_\_\_

Chair

University of California, San Diego

2003

## DEDICATION

This work is dedicated to my family.

*“The principal difficulty in your case...lay in the fact of there being too much evidence. What was vital was overlaid and hidden by what was irrelevant. Of all the facts which were presented to us we had to pick just those which we deemed to be essential, and then piece them together in their order, so as to reconstruct this very remarkable chain of events.”*

—Sherlock Holmes, “The Adventure of the Naval Treaty”, A. Conan Doyle,  
1893.

## TABLE OF CONTENTS

Signature Page . . . . .	iii
Dedication . . . . .	iv
Epigraph . . . . .	v
Table of Contents . . . . .	vi
List of Figures . . . . .	viii
List of Tables . . . . .	x
Acknowledgements . . . . .	xiii
Abstract of the Dissertation . . . . .	xiv
I Introduction . . . . .	1
A Overview . . . . .	1
B Summary and main results . . . . .	2
II <i>In silico</i> biological function attribution: a survey . . . . .	6
A Introduction . . . . .	6
B Assigning function by computer: motivation . . . . .	6
C Classification of computational approaches . . . . .	8
D Survey of methodologies . . . . .	11
E Conclusions . . . . .	29
III Interactions in a broad database . . . . .	31
A Introduction . . . . .	31
B System and Methods . . . . .	33
C Implementation . . . . .	42
D Discussion . . . . .	44
E Conclusion . . . . .	46
F Acknowledgement . . . . .	47
IV Interactions in one species . . . . .	49
A Introduction . . . . .	49
B Methods . . . . .	50
C Discussion . . . . .	58
D Conclusions . . . . .	68



V	Interactions across species . . . . .	69
A	Introduction . . . . .	69
B	System and methods . . . . .	72
C	Algorithm . . . . .	75
D	Implementation . . . . .	77
E	Discussion . . . . .	81
F	Acknowledgement . . . . .	101
VI	A new method to estimate ligand-receptor energetics . . . . .	102
A	Introduction . . . . .	102
B	System and Methods . . . . .	104
C	Implementation . . . . .	108
D	Discussion . . . . .	110
E	Conclusions . . . . .	115
F	Acknowledgement . . . . .	115
VII	Conclusions . . . . .	116
A	Conclusions . . . . .	116
B	Suggestions for future research . . . . .	118
	Appendices . . . . .	120
	Appendix A: Support vector machine . . . . .	121
	Appendix B: Fixed length vector algorithm . . . . .	125
	Bibliography . . . . .	132

## LIST OF FIGURES

III.1	Distribution of protein sequence lengths in database. At least 1,394 distinct interacting domains are represented. $\mu = 481 \pm 386$ residues. . . . .	36
III.2	Steps in the feature vector construction process. For each protein in an interaction pair, residues are encoded as features representing charge ( <i>C</i> ), hydrophobicity ( <i>H</i> ) and surface tension ( <i>T</i> ). These numbers are concatenated in the same order as their appearance in the primary structure of the protein. Next, the length of this array of numbers is normalized to a fixed length. Finally, arrays of features for two proteins are joined to form a feature vector for classification processing. . . . .	48
IV.1	Prediction performance of <i>S. cerevisiae</i> protein-protein interactions as a function of polynomial kernel order <i>d</i> for several feature sets. Colors represent different attribute sets used to encode the amino acid sequences of the constituent proteins comprising a biological interaction. <i>Blue</i> =hydrophobicity only, <i>Green</i> =hydrophobicity and surface tension, <i>Red</i> =hydrophobicity, surface tension and electric charge. Symbols correspond to performance metrics: <i>Circles</i> =accuracy ( <i>A</i> ), <i>Triangles</i> =precision ( <i>P</i> ), and <i>Boxes</i> =sensitivity ( <i>S</i> ). Each data point was obtained from 10-fold cross-validation estimates of SVM performance. Total example count: $n^+ = 2,729$ , $n^- = 3,560$ . . . . .	60
IV.2	Quasi-ROC curves for predictions of protein-protein interactions in <i>S. cerevisiae</i> . Colors represent different attribute sets used to encode the amino acid sequences of the constituent proteins comprising a biological interaction. Each data point was obtained from 10-fold cross-validation estimates of SVM performance. The arrow indicates the direction of increasing SVM polynomial kernel order $d \in \{2,3,4,5\}$ , which corresponds to increasing precision (decreasing sensitivity) as shown in Figure IV.1. . . . .	62
IV.3	Confusion matrices for protein-protein interaction predictions in <i>S. cerevisiae</i> . Features based on residue charge and surface tension (corresponding to the green data points in Figure IV.2). Each matrix represents 10-fold cross-validation performance for a given SVM polynomial kernel order <i>d</i> . Columns and rows marked “+” or “-” indicate interaction or non-interaction, respectively. Off-diagonal elements record the number of examples for which the SVM classifier made an incorrect prediction (a false positive, or a false negative decision). (i): $d = 2$ ; (ii): $d = 3$ ; (iii): $d = 4$ ; (iv): $d = 5$ . Total example count: $n^+ = 2,729$ , $n^- = 3,560$ . . . . .	64
V.1	Phylogenetic bootstrap algorithm. . . . .	78
V.2	Venn diagram depicting experimental and predicted interaction maps. . . . .	88

V.3	Predicted whole-proteome interaction map for <i>Campylobacter jejuni</i> . In this diagram, individual proteins are represented as vertices, and the interactions between pairs of proteins are indicated by edges connecting nodes. Proteins with a large number of partners ( $> 15$ ; 1% of all predictions) are colored red; green nodes signify that relatively few proteins ( $\leq 5$ ; 61% of predictions) are expected to interact with that node. Blue nodes represent proteins with 6–14 interaction partners. . . . .	96
V.4	Principal components of an hypothesized two-component thermoregulation signalling pathway in <i>C. jejuni</i> . Shown is a subnetwork of interactions comprising the primary interaction partners of the sensor (Q9PN36) and regulator (Q9PN67) proteins. Each protein node is labelled by its corresponding ORF designation. The previously uncharacterized protein Q9PMG7 may play a role in transferral of the message from sensor to regulator in the thermoregulation signalling pathway. . . . .	97
V.5	Principal components of an hypothesized ferric uptake regulation pathway in <i>C. jejuni</i> . Each protein node is labelled by its corresponding ORF designation. The figure shows a subnetwork of predicted protein interactions linking the extracellular signal (Q9PJA5, putative integral membrane protein) to the regulatory (P48796, ferric uptake regulation) and transcriptional machinery (Q9PNK3, leucyl-tRNA transferase; Q9PN44, polyribonucleotide nucleotidyltransferase). Such connection is required to respond to changing requirements for iron storage or removal. Protein Q9PMD5 (possible bacterioferritin) may participate in redox stress resistance, by storing iron in a soluble, non-toxic form. Q9PMD5 is linked to a 30S ribosomal protein (Q9PI17) suggesting that this system may be involved in protection of the ribosomal machinery from iron toxicity. . . . .	99
VI.1	Actual versus predicted binding free energy. Shown are typical results from one complete 10-fold cross validation experiment on the ligand-receptor database discussed in Section B.2. Sample size $n=2,671$ . . . . .	110
VI.2	Comparison of error and rank correlation statistics between this study and the literature. The investigations numbered along the horizontal axis appear in order of increasing normalized mean square error $nmse$ (Eq. VI.6), and correspond to the numbering appearing in Table VI.1. Notice the general trend of inverse correlation between binding energy prediction errors ( $nmse$ , $nmae$ ) and rank correlation ( $\tau$ ). The present cross validation results are represented as Investigation #4 in this figure. . . . .	114
.1	A schematic support vector machine for data falling into two classes: $A^+$ (red triangles) and $A^-$ (blue squares). In the case shown here, the classes are linearly inseparable; the SVM has been constructed using a linear kernel. Support vectors are the symbols lying on the margin containing black dots. After [42, 126]. . . . .	124

## LIST OF TABLES

II.1 Listing of the methodologies covered in this review. <i>In silico</i> functional assignment methodologies are classified according to the source of the underlying hypothesis, being generated from biological/evolutionary arguments ( <i>Biological</i> ), or numerically using machine learning techniques ( <i>Machine</i> ).	11
III.1 Organism representation by proteins found in the DIP database, circa January 2001. Frequency expressed as fraction of total number of occurrences of each organism. The top 95% most frequent organisms are listed. Number of interactions $n = 2,664$ .	34
III.2 Most frequent protein domains in the interaction dataset. Frequency expressed as fraction of total occurrences of each domain. Prediction using the Protein Families Database (Pfam v. 5.5 [18]).	35
III.3 System generalization accuracy summary. “Inductive accuracy” is the percentage of correct protein interaction predictions on test data not previously seen by the system. $N=10$ trials.	43
IV.1 Prediction performance statistics for two different SVM classifiers. Data again correspond to the feature set $\{C, T\}$ dominating the ROC space of Figure IV.2. <i>TNR</i> is the true negative rate, and <i>FNR</i> is the false negative rate. Other statistics are defined in Sections B.4 and C.2. Total example count: $n^+ = 2,729$ , $n^- = 3,560$ .	64
IV.2 Comparison of prediction performance measures between the present investigation (“Current value(s)”) and previous investigations found in the literature. References: 1. [172]; 2. [151]; 3. [130]. Notes: “accuracy” and “sensitivity” are not explicitly described in [172] or [151], respectively. Correspondence with statistics of this investigation is inferred from narrative descriptions in the respective investigations.	66
V.1 Comparison of 15 most-frequently observed protein domains in <i>C. jejuni</i> and <i>H. pylori</i> strain 26695, indexed by InterPro accession number. Comparisons of each organism are made relative to the InterPro database [99]. Frequent domains common to both proteomes are shown highlighted in boldface.	72
V.2 10-fold cross-validation performance estimate derived from classifiers trained on examples from the design organism <i>H. pylori</i> . High precision indicates the suppression of Type I (false positive) errors. High sensitivity means that Type II errors are suppressed by the decision function (i.e., low false negative rate). Numbers are expressed as percentages. Data sample size $n = 2,078$ .	83
V.3 Interactions by organism found in the DIP database, circa November 2002. Frequency expressed as fraction of total number of interactions for each organism.	89

V.4	Net charge distribution for all 20 amino acids in <i>H. pylori</i> and <i>C. jejuni</i> . “Proteome” refers to the entire proteome of each organism. “Interaction set” refers to experimental (for <i>H. pylori</i> ) and predicted (for <i>C. jejuni</i> ) protein-protein interaction maps. Numbers shown are relative to <i>E. coli</i> . . . . .	92
V.5	Hydrophobics distribution for all 20 amino acids in <i>H. pylori</i> and <i>C. jejuni</i> . “Proteome” refers to the entire proteome of each organism. “Interaction set” refers to experimental (for <i>H. pylori</i> ) and predicted (for <i>C. jejuni</i> ) protein-protein interaction maps. Numbers shown are relative to <i>E. coli</i> . . . . .	92
V.6	Cysteine residue prevalence for all 20 amino acids in <i>H. pylori</i> and <i>C. jejuni</i> . “Proteome” refers to the entire proteome of each organism. “Interaction set” refers to experimental (for <i>H. pylori</i> ) and predicted (for <i>C. jejuni</i> ) protein-protein interaction maps. Numbers shown are relative to <i>E. coli</i> . . . . .	93
V.7	Comparison of proteome-wide interaction map connectivities for different organisms found in the literature. “Proteome coverage” is the estimated number of distinct proteins involved in interactions as a fraction of either the total proteomic complement or assay depth for a given organism. “Average connectivity” refers to the average number of interaction partners per protein comprising the map. References: 1. [91]; 2. [172]; 3. [207]; 4. Present investigation; 5. [161]; 6. [188]; 7. [187]; 8. [194]. Note: in [187], a retrospective reanalysis of data originally reported in [188] resulted in an updated estimated average connectivity of 4.5–5.8 for <i>S. cerevisiae</i> . . . . .	94
V.8	Distribution of protein interaction cluster sizes compared to [93]. A cluster size represents the average number of interactions (edges) each protein (node) shares with other proteins. “Large” clusters refer to instances of proteins with a large number of partners ( $n > 15$ ); “medium” cluster nodes have $5 < n \leq 15$ , and in “small” clusters each protein has, on average, $n \leq 5$ connections to other proteins. Numbers are expressed as percentage of total number of proteins comprising the map. References: 1. [93]; 2. Present investigation. . . . .	95
V.9	Principal components of an hypothesized two-component thermoregulation signalling pathway in <i>C. jejuni</i> . “Status” refers to the functional annotation status of the ORF, with <i>H</i> =hypothetical, <i>P</i> =putative, <i>A</i> =annotated. . . . .	98
V.10	Principal components of an hypothesized ferric uptake regulation pathway in <i>C. jejuni</i> . “Status” refers to the functional annotation status of the ORF, with <i>H</i> =hypothetical, <i>P</i> =putative, <i>A</i> =annotated. . . . .	100

VI.1 Comparison of predictions of ligand-receptor binding free energies in the present investigation (boldface font) and various studies reported in the literature. Test data statistics are sample size ( $n$ ), target value mean ( $\bar{y}$ ) and standard deviation ( $\sigma_y$ ). Results are shown for normalized mean square error ( $nmse$ , Eq. VI.6), normalized mean absolute error ( $nmae$ , Eq. VI.7), and Kendall's tau ( $\tau$ , Eq. VI.8). References: 1. Head 1996, Table 3 [84]; 2. Böhm 1998, Table 3 [30]; 3. Wang 1998, Table 4 [198]; 4. Bock 2002 (Present investigation); 5. Head 1996, Table 4 [84]; 6. Wang 2002, Table 4 [197]; 7. Rarey 1996, Table 1 [162]; 8. Zhang 1996, Table 1 [212]; 9. Schapira 1999, Table 5 [167]. *Note:* results for present investigation are average values from ten 10-fold cross validation experiments. . . . . 111

## ACKNOWLEDGEMENTS

I wish to thank my wife Janet and son Alexander for their love and support. They both endured years of my intermittent attention, and sleep-deprived, occasionally temperamental behavior. I hope to make it up to both of them.

It has been a great pleasure to be associated with my thesis advisor, scientific collaborator and friend David Gough. Dave has been a consistent source of fresh ideas, enthusiasm and encouragement. His admonition that completing the doctorate while working full time would be “the hardest thing you will do in your life” was right on the mark, and it would not have been possible without his guidance. I look forward to our continued collaborations in the future.

Finally, I would like to acknowledge the contributions of the members of my doctoral committee. Phil Bourne provided advice on issues of structural bioinformatics. Charles Elkan’s contribution to the machine learning aspects of this research was significant, and his influence in this regard pervades the dissertation. Julian Schroeder posed a number of thought-provoking questions from a biologist’s perspective. Shankar Subramaniam’s insights, especially relating to establishment of the biological relevance of the machine learning predictions, were most valuable. The discussion of validation using interologs in Chapter V reflects these questions.

The text of Chapters III, V and VI, in part or in full, are reprints of the material appearing in the following publications, respectively:

1. Joel R. Bock and David A. Gough. “Predicting protein-protein interactions from primary structure”, *Bioinformatics* 17(5):455-460, 2001.
2. Joel R. Bock and David A. Gough. “Whole-proteome interaction mining”, *Bioinformatics* 19(1):125-134, 2003.
3. Joel R. Bock and David A. Gough. “A new method to estimate ligand-receptor energetics”, *Molecular & Cellular Proteomics* 1:904-910, 2002.

The dissertation author was the primary author, and the co-author listed in these publications directed and supervised the research which forms the basis for these chapters.

## ABSTRACT OF THE DISSERTATION

### **Biomolecular Interactions Using Machine Learning**

by

Joel Robert Bock

Doctor of Philosophy in Bioengineering

University of California San Diego, 2003

Professor David A. Gough, Chair

This thesis explores the automatic prediction of biomolecular interactions using machine learning. The overriding philosophy motivating these investigations is to model the interactions between biomolecules (proteins and small-molecule ligands) using simple features to represent characteristics that are hypothesized to contribute to binding.

For these investigations, I use “support vector” learning to build discrimination functions that separate input features into classes, resulting in a hypothesis as to whether or not (or how strongly) the biomolecules will interact. These discrimination functions are based on training data sets of known interactions.

Individual chapters of the thesis center on different investigations which predict protein-protein interactions in a multi-species database, within a single organism and across species. A final study focuses on the prediction of binding free energy between a receptor and ligand.

An important contribution made by this research is the demonstration that no explicit information about three-dimensional protein structure is necessary to make predictions of protein interactions. This implies that researchers may proceed directly from sequence to inference of protein function, as represented by the context of its interaction with other biomolecules.



# I

## Introduction

### A Overview

This thesis explores the automatic prediction of biomolecular interactions using machine learning. The overriding philosophy motivating these investigations is to model the interactions between biomolecules (proteins and small-molecule ligands) using simple features to represent characteristics that are hypothesized to contribute to binding.

Two types of biomolecular interactions are studied: protein-protein, and small molecule-protein. Predicting protein-protein interactions has important implications for assembling networks of interactions within living cells, which is a step toward understanding biological processes as integrated systems. Protein-small molecule prediction may someday provide the means to target pharmaceuticals to inhibit the activity of key proteins within signalling networks associated with disease states.

For these investigations, I use support vector machine (SVM) learning to build discrimination functions that separate input features into classes, resulting in a hypothesis as to whether or not (or how strongly) the biomolecules will interact. These discrimination functions are based on training data sets of known interactions.

The following learning concepts are posed as classification and regression problems, respectively:

- *Protein-protein interaction.* Given features representing amino acid sequences from each protein, construct a function indicating if they do (do not) interact.
- *Protein-small molecule interaction.* Given features representing an amino acid se-

quence for a protein and a connection table for a small molecule, construct a function indicating their binding affinity.

An important contribution made by this research is the demonstration that no explicit information about three-dimensional protein structure is necessary to make predictions of protein interactions.

## **B Summary and main results**

***In silico* biological function attribution: a survey.** In Chapter 2, a survey is presented of the scientific literature on biological function attribution by computer. For each of the investigations and methodologies reviewed, I try to present a balanced critical evaluation of its strengths and weaknesses, by referring to comments provided by peer researchers in this field. A conceptual classification scheme is proposed to compare and contrast the different methodologies that have been reported. This scheme separates computational methodologies into those based on *biological* versus *machine learning* hypotheses. This grouping places the current research in context with other prediction methodologies. It is observed that machine learning-based approaches to function attribution have been used more frequently in recent years, and it is speculated that this trend will continue as *in silico* protein functional assignments mature in reliability, and experimental affirmation of biologically relevant predictions improves our understanding of which techniques work (and which do not).

**Interactions in a broad database.** Chapter 3 represents an intellectual *entrée* into the machine learning of protein-protein interactions using strictly amino acid-based features. The material comprising this chapter was originally published in Bock and Gough, “Predicting protein-protein interactions from primary structure”, *Bioinformatics* 17(5):455-460 [25]. In this work, I trained a series of SVMs to recognize pairs of interacting proteins extracted from a heterogeneous database of experimentally verified protein-protein interactions. The performance of each SVM was evaluated using the inductive accuracy on the previously unseen test examples as the performance metric. I obtained predictive accuracy rates in excess of 80% in these experiments. In discussing these results, I note that they must be interpreted with caution; accuracy as a statistic may be misleading, and there are important issues introduced regarding the distribution of positive and negative data examples to

the learning machine. A central problem is that the distribution of positive (interacting) and negative (non-interacting) examples are highly skewed in Nature; within a given proteome we suspect that most of these are non-interactions. If the SVM is characterized by a constant false alarm rate, when faced with data containing only very few true “signals” of interest, we may observe a large increase in the number of false positive interactions. These points set the tone for subsequent analyses and the presentation of results in later chapters.

It is concluded that future proteomics studies might benefit from applying this methodology by proceeding directly from the automated identification of a cell’s gene products to prediction of protein interaction pairs. The method described in this publication has also been submitted as United States Patent Application #20020090631, “Method for predicting protein binding from primary structure data”.

**Interactions in one species.** Chapter 4 further develops this idea, however the objective here is to predict all of the protein-protein interactions within a single organism. The model organism subject to investigation is the yeast *Saccharomyces cerevisiae*. Tradeoffs that arise between the precision and sensitivity of the protein-protein interaction predictions are explored, based on results obtained from cross validation experiments on a non-redundant database. In these experiments, we observe that for certain SVM architectures the equilibrium binding of proteins was predicted with a high degree of precision ( $> 90\%$ ), but with low sensitivity (36%). Therefore, where confidence in positive predictions is high, many actual protein-protein interactions are not detected by the system. Other architectures produced classifiers characterized by a sensitivity around 64% and a precision of 68%. In this context we note again that a constant false alarm rate SVM would exhibit a sharp decline in precision performance when applied to larger, imbalanced data sets. In such cases the classifier with the highest estimated rate of sensitivity would be preferred, as the sensitivity metric, being independent of the rate of false positives, would remain unchanged.

Ideas of the costs associated with different types of incorrect predictions in the context of protein-protein interactions are introduced. Various sets of amino acid descriptors and SVM architectures are explored in a quasi-receiver operating characteristic (ROC) space, which provides an analysis of groups of classifiers on (false positive, true positive) rate coordinates. It is observed that a certain set of amino acid descriptors used to represent the interacting proteins dominates the ROC-space, providing the greatest level of sensitivity performance at a given false positive rate. Confusion matrices representing binary classifi-

cation results using this set of descriptors are presented. Several key issues are discussed, notably the question of how to properly specify the false negative examples. In addition, it is observed that without eliminating redundant examples, the apparent sensitivity rate, if obtained by indiscriminate predictions without redundancy elimination processing, may be significantly overstated.

These methods and the associated discussion represent an approach toward the engineering of classifiers that may have practical advantages in future proteomics applications. This chapter has been submitted for publication and is currently in review.

**Interactions across species.** Chapter 5 addresses the possibility of generalizing the prediction technique across species. The crux of the idea is to train a system on the known protein interactions in one species, and infer a comprehensive protein-protein interaction map of a different, related species. This idea is formalized in terms of an algorithm (the *phylogenetic bootstrap*), which suggests traversal of a phenogram, interleaving rounds of computation and experiment, to develop a knowledge base of protein interactions in genetically-similar organisms. The efficacy of this algorithm is demonstrated by building a support vector learning system based on 1,039 experimentally validated protein-protein interactions in the human gastric bacterium *Helicobacter pylori*. A complete protein-protein interaction network is predicted for enteric pathogen *Campylobacter jejuni*.

An estimate of the generalization performance of the classifier was derived from 10-fold cross-validation, which indicated expected upper bounds on precision of 80% and sensitivity of 69%. The recurring theme of problems associated with imbalanced data sets is addressed in more detail. When making predictions on all possible pairwise combinations in a different organism, if the classifier is characterized by the same false positive rate estimated from the training data set, the number of false positives would increase significantly. The precision associated with these predictions would be seriously degraded relative to the training data. The sensitivity, or true positive rate, would be expected to remain the same. Predictions made for the “minority class” (here, the interacting protein pairs) would tend to have a much higher error rate than those of the majority class. An interesting observation made here is that by extrapolating from the false positive rates observed during training, we would expect to find a significantly larger number (by a factor of 100) of positive declarations made for the predicted interaction network of *C. jejuni*. This observation suggests an interesting avenue for future research into the “needle in the haystack” problem of data

mining as applied to the biological objectives of this research.

It is further observed that the resulting network of interactions shares an average protein connectivity characteristic in common with previous investigations reported in the literature, offering strong evidence supporting the biological feasibility of the hypothesized map. Specific biological examples of two subnetworks of protein-protein interactions in *C. jejuni* resulting from the application of this approach are presented, including elements of a two-component signal transduction system for thermoregulation, and a ferritin uptake network. This chapter has been published in Bock and Gough, “Whole-proteome interaction mining”, *Bioinformatics* **19**(1):125-134 [28].

**A new method to estimate ligand-receptor energetics.** Finally, in Chapter 6 a new method is proposed to estimate the binding free energy between a small-molecule ligand and a receptor protein. Using support vector regression, a system was trained to learn the functional mapping between a set of ligand-receptor features and the total free binding energy of the complex. Ligand features used are based on the two-dimensional connectivity between constituent atoms and atomic properties. This method potentially provides the capability for large-scale “virtual screening” of receptors against a library of ligands. In cross validation experiments, it is demonstrated that objective measurements of prediction error rate and rank-ordering statistics are competitive with several other investigations, most of which depend on three-dimensional structural data. The size of the sample used ( $n = 2,671$ ) indicates that this approach is robust and may have widespread applicability beyond restricted families of receptor types. It is conjectured that this method may be especially valuable in cases where three-dimensional crystal structures of certain receptors are not easily obtained. This chapter appears in Bock and Gough, “A new method to estimate ligand-receptor energetics”, *Molecular & Cellular Proteomics* **1**:904-910 [27].

## II

# *In silico* biological function attribution: a survey

### A Introduction

Protein function is multifarious. Within the cell, proteins assemble into complex and dynamic macromolecular structures, provide structural support, recognize and degrade foreign molecules, regulate metabolic pathways, control DNA replication and progression through the cell cycle, synthesize other chemical species [2], mediate molecular recognition, organize other proteins within signal transduction cascades [149], and participate in other important functions. An essential characteristic of protein function is context; most cellular processes are carried out by multiprotein complexes [72]. Accordingly, an understanding of the semantics of protein interaction networks and cellular signalling pathways, and their correlation with normal and pathological phenotypes has profound implications for human health. One example of this great potential is in the targeted therapeutic disruption of disease-related signal transduction cascades [55].

### B Assigning function by computer: motivation

*Biological function* itself is an abstract, complex idea. A reasonable attempt at defining biological function appears in [102], where it is argued that multiple levels of context are required to adequately express biological function. These levels include intramolecular interactions within the cell, cell-cell, and tissue-organ interactions [102]. Certainly this

concept might also be extrapolated to encompass interactions between any of the sublevels and the entire organism. “Function” also has a dynamic component, varying both spatially and over time [3].

We concentrate here on methods of hypothesizing protein biological function in terms of intracellular protein-protein interactions. A variety of other approaches to functional inference are conceivable (gene-gene, gene-DNA/RNA, gene-protein, protein-small molecule, protein-epitope), but are not relevant to the present discussion.

Protein function within sequenced genomes can be obtained using experimental or computational means. Once the expressed protein complement of a genome (the *proteome* [199]) is specified, functional assignment may be developed by first elucidating individual protein-protein interactions, then constructing intracellular signalling pathways or networks of these interactions.

Experimentally, the identification of protein-protein interactions has been approached in two ways: (1) genetically, using large-scale systems representing variants of the two-hybrid assay [65, 188, 91, 161] or (2) biochemically, via, e.g., (a) microarrays [122, 213, 118], (b) proteomics technologies combined with molecular biological or immunochemical techniques to identify protein complexes [127, 210], or (c) *in vitro* combinatorial biology [154]. The cited approaches are by no means exhaustive. The two-hybrid screen is currently the most viable technique for large-scale characterization of protein interactions in complete genomes [119], but it is well-recognized that it is prone to generate both false positive and false negative results [172, 196]. In addition, false negative results might be realized due to protein misfoldings, or to insufficient screening depth [91]. Protein microarrays face serious technical problems (denaturing, substrate biocompatibility, uniformity of environment) that must be overcome to scale-up for high-throughput analysis [118]. Exclusive of microarrays, many of the biochemical methods (affinity chromatography, immunoprecipitation) are time-consuming and do not lend themselves to highly parallel experimentation. Other, pragmatic concerns with such techniques are the inability to precisely define constituents of individual protein-protein interactions within a complex of size  $> 2$ , and the detection rates associated with interaction partners present at relatively low concentrations [196].

## C Classification of computational approaches

There is a place for computational analysis to augment nascent experimental methods in the evolving disciplines of functional genomics and proteomics.

Here, we first provide a synopsis of several classification frameworks previously advanced to compare and contrast methodologies. This includes discrimination of computational techniques on the bases of (1) *homology* detection (or its absence), (2) aspects of *genomic context* and (3) *classical* versus *reverse proteomics*. Next, a new perspective is proposed, classifying functional assignment methodologies according to the source of the underlying hypothesis—those generated from biological/evolutionary assumptions and arguments, or those generated numerically, using machine learning techniques. This point of view integrates recent investigations published since the appearance of several excellent reviews on the topic of assigning gene/protein function by computer ([128], [90]).

### C.1 Homology vs. Nonhomology

Historically, the most common computational approach has been to compare the amino acid sequence of an uncharacterized protein to databases comprising (in large part) proteins whose function has been previously established. If a statistically significant similarity is detected, the functional role of the characterized protein may be transferred with confidence to the new protein. In this approach, sequence similarity is used to infer the presence of *homology*, implying “the relationship of two characters that have descended, usually with divergence, from a common ancestral character” [69]. Fitch [67] proposed subclasses of homologs (“paralogs”, and “orthologs”) to differentiate between homology arising from gene duplication events within an organism (e.g.,  $\alpha$  versus  $\beta$  hemoglobin), and that due to speciation events ( $\alpha$  hemoglobin in man versus mouse), respectively. The establishment of orthology between proteins offers a high degree of confidence in at least approximate functional assignment in this approach.

Homology methods for functional inference nevertheless have limitations. They fail when an uncharacterized query protein has no homologs in existing databases, or in cases where significant “hits” are made only to uncharacterized proteins in other organisms [71]. Also, proteins that are distant evolutionary relatives of the query sequence may be missed, where only the top-scoring similarity matches are assumed orthologous. Moreover, in [73] the authors observed that homologous enzymes often do not catalyze the same



reaction owing to divergent evolution, and recommend caution when attempting to assign function from sequence information alone.

In a previous review [128], Marcotte has classified computational functional assignment approaches along the lines of “homology-” and “-nonhomology”-based . In that work, it was noted that functional information manifested as gene fusion patterns, conservation of gene location and other evolutionary information was inherent to sequenced genomes, and that such higher-order information could be understood through cross-genomic comparisons, allowing for inference of entire networks of functionally-related proteins. Importantly, it was further argued that these so-called “nonhomology” methods might allow functional assignment, without the strict requirement for the establishment of homology with a characterized protein.

## C.2 Genomic context

Huynen *et al.* advanced a delineation of methods based on the notion of the “genomic context” of the genes for which a functional assignment is sought [90]. Genomic context methods in the literature were categorized into three distinct types, and compared against one another in terms of their coverage (i.e., how much of the genome could be predicted by each type), the relationship between context type and functional interaction type, and the amount of overlap with homology-based prediction methods.

The three types of genomic context methods reviewed included function prediction using gene fusion events, conservation of local gene context and co-occurrence of genes across genomes (“phylogenetic profiling”). The authors concluded that detection of local gene order provided the highest degree of proteomic coverage, and further that by combining all three methods, reliable functional assignment was possible for 50% of the genes in the gram-positive bacterium *Mycoplasma genitalium*. A further finding was that spatial proximity of two candidate genes was positively correlated with their functional coupling. Finally, when combining the analysis of genomic context with homology search, novel functional assignments for 10% of the *M. genitalium* genome were generated [90].

## C.3 Classical vs. Reverse Proteomics

Walhout proposes a taxonomy of proteomics strategies with two main subdivisions: “classical” and “reverse proteomics” [196]. Classical proteomics is analogous to for-

ward genetics, in that research begins with a genetic screen of the organism (protein extract) to identify phenotypes (proteins) of interest. Classical proteomics approaches include complex purification, co-immunoprecipitation and others. On the other hand, reverse genetics (proteomics) starts with the complete set of genomic (proteomic) sequence(s) and proceeds to design experiments from that position. Reverse proteomics is further partitioned into experimental and *in silico* components. The *in silico* techniques in this scheme (gene fusions, phylogenetic profiles and “interologs”) will be discussed further in the present survey.

It is argued that the reverse proteomics approach is preferred, since complete sets of open reading frames (ORFs) can be predicted and cloned into expression vectors (in the experimental methods), facilitating systematic protein-protein interaction tests, and the ability to study those proteins expressed at low concentrations [196].

#### **C.4 Biological vs. Machine Hypothesis**

We propose a different classification of *in silico* protein function assignment methodologies, based on the genesis of the underlying hypothesis facilitating numerical prediction. The scientific method prescribes three sequential steps: (1) assembling data and observations on phenomena in a physical system, (2) formation of an hypothesis to explain the genesis of the observations, and (3) testing the hypothesis. If the hypothesis is valid, i.e., if it is consistent with the observations, it has predictive utility regarding future outcomes, given additional raw data contributing to the observed phenomena. Invalid hypotheses must be revisited and modified to explain non-corroborating data, or else replaced with a better model.

In machine learning, computer algorithms are used to seek an unknown concept or function  $f$  that converts data into observations. The objective is to find an hypothesis  $h$  that is similar to this function, based on learning from available data. This is the central distinction between machine learning and the previously reviewed functional assignment techniques ([128], [90]); in the latter, hypotheses are formed based on theories regarding biological evolution, whereas in machine learning approaches, hypotheses are constructed automatically in computer memory from data examples.

An increasing number of recent investigations adopt the machine learning approach to predict the functional roles of genes and proteins. In the ensuing survey of the literature, the various methodologies are grossly classified as representing either “biological-”

or “machine-generated” hypotheses. This discrimination provides a foundation for examination of the advantages and disadvantages of these methodologies, and further, emphasizes the continuing requirement to reexamine hypotheses fundamental to any scientific investigation.

A summary list of the investigations surveyed here is presented in Table II.1.

<i>Hypothesis basis</i>	<i>Method identifier</i>	<i>Year</i>	<i>Reference</i>
Biological	Clusters of orthologous groups	97	[184]
Biological	Differential genome analysis	98	[89]
Biological	mRNA expression clustering	98	[60]
Biological	Phylogenomics	98	[58]
Biological	Gene proximity	98,99	[50, 146]
Biological	Gene/Domain fusion events	99	[129, 63]
Biological	Phylogenetic profile	99	[153]
Biological	Hybrid	99	[130]
Biological	Interologs	00	[195]
Biological	Phylogenetic tree similarity	01	[151]
Machine	Text mining	99	[24]
Machine	Map topology	00	[172]
Machine	Rules mining	00	[105]
Machine	SVM mRNA expression	00	[38]
Machine	SVM interactions	01,02	[25, 26]
Machine	Correlated sequence signatures	01	[181]
Machine	Interacting domain profile pairs	01	[207]
Machine	Probabilistic map inference	02	[80]
Machine	Probabilistic domain interactions	02	[52]
Machine	Orthogonal experiments	02	[186]
Machine	SVM map inference	03	[28]

Table II.1: Listing of the methodologies covered in this review. *In silico* functional assignment methodologies are classified according to the source of the underlying hypothesis, being generated from biological/evolutionary arguments (*Biological*), or numerically using machine learning techniques (*Machine*).

## D Survey of methodologies

In this section, we offer a brief summary of each investigation selected for review. While pure experiments to render large-scale protein interaction networks (e.g., [188, 91,

161]) are not covered, their vital importance to advancing all *in silico* techniques, through the provision of training data and ultimate validation of predictions, is recognized.

In broad terms, each functional assignment methodology class can be criticized on fundamental grounds. For example, the Biological Hypothesis-based methods may be biased due to the underlying hypothesis [207], and methods predicting functional linkage maps often require additional information to interpret the biological nature of an interaction [166]. On the positive side, the scientific line of reasoning behind the hypothesis is clearly explicated. Contrast this situation with that of the Machine Hypothesis-based methods, many of which do not explain how or why they arrive at a hypothesis or particular result [14]<sup>1</sup>. This perceived lack of comprehensibility may, however, be associated with a reduced bias, provided that the training data sets are sampled randomly enough, and the learning task is properly formulated.

It should be emphasized that success of any computational approach that leverages data from experiment or from existing sequence databases is contingent upon the quality of that data. This is true irrespective of the source of the hypothesis. For example, as noted above, two-hybrid experimental approaches tend to create false-positives [172]. Errors in data input to a computer program also produce incorrect or misleading results. Erroneous or incomplete functional annotations are still present in existing sequence databases [33], and will continue to propagate if *in silico* methods are applied without further curation of the source data.

For each investigation presently under review, we summarize the underlying concept, its advantages and disadvantages, and the potential scope of the functional predictions. Related methods are noted, based on conceptual or technical association to the particular investigation.

The discussion within each hypothesis class proceeds in approximate chronological order of year of publication.

## D.1 Biological Hypothesis-based Methods

The Biological Hypothesis-based Methods are founded upon observations of patterns observed in genomic or proteomic sequences, and some theoretical statement relating these observations to evolutionary biology. Included here are computational techniques ex-

---

<sup>1</sup>Exceptions include rule-based systems, such as [105], which generate readily intelligible rules composed of the independent variables in a decision task.

exploiting gene fusion events, gene proximity, similarity of phylogenetic patterns or trees, orthologous groups, mRNA expression patterns, and others.

### **Clusters of orthologous groups** [184]

*Concept:* The authors introduced the idea of clusters of orthologs groups (COGS), which are cross-species gene groups encoding protein families related by vertical evolutionary descent. The basic concept here is the observation that orthologs in different species often have the same function, allowing transfer of functional information from one member to an entire COG.

*Advantage:* Orthology between genes provides strong clues for functional assignment.

*Critique:* (a) Cross-lineage orthology is not necessarily a one-to-one relationship, since a single gene in species *A* may correspond to a family of paralogs (similar by gene duplication) in species *B* [71]. (b) Even where true orthologous genes are present, only approximate functional assignment is possible [67]. (c) Coverage is limited by the detection of orthology, which relies on detection of sequence similarity. Distantly-related genes may be missed as only the top-scoring matches are assumed orthologous.

*Predictive scope:* Single gene or protein.

*Related methods:* Phylogenetic profile [153].

### **Differential genome analysis** [89]

*Concept:* A specific phenotype of interest is identified. Genomic data representing two species (one displaying this phenotype, the other in which it is absent) are compared to one another, and common, homologous genes are systematically removed from consideration as potential species-specific genes. The result is a disjoint subset of genes within the organism expressing the phenotype for which functional assignment is made.

*Advantage:* It is claimed that this technique is “automated and rapidly yields a small subset of the genome” containing genes responsible for species-specific phenotypes [89].

*Critique:* This technique makes a strong assumption by directly linking a subset of species-specific genes to an observed phenotype. This ignores the reality that many functions may be associated with a single gene.

*Predictive scope:* Groups of genes or proteins.

*Related methods:* Phylogenetic profile [153]

### **Messenger RNA expression clustering [60]**

*Concept:* Using data from DNA microarray hybridization, genes with similar expression patterns are clustered hierarchically. The "co-expression" hypothesis maintains that genes of similar function will cluster together.

*Advantage:* Highly parallel, allowing for large-scale production of mRNA transcription and fast numerical analysis.

*Critique:* (a) Static views of mRNA expression are only useful for quantifying which genes are upregulated/downregulated at one instant in time. It has been argued on theoretical grounds that simultaneous mRNA expression and protein concentration data are required to enable a complete understanding of the dynamics between genomic sequence and observed phenotype [83]. This has been borne out experimentally in at least two different investigations:

- i. Anderson *et al.* [6] found poor correlation between mRNA expression and protein abundances in human liver tissue, using two-dimensional electrophoresis to analyze protein levels and transcript image methodology to measure mRNA, and
- ii. Gygi and co-workers [81] showed that correlation between mRNA and protein levels was insufficient to predict protein expression levels from mRNA transcript data in *Saccharomyces cerevisiae*.

(b) It was recently shown that in *Drosophila melanogaster*, co-expression occurs along 10-30 gene blocks of chromosomally proximal genes, accounting for 20% of the fly genome. These genes are not functionally linked. Therefore, co-expression does not imply functional similarity [180]. (c) Clustering may work well for strongly coexpressed genes, but is not necessarily good for other gene groups [128].

*Predictive scope:* Genomic scale.

*Related methods:* SVM mRNA expression [38].

### **Phylogenomics** [58]

*Concept:* This method postulates that since gene functions change in evolution, a reconstruction of the evolutionary history of a gene and its orthologs, including information regarding events such as gene duplications, lateral transfer and gene loss can be used to infer the function of uncharacterized gene products. This is accomplished by overlaying previously determined functions of the orthologous genes onto the tree, and attempting inference on the function of uncharacterized genes according to their subfamily location within this phylogenetic tree. “Phylogenomics” combines genome sequence information and phylogenetic analysis

*Advantage:* This method extends beyond the analysis of similarities or differences between genomes, adding information represented by explanations for discrete events in an organisms’ evolution.

*Critique:* (a) As with all methods relying on the detection of homology, distant orthology may remain unrecognized in the initial alignment used to reconstruct the phylogenetic tree. (b) Success of phylogenetic methods requires that protein functions change over time with only slight modifications to the corresponding amino acid sequences [58]. (c) See critiques (a,b) under the heading *Clusters of orthologous groups* above.

*Predictive scope:* Gene family.

*Related methods:* Clusters of orthologous groups [184], Phylogenetic profile [153].

### **Gene proximity** [50, 146]

*Concept:* These methods are based upon the assumption that conserved, physically proximal gene pairs comprise functional linkage between the constituent genes, and therefore are useful to predict functional coupling between the prokaryotic gene products.

*Advantage:* Each of these methods is amenable to semi-automatic processing for comparison of groups of prokaryotic genomes.

*Critique:* (a) Indiscriminate application may lead to false predictions: the constraint of proximity is not strong, and cannot predict interactions between distantly located genes [63]. (b) Gene proximity is not applicable to eukaryotes, as gene coregulation is not imposed at genome structure level [63]. (c) Composition of operons is evolutionarily variable, and one cannot count on a particular set of functionally related genes to always comprise an operon [71]. (d) The method may not extend to eukaryotes, who lack operons [128]. (e) The genomic coverage under this method is low; there is a dual requirement to identify orthologs in another genome, and to find those orthologs collocated along the genome of interest [128]. (f) See critique (b) under the heading *Messenger RNA expression clustering* above.

*Predictive scope:* Subset of genome for which proximal genes are functionally coupled.

#### **Gene/Domain fusion events** [129, 63]

*Concept:* These methods advance the hypothesis that distinct proteins which functionally interact in one organism may appear as fused together within another, multi-domain protein (the “Rosetta Stone” protein [129]) that is expressed in a different organism. Recognition of such fused proteins is used to infer the functional coupling between the two distinct proteins represented within the observed domain fusion.

*Advantage:* The methods are capable of predicting functionally linked proteins as well as physical protein-protein interactions.

*Critique:* (a) The methods are prone to false negatives; mechanisms other than Rosetta Stone may be involved in protein-protein interactions, such as gradual accumulation of mutations to evolve a binding site [129]. Also, the artifact of the fused protein may have disappeared during evolution so none exists to point to a potential interaction between other proteins. (b) False positives may be realized in cases where domains are fused but not interacting, or due to the inability to discriminate between binding/nonbinding homologs [129]. (c) “Promiscuous domains” tend to combine with variety of other domains, creating false positives. It must be shown that stand-alone



counterparts of Rosetta stone protein's components are indeed orthologs (versus paralog), or else false positives increase significantly. Non-orthologous gene displacement (NOGD) events are common [71]. (d) The method reported in [63] has been criticized for lack of coverage, as only 64 interactions in 3 bacterial genomes were presented [151]. A similar criticism was made regarding the Rosetta Stone method. For *Escherichia coli* K-12, a proteome containing (at that time) more than 4,200 proteins, only 749 interactions were found after removal of suspected false positives [151].

*Predictive scope:* Subset of proteome for which fusion events are observed.

### **Phylogenetic profiles** [153]

*Concept:* The “phylogenetic profile” of a given protein describes the presence or absence of homologous proteins appearing across organisms. The hypothesis is that functionally linked proteins evolve in a coordinated manner. Therefore, functional associations may be inferred between a pair of proteins observed to frequently co-exist (across genomes) within a structural complex or metabolic pathway. If function annotation for one of the proteins is at hand, then strong clues as to the function of the second protein are provided. The profile for a protein sequence consists of a bit string indicating its homologs across organisms (here, 16 genomes were used, and predictions were made for protein functions in *E. coli*).

*Advantage:* Information contained within the profiles will increase as more genome sequences are obtained. Presumably this will increase the method's expressive power and applicability to organisms beyond the prokaryotes.

*Critique:* (a) Like the *Gene/domain fusion* techniques, the utility of this methodology may be confounded by partial redundancy in gene functions, non-orthologous gene displacement, horizontal gene transfer and lineage-specific gene loss [71]. (b) See critique (c) under the heading *Clusters of orthologous groups* above.

*Predictive scope:* Proteome-wide.

*Related methods:* Differential genome analysis [89], Clusters of orthologous groups [184]

**Hybrid** [130]

*Concept:* In this investigation, the authors report a hybrid analytical method that combines phylogenetic profiles [153], mRNA co-expression [60] and domain fusion events [129] to predict a large-scale “functional linkage” map in *S. cerevisiae*. This map was augmented with experimental and functional annotation data from several protein interaction and mRNA expression databases. The connections within this map enable inference on uncharacterized proteins when linked to proteins of known function. Varying degrees of confidence in these inferences are assigned based on the methods used to generate the pairwise protein interactions comprising the characterized nodes.

*Advantage:* Combining different methods can reduce bias, and improve the reliability of predictions made in this hybrid methodology.

*Critique:* (a) The presence of connections in the map derived from this method may not necessarily be equated with functional prediction. It is noted that only 15% “high confidence” functional links found. Even these are ambiguous; spurious interactions may reflect a high degree of conservation of some proteins, perhaps resulting in similar phylogenetic profiles [71]. (b) Overlap of predictions with a functional linkage map developed in an independent study on *S. cerevisiae* [172] is constrained to the 15% “high quality” predictions.

*Predictive scope:* Proteome-wide.

*Related methods:* Map topology [172].

**Interologs** [195]

*Concept:* This method proposes an extension of the idea of searching across species for orthologous proteins to orthologous *interactions*. “Interologs” are inferred as follows: two proteins, say  $A$  and  $B$ , are observed or known to physically interact in species  $S_1$ . It is postulated that their respective orthologs ( $A'$ ,  $B'$ ) in another species  $S_2$  also interact, due to conserved co-evolution. Then ( $A' - B'$ ) and ( $A - B$ ) are said to be the interologs.

*Advantage:* Interologs present a powerful and expressive technique for functional inference across organisms on a proteome-wide scale.

*Critique:* The same criticisms applied to other orthology based methods (above) are relevant here.

*Predictive scope:* Proteome-wide.

*Related methods:* Clusters of orthologous groups [184], Phylogenetic profiles [153], Phylogenetic tree similarity [151]

### **Phylogenetic tree similarity [151]**

*Concept:* This investigation measures the similarity (distance) between phylogenetic trees, taken to indicate degree of correlation between the distance matrices used to build the trees. This in turn is used to predict interactions between sequences of associated protein families. The hypothesis is that phylogenetic trees of interacting proteins reflect their coordinated evolution, in particular the similar evolutionary pressures applied to all elements of a given molecular complex. A near proteome-scale set of protein-protein interaction predictions in *E. coli* was carried out, resulting in 2,700 putative interactions. It is asserted that pairs of interacting proteins can be correctly predicted at a true positive rate  $> 66\%$ , where the numerical value of this correlation index exceeds a certain threshold.

*Advantage:* Potentially, this method is capable of proteome-wide predictions. It is not predicated on the presence of fully-sequenced genomes, rather only requires data regarding protein families.

*Critique:* (a) Coverage using phylogenetic tree similarity is limited, as only 2,700 interactions were inferred from an initial alignment of 4,300 proteins representing the *E. coli* proteome. (b) Predictive success is very susceptible to quality of the multiple sequence alignments used to infer tree similarities. In particular, poor alignments adversely impact the false positive rate of the predictions [151].

*Predictive scope:* Proteome-wide.

*Related methods:* Clusters of orthologous groups [184], Phylogenomics [58], Phylogenetic profiles [153].

## D.2 Machine-Hypothesis Based Methods

Machine-generated hypotheses implicitly model biological function by learning patterns inherent in data. Applications may be formulated as machine learning tasks using a variety of techniques, including decision trees, neural networks, support vector machines, Bayesian inference, statistical estimation, and clustering algorithms.

### **Text mining** [24]

*Concept:* The text mining system described in this report automatically detects protein-protein interaction information from textual abstracts found in the literature. Sentences culled from sets of abstracts relating to a subsystem of interest are used to analyze patterns of frequently-occurring keywords relating to proteins and their mutual interaction. The trained system infers functional interactions and may be used to reconstruct the topology of interaction networks. An example reconstruction of the Toll and Pelle system in *D. melanogaster* identified 8 of 9 interactions correctly.

*Advantage:* Several different applications are proposed for this methodology, including functional annotation of genomes, automated database curation and prediction of macromolecular interaction networks.

*Critique:* (a) The success of the text mining system is dependent on the frequency of words and their linguistic context, and therefore may miss interactions that are novel or sparsely reported in the literature. (b) The heterogeneity of language comprising published scientific discourse is a significant challenge to the widespread applicability of such techniques. (c) Coverage for complex interaction networks is insufficient; for the *D. melanogaster* cell cycle control system, only 33 of 91 available proteins were identified as “significant” by the system [24]. (d) The authors note a tendency for the system to insufficiently discriminate between biologically significant interactions and other, insignificant results mentioned in the source text.

*Predictive scope:* Pathway or network.

### **Map topology** [172]

*Concept:* The approach used in this investigation is to construct a graph of experimentally confirmed protein-protein interactions, in this case the baker’s yeast *S. cere-*

*visiae*. Noting that proteins of like function and subcellular location tend to co-locate within this network, one can postulate novel functions for proteins based upon their linkages to characterized proteins. The reliability of the network was evaluated by testing its ability to correctly predict the functions of characterized proteins. The authors observed correct function predictions at rate of 72% for 1,393 proteins, compared to a rate of 12% correct for randomized links.

*Advantage:* This method has the ability to infer new functional interactions, from local subnetworks up to proteome-wide scale. This is accomplished using only incomplete knowledge of protein-protein interactions within the organism under consideration.

*Critique:* (a) It is difficult to evaluate the plausibility of predicted interactions among proteins of different functional classes. There could be false-positives, crosstalk interactions, or interactions with related pathways. Further, an “implausible” interaction may in fact represent a false negative decision in a related pathway [187]. (b) The quality of the input data determines an upper bound on the quality of the functional prediction. See the discussion in Section D.

*Predictive scope:* Proteome-wide.

*Related methods:* Hybrid [130].

### **Rules mining** [105]

*Concept:* The idea behind this investigation is that a properly constructed discrimination function might be able to directly convert amino acid sequence to protein function. The authors describe a hybrid machine-learning architecture integrating clustering and a decision tree algorithm (C4.5; [160]). They construct a database of genes labelled using known functional assignments for the bacterium *Mycobacterium tuberculosis*. For each record in this database, a set of descriptors were developed, being computed from the encoded amino acid sequence alone. These descriptors included singlet and doublet residue runs, organism phylogeny, annotation keywords, amino acid sequence length and gene molecular weight. The descriptor-supplemented database was then “mined” by randomly splitting its records into three parts; 2/3 were used to generate prediction rules, and the 1/3 partition was held aside

to test the predictive accuracy of these generated rules. In addition, the rules were applied to genes of unknown function for novel inference. New functions are assigned to 65% of test genes comprising “unknown” or “hypothetical” classes.

*Advantage:* Relying only on protein sequence based features to construct the discrimination rules, such systems may provide a means for function prediction on a proteome-wide scale in the face of negligible sequence homology, and without the requirement for information regarding three-dimensional conformation. The generated rules are comprehensible by humans, a decided benefit to experts using the output of the system to design experiments or annotate databases.

*Critique:* There is insufficient information in the presented results to fully evaluate the predictive acuity of the method. Accuracy results presented indicate 76% average correctness on test data, however this rate apparently applies to only 1% of the data. Lower rates (62-65%) were reported for larger numbers of predictions. The most voluminous group also shows that simply selecting the most populous class would be correct 48% of the time.

*Predictive scope:* Proteome-wide.

*Related methods:* SVM interactions [25, 26].

### **SVM Messenger RNA expression [38]**

*Concept:* This work reports an application of supervised learning which maps DNA microarray gene expression patterns to a functional classification. A support vector machine (SVM) is trained to discriminate between sets of genes comprising disjoint functional classes, and this machine is subsequently used to predict the functions of uncharacterized genes. To demonstrate, expression data from 2,467 *S. cerevisiae* genes, representing 79 different hybridization experiments are used to train the system to recognize patterns associated with five different functional classes. Five of the six classes were chosen because of their similar expression profiles, and the sixth (helix-turn-helix proteins) was selected as a control group, as it was believed that proteins within this class are not similarly regulated. For each of the five learnable classes, an SVM was trained to recognize members/non-members of that class, and predictions were made for a total of 3,754 genes. Three-fold cross-validation was

used to evaluate functional prediction performance. Functional prediction for 15 yeast ORFs of unknown function are presented.

*Advantage:* The advantages of SVM accrue to this investigation: sparse representation of large data sets, implicit nonlinear mapping from mRNA expression to high dimensional feature space representing protein function, noise rejection characteristics, and relatively rapid numerical convergence.

*Critique:* A systematic error in correct identification of certain protein functional classes was observed [38]. False positives may have occurred due because database annotation identifies protein complex members via biochemical co-purification, but the expression experiments highlight functional relation without the protein necessarily being physically connected. False negatives were attributed to (i) differences in database classification (by structure) versus the SVM learning, based on cell genetic response; (ii) differences in regulation context between DNA microarray (transcriptional) and genes that may be regulated by posttranslational mechanisms; (iii) genes corresponding to corrupt microarray data.

*Predictive scope:* Genome-wide.

*Related methods:* Messenger RNA expression clustering [60].

### **SVM interactions** [25, 26]

*Concept:* In these investigations, the authors formulate the mapping from amino acid sequence to function (as represented by protein-protein interactions) as a classification problem. This problem is solved using a support vector machine that learns to discriminate features within interacting protein pairs. Significantly, these features are computed only from physicochemical properties of the amino acid sequences. The labelling of these feature vectors is a binary, according to the interaction class membership (interacting, or non-interacting) of the constituent proteins. In [25], positive examples from a multiple-species database of protein-protein interactions ([209]) were combined with negative examples generated by randomizing sequences from the database at large to generate “native-like” features. Partitioning the data sets into roughly equally sized sets of 2200 training and 2200 testing examples, several SVMs were trained and tested. The inductive accuracy (no. of correct predictions as

a percentage of total predictions) averaged  $> 80\%$ . In [26], the focus is on a single organism, *S. cerevisiae*, and here negative example data are constructed by randomly sampling the balance of the yeast proteome. In 10-fold cross-validation testing, protein binding was predicted with a precision of 90% and inductive accuracy of  $> 70\%$ , at the expense of low sensitivity (about 36%).

*Advantage:* These investigations detail a methodology that allows for the direct prediction of protein function, strictly using features computed from the amino acid sequence. All that is required to implement this approach is a set of confirmed protein-protein interactions and the proteomic sequences for the organism of interest.

*Critique:* This method is data-driven, and confidence in the resulting predictions depends on the quality of the experiments and associated annotations forming the foundation for the machine hypothesis.

*Predictive scope:* Proteome-wide.

*Related methods:* Rules mining [105].

### **Correlated sequence signatures [181]**

*Concept:* The method analyzes the mutual information between amino acid sequences comprising paired interacting proteins. Frequently occurring sequence-signatures, called “correlated sequence-signatures”, are taken as characteristic motifs that are learned and used to detect interactions between other proteins of uncharacterized functional activity. Using a database of 1,274 experimentally determined interacting protein pairs in *S. cerevisiae*, signatures were constructed from regular expressions, profiles, fingerprints and hidden Markov models of the InterPro database [10]. The database proteins were characterized by 434 sequence-signatures. Leave-one-out cross validation testing on an small subset (40 proteins) indicated a sensitivity of 94%, however this subset was selected because of its favorable mutual information characteristics, introducing bias into this test.

*Advantage:* This method is easy to implement numerically, and lends itself to automation.



*Critique:* Based on the sparsity of presented results, it is unclear whether that proteomic coverage may be inherently limited using this methodology. This may be the result of insufficient training data, as the authors state that only 50% of interacting yeast protein signatures had corresponding classifications within the source database

*Predictive scope:* Proteome-wide.

### **Interacting domain profile pairs** [207, 206]

*Concept:* This work presents a method for transferring information contained in a complete protein-protein interaction map from a source species  $S_S$  to a target species  $S_T$ . This facilitates predictions of a complete set of interactions in  $S_T$ . The central method combines experimental protein interaction, interaction domain and sequence data with homology search and clustering techniques. The key step is to construct an intermediate map of domain-protein interactions, expanding the inference space to include clusters of multiple interactions with a given conserved domain. After a correspondence between this interaction map and the target proteome, novel interactions are inferred. The method was demonstrated by developing a correspondence between a source organism, *Helicobacter pylori*, and a target organism, *E. coli*. 1,524 known interactions in *H. pylori* were used to construct a domain interaction map with 1,568 vertices and 1,810 edges. Finally, 881 interactions were predicted for *E. coli*, connecting about 10% of the proteome.

*Advantage:* Interaction domain profiles appear to provide interaction predictions at low levels of sequence similarity, such that they may be missed altogether by sequence similarity techniques. Further, it is argued that including domain information reduces false positives due to multi-domain proteins using other methods.

*Critique:* The method is very strongly reliant on a complete, accurate and detailed reference data set [207].

*Predictive scope:* Proteome-wide.

*Related methods:* SVM map inference [28], Probabilistic map inference [80]

**Probabilistic map inference** [80]

*Concept:* This report proposes assembling a statistical model of a network of interacting proteins, based upon experimentally verified interactions and estimates of interaction map topology. In this approach, a posterior probability can be assigned to novel predicted interactions. The map model is constructed as a graph where vertices represent proteins and edges denote physical binding between the intersecting nodes. Edge probabilities are assigned by analysis of the binding propensities of the constituent domains located on the two proteins linked by the edge. The probability of binding between domains on two edge-linked proteins is estimated by frequency analysis of experimental protein-protein interactions and their constituent domains. Network topology is also assigned a probability, depending on the distribution of edges into and out of each vertex. Biologically realistic topologies are given a higher probability of occurrence. Yeast (*S. cerevisiae*) and *Homo sapiens* protein interaction data were aggregated from various online databases into a single training data set. For yeast, 708 interactions were represented, and the human component comprised 778 interactions. Novel predictions were attempted on a set of 40 human proteins from a known network involving apoptosis, not part of the training data.

*Advantage:* Probabilities are assigned to edges, subnetwork and full network topologies. The Bayesian framework integrates all types of data into the predictions. The general approach is not limited to protein-protein interaction networks; nucleic acids and small molecules may in principle be analyzed in a similar way.

*Critique:* While theoretically well-grounded, the predictive accuracy of the system as demonstrated is unacceptable. (a) Out of 97 edges (interactions) declared by the system, 8 of 44 true interactions present in the test data were correctly predicted with probabilities exceeding pure chance ( $> 0.5$ ). This means that the false negative rate exhibited within the predictions is quite high. (b) At the same time, there are apparently 53 false positives. Some of the difficulty may be due to the lack of adequate domain data associated with the interactions in the training set; for yeast, 40% of the interactions lacked domain annotation.

*Predictive scope:* Proteome-wide.

*Related methods:* SVM map inference [28], Interacting domain profile pairs [207]

### Probabilistic domain interactions [52]

*Concept:* A maximum likelihood approach is used to predict domain-domain interactions based on data sets of protein-protein interactions and the analysis of their sub-domains. The authors consider the prevalence of erroneous experimental data points in forming individual domain-domain interaction probabilities. The accuracy of this method is evaluated by predictions of protein-protein interactions in *S. cerevisiae*, combining interaction data collected in two independent, large-scale two-hybrid experiments [91, 188]. Reported results achieved were 39% specificity and 80% sensitivity. Novel protein-protein interactions are generated by the model.

*Advantage:* (a) The highest confidence predictions under this method can be ranked according to their expected probability of occurrence. (b) The false positive and false negative rates of experimental two-hybrid protein interaction assays are incorporated in a principled way. This provides a means to estimate the confidence individual predictions given error rates characteristic of an experimental protocol.

*Critique:* (a) The model assumes that domain-domain interactions are independent. Interaction between protein domains may depend on the presence of other domains in the same protein, or on environmental factors [52]. (b) No dynamics are present in this model. Domains present on two proteins are assumed to interact, without regard to time differences in their respective expression that may be present biologically [52].

*Predictive scope:* Proteome-wide.

*Related methods:* Interacting domain profile pairs [207], Probabilistic map inference [80]

### SVM map inference [28]

*Concept:* In this investigation, the authors expand upon the SVM interactions method [25, 26], proposing a combined computational-experimental framework for proteome-scale protein-protein interaction prediction. This framework is distilled into an algorithm the “phylogenetic bootstrap”), which suggests traversal of a phenogram, interleaving rounds of computation and experiment, to develop a knowledge base of

protein interactions in genetically-similar organisms. The approach uses analogy between the proteomes of two closely related organisms to predict protein-protein interactions. A “template” or design organism provides a network of experimentally derived interactions, and this pattern is used to infer the structure of an interaction network in a related organism. The approach is demonstrated by training a series of SVMs on interaction pairs representing *H. pylori*, and a complete, novel protein-protein network is inferred for a related bacterial organism, *Campylobacter jejuni*. 10-fold cross-validation testing during the training indicated expected upper bounds on precision of 80% and sensitivity of 69% when applied to related organisms. Specific biological examples of two predicted subnetworks of protein-protein are presented and discussed.

*Advantage:* See advantages under the heading *SVM interactions* above.

*Critique:* See critique under the heading *Interacting domain profile pairs* above.

*Predictive scope:* Proteome-wide.

*Related methods:* Interacting domain profile pairs [207], SVM interactions [25, 26].

### **Orthogonal experiments** [186]

*Concept:* The research reported here combines two different experimental approaches with computational prediction to study protein-protein interactions. There are four steps described in this investigation. (1) Libraries of peptides are randomly screened using phage display, identifying consensus sequences for cognate ligands to each peptide recognition module (here, SH3 domains) in yeast. (2) Search the yeast proteome with the consensus sequences as the query sequence. Find potential native ligands to the peptide recognition modules. Create an *in silico* interaction network connecting these SH3 domains to other proteins in the organism of interest which contain cognate ligands. (3) Using yeast two-hybrid screens, derive an experimental protein interaction network, testing 18 different SH3 domain proteins against the proteome represented computationally in the previous step. (4) Find the overlap between the *in silico* and experimental networks.

*Advantage:* (a) Both phage-display and two-hybrid analysis use full genomic information [186]. (b) The experimental approaches are orthogonal; phage-display uses *in*

*vitro* binding and short synthetic peptides, while two-hybrid uses *in vivo* binding and native proteins. This orthogonality removes systematic errors unique to each method [75]. (c) Integrating data from orthogonal sources can identify new relationships, e.g. between gene expression and subcellular localization [75].

*Predictive scope:* Genome-wide.

## E Conclusions

This survey has covered techniques for the prediction of protein function by computer that are wide-ranging in their assumptions, hypotheses, reliability and scope. Our classification of methodologies into two conceptual groupings, “Biological Hypothesis-based” and “Machine Hypothesis-based”, reflects fundamental differences between the two approaches. A quick review of the publication dates listed in Table II.1 indicates an increasing preponderance of machine learning approaches in recent years. This trend is expected to continue, as *in silico* protein functional assignments mature in reliability, and experimental affirmation of biologically relevant predictions improves our understanding of which techniques work (and which do not).

As these methods mature, they will continue to concentrate and drive experimental proteomics technologies. We anticipate the appearance of integrated technology platforms which synthesize machine learning prediction techniques with protein microarrays [122, 213, 118]. Under robotic control, such a system could be used to automate prediction, experimental validation, and subsequent refinement of key parameters of the machine learning hypothesis generators. Perhaps the vision of machine learning enabling “partial automation of every element of scientific method, from hypothesis generation to model construction to decisive experimentation” [135] is not distant in the future.

The long-term scientific objective of describing and understanding intracellular protein function remains a substantial challenge. Complete understanding reaches beyond the plateau of protein interaction networks, to include knowledge of “transcriptional, translational and posttranslational regulation, binding constants, structures, protein interactions and cellular networking” [187]. Moreover, protein interaction networks have a spatiotemporal aspect that must be explained. Details are needed regarding which proteins interact within dynamic complexes, and at what concentrations they interact, given a certain cellular

context [91]. All of these represent great challenges to the discipline of bioinformatics.

## III

# Interactions in a broad database

## A Introduction

The interaction between proteins is fundamental to a broad spectrum of biological functions, including regulation of metabolic pathways, immunologic recognition, DNA replication, progression through the cell cycle, and protein synthesis [2]. Whether or not two proteins will bind to form a stable complex that is prerequisite to biological function is dependent on the three-dimensional conformations of the proteins [97]. For a given conformation, the chemical reactivity of an individual protein is defined by the type and spatial orientation of surface-accessible amino acid side chains. Conformation therefore determines protein-ligand binding. In biology, it is virtually axiomatic that “sequence specifies conformation” [8], suggesting a provocative postulate: knowledge of the amino acid sequence alone might be sufficient to estimate the propensity for two proteins to interact and effect useful biological function.

The science of proteomics endeavors to elucidate the structures, interactions and functions of all of a cell’s or organism’s proteins [9], with the objective of understanding cellular processes and networks and, ultimately, disease processes at the protein level [23]. Current technology for cataloging the proteins contained within a cell involves (1) separation via 2-D gel electrophoresis or liquid phase chromatography, followed by (2) identification using tandem mass spectrometry [54, 175]. Experimental techniques such as two-hybrid screens [65] are often employed to study dynamic interactions between the identified cellular proteins [17, 188]. As such techniques are “tedious, labor-intensive and potentially inaccurate” [63], investigators have recently been prompted to seek computational methods

to predict whether or not two proteins will interact. Previous research groups have presented predictive methodologies based various principles, including correlated changes in amino acid sequence between interacting protein domains [150]; using genomic context to infer functional protein interactions between the gene products [90]; or inference from genome sequences, given observed homologies in other organisms, where interacting proteins have fused into a single protein chain [129, 63]<sup>1</sup>.

Earlier prediction techniques were focused on estimating the *site* of interaction, without reference to specific binding partners. These methods utilized features and properties related to interface topology, solvent accessible surface area (ASA) and hydrophobicity [98], or the recognition of specific residue or geometric motifs [106, 144]. Antigenic determinant sites in proteins were predicted using hydrophilicity profiling methods presented in [87, 204].

In contrast to the cited investigations, the methodology reported herein takes an entirely different approach to computational prediction of protein interactions. Given a database of known protein-protein interaction pairs, a machine learning system is trained to recognize interactions based *solely on primary structure and associated physicochemical properties*. Generalization of results obtained by the system upon introduction of unseen testing sequences is encouraging, given the volume of the data set. Future proteomics studies may benefit from this research by proceeding directly from the automated identification of a cell's gene products to prediction of the protein interaction pairs.

The success of the new methodology is based on the automatic recognition of correlated patterns of sequence and substructure in the interacting pairs. These patterns typically comprise a small number of functional residues in each protein [43].

Complete proteomic functional assignment requires the identification and quantitation of all contributors to dynamic multi-protein complexes. Many molecular signal transduction processes are regulated by the intermediary characteristics of discrete protein recognition "domains", evolutionarily-conserved modules of amino acid sequence found in catalytic proteins, as well as on scaffold, anchoring or adaptor proteins [149]. Protein interactions are frequently mediated by these domains, each of which bind to specific peptides. Such interactions form the basis for structural and functional organization within cells [148].

---

<sup>1</sup>A number of other approaches taken by investigators to predict protein interactions are summarized in Chapter II of this thesis.



Protein domains are often observed across genomes of multiple species [18]. While certain discrete enzymatic signaling domain families are common to all three divisions of cellular life, many non-enzymatic eukaryotic signaling domains with prokaryotic homologues have been identified [158]. Important examples include the SH3 (50 a.a) and PDZ (90 a.a) domains [64, 149]. Other domains organize into larger structural domains or families, subsequently facilitating the assembly and interaction of other proteins. For example: the tetratricopeptide repeat domain (TPR; 34 a.a.) forms a superhelical structure with an amphipathic groove for binding protein targets, and mediates protein-protein interactions [51].  $\beta$ -propeller superstructures are a common motif, comprising, e.g., NHL repeat domains (45 a.a) found on proteins involved in mediating activity of lentiviral Tat proteins *in vivo* [70], and WD40 repeat domains (40 a.a.) on G-proteins, important regulators of a host of cellular functions [142].

## B System and Methods

The protein-protein interaction prediction method is described in this section.

### B.1 Database of interacting proteins

Protein interaction data were obtained from the Database of Interacting Proteins (DIP; <http://dip.doe-mbi.ucla.edu/>). At the time of the original investigation as reported in [25], DIP contained 2,664 records; it currently comprises 18,059 entries representing pairs of proteins known to mutually bind, giving rise to a specific biological function. Each interaction pair contains fields representing accession codes linking to other public protein databases, protein name identification and references to experimental literature underlying the interactions. Alternative fields include protein interaction domains, superfamily identification, interacting residue ranges, and protein-protein complex dissociation constants.

The representation of the various biological superkingdoms in the DIP database is heavily biased towards the Eukaryotes. Table III.1 lists the top 95% most-frequently occurring organisms and their kingdom membership. Note that the budding yeast *Saccharomyces cerevisiae* accounts for 64% of the interactions, which are readily accessible online [188]. The bacterium *Escherichia coli* constitutes the most frequent non-eukaryote proteome, yet

accounts for only 1.3% of the proteins found in the database.

On the molecular level, the protein interaction substructural domain coverage within DIP is diverse. Submitting the protein sequences to the Protein Families Database [18] of protein domains and profile hidden Markov models (Pfam v. 5.5; URL: <http://pfam.wustl.edu/>), we estimated that at least 1,394 distinct domains are represented. Table B.1 lists the most frequent protein domains found in DIP, using a sequence *E*-value cutoff level of 1.0. A histogram portraying the distribution of all protein sequence lengths within the database is presented in Figure III.1. The mean and standard deviation of amino acid chain lengths are 481 and 386 residues, respectively.

<i>Organism</i>	<i>Superkingdom</i>	<i>Frequency</i>
<i>S. cerevisiae</i>	Eukaryota	0.639
<i>H. sapiens</i>	Eukaryota	0.184
<i>Mus musculus</i>	Eukaryota	0.049
<i>D. melanogaster</i>	Eukaryota	0.033
<i>R. norvegicus</i>	Eukaryota	0.020
<i>E. coli</i>	Bacteria	0.013
<i>Bos taurus</i>	Eukaryota	0.012

Table III.1: Organism representation by proteins found in the DIP database, circa January 2001. Frequency expressed as fraction of total number of occurrences of each organism. The top 95% most frequent organisms are listed. Number of interactions  $n = 2,664$ .

## B.2 Support vector machine learning

The new protein-protein interaction estimator utilizes the technique of “support vector” learning, an area of statistical learning theory subject to extensive recent research ([191, 169]). A selection of recent bioinformatic investigations utilizing Support Vector Machine (SVM) learning includes [39, 92] and [214]<sup>2</sup>. Useful for function approximation, signal processing and regression, SVM has several advantages as applied in the present context:

1. SVM generates a representation of the nonlinear mapping from residue sequence to high-dimensional protein feature space [14] using relatively few adjustable model parameters.

<sup>2</sup>Additional references may be found online at <http://www.support-vector.net/bioinformatics.html>.

<i>No.</i>	<i>Domain</i>	<i>Frequency</i>
1	WD40	0.056
2	pkinase	0.030
3	TPR	0.028
4	zf-C2H2	0.018
5	Armadillo_seg	0.016
6	EGF	0.016
7	HLH	0.013
8	spectrin	0.013
9	bZIP	0.011
10	ank	0.011
11	rrm	0.009
12	SH2	0.008
13	SH3	0.008
14	Sm	0.007
15	ras	0.007
16	fn3	0.007
17	PHD	0.006
18	efhand	0.006
19	myb_DNA-binding	0.006
20	arf	0.006

Table III.2: Most frequent protein domains in the interaction dataset. Frequency expressed as fraction of total occurrences of each domain. Prediction using the Protein Families Database (Pfam v. 5.5 [18]).

2. Based on the principle of *structural risk minimization*, SVM provides a principled means to estimate generalization performance via an *analytic* upper bound on the generalization error. This means that a confidence level may be assigned to the prediction, and alleviates problems with overfitting inherent in neural network function approximation [85].
3. SVM is readily adaptable to new data, allowing for continuous model updates in parallel with the continuing growth of biological databases.
4. An additional benefit is the fact that SVM is a *deterministic* algorithm—for a given training data set and SVM configuration, the same test data classification is derived from the solution to the quadratic optimization problem. This provides a means to systematically compare different SVM architectural parameters and protein features. In contrast, *stochastic* classification algorithms by their nature will not necessarily

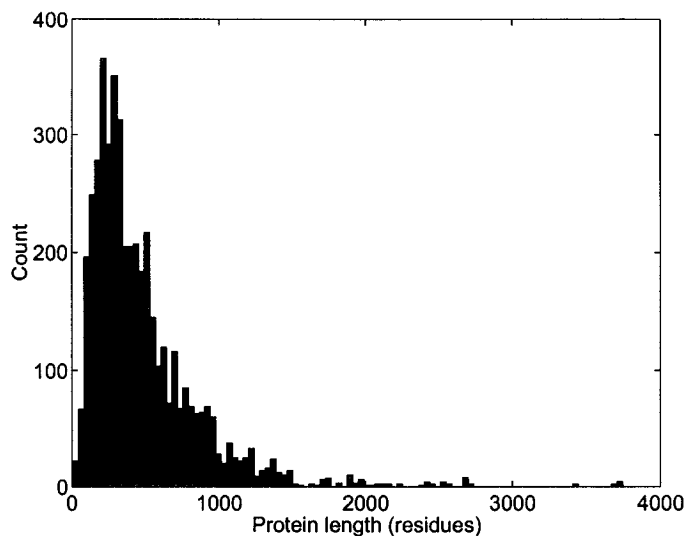


Figure III.1: Distribution of protein sequence lengths in database. At least 1,394 distinct interacting domains are represented.  $\mu = 481 \pm 386$  residues.

produce the same answer on successive processing runs.

In the present research, we train an SVM to recognize pairs of interacting proteins culled from the DIP database. The decision rules developed by the system are then used to generate a discrete, binary decision (“+” $\Rightarrow$  proteins interact; “-” $\Rightarrow$  no interaction) upon the introduction of a new feature set based on primary structure of the putative protein interaction pair.

Appendix B contains a description of the main ideas behind the support vector machine, including a graphical view that may enhance the reader’s intuition of SVM. This appendix also provides a summary of the equations which represent the optimization problem that is solved during SVM training.

### Computational complexity of SVM

There are costs and benefits associated with any given machine learning algorithm [53], and SVM is no exception to this rule. It is known that solving the quadratic programming problem arising during SVM training involves a matrix of dimension equal to the square of the number of training examples  $l$  [95, 103, 157]. For large classification problems, storing this matrix in memory may be prohibitive.

For the protein-protein and protein-ligand interaction problems as posed in this thesis, at most several thousand positive examples are available for any given organism. The number of negative examples is expected to be significantly larger. To make the present SVM approach feasible in terms of computing time, all of the potential negative examples should not be used in training; only a small sample of the negative interaction pairs are used in training a system<sup>3</sup>. The SVM training time for most of the experiments was on the order of tens of minutes on a Pentium IV class personal computer.

As positive protein interaction data accumulate in the experimental literature, or the entire set of (assumed) negative interaction pairs are to be included in the training set, online SVM training algorithms [44] or alternative machine learning techniques might be considered.

It is the mapping onto a high-dimensional feature space that allows SVM to learn patterns of amino acid sequence that are correlated with the potential for interaction between two proteins. Extremely complex interactions between attributes are represented. Here, input vector dimensions were on the order of 300 numbers. Using SVM kernel polynomials of orders 2–5, the SVM decision function is constructed in a huge dimensional feature space, at the cost of dot products in the input space dimensions. Although the *number* of data points may cause a training time bottleneck, the feature space dimension can be effectively infinite [170]. Therefore a trade between time complexity and profundity of representation exists.

### B.3 Feature representation

The problem of feature selection is to find a set of salient attributes to represent the concept that is to be learned [96]. There are literally hundreds of different metrics of residue properties available in the literature that may have been selected to represent amino acid features to the learning algorithm. One listing of the possibilities can be found online at [http://www.genome.ad.jp/dbget/AAindex/list\\_of\\_indices](http://www.genome.ad.jp/dbget/AAindex/list_of_indices). Due to the myriad possibilities, it is infeasible to search for an “optimal” set of numerical features that will produce the best classifier of protein interactions. Even if such a set of features existed, and was identified after exhaustive search, it is arguable whether or not the marginal

---

<sup>3</sup>Cauwenberghs and Poggio [44] describe an online version of the SVM algorithm that in principle could be trained on a comprehensive set of negative interaction pairs. Their algorithm is incremental, retraining on all previous data as each new data point is introduced.

improvement in predictive performance accrued would be worth the significant cost of time and computing resources.

Here, attributes representing residue charge, hydrophobicity, and surface tension were selected. For each amino acid sequence of a protein-protein complex, feature vectors were assembled from encoded representations of tabulated residue properties including charge, hydrophobicity, and surface tension for each residue in sequence. This set of features was motivated by the previous demonstration of sequential hydrophilicity profiles as sensitive descriptors of local interaction sites [87]. This concept was extended presently to integrate sequential charge and surface tension, as water molecules influence atomic packing for shape complementarity, and mediate polar interactions at protein-protein recognition sites [121]. Our postulate is that since sequentially-proximal protein secondary structure elements are often co-located in three-dimensional conformation [120], the sequential profile of these additional features (charge, surface tension) must similarly “co-locate” upon folding.

We found that the numbers used to represent residue properties needed some connection with physical characteristics that would differentiate between similar and dissimilar residues. Before arriving at this set of features, a classification experiment was carried out using only numbers (say, 1–20) to represent each amino acid. The SVM classifiers constructed using these features did not perform any better than a coin-flip, when tested on unseen data points. In retrospect, this behavior can be easily explained by the lack of correlation between features (the integers) and the physicochemical properties of the amino acid sequences they represented. Consider that within amino acid substitution matrices used for sequence alignment, residues of similar biological characteristics have similar numerical values, because in evolution their mutual substitution is more likely to be observed than would a mutation involving biochemically disparate residues [86]. Two examples of conservative (similar) substitutions are isoleucine–valine (small, hydrophobic) and serine–threonine (polar). The same principle applies to the selection of amino acid features for the numerical experiments of this thesis; random assignment of numerical values without regard to physicochemical characteristics is ineffectual.

The reader should be aware that the manner in which protein sequences are represented to the learning machine is not limited to the scheme used in these experiments. Other representations are certainly possible. For example, Hunter and Subramaniam have recently proposed a parsimonious one-dimensional structural description which uses only

a single continuous variable per amino acid to represent the  $C_{\alpha}$  backbone [88]. It would be an interesting experiment to study the use of such structural descriptors in protein interaction predictions, and to compare and contrast their predictive success with that of the physicochemical descriptors used in this thesis.

### **Hydrophobicity index**

Accepting the observation that physicochemical attributes of one form or another are required to carry out predictions discussed here, the selection of specific indices of the features identified for protein representation must be justified.

Hydrophobicity indices measure the relative tendency of a particular residue to interact with water, or the affinity for hydrophobic over hydrophilic phases in a physiological environment. The speculation is that such a metric may provide implicit information to a learning algorithm regarding paired protein conformation, and the propensity for mutual interaction. As an index of amino acid hydrophobicity, we chose to use the consensus normalized hydrophobicity scale of Eisenberg [61]. In that review, the authors identified a number of deficiencies with other contemporary scales found in the literature. Those deficiencies included: (1) a lack of account for side chain interactions, including covalent links to the rest of the protein; (2) a lack of account for all amino acid residues; (3) biased values for certain residues; and (4) the *ad hoc*, subjective adjustment of certain numerical values. Noting that no generally accepted method existed to calculate hydrophobicities, and arguing that it was unrealistic to hope to adequately express “all aspects of the interaction of a residue with water...in a single number”, Eisenberg’s resolution of the difficulties with the different existing scales was to suppress outlying values by producing a combined scale representing the numerical average of four different published indices. Eisenberg’s consensus scale continues to be cited in the literature. For these reasons, this scale was chosen for the present numerical experiments.

### **Surface tension index**

The surface tension scale used for residue features in this research is described in Bull and Breese [40]. In that investigation, the authors used a differential capillary rise experimental technique to accurately measure the surface tension of solutions of each of the amino acids. The surface tension index is actually the slope of a linear surface tension–

concentration relationship, and expresses the reduction in surface tension of the solution as additional amino acid solute is introduced.

Bull and Breese noted the inverse relationship between surface tension and hydrophobicity of the amino acids. In this sense their index may be considered to be an indirect measurement of hydrophobicity; as surface tension is reduced, the hydrophobicity is seen to increase accordingly. Therefore the hydrophobicity and surface tension are not physically orthogonal features. Receiver operating characteristic (ROC) analysis with a yeast data set in Chapter IV appear to suggest that using all three features, charge  $C$ , hydrophobicity  $H$  and surface tension  $T$ , produces a slightly lower performing classifier than simply using charge and surface tension alone.

The rationale used when conceiving the feature sets was that the surface tension index, along with electrical charge, would express properties of the protein *surface*, while the hydrophobicity scale would represent the protein's hydrophobic core in a *volumetric* sense. It is apparent in retrospect that the features  $H$  and  $T$  may present conflicting or partially redundant information, at least for the results presented for yeast in Chapter IV. When using these features together on other data sets, different results may or may not be observed.

#### B.4 Feature vector construction

This section provides a mathematical description of the construction of the feature vectors used to represent interacting proteins in the numerical experiments. Let the vector of numbers  $\{\mathbf{v}\}^i$ ,  $i \in 1, \dots, M$  in  $L$ -dimensional real space  $\mathbb{R}^L$  denote feature  $i$  for a given amino acid sequence of length  $L$  residues, where  $M$  different features are considered. Lengths of the individual feature vectors  $\mathbf{v}$  were normalized by mapping onto a fixed-length interval  $K$ , via  $\{\mathbf{y}_k\}^i = f(\{\mathbf{v}\}^i)$ , where the function  $f$  is  $f: \mathbb{R}^L \rightarrow \mathbb{R}^K$ .

We implemented the mapping  $f$  using simple linear interpolation [110]. An outline of one strategy for doing this is as follows:

1. Discretize the input and output domains:

$$\xi_{in} = (1/L) * \{1, \dots, L\}, \quad 0 \leq \xi_{in} \leq 1$$

$$\xi_{out} = (1/K) * \{1, \dots, K\}, \quad 0 \leq \xi_{out} \leq 1$$

2. For each element of the output domain  $\xi_{out,k}$ , find the indices  $(j, j + 1)$  of the input domain whose corresponding values  $\xi_{in,j}, \xi_{in,j+1}$  "bracket" it:



$$\xi_{in,j} \leq \xi_{out,k} \leq \xi_{in,j+1}, \quad j \in 1, \dots, L; k \in 1, \dots, K$$

3. Estimate the local slope  $m$ :

$$m \approx (v_{in,j+1} - v_{in,j}) / (\xi_{in,j+1} - \xi_{in,j})$$

4. Estimate the value of  $y_{out,k}$  at  $\xi_{out,k}$  by linear interpolation:

$$y_{out,k} = y_{out,k-1} + \{m * (\xi_{out,k} - \xi_{out,k-1})\}$$

Note that this procedure as summarized assumes that  $K < L$ , and should be appropriately modified for the case  $K > L$ . A detailed Java source code listing used to carry out this procedure in the numerical experiments is provided in Appendix B.

In this transformed space, the arc length coordinate  $\xi_{out}$  along the peptide sequence now varies as  $\xi_{out} \in [0, 1]$ , and each vector  $y_{out} \in \mathbb{R}^K$ . This is an essential step for representing proteins of widely varying native length (Figure III.1). The full feature vector for a particular protein  $A$  is constructed by concatenation of each feature sequence  $\mathbf{y}$ . This is written as  $\{\phi_A^+\} = \{\mathbf{y}_k\}^1 \oplus \{\mathbf{y}_k\}^2 \oplus \dots \oplus \{\mathbf{y}_k\}^M$ , where  $\mathbf{a} \oplus \mathbf{b}$  indicates simple concatenation of vectors  $\mathbf{a}$  and  $\mathbf{b}$ . Finally, a representation of an interaction pair,  $\{\phi_{AB}^+\}$  is formed by concatenating the feature vectors for proteins  $A$  and  $B$ , i.e.  $\{\phi_{AB}^+\} = \{\phi_A^+\} \oplus \{\phi_B^+\}$ . The vector  $\{\phi_{AB}^+\}$  becomes a positive training example for the SVM.

Negative examples (putative non-interacting protein pairs) must also be presented to the SVM. In this context, it may be insufficient to merely randomize the residues, a practice commonly carried out to estimate the statistical significance of biological sequence alignments as contrasted against a random control [141, 68]. Since a database of non-interacting proteins was not readily available, we chose to create negative controls by randomizing amino acids sequences sampled from DIP, while preserving both (1) amino acid composition and (2) di- and tri-peptide “ $k$ -let” frequencies [48, 100]. Presumably, where  $k > 1$ , this procedure provides more native-like artificial proteins by conserving higher-order biases. Without performing exhaustive wet experiments to prove the biological inertness of proteins encoded by negative exemplars  $\{\phi_{CD}^-\}$ , thereby proving that in fact proteins  $C$  and  $D$  do *not* interact, this must suffice to design and implement the numerical experiments. Randomized amino acid sequences were generated using Shufflet (URL: <http://www.genetique.uvsq.fr/eivind/shufflet.html>) [48]. A schematic diagram of the process used to create feature vectors for the protein interaction experiments is presented in Figure III.2.

*Note:* In subsequent investigations, our thinking on the creation of negative examples for protein-protein interaction prediction evolved. We no longer advocate the generation of “random” proteins as performed in this chapter. In fact, it is reasonable to assume that all combinations of proteins for an organism that are not represented within an experimental set are indeed negative examples. To illustrate, imagine a bacterial organism, *Bacillus X*, with a proteome of size  $n = 1000$  containing potentially 499,500 unique pairwise interactions. Let us assume that there are 5 interactions per protein, a reasonable number based on several published studies of protein interaction networks in different species (see Table V.7 in Chapter V, and the associated discussion). Then we would expect roughly 5,000 positive interactions, and 494,500 negatives for this organism.

## B.5 Data partitioning

In the experiments reported here, the DIP database entries were sampled at random, and data were partitioned into training and testing sets, at approximately a 1 : 1 ratio. Feature vectors constructed as described in Section B.4 were used as examples for training and testing the prediction system. Testing examples were not exposed to the system during SVM learning. The database is robust in the sense that it represents a compendium of protein interaction data collected from diverse experiments. As noted above, 2,664 different protein domains are represented. There is a negligible probability that the learning system will “learn its own input” (see [15]) on a narrow, highly self-similar set of data examples. This enhances the generalization potential of the trained Support Vector Machine.

## C Implementation

Software methods for parsing the DIP database, control of randomization and sampling of records and sequences, and feature vector creation were developed in Java. A new database was constructed by augmenting the original DIP records. Additional fields added included amino acid sequence data and associated residue features, generated as described in Sections B.3 and B.4.

Support Vector Machine learning was implemented using SVM<sup>light</sup> [95], available for non-commercial use on the World Wide Web at <http://svmlight.joachims.org/>. This numerical implementation was selected because of the large SVM<sup>light</sup> user

community online, and as many publications based upon studies using this code can be found the literature. The code provides analytically-motivated estimates of precision, sensitivity and leave-one-out error [94]. Most importantly, the author of the code was responsive and helpful in clarification of several technical points early in my experience with using SVM<sup>light</sup>. Alternative implementations of support vector learning are listed online at <http://www.kernel-machines.org/>.

Training and testing exemplar data files were developed using a prescribed  $k$ -let frequency ( $k \in [1, 2, 3]$ ) and ensemble sampling size as input parameters to the data preparation software. Each member of the statistical ensemble involved a random sampling of the DIP interacting proteins and newly-created “shuffled” amino acid sequences. A different SVM was trained for each  $k$ -let correlation frequency and experimental trial. The results of these trials were averaged to eliminate potential biases due to chance sampling of the data set.

The performance of each SVM was evaluated using the inductive accuracy on the previously unseen test examples as the performance metric. “Inductive accuracy” is defined here as the percentage of correct protein interaction predictions on the test set, consisting of nearly equal numbers of positive and negative interaction examples.

The main results of the protein-protein interaction predictions are summarized in Table III.3. Each row in the table corresponds to a constant  $k$ -let frequency used to generate the negative training and testing examples. Data in the column headed # *Examples* indicate the average total number of each type of examples for each case. These data have been averaged over an ensemble of 10 statistical trials, a sufficient sample as indicated by the low variance shown in Column 3.

<i>k</i> -let <i>Freq.</i>	# <i>Examples</i> (Train,Test)	<i>Inductive</i> <i>Accuracy</i>
1	(2190,2189)	80.96 ± 1.42 %
2	(2192,2192)	80.19 ± 0.86 %
3	(2203,2195)	80.13 ± 0.89 %

Table III.3: System generalization accuracy summary. “Inductive accuracy” is the percentage of correct protein interaction predictions on test data not previously seen by the system.  $N=10$  trials.

## D Discussion

The inductive accuracy of the learning machines as summarized in Table III.3 is encouraging, given the depth of the DIP database. For each statistical background comprising  $k$ -let orders 1–3, about four out of five protein interactions are correctly estimated by the system. It bears reiteration here that only primary structure data have been used to train the SVM. We submit that some implicit information regarding structural, chemical and biological affinity has been learned by virtue of the feature representation and affirmative labelling of protein interaction pairs. The implications of the results shown in Table III.3 for future proteomics research are intriguing.

However, these results must be interpreted with caution. An important objective is to ascertain the extent to which the present machine learning approach may provide utility to the proteomics community. To make this methodology genuinely useful, we need to generalize from a training set of protein interactions with some degree of confidence. Therefore several important issues must be considered.

### Problem with *accuracy* statistic

The purpose of this study was to demonstrate feasibility of predicting protein-protein binding, by posing interactions in terms of a *classification* problem. To apply this prediction methodology in more realistic proteomics experiments, additional statistics of classification performance should be computed. Prediction accuracy is tightly linked to the natural frequencies of occurrence of each data class. Provost has noted that the use of *accuracy* alone may misrepresent the generalization potential of a classifier, in particular when comparing different classification architectures against one another [159]. Accuracy as a statistic assumes (1) equal misclassification costs (for false positives and false negatives), and (2) a known class distribution in the “target environment”. If one of the classes in a two-class problem is “rare” relative to the other one, it is trivial to produce a highly accurate classifier by simply predicting the majority class. Recall the hypothetical organism *Bacillus X* from the discussion near the end of Section B.4. If 5,000 out of the possible 499,500 pairwise (+) interactions are genuinely found in Nature, a classifier always predicting the negative class (–) has 99% accuracy!

Therefore, presentation of results only in terms of the prediction accuracy may not provide sufficient information to critically evaluate one classifier over another on a given

data set. In subsequent chapters of this thesis, additional statistics are used to explore the prediction performance of SVM classifiers as applied different data sets.

### **Problem with imbalanced data**

Recall that in this investigation, the training and testing data sets were *balanced*—comprising nearly equal numbers of positive and negative examples. The true state of Nature is most likely highly *imbalanced* with respect to the distribution of classes. One strategy for training a binary classifier is to present training examples from both data classes in relative amounts not reflecting their natural frequencies of occurrence. An artificial distribution is concocted for training, because the system must learn to recognize each class independently of its prior probability. If one class is rare compared to the other, the importance of balancing the data is even more pronounced, as the examples representing the majority natural class will dominate and bias the learning process [183].

Balancing the training data solves one problem of statistical bias, while introducing another. Once trained, the classifier is used to make predictions about new data points. Since the training data were not distributed according to their expected natural rates of occurrence, generalization success on data sampled from the true probability distribution may be materially different than that realized during training experiments. For protein-protein interaction network predictions, a huge number of different combinations of proteins are possible; we suspect that most of these are non-interactions.

Essentially, this represents a problem of signal detection in noise. We may make a direct analogy to radar signal detection systems, which are designed to (1) detect objects of interest and (2) extract information from the signal representing this object [174]. Each putative protein interaction pair declared by the present classifier is a “signal detection”, and the information extraction component is represented by our belief (or disbelief) of the correctness of this decision. The problem is this: if the classification system is characterized by a constant false alarm rate (CFAR), when faced with data that has only very few true signals of interest, we may observe a large increase in the sheer number of false alarms (false positive interactions). This is the so-called “needle in a haystack” issue in data mining, and will be addressed further in subsequent discussions of this thesis (Section E.1).

## Species diversity

While the methodology presented here is generally applicable, the proteins in the interaction database predominantly represent eukaryotes, as summarized in Table III.1. This bias may also be manifested in the trained SVM, which may not immediately generalize to bacterial or archaeal organisms, although prokaryotic homologs of many non-enzymatic eukaryotic signaling domains associated with protein-protein interactions have been identified [158]. To identify conserved interactions across species, additional training based on more kingdom-diverse proteomes may be required.

## Effect of $k$ -let order

With reference to the first row of Table III.3, we observe that good predictive accuracy is achieved when amino acid composition alone is preserved during randomization ( $k=1$ ). System performance is not degraded relative to cases  $k=2,3$ . If the results indicated a predictive performance surplus where  $k=1$  (more random), one might have conjectured that the SVM had merely learned to discriminate native interactions from random, non-native proteins here. It is unclear whether this observation is an artifact of the particular bias toward *S. cerevisiae* in the database. A distinct possibility is that the randomized “proteins” were in fact substantially different from the native examples, and this difference was reflected in the high accuracy of the predictions. These questions should be addressed in future research.

## E Conclusion

In conclusion, the prediction methodology reported in this chapter generates a binary decision about potential protein-protein interactions, based only on primary structure and associated physicochemical properties. This suggests the possibility of proceeding directly from the automated identification of a cell’s gene products to inference of the protein interaction pairs, facilitating protein function and cellular signaling pathway identification.

This research represents only an initial step in the automated prediction of protein interactions. The discovery of patterns within respective primary structures of known protein interaction pairs may be subsequently enhanced by using other features (secondary and tertiary structure, binding affinities, etc.) in the learning machine.

With experimental validation, further development along these lines may produce a robust computational screening technique that narrows the range of putative candidate proteins to those exceeding a prescribed threshold probability of interaction.

## **F Acknowledgement**

The text of this chapter, in part or in full, is a reprint of the material as it appears in Joel R. Bock and David A. Gough, "Predicting protein-protein interactions from primary structure", *Bioinformatics* 17(5):455-460, 2001. The dissertation author was the primary author, and the co-author listed in this publication directed and supervised the research which forms the basis for this chapter.

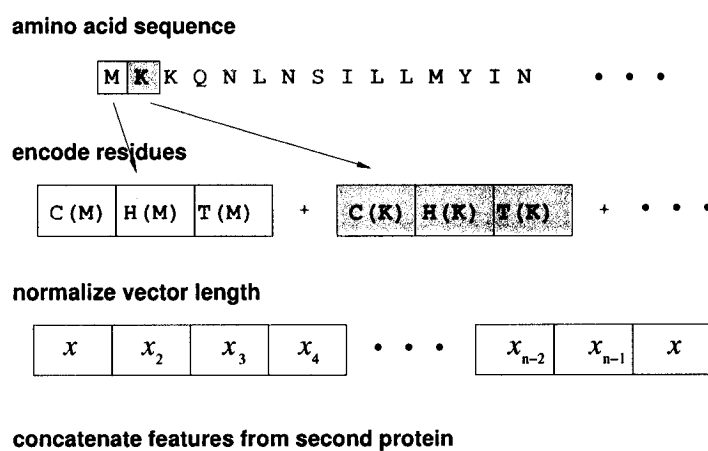


Figure III.2: Steps in the feature vector construction process. For each protein in an interaction pair, residues are encoded as features representing charge ( $C$ ), hydrophobicity ( $H$ ) and surface tension ( $T$ ). These numbers are concatenated in the same order as their appearance in the primary structure of the protein. Next, the length of this array of numbers is normalized to a fixed length. Finally, arrays of features for two proteins are joined to form a feature vector for classification processing.



## IV

# Interactions in one species

## A Introduction

In most cases, proteins perform their biological functions through specific binding with other proteins [1]. Understanding the complex roles of proteins requires experimental identification of interacting protein partners [149], which may necessitate an all-against-all screen of the proteins within a given organism. Even with large-scale automation [127], this represents a monumental undertaking. We describe here a computational method based on machine learning for prediction of equilibrium binding between proteins that may help mitigate this task.

The method makes use of a support vector machine (SVM) [191, 42], a trainable pattern recognition device that learns to classify protein-protein interactions. The present machine learning approach has two procedural steps. First, in the “training” process, experimentally derived examples from a set of proteins of known interaction are introduced to the machine. The machine learns to recognize patterns exemplifying protein interactions within the training set. Second, in “testing” mode, the trained machine is systematically applied to paired proteins of unknown interaction, producing a statistical inference as to their interaction state. In contrast to methods based upon *ab initio* calculations of protein structure, the machine learning approach requires only examples of known interacting proteins and certain biophysical properties of their constituent amino acids. No information about protein conformation is necessary. The method is validated here by 10-fold cross-validation experiments using a large empirical protein interaction data set from the yeast *Saccharomyces cerevisiae* (strain S288C). The results suggest that machine learning prediction of

protein-protein interactions may be an efficient complement to experimental techniques for studying functional proteomics [127].

## B Methods

### B.1 Interaction data set

*S. cerevisiae* was chosen as the test organism for several reasons: its genome was the first to be completely sequenced [77]; it shares a core set of conserved proteins for metabolic processes, protein folding, trafficking and degradation with many higher eukaryotes [46]; and it is an excellent experimental platform for genetic engineering [139] and targeted drug-discovery [140]<sup>1</sup>. The machine learning experiments reported here were based on protein-protein interaction data extracted from a data set of 4,549 interactions determined empirically using a comprehensive two-hybrid (Y2H) assay [91]. The Y2H assay detects the interaction of two candidate proteins (say, *A* and *B*), coupled respectively to distinct functional subunits of a transcriptional activator (GAL4) in yeast. These subunits include a DNA-binding domain and a transcription activation domain. If *A* and *B* physically interact, the resulting complex activates expression of a readily detectable yeast phenotype [65].

The data set used in this investigation was collected by Ito and co-workers [91]. These investigators used a large-scale variation of the two-hybrid technique, known as interaction mating, which exploits the fact that haploid yeast cells of opposite mating type (MAT $\alpha$ , or MAT $\alpha$ ) fuse to form diploids when brought into contact with each other. “Bait” protein *A* (fused to the DNA-binding domain) and “prey” protein *B* (fused to activation domain) are expressed in different haploid strains, each of opposite mating type. The combinatorial mating of these strains and consequent fusion of haploids indicates interaction of the corresponding fused proteins [16].

The originators of the data set used here assert that approximately 95% of the open reading frames (ORFs) in the yeast genome are represented in all possible combinations of DNA-binding and activation domain proteins [91]<sup>2</sup>. The interaction mating pro-

---

<sup>1</sup>Perhaps the most important reason for this choice is the fact that comprehensive protein interaction data for *S. cerevisiae* are readily available for machine learning studies.

<sup>2</sup>The yeast interaction data set is available from Ito and coworkers at <http://genome.c.c.kanazawa-u.ac.jp/Y2H/>.

cedure resulted in a set of 4,549 independent interactions among 3,278 distinct protein interactions after redundancy minimization. The amino acid sequences of these 4,549 positive interacting pairs were compiled from online sequence databases using their respective ORF designations.

## B.2 Redundancy minimization

There was concern that redundant interactions in the database might bias the numerical experiments. To address this concern, redundant or similar interacting pairs in the data set were identified using the Smith-Waterman dynamic programming algorithm [176] and removed prior to the study. This algorithm is based on the extreme value distribution for comparison of pairwise local sequence alignments, and includes commonly used penalty values for gap opening (12) or gap extension (2) in the alignment process<sup>3</sup>. To estimate parameters of the extreme value distribution function representative of the yeast amino acid sequences, tabulated statistics for the probabilities of amino acid substitutions were used that are consistent with an empirical affine insertion/deletion model [4]<sup>4</sup>. Sequences identified as redundant at the 99% significance level ( $p < 0.01$ ) in an all-against-all pairwise similarity analysis were removed. Although it is not known explicitly if the extreme value distribution applies to these data, it has been shown (empirically) that gapped local alignment scores tend to follow an extreme value distribution [4, 177]. Note that our objectives differ fundamentally from sequence database searches where the highest-scoring similarities are desired. Such searches typically use  $p$ -values many orders of magnitude smaller than those used here. The present stringent level of significance was chosen to reduce the sensitivity on the assumed distribution, and minimize the probability that remote similarities between proteins might bias the prediction results.

The details of the method of redundancy minimization are as follows. Let  $S$  denote the set of all amino acid sequences  $s$  that are candidates for participation in protein

<sup>3</sup>The authors showed that gaps of any length can be included in an alignment and still provide a distance metric for the alignment score, provided that the gap penalty increases as a function of the gap length. Assuming that a single mutational event involving a single gap of  $n$  residues is more likely than  $n$  single gaps, to increase the likelihood of such gaps of length  $> n$  being found, the penalty for a gap of length  $n$  is made smaller than the score for  $n$  individual gaps. An associated affine gap penalty  $w(x)$  is a linear function of gap length consisting of a larger gap opening penalty ( $g$ ) and a smaller gap extension penalty ( $r$ ) for each extra position in the gap, or  $wx = g + rx$ , where  $x$  is the length of the gap.

<sup>4</sup>Pointwise amino acid mutation probabilities were modelled using the BLOSUM62 substitution matrix [86]. This matrix assigns a probability score to each position in an alignment based on the frequency with which a given amino acid substitution is known to occur among related proteins.

interactions. We use the notation  $A$  and  $B$  to represent two proteins which come into contact, and may biologically interact (they may be identical, in case of homodimers). Sequence pairs  $(s_A, s_B)$  were first grouped together where multiple sequences  $\{s_B\}$  were identified as interaction partners with a given sequence  $s_A$ . Let this group be represented as  $G = \{(s_A, \{s_B\}_i) | i \in 1, \dots, m\}$ . Each element of the set  $G$  represents a potential protein-protein interaction. There are  $n^+$  such “clusters” in  $\mathcal{S}$ , each corresponding to an unique sequence  $s_A$ . These clusters constitute a set  $\mathcal{G}^+ = \{G_k | k \in 1, \dots, n^+\}$ <sup>5</sup>.

For each element  $G_k$ , the associated sets  $\{s_B\}_i$  were subjected to all-against-all pairwise similarity analysis, and redundant sequences found significant at the 99% level ( $p < 0.01$ ) were removed from further consideration. For each  $s_A$  and all combinations of interacting sequence pairs  $(\{s_B\}_i, \{s_B\}_j), i, j \in 1, \dots, m, i \neq j$  within its cluster, we performed Monte Carlo simulations ( $n = 100$  trials) to estimate the probability that a random rearrangement of amino acids would achieve a score  $Z$  (from sequences  $(\{s_B\}_i, s_{random})$ ) exceeding the score in question  $x$  (from native proteins  $(\{s_B\}_i, \{s_B\}_j)$  using the equation [4]

$$p(Z > x) = 1 - \exp\left\{-Km'n' \exp(-\lambda x)\right\}$$

where  $s_{random}$  is obtained by randomly permuting the amino acids in sequence  $\{s_B\}_j$ .  $p$ -values less than the probability threshold 0.01 were taken as statistically significant, and one sequence from the pair  $(\{s_B\}_i, \{s_B\}_j)$  was eliminated; i.e., the interaction  $(s_A, \{s_B\}_j)$  was excluded from the set  $G_k$ .

After comprehensive forward redundancy elimination, this entire analysis was repeated by a second pass, this time clustering and filtering groups of similar sequences in reverse order, from the set  $G = \{(\{s_A\}_i, s_B) | i \in 1, \dots, m\}$ . In this round, there are  $n^-$  distinct groups to consider (one for each sequence  $s_B$ ), comprising another set  $\mathcal{G}^- = \{G_k | k \in 1, \dots, n^-\}$ .

In this manner, similar interacting sequence pairs detected within the original set of 4,549 ORFs ( $p < 0.01$ ) were removed, leaving a positive example set of 3,011 protein interactions, roughly two-thirds the size of the original data set. Negative examples were derived from the balance of the proteome of *S. cerevisiae*. From a nominal total of 6,408 ORFs, we found 6,360 protein sequences in online databases<sup>6</sup>, which were sampled randomly to construct non-interacting pairs, and designated as “non-interacting” by virtue of

<sup>5</sup>The “+” superscript refers to the forward pass of this analysis.

<sup>6</sup>The online portal for sequence search was the Saccharomyces Genome Database (<http://genome-www.stanford.edu/Saccharomyces>). Each ORF in the yeast genome was used to download

not belonging to the positive set derived from the all-against-all Y2H screen [91]. Positive and negative amino acid interaction sequence sets were validated to ensure mutual disjointness and assembled into separate database files, keyed by corresponding ORF pairs<sup>7</sup>. The final sample contained 3,011 positive and 6,360 negative protein interaction pairs.

### B.3 Protein interaction descriptors

Next, a parsimonious set of features characterizing the design sample was developed, using only amino acid sequence descriptors. Previous investigators established the utility of amino acid sequence hydrophobicity profiles for discriminating local interaction sites on a protein [87]. Other characteristics of protein-protein interfaces have also been studied, including size and shape, electrostatic and surface shape complementarity, and hydrophobicity [97]. Guided by these and other investigations, we used a set of features described previously [25], wherein residue sequences of the proteins in a given interaction pair were encoded with numbers quantifying characteristics of electrical charge  $C$ , hydrophobicity  $H$  and solute surface tension reduction  $T$  for each residue in sequence, preserving the order of the amino acids in each protein. Charge was represented as one of  $\{+1, 0, -1\}$  according to individual residue positive, neutral or negative charge. “Surface tension” constituted an average measure of the reduction in surface tension of an aqueous solution of the amino acid [40]. Hydrophobicity values expressed a “consensus” energy required per mole of amino acid to change phase from hydrophobic to hydrophilic [61]. Taken together, these attributes relate to physical quantities important for intermolecular recognition, namely, three-dimensional conformation in the physiological environment and distribution of surface charge.

Each protein feature vector encoded in this manner was normalized to a fixed length, concatenated with features representing its (positive or negative) interaction partner, and labelled with a binary-valued classification (either +1 or -1) denoting the composite interaction status. Sequence length normalization is imperative to accommodate the wide diversity of protein lengths in *S. cerevisiae*, which range from 25 to 4,910 amino acid

---

its corresponding amino acid sequence from online databases Swiss-Prot/TrEMBL (<http://www.expasy.ch/sprot/>), The Protein Information Resource (<http://pir.georgetown.edu/pirwww/>) and NCBI Entrez Protein (<http://www.ncbi.nlm.nih.gov/entrez>).

<sup>7</sup>The non-redundant positive interactions and negative interaction database files are available on request from the authors. The positive interaction data are based on the original database described in [91]. The current database includes annotations of those interactions excluded from our experiments on the basis of sequence similarity [86].

residues<sup>8</sup>.

The features used to represent the proteins were based purely on characteristics of individual residues constituting the amino acid sequence. This is a clear departure from approaches in recent scientific literature, which have concentrated on secondary or tertiary structural information and physicochemical characteristics of the interacting surfaces [97, 121] to attempt inference of binding propensity. We were inspired by Anfinsen’s thermodynamic hypothesis:

*The native conformation [of a protein] is determined by the totality of inter-atomic interactions and hence by the amino acid sequence [8],*

and questioned whether these residue-based features could be used directly to infer equilibrium binding, in the absence of explicit secondary and tertiary structural information.

Different combinations of features associated with the amino acids were studied. We repeatedly trained and evaluated a collection of predictive systems, using different sets of features  $\mathcal{F}_i$  comprising all combinations of one, two and three physicochemical attributes in concert:

$$\begin{aligned}\mathcal{F}_1 &\subset \{\{C\}; \{H\}; \{T\}\} \\ \mathcal{F}_2 &\subset \{\{C,H\}; \{C,T\}; \{H,T\}\} \\ \mathcal{F}_3 &= \{C,H,T\}\end{aligned}$$

where  $\{*\}$  denotes a particular set of features.

The results reported here are based on numerical predictions obtained using features from the seven distinct sets  $\mathcal{F}_1$ ,  $\mathcal{F}_2$  and  $\mathcal{F}_3$  described above. Objective evaluation of the output of the trained machine was performed to determine the rates of correct and incorrect predictions, and to estimate the generalization error rate upon application of the trained system to inference on other species. The following definitions are essential in interpretation of our results:

1. A “data point” means a pair of protein feature vectors combined with its corresponding truth label (interacting, or not).
2. As only one of two classifications are possible for a data point, we define the “positive” class as denoting a protein-protein interaction in a given example, and the “neg-

---

<sup>8</sup>Source: EBI Proteome Analysis Database (<http://www.ebi.ac.uk/teome/>).

ative” class as meaning a non-interacting protein pair. These definitions apply to both observed and predicted data points.

3. A prediction made on data point assumes one of two states: correct or incorrect. If the computer decision state matches the true state of nature, the prediction is correct; it is false otherwise.

#### B.4 Prediction objectives and metrics

Several practical objectives for computational inference of protein-protein interactions are conceivable. Three cases are noteworthy, each corresponding to different goals, and each suggesting a distinct means by which prediction success should be quantified.

- a. If the objective is to detect all of the possible protein-protein interactions in a given proteome, minimizing the occurrence of false negative predictions (misses) would be important.
- b. If maximizing the correct positive prediction rate (hits) is important (for instance when scrutinizing a large set of potential drug targets to enhance the efficiency of drug discovery), confidence in the affirmative decision that a protein-protein interaction has been detected takes precedence.
- c. In general, it may be preferred to use an overall estimate of the accuracy of the system classification rate, including both positive and negative predictions.

To address each of these cases, we calculated machine learning performance statistics known as “sensitivity”, “precision”, “specificity” and “accuracy”, respectively [109]. For Case (a), the relevant metric is the *sensitivity*, which measures how many actual protein-interactions present in the data are found by the system. Sensitivity is calculated as

$$S = \frac{TP}{TP+FN} \quad (\text{IV.1})$$

where  $TP$  is the number of true positive interaction decisions, and  $FN$  is number of false negative decisions (“misses”). This is alternatively referred to as the “true positive rate” of a classifier. Case (b) calls for the use of the *precision*, which describes the rate at which a positive interaction decision is correct. Precision is computed as

$$P = \frac{TP}{TP+FP} \quad (\text{IV.2})$$

where  $FP$  is the number of false positives declared by the system.

For the general Case (c), the prediction *accuracy* might be considered. Accuracy expresses the general error of the system, computed as  $A = (TP+TN)/(TP+TN+FP+FN)$ . Here,  $TN$  represents the number of true negative classifications. Note that while accuracy provides a broad indication of the prediction performance of a classifier, its use presupposes equal costs of misclassifications (i.e., false positives and false negatives have equivalent negative utility), and that the actual class distributions are known [159]. Each of these assumptions restricts the robustness with which different classifiers may be compared in practice.

A better alternative in Case (c) is to construct a receiver operating characteristic (ROC) curve, a locus of points expressing tradeoffs between the sensitivity and the (1 minus) the specificity, as a function of variation in a detection parameter [156]. *Specificity* conveys the rate at which negative examples in the data are correctly classified, and is equal to  $1 - FP/(FP+TN)$ .

ROC curves are informative because they are based on multiple performance statistics. Baldi [15] stresses that when measuring the performance of classification systems, and using these measurements to infer the ability to generalize to new data, it may be important to present at least two statistics built from at least three elements of the set of numbers  $\{TP, FP, TN, FN\}$ . In his example, imaging that generalization performance statistical results are presented using only, say  $TP$  and  $FP$  (as in the precision, Eq. IV.2). Suppose that we are interested in comparing two different classifiers,  $C_1$  and  $C_2$ , and that each classifier's performance is characterized by different sets of prediction numbers  $r$ , say,  $r_1 = \{TP, FP, TN', FN'\}$  and  $r_2 = \{TP, FP, TN, FN\}$ , respectively. In this case, although the computed rates of precision are identical, the portrayal of relative generalization capability between the two classifiers is misleading and virtually non-informative. Consider the situation where the observed false negative count for  $C_1$  ( $FN'$ ) is significantly larger than that of  $C_2$  ( $FN$ ), while the true negative count are the same; the sensitivity (Eq. IV.1) of  $C_1$  could be much less than that of  $C_2$ . Without considering two statistics together (like sensitivity and precision), this important distinction is oblivious and the value of a classification architecture in specific applications may be seriously misrepresented.



## B.5 Experimental protocol

Using the interacting protein-encoded features (see Section B.3) as input, a series of polynomial-kernel support vector machines [191] were trained to differentiate between pairs of proteins that did (and did not) interact within a biological context. SVMs are optimal hyperplane classifiers that map input data onto a high-dimensional “feature” space. A decision surface is constructed in this feature space, where the classes become linearly separable. This surface is subsequently interrogated to classify new data points, according to their geometric location relative to the surface. During numerical optimization to construct the hyperplane, an implicit mapping of input  $n$ -dimensional data vectors  $x_i$  and  $x_j$  onto a potentially infinite-dimensional feature space is carried out using kernel functions  $K(x_i, x_j)$ , which compute the similarity between  $x_i$  and  $x_j$ . Here, we used polynomial kernels of the form  $K(x_i, x_j) = [(x_i \bullet x_j) + 1]^d$ , where  $d$  is the polynomial order, and the “ $\bullet$ ” operator denotes the vector inner product. Other kernel functions are possible, subject to certain mathematical conditions [42].

We divided a random subsample (2,729 of 3,011 positive, 3,560 of 6,360 negative) of the encoded protein interactions into 10 distinct subsets, allocating positive and negative examples to each subset in approximate proportion to their frequency in the aggregate experimental sample (2,729:3,560, or  $\approx 1:1.3$ ). This sample was used to serially train and test different SVM classifiers, and to estimate their expected generalization error rates, precision and sensitivity by the statistical technique of 10-fold cross-validation [182].

In  $k$ -fold cross-validation, the sample is randomly separated into  $k$  disjoint subsets of nearly equal size. A total of  $k$  different classifiers are generated, one for each subset, with training data comprising  $100 * (k - 1) / k\%$  of the subsets. For each data fold  $k$ , the remaining subset is used to test the system. In this manner, all the available data is used to make predictions. The average classifier error rate over the  $k$ -testing subsets is used to estimate the expected generalization performance. Averaging reduces the variance of this estimate [155]. 10-fold cross-validation has been shown to have low bias, with precision approximating that of leave-one-out error estimation [131].

The experimental cross-validation statistics enable inference regarding how well the trained system will predict novel protein-protein interactions, using amino acid sequences taken from genetically-similar species. The notion of “genetic similarity” may be quantified in a number of ways. One such method relevant here incorporates genomic

content, addition or subtraction of genes, and global similarities between two genomes [185].

## C Discussion

### C.1 Interaction prediction design curves

Figure IV.1 shows the observed 10-fold cross-validation prediction performance as a function of SVM polynomial order  $d$ , for three different feature sets  $\{C,H,T\}$ ,  $\{C,T\}$  and  $\{H,T\}$ . In this graph, colors indicate a particular set of features encoding the protein pairs, and symbols represent certain performance statistics. Precision, accuracy and sensitivity values are displayed. All data points shown are aggregate average results computed from 10-fold cross-validation estimates of the various statistics, collected on the entire sample of 2,729 positive, 3,560 negative interactions in *S. cerevisiae*.

The results summarized in Figure IV.1 show that precision varies directly with the order  $d$  of the SVM kernel, while sensitivity is inversely proportional to the value of this parameter. Precision of the system (marked  $P$ ) is greater than 90% for all feature sets used with this highest dimensioned polynomial kernel (where  $d = 5$ ). This high precision implies an extremely low incidence of false positive decisions. Recalling Case (b) of Section B.4, precision maximization is appropriate when it is important to assign a high degree of confidence in positive predictions. For example, this may have significant benefit as a numerical screening device to select the best subset of a large group of proteins participating in pathways of therapeutic interest.

We were surprised by this result, but the reader is advised to interpret high precision predictions with caution. A central objective of this research is to attempt apply this method to the discovery of protein-protein interactions in other organisms. Estimating generalization performance on the basis of precision results may be ill advised. Reprising the discussion of the “needle in a haystack” problem in the previous chapter, Section D, upon generalization to new species, we may observe a dramatic increase in the rate of false positives if the SVM generalization is characterized by a constant false alarm rate. This theme will be revisited in the discussions of Chapter V.

High precision comes at a cost—namely, a high rate of “false negatives” as represented by the sensitivity curve ( $S$ ), which suggests that only about 36% of the available

true protein-protein interactions have been detected where  $d = 5$ . Figure IV.1 highlights the stark tradeoff between these two metrics, accentuated at the higher values of  $d$ . The low polynomial order ( $d = 2$ ) SVMs are the most sensitive ones, characterized by sensitivities in the 55 – 64% range. The sensitivity is the true positive rate of a classification system. Case (a) under the various application scenarios noted in Section B.4 relates to the objective of pure discovery of novel protein-protein interactions within proteomes. With no data other than a set of amino acid sequences and a trained support vector machine, suppose that it is desired to detect as many interactions as possible, as opposed to establishing confidence in the correctness of a positive system decision. The appropriate performance metric in this instance would be the sensitivity, and SVM architectures correlated with the most sensitive predictions would be implied.

A profound and completely unexpected result is that when using only a single attribute to encode the amino acid sequence (hydrophobicity), the precision of the system decision still exceeds 90% (blue triangle,  $d = 5$ ). The sample standard deviation of this data point over the 10 data partitions was 0.0199 ( $\{H\}$ :  $P = 0.9466 \pm 0.0199$ ), fell within the error obtained from more complex residue feature sets ( $\{C, H, T\}$ :  $P = 0.9563 \pm 0.0124$ ;  $\{C, T\}$ :  $P = 0.9652 \pm 0.0093$ ). The simplest description of the paired amino acid sequence attributes considered, i.e., their hydrophobicity profiles, is found sufficient to produce an extremely precise forecast of the potential for biological interaction.

It is evident that the overall system predictive accuracy rate ( $A$ ) is at least 70% where  $d > 2$ . Accuracy describes how often the system is correct in declaring either a positive or negative protein-protein interaction, as a percentage of the total number of predictions made. While the utility of the accuracy estimate has been largely discounted in discussions of the current and preceding chapters, in the context of other statistics as shown it still conveys some indication of the fact that the desired concept is being learned by the system. Accuracy is seen to exhibit only a very weak dependence on the model order  $d$ .

## C.2 Receiver operating characteristic (ROC) analysis

Receiver operating characteristic (ROC) curves, introduced by Peterson and Birdsall [156], elucidate the relationship between true and false positive rates ( $TPR$ ,  $FPR$ ) characterizing a collection of detection systems, offering insight into the cost (in terms of false alarm probability) of a given rate of sensitivity. Different classifiers may be compared in

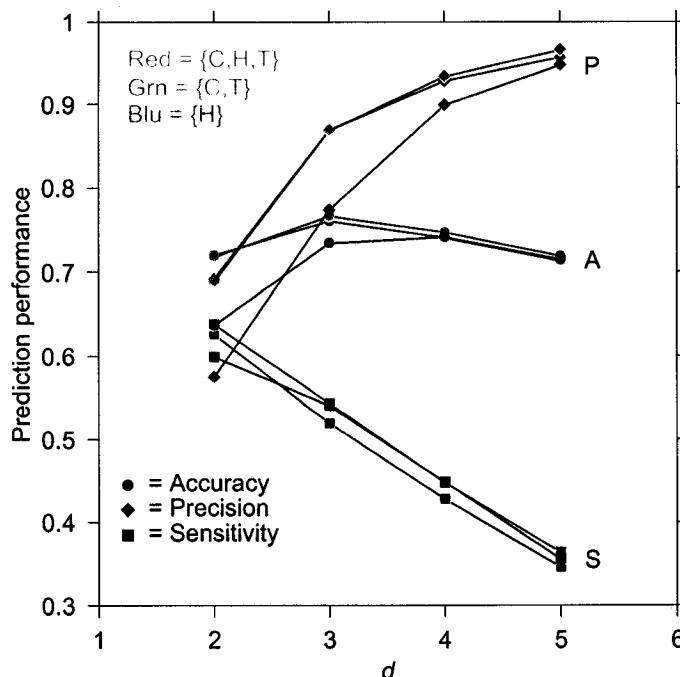


Figure IV.1: Prediction performance of *S. cerevisiae* protein-protein interactions as a function of polynomial kernel order  $d$  for several feature sets. Colors represent different attribute sets used to encode the amino acid sequences of the constituent proteins comprising a biological interaction. *Blue*=hydrophobicity only, *Green*=hydrophobicity and surface tension, *Red*=hydrophobicity, surface tension and electric charge. Symbols correspond to performance metrics: *Circles*=accuracy (*A*), *Triangles*=precision (*P*), and *Boxes*=sensitivity (*S*). Each data point was obtained from 10-fold cross-validation estimates of SVM performance. Total example count:  $n^+ = 2,729$ ,  $n^- = 3,560$ .

ROC space, where each classifier's ROC curve is typically parameterized by a constant value of the detection threshold. The threshold is selected to reject the noise background of varying levels, and its effect is to regulate the trade between precision and sensitivity, as indicated in the discussion surrounding Figure IV.1. Higher detection thresholds decrease the rate of false alarms (increase precision), but cause an increase in the rate of false negatives (decrease sensitivity). Lowering the detection threshold produces converse effects (more detections, more false positives).

We explored a "quasi-ROC" space containing the SVM classifiers of this investigation. This space is related to the ROC curves of signal detection theory in the sense that it provides an analysis of groups of classifiers in ( $FPR, TPR$ ) coordinates. However,

quasi-ROC curves here are actually loci of discrete points, each corresponding to a given SVM polynomial kernel order  $d$ . Whereas each separate curve in classical ROC space represents a different classifier, with the ratio  $FPR/TPR$  varying continuously, progressing in quantum steps along the present “curves” corresponds to changing SVM architectures. In further contrast to conventional ROC analysis, the “curves” connecting the  $d$ -values denote different feature sets used to represent proteins.

The results of this analysis are presented in Figure IV.2. Each curve plotted in the figure corresponds to a different feature set, color-coded to facilitate visual interpretation. These features are indicated in the figure legend. Points along the curves are mapped by varying the SVM polynomial order  $d$ , and recording the  $(FPR, TPR)$  values from 10-fold cross-validation<sup>9</sup>. The area under a particular ROC curve at a given value of the abscissa is an indicator of the accuracy of the associated classifier, now decoupled from *a priori* assumptions about relative distributions of the positive/negative interaction classes or misclassification costs [159].

Several interesting points arise on consideration of the family of ROC curves in Figure IV.2. The underlying trend is that sensitivity and specificity are reciprocally related. One may achieve a sensitive prediction at the risk of introducing false positives, at the rate indicated. As we have seen, such behavior is a fundamental aspect of all signal detection systems.

In this perspective on the different SVM predictors and feature sets subject to numerical experimentation, lower model kernel orders  $d$  produce more sensitive prediction results. This is consistent with results shown in Figure IV.1. The highest values of sensitivity are coupled with false positive rates between 21% ( $\{C, H, T\}$ ;  $d = 2$ ) and 33% ( $\{C\}$ ;  $d = 2$ ).

Notice that the curve corresponding to the features  $\{C, T\}$  (green curve) dominates the ROC space, that is, produces the most sensitive protein interaction predictions at any assumed  $FPR$  value. This suggests that use of parsimonious features in this methodology may be sufficient; more features per amino acid residue are not necessarily correlated with greater acuity of inference. In fact, adding one feature to the descriptor set (red curve;  $\{C, H, T\}$ ) degrades the prediction performance by a small degree. The observation that features  $\{C, T\}$  are apparently more salient than are  $\{C, H, T\}$  is undoubtedly also due in

---

<sup>9</sup>Polynomial order assumes integer values, resulting in piecewise-linear segments for these ROC “curves”.

part to the fact that the hydrophobicity and surface tension are not physically orthogonal features, as foreshadowed in the discussion of Section B.3.

The ROC curves can be seen to coalesce at  $d=5$ , indicating that predictive performance where precision is high (and sensitivity low) is relatively invariant to details of the amino acid characteristics used to represent the interacting proteins.

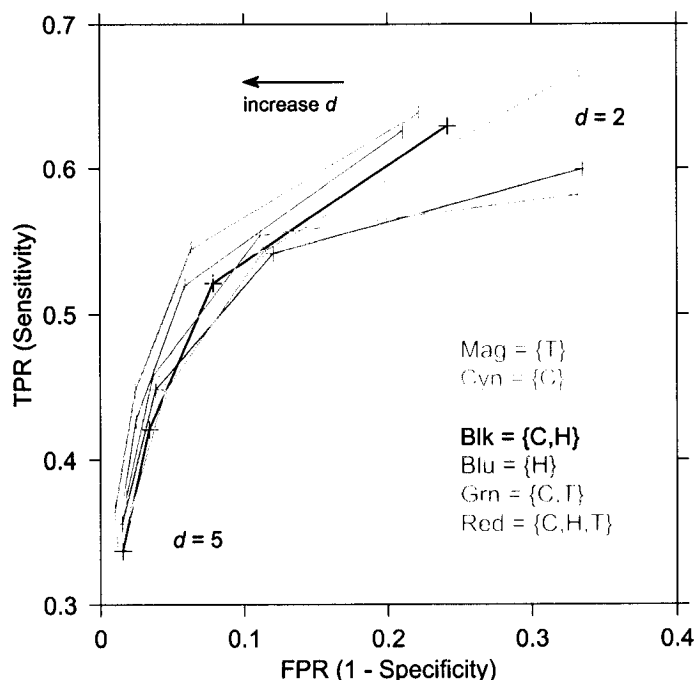


Figure IV.2: Quasi-ROC curves for predictions of protein-protein interactions in *S. cerevisiae*. Colors represent different attribute sets used to encode the amino acid sequences of the constituent proteins comprising a biological interaction. Each data point was obtained from 10-fold cross-validation estimates of SVM performance. The arrow indicates the direction of increasing SVM polynomial kernel order  $d \in \{2, 3, 4, 5\}$ , which corresponds to increasing precision (decreasing sensitivity) as shown in Figure IV.1.

### C.3 Confusion matrices

Figure IV.3 presents a complete accounting of the experimental results obtained using sequence attributes based only on residue charge ( $C$ ) and surface tension ( $T$ ). This corresponds to the best set of features as determined from the ROC analysis of Section C.2.

The data matrices found in this figure, commonly known as “confusion matrices”, may be used to compute other statistics of machine learning performance [109]. Additionally, presentation of data in this manner facilitates the independent reproduction of results, or the comparison of performance with different machine learning techniques. Each confusion matrix in Figure IV.3 represents the 10-fold cross-validation performance for a given SVM polynomial kernel order  $d$ . The columns represent predictions made by the computer, belonging to one of the possible classifications (“+” $\Rightarrow$  proteins interact; “-” $\Rightarrow$  no interaction). The rows represent the true state of nature. Along the main diagonal, the predictions agree with the true class, while off-diagonal elements record the number of examples for which the particular SVM made an incorrect prediction. This error may be either a false positive, or a false negative decision. Note again the concurrent increase of false negatives (reduced sensitivity) and decrease of false positives (increase precision) as SVM model order increases from  $d=2$  (matrix number (i)) to  $d=5$  (matrix number (iv)).

Various performance statistics computed for the two extreme SVM polynomial orders studied are summarized in Table IV.1. These data correspond to the feature set  $\{C, T\}$  found to dominate the ROC space as visualized in Figure IV.2.  $TNR$  is the true negative rate, and  $FNR$  is the false negative rate. Other statistics listed in the table are defined in Sections B.4 and C.2. It is seen that for the SVM with  $d=2$ , the sensitivity is around 64% and the precision 68%. In this context we note again that a constant false alarm rate SVM would exhibit a sharp decline in precision performance when applied to larger, imbalanced data sets. In such cases the classifier with the highest rate of sensitivity would be preferred, as the sensitivity metric, being independent of the rate of false positives, would remain unchanged.

#### C.4 Comparison with previous investigations

The machine learning approach, while using amino acid sequence descriptors, does not predict function based on homology to proteins of known functional class. Therefore it complements so-called “non-homology” based methods of functional attribution. Examples of different non-homology based methods for predicting protein-protein interactions can be found in the literature [207, 50, 153, 151, 172, 129]. Objective comparisons between the present methodology and some of these alternative prediction schemes are problematic. Many investigations offer novel interaction predictions involving hypothetical

		Predicted	
		-	+
Truth	-	2769	791
	+	988	1741

(i)

		Predicted	
		-	+
Truth	-	3334	226
	+	1245	1484

(ii)

		Predicted	
		-	+
Truth	-	3473	87
	+	1508	1221

(iii)

		Predicted	
		-	+
Truth	-	3524	36
	+	1737	992

(iv)

Figure IV.3: Confusion matrices for protein-protein interaction predictions in *S. cerevisiae*. Features based on residue charge and surface tension (corresponding to the green data points in Figure IV.2). Each matrix represents 10-fold cross-validation performance for a given SVM polynomial kernel order  $d$ . Columns and rows marked “+” or “-” indicate interaction or non-interaction, respectively. Off-diagonal elements record the number of examples for which the SVM classifier made an incorrect prediction (a false positive, or a false negative decision). (i):  $d = 2$ ; (ii):  $d = 3$ ; (iii):  $d = 4$ ; (iv):  $d = 5$ . Total example count:  $n^+ = 2,729$ ,  $n^- = 3,560$ .

Statistic	Equation	Value $d = 2$	Value $d = 5$
$TNR$	$TN/(TN + FP)$	0.778	0.990
$S(TPR)$	$TP/(TP + FN)$	0.638	0.363
$FNR$	$FN/(FN + TP)$	0.362	0.636
$FPR$	$FP/(FP + TN)$	0.222	0.010
$P$	$TP/(TP + FP)$	0.687	0.965

Table IV.1: Prediction performance statistics for two different SVM classifiers. Data again correspond to the feature set  $\{C, T\}$  dominating the ROC space of Figure IV.2.  $TNR$  is the true negative rate, and  $FNR$  is the false negative rate. Other statistics are defined in Sections B.4 and C.2. Total example count:  $n^+ = 2,729$ ,  $n^- = 3,560$ .

or putative proteins of unknown function. By definition, such predictions await experimental validation to determine their correctness. Accordingly, sensitivity, precision and



accuracy estimates of the various methodologies are not presented. Other results concentrate on a few key interactions of biological relevance, and estimates of system performance are less specific or absent completely from the discussion.

Even so, some quantitative results are available, permitting limited, direct comparisons with the prediction performance observed in the present study. These are summarized in Table IV.2. In [172], Schwikowski and co-workers describe a means to predict protein function based on relative position within a map of interactions. For proteome-wide predictions within *S. cerevisiae*, they present an estimated “reliability” of 72%, defined as a correct prediction of function for connected proteins, based on a list of its functionally-categorized partners. This is apparently equivalent to an accuracy measurement.

Pazos and Valencia [151] computed interaction predictions based upon the similarity of phylogenetic trees between interacting proteins, taken to indicate the degree of their “coordinated evolution”. In large-scale predictions for *Escherichia coli*, they assert that the tree-similarity approach can be used to “detect true positives at a rate  $> 66\%$ ”, using a particular numerical threshold value representing correlations between proteins in multiple sequence alignments [151]. No further data or equations are available to identify whether this reported detection rate corresponds to the sensitivity statistic (Eq. IV.1).

A hybrid algorithm reported by Marcotte and colleagues [130], again focused on *S. cerevisiae*, cites an overall false positive rate of 33%. When combining predictions from 2 or more different techniques, an *FPR* of 16% was reported. In compiling these statistics, the reliability of an individual “link” between two putative interacting proteins was evaluated by concordance between known functional categories for each protein. Unclassified proteins were exempt from the estimate of reliability. In the present study, false positive rates ranged between 1.0% ( $d=2$ ) and 22% ( $d=5$ ).

## C.5 Key issues

We find it interesting that protein-protein interactions in *S. cerevisiae* may be reliably inferred using only a minimal description of their sequential amino acid characteristics. This technique is exciting in its simplicity and compelling predictive performance, and may prove useful in programs of targeted pharmaceutical discovery, for example where small effector molecules are to be directed at protein interaction networks [77], or in therapeutic disruption of disease-related signal transduction cascades [55].

<i>Ref.</i>	<i>Organism</i>	<i>Statistic</i>	<i>Value(s)</i>	<i>Current value(s)</i>
1	<i>S. cerevisiae</i>	“accuracy”	72%	70%
2	<i>E. coli</i>	“sensitivity”	66%	36–64%
3	<i>S. cerevisiae</i>	<i>FPR</i>	16–33%	1–22%

Table IV.2: Comparison of prediction performance measures between the present investigation (“Current value(s)”) and previous investigations found in the literature. References: 1. [172]; 2. [151]; 3. [130]. Notes: “accuracy” and “sensitivity” are not explicitly described in [172] or [151], respectively. Correspondence with statistics of this investigation is inferred from narrative descriptions in the respective investigations.

There are, however, several important caveats associated with this computational screening approach. The experimental results indicate a high precision, in excess of 90%, may be obtained at the cost of low sensitivity (approximately 36%). This means that while confidence in positive predictions is high, many actual protein-protein interactions are not detected by the system. This may not be acceptable if the objective is to uncover all interactions in a proteome under investigation. The most sensitive predictions were observed for certain architectures, with true positive rates near 64%. This rate of sensitivity may be expected when generalizing results to other, related organisms of interest.

An important question is how to properly specify the “negative” (non-interacting) examples used to train the system. Any predictive scheme for protein-protein interaction inference will be faced with this fundamental question. Public-domain databases of protein interactions comprise only positive examples [91, 209, 12, 161], creating the requirement to define negative examples to train a machine learning system. One possibility is to manufacture proteins *in silico* that mimic native protein characteristics (amino acid composition, and perhaps short segments of contiguous residues) [25]. Another strategy is to assume that an experimental data set is comprehensive in the sense that all protein interactions detectable by the experimental system used in its derivation are represented; all protein pairs not contained within this experimental set are declared as negatives. This approach, taken here, is admittedly naïve; for example, the genomic two-hybrid screening technique for physically interacting proteins [65] has a number of limitations, most significantly a high false positive rate [172]. Further, false negatives due to protein misfoldings or insufficient screening depth were identified as particular difficulties with the comprehensive yeast interaction data subject to the current investigation [91].

Absent definitive information, the naïve approach is, on average, a reasonable approximation as a starting point. The 6,408 ORFs in *S. cerevisiae* translate into more than 20.5 million possible distinct protein-protein interactions, assuming (conservatively) that only one protein is produced per ORF (for a proteome of size  $N$  proteins, there are at least  $(N^2 + N)/2$  distinct interaction pairs.)<sup>10</sup>

The positive interaction data set used here consisted of 4,459 interactions, or about 2% of those possible based on the one gene-one protein scenario. Accordingly, even if the positive interaction set underrepresented the true biological state by as much as a factor of 10, the error rate expected by randomly mislabelling uncharacterized protein interactions as “negative” would be only 11%.

Our approach to minimizing redundant protein interaction pairs (described in Section B.2) was carried out to eliminate bias in the predictor due to similar training and testing examples. To study the effect on predictive acuity without any such filtering, cross-validation experiments were conducted on the complete design sample from [91], intentionally using all available interactions in cross-validation experiments. Here, each of the feature sets  $\mathcal{F}_i$  were scrutinized. As anticipated, prediction success as quantified by the objective measures discussed in Section B.4 was enhanced relative to the nonredundant data set results summarized in Figures IV.1 and IV.2. While both precision and accuracy were characterized by modest improvements (3.9%, 5.2%, respectively), the observed sensitivity rate increased by 20.6%, averaged over all feature sets. This result suggests that an apparent sensitivity rate, if obtained by indiscriminate predictions without redundancy elimination processing, may be significantly overstated.

Our method implicitly assumes a static intracellular state. If proteins  $A$  and  $B$  interact in a design species, say  $S_d$ , the proteome of which is sampled to obtain training data, it is assumed that the same (or a similar) protein pair will also interact when generalizing to a novel species  $S_n$ . This assumption may be invalid if the physiological *milieu* in  $S_n$  is different than that of  $S_d$ . The method and results are only pertinent for simple, binary interactions between physically proximal proteins; dynamically assembling multi-protein complexes cannot be resolved [72]. Post-translational modifications to a protein  $A$  prerequisite to its recognition by protein  $B$  are not identified.

---

<sup>10</sup>Due to alternative RNA splicing, the actual number of proteins produced by a gene is likely to be much higher. Genes with dozen or more transcripts are commonly observed [22]. In one dramatic example, over 38,000 different isoforms of Down syndrome cell adhesion molecule (DSCAM) were observed in *Drosophila melanogaster* [168].

The Y2H *in vivo* assay is predisposed to both false positive and false negative interaction detections [196]. False positives occur from self-transcriptional activation events, or from chance contacts devoid of biological relevance. Two putative interacting proteins may be linked via a third protein located with the yeast nucleus, initiating the transcriptional machinery or the reporter gene. Mutations in bait or prey plasmids may also contribute to false positive results. False negatives might be encountered because of transient or low-affinity binding, insufficient localization to or conditions within the nucleus, or misfoldings of the fusion proteins. One estimate of false negatives for a Y2H system was as high as 45% [195].

Therefore, it is quite possible that the data used to train the system in this investigation may contain incorrect labellings. This fact does not diminish the efficacy of the present approach as demonstrated. As the quality of experimental interaction data improves with technological advances, correctness probabilities of the machine learning predictions will improve in parallel.

## D Conclusions

We conclude that protein-protein interactions can be predicted with computationally efficient machine learning without requiring information about protein conformation. Cross validation experiments on protein interactions within *S. cerevisiae* showed that these predictions can be made at a high rate of precision. However, precision is only obtained the expense of an increase in the rate of false negatives, or reduced sensitivity. This dichotomy must be weighed against the objectives and expectations underlying the application of this system for protein-protein interaction prediction. Certain SVM architectures predicted test data with rates of sensitivity in the 55 – 64% range. This rate of sensitivity may be expected when generalizing to other, related organisms of interest. The precision estimates may not necessarily be applicable, and in fact may be seriously degraded if the observed false positive rate of the trained machine is maintained when generalizing to all possible protein pairs in an organism.

Our results demonstrate that only amino acid sequence and residue properties are required as training information, suggesting some practical advantages of this approach.

# V

## Interactions across species

### A Introduction

The recent publication of the Human Genome Working Draft Sequence [115, 192] is an unequivocal landmark in the advancement of biological knowledge. However, even a completely-sequenced genome presents only a coarse specification for an organism's proteomic complement, and cannot provide understanding of biological function. A major post-genomic scientific and technological pursuit is to describe the exceedingly diverse functions performed by the proteins encoded by the genome. Within the cell, proteins assemble into complex and dynamic macromolecular structures, recognize and degrade foreign molecules, regulate metabolic pathways, control DNA replication and progression through the cell cycle, synthesize other chemical species [2], facilitate molecular recognition, localize and "scaffold" other proteins within signal transduction cascades [149], and participate in other important functions.

To appreciate the role of protein function, a description of protein-protein interactions is a necessary first step. After identifying the proteomic constituents, a rational research strategy should then proceed in the direction of information flow represented by [101]

Interaction  $\rightarrow$  Network  $\rightarrow$  Function

The combinatorial expansion of information advancing along this pathway is enormous. Given the volume of proteomic data generated by high-throughput technologies [189], description of protein function must rely on the integration of empirical data with bioinfor-

matic comparative and predictive analyses.

The workhorse of experimental proteomics has been the two-hybrid screen [65]. Although criticized based on the accuracy of results and its labor-intensive nature [63, 91], it presently stands as the most viable technique for large-scale characterization of protein interactions in complete genomes [119]. Protein chips may eventually provide large-scale simultaneous protein-protein interaction data [122], but technical problems (denaturing, substrate biocompatibility) must be overcome to scale-up for high-throughput analysis. Other approaches will undoubtedly become prominent as proteomics technology continues to evolve. A review of technological advances on this front can be found in [127, 175, 210].

In the meantime, bioinformatics approaches may help bridge the information gap required for inference of protein function.

### **A.1 Bioinformatic approaches to protein-protein interactions**

As discussed in Chapter II, a number of different strategies have been proposed, including network inference based on a reference map of interacting domain profile pairs [207, 206], conserved gene-pairs and correlated prokaryotic interacting gene products [50], clusters of orthologous proteins [184], phylogenetic profile [152] or tree similarity [151], gene fusion events [129], location within a functional cluster map [172], and others. Because investigators concentrate on different organisms, or reporting is confined to partial hypothesized interaction results, it is difficult to compare the predictive power of these various computational methods on an objective basis.

We previously reported a data mining technique [25] wherein a Support Vector Machine (SVM) learning system was trained on a limited, heterogeneous data set to recognize and predict protein interactions based solely on primary structure and associated physicochemical properties. Testing against previously unseen test samples, the system predictive accuracy exceeded 80% over the ensemble of statistical experiments. It was argued that such a system might be used as a screening method to focus experimental assessment of protein interactions. The remarkable success of the methodology reported in [25] has provided motivation for the present work, which is more ambitious in scope. Our present objective is to expand the range of prediction to whole-proteome “interaction mining” using computational statistical learning theory.

Interaction mining uses analogy between the proteomes of two closely related or-

ganisms to predict protein-protein interactions. A “template” or design organism provides a network of experimentally derived interactions, and this pattern is used to infer the structure of an interaction network in a related organism.<sup>1</sup> Given a list of experimental interactions, all that is required to infer the proteome-wide interaction map are the amino acid sequences of the target organism. We refer to this approach as “interaction mining”, in association with the concept of data mining, which concentrates on the application of specific algorithms for extracting structure from data [35].

To demonstrate the approach, we trained a learning system to recognize correlated patterns of primary structure within protein interaction pairs taken from the human gastric bacterium *Helicobacter pylori*, associated with peptic ulcers. A compendium of over 1,200 *H. pylori* interactions were recently reported [161]. This publication is thought to represent the first collective protein interaction map for a human pathogen. The *H. pylori* data are publicly available online at <http://pim.hybrigenics.com>. The provision of this data set for academic research is an important scientific contribution, since few representatives from the Prokaryotes have been widely available to date.<sup>2</sup>

*Helicobacter pylori* interaction data are used to train the system, and to estimate the standard error of its generalization capability. Primary structure data from a close phylogenetic neighbor within the Bacteria Kingdom, *Campylobacter jejuni*, comprise the prediction data set. *C. jejuni* is an enteric pathogen causing common symptoms of food poisoning. Its infection is a precursor to a form of neuromuscular paralysis known as Guillain-Barre syndrome [147]. Both *H. pylori* and *C. jejuni* are microaerophilic, gram-negative, flagellate, spiral bacteria. Analysis of their major constituent protein domains shows a high degree of similarity (see Table V.1). These orthologous bacteria represent model systems for demonstration of the proteome-wide interaction mining approach.

---

<sup>1</sup>After the original submission of this manuscript, the authors were made aware of conceptually similar work reported in [207]. In that investigation, a reference map of interacting protein domains was combined with sequence similarity and clustering analysis to predict a new interaction map in another organism.

<sup>2</sup>In the Database of Interacting Proteins (DIP; [209]) circa November 2002, the most frequent non-eukaryotic proteome (outside of *H. pylori*) is *E. coli*, accounting for only about 1.6% of all interactions found in the database. The current release of DIP has been significantly expanded, now containing over 18,000 protein-protein interactions. See Table V.3.

<i>C. jejuni</i>	<i>H. pylori</i> 26695
IPR000345	<b>IPR001450</b>
<b>IPR001450</b>	<b>IPR001789</b>
<b>IPR001789</b>	<b>IPR005225</b>
<b>IPR005225</b>	IPR001650
<b>IPR003594</b>	<b>IPR002942</b>
IPR001064	<b>IPR003009</b>
<b>IPR003009</b>	<b>IPR003594</b>
<b>IPR000205</b>	IPR000055
<b>IPR000531</b>	<b>IPR000205</b>
<b>IPR002942</b>	IPR000713
IPR004359	<b>IPR001230</b>
<b>IPR001230</b>	IPR002545
IPR002912	IPR004087
<b>IPR004161</b>	<b>IPR004161</b>
IPR000063	<b>IPR000531</b>

Table V.1: Comparison of 15 most-frequently observed protein domains in *C. jejuni* and *H. pylori* strain 26695, indexed by InterPro accession number. Comparisons of each organism are made relative to the InterPro database [99]. Frequent domains common to both proteomes are shown highlighted in boldface.

## B System and methods

The Support Vector Machine [191, 42] can be trained to classify labelled empirical data points by constructing an optimal high-dimensional decision surface that simultaneously maximizes the separation between data classes, and minimizes the “structural risk”

$$R(\alpha) = \int_Z Q(z, \alpha) dF(z), \quad \alpha \in \Lambda \quad (\text{V.1})$$

with respect to parameters  $\alpha$  using an independent, identically distributed (i.i.d.) sample  $Z = \{z_1, z_2, \dots, z_l\}$  generated by an (unknown) underlying probability distribution  $F$ , where  $Q$  is an indicator function, and  $\Lambda$  is a set of parameters.

The sample points  $z_i = (x_i, y_i)$  comprise protein features  $x_i \in \mathbb{R}^n$  and their classifications  $y_i \in \{-1, +1\}$ . In practice, the learning task converges rapidly as a constrained quadratic programming is solved. The resultant decision function  $h$  represents an hypothesis generator for inference on novel data points, mapping them onto the discrete set  $y$ , or  $h : x \rightarrow y$ . This is a binary decision ( $+1 \Rightarrow$  interaction,  $-1 \Rightarrow$  no interaction).



## B.1 Phylogenetic bootstrap

In previous work [25], we trained an SVM to recognize pairs of interacting proteins found in the publicly-accessible Database of Interacting Proteins (DIP) [209]. The system learned to predict interactions of previously-unseen protein pairs. Results presented therein indicated that inference made on a protein-protein interaction data set might facilitate the functional annotation of uncharacterized proteins (where significant homologies to other proteins of known function do not exist). While heterogeneous in terms of the number of different conserved protein domains represented, the distribution of organisms found in the database at the time of the cited investigation was overwhelmingly biased towards Eukaryotes, in particular the yeast *Saccharomyces cerevisiae*.

Building on previous work [25], we propose that the support vector machine-learning approach may be used to extrapolate from a protein interaction map in one organism to a complete map in a related organism, for which only the proteomic sequences have been identified.

Let us establish a framework for prediction of whole-proteome interaction maps. The assumption in Eq. V.1 of a fixed generative probability distribution  $F(Z)$  is a key issue in the design of this data mining application. A direct consequence of this assumption is that a decision function  $h$ , developed from a training sample  $Z_a$  taken from species  $S_a$ , may be used to predict protein-protein interactions on a sample  $Z_b$  from another species  $S_b$ , provided that features of their respective proteomes are not too dissimilar in some sense, or

$$\rho(F(Z_a), F(Z_b)) \leq \delta \quad (\text{V.2})$$

where  $\rho$  is a measure of distance between its arguments, and  $\delta$  is a constant. The statistic  $\rho$  is general, and may be taken to signify cross-species similarity based on genome-level “edit distance” [165], whole-proteomic content [185], or proximity within phylogenies constructed from multi-domain orthologous protein sequences [37], to cite only three of many possibilities. For this discussion, it is assumed that  $\delta$  varies as  $0 \leq \delta < \infty$ , where  $\delta = 0$  is a proteome’s self-distance, and extreme mutual divergence between two organisms is expressed in the limit as  $\delta \rightarrow \infty$ .

We introduce here the *phylogenetic bootstrap* algorithm. Bootstrap methods in applied statistical inference are numerical techniques for estimating the standard error of arbitrary test statistics [57]. The phylogenetic bootstrap for protein-protein interaction mining

does not compute a statistic *per se*, but suggests a method for incrementally “walking” laterally across a phenogram, interleaving rounds of computation and experiment, to develop a knowledge base of protein-protein interactions in genetically related organisms. Using the hypothesis  $h : x \rightarrow y$  (based on an assumed common probability distribution  $F(Z)$ ), we infer the interactions within a sample taken from a distinct, evolutionarily similar proteome. These predictions are a function of the generalization confidence level derived from 10-fold cross-validation error estimation [182]. The probability of correctness of a novel prediction may be estimated by

$$\Pr\{\hat{y} = y \mid h\} = g(\delta)(1 - \epsilon_{cv}) \quad (\text{V.3})$$

where  $\hat{y}$  is the predicted interaction for a putative interacting protein pair,  $y$  is the true state of nature,  $\epsilon_{cv}$  is the cross-validation error rate, and  $g(\delta)$  is a decreasing function of the interproteomic distance (Eq. V.2). A simple plausible (and conservative) form for the function  $g$  is an exponential

$$g(\delta) = e^{-\lambda\delta} \quad (\text{V.4})$$

where  $\lambda$  is the rate of decay. Substituting this function in Eq. V.3, the prediction confidence becomes

$$\Pr\{\hat{y} = y \mid h\} = e^{-\lambda\delta}(1 - \epsilon_{cv}), \quad \lambda > 0, \delta \in [0, \infty) \quad (\text{V.5})$$

Note that this representation is schematic. The value of the decay parameter  $\lambda$  and calibration of the distance in Eq. V.2 can only be determined after experimental validation of the numerical predictions.

Upon completion of this process, predicted protein-protein interactions in the novel organism may be used to design successive genetic or biochemical experiments. The results of these selected experiments are fed-back to refine the current model, and flesh out empirical protein interactions within the new proteome. This iterative process may continue as long as certain criteria on acceptable estimated prediction error rate and proteome similarity remain satisfied. The steps comprising the phylogenetic bootstrap as proposed in this investigation may be distilled into an algorithm, described. This algorithm is described in greater mathematical detail in Section C.

## B.2 Generalization potential

We estimate the expected value of the error rate of the classifier  $h(\alpha, x)$  using  $k$ -fold cross-validation on the training sample  $Z_a$ . Here, we take  $k = 10$ , producing a 10-

fold cross-validation prediction error estimate. The expected generalization error is taken as the average of the classification error observed on each of the  $k$  data folds. Averaging reduces the variance of this estimate [155]. The prediction error derived from 10-fold cross-validation is known to have low bias, and precision approximating that of leave-one-out error estimation, at lower computational cost [131].

In this procedure, an SVM decision rule  $h(\alpha, x)$  is constructed  $k$  times, each time training on a different set of example data points  $\{Z_m \mid Z_m \subset Z_a, m \in 1, \dots, (k-1)\}$ , and testing prediction accuracy on the omitted set  $\{Z_n \mid Z_n \subset Z_a, n \neq m\}$ , where  $Z_m \cup Z_n = Z_a$ . The number of prediction errors for each model is accumulated, and the  $k$ -averaged expected value of the individual data sets' inferred classifiers is taken as the system error rate estimate  $\epsilon_{cv}$ . Note that the statistic  $\epsilon_{cv}$  is an estimate of the expected prediction error rate, and is itself a random function of population, the sample taken from that population, and the inference method. [131].

“Prediction accuracy” as used here means that a correct declaration is made by the decision rule, or  $\hat{y} = y \mid h$ . This can represent either a positive or a negative predicted protein interaction. If the cross-validation error rate is expressed as a fraction assuming values  $0 \leq \epsilon_{cv} \leq 1.0$ , the confidence level expected for predictions of putative protein-protein interactions is given by the probability expression of Eqs. V.3-V.5.

## C Algorithm

The phylogenetic bootstrap algorithm is summarized in this section.

1. *Input.* First, it is necessary to specify the species  $S_a, S_b$  subject to investigation. In general, some existing protein interaction data may be at hand for each proteome, although their relative cardinality may be quite skewed. Our line of thought assumes that no interaction data are available for  $S_b$ ; we have only a set of labels  $\{Y_a\}$  corresponding to experimentally verified interactions sampled from the proteome of species  $S_a$ . These labels, along with the amino acid sequence sets  $\{s_a\}$  and  $\{s_b\}$  comprising the species respective proteomes, are inputs to the algorithm. Other inputs required are the inter-proteome distance  $\delta$  (Eq. V.2), and the maximum acceptable rate of generalization error,  $\epsilon_{cv}^{max}$ , where  $0 < \epsilon_{cv}^{max} < 0.5$ .
2. *Construct features from training sample,* based on attributes of the primary structure

sequences  $s_a$  from the training data set. Encoded attributes  $X_a$  for entire proteomes may be derived from tabulated residue properties including charge, hydrophobicity, and surface tension as described previously [25]. At this stage, data preprocessing including normalization and filtering should be performed to produce a useful sampled attribute set  $\{x|x \in \mathbb{R}^n, x \subset X\}$ . A total of  $l$  data points  $z$  are constructed by adding labels  $y$  to the accepted feature vectors  $\{x\}$ , or  $z_i = (x_i, y_i), i = 1, \dots, l$ . The union of positively- and negatively-labelled examples constitutes the training sample  $\{Z_a\}$ .

3. *Compute decision rule.* Design an optimal support vector machine to classify data points in the sample  $\{Z_a\}$ . After learning, the system builds a decision rule  $h$  that maps input data vectors  $x_i$  onto the classification space  $y_i \in [+1, -1]$ . The numerical sign of  $y_i$  is interpreted as the likelihood that the two proteins represented by  $x_i$  will interact.
4. *Estimate CV error.* Perform  $k$ -fold cross-validation experiments on the training set. Segregate the observations  $\{z^k\}$  within each data fold  $k$ , and train a different SVM using data  $\{z^m\}$  from each of the  $k-1$  disjoint data folds  $\{z^m|z^m \in Z_a, m \neq k\}$ . Predict the class membership of the omitted points  $\{z^k\}$ . Accumulate the total number of misclassifications observed in this process. Take the final  $k$ -fold average cross-validation error as the estimated expectation of generalization error rate  $\epsilon_{cv}$  of the learner  $h$ . The magnitude of this error estimate in practice will be extended by some function of inter-proteomic distance, say  $g(\delta)$ .
5. *Construct features from novel sample.* Construct features  $\{X_b\}$  from sequences  $\{s_b\}$  for the unlabelled proteome  $S_b$ . All-vs-all pairwise interactions may be represented in the prediction set. The same data preparation process should be applied as carried out in Step 1.
6. *Predict novel interaction network.* Predict a new network of protein-protein interactions  $\{\hat{Y}_b\}$  via the trained system  $h(\alpha) : x_b \rightarrow Y_b$ , where  $\alpha$  are parameters of the model. To the extent that the assumption of proteomic similarity  $\rho(F(Z_a), F(Z_b)) < \delta$  is satisfied, each point estimate is expected to be accurate with a probability  $g(\delta)(1 - \epsilon_{cv})$ , or  $\Pr\{\hat{y} = y | h\} = g(\delta)(1 - \epsilon_{cv})$ .
7. *Validate sample experimentally.* Take a random sample from the protein interaction prediction set  $Z_b = \{(x, \hat{y})|x \subset X_b, \hat{y} \subset Y_b\}$  and verify the predicted protein interac-

tions (both positive and negative) using experimental proteomics techniques. Compare the experimentally observed and calculated estimated prediction error rates. Assert that the following statement holds true:  $\epsilon_{cv}^v \leq \epsilon_{cv} < \epsilon_{cv}^{max}$ , where the superscript  $v$  denotes validation by biological experiment.

8. *Input.* Select sequences  $\{s_c\}$  from a new, related organism  $\{S_c\}$ . The similarity assumption  $\rho(F(Z_a), F(Z_b)) < \delta$  must still be maintained.
9. *Update training sample.* Add sequences from the validated prediction set to the training set, and consider this expanded set as the training set for the next iteration:  $\{s_a\} = \{s_a\} + \{s_b\}$ . Update the class labels by adding the prediction label set  $\{Y_a\} = \{Y_a\} + \{\hat{Y}_b\}$ . Protein interactions for organism  $\{S_c\}$  can now be computed.
10. *Iterate.* Return to Step 1 and repeat the process. The stopping condition for this iteration is violation at any time of the assertions regarding the generalization error rate, i.e. when the error rate from cross-validation,  $\epsilon_{cv}$ , exceeds the specified limit  $\epsilon_{cv}^{max}$ , or when the experimental observations contain more frequent errors than the calculated rate, or  $\epsilon_{cv}^v \geq \epsilon_{cv}$ .

## D Implementation

### D.1 Generalization potential

A fundamental premise of our methodology for whole-proteome interaction mining is that a learning system trained on a finite set of attributes from species  $S_a$  may be used to predict protein interactions in a different species  $S_b$  (see discussion in Section B). If experimental protein-protein interaction data are only available for  $S_a$ , how can we assign confidence in the predictions made for  $S_b$ ? The “No Free Lunch” theorems introduced in [208] state that any learning algorithm can be expected to perform “only as well as the knowledge concerning the cost function put into the cost algorithm”. In the notation used here (Eq. V.1), the cost function is the risk  $R(\alpha)$  used to derive the soft-margin classifier  $h$  in terms of the distribution  $F$  based on the salient features  $Z$ , and the cost algorithm is the SVM optimization procedure.

**INPUT** Proteome sequences  $s_a, s_b$ , labels  $Y_a$   
**INPUT** Parameters  $\delta, \lambda, \epsilon_{cv}^{max}$   
**ASSUME** Similarity  $\rho(F(Z_a), F(Z_b)) \leq \delta$  (Eq. V.2)  
**COMPUTE** feature set  $X_a$ , sample  $Z_a$   
 $X_a \leftarrow \text{getFeatures}(s_a)$   
 $Z_a^+ \leftarrow \{(x, y) \mid x \subset X_a, y \subset Y_a, y = +1\}$   
 $Z_a^- \leftarrow \{(x, y) \mid x \subset X_a, y \subset Y_a, y = -1\}$   
 $Z_a \leftarrow Z_a^+ \cup Z_a^-$   
**COMPUTE** decision rule on sample  
 $h(\alpha, x) \leftarrow \text{SVM}(Z_a)$   
**COMPUTE** C.V. generalization error estimate  
 $\epsilon_{cv} \leftarrow \text{CV}(\{h\})$   
 $\Pr\{\hat{y} = y \mid h\} = g(\delta)(1 - \epsilon_{cv})$  **ASSERT**  $\epsilon_{cv} \leq \epsilon_{cv}^{max}$ ?  
**COMPUTE** feature set  $X_b$   
 $X_b \leftarrow \text{getFeatures}(s_b)$   
**COMPUTE** predicted interactions  
 $\hat{Y}_b \leftarrow h(\alpha, X_b)$   
**ASSERT** validate sample experimentally  
 $Z_b \leftarrow \{(x, \hat{y}) \mid x \subset X_b, \hat{y} \subset \hat{Y}_b\}$   
**ASSERT**  $\epsilon_{cv}^v \leq \epsilon_{cv}$ ?  
**INPUT** new proteome sequences  $s_c$   
**UPDATE**  $s_a, s_b$ , labels  $Y_a$   
 $s_a \leftarrow s_a + s_b; Y_a \leftarrow Y_a + \hat{Y}_b; s_b \leftarrow s_c$   
**GOTO** Step 1; iterate while  $\epsilon_{cv}^v \leq \epsilon_{cv} \leq \epsilon_{cv}^{max}$

Figure V.1: Phylogenetic bootstrap algorithm.

*No Free Lunch* implies that in order to make meaningful generalizations, it is essential that all priors or assumptions about the data and classifier be applicable to the prediction data set. In other words, a generic “black-box” generalization machine cannot perform any better than a random guess on other priors. In terms of the present research, the strongest assumption advanced is that of proteomic similarity,  $\rho(F(Z_a), F(Z_b)) \leq \delta$  (Eq. V.2). Other underlying assumptions are stated explicitly in Section D.4 below.

## D.2 Primary structure features

Our objective is to gain insight into protein interactions, if possible using strictly amino acid sequence information. To teach a learning machine, it is necessary to portray salient aspects of the data (the “features”) that intuition or hypotheses suggest will contribute to effective learning of the concept. The problem of feature selection is to define descriptors which discriminate between two classes of data, while inhibiting the irrelevant and redundant features [124]. Here, we sought to find the interacting protein pairs within a complete proteome, for which experimental data representing a negligible percentage of the total possible pairwise interactions are available. We built feature vectors for SVM training as described previously [25], using native proteins directly sampled from the proteome of *Helicobacter pylori*. The protein interaction data were obtained from the online resource as described in Section B. Construction of the negative examples was carried out following Assumption 2 (see Section D.4), which maintains that any pair of proteins not labelled as mutually interacting in the design sample  $Z$  are assumed to not interact. This represents another strong assumption: we assume that the *H. pylori* design sample reported in [161] is complete in the sense that all possible protein-protein interactions comprising the proteome were discovered. Non-interacting protein pairs are designated as negative interactions. In the absence of further information, we must make this assumption, cognizant that by labelling the sample in this manner we may inadvertently commit a logical fallacy of *argumentum ad ignorantiam* (argument from ignorance).

## D.3 Proteome data quality control

Protein interaction examples are filtered to ensure high-quality representation in the learning machine. In Step 1 of the phylogenetic bootstrap algorithm (cf. Section C), data preprocessing is performed. This preprocessing typically includes (1) scaling the fea-

ture vectors to equalize relative numerical magnitudes of the disparate features, and may be followed by (2) curation based on predefined criteria or prior knowledge impacting confidence in the data set. Scaling techniques are well-documented in the machine learning literature, and will not be further discussed here (a succinct summary for applications can be found in [183].)

With regard to the second cited aspect of preprocessing, we selected only positive samples for *H. pylori* interactions where the estimated probability that the observed interaction was found purely by chance (as a two-hybrid artifact) was at most  $1.0E-6$ . In this case the originators of the data set assigned degrees of confidence to the various interactions comprising the sample, according to a model of competition for bait-binding between prey fragments [161].

Commonly, a large percentage of the open reading frames (ORFs) in a given genome remain experimentally unobserved, and if sequential homology to a protein of known function is not discovered, these proteins are labelled as “hypothetical”.

Further complications include the lack of solubility and/or native conformational stability of newly-expressed proteins. In a wide-ranging study of current structure-determination technologies [47], investigators began with 1,871 ORFs from the thermophilic archaeon *Methanobacterium thermoautotropicum*. After exclusion of membrane-bound proteins and others with clear structural homologs, 424 ORFs were chosen for cloning, expression and structural analysis. Experimental observations indicated that over 50% of the proteins studied were either insoluble or misfolded. It has been suggested that using such proteins in biochemical assays will contribute to false positive or false negative results [56].

In light of these facts, the machine learning investigator might be tempted to consider excluding such sequences from the design sample. An overriding argument against such action is the recognition of the fundamental objective of assigning functional roles to the so-called “hypothetical” protein sequences. Consequently, a concession must be made to incorporate possible numerical artifacts, learned from experimental data which may be fraught with false positive and false negative interaction data. As structural proteomics continues to fill in the gaps in our knowledge in the future, these hypothetical proteins will eventually be confirmed or invalidated experimentally.



## D.4 Assumptions

Interaction mining analysis makes certain assumptions about the distributions of proteomic data in the design sample  $Z$  (recall discussions in the context of Eq. V.2). Other assumptions inherent in this approach include [26]:

1. *Static intracellular state.* If proteins  $A$  and  $B$  interact in the design species  $S_d$ , they will also interact if co-occurring in a novel species  $S_n$ . This assumption may not be generally valid for where physiological conditions present in  $S_n$  differ relative to  $S_d$ .

2. *Coverage of design sample.* Any pair of proteins ( $A, B$ ) not labelled as interactors in the design sample  $Z$  are assumed to not interact. This is a subtle but significant point that must be held in mind when interpreting prediction results.

3. *Physical proximity.* The all-vs.-all interacting mining technique selects interaction pairs based on correlated patterns of primary structure, and does not discriminate protein subcellular location. In particular cases, additional information regarding subcellular location might offer insight regarding prediction practicability. Such analysis could be done in a separate post-mining filtering step.

4. *Simple interactions.* Only binary interactions are represented; complexes of proteins with more than two components are only inferred indirectly in post-mining analysis. Dynamic multiprotein complexes [72] are not directly resolved (but, may be inferred after the fact, with details of each component protein's interaction surface characteristics [66]). Also, pairwise interactions predicated upon modifications to protein  $A$  (e.g., phosphorylation, glycosylation, proteolytic cleavage) prerequisite to its recognition by  $B$  are excluded from the prediction space.

## E Discussion

For the design organism *Helicobacter pylori* strain 26695, a total of 1,039 protein interactions were selected for analysis. Interactions were identified from the database provided online at <http://pim.hybrigenics.com>. From the nominal *H. pylori* proteomic complement of  $N = 1,555$  sequences, a sample of 1,039 non-interacting sequences was selected according to the various data filtering procedures described in Section D, and following the assumption of comprehensive coverage in the positive design sample (Section D.4 ). This created a balanced representation of each data class to train the

learning system, the total sample length being  $l=2,078$  observations. Each sample point  $z_i = (x_i, y_i)$ ,  $i = 1, \dots, l$  was constructed from primary structure features  $x_i \in \mathbb{R}^n$  and their interaction class labels  $y_i \in \{-1, +1\}$  (see Section B).

### E.1 Cross validation estimates from *H. pylori*

The learning machine generates an interaction hypothesis  $\hat{y}$  for each data point  $x$  via the computed decision surface  $h : x \rightarrow y$ . Define the null  $H_0$  and alternative hypotheses  $H_A$  as:

$$\begin{aligned} H_0: & y | x = -1, \text{ (no interaction),} \\ H_A: & y | x = +1, \text{ (interaction present)} \end{aligned}$$

There are two types of statistical errors that may occur on each decision  $\hat{y}$ . (1) If  $H_0$  is true and is rejected ( $\hat{y} = +1, y = -1$ ), the machine commits a Type I error, or “false positive” decision. (2) If  $H_0$  is false (interaction present) and is not rejected ( $\hat{y} = -1, y = +1$ ), a Type II, or “false negative” error, is made.

The 10-fold cross-validation prediction error estimates obtained on the design sample are presented in Table V.2. Results are shown for three conventional statistical instruments used to evaluate the performance of classifiers in machine learning applications<sup>3</sup>. These include the *sensitivity*, *precision* and *accuracy* [109]. Sensitivity is calculated as  $S = TP / (TP + FN)$ , where  $TP$  = number of true positive interaction decisions, and  $FN$  = number of Type II errors. Precision is computed as  $P = TP / (TP + FP)$ , where  $FP$  is the number of Type I errors made by the system. Accuracy expresses an overall correctness rate of the system, and is computed as  $A = (TP + TN) / (TP + TN + FP + FN)$ . Here,  $TN$  represents the number of true negative classifications.

The cross-validation measurements summarized in Table V.2 are comparable to previously published predictive results [25]. On average, three of four SVM predictions were correct when applied to the unseen data partition. The precision was 80%, a result which seems to suggest a strong level of confidence in positive interactions detected by the system. Precision indicates the rate of Type I error suppression. Sensitivity observed

<sup>3</sup>As Baldi has pointed out [15], it is important to present multiple statistics of predictive performance. If the performance statistics are constructed from elements of the set  $\{TP, TN, FP, FN\}$ , a high bias is possible if a single statistic is presented using only two of the statistics comprising this set.

<i>Precision</i>	<i>Sensitivity</i>	<i>Accuracy</i>
80.2	68.6	75.8

Table V.2: 10-fold cross-validation performance estimate derived from classifiers trained on examples from the design organism *H. pylori*. High precision indicates the suppression of Type I (false positive) errors. High sensitivity means that Type II errors are suppressed by the decision function (i.e., low false negative rate). Numbers are expressed as percentages. Data sample size  $n = 2,078$ .

in cross-validation was 69%, indicating the true positive rate expected for inference about a different organism. The sensitivity measures the rate of Type II (false negative) error suppression in a classifier. We cannot necessarily directly apply the precision statistic to estimate the generalization performance on novel organisms; the reasons for this are discussed in Section E.1.

We found, as elsewhere [26], that precision and sensitivity can be interchanged by suitable tuning of the parameters of the learning machine. By analogy to digital signal processing, this effect corresponds to a noise filter threshold (reduced false positives) at the expense of lost detection of weak signals (increased false negatives). For whole-proteome interaction mining, false positive (Type I) error minimization may be the preferred mode of operation. In this case, the interactions in the generated map must be associated with a high degree of confidence to warrant closer scrutiny and the expenditure of resources associated with biological or biochemical validation experiments. On the other hand, if the *detection* of all potential protein interactions is given high priority, then the sensitivity metric is most informative.

Recalling Eqs. V.3- V.5, the expected precision of the classifier's performance in the novel organism will be less than 80%, and the sensitivity will be less than 69%. The actual performance decrement cannot be evaluated until biological experiments validate or invalidate the testable hypotheses comprising the network of interactions. At present we can only estimate upper bounds on performance for this set of generated hypotheses.

### **Needle in the Haystack**

An important complication concerns the distribution of *actual* positive and negative examples in Nature. It is expected that the number of "non-interactions" greatly outweighs the number of actual interactions when considering all pairwise combinations of

proteins for a given proteome. The majority class in this context is the set of non-interacting protein pairs; actual interactions are in the minority.

When data classes are highly imbalanced, replicating the naturally-occurring proportions of each class within the training set produces classifiers that perform poorly on minority-class examples [203]. This observation may be attributed to the influence of noisy or unreliable examples present within the majority data class [112], or to the overlapping of data classes in feature space, a situation that may increase the likelihood of misclassified positive class examples by nearest-neighbor classification methods [111]. The latter may be encountered using support vector learning methods when a large percentage of the training examples lie very close to the decision surface.

There are several ways to deal with unbalanced data classes in machine learning applications. One can present both positive and negative examples in a nearly-balanced proportion, so that new data points are recognized as members of the correct class in generalization [183]. This should be done without regard to the prior probabilities associated with positive and negative data points in the biological sample. However, in designing a balanced sampling within the artificial (machine learning) environment, a variance with respect to the distribution of data classes in the real, biological world is generated.

As our ultimate objective is to make useful predictions in biology, we are forced to deal with this dilemma, which is sometimes referred to as a “needle in a haystack” problem in data mining. When making predictions on all possible pairwise combinations in a different organism, if the classifier is characterized by the same false positive rate estimated from the training data set, the number of false positives would increase significantly. The precision associated with these predictions would be seriously degraded relative to the training data. The sensitivity, or true positive rate, would be expected to remain the same. Predictions made for the “minority class” (interacting protein pairs) would tend to have a much higher error rate than those of the majority class.

The “cost” of making a false positive decision is generally different than the cost of a false negative one. In whole-proteome interaction mining, it may be argued that false negatives are inherently more costly than false positives, due to their relative scarcity. A commercial drug discovery entity might take the opposite position, as false positive leads used to initiate expensive lead validation wet chemistry experiments carry potentially significant economic costs. Therefore, it may be appropriate to take prediction *probabilities* into account, since we envision that the predictions should be subjected to some degree

of manual curation by human experts. SVM classifiers produce a binary decision<sup>4</sup>, so doing this would require transforming the outputs to produce continuous-valued, *a posteriori* probabilities (e.g., see [157, 114]).

One way to account for cost-sensitivity in learning is to analyze a group of classifiers using lift charts, recall-precision charts or receiver operating characteristic (ROC) curves [205]. ROC analysis (see Section C.2 of Chapter IV) makes no assumptions about relative distributions of the data classes, or about misclassification costs [159]. Unfortunately, ROC curves for protein-protein interactions are not likely to be as widely applicable as in the fields of radar [174] or sonar signal processing [190], where the signal and noise backgrounds have been thoroughly characterized. The protein interaction ROC space will appear differently, in terms of curve shape and absolute magnitudes, under different relative distributions of data used in their creation. Moreover, this effect might be encountered for each different organism under investigation.

Another possibility when data classes are expected to be largely imbalanced is to equalize the cost basis for the associated disproportionate misclassification costs. Practical strategies to accomplish this are presented by Elkan [62], who proposes (1) changing the proportion of examples of the majority class (here, the non-interacting protein-protein pairs), then retraining the classifier using estimated costs; or (2) applying a classification rule involving the cost-weighted posterior probabilities of class membership for each training pattern. In the second case the (artificial) even class distribution is maintained.

Alternatively, nonsymmetric costs for false positive and false negative errors during training may be explicitly embedded within the learning algorithm itself. This approach was taken by Morik and co-workers, who extended the SVM algorithm to incorporate different cost penalties for each error type independently [137].

Results of the investigations presented here were obtained by taking the numerical sign of the SVM output to indicate class membership (see Section E.1). Clearly, there are a large number of alternative strategies to deal with the imbalanced data set problem as it relates to predicting protein-protein interactions. The issues and methodologies mentioned here should be kept mind when considering statistics of cross validation performance, and their association with the expected performance on extrapolation to new organisms.

---

<sup>4</sup>That is, after thresholding of a continuous-valued output; see Eq. A-6. In principle the thresholding operation could be foregone and the real-valued magnitude of the SVM output, geometrically representing the distance of a data point from the decision surface, would then be correlated with a confidence measure: the larger the absolute value of the output, the greater the confidence in the decision.

## E.2 *C. jejuni* interaction hypotheses

The estimated generalization performance from leave-one-out experiments on the *H. pylori* proteome (Table V.2) supports confidence in the prediction of protein-protein interactions in *Campylobacter jejuni*. *C. jejuni* and *H. pylori* are close phylogenetic relatives (see, e.g., Figure 1 in [59]), displaying highly-similar constituent protein domains<sup>5</sup> (Table V.1) and genomic content ([185], Figure 2). The *C. jejuni* proteome contains 1,613 proteins, of which all possible unique pairwise protein-protein interactions (1,300,078 pairs) were encoded as features and added to the sample  $X_b$  for interaction mining. Using one of the 10 classifiers  $h(\alpha, x)$  developed during cross-validation analysis on the design organism, an interaction hypothesis was generated for each data point in this sample. A total of 5,367 distinct protein-protein interactions were declared by the decision function. Each protein comprising the *C. jejuni* interaction map was predicted to have, on average, biological connections with 3.33 other proteins.

By way of examination of the predicted *C. jejuni* protein interaction network, we first investigate the possibility of alternative automatic map inference using conserved interactions, or interologs. Secondly, we will look at some of the gross physical characteristics of the predicted interactions. After this the discussion covers general scaling properties of the map, comparing these to investigations appearing in the literature. Finally, some specific biological examples produced by the interaction mining procedure will be examined in greater detail.

## E.3 Interologs: prediction using sequence similarity

In Section C.1 of this thesis, some of the advantages and limitations of the homology approach to protein functional assignment were discussed. Despite serious limitations, the identification of protein sequence similarity remains an effective means by which function may be transferred to a query sequence from a previously characterized protein in an amino acid sequence database. It is therefore reasonable to ask of the predicted interactions:

*“How many of the predictions made for C. jejuni could have been produced on the basis of sequential similarity to the experimental interaction network of H. pylori alone?”*

---

<sup>5</sup>Source: EBI Proteome Analysis Database <http://www.ebi.ac.uk/proteome/comparisons.html>.

The idea is to find orthologous proteins within each organism, then assemble sets of conserved protein-protein interactions between each member of a known *H. pylori* interaction pair and their respective orthologs in *C. jejuni*. This is the concept of “interologs” (conserved interactions), proposed by Walhout and co-workers [195] as a means to functionally annotate uncharacterized proteins. The interolog hypothesis is that physically interacting proteins in one organism co-evolve such that their respective orthologs also interact within another organism.

To investigate the overlap between interologs and the predicted interaction map, I first estimated the orthologous amino acid sequences between *C. jejuni* (*CJ*) and *H. pylori* (*HP*) using the program InParanoid [163]. For each pairwise sequence comparison, InParanoid searches for possible orthology using BLAST [5]. BLAST (Basic Local Alignment Search Tool) is actually a series of programs for heuristic local alignment of biological sequences. This search was performed using the program `blastp` against the non-redundant protein sequence database (`nr`) using default parameters. The amino acid substitution matrix used was BLOSUM62 [86]. InParanoid input parameters used included the minimum desired percentage overlap of amino acids in the match region (90%) and a minimum bit score (100 bits). The length requirement precludes short, domain-level matches [163, 164] while the stringent score cut-off value reduces spurious local similarities in amino acid sequence.

After the BLAST search, a new interaction network of “interologs” is constructed, and we analyze its correspondence with the set of predicted interactions. The strategy is depicted in Figure V.2 in the form of a Venn diagram.

This figure shows the experimental map of *H. pylori* on the left, which is linked to the SVM-predicted and BLAST-derived maps on the right. The topmost circle denotes a collection of experimental interactions that may exist in heterogeneous databases such as DIP [209], BIND [12], MINT [211], or any number of organism-specific protein interaction databases in the public or private domain. Of particular interest are the regions of overlap indicated in the figure—these might offer evidence supporting the biological relevance of the predicted interaction map. Overlap *Region I* represents the set of interactions obtainable by sequence similarity search alone; in this region the predictions correspond to interologs. *Region II* interactions represent the intersection of hypothesized and experimental interactions. DIP contains a large number of the known protein interactions in the literature. Its current utility for our purpose here is limited; as can be seen from Table V.3,

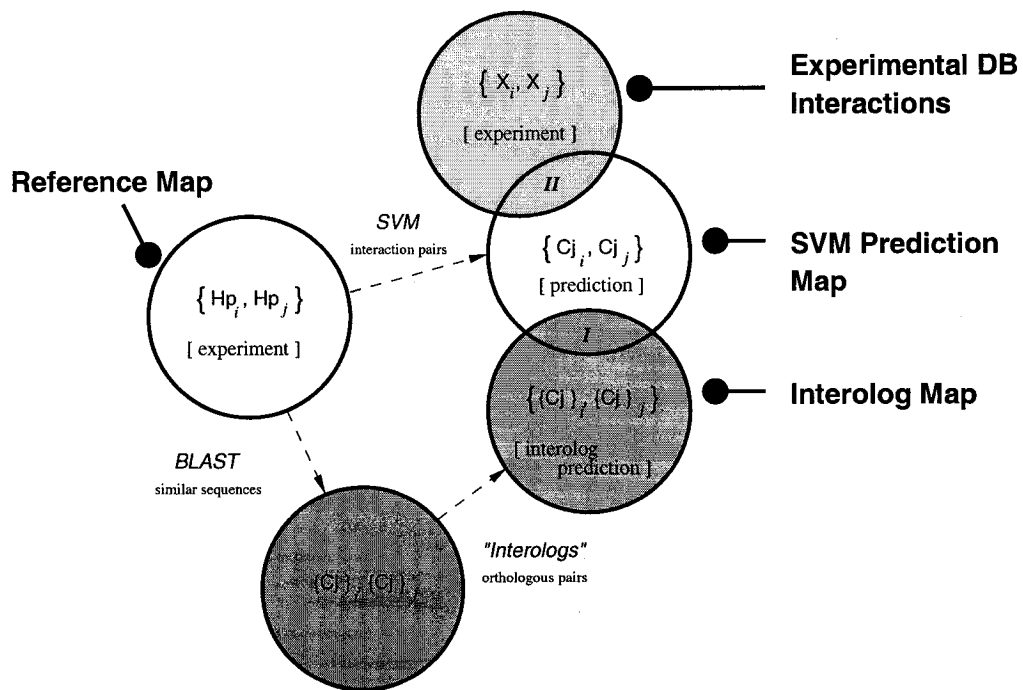


Figure V.2: Venn diagram depicting experimental and predicted interaction maps.



82% of the data in DIP represent yeast, and (excluding *H. pylori*) the most-similar organism to *C. jejuni* found therein would be *E. coli*. Although it would be interesting to investigate the similarities and differences between *E. coli* and the *C. jejuni* interaction networks in a future investigations, for the present purpose such a comparison would not necessarily provide additional insight. The strongest evidence supporting the predictions would be *Region II* protein interactions developed experimentally from the *C. jejuni* proteome.

<i>Organism</i>	<i># Interactions</i>	<i>Frequency</i>
<i>S. cerevisiae</i>	14,941	0.8273
<i>H. pylori</i>	1,415	0.0784
<i>H. sapiens</i>	717	0.0397
<i>E. coli</i>	286	0.0158
<i>M. musculus</i>	97	0.0054
<i>Others</i>	603	0.0334
<i>Total</i>	18,059	1.00

Table V.3: Interactions by organism found in the DIP database, circa November 2002. Frequency expressed as fraction of total number of interactions for each organism.

## Results

The main result from this analysis was that from a set of 125 putative interologs in *C. jejuni*, only 4 might have been predicted on the basis of sequential similarity to *H. pylori* interactions alone. This total represents 3.2% of the predicted orthologous interactions, an unexpectedly low result. Biological intuition might lead one to expect a larger percentage of common protein interactions, considering the high degree of proteomic overlap between the two organisms. If one accepts the interolog presumption of co-evolution, given the strict parameters used to construct the interologs in this analysis, SVM predictions disjoint from the interolog set are potentially false positives. More importantly, the low percentage of interologs recovered by the prediction method suggest that the SVM method is very insensitive, characterized by a high incidence of false negatives.

Compare the present 3.2% interolog “recovery rate” to that of Matthews and co-workers [133]. Using yeast two-hybrid assays, these investigators were able to experimentally verify only 16% of the predicted interologs in the nematode *Caenorhabditis elegans* when extrapolating from *Saccharomyces cerevisiae*. This rate of validation of predicted interologs precludes its automatic application to the generation of complete protein-protein

interaction maps.

It is possible that some of the “interologs” produced by selecting one sequence from each of two lists of BLAST hits were not biologically relevant. Wojcik and Schächter argued that simple similarity search may be insufficient to detect some interactions, as similarity is a property of individual proteins, and not of protein *pairs* [207].

Interolog analysis remains an active research topic. For example, unpublished work by Marc Vidal has verified some interologs between *C. elegans* and other species such as humans and *Drosophila melanogaster* [193]. This research is ongoing, and the rates of recovery of these interologs by experimentation are not available.

At present, we cannot assess the validity of the putative interologs between *H. pylori* and *C. jejuni* without experimental studies. Until definitive experimentation is performed, one likewise cannot quantitatively estimate the rates of false positives within the prediction map. This is an important area for future investigation.

#### **E.4 Physical characteristics of predicted interaction pairs**

To further examine the validity of the protein interactions within the predicted *C. jejuni* interaction map, we analyzed some of the charge and amino acid compositional statistics of the map, and compared these properties to the corresponding quantities in *H. pylori*. Such analysis may help in the assessment of the biological relevance of the predictions, and perhaps offer insight into particular biological characteristics associated with successful and unsuccessful interaction predictions. We posed three specific questions in this regard. These were as follows:

1. *What is the distribution of charged residues for each set of interactions?*
2. *What percentage of hydrophobic residues are found within the predicted interactions?*
3. *Is there any significant trend showing a predominance of cysteine residues in the interactions?*

To address these questions, protein sequence statistics for *H. pylori* and *C. jejuni* were analyzed using the program SAPS (Statistical Analysis of Protein Sequences) [36] which describes protein sequence properties for a protein or group of proteins relative to a

reference set<sup>6</sup>. The reference organism must be chosen from a select few; for this analysis *E. coli* strain K-12 was chosen because of its similarity to both *HP* and *CJ* (although in principle any reference organism would suffice, since the comparisons of interest are between *HP* and *CJ*).

In this analysis, four different amino acid sequence sets were considered: (i) all *HP* amino acid sequences, (ii) the set of interacting *HP* sequences, (iii) all *CJ* sequences and (iv) the predicted interacting sequences for *CJ*.

Not surprisingly, the resulting compositional distributions of each of the 20 amino acid types within *HP* and *CJ* were found to be highly similar to one another relative to the reference organism.

## Results

*Charge:* Results of the electrical charge distribution analysis are summarized in Table V.4. Multiple observations can be made the data in this table. In a given column, the attribute of an interaction network (predicted, or experimental) is compared to the corresponding value for the proteome at large, for a given organism. Data appearing across a row provides a comparison of the predicted and experimental interaction sets representing the two species.

One observation drawn from Table V.4 is that the net electrical charge within the interacting sequence sets was positive, whereas the net charge over the entire proteomes was negative<sup>7</sup>. This may reflect the fact that elements of the set of interacting proteins tend, on average, to be more positively charged than their “population” as represented by the proteome at large. Differential net charge between proteome and interacting protein set is seen to be larger in *HP* than in *CJ*; this may be due to the influence of a larger set of interacting proteins within *CJ* than in the *HP* experimental set (5,367 versus 1,039). The data in the table suggest that the predicted interaction set has similar electrical characteristics to that of the template organism.

*Hydrophobic clusters:* The SAPS analysis of amino acid sequences indicated that the composition of statistically significant hydrophobic amino acid segments is larger

---

<sup>6</sup>SAPS is available as an online service at <http://bioweb.pasteur.fr/seqanal/interfaces/saps.html>.

<sup>7</sup>Recall that the compositional analysis is carried out relative to a baseline organism, in this case *E. coli*.

	<i>H. pylori</i>	<i>C. jejuni</i>
<i>Proteome</i>	-1.2%	-1.2%
<i>Interaction set</i>	+0.7%	+0.2%

Table V.4: Net charge distribution for all 20 amino acids in *H. pylori* and *C. jejuni*. “Proteome” refers to the entire proteome of each organism. “Interaction set” refers to experimental (for *H. pylori*) and predicted (for *C. jejuni*) protein-protein interaction maps. Numbers shown are relative to *E. coli*.

in both interaction sets than found in their proteomic supersets. The numerical results are shown in Table V.5. From these data, it may be conjectured that the interacting proteins appear to be generally more hydrophobic than the general population from which they are extracted. Again, the level of agreement between *CJ* and *HP* inherent in these statistics suggests that the predicted interaction set has qualitatively similar hydrophobic characteristics to that of the template organism.

	<i>H. pylori</i>	<i>C. jejuni</i>
<i>Proteome</i>	30.4%	31.8%
<i>Interaction set</i>	31.7%	33.0%

Table V.5: Hydrophobics distribution for all 20 amino acids in *H. pylori* and *C. jejuni*. “Proteome” refers to the entire proteome of each organism. “Interaction set” refers to experimental (for *H. pylori*) and predicted (for *C. jejuni*) protein-protein interaction maps. Numbers shown are relative to *E. coli*.

*Distribution of cysteine residues:* Cysteine residues are known to be important because they form strong, covalent disulfide bonds with other cysteines in a protein, causing sequentially distant segments of the protein to come together in the folded, three-dimensional conformation [2]. Cysteine-rich, zinc-binding motifs along certain conserved protein domains (e.g., the RING and B-Box domains) mediate protein interactions in the context of a number of different biological functions, suggesting their fundamental importance to cellular molecular function [32]. An interesting question is whether or not cysteine residues are over/underrepresented in the set of protein-protein interactions.

Table V.6 summarizes a comparison between *H. pylori*, *C. jejuni* in the same manner as above, showing frequencies of occurrence of cysteine residues in the set of protein-protein interactions (predicted and empirical), and their occurrence rates in the correspond-

ing proteomes. It can be seen from the table that cysteines are observed less frequently in the interacting protein samples than in their proteomic representation. One possible explanation for this may be that the yeast two-hybrid assay used in obtaining the interaction map [161] cannot adequately detect certain protein interactions, including those that require post-translational modifications such as glycosylation, or disulfide bond formation. This is because the “prey”-“bait” interaction fundamental to the experimental technique must take place in the nucleus [65]. This might account for the discrepancy in cysteine representation relative to the proteome; that is, perhaps it is an artifact of the data collection method as opposed to some fundamental biological phenomenon.

	<i>H. pylori</i>	<i>C. jejuni</i>
<i>Proteome</i>	1.5%	1.6%
<i>Interaction set</i>	1.1%	1.2%

Table V.6: Cysteine residue prevalence for all 20 amino acids in *H. pylori* and *C. jejuni*. “Proteome” refers to the entire proteome of each organism. “Interaction set” refers to experimental (for *H. pylori*) and predicted (for *C. jejuni*) protein-protein interaction maps. Numbers shown are relative to *E. coli*.

## E.5 Scaling properties of predicted interaction map

Objects or processes in Nature which are invariant with respect to mathematical transformations are said to *scale* [123]. Networks of interactions in a number of natural and man-made system display conserved motifs of substructural connections, suggesting universal design patterns that correlate with successful information processing or evolutionary fitness [134]. We observed here that the inferred *C. jejuni* protein-protein interaction map shares a key topological scaling property in common with previous proteome-wide investigations: the average connectivity of the interaction network. The agreement between the present results and the cited works, which represent a variety of investigations on different organisms, offers strong evidence supporting the biological feasibility of the hypothesized map. Another scaling property, namely the distribution of sizes of “clusters” of binary protein-protein interactions, varied significantly between the present investigation and a previous study [93].

*Network connectivity.* A basic, large-scale architectural statistic describing a pro-

<i>Refs.</i>	<i>Organism</i>	<i>Method</i>	<i>Proteomic coverage</i>	<i>Average connectivity</i>
1	<i>S. cerevisiae</i>	experiment	0.55	1.388
2	<i>S. cerevisiae</i>	experiment	0.26	1.523
3	<i>E. coli</i>	prediction	0.10	2.14
4	<i>C. jejuni</i>	prediction	1.00	<b>3.33</b>
5	<i>H. pylori</i>	experiment	0.47	3.36
6,7	<i>S. cerevisiae</i>	experiment	0.17	3.2, 4.5–5.8
8	<i>C. elegans</i>	experiment	??	5.4

Table V.7: Comparison of proteome-wide interaction map connectivities for different organisms found in the literature. “Proteome coverage” is the estimated number of distinct proteins involved in interactions as a fraction of either the total proteomic complement or assay depth for a given organism. “Average connectivity” refers to the average number of interaction partners per protein comprising the map. References: 1. [91]; 2. [172]; 3. [207]; 4. Present investigation; 5. [161]; 6. [188]; 7. [187]; 8. [194]. Note: in [187], a retrospective reanalysis of data originally reported in [188] resulted in an updated estimated average connectivity of 4.5–5.8 for *S. cerevisiae*.

tein interaction map is the average number of connections between a given protein and other proteins in the map. Let us call this the “average connectivity” of the map. Table V.7 lists data collected from several different proteome-scale investigations on different organisms. It can be seen that on average, 3.33 proteins are linked to each protein in the *C. jejuni* interaction map. This level of connectivity compares favorably to the other investigations cited in the table, especially to the experimental data from [161], which provided the design sample for training the learning system in the present investigation.

Table V.7 contains a column entitled “Proteome coverage”, defined here as the estimated number of distinct proteins involved in interactions as a fraction of either the total proteomic complement or assay depth for a given organism. Note that the inferred network of interactions in this investigation has full coverage, that is, each protein is expected to participate in at least one biological interaction. Although this level of coverage is higher when compared to estimates made from other investigations in the table, a recent investigation focused on elucidating multiprotein complexes in *S. cerevisiae* indicates higher connectivity densities (0.78) than previously observed [72].

*Cluster size distribution.* In [93], it is argued that the most highly-connected proteins within a cell are also the most critical for its survival. In studies involving the protein

<i>Ref.</i>	<i>Large clusters</i> %	<i>Medium clusters</i> %	<i>Small clusters</i> %
1	0.7	6.3	93
2	1.054	38.0	60.9

Table V.8: Distribution of protein interaction cluster sizes compared to [93]. A cluster size represents the average number of interactions (edges) each protein (node) shares with other proteins. “Large” clusters refer to instances of proteins with a large number of partners ( $n > 15$ ); “medium” cluster nodes have  $5 < n \leq 15$ , and in “small” clusters each protein has, on average,  $n \leq 5$  connections to other proteins. Numbers are expressed as percentage of total number of proteins comprising the map. References: 1. [93]; 2. Present investigation.

interaction network of *Saccharomyces cerevisiae*, they derived scaling laws describing the distribution of numbers of connections between proteins in the network. Power-law scaling characteristics were found common to both *S. cerevisiae* and *H. pylori*, indicating the possibility of a universal large-scale structure in biological networks.

In that investigation, network architectural details for *S. cerevisiae* showed that the largest and smallest clusters of connected proteins constituted 0.7% and 93% of the total number of proteins comprising the map, respectively. A large interaction cluster was defined as one with  $> 15$  links, while small clusters had  $\leq 5$  binary connections to other proteins. In the present investigation, we found similar connectivity distribution properties in the predictions for *C. jejuni* only for the largest clusters, i.e. those where  $n > 15$  partners per protein node were predicted. The inferred map has a much larger distribution of small- to medium-sized clusters by comparison, as summarized in Table V.8. One explanation for this variance might be represented in arguments put forth in [82], where it is noted that the power-law cluster size distribution is characteristic of networks in a state of transitory expansion. It follows that protein interaction network connectivity is a dynamic feature; different connection properties would be expected at different states in an organisms’ evolution.

## E.6 Map visualization

We present a visualization of the complete hypothesized protein interaction map for *C. jejuni* in Figure V.3. In the figure, individual proteins are represented as vertices, and the interactions between pairs of proteins are indicated by edges connecting nodes. Proteins with a large number of partners ( $> 15$ ; 1% of all predictions) are colored red, while green



Figure V.3: Predicted whole-proteome interaction map for *Campylobacter jejuni*. In this diagram, individual proteins are represented as vertices, and the interactions between pairs of proteins are indicated by edges connecting nodes. Proteins with a large number of partners ( $> 15$ ; 1% of all predictions) are colored red; green nodes signify that relatively few proteins ( $\leq 5$ ; 61% of predictions) are expected to interact with that node. Blue nodes represent proteins with 6–14 interaction partners.

nodes signify that relatively few proteins ( $\leq 5$ ; 61% of predictions) are expected to interact with that node. Blue nodes represent proteins with 6–14 interaction partners<sup>8</sup>.

### E.7 Selected biological examples

In this section, we present specific biological examples of protein-protein interactions predicted for *C. jejuni*, exemplifying the type of information that may be extracted from the application of this approach. This represents only a sampling of the subnetworks

<sup>8</sup>The figure was generated using the graph visualization program aiSee, available online at <http://www.AbsInt.de>.



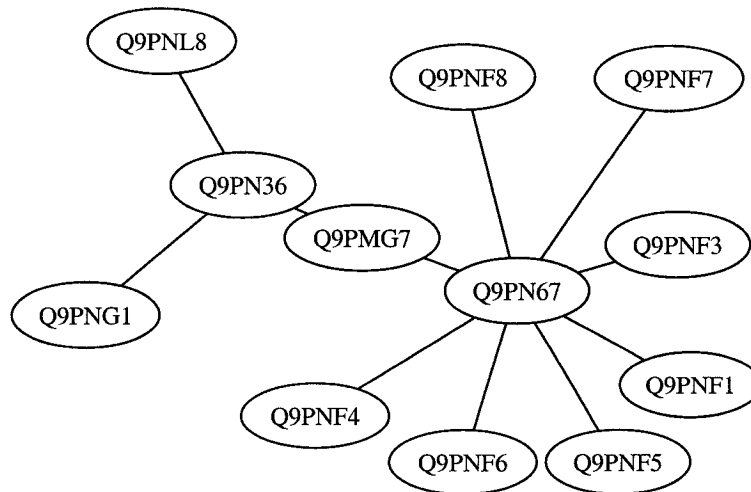


Figure V.4: Principal components of an hypothesized two-component thermoregulation signalling pathway in *C. jejuni*. Shown is a subnetwork of interactions comprising the primary interaction partners of the sensor (Q9PN36) and regulator (Q9PN67) proteins. Each protein node is labelled by its corresponding ORF designation. The previously uncharacterized protein Q9PMG7 may play a role in transferral of the message from sensor to regulator in the thermoregulation signalling pathway.

automatically generated by the interaction mining procedure.

1. *Thermoregulation.* Two-component signal transduction systems are essential in regulation of many bacterial functions, including chemotaxis, metabolism, and the response to environmental stress. The two-component mechanism constitutes a membrane environmental sensor and a cytoplasmic regulator. This mechanism typically involves autophosphorylation of histidine residues on the sensor protein, which then acts as a kinase for the regulator, the phosphorylation of which induces transcriptional activation appropriate to the chemical or thermal stimulus [107].

Elements of an hypothesized a two-component thermoregulation signalling pathway in *C. jejuni* are presented in Figure V.4 and Table V.9. The figure displays only a subnetwork of interactions comprising the primary interaction partners of the sensor and regulator proteins. Each protein node is labelled by its corresponding ORF designation. The two-component sensor (Q9PN36) is functionally linked to the putative heat-shock regulator (Q9PN67) via an intermediary protein Q9PMG7. Heat-shock proteins are known to solubilize misfolded or denatured proteins in case of extreme thermal insult to the cell [2].

The intermediate protein Q9PMG7 is designated as “hypothetical”, meaning it

has sequential similarity to other proteins of unknown function. This 180-residue protein contains two possible sites for phosphorylation (casein kinase II, tyrosine) as detected by PROSITE search [13]. It is feasible hypothesis that this previously uncharacterized protein may play a role in transferral of the message from sensor to regulator in the *C. jejuni* thermoregulation signalling pathway.

If elements of this inferred pathway are validated in wet biological studies, we suggest the possibility of its manipulation or obstruction using antibiotic agents. As recently noted, targeted inhibition of histidine kinase signal transduction pathways in bacteria may have beneficial effects for host mammals, in which cellular signal transduction proceeds according to a different mechanism [132].

<i>ORF</i>	<i>Status</i>	<i>Annotation</i>	<i>Partners</i>
Q9PN36	A	Two-component sensor	Q9PNL8,Q9PNG1,Q9PMG7
Q9PN67	P	Heat shock regulator	Q9PMG7,Q9PNF8,Q9PNF7, Q9PNF3,Q9PNF1,Q9PNF5, Q9PNF6,Q9PNF4
Q9PMG7	H	Protein Cj1495c	Q9PN36,Q9PN67

Table V.9: Principal components of an hypothesized two-component thermoregulation signalling pathway in *C. jejuni*. “Status” refers to the functional annotation status of the ORF, with *H*=hypothetical, *P*=putative, *A*=annotated.

2. *Ferric uptake and regulation.* The storage and regulation of iron levels is a fundamental aspect of cellular survival for gram-negative bacteria. Iron is a nonabundant essential nutrient that is toxic in excessive concentrations, necessitating its regulation within the cell. In *C. jejuni*, ferritins (iron-storage proteins) are also involved in oxidative stress resistance [7].

A subnetwork of putative protein interactions integral to ferric uptake and regulation processes is shown in Figure V.5. This interaction group comprises proteins linking the extracellular signal (Q9PJA5, putative integral membrane protein) to the regulatory (P48796, ferric uptake regulation) and transcriptional machinery (Q9PNK3, leucyl-tRNA transferase; Q9PN44, polyribonucleotide nucleotidyltransferase) within the cell. Such a connection is required to respond to dynamically changing requirements for iron storage or removal. Q9PNK3 is predicted to interact with Q9PMS3, a putative ferredoxin that may play a role in the intracellular redox system.

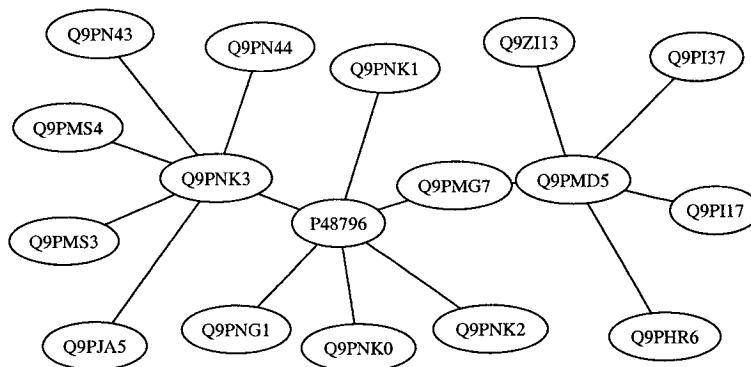


Figure V.5: Principal components of an hypothesized ferric uptake regulation pathway in *C. jejuni*. Each protein node is labelled by its corresponding ORF designation. The figure shows a subnetwork of predicted protein interactions linking the extracellular signal (Q9PJA5, putative integral membrane protein) to the regulatory (P48796, ferric uptake regulation) and transcriptional machinery (Q9PNK3, leucyl-tRNA transferase; Q9PN44, polyribonucleotide nucleotidyltransferase). Such connection is required to respond to changing requirements for iron storage or removal. Protein Q9PMD5 (possible bacterioferritin) may participate in redox stress resistance, by storing iron in a soluble, non-toxic form. Q9PMD5 is linked to a 30S ribosomal protein (Q9PI17) suggesting that this system may be involved in protection of the ribosomal machinery from iron toxicity.

Another key protein in this figure is Q9PMD5 (possible bacterioferritin) that may be instrumental in redox stress resistance, by storing iron in a soluble and non-toxic form. Q9PMD5 is linked to a 30S ribosomal protein (Q9PI17) which may suggest that this system is also involved in protection of the ribosomal machinery from iron toxicity. It is of interest to note that the hypothetical protein Q9PMG7 appears again in this inferred scenario of iron regulation. While a functional role has not been assigned for this protein, is it possible that it participates in many pathways within the cell. Recall [93], where it was argued that the most highly-connected proteins in protein interaction networks are most crucial to a cell's viability. Perhaps this protein carries such significance within *C. jejuni*. This question awaits further proteomic study and validation.

The protein components central to the hypothesized ferric uptake interaction cluster are summarized in Table V.10.

## E.8 Postscript

There is a great amount of future research and experimentation required to fully assess the biological relevance of predicted whole-proteome interaction maps using the

<i>ORF</i>	<i>Status</i>	<i>Annotation</i>	<i>Partners</i>
P48796	A	Ferric uptake regulation protein	Q9PNK3,Q9PNK2,Q9PNK1, Q9PNG1,Q9PMG7
Q9PNK3	A	Leucyl-tRNA synthetase	Q9PMS3,Q9PN43,Q9PMS4, Q9PN44,Q9PJA5
Q9PMD5	A	Possible bacterioferritin	Q9PI17,Q9PHR6,Q0ZI13, Q9PI37,Q9PMG7

Table V.10: Principal components of an hypothesized ferric uptake regulation pathway in *C. jejuni*. “Status” refers to the functional annotation status of the ORF, with *H*=hypothetical, *P*=putative, *A*=annotated.

methodology proposed here. The fact is that we do not know whether or not each of the predictions are useful until genetic or biochemical experiments are performed to substantiate or invalidate them. Until that time, they may be considered as valid hypotheses awaiting empirical confirmation or falsification.

We predict 5,367 protein-protein interactions in *C. jejuni* using training data from *H. pylori*, which contains 1,039 data points. Are the 4,000 or so novel *C. jejuni* predictions false positives? From the perspective of classical computational biology, one conclusion that might be drawn from the interolog analysis discussed in Section E.3 is that there are *too few* homologous pairs (from sequence similarity search) found within the predicted interaction network. To the extent that many interologs truly exist between the two organisms, the predictions are insensitive and imprecise, characterized by a large rate of both false negatives and false positives.

From another angle, it is possible to argue the opposite position—that in fact the false positive rate is not compromised. Let us return momentarily to the discussion of Section E.1. For the sake of argument, assume that the experimentalists have diligently performed their duties, and that the interactions they have measured in *H. pylori* are indeed comprehensive and biologically relevant. The SVM trained on these interactions in turn predicted 5,367 protein-protein interactions in *C. jejuni*. This result appears to agree, at least in order of magnitude, with the “true state of nature”. Since the two organisms have highly-similar protein distributions, and are similar in terms of biological environment and function, it would be reasonable, intuitively, to expect that the number of protein-protein interactions in each should be roughly similar as well.

However, rigorously following the “needle in the haystack” line of reasoning pre-

sented in Section E.1, only finding 5,367 interactions in *C. jejuni* is a surprising result, as one might expect a much larger number of predicted interactions ( $O(100,000)$ !) to occur by extrapolation from raw numbers calculated from the precision and sensitivity cross validation estimates. Does the SVM learn to “detect” an interaction on a constant percentage of data points, or only when a data point representing patterns correlated with a true protein-protein interaction is presented?

These are important questions that need to be addressed in future research. The only conclusion that may be advanced with certainty at the present time is that verification (or refutation) of the hypothetical predictions requires experimentation. The methodology presented in this chapter offers one framework in which the biological utility of proteome interaction mining may be explored.

## **F Acknowledgement**

The text of this chapter, in part or in full, is a reprint of the material as it appears in Joel R. Bock and David A. Gough, “Whole-proteome interaction mining”, *Bioinformatics* 19(1):125-134, 2003. The dissertation author was the primary author, and the co-author listed in this publication directed and supervised the research which forms the basis for this chapter.

## VI

# A new method to estimate ligand-receptor energetics

## A Introduction

The process of developing a new drug involves seven major steps [11]. (i) First, a disease is identified, then (ii) drug targets (usually, proteins) within the cell are hypothesized, the activation or inhibition of which it is thought to alter the disease state. Once targets are identified, the next task is to (iii) identify potential lead compounds that will bind to the target. These leads are subsequently (iv) optimized with respect to their structural characteristics in the context of the target binding site, then subjected to (v) preclinical and (vi) clinical trials to determine their bioavailability and therapeutic potential. The final step is to (vii) optimize efficacy, toxicity and pharmacokinetic properties. This may involve the use of pharmacogenomics techniques to tailor compounds to a subset of the patient population that is predisposed to a disease.

Pharmaceutical companies are exposed to great financial risk in the course of identifying viable drugs to treat a certain condition or disease. There are also tremendous direct and indirect (opportunity) costs associated with delaying the removal of non-viable drugs from this drug discovery “pipeline” until the latest stages of the process.

A huge number of drug targets have been generated from genetics, genomics and proteomics technologies. Accordingly, the lead identification and optimization steps have assumed critical importance. High-throughput experimental screening assays [49]

have been complemented recently by computational (“virtual screening”) approaches to identify and filter potential ligands, given the characteristics of the target receptor structure of interest [21, 200]. In virtual screening, databases of compound libraries are searched, and scoring or discrimination functions are used to select the “best” candidate compounds for biological activity analysis [116].

The scoring of ligands in virtual screening is often associated with computational docking simulations that mate receptor and cognate small-molecule ligand in three-dimensional space. To provide broad generalization in “chemical diversity” space, computing this score requires the accurate prediction of binding affinities of many structurally distinct ligands [78]. Three main methodologies have been identified for free binding energy calculations. In order of computational complexity, these are: (1) knowledge-based scoring functions, (2) partitioning the binding energy into biophysical energy terms and (3) molecular dynamics [167]. The most accurate computations are represented by molecular dynamics techniques, but their inherent computational intensity precludes their application to industrial-size chemical databases.

Regression-based scoring functions, as exemplified by the work of Böhm [30], are fast but require a three-dimensional structure of the receptor. This prohibits their use in cases where the structure is difficult to obtain, such as with transmembrane proteins. The accuracy of such methods has also been called into question. A recent investigation concluded that “no significant correlation” existed between Böhm-type scores and experimentally-determined binding affinities for a group of fifteen complexes [136].

In this research, we propose a new method to estimate the free binding energy between a ligand and receptor. We extend a central idea developed in previous investigations [25, 26, 28] that uses simple descriptors to represent biomolecules as input examples to train a support vector machine (SVM) [191], and the application of the trained system to previously unseen pairs, estimating their propensity for interaction. Here, we seek to learn the function that maps features of a receptor-ligand pair onto their equilibrium free binding energy.

## B System and Methods

### B.1 Thermodynamics of binding

For our purposes, consider that a single protein  $P$  binds a single small molecule ligand  $L$  to form complex  $C$ , or



Assuming that this reaction is in thermodynamic equilibrium, the Gibbs free energy change on binding  $\Delta G^0$  is written

$$\Delta G^0 = -RT \ln(K_a) \quad (\text{J/mol}) \quad (\text{VI.2})$$

where  $R$  is the gas constant,  $T$  is the temperature ( $^{\circ}\text{K}$ ) and  $K_a$  is the equilibrium binding constant between protein and ligand<sup>1</sup>.  $K_a$  is defined as

$$K_a = [C]/[P][L] \quad (\text{M}^{-1}) \quad (\text{VI.3})$$

where  $[C]$ ,  $[P]$  and  $[L]$  are molar concentrations of complex product, protein and ligand reactants, respectively. Often the equilibrium dissociation constant  $K_d$  is used to quantify ligand binding strength. It is simply the inverse of the binding constant, or

$$K_d = \frac{1}{K_a} = [P][L]/[C] \quad (\text{M}) \quad (\text{VI.4})$$

and represents the concentration of ligand required to saturate half of the protein's available binding sites.

Calculation of  $\Delta G^0$  usually entails its partitioning into various energetic components accounting for rotatable bond entropy, hydrogen bonds and ionic interaction forces, lipophilic protein-ligand contact surface, and others [29].

### B.2 Database of ligand-receptor objects

The data set used in this investigation was aggregated automatically using information located in a number of disparate online resources, coupled with local computations. An object data base was constructed from this data, and subsequently sampled to generate examples for training and testing the performance of the regression estimation system. The experimental database consisted of 2,956 objects, each having attributes as summarized in this section.

---

<sup>1</sup>Under physiological conditions (310  $^{\circ}\text{K}$ , 1 atm, 1.0 M), the value of  $RT$  is about 2.577 kJ/mol or 0.616 kcal/mol.



**Ligand-receptor complex.** Ligand-receptor data were extracted from the Computed Ligand Binding Energy (CLiBE)<sup>2</sup> database, a compendium of information on complexed receptors and ligands. Each record in CLiBE contains computed values for the total ligand-receptor potential energy field  $\Delta G^0$ , given by

$$\Delta G^0 = \Delta G_v + \Delta G_h + \Delta G_e + \Delta G_s \quad (\text{VI.5})$$

where the right-hand-side partitioning represents energy contributions due to non-bonded van der Waals interactions, hydrogen bonds, electrostatic forces and ligand desolvation energies, respectively [145]. Methods underlying the computation of binding energies comprising the database subject to this investigation are described in [45].

The complexes within this resource are themselves based on “heterogen” records found in the Protein Data Bank (PDB) [20]<sup>3</sup> for which a chemical identity has been assigned to the ligand. PDB is a public domain repository of experimentally determined structures of biological macromolecules.

**Ligand structures and chemical names.** Data files with entries representing ligand structures and their associated chemical names were obtained from the National Cancer Institute (NCI) Open Database of Compounds<sup>4</sup>. The data entries were represented as “SMILES” strings, where SMILES (Simplified Molecular Input Line Entry System) is a specification and nomenclature for describing molecules as a compact, one-dimensional string of characters, including atoms, bonds, aromatic rings and branches [202].

**Molecular connectivity.** The SMILES representation for each ligand molecule was converted to a two-dimensional connectivity matrix using a computational chemistry package (JOelib; [201])<sup>5</sup>. The rows and columns of this matrix reflect the cardinality of constituent atoms established by the SMILES representation. At row  $i$  and column  $j$ , a unit-valued entry is made if the corresponding atoms in the molecule are covalently connected; otherwise the value of that matrix element is zero. Diagonal elements of this matrix store the appropriate atomic number, as suggested previously [41].

<sup>2</sup>CLiBE circa August 2002 has 14,731 records, with 2,803 distinct ligands and 2,256 distinct receptors. See <http://xin.cz3.nus.edu.sg/group/clibe/clibe.asp>.

<sup>3</sup>PDB contains 18,294 structures as of 23-Jul-2002. See <http://www.rcsb.org/pdb/>.

<sup>4</sup>Available at <http://cactus.cit.nih.gov/ncidb2/download.html>, this resource currently contains over 250,000 compounds.

<sup>5</sup>Open source, available at <http://joelib.sourceforge.net/>.

**Molecular synonyms.** To maximize the chemical diversity of objects potentially available for numerical experiments, a list of common chemical synonyms corresponding to each ligand were obtained using the online ChemFinder service<sup>6</sup>. Each ligand synonym within its list was used in a lexical similarity search of the NCI compound files to obtain SMILES representations in cases where different chemical names were used for identical ligands across databases.

### B.3 Support vector regression

The support vector algorithm, based on statistical learning theory, is applicable to both (1) binary classification and (2) regression estimation [191]. In previous work, we developed methods to train a support vector machine (SVM) classifier to learn to predict protein-protein interactions using descriptors based on physicochemical properties of paired amino acid sequences [25, 26, 28]. In the present application, we propose to exploit the SV algorithm to solve a regression problem. The concept to be learned is the functional mapping between a set of ligand-receptor features and the total free binding energy of the complex. The basic idea in support vector regression (SVR) is to map a set of input patterns  $X = \{x_1, x_2, \dots, x_l\} \in \mathbb{R}^n$  onto a high-dimensional feature space  $\mathcal{F}$  via a nonlinear mapping  $\Phi: \mathbb{R}^n \mapsto \mathbb{R}^D$  ( $D \gg n$ ), and then perform linear regression in  $\mathcal{F}$ . Each pattern vector  $x_i$  has a matching target value  $y_i \in \mathbb{R}$ . The goal is to find a function  $y = f(x)$  representing the real-valued pairs  $\{z_i \mid z_i = (x_i, y_i), i \in 1, \dots, l\}$  within a certain acceptable maximum deviation level  $\varepsilon$  [179]. Practical implementation issues with SVR are presented in [179, 138], and theory and algorithms for extension to regression estimation with noisy data appear in [178].

### B.4 Feature representation

Each ligand-receptor complex was transformed into a vector of numerical features presumed salient for learning the target concept. Receptor and ligand feature vectors constructed as outlined in this section are concatenated and labelled with the value of their total free binding energy. These vectors are subjected to SVM regression training and cross-validation testing to evaluate how keenly the system learned the concept as posed.

---

<sup>6</sup>See <http://chemfinder.cambridgesoft.com/result.asp>.

**Receptor.** Receptor protein features were generated as described previously [25], considering tabulated physicochemical properties (charge, hydrophobicity and surface tension) of the amino acid sequence to be prototypical of binding characteristics of the receptor. Each residue in sequence was replaced by floating point numbers with values corresponding to these physical properties. This vector of numbers was then mapped onto a fixed-length interval, to provide a basis for comparison between receptor proteins of varying sequence length.

**Ligand.** Exemplars for the ligand component of each molecular complex required a novel approach. The design ethos followed here dictates beginning with a minimal, elemental group of features, in order to develop intuition regarding the feature space.

In accordance with this approach, the two-dimensional molecular connection matrix described in Section B.2 was supplemented by additional arrays, each of which contained numerical values for fundamental, measurable chemical properties characterizing the atoms comprising the molecule. These properties included the atomic *ionization potential energy*, which represents the energy necessary to remove the outermost electron from the ground state of a neutral atom, and the *electron affinity*, which is a measure of energy change upon adding an electron to a neutral atom [31]. Ionization energies are always positively-valued, while electron affinities may assume either positive or negative numerical values.

For each small molecule ligand, three two-dimensional arrays representing molecular topology, electronic structure and chemical behavior of the component elements, were concatenated into a single, wide matrix. The resulting aggregate data matrix was then factorized using the singular value decomposition [79]. The singular values computed in this factorization are extracted, representing a projection onto one-dimensional space of the essential characteristics of molecular bond topology, and, it is hypothesized, the spatial distribution of molecular properties important for binding with a receptor.

Burden [41] introduced the idea of computing the eigenvalues of a hydrogen-suppressed molecular bond graph with atomic number on the diagonal and numbers indicating bond presence and type at off-diagonal positions. This matrix was used as a means to group substructures for chemical similarity search. In that work, it was maintained that the smallest eigenvalue embodied information on *all* molecules, and therefore was sufficient as a topological descriptor. Here, all singular values are retained, regardless of their relative magnitudes, as discarding the entire set is not justifiable. This vector is finally stretched (or

compressed) onto a fixed length interval, as was performed for the receptor features.

## C Implementation

### C.1 Learning concept

The concept to be learned is the function  $y = f(x)$  that maps ligand-protein feature vectors  $x$  to the corresponding free energy of binding  $y$ . How well the SVR machine learns this concept will be quantified using the statistics described in Section C.2, collected from observations of the cross validation protocol as described in Section C.3.

### C.2 Evaluation of machine learning

One measure of effectiveness for regression estimation is the normalized mean squared error, given by

$$nmse = \frac{1}{\sigma^2} \frac{1}{N} \sum_{k=1}^N (y_k - \hat{y}_k)^2 \quad (\text{VI.6})$$

where  $N$  is the number of target points predicted,  $\sigma^2$  is the actual sample variance,  $y_k$  and  $\hat{y}_k$  are the actual and estimated target values of the  $k$ -th data point, respectively [74]. Because  $nmse$  is normalized by the sample variance, it may be used to compare different regression studies on a more equitable basis than would be possible using the conventional rms error; intuitively, a given prediction experiment is less challenging where the variance in the data is small. Notice that if we replace the prediction terms  $\hat{y}_k$  with the arithmetic mean  $\bar{y}$  in Eq. VI.6, the value of the statistic is 1. This trivial case results when the predictor simply outputs the mean value of the data. Low values of  $nmse$  indicate good overall predictive acuity.

Pointwise predictions of ligand binding may be evaluated using the normalized mean absolute error, defined by

$$nmae = \frac{1}{\sigma} \frac{1}{N} \sum_{k=1}^N |y_k - \hat{y}_k| \quad (\text{VI.7})$$

This statistic is normalized by the sample variance for the same reasons as were cited for  $nmse$  above. Furthermore, its value may be interpreted as the number of standard deviations, on average, that predictions differ from the target values across the test set. The lower the value of  $nmae$ , the better the system pointwise predictive ability.

In some ligand screening situations (such as “virtual screening” [21]), predicting the relative ranking of binding strengths among a set of ligand-receptor pairs may be desired. The output of such an analysis would be a list of predicted binding energies, sorted according to predicted magnitudes  $\Delta\hat{G}^0$ . In such cases a measurement of non-parametric or rank correlation, such as represented by Kendall’s  $\tau$  coefficient [104], is informative. In cross-validation, given an ordered array of  $N$  “(actual,predicted)” values  $(y_1, \hat{y}_1), \dots, (y_N, \hat{y}_N)$ , we systematically compare the numerical signs of individual bivariate pairs  $X = (y_i, \hat{y}_i)$  and  $Y = (y_j, \hat{y}_j)$  for  $i = 1, \dots, N, j = (i+1), \dots, N$ .

If either (a)  $y_i > y_j$  and  $\hat{y}_i > \hat{y}_j$ , or (b)  $y_i < y_j$  and  $\hat{y}_i < \hat{y}_j$  is observed,  $X$  and  $Y$  are said to be “concordant”. Otherwise, the points are “discordant”. Kendall’s  $\tau$  expresses the tendency of two ordered lists  $y$  and  $\hat{y}$  to coordinately increase or decrease, and is computed as

$$\tau = \frac{N_C - N_D}{\sqrt{N_C + N_D + T_X} \sqrt{N_C + N_D + T_Y}}, \quad -1 \leq \tau \leq +1. \quad (\text{VI.8})$$

where  $N_C$  is the total number of concordant pairs,  $N_D$  is the number of discordant pairs, and  $T_X, T_Y$  are counts of the “ties” found in  $X$  and  $Y$  pairs, respectively. A large positive(negative) value of  $\tau$  that the rank ordered values  $y$  and  $\hat{y}$  are positively(negatively) correlated.

### C.3 Cross validation experiments

To estimate the generalization error of the trained support vector regression system, we averaged the results of ten separate 10-fold cross validation experiments. In  $k$ -fold cross validation,  $k$  random, equal-sized, disjoint partitions (folds) of the example data are constructed, and an “inducer” (here, an SVR engine) is trained on  $(k-1)$  folds, with the excluded fold being used to test the trained system performance. After  $k$  such experiments, the results are averaged, and the observed error rate may be taken as an estimate of the error rate expected upon generalization to new data [108]. To reduce further the effects of chance in randomly sampling the data, we averaged the results of 10 different 10-fold cross validation experiments, performing 100 different training/testing procedures. The results we present are cross validation averages for the statistics  $nmse$ ,  $nmae$  and  $\tau$  as described in Section C.2.

The total sample used in these experiments comprised 2,671 distinct ligand-receptor complexes. The output of the trained system is a predicted level of binding free

energy  $y$  (in kcal/mol) given a set of features abstracted from a given input complex  $x$ . A qualitative glimpse of typical results from one complete 10-fold cross-validation test is offered in Figure VI.1, which shows a scatter plot of actual versus predicted binding energy. The figure shows that some degree of correlation between prediction and truth exists. This correlation will be examined on an objective basis in the discussion of Section D.1.

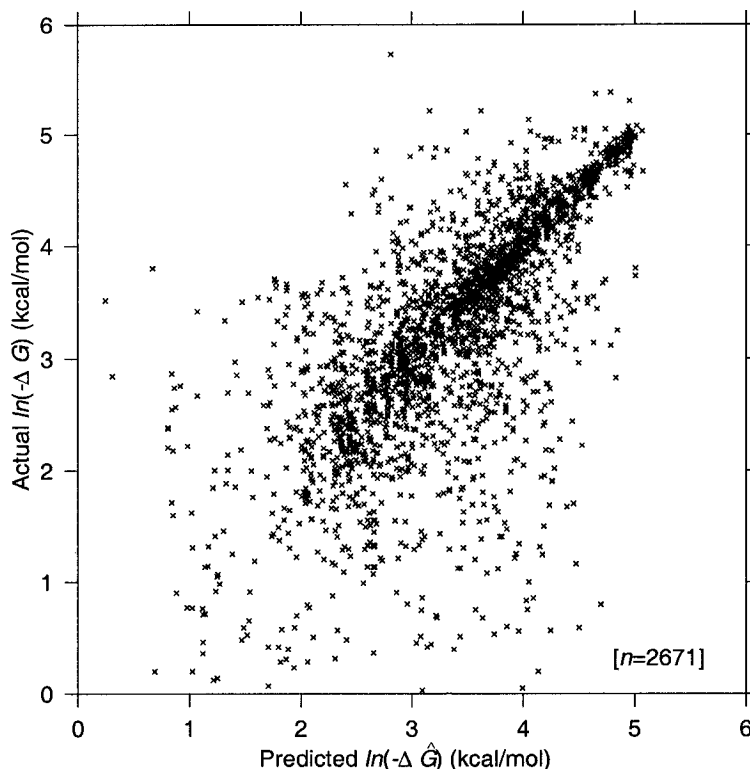


Figure VI.1: Actual versus predicted binding free energy. Shown are typical results from one complete 10-fold cross validation experiment on the ligand-receptor database discussed in Section B.2. Sample size  $n = 2,671$ .

## D Discussion

### D.1 Cross validation results

The principal results obtained in this investigation are summarized in Table VI.1 and in Figure VI.1. The table compares the ten 10-fold cross validation error estimates to a number of studies reported in the literature. In contrast to the present results (shown in boldface), all of the competing methodologies shown in the table are derived from scoring

functions or simulations predicated upon knowledge of the three-dimensional structure of receptor and ligand complex. The columns in Table VI.1 comprise test sample size  $n$ ; the mean target binding energy  $\bar{y}$  and standard deviation ( $\sigma_y$ ), in kcal/mol; normalized mean square error ( $nmse$ , Eq. VI.6); normalized mean absolute error ( $nmae$ , Eq. VI.7); and Kendall's tau ( $\tau$ , Eq. VI.8).

<i>Ref.</i>	$n$	$\bar{y}$ (kcal/mol)	$\sigma_y$ (kcal/mol)	$nmse$	$nmae$	$\tau$
1	14	-4.09	1.179	0.198	0.344	0.753
2	12	-0.98	0.332	0.271	0.401	0.667
3	11	-4.25	0.711	0.377	0.466	0.455
<b>4</b>	<b>2671</b>	<b>-37.76</b>	<b>35.106</b>	<b>0.419</b>	<b>0.377</b>	<b>0.552</b>
5	13	-3.93	0.796	0.440	0.497	0.632
6	30	-8.897	2.591	0.720	0.661	0.418
7	17	-8.17	3.785	0.789	0.621	0.358
8	63	-1.45	0.560	1.342	0.836	0.307
9	13	-10.27	6.683	1.466	0.511	0.533

Table VI.1: Comparison of predictions of ligand-receptor binding free energies in the present investigation (boldface font) and various studies reported in the literature. Test data statistics are sample size ( $n$ ), target value mean ( $\bar{y}$ ) and standard deviation ( $\sigma_y$ ). Results are shown for normalized mean square error ( $nmse$ , Eq. VI.6), normalized mean absolute error ( $nmae$ , Eq. VI.7), and Kendall's tau ( $\tau$ , Eq. VI.8). References: 1. Head 1996, Table 3 [84]; 2. Böhm 1998, Table 3 [30]; 3. Wang 1998, Table 4 [198]; 4. Bock 2002 (Present investigation); 5. Head 1996, Table 4 [84]; 6. Wang 2002, Table 4 [197]; 7. Rarey 1996, Table 1 [162]; 8. Zhang 1996, Table 1 [212]; 9. Schapira 1999, Table 5 [167]. *Note:* results for present investigation are average values from ten 10-fold cross validation experiments.

The records in the table are listed in order of increasing  $nmse$ . This statistic is proposed as the primary objective indicator of accuracy for direct prediction of binding free energy.

Of particular note on consideration of Table VI.1 are the sample size and mean free binding energies characterizing the ligand-receptor data used here, when contrasted to the other investigations. The current sample size ( $n = 2,671$ ) is a factor of 42 times larger than the next largest data set. The mean free binding energy is seen to be  $-38$  kcal/mol, significantly stronger than the other data summarized in the table. Moreover, it can be seen that the present data set is highly variable, as the standard deviation (35 kcal/mol) is on the

same order as the mean.

Recall from the previous discussion that *nmse* values on the order of 1 are tantamount to trivial prediction of the mean value of a test data set. Lower values of *nmse* are associated with genuine learning of underlying patterns in the data, and effective generalization. On this basis, the highest predictive accuracy (#1;  $nmse = 0.198$ ) observed in this comparative study was realized by Head and co-workers [84], who present a hybrid approach combining ligand-receptor 3D-structural information and parameters derived from molecular mechanics. The test set comprised 14 ligand-receptor complexes.

The second best *nmse* in this group was achieved by Böhm [30] using a regression-based empirical scoring function based on hydrogen bonds, electrostatics, complementary surface areas and other characteristics of receptor-ligand pairs where the 3D structure has been previously determined.

Next in our list of prediction results is the investigation reported in Wang *et al.* [198]. Their approach uses another empirical scoring function for binding free energy that explicitly accounts for contributions due to Van de Waals interactions, metal-ligand bonding, hydrogen bonds, desolvation energies and different kinematic effects. A regression equation is developed using these terms derived from known receptor-ligand complexes. All 11 data points in the test sample were based on endothiapepsin receptor complexes.

The current method, based on support vector regression, obtained the fourth-best prediction error ( $nmse = 0.419$ ) averaged over ten different 10-fold cross validation tests. We suggest that this error rate represents a significant step, for the following reasons:

1. The error rate and rank correlation value are surprisingly competitive with other investigations, in light of the relatively large variance and extremely large sample size of the underlying data set. Note that the fifth-lowest *nmse* value in Table VI.1 was also obtained by Head *et al.* [84], for a different data set than they used in entry #1. Group #5 comprised 13 HIV-1 protease/HIV protease inhibitor complexes, and showed a value of  $nmse = 0.440$ . So the same methodology by the same research group, applied on a different data set, realized much different predictive results. This demonstrates the variability in results that are possible when using small sample sizes, while providing confidence in the robustness of our current method and results, which were based on a sample size  $n = 2,671$ .
2. The features used to represent the ligand-protein complexes in the support vector



regression do not require any information about three-dimensional structure. All that is required as input data are the amino acid sequence of the receptor, and a connection table representing the ligand structure in two dimensions and the atom characteristics at the nodes of this connection table.

3. There is no limitation on the protein family membership of the putative receptor(s), or on the type (organic, synthetic) or size of ligand used.
4. The results obtained in this study suggest that it may be possible to infer binding energies for complexes involving newly-sequenced or difficult-to-crystallize proteins, or for ligands that only exist in computer memory, awaiting synthesis upon successful *in silico* screening.

**Rank correlation.** We draw the reader's attention to the trend in Kendall's rank correlation statistic  $\tau$  in Table VI.1. It is apparent that there is a general inverse correlation between the magnitude of binding energy prediction errors (*nmse*, *nmae*) and the value of  $\tau$ . That is, low values of prediction error are associated with high values of the correlation coefficient.  $\tau$  measures the tendency of two ordinal random variables (here, actual and predicted binding energy rank) to increase or decrease coordinately. If direct prediction of the physical binding energy is reasonably accurate, we would expect to see a positive and non-trivial correlation between the corresponding rank-ordered variables.

Computing biomolecular binding energies to higher accuracy remains a challenging problem [76]. One author recently noted that current computational docking simulations, used to search for the best (lowest energy) "fit" of ligand into a target receptor cavity, still "suffer from insufficient precision of the scoring functions" [113]. In [117], molecular dynamics simulations focused on biotin binding to avidin and streptavidin indicated that the energies of protein and ligand reorganization were found to be significant contributors to protein-ligand binding free energy in molecular dynamics simulations. These reorganization energies were estimated to be on the order of 10–30 and 4.5–6 kcal/mol for protein and ligand, respectively. Because of the large variance in protein reorganization energy, the authors concluded that precise predictions of binding free energy were suspect.

Given these difficulties, the ability to reliably rank a set of ligand-receptor complexes during lead optimization (versus directly computing binding energy) remains important in the area of drug discovery. Such a procedure may add value, for example, as a

decision aid when down-selecting a set of ligands for chemical synthesis. In connection with the current methodology, we recognize that training the SVR requires example data representing estimated or measured values of binding free energy. The output of a computational technique cannot exceed the accuracy of its input; this is especially true with systems that learn from examples. Therefore, at present the qualitative analysis or ranking of potential ligands may be the main utility of the SVR technique.

The prediction evaluation statistics appearing in Table VI.1 are presented in the form of a bar chart in Figure VI.2. The investigations numbered along the horizontal axis appear in order of increasing  $nmse$ , and corresponding to the numbering in Table VI.1. This visualization provides a different perspective on the opposing trends of  $nmse$ ,  $nmae$  and  $\tau$  as discussed above.

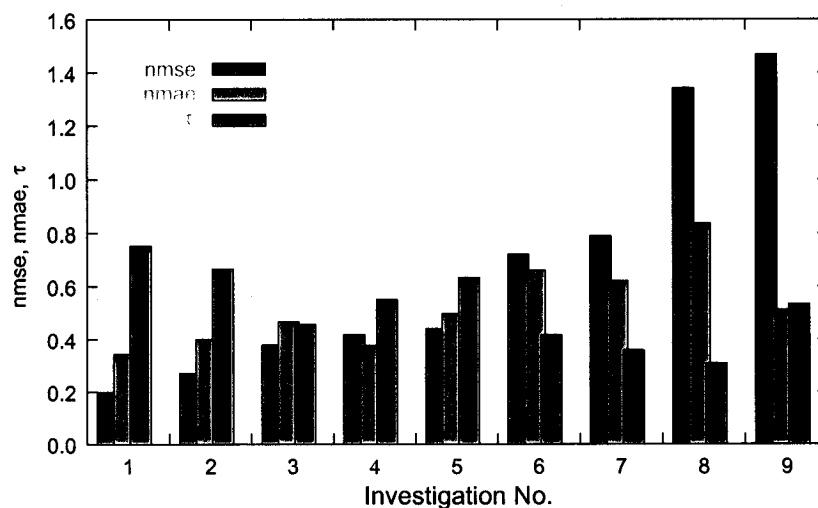


Figure VI.2: Comparison of error and rank correlation statistics between this study and the literature. The investigations numbered along the horizontal axis appear in order of increasing normalized mean square error  $nmse$  (Eq. VI.6), and correspond to the numbering appearing in Table VI.1. Notice the general trend of inverse correlation between binding energy prediction errors ( $nmse$ ,  $nmae$ ) and rank correlation ( $\tau$ ). The present cross validation results are represented as Investigation #4 in this figure.

## E Conclusions

In this work, we have introduced a new methodology, showing that it is possible to predict the binding free energy between ligand and receptor without direct information about their three-dimensional structures.

In cross validation experiments, we have demonstrated that objective measurements of prediction error rate and rank-ordering statistics are competitive with several other investigations, most of which depend on three-dimensional structural data. The size of the sample used ( $n = 2,671$ ) indicates that this approach is robust and may have widespread applicability beyond restricted families of receptor types.

Newly-sequenced proteins, or those for which three-dimensional crystal structures are not easily obtained, can be rapidly analyzed for their binding potential against a library of ligands using this methodology.

## F Acknowledgement

The text of this chapter, in part or in full, is a reprint of the material as it appears in Joel R. Bock and David A. Gough, "A new method to estimate ligand-receptor energetics", *Molecular & Cellular Proteomics* 1:904-910, 2002. The dissertation author was the primary author, and the co-author listed in this publication directed and supervised the research which forms the basis for this chapter.

## VII

# Conclusions

### A Conclusions

1. In this thesis, I have demonstrated that explicit information about three-dimensional biomolecular structure is not necessary to make predictions of protein-protein and protein-ligand interactions. Using simple descriptors of physicochemical characteristics of amino acid sequences (for proteins) and molecular connection tables (for small-molecule ligands), the techniques introduced and developed here have been shown to successfully predict these interactions at rates greatly exceeding chance. This is a significant contribution, because it implies that researchers may proceed directly from sequence to inference of protein function, as represented by the context of a protein's interactions with other biomolecules. Newly-sequenced proteins, or those for which three-dimensional crystal structures are not easily obtained, can be rapidly analyzed using this methodology.
2. It is possible to predict a complete protein-protein interaction map within a single organism, as shown in Chapter IV. Using receiver operating characteristic (ROC) analysis, I have demonstrated that the precision and sensitivity of these predictions can be traded against one another, and "engineered" to some extent by the choice of parameters used to implement the support vector learning machine (SVM). In experiments on data representing protein interactions in *Saccharomyces cerevisiae*, certain SVMs were characterized by high degree of precision (> 90%) and low sensitivity (36%); others produced classifiers that were characterized by more moderate cross

validation sensitivities (64%) and precision (68%).

3. The investigation of Chapter IV showed that it is essential to eliminate redundant data examples before training the learning machine. If this is not done, an artificially-high sensitivity rate may be realized. I found that the observed sensitivity rate for *S. cerevisiae* interaction data was overstated by more than 20% if redundancy-reducing processing was not performed in advance.
4. This methodology may be used to infer a complete protein-protein interaction map in a novel organism. In Chapter V, I presented an algorithm for systematically training an SVM learner on protein interactions in species *A*, and predicting a complete interaction network in species *B*. The training data set does not require any examples of interactions within *B*, only that the two organisms have sufficient genetic similarity that useful rates of prediction may be expected. This idea has been demonstrated by training a learning machine on experimental interactions in the bacterium *Helicobacter pylori*, and computing a complete protein-protein interaction network for the enteric pathogen *Campylobacter jejuni*.
5. The study presented in Chapter V considers the possibility of inferring interaction networks across organisms, making predictions where no experimental interaction data are yet available. When training a learning system on interactions found experimentally in one organism and generalizing to other organisms not represented in training data, precision rates may be significantly less than those estimated from cross validation errors. The true distribution of positive and negative examples in Nature is expected to be highly skewed—the number of “non-interactions” greatly outweighs the number of actual interactions when considering all pairwise combinations of proteins for a given proteome. When making predictions, a constant false alarm rate classifier will exhibit much lower rates of precision than those indicated by cross validation errors in training. In such cases, a classifier architecture characterized by a high rate of sensitivity would be preferred, as the sensitivity metric is independent of the rate of false positives, and would be expected to remain unchanged.
6. The method and results described in Chapter VI establish the feasibility of predicting the binding free energy of a ligand-receptor complex, without knowledge of the three-dimensional structural configuration of either the ligand or the receptor. I showed that

objective measurements of prediction error rate and rank-ordering statistics are competitive with methods presented in several published investigations, most of which depend on three-dimensional structural data. The size of the sample indicates that this approach is robust, and may have widespread applicability beyond restricted families of receptor types.

7. The objective of this research was to develop methods to automatically generate hypothesized protein-protein interactions within a proteome. Machine learning approaches generate empirically falsifiable hypotheses to be subsequently supported (or not supported) experimentally; therefore currently they may complement– but cannot supplant– biological experiments. An iterative coupling between successive rounds of computer prediction and experimental validation must be accomplished. Only in so doing can the regions of applicability and limitations of the present approach be discovered. Further development along these lines may produce a robust computational screening technique that may be useful to reduce the set of putative candidate protein-protein or protein-ligand interactions within an organism, tissue or physiological state of interest.

## **B Suggestions for future research**

1. *Negative examples.* The specification of “negative” examples in designing machine learning experiments such as this is fraught with simplifying assumptions. In Chapter III, I used randomized amino acid sequences derived from native proteins to represent the “non-interacting” class to a learning machine. Subsequently, my thinking on how such examples should be properly constructed evolved: if the experimentally-derived training data set is comprehensive, we can make some educated guesses about the prevalence of protein interactions in Nature, and label the balance of the protein pairs within a proteome as belonging to the negative class. This is a good starting position, one that is reasonable based upon our present expectations given the limited amount of experimental protein-protein interaction data currently available. Unfortunately, this approach makes an assumption regarding the distribution of positive and negative examples in Nature. At present there simply is not sufficient experimental data available (in either volume or species diversity) to firmly solidify our confidence in

this assumption. Does the number of protein-protein interactions scale directly with the size of a proteome? What are the differences in numbers of protein interactions observed between, say, bacteria and the eukaryotes? When more empirical data becomes available, we may begin to sketch error bars on these quantities, and perhaps begin to realize a better means to specify the negative class to a learning machine.

2. *Optimization via interologs.* In Chapter V, the predicted protein interaction map for *C. jejuni* was evaluated using the method of “interologs” [195]. Interologs are conserved interactions between species that are postulated on the basis of high sequence similarity between paired proteins in an experimental map and their corresponding orthologs in a second proteome. It would be very interesting to use the interolog recovery rate in a predicted protein interaction network as the criterion to optimize the parameters of a support vector machine—the architecture and learning parameters associated with the highest interolog recovery would be associated with the most sensitive discrimination in this scheme. A specific question to be addressed would be the degree to which an SVM trained in such a scheme would learn to detect orthology as opposed to more subtle patterns relating sets of amino acid sequence features.
3. *Features.* A natural extension of this research would be to optimize the 1-D feature representation. In Chapter III it was noted that there are hundreds of different tabulated metrics of residue properties available in the literature; many of these would undoubtedly be useful to represent amino acid features to the learning algorithm. This research only barely scratched the surface of the possible one-dimensional representations of a protein. Further, the premise that “sequence determines structure determines function” implies that higher-order structural information may increase the prediction performance. Protein secondary structure is expressible as a string of characters, for example. Hunter and Subramaniam [88] recently proposed a parsimonious one-dimensional structural description which uses only a single continuous variable per amino acid to represent the  $C_\alpha$  backbone. It would be an interesting experiment to study the use of such structural descriptors in protein interaction predictions, and to compare and contrast their predictive success with that of the physicochemical descriptors used in this thesis.

# Appendices



# Appendix A: Support vector machine

A support vector machine (SVM) is a classification device constructed by building a decision surface in feature space to optimally separate classes of data [191]. One way to do this is to locate this surface such that the closest point of approach of data points representing each class is maximized.

A schematic SVM is depicted in Figure .1 for two classes, as represented by the blue squares (Class  $A^-$ ) and red triangles (Class  $A^+$ ). These objects represent  $n$ -dimensional, real-valued data vectors  $x$ . The separating hyperplane  $x^T w = \gamma$  (indicated by a solid line in the figure) is offset by a distance  $\gamma/\|w\|$  from the origin of coordinates, and its orientation is defined by the unit normal vector  $w/\|w\|$ , where  $w \in \mathbb{R}^n$  and  $\gamma$  is a constant. The *margin* surrounding the separating hyperplane is defined by the (dashed) parallel bounding planes  $x^T w = \gamma \pm 1$ . The margin width is  $2/\|w\|$ . If the data were linearly separable, no violations of the margin would be observed; however the figure shows several instances of data points that have exceeded the soft margin. This means that accurate class discrimination for this case dictates a nonlinear separating plane.

For a given set of data, the objective is to find the parameters  $w, \gamma$  defining the hyperplane which optimally separates classes  $A^+$  and  $A^-$ . Numerically, these parameters may be computed using a variety of algorithmic strategies (e.g., see [95, 103, 157, 171, 126, 173]). Once this is accomplished, novel data points  $x$  can be classified according to their location relative to this decision surface in feature space. Notice from Figure .1 that only a subset of the training data essentially define the margin; these are the *support vectors*, and appear as the symbols containing black dots.

The SVM is simply a linear combination of kernel function evaluations  $K(x, x_i^T)$ ,  $i = 1, \dots, k$ , where  $x$  is the input vector,  $x_i$  are the support vectors and  $k$  is their cardinality. The kernel function  $K$  measures the similarity of  $x$  to each support vector  $x_i$ . SVM maps input

data  $x$  into a feature space  $F$  via a nonlinear map

$$\Phi : x \in \mathbb{R}^n \mapsto F \in \mathbb{R}^D \quad (\text{A-1})$$

(where in general,  $D \gg n$ ) and constructs a linear separation in this high(possibly infinite)-dimensional space. In [34] it was observed that because  $\Phi$  enters the optimization and classification problems (see Eqs. A-5, A-6 below) as inner products, finding an expression for inner products in feature space  $F \in \mathbb{R}^D$  in terms of input data points  $x \in \mathbb{R}^n$  would obviate the requirement to discover and compute the feature map  $\Phi$ . Symmetric functions  $K$  with certain properties were proposed to implement this idea:

$$K(x_1, x_2) = \Phi(x_1) \cdot \Phi(x_2) \quad (\text{A-2})$$

This equivalence facilitates computational efficiency as represented by the dot product evaluations, and an implicit mapping without the need to specify  $\Phi$ . Any linear algorithm computable in terms of dot products can be made nonlinear in this manner by substitution of an appropriate kernel [170]. In the case of a gaussian kernel

$$K(x_1, x_2) = \exp(-||x_1 - x_2||^2 / (2\sigma^2)) \quad (\text{A-3})$$

$F$  has infinite dimension, however an SVM can be readily computed to construct a linear separation of data classes within this space [42].

To describe the basic equations that must be solved to construct the SVM, we use the compact matrix notation as in [126] and [19]. Suppose that the training examples are assembled in a matrix  $A$ , where  $A \in \mathbb{R}^{l \times n}$ , and  $l$  is the number of examples used to train the system. Each row of  $A$  contains a vector of features  $x \in \mathbb{R}^n$ . The class labels corresponding to these examples are appear in the diagonal matrix  $D \in \mathbb{R}^{l \times l}$ , with  $D_{ii} \in \{+1, -1\}$ .

The constrained quadratic optimization problem to be solved is [191]

$$\begin{aligned} \min_{w, \gamma, \xi} \quad & C e^T \xi + \frac{1}{2} ||w||^2 \\ \text{s.t.} \quad & D \{ A w + e \gamma \} \geq e - \xi \\ & \xi \geq 0, \quad C > 0 \end{aligned} \quad (\text{A-4})$$

where  $\xi$  is a vector of “slack” variables allowing for margin errors,  $e$  is a vector of ones, and both  $\xi, e \in \mathbb{R}^l$ . The user-selected constant  $C$  in the objective function controls the amount of penalty assigned to training example errors during optimization. If  $C$  is small,

the margin is maximized, but many points step over the margin; if  $C$  is large, a narrower margin is produced, with a minimum number of errors during training. Anecdotal evidence suggests that enhanced generalization capability may be realized if allowance is made for some training errors during the construction of the SVM [42].

In practical application the following dual formulation of Equation A-4 is solved for  $u \in \mathbb{R}^l$

$$\begin{aligned} \max_u \quad & e^T u - \frac{1}{2} u^T D A A^T D u \\ \text{s.t.} \quad & e^T D u = 0 \\ & 0 \leq u \leq C e \end{aligned} \tag{A-5}$$

and the primal variables defining the decision surface  $(w, \gamma)$  and the slack values  $\xi$  are obtained after subsequent processing steps [125]. Once this is done, the nonlinear classification decision for input vector  $x$  is

$$h(x) = \text{sgn} \left\{ K(x^T, A^T) D u - \gamma \right\} \tag{A-6}$$

where the signum function is computed by  $\text{sgn}(z) = z/|z|$ . Note that the kernel function  $K$  in Eq. A-6 is only evaluated for training patterns (rows of  $A$ ) corresponding to nonzero dual variables  $u$  from optimization Eq. A-5. These patterns are the support vectors defining the decision boundary.

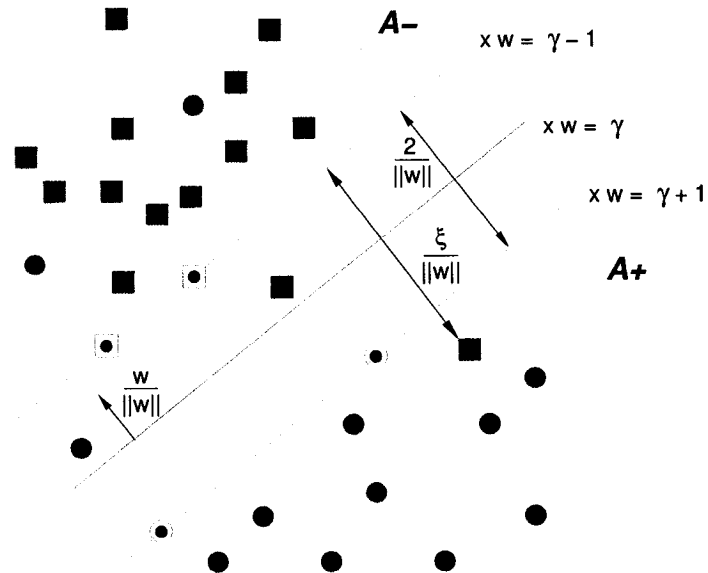


Figure .1: A schematic support vector machine for data falling into two classes:  $A^+$  (red triangles) and  $A^-$  (blue squares). In the case shown here, the classes are linearly inseparable; the SVM has been constructed using a linear kernel. Support vectors are the symbols lying on the margin containing black dots. After [42, 126].

# Appendix B: Fixed length vector algorithm

This Appendix presents source code for mapping variable-length protein features onto a fixed-length interval for SVM classification and regression analysis. A number of investigators requested explicit details on how this mapping was carried out, as insufficient mathematical details appeared in the original reference [25].

The code is written in Java, but should be easily transcribed to other programming languages of choice. I make no claims about the numerical efficiency or precision of the results of using these methods; they are admittedly inelegant, inefficient and brute-force.

```
import java.math.*;
import java.util.Vector;

/**
 * This class contains methods used to interpolate
 * arrays of data.
 *
 * @author Joel R. Bock
 * @copyright (c) 2002 by Joel R. Bock.
 */
public class Interp
{
    public static final String cid="INTERP: ";
    static String DASH="--";

    /**
     * Interpolate(extrapolate) input signal "yin" onto
     * smaller(bigger) length signal "yout".
     * Output signal length is "oLen"
     * @param yin Original data array
```

```

* @param oLen The desired length of the
*     interpolated array, on output.
*/
public static double[] interp(double[] yin,int oLen)
{
    String mid="interp(): ";
    String serr=null;
    int i=0;
    int j=0;
    int iLen=0;
    double yInVal=0.0f;
    double yOutVal=0.0f;
    double yInValOld=0.0f;
    double yOutValOld=0.0f;
    double xVal=0.0f;
    double xValOld=0.0f;
    double m=0.0f;
    double[] yout=null;
    double[] XiOut=null;
    double[] XiIn=null;
    double XiInOld=0.0f;
    double XiOutOld=0.0f;
    double dXi=0.0f;
    double dY=0.0f;
    boolean found=false;
    boolean shrink=false;
    double x0=0.0f;
    double x1=0.0f;
    double y0=0.0f;
    double y1=0.0f;
    Vector iBinV=new Vector();
    Vector oBinV=new Vector();
    int bin=0;

    if((yin==null)|| (yin.length<1))
        prex(cid+mid+"Input array 'yin' is invalid...");
    if(oLen<1)
        prex(cid+mid+"Input param 'oLen' is invalid...");
    iLen=yin.length;           //--- input length
    yout=new double[oLen];     //--- output range
    XiOut=new double[oLen];    //--- output domain
    XiIn=new double[iLen];     //--- input domain

    //--- is output array smaller (default) or larger

```

```

//--- than the input array?
shrink=((oLen>iLen)?false:true);

//--- discretize output domain: 0 <= XiOut <= 1
dXi=1.0f/(oLen-1);
XiOut[0]=0.0f;
XiOut[oLen-1]=1.0f;
for (i=1;i<(oLen-1);i++)XiOut[i]=XiOut[i-1]+dXi;

//--- discretize input domain: 0 <= XiIn <= 1
dXi=1.0f/(iLen-1);
XiIn[0]=0.0f;
XiIn[iLen-1]=1.0f;
for(i=1;i<(iLen-1);i++)XiIn[i]=XiIn[i-1]+dXi;

if(shrink)
{
    //--- loop over output domain
    for(i=0;i<oLen;i++)
    {
        found=false;
        while(!found)
        {
            XiInOld=0.0f;
            //--- loop over input domain
            for(j=0;j<iLen;j++)
            {
                yInVal=yin[j];
                yInValOld=yInVal;
                //--- find bracket
                if((XiOut[i]>=XiInOld)&&(XiOut[i]<=XiIn[j]))
                {
                    if((XiIn[j]-XiInOld)!=0.0f)
                        m=(yInVal-yInValOld)/(XiIn[j]-XiInOld);
                    else
                        m=0.0f;
                    yout[i]=yInValOld+(m*(XiOut[i]-XiInOld));
                    found=true;
                    yInValOld=yInVal;
                    break;
                }
                XiInOld=XiIn[j];
                yInValOld=yInVal;
            } // End: for(j=0;j<iLen;j++)
        }
    }
}

```

```

        } // End: while
    } // End: for(i=0;i<oLen;i++)
}
else
{
    //--- shrink=false
    //--- "bin" input domain
    for(i=0;i<(iLen-1);i++)
        iBinV.add(XiIn[i]+DASH+XiIn[i+1]);
    //--- assign output domain to bins
    for(i=0;i<oLen;i++)
        oBinV.add(""+getBin(iBinV,XiOut[i]));
    //--- Knowing domain bins for each output
    //--- domain, interpolate to find their range
    //--- values
    //--- Set endpoints
    yout[0]=yin[0];
    yout[oLen-1]=yin[iLen-1];
    for(i=1;i<(oLen-1);i++)
    {
        bin=Integer.parseInt((String)oBinV.elementAt(i));
        //--- last edge
        x0=Double.parseDouble(
            getTokByIndex((String)iBinV.elementAt(bin),
                1,DASH));
        //--- next edge
        x1=Double.parseDouble(
            getTokByIndex((String)iBinV.elementAt(bin),
                2,DASH));
        //--- local slope
        m=(yin[bin+1]-yin[bin])/(x1-x0);
        dXi=XiOut[i]-x0;
        yout[i]=yin[bin]+m*dXi;
    }
} // End: if(shrink)

//--- return array
return yout;
} // End: interp

/**
 * Assign "xi" to a bin in "v"
 * Elements of "v" are Strings composed
 * of dash-separated bin edges,
 * e.g. "0-0.333"

```



```

*/
private static int getBin(Vector v,double xi)
{
    String mid="getBin(): ";
    String serr=null;
    double x0=0.0f;
    double x1=0.0f;
    double xtest=0.0f;
    int bin=0;
    String s=null;
    boolean found=false;

    try
    {
        for(int i=0;i<v.size();i++)
        {
            s=(String)v.elementAt(i);
            x0=Double.parseDouble(getTokByIndex(s,1,DASH));
            x1=Double.parseDouble(getTokByIndex(s,2,DASH));
            if(between(x0,x1,xi))
            {
                found=true;
                bin=i;
                break;
            }
        }
    }
    catch(NumberFormatException e)
    {
        prex(cid+mid+"NumberFormatException: "+e.getMessage());
    }
    //--- return value
    return bin;
} // End: getBin

/**
 * "Between" function
 * Test whether the input value lies
 * within indicated limits, i.e.
 *  $x_0 \leq x \leq x_1$ 
 */
public static boolean between(double x0,
    double x1,double xtest)
{

```

```

String mid="between(): ";
boolean isBetween=false;

if((xtest>x1)|| (xtest<x0))return false;

if((x0<=xtest)&&(x1>=xtest))isBetween=true;

//--- return value
return isBetween;
} // End: isBetween

/**
 * Get String from input line by token index.
 * @param line The String to parse.
 * @param idx The index of the token desired.
 * @param delim A list of delimiters to use in tokenizing.
 * @return The indicated token on success; else null.
 */
public static String getTokByIndex(String line,
    int idx,String delim)
{
    String mid="getTokByIndex(): ";
    StringTokenizer st=null;
    boolean done=false;
    String tok=null;
    String theTok=null;
    int count=0;

    if(delim!=null)
        st=new StringTokenizer(line,delim);
    else
        st=new StringTokenizer(line);

    while((!done)&&(st.hasMoreTokens()))
    {
        tok=st.nextToken();
        count++;
        if(count==idx)
        {
            theTok=tok;
            break;
        }
    }
    //--- return object

```

```

        if(theTok!=null)
            return theTok;
        else
            return ("");
    } // End: getTokByIndex

/**
 * Print to System.out and exit
 */
public static void prex(String s)
{
    String mid="prex(): ";
    String serr=null;
    if(s==null)
    {
        serr=cid+mid+"Null input 's'; must be set...";
        pr(serr);
        return;
    }
    pr(s+"\n...[System.exit(0) called]");
    ex();
} // End: prex

/** System.exit(0) */
public static void ex()
{
    String mid="ex(): ";
    System.exit(0);
} // End: ex

/** System.exit(0) with message */
public static void ex(String msg)
{
    String mid="ex(): ";
    System.out.println(msg);
    System.exit(0);
} // End: ex

} // End: Interp

```

# Bibliography

- [1] B Alberts. The cell as a collection of protein machines: Preparing the next generation of molecular biologists. *Cell*, 92:291–294, February 6 1998.
- [2] B Alberts, D Bray, J Lewis, M Raff, K Roberts, and JD Watson. *Molecular Biology of the Cell*. Garland Publishing, New York, NY, 2nd edition, 1989.
- [3] R B Altman. Challenges for intelligent systems in biology. Technical Report SMI-2002-0913, Stanford Medical Informatics, 2001.
- [4] SF Altschul and W Gish. Local alignment statistics. *Methods in Enzymology*, 266:460–480, 1996.
- [5] SF Altschul, TL Madden, AA Schaffer, J Zhang, Z Zhang, W Miller, and DJ Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, September 1 1997.
- [6] L Anderson and J Seilhamer. A comparison of selected mRNA and protein abundances in human liver. *Electrophoresis*, 18(3-4):533–537, March-April 1997.
- [7] SC Andrews. Iron storage in bacteria. *Advances in Microbial Physiology*, 40:281–351, 1998.
- [8] CB Anfinsen. Principles that govern the folding of protein chains. *Science*, 181:223–230, 1973.
- [9] Anonymous. The promise of proteomics. *Nature*, 402:703, 1999.
- [10] R Apweiler, TK Attwood, A Bairoch, and A Bateman *et al.* The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Research*, 29(1):37–40, January 2001.
- [11] J Augen. The evolving role of information technology in the drug discovery process. *Drug Discovery Today*, 7(5):315–323, March 1 2002.
- [12] GD Bader and CW Hogue. BIND—a data specification for storing and describing biomolecular interactions, molecular complexes and pathways. *Bioinformatics*, 16(5):465–477, May 2000.

- [13] A Bairoch, P Bucher, and K Hofmann. The PROSITE database, its status in 1997. *Nucleic Acids Research*, 25(1):217–221, January 1 1997.
- [14] P Baldi and S Brunak. *Bioinformatics: The machine learning approach*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 1998.
- [15] P Baldi, S Brunak, Y Chauvin, C Andersen, and H Nielsen. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16(5):412–424, May 2000.
- [16] P Bartel, CT Chien, R Sternglanz, and S Fields. Elimination of false positives that arise in using the two-hybrid system. *Biotechniques*, 14(6):920–924, June 1993.
- [17] PL Bartel and S Fields, editors. *The Yeast Two-Hybrid System*. Advances in Molecular Biology. Oxford University Press, 1997.
- [18] A Bateman, E Birney, R Durbin, SR Eddy, KL Howe, and EL Sonnhammer. The Pfam protein families database. *Nucleic Acids Research*, 28(1):263–266, January 1 2000.
- [19] KP Bennett and EJ Bredensteiner. Duality and geometry in SVM classifiers. In P Langley, editor, *Proceedings of the Seventeenth International Conference on Machine Learning (ICML'2001)*, pages 57–64, Stanford University, Palo Alto, CA, June 29-July 2 2000. Morgan Kaufmann.
- [20] HM Berman, J Westbrook, Z Feng, G Gilliland, TN Bhat, H Weissig, IN Shindyalov, and PE Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, January 2000.
- [21] C Bissantz, G Folkers, and D Rognan. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *Journal of Medicinal Chemistry*, 43(25):4759–4767, December 14 2000.
- [22] DL Black. Protein diversity from alternative splicing: A challenge for bioinformatics and post-genome biology. *Cell*, 103:367–370, October 27 2000.
- [23] WP Blackstock and MP Weir. Proteomics: quantitative and physical mapping of cellular proteins. *Trends in Biotechnology*, 17(3):121–127, March 1999.
- [24] C Blaschke, MA Andrade, C Ouzounis, and A Valencia. Automatic extraction of biological information from scientific text: protein-protein interactions. In T Lengauer, R Schneider, P Bork, D Brutlag, J Glasgow, H-W Mewes, and R Zimmer, editors, *Proceedings of The Seventh International Conference on Intelligent Systems for Molecular Biology (ISMB-99)*, page 60, Menlo Park, California, August 6-10 1999. AAAI Press.
- [25] JR Bock and DA Gough. Predicting protein-protein interactions from primary structure. *Bioinformatics*, 17:455–460, 2001.

- [26] JR Bock and DA Gough. Machine learning inference of protein-protein binding in *Saccharomyces cerevisiae*. In review, August 2002.
- [27] JR Bock and DA Gough. A new method to estimate ligand-receptor energetics. *Molecular and Cellular Proteomics*, 1:904–910, November 2002.
- [28] JR Bock and DA Gough. Whole-proteome interaction mining. *Bioinformatics*, 19(1):125–134, January 2003.
- [29] HJ Böhm. The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *Journal of Computer Aided Molecular Design*, 8(3):243–256, June 1994.
- [30] HJ Böhm. Prediction of binding constants of protein ligands: a fast method for the prioritization of hits obtained from *de novo* design or 3D database search programs. *Journal of Computer Aided Molecular Design*, 12(4):309–323, 1998.
- [31] RS Boikess and E Edelson. *Chemical Principles*. Harper & Row, New York, NY, 2nd edition, 1981.
- [32] KL Borden. RING fingers and B-boxes: zinc-binding protein-protein interaction domains. *Biochemistry and Cell Biology*, 76(2-3):351–358, 1998.
- [33] P Bork and EV Koonin. Predicting functions from protein sequences—where are the bottlenecks? *Nature Genetics*, 18:313–318, April 1998.
- [34] BE Boser, IM Guyon, and VN Vapnik. A training algorithm for optimal margin classifiers. In D Haussler, editor, *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory*, pages 144–152, Pittsburgh, PA, 1992. ACM Press.
- [35] PS Bradley, UM Fayyad, and OL Mangasarian. Mathematical programming for data mining: Formulations and challenges. Technical Report MSR-98-01, University of Wisconsin, Data Mining Institute, Madison, WI, January 1998.
- [36] V Brendel, P Bucher, I Nourbakhsh, BE Blaisdell, and S Karlin. Methods and algorithms for statistical analysis of protein sequences. *Proceedings of the National Academy of Sciences USA*, 89(6):2002–2006, March 15 1992.
- [37] JR Brown, CJ Douady, MJ Italia, WE Marshall, and MH Stanhope. Universal trees based on large combined protein sequence data sets. *Nature Genetics*, 28:281–285, July 2001.
- [38] MP Brown, WN Grundy, D Lin, N Cristianini, CW Sugnet, TS Furey, M Ares, and D Haussler. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences USA*, 97(1):262–267, January 4 2000.
- [39] PO Brown and D Botstein. Exploring the new world of the genome with DNA microarrays. *Nature Genetics*, 21(1 Suppl.):33–37, January 1999.

- [40] HB Bull and K Breese. Surface tension of amino acid solutions: a hydrophobicity scale of the amino acid residues. *Archives of Biochemistry and Biophysics*, 161(2):665–670, April 2 1974.
- [41] FR Burden. Molecular identification number for substructure searches. *Journal of Chemical Information and Computer Sciences*, 29(3):225–227, August 1989.
- [42] C Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:121–167, 1998.
- [43] G Casari, C Sander, and A Valencia. A method to predict functional residues in proteins. *Nature Structural Biology*, 2(2):171–178, February 1995.
- [44] G Cauwenberghs and T Poggio. Incremental and decremental support vector learning. In *Advances in Neural Information Processing (NIPS 2000)*, Cambridge, MA, 2001. MIT Press.
- [45] YZ Chen and DG Zhi. Ligand-protein inverse docking and its potential use in the computer search of protein targets of a small molecule. *Proteins: Structure, Function, and Genetics*, 43:217–226, 2001.
- [46] SA Chervitz, L Aravind, G Sherlock, CA Ball, EV Koonin, and SS Dwight. Comparison of the complete protein sets of worm and yeast: Orthology and divergence. *Science*, 282:2022–2028, December 11 1998.
- [47] D Christendat, A Yee, A Dharamsi, and Y Kluger *et al.* Structural proteomics of an archaeon. *Nature Structural Biology*, 7(10):903–909, October 2000.
- [48] E Coward. Shufflet: shuffling sequences while conserving the k-let counts. *Bioinformatics*, 15(12):1058–1059, December 1999.
- [49] MJ Cunningham. Genomics and proteomics: The new millennium of drug discovery and development. *Journal of Pharmacological and Toxicological Methods*, 44(1):291–300, July-August 2000.
- [50] T Dandekar, B Snel, M Huynen, and P Bork. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends in Biochemical Sciences*, 23(9):324–328, 1998.
- [51] AK Das, RW Cohen, and D Barford. The structure of the tetratricopeptide repeats of protein phosphatase 5: implications for TPR-mediated protein-protein interactions. *The EMBO Journal*, 17(5):1192–1199, March 2 1998.
- [52] M Deng, S Metah, F Sun, and T Chen. Inferring domain-domain interactions from protein-protein interactions. Extended abstract, April 18-21 2002. ACM-SIGACT Sixth Annual International Conference on Computational Molecular Biology (RECOMB02).
- [53] T Dietterich. Lecture notes for CS534: Machine Learning. Computer Science Department, Oregon State University, Corvallis, OR, Spring 2001.

- [54] A Dove. Proteomics: translating genomics into products? *Nature Biotechnology*, 17(3):233–236, March 1999.
- [55] BJ Druker, MT Talpaz, DJ Resta, and B Peng *et al.* Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia and acute lymphoblastic leukemia. *New England Journal of Medicine*, 344(14):1031–1037, April 5 2001.
- [56] AM Edwards, CH Arrowsmith, and B des Pallieres. Proteomics: New tools for a new era. *Modern Drug Discovery*, 5(7):35, September 2000.
- [57] B Efron and G Gong. A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, 37:36–48, 1983.
- [58] JA Eisen. Phylogenomics: Improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Research*, 8(3):163–167, March 1998.
- [59] JA Eisen. Assessing evolutionary relationships among microbes from whole-genome analysis. *Current Opinion in Microbiology*, 3:475–480, 2000.
- [60] MB Eisen, PT Spellman, PO Brown, and D Bottstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences USA*, 95:14863–14868, December 1998.
- [61] D Eisenberg. Three-dimensional structure of membrane and surface proteins. *Annual Review of Biochemistry*, 53:595–623, 1984.
- [62] C Elkan. The foundations of cost-sensitive learning. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 973–978, Seattle, WA, USA, August 2001.
- [63] AJ Enright, I Iliopoulos, NC Kyrpides, and CA Ouzounis. Protein interaction maps for complete genomes based on gene fusion events. *Nature*, 402:86–90, November 4 1999.
- [64] AS Fanning and JM Anderson. Protein-protein interactions: PDZ domain networks. *Current Biology*, 6(11):1385–1388, November 1 1996.
- [65] S Fields and O-K Song. A novel genetic system to detect protein-protein interactions. *Nature*, 340(6230):245–246, July 20 1989.
- [66] RL Finley and R Brent. Interaction mating reveals binary and ternary connection between *Drosophila* cell cycle regulators. *Proceedings of the National Academy of Sciences USA*, 91:12980–12984, December 1994.
- [67] WM Fitch. Distinguishing homologous from analogous proteins. *Systematic Zoology*, 19(2):99–113, June 1970.
- [68] WM Fitch. Random sequences. *Journal of Molecular Biology*, 163:171–176, 1983.



- [69] WM Fitch. Homology: a personal view on some of the problems. *Trends in Genetics*, 16(5):227–31, May 2000.
- [70] RA Fridell, LS Harding, HP Bogerd, and BR Cullen. Identification of a novel human zinc finger protein that specifically interacts with the activation domain of lentiviral Tat proteins. *Virology*, 209(2):347–357, June 1 1995.
- [71] MY Galperin and EV Koonin. Who’s your neighbor? New computational approaches for functional genomics. *Nature Biotechnology*, 18(6):609–613, June 2000.
- [72] AC Gavin, M Bosche, R Krause, and P Grandi *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868):141–147, January 10 2002.
- [73] JA Gerlt and PC Babbitt. Can sequence determine function? *Genome Biology*, 1(5):REVIEWS0005, 2000.
- [74] NA Gershenfeld and AS Weigend. *The future of time series: Learning and understanding*, volume XV of *Sante Fe Institute Studies in the Sciences of Complexity*, pages 1–70. Addison-Wesley, Reading, MA, 1993.
- [75] M Gerstein, N Lan, and R Jansen. Proteomics: Integrating interactomes. *Science*, 295(5553):284–287, January 11 2002.
- [76] M B Gillies. *Computational studies of protein-ligand molecular recognition*. PhD thesis, Universiteit Utrecht, The Netherlands, April 19 2001.
- [77] A Goffeau, BG Barrell, H Bussey, and RW Davis *et al.* Life with 6000 genes. *Science*, 274(5287):563–567, October 25 1996.
- [78] H Gohlke and G Klebe. Statistical potentials and scoring functions applied to protein-ligand binding. *Current Opinion in Structural Biology*, 11(2):231–235, April 2001.
- [79] GH Golub and CF van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, MD, 2nd edition, 1989.
- [80] SM Gomez and A Rzhetsky. Towards the prediction of complete protein-protein interaction networks. In RB Altman, AK Dunker, L Hunter, K Lauderdale, and TE Klein, editors, *Biocomputing 2002: Proceedings of the Pacific Symposium*. World Scientific, January 2002.
- [81] SP Gygi, Y Rochon, BR Franza, and R Aebersold. Correlation between protein and mRNA abundance in yeast. *Molecular and Cellular Biology*, 19(3):1720–1730, March 1999,.
- [82] J Hasty and JJ Collins. Protein interactions: Unspinning the web. *Nature*, 411(6833):30–31, May 3 2001.

- [83] V Hatzimanikatis and KH Lee. Dynamical analysis of gene networks requires both mRNA and protein expression information. *Metabolic Engineering*, 1(4):275–281, October 1999.
- [84] RD Head, ML Smythe, TI Oprea, CL Waller, SM Green, and GR Marshall. VALIDATE: a new method for the receptor-based prediction of binding affinities of novel ligands. *Journal of the American Chemical Society*, 118:3959–3969, 1996.
- [85] R Hecht-Nielsen. *Neurocomputing*. Addison-Wesley, Reading, MA, 1989.
- [86] S Henikoff and JG Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences USA*, 89:10915–10519, 1992.
- [87] TP Hopp and KR Woods. Prediction of protein antigenic determinants from amino acid sequences. *Proceedings of the National Academy of Sciences, USA*, 78(6):3824–3828, June 1981.
- [88] CG Hunter and S Subramaniam. Natural coordinate representation for the protein backbone structure. *Proteins: Structure, Function, and Genetics*, 49(2):206–215, November 1 2002.
- [89] M Huynen, T Dandekar, and P Bork. Differential genome analysis applied to the species-specific features of *Helicobacter pylori*. *FEBS Letters*, 426(1):1–5, April 10 1998.
- [90] M Huynen, B Snel, W Lathe, and P Bork. Predicting protein function by genomic context: Quantitative evaluation and qualitative inferences. *Genome Research*, 10(8):1204–1210, August 2000.
- [91] T Ito, T Chiba, R Ozawa, M Yoshida, M Hattori, and Y Sakaki. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences USA*, 98(8):4569–4574, April 10 2001.
- [92] T Jaakkola, M Diekhans, and D Haussler. A discriminative framework for detecting remote protein homologies. *Journal of Computational Biology*, 7(1-2):95–114, 2000.
- [93] H Jeong, SP Mason, A-L Barabási, and ZN Oltvai. Lethality and centrality in protein networks. *Nature*, 411:41–42, May 3 2001.
- [94] T Joachims. Estimating the generalization performance of an SVM efficiently. Technical Report LS-8 Report 25, Universität Dortmund Fachbereich Informatik, Dortmund, December 29 1999.
- [95] T Joachims. *Making Large-Scale Support Vector Machine Learning Practical*. In: *Advances in Kernel Methods: Support Vector Learning*, chapter 11, pages 169–184. MIT Press, Cambridge, MA, 1999.
- [96] GH John, R Kohavi, and K Pfleger. Irrelevant features and the subset selection problem. In *Machine Learning: Proceedings of the Eleventh International Conference*

- (*ICML-94*), pages 121–129, New Brunswick, NJ, July 10-13 1994. Morgan Kaufmann.
- [97] S Jones and JM Thornton. Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences USA*, 93:13–20, 1996.
- [98] S Jones and JM Thornton. Prediction of protein-protein interaction sites using patch analysis. *Journal of Molecular Biology*, 272(1):133–143, September 12 1997.
- [99] A Kanapin, R Apweiler, M Biswas, and W Fleischmann *et al.* Interactive InterPro-based comparisons of proteins in whole genomes. *Bioinformatics*, 18(2):374–375, February 2002.
- [100] D Kandel, Y Mathias, R Unger, and P Winkler. Shuffling biological sequences. *Discrete Applied Mathematics*, 71:171–185, 1996.
- [101] M Kanehisa. *Post-genome Informatics*. Oxford University Press, Oxford, UK, 2000.
- [102] PD Karo. An ontology for biological function based on molecular interactions. *Bioinformatics*, 16(3):269–285, March 2000.
- [103] L Kaufman. *Solving the quadratic programming problem arising in support vector classification*. In: *Advances in Kernel Methods: Support Vector Learning*, chapter 10, pages 147–167. MIT Press, Cambridge, MA, 1999.
- [104] M G Kendall. A new measure of rank correlation. *Biometrika*, 30:81–93, 1938.
- [105] RD King, A Karwath, A Clare, and L Dehaspe. Genome scale prediction of protein functional class from sequence using data mining. In *Proceedings of The Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2000)*, pages 384–389, Boston, MA, August 20-23 2000.
- [106] RM Kini and HJ Evans. Prediction of potential protein-protein interaction sites from amino acid sequence. identification of a fibrin polymerization site. *FEBS Letters*, 385(1-2):81–86, April 29 1996.
- [107] S Klumpp and J Krieglstein. Phosphorylation and dephosphorylation of histidine residues in proteins. *European Journal of Biochemistry*, 269(4):1067–1071, February 2002.
- [108] R Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1137–1145, Montreal, Quebec, August 20-25 1995.
- [109] R Kohavi and F Provost. Glossary of terms. *Machine Learning*, 30:271–274, 1998.
- [110] E Kreysig. *Advanced Engineering Mathematics*. Wiley & Sons, New York, NY, 5th edition, 1983.

- [111] M Kubat, R Holte, and S Matwin. Learning when negative examples abound. In M van Someren and G Widmer, editors, *Proceedings of the European Conference on Machine Learning (ECML'97)*, pages 146–153, April 23-25 1997.
- [112] M Kubat and S Matwin. Addressing the curse of imbalanced training sets: One-sided selection. In *Proceeding of the 14th International Conference on Machine Learning*, pages 179–186. Morgan Kaufmann, 1997.
- [113] H Kubinyi. The design of combinatorial libraries. *Drug Discovery Today*, 7(9):503–504, May 1 2002.
- [114] JT Kwok. Moderating the outputs of support vector machine classifiers. *IEEE Transactions on Neural Networks*, 10(5):1018–1031, September 1999.
- [115] ES Lander, LM Linton, B Birren, and C Nusbaum *et al.* Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, February 15 2001.
- [116] T Langer and RD Hoffmann. Virtual screening: An effective tool for lead structure discovery? *Current Pharmaceutical Design*, 7(7):509–527, May 2001.
- [117] T Lazaridis, A Masumov, and F Gandolfo. Contributions to the binding free energy of ligands to avidin and streptavidin. *Proteins: Structure, Function, and Genetics*, 47(2):194–208, May 1 2002.
- [118] Y-S Lee and M Mrksich. Protein chips: from concept to practice. *Trends in Biotechnology*, 20(12 (Suppl.)):S14–S18, 2002.
- [119] L LeGrain and L Selig. Genome-wide protein interaction maps using two-hybrid systems. *FEBS Letters*, 480:32–36, August 25 2000.
- [120] M Levitt and C Chotia. Structural patterns in globular proteins. *Nature*, 261(5561):552–558, June 17 1976.
- [121] L Lo Conte, C Chothia, and J Janin. The atomic structure of protein-protein recognition sites. *Journal of Molecular Biology*, 285(5):2177–2198, February 5 1999.
- [122] G MacBeath and SL Schreiber. Printing proteins as microarrays for high-throughput function determination. *Science*, 289(5485):1760–1763, 8 September 2000.
- [123] BB Mandelbrot. *The Fractal Geometry of Nature*. W.H. Freeman, New York, NY, 1977.
- [124] OL Mangasarian. Mathematical programming in data mining. Technical Report 96-05, University of Wisconsin, Madison, WI, August 1996.
- [125] OL Mangasarian and DR Musicant. Successive overrelaxation for support vector machines. *IEEE Transactions on Neural Networks*, 10(5):1032–1037, September 1999,.

- [126] OL Mangasarian and DR Musicant. Lagrangian support vector machines. *Journal of Machine Learning Research*, 1:161–177, March 2001.
- [127] M Mann, RC Hendrickson, and A Pandey. Analysis of proteins and proteomes by mass spectrometry. *Annual Reviews in Biochemistry*, 70:437–73, 2001.
- [128] EM Marcotte. Computational genetics: finding protein function by nonhomology methods. *Current Opinion in Structural Biology*, 10(3):359–365, June 2000.
- [129] EM Marcotte, M Pellegrini, HL Ng, DW Rice, TO Yeates, and D Eisenberg. Detecting protein function and protein-protein interactions from genome sequences. *Science*, 285(5428):751–753, July 30 1999.
- [130] EM Marcotte, M Pellegrini, MJ Thompson, TO Yeates, and D Eisenberg. A combined algorithm for genome-wide prediction of protein function. *Nature*, 402:83–86, November 4 1999.
- [131] JK Martin and DS Hirschberg. Small sample statistics for classification error rates I: Error rate measurements. Technical Report ICS-TR-96-21, Department of Information and Computer Science, University of California Irvine, July 2 1996.
- [132] M Matsushita and KD Janda. Histidine kinases as targets for new antimicrobial agents. *Bioorganic and Medicinal Chemistry*, 10(4):855–867, April 2002.
- [133] LR Matthews, P Vaglio, J Reboul, H Ge, BP Davis, J Garrels, S Vincent, and M Vidal. Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". *Genome Research*, 11(12):2120–2126, December 2001.
- [134] R Milo, S Shen-Orr, S Itzkovitz, N Kashtan, D Chklovskii, and U Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, October 25 2002.
- [135] E Mjolsness and D DeCoste. Machine learning for science: State of the art and future prospects. *Science*, 293:2051–2055, September 14 2001.
- [136] EE Moret, MC van Wijk, AS Kostense, and MB Gillies. Scoring peptide(mimetic)-protein interactions. *Medicinal Chemistry Research*, 9:604–620, 1999.
- [137] K Morik, P Brockhausen, and T Joachims. Combining statistical learning with a knowledge-based approach: A case study in intensive care monitoring. In I Bratko and S Dzeroski, editors, *Proceedings of the 16th International Conference on Machine Learning (ICML '99)*. Morgan Kaufmann, 27-30 June 1999.
- [138] K-R Müller, AJ Smola, G Rätsch, B Schölkopf, J Kohlmorgen, and V Vapnik. *Advances in Kernel Methods*, chapter 14, pages 243–253. MIT Press, Cambridge, MA, 1999.

- [139] D Mumberg, R Muller, and M Funk. Yeast vectors for the controlled expression of heterologous proteins in different genetic backgrounds. *Gene*, 156(1):119–22, April 14 1995.
- [140] T Munder and A Hinnen. Yeast cells as tools for target-oriented screening. *Applied Microbiology and Biotechnology*, 52(3):311–320, 1999.
- [141] SB Needleman and CD Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48:443–453, 1970.
- [142] EJ Neer, CJ Schmidt, R Nambudripad, and TF Smith. The ancient regulatory-protein family of WD-repeat proteins. *Nature*, 371(6495):297–300, September 22 1994.
- [143] JH Ng and LL Ilag. Functional proteomics: separating the substance from the hype. *Drug Discovery Today*, 7(9):504–505, May 1 2002.
- [144] JWM Nissink, ML Verdonk, and G Klebe. Simple knowledge-based descriptors to predict protein-ligand interactions. methodology and validation. *Journal of Computer Aided Molecular Design*, 14(8):787–803, November 2000.
- [145] AR Ortiz, MT Pisabarro, F Gago, and RC Wade. Prediction of drug binding affinities by comparative binding energy analysis. *Journal of Medicinal Chemistry*, 38(14):2681–2691, July 7 1995.
- [146] R Overbeek, M Fonstein, M D’Souza, GD Pusch, and N Maltsev. The use of gene clusters to infer functional coupling. *Proceedings of the National Academy of Sciences USA*, 96:2896–2901, March 1999.
- [147] J Parkhill, BW Wren, K Mungall, and JM Ketley *et al.* The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. *Nature*, 403(6770):665 – 668, February 10 2000.
- [148] T Pawson. Protein modules and signalling networks. *Nature*, 373(6515):573–580, February 16 1995.
- [149] T Pawson and JD Scott. Signaling through scaffold, anchoring, and adaptor proteins. *Science*, 278(5346):2075–2080, December 19 1997.
- [150] F Pazos, M Helmer-Citterich, G Ausiello, and A Valencia. Correlated mutations contain information about protein-protein interaction. *Journal of Molecular Biology*, 271(271):511–523, August 29 1997.
- [151] F Pazos and A Valencia. Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Engineering*, 14(9):609–614, September 2001.
- [152] M Pellegrini. Computational methods for protein function analysis. *Current Opinion in Chemical Biology*, 5(1):46–50, February 2001.

- [153] M Pellegrini, EM Marcotte, MJ Thompson, D Eisenberg, and TO Yeates. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proceedings of the National Academy of Sciences USA*, 96:4285–4288, April 1999.
- [154] J Pelletier and S Sidhu. Mapping protein-protein interactions with combinatorial biology methods. *Current Opinion in Biotechnology*, 12(4):340–347, August 2001.
- [155] M Perrone. *Improving Regression Estimation: Averaging methods for variance reduction with extensions to general convex measure optimization*. PhD thesis, Brown University, May 1993.
- [156] WW Peterson and TG Birdsall. The theory of signal detectability. Technical Report TR-13, Communications and Signal Processing Laboratory, University of Michigan, 1953.
- [157] J C Platt. *Fast Training of Support Vector Machines using Sequential Minimal Optimization*. In: *Advances in Kernel Methods: Support Vector Learning*, chapter 12, pages 185–208. MIT Press, Cambridge, MA, 1999.
- [158] CP Ponting, L Aravind, J Schultz, P Bork, and EV Koonin. Eukaryotic signalling domain homologues in Archaea and Bacteria. Ancient ancestry and horizontal gene transfer. *Journal of Molecular Biology*, 289(4):729–745, June 18 1999.
- [159] F Provost, T Fawcett, and R Kohavi. The case against accuracy estimation for comparing induction algorithms. In *Proceedings of the Fifteenth International Conference on Machine Learning (IMLC-98)*, pages 445–453. Morgan Kaufmann, San Francisco, CA, 1998.
- [160] JR Quinlan. *C4.5: Programs for machine learning*. Morgan Kaufmann Publishers, San Mateo, CA, 1993.
- [161] JC Rain, L Selig, H De Reuse, V Battaglia, C Reverdy, S Simon, G Lenzen, F Petel, J Wojcik, V Schächter, Y Chemama, A Labigne, and P Legrain. The protein-protein interaction map of *Helicobacter pylori*. *Nature*, 409:211–215, January 2001.
- [162] M Rarey, B Kramer, C Bernd, and T Lengauer. Time-efficient docking of similar flexible ligands. In L Hunter and T Klein, editors, *Biocomputing: Proceedings of the 1996 Pacific Symposium*, Singapore, January 3-6 1996. World Scientific Publishing.
- [163] M Remm, CE Storm, and EL Sonnhammer. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *Journal of Molecular Biology*, 314(5):1041–1052, December 14 2001.
- [164] GM Rubin, MD Yandell, JR Wortman, and GL Gabor Miklos *et al.* Comparative genomics of the eukaryotes. *Science*, 287(5461):2204–2215, March 24 2000.
- [165] D Sankoff, G Leduc, B Paquin, BF Lang, and R Cedergren. Gene order comparisons of phylogenetic inference: Evolution of the mitochondrial genome. *Proceedings of the National Academy of Sciences USA*, 89(14):6575–6579, July 15 1992.

- [166] V Schächter. Protein interaction networks: from experiments to analysis. *Drug Discovery Today*, 7(11):S48–S54, May 6 2002.
- [167] M Schapira, M Totrov, and R Abagyan. Prediction of the binding energy for small molecules, peptides and proteins. *Journal of Molecular Recognition*, 12(3):177–190, May-June 1999.
- [168] D Schmucker, JC Clemens, H Shu, CA Worby, J Xiao, M Muda, JE Dixon, and SL Zipursky. *Drosophila* Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell*, 101(6):671–684, June 9 2000.
- [169] B Schölkopf, CJ Burges, and AJ Smola, editors. *Advances in Kernel Methods*. MIT Press, Cambridge, MA, 1999.
- [170] B Schölkopf, S Mika, CJ Burges, P Knirsch, K-R Müller, G Raetsch, and AJ Smola. Input space versus feature space in kernel-based methods. *IEEE Transactions On Neural Networks*, 10(5):1000–1017, September 1999.
- [171] B Schölkopf, AJ Smola, R Williamson, and P Bartlett. New support vector algorithms. *Neural Computation*, 12:1083–1121, 2000.
- [172] B Schwikowski, P Uetz, and S Fields. A network of protein-protein interactions in yeast. *Nature Biotechnology*, 18:1257–1261, December 2000.
- [173] F Sha, LK Saul, and DD Lee. Multiplicative updates for nonnegative quadratic programming in support vector machines. Technical Report MS-CIS-02-19, University of Pennsylvania, Philadelphia, PA, 2002.
- [174] MI Skolnik. *Introduction to Radar Systems*. McGraw-Hill, New York, NY, 2nd edition, 1980.
- [175] RD Smith. Trends in mass spectrometry instrumentation for proteomics. *Trends in Biotechnology*, 20(12 (Suppl.)):S3–S7, 2002.
- [176] TF Smith and MS Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147:195–197, March 25 1981.
- [177] TF Smith, MS Waterman, and C Burks. The statistical distribution of nucleic acid similarities. *Nucleic Acids Research*, 13:645–656, 1985.
- [178] AJ Smola. Regression estimation with support vector learning machines. Master’s thesis, Physik Department, Technische Universität München, Germany, December 31 1996.
- [179] AJ Smola and B Schölkopf. A tutorial on support vector regression. Technical Report NC-TR-98-030, Royal Holloway College, University of London, October 1998.
- [180] PT Spellman and GM Rubin. Evidence for large domains of similarly expressed genes in the *Drosophila* genome. *Journal of Biology*, 1:5.1–5.8, June 18 2002.



- [181] E Sprinzak and H Margalit. Correlated sequence-signatures as markers of protein-protein interaction. *Journal of Molecular Biology*, 311(4):681–692, August 24 2001.
- [182] M Stone. Cross-validatory choices and assessment of statistical predictions. *Journal of the Royal Statistical Society*, 36(1):111–147, 1974.
- [183] K Swingler. *Applying Neural Networks: A Practical Guide*. Academic Press Limited, London, UK, 1996.
- [184] RI Tatusov, EV Koonin, and DJ Lipman. A genomic perspective on protein families. *Science*, 278(5338):631–637, October 24 1997.
- [185] F Tekaia, A Lazcano, and B Dujon. The genomic tree as revealed from whole proteome comparisons. *Genome Research*, 9(6):550–557, 1999.
- [186] AH Tong, B Drees, G Nardelli, GD Bader, B Brannetti, L Castagnoli, M Evangelista, S Ferracuti, B Nelson, S Paoluzi, M Quondam, A Zucconi, CW Hogue, S Fields, C Boone, and G Cesareni. A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science*, 295(5553):321–324, January 11 2002.
- [187] CL Tucker, JF Gera, and P Uetz. Towards an understanding of complex protein networks. *Trends in Cell Biology*, 11(3106):102–106, March 2001.
- [188] P Uetz, L Goit, G Cagney, and TA Mansfield *et al.* A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature*, 403:623–627, February 10 2000.
- [189] P Uetz and RE Hughes. Systematic and large-scale two-hybrid screens. *Current Opinion in Microbiology*, 3(3):303–308, 2000.
- [190] RJ Urick. *Principles of Underwater Sound*. McGraw-Hill, New York, NY, 3rd edition, 1983.
- [191] VN Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, Heidelberg, Germany, 1995.
- [192] JC Venter, MD Adams, EW Myers, and PW Li *et al.* The sequence of the human genome. *Science*, 291(5507):1304–1351, February 16 2001.
- [193] M Vidal. Personal communication. December 2002.
- [194] A Walhout, S Boulton, and M Vidal. Yeast two-hybrid systems and protein interaction mapping projects for yeast and worm. *Yeast*, 17:88–94, 2000.
- [195] A Walhout, S Sordella, X Lu, JL Hartley, GT Temple, MA Brasch, N Thierry-Mieg, and M Vidal. Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science*, 287:116–122, 2000.

- [196] A Walhout and M Vidal. Protein interaction maps for model organisms. *Nature Reviews Molecular Cell Biology*, 2:55–63, January 2001.
- [197] R Wang, L Lai, and S Wang. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *Journal of Computer Aided Molecular Design*, 16(1):11–26, January 2002.
- [198] R Wang, L Liu, L Lai, and Y Tang. SCORE: A new empirical method for estimating the binding affinity of a protein-ligand complex. *Journal of Molecular Modeling*, 4:379–394, 1998.
- [199] VC Wasinger, SJ Cordwell, A Cerpa-Poljak, JX Yan, AA Gooley, MR Wilkins, MW Duncan, R Harris, KL Williams, and I Humphery-Smith. Progress with gene-product mapping of the Mollicutes: *Mycoplasma genitalium*. *Electrophoresis*, 16(7):1090–1094, July 1995.
- [200] B Waszkowycz. Structure-based approaches to drug design and virtual screening. *Current Opinion in Drug Discovery and Development*, 5(3):407–413, May 2002.
- [201] J Wegner and A Zell. JOELib: A Java based computational chemistry package. In *6th Darmstädter Molecular-Modelling Workshop*, Technische Universität, Darmstadt, Germany, 2002.
- [202] D Weininger. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, 1988.
- [203] GM Weiss and F Provost. The effect of class distribution on classifier learning: An empirical study. Technical Report ML-TR-44, Department of Computer Science, Rutgers University, August 2 2001.
- [204] GW Welling, WJ Weijer, R van der Zee, and S Welling-Wester. Prediction of sequential antigenic regions in proteins. *FEBS Letters*, 188(2):215–8, September 2 1985.
- [205] IH Witten and E Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Francisco, CA, USA, 1999.
- [206] J Wojcik, IG Boneca, and P Legrain. Prediction, assessment and validation of protein interaction maps in bacteria. *Journal of Molecular Biology*, 323(4):763–770, November 1 2002.
- [207] J Wojcik and V Schächter. Protein-protein interaction map inference using interacting domain profile pairs. *Bioinformatics*, 17(Suppl. 1):S296–S305, 2001.
- [208] DH Wolpert and WG Macready. No free lunch theorems for search. Technical Report SFI-TR-95-02-010, The Santa Fe Institute, Santa Fe, NM, February 6 1995.

- [209] I Xenarios, DW Rice, L Salwinski, MK Baron, EM Marcotte, and D Eisenberg. DIP: The database of interacting proteins. *Nucleic Acids Research*, 28(1):289–291, January 1 2000.
- [210] ML Yarmush and A Jayaraman. Advances in proteomic technologies. *Annual Reviews in Biomedical Engineering*, 4:349–373, 2002.
- [211] A Zanzoni, L Montecchi-Palazzi, M Quondam, G Ausiello, M Helmer-Citterich, and G Cesareni. MINT: a molecular INTERaction database. *FEBS Letters*, 513(1):135–140, February 20 2002.
- [212] T Zhang and DE Koshland. Computational method for relative binding energies of enzyme-substrate complexes. *Protein Science*, 5(2):348–356, February 1996.
- [213] H Zhu, M Bilgin, R Bangham, D Hall, A Casamayor, P Bertone, N Lan, R Jansen, S Bidlingmaier, T Houfek, T Mitchell, P Miller, RA Dean, M Gerstein, and M Snyder. Global analysis of protein activities using proteome chips. *Science*, 293(5537):2101–2105, September 14 2001.
- [214] A Zien, G Rätsch, S Mika, B Schölkopf, C Lemmen, A Smola, T Lengauer, and K-R Müller. Engineering support vector machine kernels that recognize translation initiation sites. In *Proceedings of the German Conference on Bioinformatics 1999*, pages 37–43, Hannover, Germany, October 1999. GBF.