

**REVEALING THE CONFORMATION AND PROPERTIES
OF HUMAN GENOME, PROTEIN MOLECULES AND
PROTEIN DOMAIN CO-OCCURRENCE NETWORK**

A Dissertation presented to
the Faculty of the Graduate School
at the University of Missouri

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

by
ZHENG WANG
Dr. Jianlin Cheng, Dissertation Supervisor
July 2012

The undersigned, appointed by the Dean of the Graduate School, have examined the dissertation entitled:

**REVEALING THE CONFORMATION AND PROPERTIES
OF HUMAN GENOME, PROTEIN MOLECULES AND
PROTEIN DOMAIN CO-OCCURRENCE NETWORK**

presented by Zheng Wang,

a candidate for the degree of Doctor of Philosophy and hereby certify that, in their opinion, it is worthy of acceptance.

Dr. Jianlin Cheng

Dr. Dong Xu

Dr. Yi Shang

Dr. Gary Stacey

To My Parents. Thank You, Mom and Dad!

ACKNOWLEDGMENTS

I appreciate my Ph.D. advisor, Dr. Jianlin Cheng, for imparting knowledge and enlightening research ideas.

I appreciate the co-authors of the published and unpublished manuscripts included in this dissertation: Renzhi Cao, Dr. Jianlin Cheng, Dr. Dong Xu, Dr. Gary Stacey, Dr. Kristen Taylor, Aaron Briley, Dr. Charles Caldwell, Dr. Marc Libault, Dr. Xue-Cheng Zhang, Jesse Eickholt, Trupti Joshi, Dr. Henry T. Nguyen, Dr. Babu Valliyodan, and Mi Ha Le. I appreciate Dr. GuanNing Lin for sharing phylogeny benchmark datasets and providing ComPhy software.

I appreciate my committee members, Dr. Dong Xu, Dr. Gary Stacey, and Dr. Yi Shang, for providing suggestions and comments.

Contents

ACKNOWLEDGMENTS	ii
LIST OF TABLES	ix
LIST OF FIGURES	xii
ABSTRACT	xxii
CHAPTER	
1 Introduction	1
2 The Properties of Human Genome Conformation and Spatial Gene Interaction and Regulation Networks	5
2.1 Abstract	5
2.2 Introduction	6
2.3 Methods	9
2.3.1 Hi-C library preparation and sequencing of the primary human acute lymphoblastic leukemia B-cell (B-ALL), MHH-CALL-4 B-ALL cell line (CALL4), and lymphoma RL cell-line (RL)	9
2.3.2 Mapping Hi-C sequence reads to the reference genome	10
2.3.3 Generating intra- and inter-chromosomal contact matrices	12
2.3.4 Statistical significance analysis of chromosomal contact matrix	13
2.3.5 Detection of inter-chromosomal translocation boundaries and reconstruction of translocated chromosomes	13
2.3.6 Construction of Pearson's correlation matrix from chromosomal contact matrix	14
2.3.7 Calculation of observed/expected numbers of contacts between all pairs of chromosomes	14
2.3.8 Construction of gene-gene interaction networks from Hi-C chromosomal contact data	15

2.3.9	Construction of interaction network of transcription factor binding sites (TFBS)	16
2.3.10	Construction of interaction networks of transcription factor binding sites (TFBS) and genes	17
2.4	Results	17
2.4.1	Hi-C Read Mapping	17
2.4.2	Intra-chromosomal contacts of different cells or cell lines	18
2.4.3	Inter-chromosomal contacts and chromosomal translocations	19
2.4.4	Spatial gene-gene interactions in the HOXA gene cluster in chromosome 7 of four cell lines	25
2.4.5	Chromosome- or genome-wide spatial gene-gene interaction networks	25
2.4.6	Interaction networks of transcription factor binding sites (TFBS)	27
2.4.7	Interaction networks of transcription factor binding sites and genes	27
2.5	Conclusions	29
2.6	Acknowledgement	30
2.7	Author Contributions	30
3	APOLLO: A Quality Assessment Service for Single and Multiple Protein Models	31
3.1	Abstract	31
3.2	Introduction	32
3.3	Methods	33
3.3.1	Input and Output	33
3.3.2	Algorithms	33
3.4	Results	36
4	An Iterative Self-Refining and Self-Evaluating Approach for Protein Model Quality Estimation	38

4.1	Abstract	38
4.2	Introduction	39
4.3	Materials and Methods	40
4.4	Results and Discussions	42
4.4.1	Evaluation datasets and criteria	42
4.4.2	Evaluation in terms of correlation and loss	43
4.4.3	Evaluation in terms of ranking correlation	47
5	SoyDB: A Knowledge Database of Soybean Transcription Factors	61
5.1	Abstract	61
5.2	Background	62
5.3	Construction and Content	65
5.3.1	Database Overview	65
5.3.2	Data Source	66
5.3.3	Transcription Factor Identification	66
5.3.4	Transcription Factor Family Prediction Using SAM Hidden Markov Models	66
5.3.5	Annotations Using Bioinformatics Tools	67
5.3.6	Links to External Databases and Datasets	69
5.3.7	Database and Website Implementation	71
5.4	Utility and Discussion	71
5.4.1	Protein Information	71
5.4.2	Family Information	72
5.4.3	Protein Browsing	72
5.4.4	Family Browsing	75
5.4.5	Full Text Search	75

5.4.6	PSI-BLAST Sequence Search	75
5.4.7	Family Classification by Hidden Markov Model	77
5.4.8	FTP Site	77
5.4.9	Comparisons and Overlapping between SoyDB and PlantTFDB	77
5.4.10	Comparisons of Soybean Transcription Factor Family Distribu- tions with Other Plants	78
5.5	Future Development Plan	79
5.6	Conclusions	81
5.7	Availability and Requirements	81
6	Three-Level Prediction of Protein Function by Integrating Profile- Sequence Search, Profile-Profile Search, and Domain Co-Occurrence Networks	83
6.1	Abstract	83
6.2	Introduction	84
6.3	Methods	87
6.3.1	Overview	87
6.3.2	Predictor-1	88
6.3.3	Predictor-2	90
6.3.4	Predictor-3	91
6.4	Results and Discussion	91
6.4.1	Overview	91
6.4.2	Precision and Recall of Top n Predictions	92
6.4.3	Precision and Recall Under a Sliding Threshold on Confidence Scores	94
6.4.4	Evaluations by Semantic Similarity	98
6.4.5	An Example Illustrating the Effectiveness of Domain Co-Occurrence Networks for Protein Function Prediction	100

6.5	Conclusions	105
7	A Protein Domain Co-Occurrence Network Approach for Predicting Protein Function and Inferring Species Phylogeny	107
7.1	Abstract	107
7.2	Introduction	108
7.3	Materials and Methods	113
7.3.1	Construction of Domain Co-occurrence Networks	113
7.3.2	Domain function prediction - GO terms	114
7.3.3	Domain function prediction - enzyme family	120
7.3.4	Protein function prediction	121
7.3.5	Phylogenetic tree construction and its evaluation	122
7.4	Results and Discussion	126
7.4.1	Statistical Properties of Domain Co-occurrence Networks	126
7.4.2	Error and attack robustness of DCNs	130
7.4.3	Domain function prediction - GO terms	132
7.4.4	Domain function prediction - Enzyme	138
7.4.5	Protein function prediction	140
7.4.6	Evaluation of DCN-Inferred Phylogeny	141
7.5	Conclusions and Future Work	145
A	Supplementary Documents for Chapter 2	155
B	Statistical Properties of <i>H. sapiens</i>, <i>S. cerevisiae</i>, <i>C. elegans</i>, <i>D. melanogaster</i>, and 15 Plant Genomes	181
C	Web-based Bioinformatics Tools and Services	185
C.1	APOLLO: A Quality Assessment Service for Single and Multiple Protein Models	185
C.1.1	Overview	185

C.1.2	URL	185
C.1.3	Input	186
C.1.4	Output	186
C.1.5	Software Architecture	186
C.2	Automated Assessment of CASP8 (2008) and CASP9 (2010) Predictions	186
C.2.1	Overview	186
C.2.2	URL	186
C.2.3	Descriptions	187
	BIBLIOGRAPHY	188
	VITA	214

List of Tables

Table	Page
3.1 Results of global quality assessment methods used by APOLLO server on 107 CASP9 targets	37
4.1 Performances of CASP8 MQAPs before and after iterative refinements.	50
4.2 Performances of CASP9 MQAPs before and after iterative refinements - 1.	51
4.3 Performances of CASP9 MQAPs before and after iterative refinements - 2.	52
4.4 The average Kendall tau ranking correlation and average Spearman's ranking correlation before and after the first and last iteration tested on CASP8 MQAPs.	58
4.5 The average Kendall tau ranking correlation and average Spearman's ranking correlation before and after the first and last iteration tested on CASP9 MQAPs - 1.	59
4.6 The average Kendall tau ranking correlation and average Spearman's ranking correlation before and after the first and last iteration tested on CASP9 MQAPs - 2.	60
5.1 Distributions of transcription factor families across major plant species	82

6.1	The minimum, maximum, and average number of predictions per target of our three models and the three baseline methods.	98
6.2	The break-even values between precision and recall (i.e., when precision = recall) of the six predictors. Average values are the averages of precisions and recalls at decision thresholds yielding the closest precision and recall values.	98
7.1	The prediction accuracy of using neighbor counting, χ^2 , and SVM when predicting GO terms	134
7.2	The average semantic similarity scores of the best predictions, among “Top 1” or “Top 3” GO term(s)	134
7.3	The average recall value of using neighbor counting, χ^2 , and SVM when predicting GO terms	136
7.4	The average precision value of using neighbor counting, χ^2 , and SVM for top 3 prediction	137
7.5	Performances of neighbor-counting method on promiscuous domains .	147
7.6	Performances of χ^2 method on promiscuous domains	148
7.7	Performances of SVM-based method on promiscuous domains	148
7.8	The leave-one-out cross-validation results of the SVM-based enzyme “yes or no” predictions	149
7.9	The accuracies of using “neighbor-inference” method for enzyme “yes or no” predictions	149
7.10	Prediction accuracy of the neighbor-counting and SVM-base method when predicting EC families	150
7.11	The results of DCN-based aggregated neighbor-counting method on 66 randomly selected proteins	151
7.12	The results of DCN-based aggregated neighbor-counting method on 9 single-domain proteins	152

7.13	Accuracy comparisons between our DCNs-based method and other methods for phylogeny inference	153
7.14	The composition of dataset 6 containing 54 single-chromosome prokaryotic organisms	154
A.1	Reads coverage of the gene regions and non-gene regions. The reads coverage of gene region was calculated as the reads length multiply the number of contact in gene region / total length of gene region. The coverage of non-gene region was calculated as the read length * number of contact not in gene region / total length of non-gene region. Here the gene length was calculated according to the gene start and end information.	180
A.2	Total number of reads for all samples mentioned in this work. The data of the normal B cell was downloaded from the publication [1]. The others were generated by us. One pair-end read pair contains two ends of reads. This table shows the total number of ends. For some cell / cell lines, we sequenced them more than one times and selected the one with the best quality to use in this work.	180

List of Figures

Figure	Page
2.1 The original contact matrices (A-D), Pearson's correlation matrices (E-H), and difference matrices (I-L) of chromosome 14 for the normal human B-cell, human acute lymphoblastic leukemia B-cell, human MHH-CALL-4 B-ALL cell line, and human lymphoma RL cell-line.	20
2.2 (A) Inter-chromosomal contact heat map between chromosome 11 and 14 for the primary ALL B-cell. Three resolution thresholds (1000Kb, 100Kb, and 10Kb) are used and the chromosomal translocation boundaries are illustrated. (B-E) The observed numbers of inter-chromosomal contacts divided by the expected numbers of inter-chromosomal contacts between all pairs of human chromosomes are visualized.	23
2.3 The contact profile between genes in the HoxA gene cluster in human chromosome 7. Numbers in cells of the contact matrices are the observed number of contacts.	24

2.4	The spatial gene-gene interaction networks and the analysis of its properties. (A) The gene-gene interaction network of genes residing in chromosome 14 for the CALL-4 cell line. (B) The distribution of node degrees. (C) The histogram of shortest path lengths. (D) The plot of average clustering coefficient against of the degree (the number of neighbors) of a node. (E) The plot of closeness centralities against the degree of a node. (F) The stress distribution. (G) The plot of topological coefficients against the degree of a node. The network and its properties were visualized and analyzed by Cytoscape [2].	28
3.1	A local quality example for CASP9 target T0563	35
4.1	The average correlation, overall correlation, average loss, and average computational time under different numbers of reference models. . . .	41
4.2	The plot of the average losses against iterations for CASP8 MQAPs. .	45
4.3	The plot of the average losses against iterations for CASP9 MQAPs. .	46
4.4	The Kendall tau rank correlations of the rankings before and after each round of refinement for CASP8 MQAPs.	49
4.5	The Spearman's rank correlations of the rankings before and after each round of refinement for CASP8 MQAPs.	53
4.6	The Kendall tau rank correlations of the rankings before and after each round of refinement for CASP9 MQAPs.	54
4.7	The Spearman's rank correlations of the rankings before and after each round of refinement for CASP9 MQAPs.	55
4.8	The plot of the Spearman's RCBAF values against the average per-target correlation of the 30 CASP8 MQAPs.	56
4.9	The plot of the Kendall tau RCBAF values against the average per-target correlation of the 30 CASP8 MQAPs.	57

5.1	Architecture of SoyDB website	65
5.2	The predicted structure of a transcription factor in SoyDB	70
5.3	Information page for a transcription factor	73
5.4	Family information page in SoyDB	74
5.5	Transcription factor browsing page in SoyDB	76
5.6	Distributions of transcription factor families across major plant species	80
6.1	The architecture of the three-level prediction methodology used in our predictor-1	88
6.2	Precision and recall of our three models and three baseline methods when considering top n , $1 \leq n \leq 20$, predictions ranked by confidence scores.	95
6.3	Precision and recall of our three models and three baseline methods when considering predictions with confidence score above a threshold t , $0 \leq t \leq 1$	97
6.4	Precision and recall when progressively considering predictions with confidence score in three ranges for our model 1 predictions.	99
6.5	Average similarity scores of our three models and three baseline meth- ods for top 1-20 predictions.	101
6.6	Best similarity scores of our three models and three baseline methods for top 1-20 predictions.	102
6.7	An example (CAFA target T30248) showing how DCN-based “aggre- gated neighbor-counting” method works.	104
7.1	A small DCN consisting of two Arabidopsis proteins	111
7.2	Domain co-occurrence network of Arabidopsis thaliana	115
7.3	The relationship between sub-graphs and protein function	116
7.4	Predicting the GO terms of domain a using SVM classification	119

7.5	An example showing the DCN-based aggregated neighbor-counting method for protein function prediction	123
7.6	An example illustrating the graph alignment algorithm we utilized . .	125
7.7	Composition details of the 398 single-chromosome prokaryotic genomes (strains)	127
7.8	Statistical properties of DCNs of four representative species	129
7.9	The number of neighbors with known functions versus the prediction accuracy of the neighbor-counting method	135
7.10	The phylogenetic tree generated on 54 single-chromosome prokaryotic taxa by our DCNs-alignment-based inference method	144
A.1	An overview of the bioinformatics pipeline of analyzing Hi-C experimental reads data.	156
A.2	The visualization of reads mapped to the HoxA gene region (27,104,502 - 27,212,501) on chromosome 7 of the human genome by the UCSC genome browser. The vertical line segments under the label “chromosome contact” denote the locations where the reads were mapped to. The reads data of the MHH-CALL-4 cell line was used.	157
A.3	The distribution of the sequencing qualities (Solexa-scale) of paired-end reads of the two malignant primary ALL B-cell data sets (i.e. quality scores versus nucleotide positions). The sequencing quality score at a position is calculated as $Q_{Solexa} = -10 \log_{10} \frac{p}{1-p}$, where p is the probability of a sequencing error at the position. A score 30 means the probability of a sequencing error at the position is ~ 0.001 . A score 20 or above may be considered acceptable. The plots show the median (the black curve), 1st and 91st percentiles, 2nd and 3rd quartiles from positions 1 to 120 in the reads data.	158

A.4	The distribution of the sequencing qualities of paired-end reads of the two malignant MHH-CALL-4 cell line data sets.	159
A.5	The distribution of the sequencing qualities of paired-end reads of the two malignant lymphoma RL cell line data sets.	160
A.6	The plots of contact numbers against regions of chromosome 7, 11 and 14 of four cell samples and the plots of gene numbers against regions of chromosome 7, 11 and 14. The X-axis in Plots A-L denotes chromosomal region index at resolution 1Mb and the Y-axis denotes the number of intra- and inter-chromosomal contacts in each region. An inter-chromosomal contact is a spatial contact between two different chromosomes, and an intra-chromosomal contact a contact within the same chromosome. A, B and C are the plots of chromosomes 7, 11, and 14 for the MHH-CALL-4 cell line respectively, D, E and F for the RL cell line, G, H and I for the normal B-Cell, and J, K and L for the Primary B-ALL cell. The plots show that the number of contacts generated from the sequence data is not evenly distributed along the chromosomes. The extra M, N and O plots show the number of genes in each region against the regions of chromosome 7, 11 and 14 separately.	161
A.7	The intra-chromosomal contact heat maps for all chromosomes of the primary ALL B-cell. Interested readers may contact us for images with higher resolution and for contact matrix data.	162
A.8	The intra-chromosomal contact heat maps for all chromosomes of the MHH-CALL-4 cell line. Interested readers may contact us for images with higher resolution and for contact matrix data.	163
A.9	The intra-chromosomal contact heat maps for all chromosomes of the RL cell line. Interested readers may contact us for images with higher resolution and for contact matrix data.	164

A.10 The intra-chromosomal contact heat maps for all chromosomes for the normal B-cell line. Sequence reads data were downloaded from Lieberman-Aiden etc [1]. Mapping and construction of contact maps were carried out by our pipeline. 165

A.11 The Pearson’s correlation matrix for intra-chromosomal contact numbers between the normal B cell, primary ALL B-cell, MHH-CALL-4 cell line, and RL cell line. For each cell, the number of intra-chromosomal contacts for each of 23 pairs of chromosomes was calculated and was put into a vector. Thus, each cell sample has one intra-chromosomal contact vector. The matrix below shows the Pearson’s correlation between each pairs of vectors of two cell samples. 166

A.12 Contact significance analysis of selected chromosomes. In order to check if the number of contacts between two specific chromosome regions is significantly large, we calculated the significance score (i.e. the probability of receiving this number of contacts or more) in each cell of an intra-chromosome contact matrix at 1Mb resolution, assuming the background distribution of contact numbers follows the Poisson distribution. The parameter (lamda: mean contact number) of the background distribution was set to the average of number of contacts in the matrix excluding contacts within the same region (i.e. diagonal line in a matrix). Sub-figures A, B, C and D illustrate the contact significant scores of the intra-chromosomal contract matrices of chromosome 7 of the MHH-CALL-4 cell line, RL cell line, normal B-cell and the Primary B-ALL cell, respectively. Sub-figures E, F, G and H depict the significance scores of intra-chromosomal contact matrices of chromosome 14 of the MHH-CALL-4 cell line, RL-cell line, normal B-cell line and the primary B-ALL cell, respectively. Darker red indicates higher significant score. 167

A.13 The corrected inter-chromosomal contact map between translocated chromosomes 11 and 14 for the primary ALL B-cell. The method of calculating it can be found in Figure A.15. 168

A.14 The method of calculating the corrected inter-chromosomal contact matrix of translocated chromosomes. (A) Division of the corrected inter-chromosomal contact matrix between chromosomes 11 and 14 into three regions to be reconstructed separately. Region A contains the contacts between non-translocated segments in chromosome 11 (i.e. 11(c) and 11(d) in (B)) and non-translocated segments in chromosome 14 (i.e. 14(c) and 14(d) in (B)). Region B contains the contacts between translocated segments in chromosome 11 (i.e. 11(b) in (B)) and translocated segments in chromosome 14 (i.e. 14(a) in (B)). Region C contains the contacts between non-translocated segments in chromosome 11 (i.e. 11(c) and 11(d) in (B)) and translocated segments in chromosome 14 (i.e. 14(a) in (B)). Region D contains the contacts between non-translocated segments in chromosome 14 and translocation segment in chromosome 11. For the contacts in regions A and B, we divided the original contact numbers by 2 in order to estimate the inter-chromosome contacts. For region C, we normalized the value of each cell $C_{ij} = \max(0, C_{ij} - \text{average num of row } i \text{ in region A})$. For region D, we normalized the value of each cell $D_{ij} = \max(0, D_{ij} - \text{average num of column } j \text{ in region A})$ 169

A.15 The Pearson's correlation matrix for inter-chromosomal contact numbers between the normal B cell, primary ALL B-cell, MHH-CALL-4 cell line, and RL cell line. For each cell, the number of inter-chromosomal contacts between chromosomes were calculated and put into a vector. Thus, each cell has one vector to represent all its inter-chromosomal contact numbers. The matrix below shows the Pearson's correlation between each pairs of vectors of two cell samples. 170

A.16	The interaction network between transcription factor binding sites (TBSs) in the entire genome of the MHH-CALL-4 cell line.	171
A.17	The distribution of node degree of the TBS-TBS interaction network shown in Figure A.16.	172
A.18	The histogram of lengths of the shortest paths between any two nodes in the TBS-TBS interaction network shown in Figure A.16.	172
A.19	The distribution of topological coefficients the TBS-TBS interaction network shown in Figure A.16.	173
A.20	The distribution of node stresses of the TBS-TBS interaction network shown in Figure A.16.	174
A.21	The spatial interaction networks between genes and transcription factor binding sites (TFB) in chromosome 14 for the CALL-4 cell line. A node in the network denotes a gene or a TFB. Two nodes are connected by an edge if they are spatially contacted.	175
A.22	The node degree distribution of the network shown in Figure A.21. It is shown that the frequency (number of nodes) is largely linear to the degree of the nodes on the log-log scale. This suggests that the network is likely a scale-free network.	176
A.23	The histogram of lengths of the shortest path between any two nodes in the network shown in Figure A.21.	177
A.24	The distribution of stress values of the network shown in Figure A.21.	177
A.25	The distribution of topological coefficients of the network shown in Figure A.21.	178

A.26	(A) The chromosomal region of the transcription factor binding site on chromosome 14 of the MHH-CALL-4 cell line that has the highest contacts with other genes is visualized by the UCSC genome browser. This transcription factor binding site contacted 1460 times with GeneID:145508 (starting from the position 61002767 and ending at 61445813), 1 time with GeneID:7253, 2 times with GeneID:6710, 2 times with GeneID:56659, and 1 time with GeneID:9369. (B) The chromosomal region of the gene (GeneID:145508) that encodes a centrosomal protein (128kDa). . . .	179
B.1	The node degree distributions of <i>H. sapiens</i> , <i>S. cerevisiae</i> , <i>C. elegans</i> , <i>D. melanogaster</i> , and 15 plant genomes	182
B.2	The shortest path length distributions of <i>H. sapiens</i> , <i>S. cerevisiae</i> , <i>C. elegans</i> , <i>D. melanogaster</i> , and 15 plant genomes	183
B.3	The average clustering coefficient distributions of <i>H. sapiens</i> , <i>S. cerevisiae</i> , <i>C. elegans</i> , <i>D. melanogaster</i> , and 15 plant genomes	184

ABSTRACT

Deoxyribonucleic acid, or DNA, encodes genetic instructions for the functionalities of organisms. For human beings, 23 pairs of chromosomes, containing DNA strands, form a globule structure in the nucleus. This chromosomal conformation influences the subsequent biological processes including transcription and translation by positioning sequentially remote genes spatially close. Chapter 2 reveals human chromosomal conformation and studies gene-gene interactions and “transcription factor binding site” interactions based on chromosomal spatial proximity.

During the transcription process, an mRNA chain is produced from decoding DNA, followed by translation when protein molecules are synthesized. Proteins are the biological units that conduct biological functions. The three-dimensional structure of a protein molecule determines its particular functions. Predicting protein three-dimensional structures and functions from amino acid sequence has drawn substantial attention because it is an essential step for thoroughly understanding biological processes. Chapters 3 and 4 discuss research in predicting protein tertiary structures. Algorithms that can predict residue-specific qualities of predicted structures were constructed and benchmarked. A knowledge database of soybean transcription factors is presented in Chapter 5, which contains predicted protein tertiary structures. Chapter 6 shows a computer system predicting protein functions using profile-sequence alignment, profile-profile alignment, and protein domain co-occurrence network.

A biological process is usually performed by multiple proteins. Biological network provides a global perspective of studying lives, which usually considers the entire set of the same type of biological molecules of the target organism. Chapter 7 introduces a novel biological network, protein Domain Co-occurrence Network (DCN), and demonstrates that DCN has great potentials in inferring species phylogenies and predicting protein functions.

Chapter 1

Introduction

This dissertation includes my research in genome conformation, protein structural and functional prediction, and a biological network, protein Domain Co-occurrence Network (DCN).

The genome of a species is the complete set of double-helix DNA strands, which is the blueprint of the whole set of biological features including genotype and phenotype of an organism. The uncoiled human genome is around 1.02 meters long [3], which is wrapped inside a tiny nucleus of a eukaryotic cell. The spatial conformation of a genome, which is the way a genome folds inside a nucleus, facilitates the genetic activation and transcription in a specific cell type, cell cycle, and biological condition because the spatial proximity between genes can be achieved by folding a linear sequence of DNA strands in a three-dimensional space [4, 5, 6, 7, 8, 9]. Moreover, genome conformation provides a novel perspective of studying gene-gene interactions and “transcription factor binding site” interactions, which is a novel approach in the scientific community. Chapter 2 and Appendix A of this dissertation include the above-mentioned research, whose content is from an unpublished manuscript:

Wang, Z., Cao, R., Taylor, K., Briley, A., Caldwell, C., and Cheng, J. The properties of human genome conformation₁ and spatial gene interaction and regulation

networks. submitted.

Protein is synthesized from protein coding regions, or genes, along the DNA strands. It is the entity that physically carries out biological functions and makes the whole life system work. For example, a transcription factor is a type of protein that binds to DNA strands and controls the synthesis of mRNA; and a ribosome, the biological complex that synthesizes protein molecules, also consists of proteins. The specific functions of a protein molecule are determined by the particular tertiary structure of the protein. The backbone conformation of the structure is determined by the amino acid composition of the protein, which provides theoretical possibility for protein structural prediction. With the development of next-generation sequencing technology, more and more species have been annotated, which generates a large amount of protein sequences (amino acid sequences). However, detecting protein structures by experimental methods, such as X-ray crystallography and Nuclear Magnetic Resonance (NMR) spectroscopy, is expensive and cannot catch up with the pace of increasing protein sequences. Therefore, utilizing computer algorithms to predict protein structures initially became a reasonable alternative, and then an indispensable approach.

The CASP competition (Critical Assessment of Techniques for Protein Structure Prediction) witnesses the efforts dedicated to, and the development of, protein structural prediction. The first CASP, i.e. CASP1, was held in 1994 when there were 33 targets available and 35 prediction groups participated, whereas after 18 years, the latest CASP10 in 2012 has 260 groups participating world-wide and 102 targets in 6 prediction categories available including tertiary structure, model quality assessment, disordered region prediction, binding site prediction, residue-residue contacts prediction, and model refinement. Current computer systems can generate high quality models (predicted structures) with GDT-TS [10] score > 0.95 (GDT-TS score 1 would be the native structure) when proper templates can be found, which is a great

achievement. However, *ab initio* prediction when no templates are available is still a challenge to the scientific community.

During my PhD study, I participated in building a computer system predicting protein tertiary structures [11] and developed automatic evaluation software (Appendix C). Chapters 3 and 4 of this dissertation include my research related to protein model quality assessment, one important component of protein structural prediction, which aims to predict the qualities of protein models before knowing native structure. Nowadays, computer systems can generate hundreds or even thousands of models for one query within days, which makes it an essential task to select models with good qualities. The algorithms shown in this dissertation can successfully predict the absolute global quality of a single protein model and also the residue-specific qualities of an individual protein model. The content of Chapters 3 and 4 are mostly from the manuscripts published as:

Wang, Z., Eickholt, J., and Cheng, J. (2011) APOLLO: A Quality Assessment Service for Single and Multiple Protein Models. *Bioinformatics*, 27(12), 1715-1716. [2012 Impact Factor: 5.468]

Wang, Z. and Cheng, J. (2012) A Hybrid and Iterative Self-Refining Approach for Protein Model Quality Assessment. *Protein Science*, 21:142-151. [2010 Impact Factor: 2.741]

Chapter 5 and Appendix B describe a knowledge database for soybean transcription factors. This database contains protein sequences, predicted tertiary structures, putative DNA binding sites, protein domains, homologous templates in the Protein Data Bank (PDB), protein family classifications, multiple sequence alignments, consensus protein sequence motifs, the web logo of each family, and web links to 12 other popular databases or datasets. This database is an application of protein tertiary structure prediction as the soybean was newly sequenced at the time when this database was built. The content of Chapter 5 has been published as:

Wang, Z., Libault, M., Joshi, T., Valliyodan, B., Nguyen, H., Xu, D., Stacey, G., and Cheng, J. (2010) SoyDB: A Knowledge Database of Soybean Transcription Factors. *BMC: Plant Biology*, 10:14. [2012 Impact Factor: 4.09]

Predicting the biological functions of a protein, to the specificity of GO terms [12], directly from amino acid sequence is very useful but also difficult. One of the reasons is that a large number of GO terms form a hierarchical tree (directed acyclic graph), which makes it difficult to hit the specific GO terms farthest to the root. Chapter 6 describes a protein functional prediction system using profile-sequence alignment, profile-profile alignment, and protein domain co-occurrence network. The predictions of this system have better specificity and accuracy compared with baseline methods. Chapter 6 is based on a manuscript that has been accepted as:

Wang, Z., Cao, R., Cheng, J. (2012) Three-Level Protein Function Prediction by Integrating Homology Search and Domain Co-Occurrence Networks. *BMC Bioinformatics*, accepted. [2012 Impact Factor: 3.03]

A biological process is conducted by multiple interacting proteins and other biological molecules. Therefore, besides studying structure and functions of an individual protein, it is requisite to investigate biological networks, for example, protein-protein interaction networks. A biological network usually contains the same type of biological molecules existent in a species, which provides a global perspective of studying an organism. Chapter 7 summarizes my research related to a novel biological network, protein domain co-occurrence network. For the first time in the scientific community, we demonstrated its capabilities of predicting protein functions and inferring species phylogeny. The content of Chapter 7 has been published as:

Wang, Z., Zhang, X. C., Le, M. H., Xu, D., Stacey, G., and Cheng, J. (2011) A Protein Domain Co-Occurrence Network Approach for Predicting Protein Function and Inferring Species Phylogeny. *PLoS ONE*, 6(3): e17906. [2012 Impact Factor: 4.411]

Chapter 2

The Properties of Human Genome Conformation and Spatial Gene Interaction and Regulation Networks

2.1 Abstract

The spatial conformation of a genome plays an important role in the long-range regulation of genome-wide gene expression and methylation, but has not been extensively studied due to lack of genome conformation data. The recently developed chromosome conformation capturing techniques such as the Hi-C method empowered by next generation sequencing can generate unbiased, large-scale, high-resolution chromosomal interaction (contact) data, providing an unprecedented opportunity to investigate the spatial structure of a genome and its applications in gene regulation, genomics, epigenetics, and cell biology. In this work, we conducted a comprehensive, large-scale computational analysis of this new stream of genome conformation data generated for three different human leukemia cells or cell lines by the Hi-C technique. We

developed and applied a set of bioinformatics methods to reliably generate spatial chromosomal contacts from high-throughput sequencing data and to effectively use them to study the properties of the genome structures in one-dimension (1D) and two-dimension (2D). Our analysis demonstrates that Hi-C data can be effectively applied to study tissue-specific genome conformation, chromosome-chromosome interaction, chromosomal translocations, and spatial gene-gene interaction and regulation in a three-dimensional genome of primary tumor cells. Particularly, for the first time, we constructed genome-scale spatial gene-gene interaction network, transcription factor binding site (TFBS) - TFBS interaction network, and TFBS-gene interaction network from chromosomal contact information. Remarkably, all these networks possess the properties of scale-free modular networks.

2.2 Introduction

The genome of a species is the complete set of linear DNA sequences of double-stranded nucleotides. It consists of protein coding regions (genes), gene regulatory elements (e.g. promoter, enhancer), and non-coding functional or nonfunctional regions, which encode the whole set of features (e.g. genotype and phenotype) of the species. In order to physically store a long genomic DNA sequence in a small nucleus of a eukaryotic cell, every 145-147 base pairs of DNA are wrapped around a protein complex (histone octamer) to form a nucleosome achieving higher compaction. Tens of nucleosomes are joined by linker DNA sequences to form a larger dense structural unit (chromatin fiber) of several kilobase (Kb) pairs [13, 14]. Chromatin fibers are further folded into higher-order modules of megabase pairs (Mb), such as domains / globules / gene loci / chromatin clusters, which eventually aggregate into a large standalone physical chunk, i.e. chromosome [15, 16]. A eukaryotic genome is distributed on a number of discrete chromosomes. For instance, the human genome of \sim

3 billion base pairs (Gb) resides on 23 pairs of chromosomes. Instead of intermingling together, each chromosome occupies a particular space in a nucleus called a chromosome territory [4, 5], while they also interact with each other at their boundaries to constitute the spatial, higher order (i.e. three dimensional) conformation of the entire genome during interphase.

Although the linear sequence of a genome encodes all the genetic information for a species, it is the spatial, high-order conformation of the genome that largely determines and facilitates which sets of genes are activated and transcribed in a specific cell type, cell cycle, and biological condition by bringing distant genes and regulatory elements located on the linear genome into spatial proximity in order to achieve coordinated or efficient gene transcription to carry out appropriate cellular functions [4, 5, 6, 7, 8, 9]. Therefore, studying genome spatial conformation in either two dimensional (2D) or three dimensional (3D) space is crucial for elucidating the complex multi-dimensional logic of gene transcription, involving long-range gene-gene interactions, spatial gene and regulatory element interactions, epigenetic DNA methylation, and chromatin modification and remodeling.

In contrast to the extensive research on genome-wide gene expression and DNA methylation in a linear genome facilitated by whole genome sequencing, the examination of the spatial conformation of a genome is still at an early stage as only a few experimental methods have been available to interrogate structure at the genome scale until recently [17, 18]. An early, but still widely used method, fluorescence in-situ hybridization (FISH) [19], can selectively measure the physical distances between a number of genetic markers (e.g. a marked position on a chromosome) at a resolution of ~ 50 -200 nanometers. More recently, with the assistance of next generation sequencing, 3C [20], 4C [21], and 5C [22] methods have determined most chromosomal regions in spatial proximity (or contact) with a pre-marked genomic region of several hundred Kb to a few Mb. This provides abundant information regarding

chromosomal interactions and the spatial conformation of the marked region. Most recently, an exciting, revolutionary Hi-C technique [1] was developed to determine chromosomal contacts in an unbiased manner at genome scale. The Hi-C technique uses proximity based ligation to join two DNA fragments within a chromosome or between two chromosomes that are spatially close. The combined fragment is labeled with biotin at the ligation junction to enable recovery of the fragment and then sequenced using next-generation sequencing technologies. The sequence of the fragment can then be mapped to two unique locations in the genome, which correspond to the positions of the two DNA fragments in contact. The technique can identify chromosomal sub-region contacts at an arbitrary coverage as long as sufficient sequencing and ligation experiments are conducted. Theoretically, with access to high-throughput sequencing equipment, many biology labs can carry out Hi-C experiments to generate chromosomal contact data [23, 24]. Chromosomal contacts can be used to study gene / regulatory element interactions and even as distance restraints to construct the 3D structure through computational modeling. Computational methods need to be developed to generate, analyze, and model these new sources of data in a large scale manner in order to study the structural and functional properties of a genome in a spatial context.

In this work, we developed a bioinformatics software pipeline to process hundreds of millions of Hi-C paired-end sequence reads generated for three different human cells (RL follicular lymphoma cell line, primary tumor B-cells from an acute lymphoblastic leukemia patient, and MHH-CALL-4 B-cell acute lymphoblastic leukemia cell line) by the Hi-C experimental techniques. Our method successfully mapped a majority of sequence reads to the human reference genome and generated a large data set of high-quality and high-resolution chromosome contacts. The contact data was effectively used to study and discover the properties of human genome conformation, including conservation and variation of genomes of different cell types, intra- and inter-

chromosomal interactions, abnormal chromosomal translocation, spatial chromosomal contact clusters, spatial gene-gene interactions, and spatial gene-regulatory-element interaction. The analysis yields a number of new insights into the spatial structure of human genome conformation and abnormal inter-/intra-chromosomal interactions associated with chromosomal translocation. To the best of our knowledge, we have constructed chromosome-/genome-wide gene-gene interaction networks for the first time, as well as transcription factor binding site (TFBS) - TFBS networks, and gene-TFBS networks from chromosomal contact data. Interestingly, each of the genetic networks has the same properties of other non-spatial scale-free biological networks such as protein interaction networks.

2.3 Methods

2.3.1 Hi-C library preparation and sequencing of the primary human acute lymphoblastic leukemia B-cell (B-ALL), MHH-CALL-4 B-ALL cell line (CALL4), and lymphoma RL cell-line (RL)

The primary ALL patient sample was obtained from the Ellis Fischel Cancer Center (Columbia, MO) following diagnostic evaluation and in compliance with the local Institutional Review Board. Human cell lines RL and MHH-CALL4 were maintained at 37°C with 5% CO_2 . MHH-CALL4 is a human B cell precursor leukemia cell line established from the peripheral blood of a 10-year-old Caucasian boy with acute lymphoblastic leukemia at diagnosis in 1993 [25].

Library preparation was adapted from [1]. Cells from a patient with acute lymphoblastic leukemia, a lymphoma cell line (RL), and an acute lymphoblastic leukemia cell line (MHH-CALL4) were cross linked using formaldehyde. Once cross-linked, the cells were lysed and the chromatin was digested using a restriction enzyme (HindIII).

The ends of the fragmented DNA were repaired using biotinylated cCTP and subjected to blunt-end ligation. This process marked the DNA with biotin and an NheI recognition sequence was formed at the ligation junction. The DNA was then purified by degrading the remaining proteins with proteinase K and performing phenol-chloroform extractions.

To verify ligation efficiency, PCR was performed using quality control primers [1]. The PCR products were then digested with HindIII or NheI. As in [1], we detected an approximate 70% ligation efficiency in our Hi-C libraries. After the ligation efficiency was validated, biotin was removed from unligated DNA fragments using T4 DNA polymerase. The DNA was then sheared to a size of 300-500 base pairs and streptavidin beads were used to collect the remaining biotin labeled DNA fragments. The three Hi-C libraries were then subjected to paired-end high-throughput sequencing on the Illumina HiSeq 2000 by core facilities at the University of Missouri-Columbia.

2.3.2 Mapping Hi-C sequence reads to the reference genome

The sequencing generated millions of pair-end reads for each sample above. Each read end has a length of 100 or 120 nucleotides (Supplementary Table A.1 and A.2). The Hi-C sequence reads of the normal B-cell line were downloaded from [1] to test and ensure the correctness of our methods. The sequence reads of the four cell lines were mapped to the human genome according to the protocol illustrated in Supplementary Figure A.1. Briefly speaking, software Maq (<http://maq.sourceforge.net/>) was used to map each read-pair to the reference human genomes (NCBI build 36.3), where parameter “sum of mismatching base qualities (-e)” controlling the tolerance of mismatches was set to 150 in most experiments. Maq outputs the base pair positions in the reference genomes where each DNA read is mapped to. The mapped positions were analyzed by our method to generate chromosomal contacts. Although one read may be mapped to multiple locations due to inexact match of Maq, only the location

with the highest mapping quality was likely the correct one as the read of 70-120 bp long can most likely be mapped to one unique location on the genome. This strict strategy of handling multiple mapping locations reduced noise in the data and ensured the high quality of contacts. Moreover, we only kept the reads-pairs whose two ends are either mapped to two different chromosomes or ≥ 2 K bp away on the same chromosome.

Supplementary Figure A.1 illustrates the bioinformatics pipeline of our Hi-C experiment. The Hi-C wet lab experiment is similar to the method described in [1], in which chromosome DNA is cross-linked, ligated and then sheared. Each of the reads-pairs was mapped to the human genome by the tool `maq` (<http://maq.sourceforge.net/>) with the mistake threshold (-e) set to 150. Our computer programs analyzed the mapping output and handled the four different cases, in which the reads may cover different portions of the two chromosomes. Case one is that each of the two ends can only be mapped to one location; and the two mapped locations of the two ends are 2000b away. Case two is that one end can be mapped to one location (e.g. location A), but the other end to two locations (e.g. B and C). In this case, we checked whether one of the two locations (B or C) is within 2000b of A. If not, the case is considered invalid and is discarded. 2000bp was used as the threshold because the average length of the DNA insert is 2000bp long. Case three is the same as Case two except that the first end was mapped to two locations and the second to one location. Case four is that both two ends can be mapped to two locations (e.g. one end to A and B, and the other to C and D. A, B, C, and D are the starting positions of the mapping locations). In this case, we checked whether the distance between A and C is less than the read length and whether the distance between B and D is less than the read length. If yes, they were kept. In our first mapping strategy, only these four cases were considered and processed to generate contacts and all the other cases were discarded. For example, if one end of a pair of ends can be mapped

to ≥ 3 locations, they were discarded. This process was able to reduce noise (e.g. wrongly-aligned reads) and ensure the quality of contact parsing. We also developed another simplified strategy to control mapping quality and minimize mapping ambiguities. That is, only keep the reads that can be uniquely mapped to one location of the chromosome. If one of the two ends was mapped to ≥ 2 locations, these pair of ends were discarded. We found these two strategies generated similar contacts. The results presented in this paper were based on contacts generated by the second strategy that is more stringent. The intra- and inter-chromosomal contact matrices for chromosomes and chromosome pairs were visualized as heat maps by the statistical package R. As an example, the reads mapped to the HoxA gene cluster region in chromosome 7 were visualized in Supplementary Figure A.2.

2.3.3 Generating intra- and inter-chromosomal contact matrices

Intra-chromosomal and inter-chromosomal contact matrices were generated by counting numbers of mapped read contacts falling into each pair of equal-length regions or segments of chromosome(s). The length of the segment (i.e. the resolution of a contact matrix) can be adjusted according to research goals. We used 1Mb resolution for comparing interaction patterns between different cell lines, but used much smaller resolutions, 0.1Mb and 0.01Mb, in order to identify the boundaries of chromosomal translocations. Similarly as in [1], the number of short-range contacts is much larger than long-range ones. In order to make the long-range contacts easy to observe in the visualized heat maps of contact matrices, we set a maximum cap (e.g. 50) on the number of contacts between two regions in contact matrices for visualization.

We normalized the chromosomal contact matrices in order to (1) discover the statistically enriched and depleted regions within one contact matrix and (2) compare two contact matrices to recognize the differences in their interaction patterns. For

example, by comparing the normalized inter-chromosomal contact matrices of the healthy and Leukemia cell samples, we discovered the translocation between chromosome 11 and 14 in the Leukemia cell line. A variety of normalization methods were tested, including x/avg , $(x - min)/max$, and $(x - mean)/sd$, where x , max , min , $mean$, and sd are the number of contacts between two regions i and j , maximum, minimum, mean, and standard deviation of contact numbers in the matrix, respectively. The heat maps were generated using the “heatmap.2” package in R.

2.3.4 Statistical significance analysis of chromosomal contact matrix

To infer the statistical significance of the number of contacts, we assumed that contact values excluding local contacts on the diagonal line in a contact matrix follow a Poisson distribution. Each value in the statistical significance matrices is the probability of observing a higher contact number for a pair of locations by chance. The smaller probability value in the significance matrix indicates the higher statistical significance.

2.3.5 Detection of inter-chromosomal translocation boundaries and reconstruction of translocated chromosomes

We identified the regions with the unusually high number of inter-chromosomal contacts between chromosome 11 and 14 in the Primary ALL sample, and then zoomed in the targeted regions as shown in Figure 2.2. After zoomed in 100 times, the boundaries were detected when the contact value or the significant value has a sudden change.

2.3.6 Construction of Pearson’s correlation matrix from chromosomal contact matrix

The correlation matrices were generated based on the contact matrices mentioned above. The value $C(i, j)$ of a cell in a correlation matrix is the Pearson correlation between the values in the i th and j th rows in the contact matrix containing absolute contact numbers between region i and j . If the i th and j th regions are in contact, they likely share similar contact partners, leading to a higher correlation. The Pearson’s correlation matrices can effectively reduce the noise in contact data to amplify the plaid contact patterns corresponding to open and closed chromatin conformations. The contact correlation matrices were also used to discern the genome conformation difference between different cell lines. The difference matrix between two correlation matrices $C1$ and $C2$ equals to $|C1 - C2|$, which can be visualized as heat maps to show differences in chromosomal conformations.

2.3.7 Calculation of observed/expected numbers of contacts between all pairs of chromosomes

We generated the observed/expected number of inter-chromosomal contacts between each pair of chromosomes for all four cell lines. The contact numbers between 23 pairs of chromosomes provide a global view that can be used to distinguish conservation and variation in genome conformation between different cell samples as shown in Figure 2.2B-E. As in [1], the expected number of contacts between chromosome i and j was calculated by:

$$E_{i,j} = R_i \times R_j \times N_{inter}$$

where R_i and R_j are the fractions of inter-chromosomal reads associated with i and j , respectively; and N_{inter} is the total number of inter-chromosomal reads for

a cell sample. The actual observed number of inter-chromosomal contacts between chromosomes i and j divided by the expected number indicates the enrichment or depletion of inter-chromosomal contacts between them.

2.3.8 Construction of gene-gene interaction networks from Hi-C chromosomal contact data

The gene definitions of the human genome (build 36.3) were downloaded from the NCBI website. We only kept the “GENE” entries excluding the other types including “PSEUDO”, “RNA”, “CDS”, and “UTR”. A gene is denoted by a node or vertex in the gene-gene interaction network that represents all the spatial gene-gene interactions in a chromosome or a genome. An edge between two nodes is added if there is a Hi-C contact between the two genes. Figure 2.4 shows the intra-chromosomal gene-gene interaction network and its properties (node degree distributions, shortest path length distribution, average clustering coefficient distribution, closeness centrality, stress distribution, and topological coefficients) for the chromosome 14 of the CALL-4 cell line. The degree of a node is the number of edges linked to it. The node degree distribution depicts the number of nodes having various degree values. The shortest path length distribution indicates the number of node pairs who have a shortest path k between them, $k = 1, 2, 3 \dots$. The clustering coefficient of a node n is calculated by:

$$C_n = \frac{2S_n}{K_n(K_n - 1)}$$

where S_n is the number of edges among immediate (one edge away) neighbors of node n ; and K_n is the number of immediate neighbors of the node n . Clustering coefficient is between 0 and 1 indicating the tendency of immediate neighboring nodes to be clustered. Average clustering coefficient is calculated by averaging the clustering coefficients of all nodes in the network.

The closeness centrality $C_{c(n)}$ of a node n is calculated as:

$$C_{c(n)} = \frac{1}{\text{avg}(L(n, m))}$$

where $L(n, m)$ is the length of the shortest path between the nodes n and any other node m . Closeness centrality of a node is between 0 and 1 measuring how fast information can be broadcasted from one node to other reachable nodes. Isolated nodes having a closeness centrality 0 were not considered in Figure 2.4. The stress of a node n is the number of shortest paths traversing through node n ; and the stress distribution indicates the number of nodes with specific stress values. The stress values are grouped into bins whose sizes are factor of 10, i.e., $\{0\}$, $[1, 10]$, $[10, 100]$, ... The topological coefficient T_n of a node n is calculated as:

$$C_n = \frac{\text{avg}(S(n, m))}{g(n)}$$

where m is a node that shares at least one neighbor with node n or there is a direct link between node m and node n ; function $S(n, m)$ returns the number of neighbors shared between node m and node n , with 1 added if there is a direct link between node m and node n ; and $g(n)$ is the number of immediate neighbors node n . The topological coefficient of a node measures the extent to which a node shares neighbors with other nodes. The topological coefficient of a node having zero or one neighbor is assigned to 0.

2.3.9 Construction of interaction network of transcription factor binding sites (TFBS)

The definitions and coordinates of transcription factor binding sites were downloaded from Yale TFBS [26], which were identified by ChiP-seq experiments. In the TFBS networks, a node denotes a TFBS. Two TFBS nodes are connected by an edge if

there is a Hi-C contact between them. The weight of the edge is the number of Hi-C contacts between the two nodes. We generated a genome-wide TFBS-TFBS interaction network including both intra- and inter-chromosomal contacts of all 23 pairs of chromosomes.

2.3.10 Construction of interaction networks of transcription factor binding sites (TFBS) and genes

Based on the definitions of TFBS downloaded from Yale TFBS [26] and the NCBI gene definitions, we constructed the TFBS-gene interaction networks from the Hi-C contact data using the same approach described above.

2.4 Results

2.4.1 Hi-C Read Mapping

We created Hi-C libraries for a case of primary human B-acute lymphoblastic leukemia (B-ALL), the MHH-CALL-4 B-ALL cell line (CALL4), and the follicular lymphoma cell-line (RL). These libraries were sequenced using an Illumina HiSeq 2000. High-quality paired-end reads of 39M, 79M, and 33M were obtained for these cells, respectively. The quality distributions of 100-120 bp reads are shown in Supplementary Figures A.3-A.5. The read number distributions and gene number distributions along selected chromosomes are reported in Supplementary Figure A.6. The paired-end DNA-reads of the normal human B-cell line (GM06990) (7M pair of reads) were downloaded from [1] as a reference benchmark to test our in-house read mapping method. Paired-end reads for both the reference data and our own Hi-C data were mapped to the human genome and the chromosomal contact information was generated (see details in "Methods" section). On the reference data, 98.3% of the contacts

generated by our method were identical with the contacts produced in [1]; and 83.2% of contacts in [1] were also reproduced by our method. This high consistency supported the validity of our system, where the minor difference was likely caused by the higher stringency adopted in our method of handling single-end reads mapped to multiple locations in the genome. In this case, our system only used the mapped location with the highest mapping quality score. In terms of sequencing depth, on average, each gene region in our Primary ALL, MHH-CALL-4, RL cell lines has about 1.8, 2.8, and 1.5 mapped reads (Supplementary Table A.1), which is much higher than 0.17 in [1] likely because of our higher level of sequencing reads.

2.4.2 Intra-chromosomal contacts of different cells or cell lines

We constructed the intra-chromosomal contact matrices for the normal B-cell, primary B-ALL cells, MHH-CALL-4 cell line, and lymphoma RL cell line and visualized them as heat maps (Figure 2.1A-D for chromosome 14, supplementary Fig A.7-A.10 for all chromosomes). In a heat map matrix (M), a chromosome is divided into a number of 1 Mb regions, where the value of a cell $M(i, j)$ is the number of contacts between regions i , and j . Although 1Mb resolution contact matrices were generated here for comparison with the reference data [1], higher resolution matrices or maps (e.g. 10Kb resolution) can also be generated for our cell samples or at least some chromosomal regions since many more reads were collected. In the heat maps of visualizing contact matrices, the intensity of color at position i, j is set proportional to the contact frequency between two regions i, j . The heat maps we constructed for the reference normal B-cell (Figure 2.1A) are almost identical to those in [1], supporting the validity of our method. As in [1], in order to sharpen contact patterns, we generated Pearson’s correlation maps (C) from initial contact maps (Figure 2.1E-H), in which the value of a cell $C(i, j)$ was equal to the Pearson’s correlation between the

ith and jth rows in the original contact matrix (M). The assumption is that if two regions are spatially close, they should share similar contact neighbors, thus have a higher correlation between their contact profiles. Indeed, the correlation maps clearly reveal the plaid contact patterns likely corresponding to open or closed euchromatin and heterochromatin compartments than in the initial contact matrices by reducing noise in the data. The results suggest that the correlation maps of the normal B-cell, primary B-ALL, and CALL4 (B-ALL cell line) (Figure 2.1E-G) are more similar to each other than to the RL cell line (follicular lymphoma cell line) (Figure 2.1H), even though there are also some differences in the maps of normal and malignant B-cells (Figures 2.1E-F). To reveal the differences in contact profiles between healthy and malignant B-cells and between different cell types, we constructed difference matrices showing the absolute difference between their correlation maps. Figure 2.1I-K illustrates the differences in the correlation maps of chromosome 14 between normal B-cells and the primary B-ALL cells, MHH-CALL-4 cell line, and lymphoma RL cell line, respectively. Figure 2.1L shows the difference between two malignant cell lines MHH-CALL-4 and follicular lymphoma RL. Higher color intensity indicates larger absolute difference and white indicates the same. It seems that the difference between the normal and malignant B-cells is more obvious than that between the other pairs. The Pearson's correlations between the vectors of intra-chromosomal contact numbers of 23 pairs of chromosomes of these four samples (Supplementary Fig. A.11) also show similar relationships. Supplementary Figure A.12 shows the significance analysis of intra-chromosomal contact matrices.

2.4.3 Inter-chromosomal contacts and chromosomal translocations

Inter-chromosomal contacts exist between two spatially close regions from two different chromosomes. We generated inter-chromosomal contact matrices for all pairs

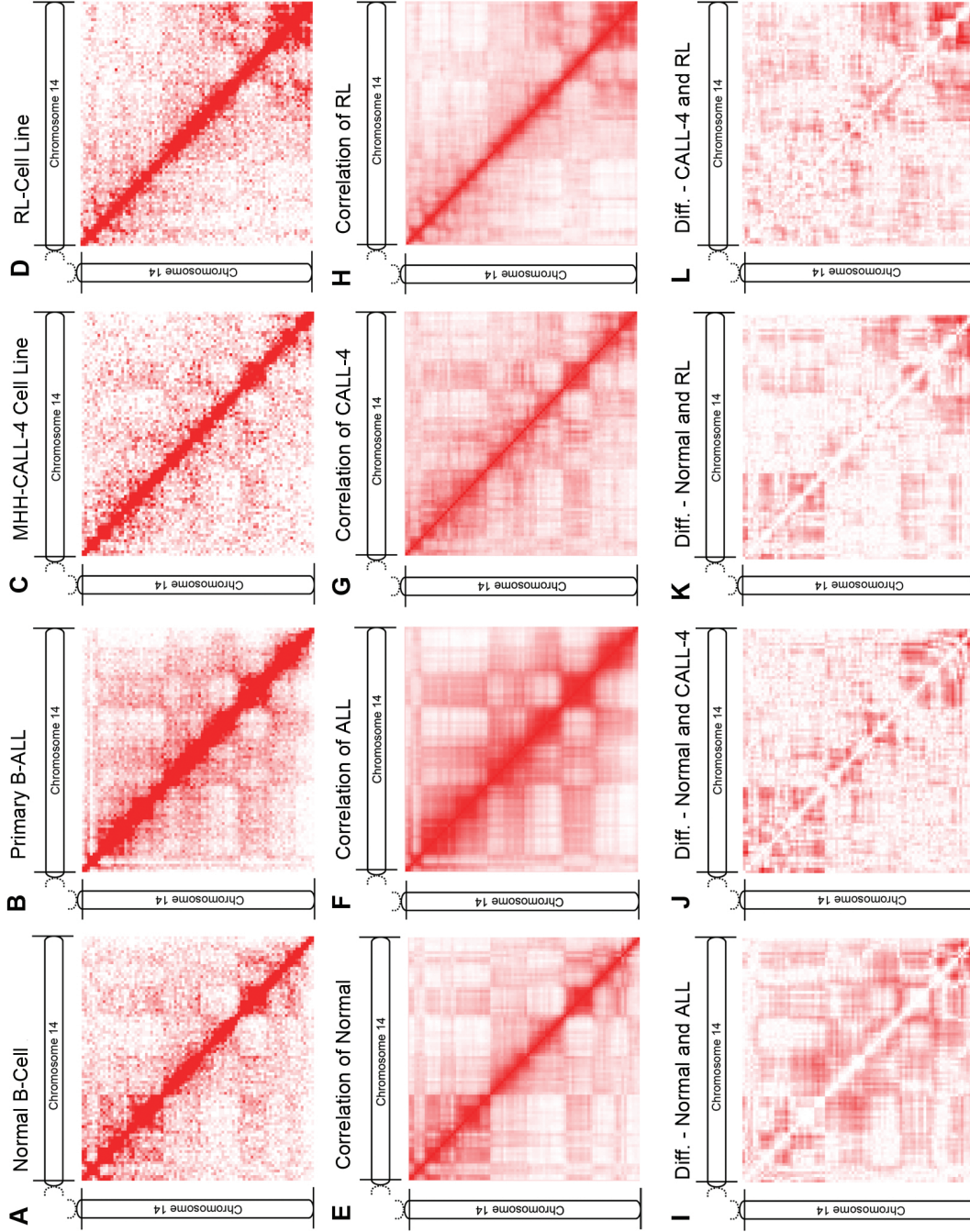


Figure 2.1: The original contact matrices (A-D), Pearson's correlation matrices(E-H), and difference matrices (I-L) of chromosome 14 for the normal human B-cell, human acute lymphoblastic leukemia B-cell, human MHH-CALL-4 B-ALL cell line, and human lymphoma RL cell-line.

of chromosomes for all cellular samples (data not shown). In comparison with plaid patterns in intra-chromosomal contact maps, inter-chromosomal contacts are much less frequent and more uniformly distributed in all but one case. The unusually dense inter-chromosomal contacts were found between chromosomes 11 and 14 for the primary B-ALL cells (Figure 2.2A). The bottom part of the contact map is significantly more intense compared with the other regions (Figure 2.2A), suggesting the telomeric end of chromosome 11 is very spatially close to chromosome 14. The most plausible explanation for this is that a reciprocal translocation of chromosome 11 and chromosome 14 was present. By zooming in on the map from 1Mb resolution to 10Kb, the boundaries of the translocation between chromosomes 11 and 14 were identified between genomic regions at 121.48Mb - 121.49Mb along chromosome 11 and between 105.42Mb - 105.47Mb along chromosome 14 (Figure 2.2A) by detecting the locations with a sudden, drastic increase of inter-chromosomal contacts (see details in Methods section). The translocation was later confirmed by an independent oligonucleotide tiling array experiment specialized at detecting chromosome rearrangements and was shown to be a cancer causing factor (manuscript in preparation). Based on the translocation, our method was able to reconstruct the translocated chromosomes 11 and 14 (t;11:14), and to estimate their inter-chromosomal contacts (Supplementary Fig. A.13-A.14).

In order to study the genome-wide inter-chromosomal contact profiles between healthy and malignant cells and between different cell types, we calculated the ratio between the observed and the expected number of contacts for all pairs of human chromosomes for the normal B-cell, primary B-ALL cells, the MHH-CALL-4 cell line, and the follicular lymphoma RL cell-line (Figure 2.2B-E). The higher ratios indicate more enrichment of inter-chromosomal contacts between two chromosomes. As in [1], small gene-rich chromosomes have more interactions than large gene-sparse chromosomes in the normal B-cell line (bottom right in Figure 2.2B). While this is still largely true

for primary malignant B-ALL cells and the MHH-CALL-4 cell line, unusual dense inter-chromosomal contacts were also found occurring in large or small chromosomes in these cells or cell lines. For example, in comparison with the normal B-cells, more contacts were found between chromosome 1 and chromosome 19, chromosome 11 and 14 in primary B-ALL tumor cells; chromosome 5 and chromosome 6, chromosome 3 and chromosome 11, chromosome 10 and 17, chromosome 9 and chromosome 19, chromosome 16 and chromosome 21 in the MHH-CALL-4 cell line. And a substantial difference in inter-chromosomal contact density had been observed between the normal or malignant B-cells and lymphoma RL-cell lines. For instance, there are unexpectedly more contacts between chromosome 2 and chromosome 8, chromosome 17 and chromosome 20, chromosome 13 and chromosome 18 in the RL-cell line. In comparison with the normal B-cells, the small and gene-rich chromosomes appear to have fewer contacts in leukemia cells or cell lines, and the least in the lymphoma sample. The Pearson's correlations between total numbers of inter-chromosomal contacts of four cells or cell lines are reported (Supplementary Fig. A.15). The difference in inter-chromosomal interaction patterns may shed light on the biology of different cell types and potential causes of diseases.

In addition to the conservation and variation in inter-chromosomal interactions described above, one remarkable conserved interaction among all four cells or cell lines is that one region [18Mb, 19Mb] in chromosome 14 always has the maximum number of inter-chromosomal contacts with other chromosomes. The most enriched gene function of the genes in this region is ubiquitin-protein ligase activity (Gene Ontology term, GO:0004842), but the importance of this function awaits further investigation.

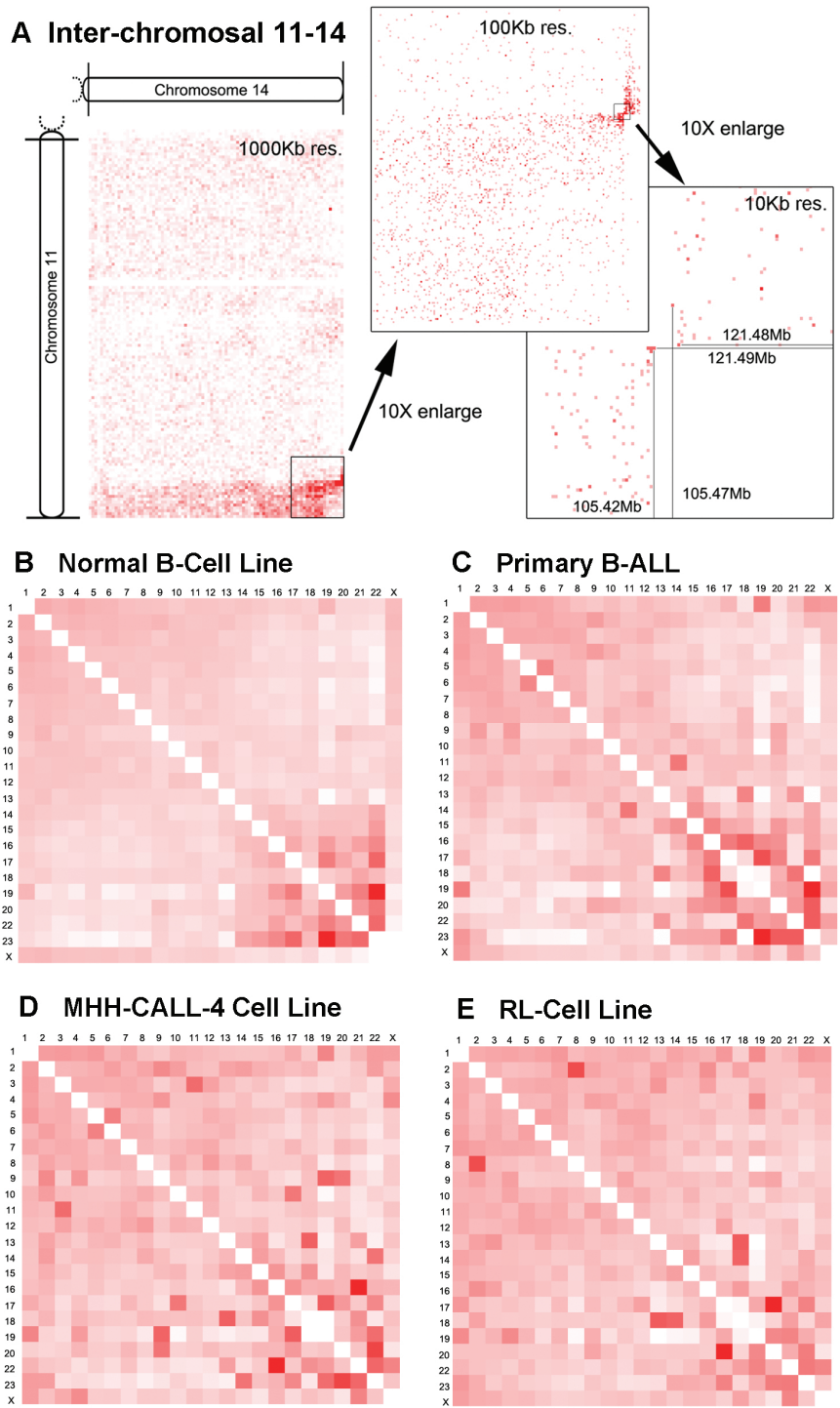


Figure 2.2: (A) Inter-chromosomal contact heat map between chromosome 11 and 14 for the primary ALL B-cell. Three resolution thresholds (1000Kb, 100Kb, and 10Kb) are used and the chromosomal translocation boundaries are illustrated. (B-E) The observed numbers of inter-chromosomal contacts divided by the expected numbers of inter-chromosomal contacts between all pairs of human chromosomes are visualized.

2.4.4 Spatial gene-gene interactions in the HOXA gene cluster in chromosome 7 of four cell lines

We chose the HoxA gene cluster 18 in chromosome 7 (location 27,095,000 - 27,215,000) to compare interaction patterns of different cell lines in greater detail. Supplementary Figure A.2 shows the reads mapped to this cluster visualized by the UCSC Genome Browser [27]. The numbers of contacts between 13 genes in the cluster in each cell line are shown in Figure 2.3 at 10Kb resolution. Gene-gene contacts in the normal B-cells are less than in malignant cells or cell lines, which might be due to the introduction of abnormal gene interactions within the cluster in malignant cells or cell lines, the higher read coverage or deeper sequencing for the gene cluster in our Hi-C samples leading to detecting more gene-gene interactions, or technical differences in read quality control. It seems that our deeper sequencing of the region enabled the detection of differences in gene-gene interactions among different cells or cell lines. For example, the RL-cell line had contacts between HOXA9 and HOXA3, between HOXA5-A6 and HOXA11, between HOXA7 and HOXA2-A3, which were not present in the MHH-CALL-4 cell line. The results demonstrate that our method is sufficiently sensitive to detect gene-gene interactions in cells of different types and states, which could be useful in elucidating mechanisms related to specific diseases.

2.4.5 Chromosome- or genome-wide spatial gene-gene interaction networks

The large amount of Hi-C chromosomal interaction data provided an unprecedented opportunity to study spatial gene-gene interactions. To the best of our knowledge, for the first time, we constructed chromosome- or genome-wide spatial gene-gene interaction networks for the human genome. In these networks, each node represents a gene; and an edge is used to connect two genes if there is at least one Hi-C read showing they are in spatial contact. The weight of the edge is the number of observed

contacts between two genes. Figure 2.4A illustrates the intra-chromosomal gene-gene interaction network of the genes on chromosome 14 for the MHH-CALL-4 cell line. The isolated genes without any spatial contact with other genes and the genes with types of “PSEUDO”, “RNA”, “CDS”, and “UTR” were not included. We analyzed a number of properties of the network, including node degree distribution (Figure 2.4B), shortest path length distribution (Figure 2.4C), average clustering coefficient distribution (Figure 2.4D), closeness centrality (Figure 2.4E), stress distribution (Figure 2.4F), and topological coefficients (Figure 2.4G). The linear relationship in the log-log plot of node-degree distribution (Figure 2.4B) shows that the gene-gene interaction network is very likely a scale-free network, like human social networks, world wide web, protein-protein interaction networks [28, 29, 30, 31, 32], and protein domain co-occurrence networks [33], where most nodes have few connections and some hub nodes have many connections. For example, a hub gene (GeneID:9369) has interactions with 111 other genes in chromosome 14. The gene-gene interaction networks of several chromosomes of other cells or cell lines that we investigated also possess the same scale-free property (data not shown). Figure 2.4C further depicts the small-world phenomenon associated with a scale-free network, i.e. most of genes are three or four steps away from each other through the shortest path between them. Figure 2.4D shows that the hub nodes with many interactions tend to have small clustering coefficients, whereas nodes with fewer connections tend to cluster with others to form densely connected sub-graphs. These clusters are connected through hub nodes. We also found that the nodes with smaller degree have the less average closeness centrality that measures how fast information can be spread to other reachable nodes, whereas hub nodes usually have a higher closeness centrality, which might suggest their importance in maintaining the connectivity of the networks (Figure 2.4E).

The stress value of a node is the number of shortest paths passing through it. A higher stress value may imply a more important role of the node in the network. A

small portion of nodes have a very high or very low stress value and most nodes have a middle stress value (Figure 2.4F), indicating that a small number of nodes may be extremely important to the network. The topological coefficient of a node measures the extent to which it shares neighbors with others. Figure 2.4G shows that hub nodes usually tend not to share neighbors. Instead, they serve as a center of a group of nodes (i.e. a module or clique), and the modules are connected through their hub nodes indirectly.

2.4.6 Interaction networks of transcription factor binding sites (TFBS)

In order to study interactions between TFBS that play an important role in spatial gene regulation, we constructed interaction networks of transcription factor binding sites (TFBS) from the Hi-C data. The whole-genome TFBS-TFBS interaction network for the MHH-CALL-4 cell line is shown in Supplementary Figure A.16. Like spatial gene-gene interaction networks, the TFBS-TFBS interaction networks also have the hallmark features of scale-free networks (Supplementary Fig.A.17-A.20). This de novo network may help study how spatially distal genes may be brought together to share the same transcription machinery.

2.4.7 Interaction networks of transcription factor binding sites and genes

To investigate interactions among TFBS and genes, we constructed networks showing TFBS-TFBS-gene interaction relationship for chromosome 14 of the MHH-CALL-4 cell line (Supplementary Fig. A.21). The statistical properties of TFBS-TFBS-gene interaction network suggest it is a scale-free network (Supplementary Figures A.22-A.25) that is very different from a random network. One TFBS on chromosome 14 that has a lot of contacts with a gene (GeneID: 145508) along its location was

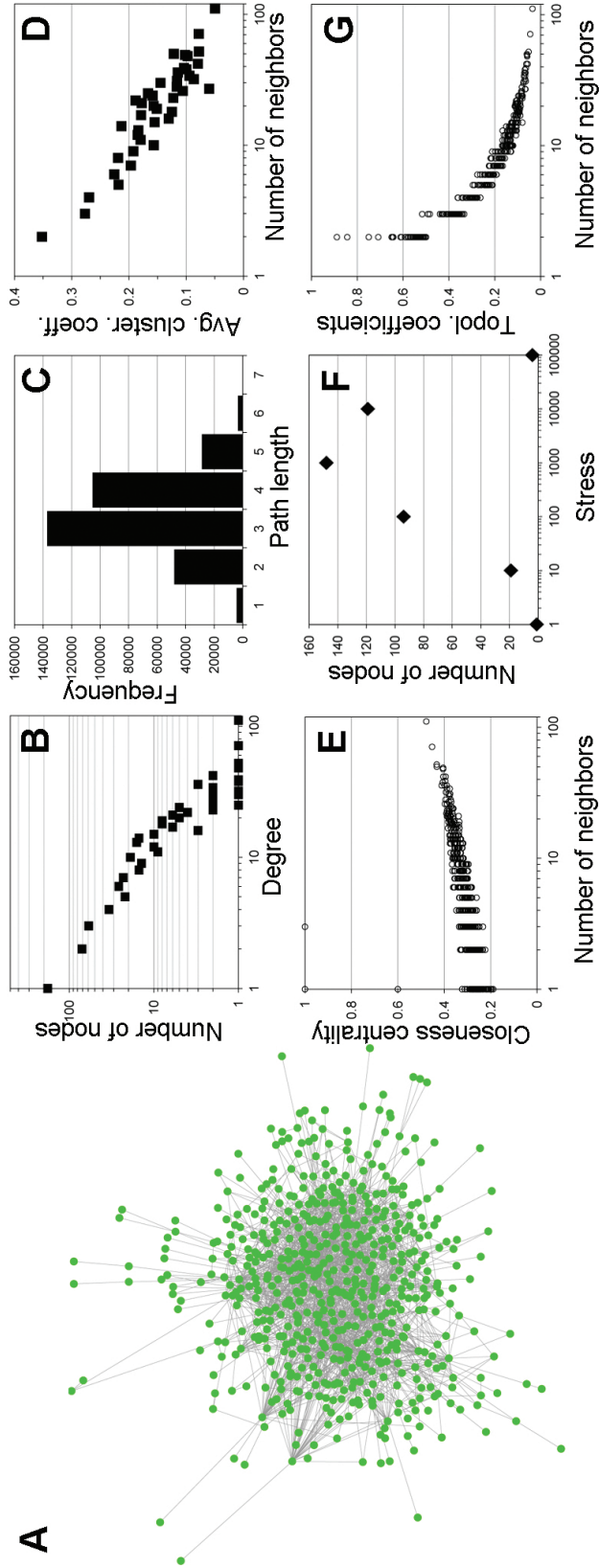


Figure 2.4: The spatial gene-gene interaction networks and the analysis of its properties. (A) The gene-gene interaction network of genes residing in chromosome 14 for the CALL-4 cell line. (B) The distribution of node degrees. (C) The histogram of shortest path lengths. (D) The plot of average clustering coefficient against of the degree (the number of neighbors) of a node. (E) The plot of closeness centralities against the degree of a node. (F) The stress distribution. (G) The plot of topological coefficients against the degree of a node. The network and its properties were visualized and analyzed by Cytoscape [2].

visualized by the UCSC Genome Browser (Supplementary Figure A.26).

2.5 Conclusions

We developed a bioinformatics pipeline to study the properties of human genome conformation and spatial gene interaction and regulation networks by analyzing Hi-C data. Our computational method can reliably generate intra- and inter-chromosomal contact matrices according to a standard Hi-C benchmark [1]. The chromosomal contact matrices built on the Hi-C data of three malignant cells or cell lines (B-ALL CALL-4, and RL) and a normal B-cell line demonstrates both the conservation in the genome conformation across different cells and the cell-type-specific or cell-state- (i.e. disease versus normal)-specific variation. For instance, smaller chromosomes had more contacts than expected compared with large chromosomes in the normal B-cell, but the pattern became less obvious in the malignant B-cells or cell lines, especially in the follicular lymphoma RL cell line. The conformational difference between different cell lines may be used to help understand 3D nuclear conformation and disease associations in different malignancies.

Furthermore, our method successfully used our high-resolution (e.g. 10Kb) inter-chromosomal contact matrix built on a case of primary B-ALL to identify a cancer-related chromosomal translocation between chromosomes 11 and 14 based on abnormal intensive interactions between the two ends of the two chromosomes. The boundaries of the translocation were accurately predicted and then used to construct the *in silico* translocated chromosomes. This is probably one of the first few examples that the Hi-C method can be used to accurately pinpoint and re-construct clinically important chromosomal translocations.

In addition to studying the properties of the genome and chromosome conformation as a whole, we also developed methods to use Hi-C data to investigate both

the gene-gene interactions in the HoxA gene cluster on chromosome 11 and the chromosome-wide spatial gene-gene interactions. To the best of our knowledge, this is the first demonstration of chromosome-/genome-wide spatial interaction networks between genes and transcription-factor-binding-sites (TFBS). Our experiments show that these gene interaction and regulation networks have the properties of modular scale-free networks similar to other biological networks. These discoveries shed new light on the study of 3D nuclear gene interactions and regulation.

2.6 Acknowledgement

The research was partially supported by a NIH grant (R01GM093123) to JC and a NIH grant (K99CA132784) to KT. ZW is partially supported by Shumaker bioinformatics fellowship.

2.7 Author Contributions

JC and BC conceived the project. JC and ZW designed the computational experiment. ZW and RZ implemented the computational methods and carried out the computational experiment. ZW, JC, RZ, BC, and KT analyzed the data. KT and CW designed the Hi-C experiment and AB implemented the Hi-C methods and performed the experiment. ZW, JC, RZ, and KT wrote the manuscript. All authors edited the manuscript and approved it.

Chapter 3

APOLLO: A Quality Assessment Service for Single and Multiple Protein Models

3.1 Abstract

We built a web server named APOLLO which can evaluate the absolute global and local qualities of a single protein model using machine learning methods or the global and local qualities of a pool of models using a pairwise comparison approach. Based on our evaluation of 107 CASP9 targets, the predicted quality scores generated from our machine learning and pair-wise methods have an average per-target correlation of 0.671 and 0.917, respectively, with the true model quality scores. Based on our test on 92 CASP9 single-domain targets, our predicted absolute local qualities have an average difference of 2.60 Å with the actual distances to native structure. APOLLO is freely accessible at <http://sysbio.rnet.missouri.edu/apollo/>.

3.2 Introduction

Protein model quality assessment plays an important role in protein structure prediction and application. Assessing the quality of protein models is essential for ranking models, refining models and using models [34, 35, 36, 37]. Model Quality Assessment Programs (MQAPs) predict model qualities from two perspectives: the global quality of an entire model and the residue-level local quality. The techniques often used by MQAPs include multiple-model (clustering) methods [38, 39, 40, 41, 42, 43], single-model methods [44, 45, 46, 47, 48], and hybrid methods [49].

Multiple-model methods assess the quality of a model by assessing its similarity with other models for the same protein target through pair-wise structure comparison. Single-model methods directly assess the quality of a model from its structural features using machine learning, statistical or physical methods. According to the Critical Assessment of Techniques for Protein Structure Prediction (CASP) experiments [36], multiple-model quality assessment methods are currently more accurate than single-model methods. However, they cannot be applied to a single model and do not work well if only a few models are available or the structures of the input models are very different. To address the problem, a hybrid quality assessment method [49] was recently developed to combine the two approaches and integrate their respective strengths. Here we build a web server to provide the community with these three complementary model quality assessment services which have been rigorously and successfully tested during CASP9, 2010.

3.3 Methods

3.3.1 Input and Output

Users only need to upload or paste a single model file in PDB format or a zipped file containing multiple models. If a single model is submitted, APOLLO only predicts the absolute global quality of the model. If multiple models are submitted, APOLLO outputs the absolute global qualities, average pair-wise GDT-TS scores, refined average pair-wise Q scores, and refined global and local qualities. All of the global qualities range between (0, 1] where 1 indicates a perfect model that is the same as the native structure and 0 indicates the worst case. The local quality scores are also visualized in a plot. Detailed descriptions and examples can be found in the help page for the APOLLO server: <http://sysbio.rnet.missouri.edu/apollo/help.html>.

3.3.2 Algorithms

The absolute global quality score is generated based on our ab initio QA predictor - ModelEvaluator [48]. Given a single model, ModelEvaluator (as MULTICOM-NOVEL server in CASP9) extracts secondary structure, solvent accessibility, beta-sheet topology, and a contact map from the model and then compares these items with those predicted from the primary sequence using the SCRATCH program [50]. These comparisons generate match scores which are then fed into a SVM model trained on CASP6 and CASP7 data to predict the absolute quality of the model in terms of GDT-TS scores. To predict absolute local quality scores, the secondary structure and solvent accessibility predicted from primary sequence are compared with the ones parsed from the model. From the model, we select all the pairs of residues, with 6 residues away in primary sequence, that have a Euclidean distance $\leq 8 \text{ \AA}$, and then gather their predicted probabilities of being contact from our predicted

contact map. The averaged contact probabilities, the match scores of comparing secondary structure and solvent accessibility, and the amino acids (within a 15-residue input window) are fed into a SVM model (RBF kernel) trained and optimized on 30 CASP8 single-domain proteins by 5-fold cross-validation. The accuracy we got from 5-fold cross-validation is: on the residues that have a real distance with the native ≤ 10 and 20 \AA , the average absolute difference between our predicted distance to the actual distance is 1.79 and 2.34 \AA , respectively.

The average pair-wise GDT-TS score is generated using our latest implementation (as MULTICOM-CLUSTER server in CASP9) of the widely-used pairwise comparison approach [51]. Taking a pool of models as input, it first filters out illegal characters and chain-break characters in their corresponding PDB files. It then uses TM-Score [52] to perform a full pair-wise comparison between these models. The average GDT-TS score between a model and all other models is used as the predicted GDT-TS score of the model. One caveat is that the GDT-TS score of a partial model is scaled down by the ratio of its length divided by the full target length.

The refined global and local quality scores are generated using a hybrid approach (as MULTICOM-REFINE server in CASP9) [49] that integrates ab initio model ranking methods with structural comparison-based methods. It first selects several top models (i.e. top five or top ten) as reference models. Each model in the ranking list is superposed with the reference models by TM-score. The average GDT-TS score of these superpositions is considered as the predicted quality score. The superimpositions with the reference models are also used to calculate Euclidean distances between the same residues in the superposed models. The average distance is used as the predicted local quality of the residue (Figure 3.1). Higher distances correspond to poorer local quality.

The refined average pair-wise Q scores are generated using a consensus approach (as MULTICOM-CONSTRUCT server in CASP9). APOLLO first uses the average

pair-wise similarity scores, calculated in terms of Q-score [53, 54], to generate an initial ranking of all the models. The Q-score between a pair of residues (i, j) in two models is computed as:

$$Q_{i,j} = \exp[-(r_{i,j}^a - r_{i,j}^b)^2]$$

where $r_{i,j}^a$ and $r_{i,j}^b$ are the distance between C_α atoms at residue position i and j in model a and b , respectively. The overall Q-score between model a and b is equal to the average of all $Q_{i,j}$ scores of all residue pairs in the entire model. The average Q-score between a model and all other models is used as the predicted quality score of the model. The initial quality scores are refined by the same refinement process used by our hybrid method in MULTICOM-REFINE.

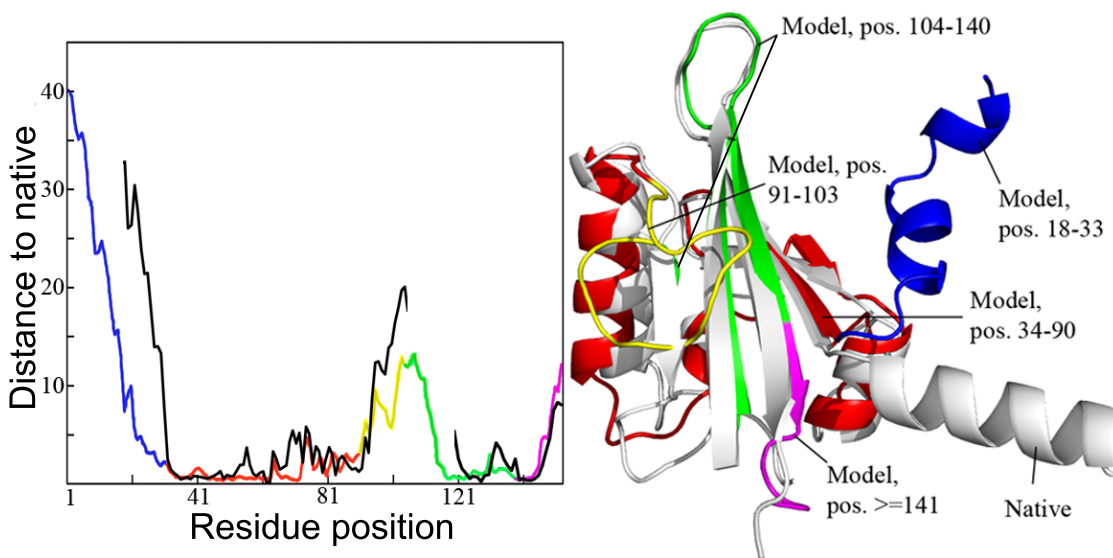


Figure 3.1: A local quality example for CASP9 target T0563. On the left is a plot of predicted local quality scores (colorful line) and actual distance (black line) against residue positions. On the right is the superposition between native structure (grey) and the model. The regions of the model with different local quality are visualized in different colors corresponding to the color of line segments in the plot on the left. Disordered regions are not plotted in the actual distance line.

3.4 Results

We assessed the performance of all the methods used by APOLLO on 107 valid CASP9 targets. We downloaded all the CASP9 tertiary structure (TS) models from the CASP9 web site and the experimental structures from the Protein Data Bank (PDB). These PDB files were preprocessed in order to select correct chains and residues that match the CASP9 target sequences. TM-Score was used to align each model with the corresponding native structure and generate its real quality score (GDT-TS). The CASP9 QA predictions which correspond to the exact QA methods employed by APOLLO were downloaded and evaluated against the actual quality scores by four criteria: average per-target correlation [36], the average sum of the GDT-TS scores of the top one ranked models, the overall correlation on all targets, and the average loss - the difference in score between the top ranked model and the best model [36, 49] (Table 3.1). The results show that the average correlation can be as high as 0.92 (resp. 0.67) and the average loss can be as low as 0.057 (resp. 0.095) for multiple-model (resp. single-model) quality assessment. The multiple-model global quality assessment and single-model global quality assessment methods were ranked among the most accurate QA methods of their respective kind according to the CASP9 official assessment (http://www.predictioncenter.org/casp9/doc/presentations/CASP9_QA.pdf). The average per-target correlation of our local quality predictions is 0.53, which is also among the top local quality predictors according to the CASP9 official assessment. Figure 3.1 visualizes one local quality assessment example. We conducted a blind test of the absolute local quality predictions on 92 CASP9 single-domain proteins. On the residues whose actual distances to the native are ≤ 10 and 20 \AA , the average absolute difference between our predicted distance and the actual distance is 2.60 and 3.18 \AA , respectively.

Table 3.1: Results of global quality assessment methods used by APOLLO server on 107 CASP9 targets.

Method	Avg. Corr.	Avg. Top 1	Over. Corr.	Avg. Loss
Abs. score	0.671	0.552	0.767	0.095
Avg. pair-wise GDT-TS	0.917	0.591	0.943	0.057
Refn. abs. score	0.870	0.567	0.928	0.081
Refn. pair-wise Q score	0.835	0.572	0.904	0.076

Chapter 4

An Iterative Self-Refining and Self-Evaluating Approach for Protein Model Quality Estimation

4.1 Abstract

Evaluating or predicting the quality of protein models (i.e. predicted protein tertiary structures) without knowing their native structures is important for selecting and appropriately using protein models. We describe an iterative approach that improves the performances of protein Model Quality Assurance Programs (MQAPs). Given the initial quality scores of a list of models assigned by a MQAP, the method iteratively refines the scores until the ranking of the models does not change. We applied the method to the model quality assessment data generated by 30 MQAPs during the Eighth Critical Assessment of Techniques for Protein Structure Prediction. To various degrees, our method increased the average correlation between predicted and real quality scores of 25 out of 30 MQAPs and reduced the average loss (i.e. the difference between the top ranked model and the best model) for 28 MQAPs. Particularly, for MQAPs with low average correlations (< 0.4), the correlation can be

increased by several times. Similar experiments conducted on the CASP9 MQAPs also demonstrated the effectiveness of the method. Our method is a hybrid method that combines the original method of a MQAP and the pair-wise comparison clustering method. It can achieve a high accuracy similar to a full pair-wise clustering method, but with much less computation time when evaluating hundreds of models. Furthermore, without knowing native structures, the iterative refining method can evaluate the performance of a MQAP by analyzing its model quality predictions.

4.2 Introduction

Nowadays, computer programs can generate a large number of protein models in a relatively short time, which makes protein model quality evaluation / assessment indispensable. Protein model quality assessment programs (MQAPs) can predict the qualities of protein models before knowing the experimental structures, which is essential to the proper usage of the models [34, 35, 55]. Current model quality assessment programs can predict both global and local qualities of one or multiple models. The methods used to predict global qualities can be categorized as multiple-model (clustering) methods and single-model methods.

Multiple-model methods assess the quality of a model by assessing its similarity with other models for the same protein target through full pair-wise structure comparisons [38, 39, 40, 42, 43, 56, 57]. Single-model methods directly assess the quality of a model from its structural features using machine learning, statistical, or physical methods [44, 45, 46, 47, 41, 58, 59, 60]. According to recent CASP experiments [61], multiple-model methods are currently more accurate than single-model methods, although they do not work well if only a small number of models are available or the structures of input models are largely different. Another drawback is that clustering method usually needs a relatively long computational time that makes it less efficient

and less feasible to be used in daily research. To address these problems, recently a hybrid quality assessment method [49] was developed to integrate the strengths of the two approaches. The hybrid method at first uses a single-model quality assessment method [58] to generate initial quality scores of input models, and then compares the structure of each model with those of the top ranked models. It uses the average structural similarity score with the top ranked models as predicted quality score.

Here we generalize the hybrid approach and use it to refine the quality scores predicted by any MQAPs. The iterative self-refining approach can consistently improve single-model MQAPs in almost all situations in just a few iterations. Our results showed that instead of performing full pair-wise comparisons between models, partial pair-wise comparisons against a few top models can achieve similarly high accuracy, but with much less computational time. Moreover, for the first time, the iterative method can help evaluate the performance of a MQAP before knowing the experimental structures. Although our algorithm can also generate local quality scores, in this article, we mainly focus on discussing its performances in improving global quality assessment.

4.3 Materials and Methods

The iterative quality assessment (IQA) method starts from the initial quality scores of a set of protein models. In the first round of refinement, the initial scores are used to rank all models. The top n models are selected as reference models and used to compare with every other models by a structural comparison tool TM-Score [52], which generates a GDT-TS score [10] for each comparison. The average GDT-TS score over the n reference models is used as the refined global quality score of a model. The new, presumably better, quality scores are then used to generate a new ranking of the models for the next round of refinement. The same refinement

process is executed iteratively until it converges, i.e., the ranking of models does not change any more. The average GDT-TS scores generated in the last round are used as the final global quality scores. When comparing a model to each of the n reference models in each round, TM-Score superimposes two models and outputs the superimposed coordinates of each pair of residues. These coordinates are used to calculate the residue-specific distances. The averaged residue-specific distances over the n reference models are used as the refined local quality scores. The average residue-specific distances generated in the last round are used as the final local quality scores. The only parameter of the iterative quality assessment is n , the number of reference models, which is set to five during most of our experiments except for Figure 4.1.

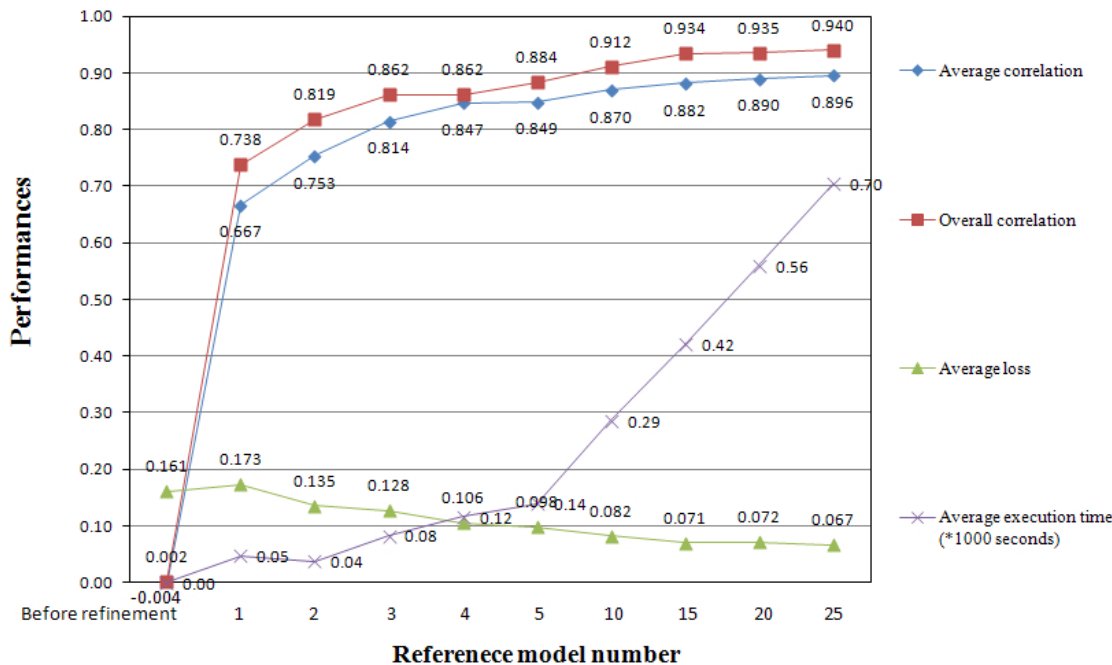


Figure 4.1: The average correlation, overall correlation, average loss, and average computational time under different numbers of reference models. This experiment was conducted on a MQAP whose predicted quality scores of CASP9 models were randomly generated. The predicted model quality scores had an average correlation of -0.0036 with the true model quality scores. Different numbers of reference models were tested under a single round of refinement.

4.4 Results and Discussions

4.4.1 Evaluation datasets and criteria

We applied our iterative refinement approach to each of the MQAPs that participated in the Eighth Critical Assessment of Techniques for Protein Structure Prediction (CASP8, 2008) and the Ninth Critical Assessment of Techniques for Protein Structure Prediction (CASP9, 2010). Taking CASP8 as an example, we downloaded the predicted quality scores of more than 50,000 tertiary structure (TS) models associated with 120 CASP8 targets from the CASP8 web site. We also downloaded all the TS models and compared each of them with its true experimental structure using the tool TM_Score [52]. The GDT-TS [10] score resulted from comparison is considered as the real quality score of the model. The real quality scores were used to evaluate whether the iterative quality assessment method improved the initial quality scores predicted by CASP8 MQAPs.

We evaluated the iterative quality assessment method using the following criteria: average and overall correlation of predicted and real GDT-TS scores, and average loss of the GDT-TS scores on top one ranked models. The average correlation is the average of the per-target Pearson correlations between predicted quality scores and real GDT-TS scores. The overall correlation is the Pearson correlation of predicted quality scores and real GDT-TS scores of all models of all CASP8 or CASP9 targets. The loss on a target is the difference between the real GDT-TS score of the top one ranked model and the real GDT-TS score of the best model. The average loss over all targets measures the ranking ability of a MQAP, which ideally equals to zero indicating the program can always rank the best model as the top one model.

4.4.2 Evaluation in terms of correlation and loss

Table 4.1 reports the average correlation, overall correlation, and average loss of 30 CASP8 MQAPs before and after applying our refinement algorithm. The method “ModFOLDclust” [57] is a full pair-wise clustering method that can serve as a baseline predictor for reference purpose. Our refinement method improved the performance of some single-model MQAPs, such as QMEAN, to a level close to that of ModFOLDclust. The average (overall) correlations of 25 (24) out of 30 MQAPs were increased. The average losses of 28 MQAPs were reduced. According to t-tests, the p-value of observing the difference before and after refinements for average correlation, overall correlation, and average loss is less than 0.0001, 0.0001, and 0.01, respectively. The correlations of MQAPs with low initial correlation scores (<0.4), such as qa-ms-torda-server and ProtAnG_s, were increased by up to 60 times. After refinement, the correlations of all MQAPs except one are improved to above 0.80; and the average losses of all the MQAPs except two are reduced to below 0.10. One extreme example is qa-ms-torda-sever, whose average correlation was improved from 0.012 to 0.767. However, we noticed that the refinement method did not improve the correlation of several clustering-based methods probably because they had already used structural comparisons in their model evaluation process. In contrast, all the single-model methods that do not utilize structural comparisons were improved by the iterative refinement method.

The same experiment was performed on 107 valid CASP9 targets (Table 4.2 and 4.3). Average correlation, overall correlation, and average loss of CASP9 MQAPs before and after iterative refinements. Bold fonts denote improvements. According to t-tests, the p-values of observing differences in average correlation, overall correlation, and average loss are less than 0.1, 0.1, and 0.05, respectively. The method “MULTICOM-CLUSTER” [62] is a full pair-wise clustering method that can serve as a baseline predictor for reference purpose. Our method improved the average cor-

relation, overall correlation, and average loss of almost all CASP9 MQAPs that did not use structural comparisons, such as PconsR, PconsD, PRECORES-QA, ProQ, MetaMAQP, Batymus, DistillNNPIE, ProQ2, ConQuass, MULTICOM-NOVEL, and QMEAN. However, our method rarely improved clustering-based MQAPs that used structural comparisons, such as MULTICOM-CLUSTER, MUFOLD-QA, and QMEAN-clust, although it slightly reduced the average loss of ModFOLDclust2, one of the top pair-wise comparison methods. According to t-test, the p-value of the improvements on average correlation and overall correlation is 0.1 for all CASP9 MQAPs, which is less significant than the ones on CASP8 data. This may be because a larger portion of CASP9 MQAPs used structural comparisons. However, the p-value of the improvements on loss is still at a significant level 0.05.

To investigate how fast the iterative QA method converged, we plotted the average loss against iterations for each CASP8 MQAP (Figure 4.2) and CASP9 MQAP (Figure 4.3). Most methods converged in the first one or two iterations (Figure 4.2 and 4.3). On average, it takes up to about five iterations to converge. The number of iterations depends on the quality of initial ranking. Better initial rankings require fewer iterations of refinement.

To investigate how “the number of reference models” influences the refinement performance and also the efficiency of our method, we created a random MQAP on CASP9 dataset (Figure 4.1). The predicted model quality scores of this random MQAP were randomly generated, which had an average correlation -0.00357, an overall correlation 0.0021, and a loss 0.161 compared with the true model quality scores. Models were then initially ranked by these randomly generated quality scores. After applied a single iteration of refinement using top 1 ranked model as reference model, our method substantially improved the average correlation to 0.667 and overall correlation to 0.738. Moreover, by using top 3 ranked models as reference models, both average correlation and overall correlation were improved to 0.814 and

0.862, respectively, after only one round of iteration. The improvement continued to increase as the number of reference models increased and started to saturate after using 15-25 reference models. By using 25 top models as reference models, the average correlation, overall correlation, and average loss were improved to 0.896, 0.940, and 0.067 respectively, which were much better than the initial ranking generated by the random MQAP. Its performance was also close to the average correlation 0.916, overall correlation 0.947, and average loss 0.059 of a full pair-wise comparison method MULTICOM-CLUSTER [62], which was developed by our group and was ranked as one of the top MQAPs in CASP9 (see Table 4.2 and 4.3).

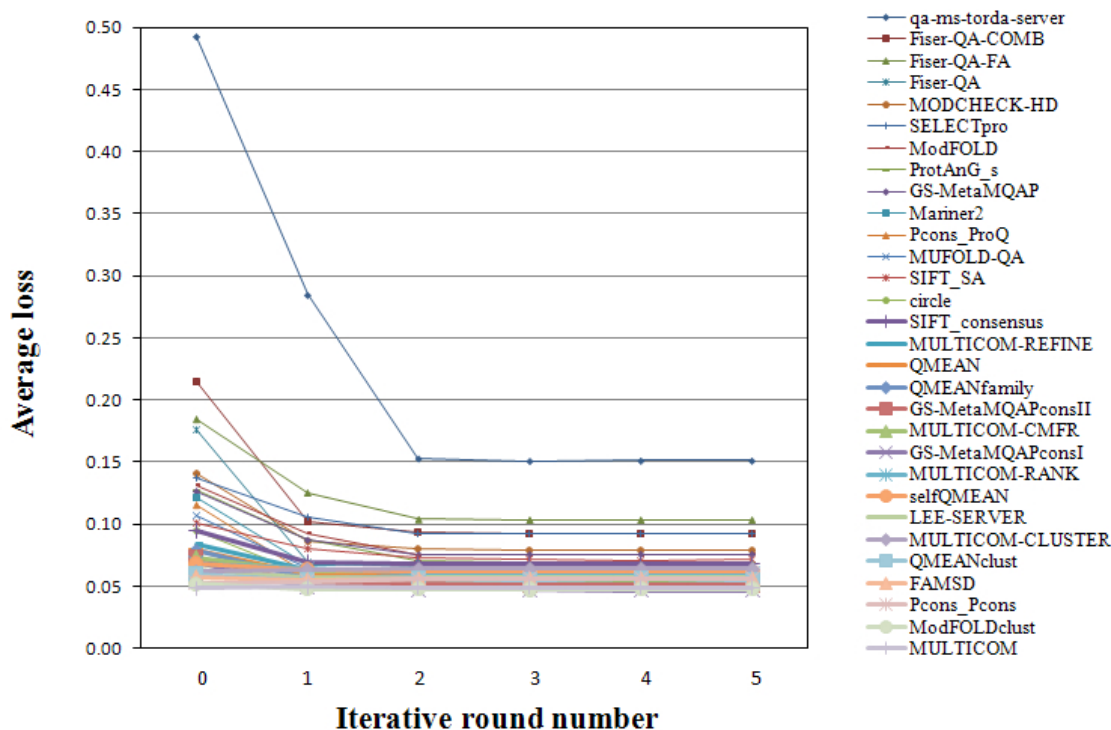


Figure 4.2: The plot of the average losses against iterations for CASP8 MQAPs. The method “ModFOLDclust” [57] is a full pair-wise clustering method that can serve as a baseline predictor for reference purpose.

We studied some cases in which our refinement method worked well or failed. Using the random MQAP mentioned above as an example, we noticed that it worked well on template-based modeling (TBM) targets whose model pool largely contains

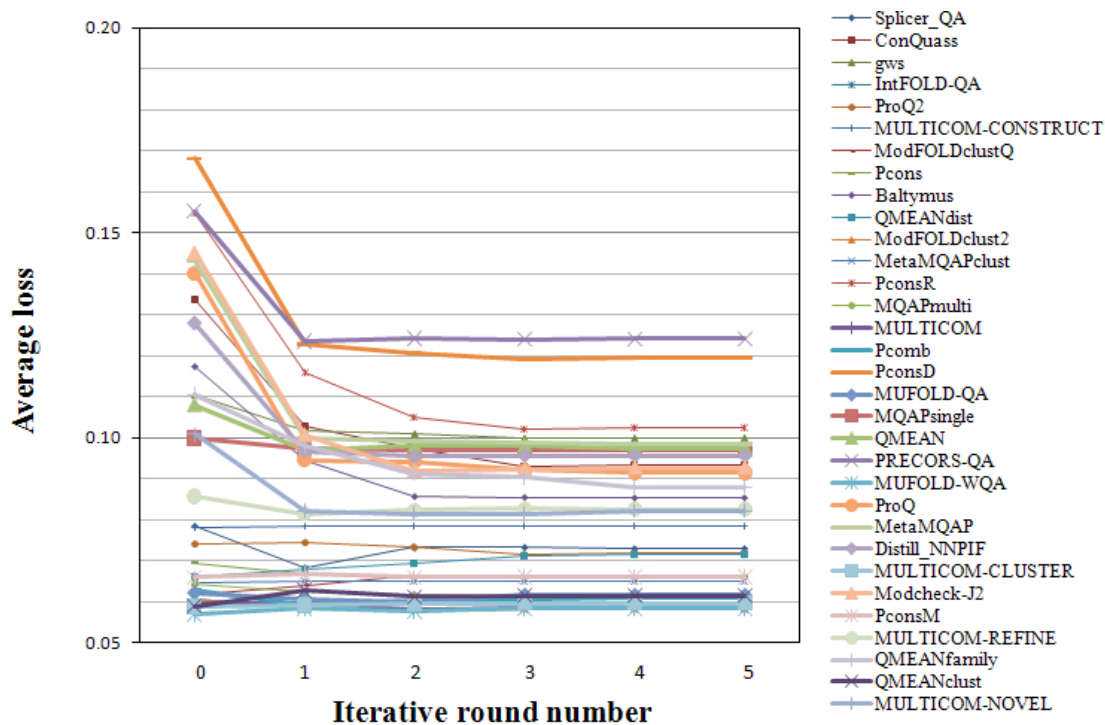


Figure 4.3: The plot of the average losses against iterations for CASP9 MQAPs. The method “MULTICOM-CLUSTER” [62] is a full pair-wise clustering method that can serve as a baseline predictor for reference purpose.

high quality models. For example, the predictions of the random MQAP had an average correlation -0.071 on an easy TBM target T0522; and 218 out of 371 models of that target have true GDT-TS scores > 0.9 . GDT-TS score is a structural similarity score that ranges from 0 to 1, whereas 1 indicates the model is the same as the native structure and 0 completely different [10]. After one round of refinement using the top one ranked model as reference model, the average correlation was improved to 0.985. In contrast, our refinement method did not work well on some of the hard targets whose models are mostly of low quality. For example, the random MQAP has an original correlation -0.013 for target T0537. After one round of refinement using the top one ranked model as reference model, the correlation became -0.067. T0537 is a hard target that contains two free modeling (FM) domains. The best model generated by CASP9 tertiary structure predictors has a GDT-TS score 0.32 whereas all other models have a GDT-TS score < 0.3 . These two extreme examples may suggest that, similarly as clustering method, our iterative refinement method works better if a large portion of input models have reasonable qualities.

4.4.3 Evaluation in terms of ranking correlation

Moreover, to investigate how model rankings are changed during the refinement process, we calculated the average Kendall tau rank correlation and Spearman’s rank correlation. Kendall tau rank correlation coefficient is defined as:

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n - 1)}$$

where n_c is the number of concordant pair of models whose ranking orders are not changed in two rankings; n_d is the discordant pairs; and n is the total number of models in the ranking. Kendall tau rank correlation measures the agreement level between two rankings and ranges from -1 and 1, while 1 indicates the two rankings

are the same, -1 one ranking is the reverse of the other, and 0 the two rankings are completely independent.

The Spearman’s rank correlation coefficient is defined as:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where $d_i = x_i - y_i$, which is the difference between the ranking orders of a model in two rankings; and n is the number of models in the rankings.

The average Kendall tau and Spearman’s rank correlations were plotted against iteration numbers in Figures 4.4 and 4.5 for CASP8, and Figures 4.6 and 4.7 for CASP9. Similarly as for the average correlation, it took about three iterations to converge on average. For almost all the cases, the biggest increase happened after the first iteration of refinement. The “rank correlations between the rankings before and after the first iteration of refinement” (RCBAF) is particularly interesting since it reports the degree a ranking is changed by the refinement. The RCBAF of initially less accurate MQAPs (e.g. qa-ms-torda-server) is much lower than that of initially more accurate MQAPs (e.g., Pcons-Pcons, ModFOLDclust, QMEANclust, and MULTICOM). Table 4.4, 4.5 and 4.6 report the average Spearman’s and Kendall tau rank correlations before and after the first and last iteration. The RCBAF for a less accurate MQAP is relatively low (e.g. < 0.5 for Spearman’s and < 0.4 for Kendall tau). These suggest that the RCBAF can be used to assess the quality of a MQAP.

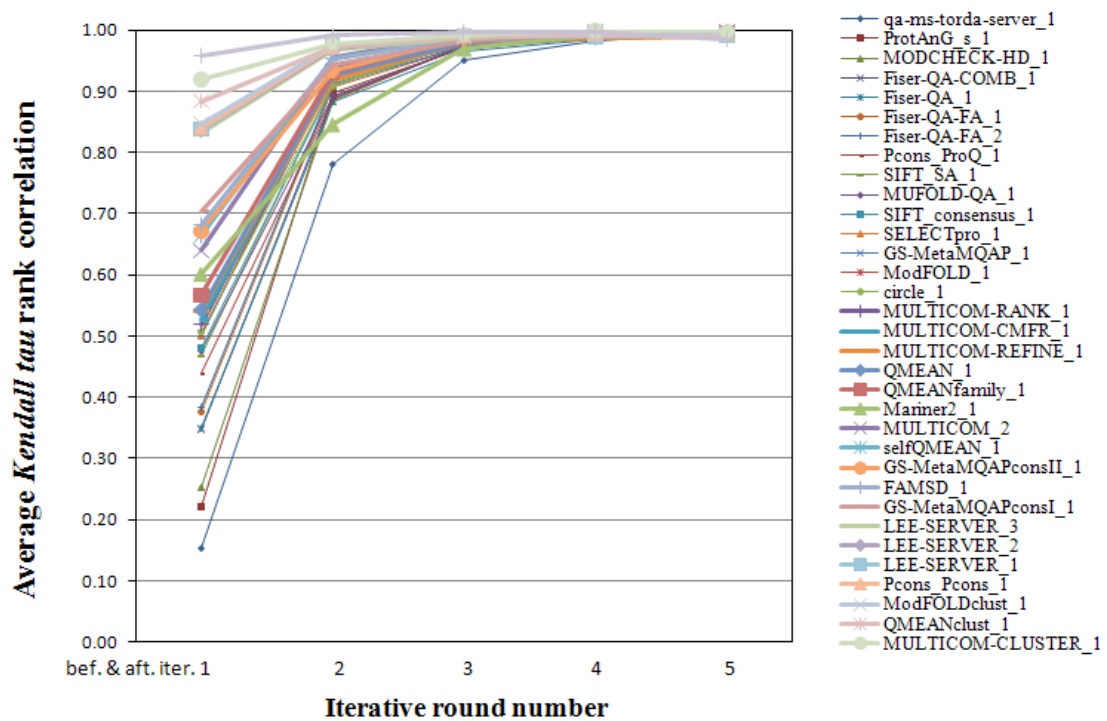


Figure 4.4: The Kendall tau rank correlations of the rankings before and after each round of refinement for CASP8 MQAPs.

Table 4.1: Performances of CASP8 MQAPs before and after iterative refinements.

	Avg. corr.		Over. corr.		Avg. loss	
	Bef.	Aft.	Bef.	Aft.	Bef.	Aft.
qa-ms-torda-server	0.012	0.767	0.110	0.730	0.483	0.149
ProtAnG _s	0.145	0.823	0.100	0.878	0.130	0.070
MODCHECK-HD	0.284	0.826	0.501	0.858	0.141	0.081
Fiser-QA-COMB	0.476	0.836	0.484	0.856	0.214	0.092
Fiser-QA-FA	0.485	0.822	0.287	0.834	0.183	0.105
Fiser-QA	0.523	0.857	0.506	0.879	0.176	0.063
ModFOLD	0.597	0.835	0.681	0.868	0.132	0.076
SELECTpro	0.608	0.805	0.432	0.844	0.138	0.093
SIFT_SA	0.623	0.840	0.459	0.858	0.102	0.074
MUFOLD-QA	0.633	0.832	0.576	0.872	0.108	0.067
Pcons_ProQ	0.652	0.860	0.652	0.882	0.114	0.055
SIFT_consensus	0.658	0.850	0.673	0.869	0.097	0.068
MULT.-RANK	0.665	0.838	0.705	0.867	0.069	0.061
QMEANfamily	0.678	0.847	0.733	0.869	0.080	0.058
GS-MetaMQAP	0.681	0.843	0.771	0.856	0.124	0.079
Circle	0.683	0.862	0.658	0.881	0.098	0.055
QMEAN	0.699	0.859	0.740	0.877	0.081	0.060
MULT.-REFINE	0.710	0.848	0.772	0.871	0.085	0.061
MULT.-CMFR	0.721	0.836	0.734	0.869	0.075	0.066
Mariner2	0.730	0.813	0.877	0.889	0.126	0.068
FAMSD	0.825	0.856	0.661	0.880	0.060	0.058
selfQMEAN	0.833	0.842	0.893	0.892	0.071	0.063
GS-M.PconsII	0.838	0.866	0.829	0.882	0.074	0.053
GS-M.PconsI	0.860	0.870	0.855	0.883	0.072	0.051
MULT.-CLUSTER	0.865	0.847	0.878	0.871	0.064	0.066
LEE-SERVER	0.866	0.882	0.778	0.878	0.062	0.056
MULTICOM	0.879	0.869	0.891	0.886	0.050	0.049
QMEANclust	0.886	0.864	0.919	0.909	0.062	0.056
ModFOLDclust	0.894	0.856	0.891	0.878	0.053	0.049
Pcons_Pcons	0.900	0.840	0.886	0.870	0.055	0.057

Table 4.2: Performances of CASP9 MQAPs before and after iterative refinements - 1.

	Avg. corr.		Over. corr.		Avg. loss	
	Bef.	Aft.	Bef.	Aft.	Bef.	Aft.
ModFOLDclustQ	0.832	0.849	0.929	0.898	0.062	0.066
QMEANdist	0.833	0.854	0.788	0.863	0.066	0.071
MULT.REFINE	0.866	0.821	0.929	0.918	0.086	0.083
Pcomb	0.870	0.862	0.929	0.892	0.063	0.061
MULTICOM	0.885	0.860	0.933	0.925	0.060	0.059
PconsM	0.885	0.838	0.930	0.893	0.066	0.066
IntFOLD-QA	0.887	0.870	0.940	0.912	0.060	0.058
ModFOLDclust2	0.888	0.863	0.944	0.915	0.061	0.058
Pcons	0.893	0.851	0.933	0.881	0.069	0.066
MQAPmulti	0.895	0.855	0.932	0.920	0.064	0.061
MetaMQAPclust	0.896	0.835	0.936	0.919	0.064	0.065
MULT.-CLUSTER	0.916	0.872	0.947	0.912	0.059	0.060
MUFOLD-QA	0.920	0.874	0.941	0.914	0.062	0.062
MUFOLD-WQA	0.920	0.865	0.896	0.888	0.057	0.058
QMEANclust	0.921	0.865	0.950	0.917	0.059	0.061

Table 4.3: Performances of CASP9 MQAPs before and after iterative refinements - 2.

	Avg. corr.		Over. corr.		Avg. loss	
	Bef.	Aft.	Bef.	Aft.	Bef.	Aft.
PconsR	0.052	0.629	0.026	0.743	0.155	0.102
PconsD	0.119	0.649	-0.158	0.605	0.168	0.120
PRECORS-QA	0.260	0.676	0.065	0.694	0.155	0.124
ProQ	0.415	0.777	0.665	0.684	0.140	0.092
MetaMQAP	0.583	0.783	0.744	0.883	0.143	0.098
Baltymus	0.586	0.810	0.573	0.888	0.117	0.085
Dist._N.	0.601	0.757	0.626	0.833	0.128	0.096
ProQ2	0.627	0.798	0.781	0.901	0.074	0.072
ConQuass	0.656	0.837	0.722	0.853	0.134	0.093
MULT.-NOVEL	0.662	0.795	0.767	0.890	0.101	0.082
QMEAN	0.685	0.777	0.808	0.889	0.108	0.097
QMEANfamily	0.697	0.805	0.805	0.904	0.111	0.088
Modcheck-J2	0.730	0.799	0.820	0.884	0.145	0.093
Gws	0.769	0.772	0.868	0.893	0.110	0.100
MQAPsingle	0.810	0.766	0.926	0.906	0.100	0.097
Splicer_QA	0.818	0.827	0.885	0.914	0.079	0.073
MULT.-CONSTRUCT	0.832	0.806	0.903	0.898	0.078	0.078

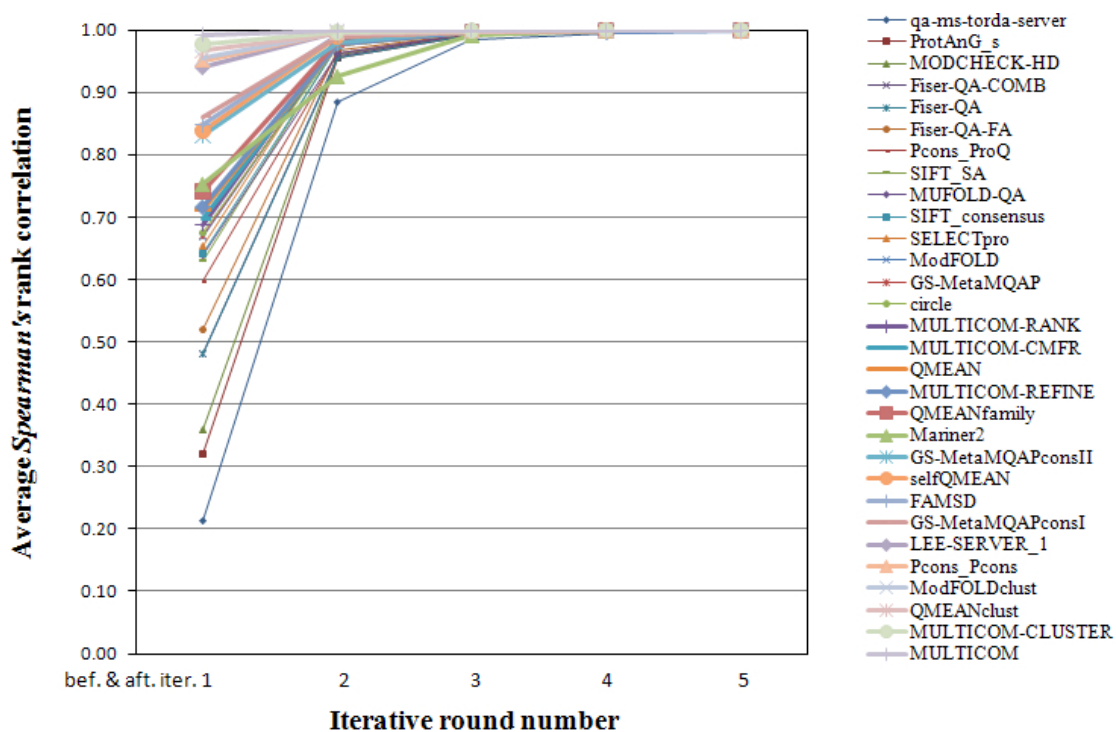


Figure 4.5: The Spearman's rank correlations of the rankings before and after each round of refinement for CASP8 MQAPs.

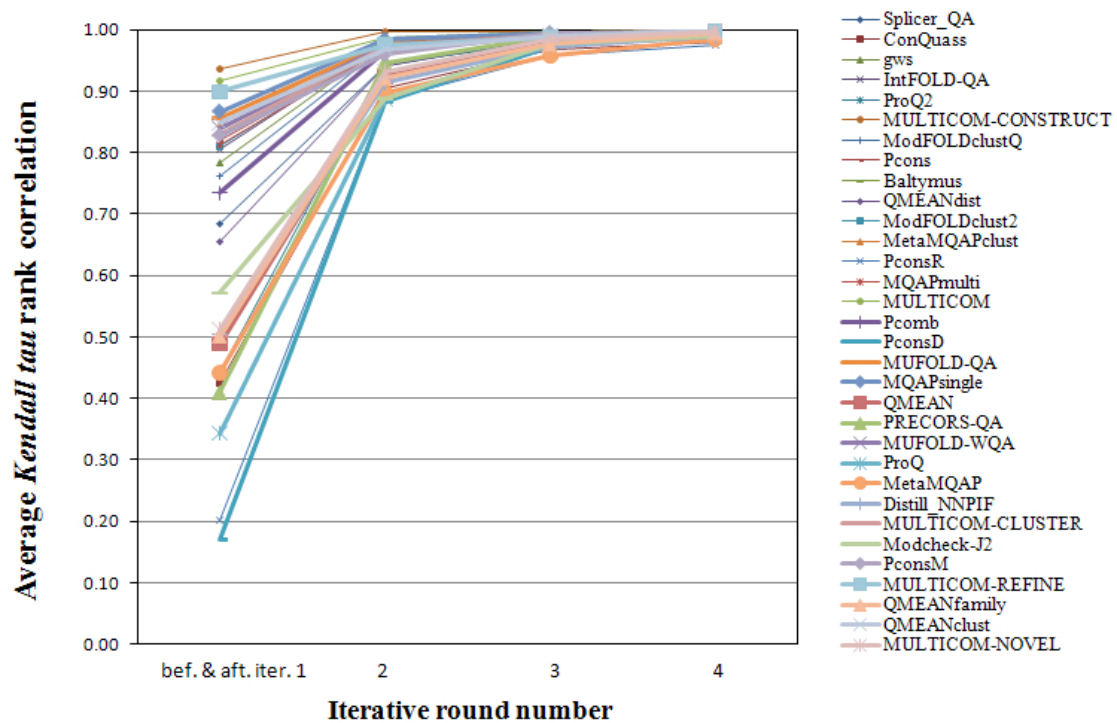


Figure 4.6: The Kendall tau rank correlations of the rankings before and after each round of refinement for CASP9 MQAPs.

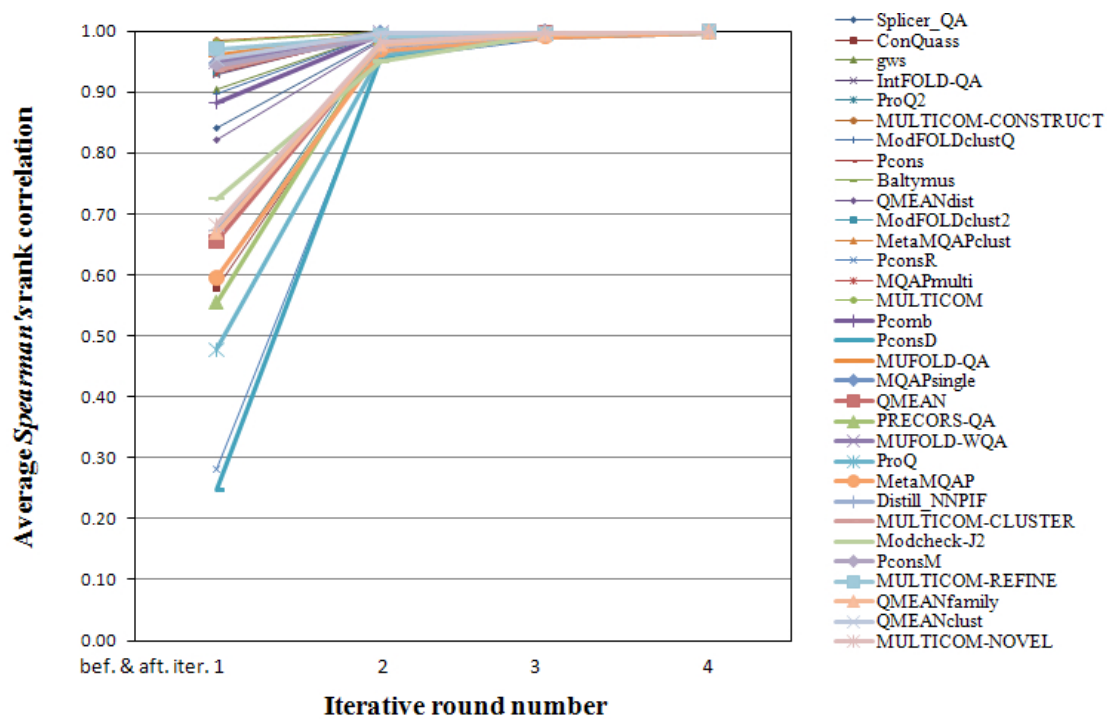


Figure 4.7: The Spearman's rank correlations of the rankings before and after each round of refinement for CASP9 MQAPs.

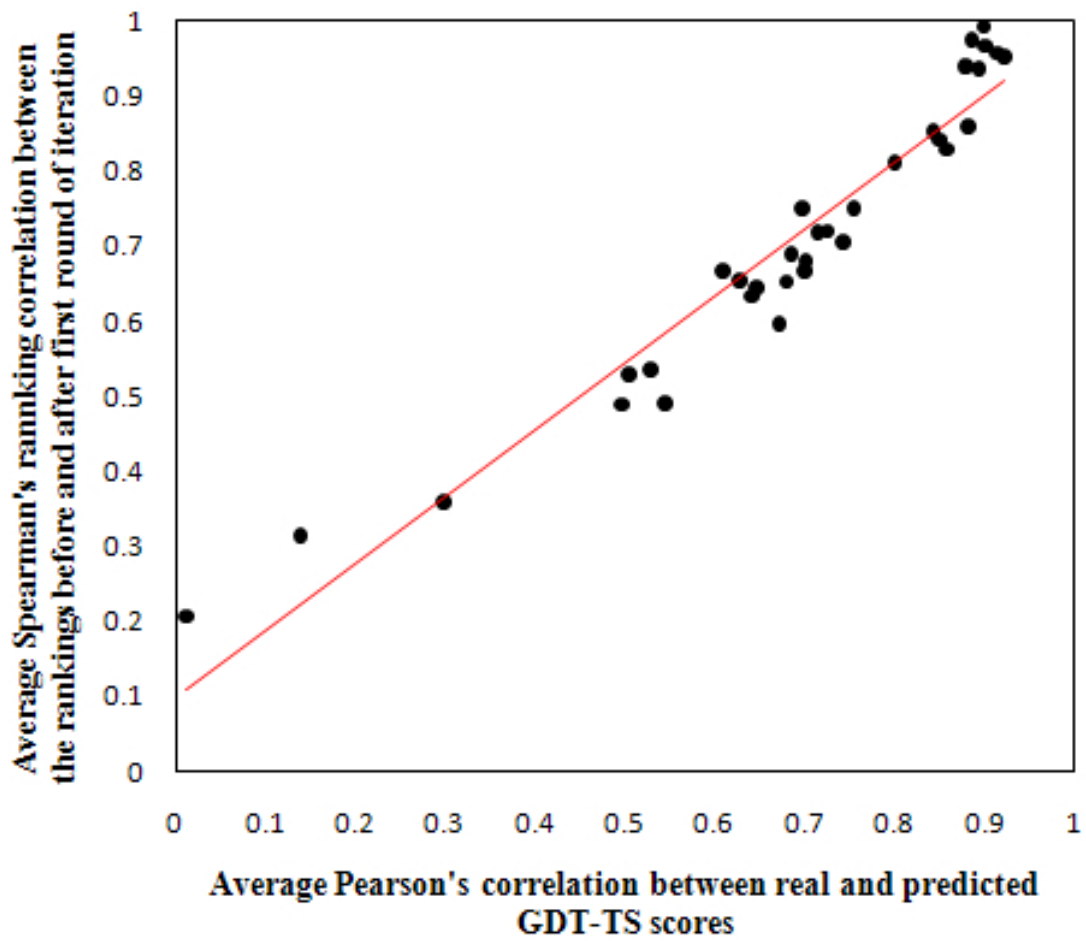


Figure 4.8: The plot of the Spearman's RCBAF values against the average per-target correlation of the 30 CASP8 MQAPs. Their Pearson's correlation is 0.965.

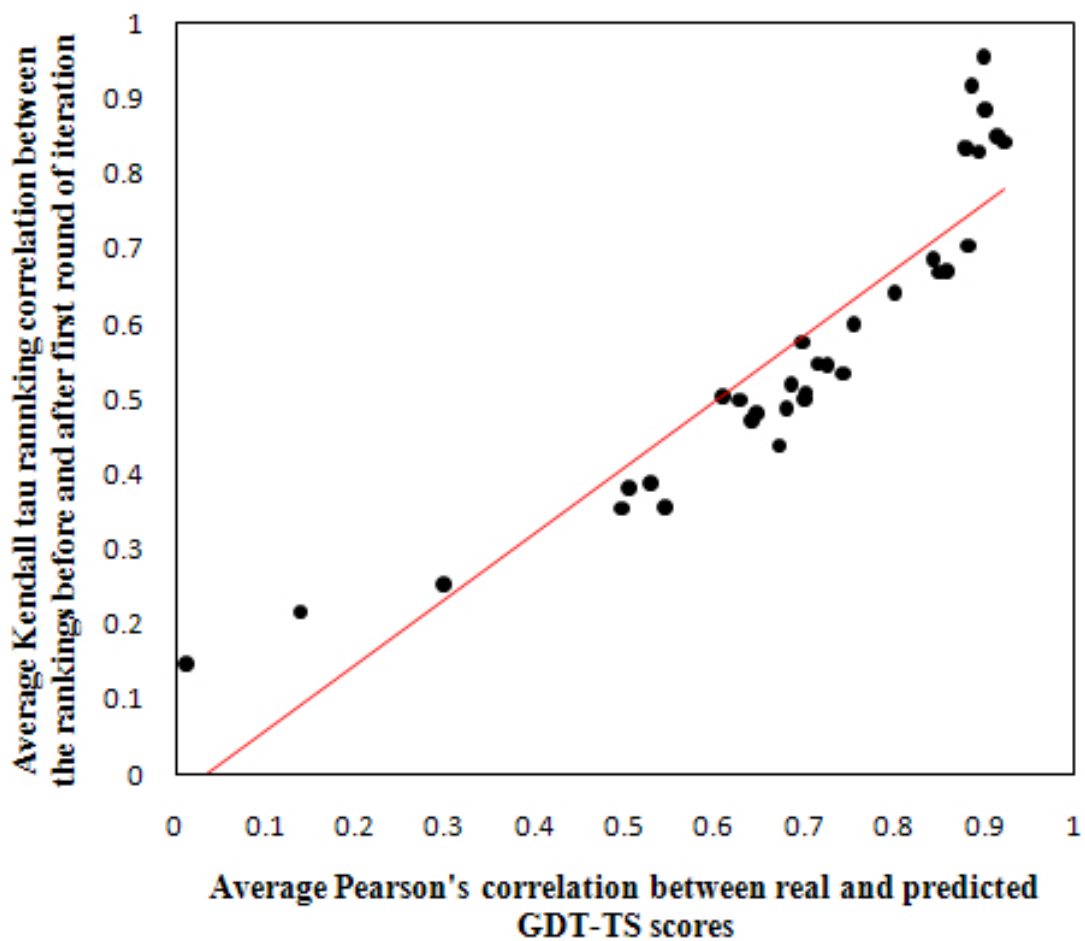


Figure 4.9: The plot of the Kendall tau RCBAF values against the average per-target correlation of the 30 CASP8 MQAPs. Their Pearson's correlation coefficient is 0.911.

Table 4.4: Ranking correlations tested on CASP8 MQAPs.

	Avg. Kendall corr.		Over. Spearman's corr.	
	Bef.	Aft.	Bef.	Aft.
qa-ms-torda-server	0.153	0.991	0.214	0.999
ProtAnG_s	0.221	0.993	0.320	0.999
MODCHECK-HD	0.254	1.000	0.358	1.000
Fiser-QA-COMB	0.348	0.998	0.482	1.000
Fiser-QA	0.348	0.989	0.482	0.999
Fiser-QA-FA	0.375	0.991	0.521	0.999
Pcons_ProQ	0.438	0.992	0.597	0.999
SIFT_SA	0.468	0.987	0.630	0.998
MUFOLD-QA	0.476	0.996	0.638	1.000
SIFT_consensus	0.479	0.989	0.642	0.998
SELECTpro	0.499	0.987	0.654	0.999
GS-MetaMQAP	0.503	0.988	0.670	0.999
ModFOLD	0.504	0.991	0.668	0.999
circle	0.506	0.988	0.676	0.998
MULTICOM-RANK	0.518	0.990	0.689	0.998
MULTICOM-CMFR	0.528	1.000	0.696	1.000
MULTICOM-REFINE	0.541	0.983	0.717	0.997
QMEAN	0.542	0.992	0.713	0.999
QMEANfamily	0.567	0.984	0.741	0.998
Mariner2	0.601	0.990	0.753	0.998
selfQMEAN	0.665	0.992	0.830	0.998
GS-MetaMQAPconsII	0.671	0.986	0.838	0.999
FAMSD	0.681	0.991	0.848	0.999
GS-MetaMQAPconsI	0.705	0.999	0.862	1.000
LEE-SERVER	0.837	1.000	0.941	1.000
Pcons_Pcons	0.840	0.989	0.951	0.998
ModFOLDclust	0.847	0.969	0.955	0.995
QMEANclust	0.884	0.961	0.967	0.994
MULTICOM-CLUSTER	0.919	0.978	0.977	0.999
MULTICOM	0.958	0.970	0.993	0.998

Table 4.5: Ranking correlations tested on CASP9 MQAPs - 1.

	Avg. Kendall corr.		Over. Spearman's corr.	
	Bef.	Aft.	Bef.	Aft.
PconsD	0.170	0.992	0.246	0.999
PconsR	0.201	0.975	0.282	0.994
ProQ	0.343	0.989	0.478	0.998
PRECORS-QA	0.411	0.996	0.555	1.000
ConQuass	0.424	0.981	0.578	0.997
Baltymus	0.438	0.984	0.595	0.997
ProQ2	0.441	0.987	0.596	0.998
MetaMQAP	0.442	0.984	0.595	0.998
QMEAN	0.489	0.998	0.655	1.000
QMEANfamily	0.500	0.995	0.669	0.999
Distill_NNPIF	0.505	0.989	0.673	0.998
MULTICOM-NOVEL	0.512	0.996	0.682	1.000
Modcheck-J2	0.571	0.990	0.725	0.999
QMEANdist	0.654	0.999	0.822	1.000
Splicer_QA	0.684	0.996	0.842	0.999
Pcomb	0.734	0.995	0.884	0.999
ModFOLDclustQ	0.763	0.997	0.897	1.000

Table 4.6: Ranking correlations tested on CASP9 MQAPs - 2.

	Avg. Kendall corr.		Over. Spearman's corr.	
	Bef.	Aft.	Bef.	Aft.
gws	0.783	0.999	0.905	1.000
IntFOLD-QA	0.806	1.000	0.929	1.000
ModFOLDclust2	0.810	0.999	0.932	1.000
MQAPmulti	0.812	0.996	0.932	1.000
Pcons	0.820	0.996	0.941	0.999
PconsM	0.828	0.996	0.944	1.000
MULTICOM-CLUSTER	0.830	0.999	0.936	1.000
MUFOLD-WQA	0.839	0.998	0.951	1.000
MetaMQAPclust	0.847	0.996	0.951	1.000
QMEANclust	0.848	0.998	0.954	1.000
MUFOLD-QA	0.856	0.999	0.961	1.000
MQAPsingle	0.867	0.991	0.948	0.999
MULTICOM-REFINE	0.899	0.998	0.970	1.000
MULTICOM	0.918	1.000	0.983	1.000
MULTICOM-CONSTRUCT	0.936	1.000	0.984	1.000

Chapter 5

SoyDB: A Knowledge Database of Soybean Transcription Factors

5.1 Abstract

Transcription factors play the crucial role of regulating gene expression and influence almost all biological processes. Systematically identifying and annotating transcription factors can greatly aid further understanding their functions and mechanisms. In this article, we present SoyDB, a user friendly database containing comprehensive knowledge of soybean transcription factors. The soybean genome was recently sequenced by the Department of Energy-Joint Genome Institute (DOE-JGI) and is publicly available. Mining of this sequence identified 5,671 soybean genes as putative transcription factors. These genes were comprehensively annotated as an aid to the soybean research community. We developed SoyDB - a knowledge database for all the transcription factors in the soybean genome. The database contains protein sequences, predicted tertiary structures, putative DNA binding sites, domains, homologous templates in the Protein Data Bank (PDB), protein family classifications, multiple sequence alignments, consensus protein sequence motifs, web logo of

each family, and web links to the soybean transcription factor database PlantTFDB, known EST sequences, and other general protein databases including Swiss-Prot, Gene Ontology, KEGG, EMBL, TAIR, InterPro, SMART, PROSITE, NCBI, and Pfam. The database can be accessed via an interactive and convenient web server, which supports full-text search, PSI-BLAST sequence search, database browsing by protein family, and automatic classification of a new protein sequence into one of 64 annotated transcription factor families by hidden Markov models. A comprehensive soybean transcription factor database was constructed and made publicly accessible at <http://casp.rnet.missouri.edu/soydb/>.

5.2 Background

Soybean is a great source of protein, as it contains significant amounts of all the essential amino acids, including some that cannot be synthesized by the human body [63]. Soybean has been used as a food and a drug component in China for thousands of years [64] and over the past 60 years has become a leading crop in many nations around the world [65]. Because of its high value in the agricultural and food industry, soybean has received greater and greater research attention, both to improve soybean agronomic performances and as a model for basic biological studies. In early 2008, the Department of Energy-Joint Genome Institute (DOE-JGI) finished sequencing the soybean genome using a whole-genome shotgun approach [66], which makes soybean the most complex plant so far ever sequenced (<http://www.phytozome.net/soybean>). The homology-based gene prediction and annotation produced putative protein sequences [66](<http://www.phytozome.net/soybean>), which makes it feasible to identify and annotate soybean transcription factors.

Transcription factors (TF) are proteins that bind to DNA sequences (i.e. promoters) and regulate gene expression by one or more DNA binding domains. Virtually all

biological processes are directly regulated or influenced by transcription factors [67]. For example, the transcription process in eukaryotes would not occur in the absence of a specific class of transcription factors named “general transcription factors” [68, 69]. Studies have shown that transcription factors are closely involved in the process of cell development, such as cellular division, migration, and differentiation [70]. Transcription factors of *Arabidopsis thaliana* have been well studied since its genome has been fully sequenced as a model specie [67, 71, 72, 73, 74, 75]. This makes it possible to identify and study transcription factors of other newly sequenced species, such as soybean, by homology searching and comparative analysis.

Several databases for soybean genome analysis have been built and made publicly available, such as SoyGD [76], SoyBase (<http://soybase.org/>), and SoyXpress [77]. These databases contain a variety of information, such as soybean genome sequences, bacterial artificial chromosome (BAC), expressed sequence tags (EST), and some useful tools including genome browsers, BLAST searching, and pathway searching. However, these databases only contain general annotations for the soybean genome, instead of knowledge specifically targeting the transcription factors. For example, none of them systematically organizes transcription factors into families or clearly points out the DNA binding domains. PlantTFDB [78] and DBD [79] are two existent transcription factor databases, which contain knowledge about transcription factors from multiple species. For each soybean transcription factor, PlantTFDB contains information including protein sequence, Gene Ontology annotation [12], putative binding domains found by InterProScan [80], and cross-links to external databases, including EMBL [81], UniProt [82], RefSeq [83], and TRANSFAC [84]. DBD contains the amino acid sequence of each transcription factor and external links to Ensembl [85], Pfam [86], and SUPERFAMILY [87]. Compared to PlantTFDB, DBD has less external database links, but DBD claims to contain the transcription factors of 927 completely sequenced genomes whereas PlantTFDB cov-

ers 22 species. The knowledge in these two databases is very useful; however, they were built based on a relatively older version of soybean sequence data and their annotations are still incomplete. The most important component they lack is the three dimensional structure for each transcription factor, because the visualization of the transcription factor, especially binding sites, can help further understanding the mechanism and functions of transcription factors, which is indispensable to structural genomics [88, 89]. Furthermore, with the complete genome sequences of more and more species available, a computer system is needed that can automatically generate a knowledge database and publish it with a user-friendly interface.

To fill the gap, we developed SoyDB - a comprehensive and integrated database for soybean transcription factors. This database not only contains most of the content and features already existed in PlantTFDB and DBD, but also extends them by containing more comprehensive knowledge and links to more versatile external datasets. The annotations in SoyDB include predicted tertiary structures, protein domains, multiple sequence alignments, DNA binding sites, and web logos and consensus sequences for each family. The SoyDB database also contains links to the homologous EST sequences, and the same or homologous proteins in other databases including PlantTFDB [78], PDB [90], Swiss-Prot [91], TAIR [92], RefSeq [83], SMART [93], Pfam [86], KEGG [94], SPRINTS [95], EMBL [96], InterPro [97], PROSITE [98], and Gene Ontology [12].

Moreover, our system can automatically execute bioinformatics tools and generate annotations, link to other well-known protein databases, construct MySQL databases, and generate PHP scripts to build its website. This fully automated approach can be used to create a protein annotation database and website for any sequenced organism in the future.

5.3 Construction and Content

5.3.1 Database Overview

SoyDB contains the annotations of 5,671 putative transcription factors. These proteins were classified into 64 families (for details see the section “Transcription Factor Family Prediction Using SAM Hidden Markov Models”). Figure 5.1 illustrates the architecture of the SoyDB website. Users can access the main components from the home page: full-text search, PSI-BLAST sequence search, family classification by hidden Markov model, family browsing, protein browsing, family information, protein information, and FTP site.

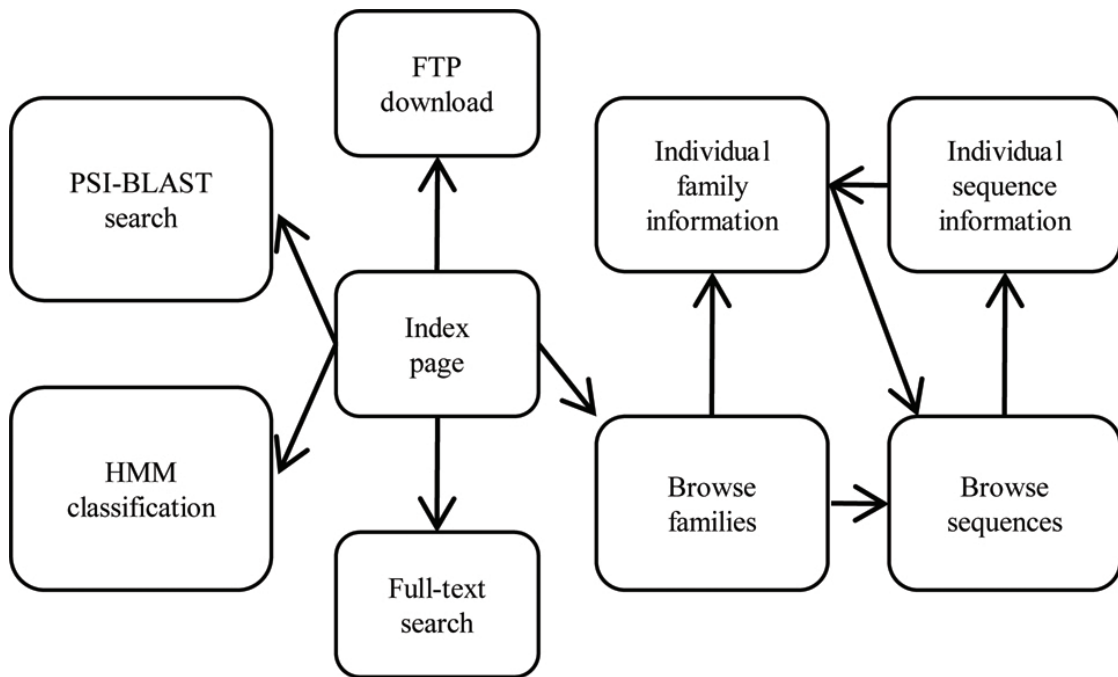


Figure 5.1: Architecture of SoyDB website. Main annotation components include family information page and sequence information page. Main functional components include full-text search, HMM family classification, PSI-BLAST sequential search, and FTP downloading.

5.3.2 Data Source

The soybean genome sequences and gene model predictions used in this study were acquired from the publicly available database Phytozome (<http://www.phytozome.net/soybean>). These sequences were generated by the preliminary GenomeScan [99], FgenesH (<http://linux1.softberry.com/berry.phtml>), and PASA [100] gene annotations based on the Gm1.01 version of soybean assembly data [66].

5.3.3 Transcription Factor Identification

We used the standalone versions of InterProScan [80] to search all the soybean protein sequences against 11 databases integrated in InerPro [97]. These databases and their corresponding scanning methods include: PROSITE (pfscan) [98], PRINTS (FingerPRINTSscan) [101], Pfam (HMMPfam) [86], ProDom (ProDomBlast3i) [102], SMART (HMMSmart) [93], TIGRFAMs (HMMTigr) [103], PIR SuperFamily (HMM-PIR) [104], SUPERFAMILY (superfamily) [105], Gene3D (gene3d) [106], PANTHER (HMMPanther) [107], and HAMAP (pfscan) [108]. InterProScan systematically searches each of these databases using their corresponding scanning methods to find domains. The proteins predicted to contain TF related domain(s) were considered as putative transcription factors. Using the Plant Transcription Factor Database (PlnTFDB) [109] and the classification of *Medicago truncatula* TF genes [110] as references, we manually curated the list of putative transcription factors and eliminated any mistakenly identified sequences. In this way, we identified 5,671 putative TF sequences.

5.3.4 Transcription Factor Family Prediction Using SAM Hidden Markov Models

The transcription factors of *Arabidopsis thaliana* have been well studied and classified into 64 families [92]. This provides a model for us to classify soybean transcription

factors. We used MUSCLE [111] to generate a multiple sequence alignment for each *Arabidopsis thaliana* TF family. The multiple sequence alignment was then input into SAM 3.5 [112] to build a hidden Markov model (HMM) for each family. Every soybean TF sequence was aligned with each of the 64 HMMs, which outputs an e-value. This e-value can be considered as a fitness score between a TF sequence and a hidden Markov model: lower e-value indicates better fitness. Finally, a transcription factor was classified into the family whose HMM yields the lowest e-value.

5.3.5 Annotations Using Bioinformatics Tools

The standalone versions of several bioinformatics tools were locally installed and executed to generate annotations for soybean transcription factors. An accurate protein structure prediction tool MULTICOM [11] was used to predict the tertiary structure of each transcription factor when homologous template proteins could be found in PDB. According to the official evaluations of the 8th community-wide Critical Assessment of Techniques for Protein Structure Prediction (CASP8) (<http://www.predictioncenter.org/casp8/results.cgi>), MULTICOM was able to predict high-accuracy tertiary structures with an average GDT-TS score [10] 0.87 if suitable templates can be found. GDT-TS score ranges from 0 to 1 measuring the similarities between the predicted and experimental structures, whereas 1 indicates completely the same and 0 completely different.

Figure 5.2 illustrates the predicted tertiary structure of a transcription factor in SoyDB with ID GM00002, and the electrostatic polarization of the predicted structure. The electrostatic polarization Figure 5.2(a) (blue, positive; red, negative), sphere Figure 5.2(b) and ribbon Figure 5.2(c) visualizations of MULTICOM predicted structure for GM00002. Figure 5.2(d) a segment of the pair-wise alignment between 1WID (PDB template of GM00002) and GM00002, and, below, the DNA-binding site predictions from three independent tools: DNABindR [113], BindN [114], and DP-

Bind [115] (“+” indicates predicted DNA-binding positions, “-” indicates gap or no prediction). The green regions in the sequence of 1WID are the DNA-binding regions identified by experimental methods [116]. The green regions in GM00002 sequence are the two DNA-binding regions derived from the alignment with 1WID. The predicted DNA-binding regions in GM00002 are illustrated in green in Figure 5.2(b) and Figure 5.2(c). Figure 5.2(c) the side chains of the predicted binding regions. Figure 5.2(a), Figure 5.2(d), and Figure 5.2(c) are in the same orientation. The electrostatic polarization (a) was computed and mapped to protein surface by Swiss-PDB viewer (deep view) [117], and the structures in sphere Figure 5.2(b) and ribbon styles Figure 5.2(c) were made with PyMol (<http://pymol.sourceforge.net/>). The blue area in the electrostatic polarization shows residues positively charged, which is found to be highly identical to the green area in Figure 5.2(b), which is the putative DNA-binding sites identified by a pair-wise alignment between GM00002 and its template protein 1WID (Figure 5.2(d)). Since it has been studied and found that the DNA-binding area is positively charged if analyzed by electrostatic potentials [118], the highly identical area in Figure 5.2(a) and (b) strongly confirms that the predicted structure has the electrostatic properties of a transcription factor. This further confirms the qualities of MULTICOM predictions and the correctness of the predicted binding sites derived from the homology alignment.

In SoyDB, a predicted tertiary structure is visualized by Jmol (<http://jmol.sourceforge.net/>). In order to clearly visualize the tertiary structure of the DNA-binding region, only the segments containing homologous DNA binding domains are visualized by Jmol. Users can view a TF structure from various perspectives in a three-dimensional way and perform many operations including selecting and highlighting interested regions, changing view styles and colors, and measuring atom distances and angles by right clicking on the Jmol console and selecting corresponding menus. Detailed instructions about Jmol menus can be found at Jmol website

(<http://jmol.sourceforge.net/>). During the structure prediction process, MULTICOM generates the sequence alignments between the transcription factor and its homologous templates using PSI-BLAST. These sequence alignments can be used to predict the binding sites of a transcription factor based on the experimentally determined binding sites of its template as shown in Figure 5.2.

A predicted structure was parsed into domains by Protein Domain Parser (PDP) [119]. Since some transcription factors did not have homologous templates found in PDB, DOMAC [120], an accurate ab initio domain prediction tool, was also used to predict domains for each transcription factor.

The protein sequences in the same family were aligned by MUSCLE [111] and visualized by WebLogo [121]. A consensus sequence was derived from the multiple sequence alignment by selecting the most frequently appeared amino acid at each position. The multiple sequence alignments were also used to identify the conserved signatures (likely the DNA binding domains) for each family.

All of the bioinformatics tools incorporated to construct SoyDB can be used to automatically annotate other species in the future.

5.3.6 Links to External Databases and Datasets

In order to annotate the functions of soybean transcription factors, each TF protein sequence was searched against the soybean TF database PlantTFDB, NCBI known EST sequences, and other general protein databases by PSI-BLAST or TBLASTN. The external protein databases include Swiss-Prot [91], TAIR [92], RefSeq [83], SMART [93], Pfam [86], KEGG [94], SPRINTS [95], EMBL [96], InterPro [97], PROSITE [98], and Gene Ontology [12]. Web links to these databases were created when the same transcription factor or its homologous proteins were found in them; and for each database or EST dataset only the PSI-BLAST or PBLASTN hit with the smallest e-values was listed in SoyDB. To search the known EST sequences, PSI-BLAST

was first used to build a position-specific score matrix for each transcription factor. TBLASTN was then used to search each protein sequence against three known EST datasets: EST human, EST mouse, and EST others. GenBank [122] web page of each EST hit was linked to SoyDB website. The gene expression of a subset of TF genes (about 1,000 TF genes) was recently published [123]. Transcription profile of all soybean TF genes in various conditions is under investigation.

These external links greatly expand the annotation scope of SoyDB providing related knowledge from various perspectives. SoyDB provides a systematic view of a transcription factor - from the features of the protein itself, to the biological pathway it locates in. The links to the external databases and datasets can be updated by a re-run of PSIBLAST and TBLASTN. Currently, these links are scheduled to be updated once every six months. This time interval can be changed if necessary.

5.3.7 Database and Website Implementation

The programs used to automatically annotate proteins were written in PERL. The relational database was built on MySQL with database schemas automatically generated by programs written in PERL. The website was implemented in PHP. The database and web site were automatically constructed by computer programs with little human intervention.

5.4 Utility and Discussion

5.4.1 Protein Information

This component contains the complete annotations for each transcription factor, including protein ID, protein name and description, tools used for TF identification, family ID, family name and description, amino acid sequence, homology domain pre-

diction, ab initio domain prediction, PDB homologous templates, and predicted tertiary structure. This component can be reached by clicking the sequence ID, such as GM00001, or the Phytozome protein name, such as Glyma01g11670.1, at the “Protein Browsing” webpage (for details see the following “Protein Browsing” section). Figure 5.3 illustrates the protein information page. The sequence ID and family ID, such as GM00001 and GMF0001, are internal indices used by the SoyDB, and the sequence name is the standard soybean TF name used by the soybean genome database Phytozome (<http://www.phytozome.net/soybean>). We noticed the trend of unifying annotation formats within the soybean community. Therefore, the commonly used TF ID format, such as PTGm00009.1, is also compatible in SoyDB. Details are described in the “Full-Text Search” section below.

5.4.2 Family Information

This component contains the complete annotations for each TF family, including family ID, family name and description, number of sequences within the family, consensus sequence, consensus signatures (likely the DNA binding regions), web logo of the signature profile, and multiple sequence alignment of the protein sequences within the family. Figure 5.4 demonstrates a family information web page. This component can be reached from the “Family Browsing” web page.

5.4.3 Protein Browsing

The transcription factors within a family are listed in the order of sequence IDs. The list contains the thumbnail of tertiary structure, protein ID and name, family ID, and family name of each transcription factor (Figure 5.5). Users can further view the complete annotations by clicking its sequence ID or the Phytozome protein name. This component can be reached by clicking the number of sequences in the “Family

[Home Page](#)
[Text Search](#)
[PSI-BLAST Search](#)
[Browse Database](#)
[Family Prediction by HMM](#)
[FTP](#)
[People](#)

Search for

Sequence ID

GM00001

Gene name and description

Glyma01g11670.1

Identified as a transcription factor by InterproScan, HMMPfam.

Family ID

GMF0001

Family name and description

ABI3-VP1

Amino acid sequence

```

IIIIAPSLQEGKLMLENKFVVEKYGEGLDNTLFLKADNGAEWKLTLKRRDDRHWFGWRFPAKHHSLDHGHLLEFRYQRTSHFQVHIFDGSGLLEYPLK
VEGRMTSNYQKRNKRNKLEVEFLQPCMGSRKCKVVDNMTKPKLGGCSACASRYRQKGRYITLSQLGHSPLYLTKMTTTEHVTAFDRASYRPNPSELV
VIYPSNARSRGGL

```

Homology domain

domain num: 2
domain 1: 1-96
domain 2: 97-213

Ab initio domain(s)

domain number: 2
domain 1: 1-119
domain 2: 120-213

Template(s)

PDB template: [1YELA](#)
Region aligned with template is [3, 92]

Template(s) alignment

[Click to view the alignment.](#)

Links to other databases

Data Source	Accession ID	Query Start	Query End	Sbjct Start	Sbjct End	e-value
PlantTFDB	PTGm00009.1	6	139	2	148	3e-52
SWISS-Prot	Q9FX77	12	96	22	105	2e-40
EMBL	AC011808	-	-	-	-	2e-40
EMBL	BT012511	-	-	-	-	2e-40
EMBL	BT014811	-	-	-	-	2e-40
RefSeq	NP_173109	-	-	-	-	2e-40
KEGG	ath_AT1G16640	-	-	-	-	2e-40
TAIR	At1g18640	-	-	-	-	2e-40
InterPro	IPR003340	-	-	-	-	2e-40
Pfam	PF02362	-	-	-	-	2e-40
PROSITE	PS50863	-	-	-	-	2e-40
EST Others	48806906	1	206	2	739	1e-47
EST Human	12431348	99	192	1263	1550	0.002
EST Mouse	24191927	87	156	54	263	0.19

[Home Page](#)
[Text Search](#)
[PSI-BLAST Search](#)
[Browse Database](#)
[Family Prediction by HMM](#)
[FTP](#)
[People](#)

Figure 5.3: Information page for a transcription factor. This example web page shows the knowledge for each transcription factor, which includes amino acid sequence, predicted tertiary structure, domain(s) found by homologous search and ab initio prediction, PDB template and alignment, and links to other protein databases and EST datasets.

Home Page	Text Search	PSI-BLAST Search	Browse Database	Family Prediction by HMM	FTP	People
---------------------------	-----------------------------	----------------------------------	---------------------------------	--	---------------------	------------------------

Search for

Family ID
GMF0008

Family Name and Description
BBR-BPC

Number of sequences
19

Consensus sequence

```

consensus:
-----GDFDAAGAEFDDKMALIFNLQ--GVDNNIRRVFDELAMFAPALA-----
-----AVLQTGLASAKNAAIR-----
-----APPTARAILKHRDFVQDLQDRLKERR-KFNLRAIRSDFKALCBLCNVAVIKD-----ANYFDHLK
GEVHLK-----AKGMALIVEGLTLEAIS-----
-----EAFELLD-DLKDRGSLALAK-----RGDLD-----RRIKLLQAEBAKRLELKLKDLKD-----SKK-----
-----NRRRKALEKD--LSEKPCNEDDALGSRLEFE-----EAVEEPPRDAKAAEDLGSFYV-GADGDKKGTALKN-----PVLGQILE-----DFR
SGGRKA-----total
score = 491.095283986217 average score = 1.79231855469422

```

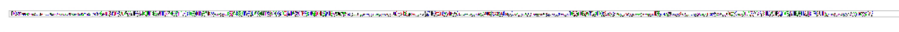
Consensus raw sequences

[Click to view the consensus sequence raw file.](#)

Signature of each sequence

[Click to view the signature of each sequence.](#)

Web LOGO



Multiple sequence alignment

[Click to view the multiple sequence alignment.](#)

Home Page	Text Search	PSI-BLAST Search	Browse Database	Family Prediction by HMM	FTP	People
---------------------------	-----------------------------	----------------------------------	---------------------------------	--	---------------------	------------------------

Figure 5.4: Family information page in SoyDB. This example web page shows the knowledge for each TF family, which includes number of sequences in the family, consensus sequence of the family, signature of sequences in the family, web logo, and multiple sequence alignment of the sequences in the family.

Browsing” or the “Family Information” web page.

5.4.4 Family Browsing

A user can browse SoyDB from TF family perspective. The 64 TF families are listed in the order of family IDs. The family ID, family name, and the number of transcription factors within each family are listed. By clicking the family ID or name, users can view the complete annotations for a family, or further browse the sequences within a family by clicking the number of sequences. This component can be reached by clicking “Browse Database” in both the top and bottom menu bars from any SoyDB web pages.

5.4.5 Full Text Search

This component allows users to search the entire SoyDB database by a query text, such as protein name or family name. Given input keywords, SoyDB searches all the fields in the database and returns matched transcription factors with links to their annotations. Users can also search SoyDB by the TF IDs used in PlantTFDB. The search component will return the homologous soybean TFs found in SoyDB.


5.4.6 PSI-BLAST Sequence Search

This component allows users to search a query sequence against every TF sequence stored in SoyDB. Users can submit a query sequence and adjust PSI-BLAST parameters from a web page. After a PSI-BLAST search is performed, the significant hits, with links to their annotation web pages, are ranked based on the e-values generated by PSI-BLAST.

Home Page	Text Search	PSI-BLAST Search	Browse Database	Family Prediction by HMM	FTP	People
---------------------------	-----------------------------	----------------------------------	---------------------------------	--	---------------------	------------------------

Search for

GM00001




Description: Glyma01g11670.1

Family ID: [GMF0001](#)

Family name: [ABI3-VP1](#)

GM00002




Description: Glyma01g32810.1

Family ID: [GMF0001](#)

Family name: [ABI3-VP1](#)

GM00003




Description: Glyma01g45640.1

Family ID: [GMF0001](#)

Family name: [ABI3-VP1](#)

GM00004




Description: Glyma02g36090.1

Family ID: [GMF0001](#)

Family name: [ABI3-VP1](#)

GM00005



Description: Glyma02g40280.1

Family ID: [GMF0001](#)

Family name: [ABI3-VP1](#)

1 - 5 of 99 [Next](#) [Last](#)

Home Page	Text Search	PSI-BLAST Search	Browse Database	Family Prediction by HMM	FTP	People
---------------------------	-----------------------------	----------------------------------	---------------------------------	--	---------------------	------------------------

Figure 5.5: Transcription factor browsing page in SoyDB. This page lists the transcription factors in a TF family. The tertiary structure of each sequence is displayed in an interactive way, i.e. users can zoom in/out and rotate the structure. Users can further view sequence annotations by clicking the TF IDs, and view family annotations by clicking family names.

5.4.7 Family Classification by Hidden Markov Model

This component classifies a query protein sequence into one of the 64 TF families. A submitted query sequence is aligned with each of the 64 hidden Markov models built based on the 64 *Arabidopsis thaliana* TF families. The query sequence is classified into a family whose hidden Markov model outputs the lowest e-value (correspondingly the highest alignment score or fitness score). More details about family classification can be found at the “Transcription Factor Family Prediction Using SAM Hidden Markov Models” section under “Construction and Content”.

5.4.8 FTP Site

All of the information in SoyDB is available for users to download from an FTP site. For example, users can download all of the TF protein sequences in the FASTA format and the multiple sequence alignments for each family in plain text. This makes it possible for other websites to link to SoyDB by performing PSI-BLAST searches on SoyDB sequences, similarly as SoyDB links with other external databases.

5.4.9 Comparisons and Overlapping between SoyDB and PlantTFDB

In total, SoyDB has 5,671 transcription factors - 4,306 of them (75.9%) have hits found in PlantTFDB identified by PSI-BLAST with an e-value threshold 10^{-3} . PlantTFDB has 1,891 soybean transcription factors (based on the FASTA file downloaded from PlantTFDB FTP site), and 1,805 of them (95.5%) have hits found from SoyDB based on a PSI-BLAST search with an e-value threshold 10^{-3} .

5.4.10 Comparisons of Soybean Transcription Factor Family Distributions with Other Plants

The collection and analyses in SoyDB allows us to perform comparisons of TF family distributions across the plant kingdom. The large number of soybean TF genes (5,671) described in this study is likely due to the two soybean whole genome duplication events; one estimated to have occurred 40-50 million years ago (mya) and the most recent one approximately 10-15 million years ago [124, 125]. By comparing the total number of genes in different organisms, it was found that the increase of plant gene number is related to multicellularity and ploidy. For example, compared to the unicellular eukaryote *Chlamydomonas reinhardtii* where 15,143 genes were predicted [126], larger numbers of protein-encoding genes were reported in multicellular plant organisms, e.g. *Physcomitrella patens* (35,938; [127]), *Arabidopsis thaliana* (32,944; TAIR [92]), and tetraploid *Glycine max* (66,153, Phytozome). We hypothesize that TF gene number also follows the same trend as land plants, which have a larger number of TF genes compared to algae. To perform comparisons of plant TF genes and their distributions across TF gene families, we mined the last updated DBD database [79] for 11 plant species (*C. reinhardtii*, *P. patens*, *Oryza sativa*, *Zea mays*, *Sorghum bicolor*, *Lotus japonicum*, *Medicago truncatula*, *A. thaliana*, *Vitis vinifera*, *Ricinus communis*, and *Populus trichocarpa*). These species were then compared with the soybean TF genes stored in our SoyDB database.

Our analysis showed that *unicellular C. reinhardtii* has the lowest number of TF genes compared to multicellular land plants (the exceptions are *L. japonicus* and *M. truncatula* where only a partial genome sequence is available). This trend also reflects the differences of total gene number in the organisms shown in Figure 5.6. For example, it is interesting to note that homeobox, MYB, NAC, and WRKY TF genes in *C. reinhardtii* lack or have very low representations compared to the 11 other plant models (Table 1). Previous studies defined a role for homeobox [128] and

WRKY genes [74] in plant development. Therefore, the occurrence of these genes only in multicellular plants may reflect their special roles in development. In addition, a close relationship between TF gene number and total gene number [129] is observed when comparing the TF gene numbers of *G. max* and *A. thaliana* with their total gene numbers (i.e. *G. max* encodes 66,153 protein-coding genes including 5,683 TF genes; *A. thaliana* encodes 32,944 protein-coding genes and 1,738 TF genes). Thus, the family distribution of soybean TF genes is similar to other land plant species, except for *P. patens* (e.g. AP2 represents 7% of total TF genes in soybean vs. 8-12% for other land plants; bZIP: 3% vs. 3-7%; bHLH: 7% vs. 8-11%; homeobox: 6% vs. 4-7%; MYB: 14% vs. 7-14%; NAC: 4% vs. 4-9%; WRKY: 3% vs. 4-7%; ZF-C2H2: 7% vs. 5-9%) (Figure 5.6 and Table 1).

Collectively, these results suggest that soybean TF genes were not lost following soybean genome duplication, and may have evolved for specialized functions in plant development or response to the environment.

5.5 Future Development Plan

In the future, we plan to link to more soybean database, such as SoyBase, and add a human expert discussion section for each transcription factor where biologists can register, log in, and make comments on any annotation items. Also, we plan to link the protein name, such as Glyma01g11670.1, listed in each protein information page to its entry in Phytozome. By doing this, SoyDB can be linked with other soybean genome annotations. Furthermore, we may identify the binding regions on the soybean DNA sequences, which can further help biologists target the regulated regions on soybean genome.

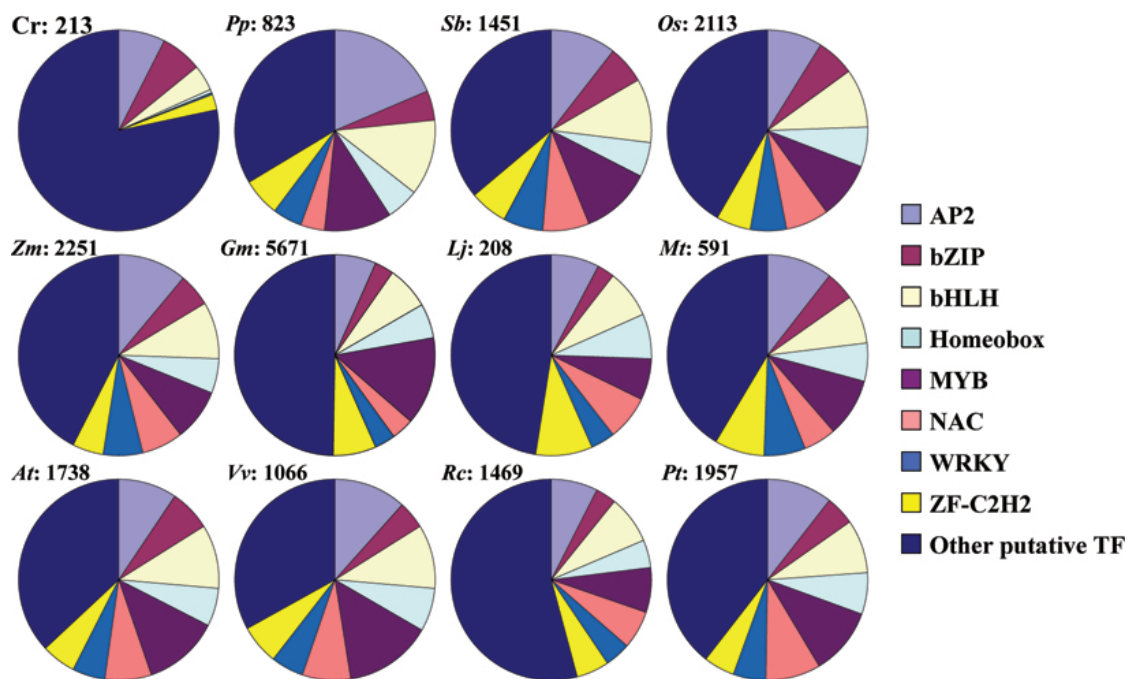


Figure 5.6: Distributions of transcription factor families across major plant species. Phytozome and DBD databases were mined to identify transcription factor genes in soybean (Gm: *Glycine max*) and in the 11 remaining plant species, respectively (Cr: *Chlamydomonas reinhardtii*; Pp: *Physcomitrella patens*; Sb: *Sorghum bicolor*; Os: *Oryza sativa*; Zm: *Zea mays*; Lj: *Lotus japonicus*; Mt: *Medicago truncatula*; At: *Arabidopsis thaliana*; Vv: *Vitis vinifera*; Rc: *Ricinus communis*; Pt: *Populus trichocarpa*). After being classified based on their family membership, nine major TF families are represented for each plant species. Numbers next to the plant name abbreviation are the total number of TF genes available in DBD. Details are available in Table 5.1.

5.6 Conclusions

SoyDB is a comprehensive database for soybean transcription factors. It integrates bioinformatics tools and various external databases to provide rich annotations, which can be browsed and retrieved through convenient web interfaces. The automated process generates annotations and creates database and website, and can be used to annotate other sequenced species.

5.7 Availability and Requirements

SoyDB is freely available at <http://casp.rnet.missouri.edu/soydb/> for academic use. Based on our test, SoyDB is fully functional with three web browsers: Mozilla Firefox, Internet Explorer, and Safari, and four operating systems: Windows XP, Windows Vista, Linux (Red Hat), and Mac OS. The only system requirement for SoyDB is that JAVA runtime environment (JRE) needs to be installed and set fully functional in order to make Jmol work.

Table 5.1: Distributions of transcription factor families across major plant species.

	At	Zm	Os	Gm	Lj	Mt	Sb	Pt	Pp	Cr	Vv	Rc
AP2	162	251	186	381	16	63	153	207	153	16	124	111
bZIP	116	119	130	176	6	27	89	90	38	14	48	50
bHLH	183	207	203	393	16	47	148	172	102	9	110	112
homeobox	105	121	132	319	15	35	81	129	44	1	74	66
MYB	212	192	193	791	14	56	165	210	89	0	151	105
NAC/NAM	132	149	146	208	15	31	111	174	32	0	81	92
WRKY	89	141	123	197	8	39	92	103	37	1	59	58
ZF-C2H2	98	114	117	395	19	49	88	101	51	5	67	78
Other TF	641	957	883	2823	99	244	524	771	277	167	352	797
Total	1738	2251	2113	5671	208	591	1451	1957	823	213	1066	1469

%	At	Zm	Os	Gm	Lj	Mt	Sb	Pt	Pp	Cr	Vv	Rc
AP2	9%	11%	9%	7%	8%	11%	11%	11%	19%	8%	12%	8%
bZIP	7%	5%	6%	3%	3%	5%	6%	5%	5%	7%	5%	3%
bHLH	11%	9%	10%	7%	8%	8%	10%	9%	12%	4%	10%	8%
homeobox	6%	5%	6%	6%	7%	6%	6%	7%	5%	0%	7%	4%
MYB	12%	9%	9%	14%	7%	9%	11%	11%	11%	0%	14%	7%
NAC/NAM	8%	7%	7%	4%	7%	5%	8%	9%	4%	0%	8%	6%
WRKY	5%	6%	6%	3%	4%	7%	6%	5%	4%	0%	6%	4%
ZF-C2H2	6%	5%	6%	7%	9%	8%	6%	5%	6%	2%	6%	5%
Other TF	37%	43%	42%	50%	48%	41%	36%	39%	34%	78%	33%	54%
hline Total	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

Chapter 6

Three-Level Prediction of Protein Function by Integrating Profile-Sequence Search, Profile-Profile Search, and Domain Co-Occurrence Networks

6.1 Abstract

Predicting protein function from sequence is useful for biochemical experiment design, mutagenesis analysis, protein engineering, protein design, biological pathway analysis, drug design, disease diagnosis, and genome annotation as a vast number of protein sequences with unknown function are routinely being generated by DNA, RNA and protein sequencing in the genomic era. However, despite significant progresses in the last several years, the accuracy of protein function prediction still needs to be improved in order to be used effectively in practice, particularly when little or no homology exists between a target protein and proteins with annotated function. Here, we developed a method that integrated profile-sequence alignment, profile-profile align-

ment, and Domain Co-Occurrence Networks (DCN) to predict protein function at different levels of complexity, ranging from obvious homology, to remote homology, to no homology. We tested the method blindly in the 2011 Critical Assessment of Function Annotation (CAFA). Our experiments demonstrated that our three-level prediction method effectively increased the recall of function prediction while maintaining a reasonable precision. Particularly, our method can predict function terms defined by the Gene Ontology more accurately than three standard baseline methods in most situations, handle multi-domain proteins naturally, and make ab initio function prediction when no homology exists. These results show that our approach can combine complementary strengths of most widely used BLAST-based function prediction methods, rarely used in function prediction but more sensitive profile-profile comparison-based homology detection methods, and non-homology-based domain co-occurrence networks, to effectively extend the power of function prediction from high homology, to low homology, to no homology (ab initio cases).

6.2 Introduction

In the genome era, high-throughput genome, transcriptome, and proteome sequencing is generating an enormous amount of omics data such as gene and protein sequences. Since experimental characterization of these proteins can only be carried out at a selectively small scale, large-scale and high-throughput computational prediction methods are needed to annotate the structure and function of most of these proteins in order for the biomedical research to effectively utilize this vast resource to study genotype - phenotype relationships. To fill the gap, a variety of computational methods have been developed to predict protein function from protein sequence from different perspectives.

The most commonly used approach to function prediction is based on sequence

homology. It uses a sequence comparison / alignment tool to search a target protein sequence against protein sequences with known function annotations in a protein database, and if some homologous hits are found, their function annotations may be transferred to the target protein as predictions. GOtcha [130], OntoBlast [131], and Goblet [132] are tools that use BLAST [133] to search for homologues and then combine the Gene Ontology function terms [12] of homologous hits based on BLAST e-values. PFP [134] uses a more sensitive profile-sequence alignment tool PSI-BLAST [133] to search for remote homologues, and also considers co-occurrence relationships between GO [12] terms in order to improve the sensitivity of prediction.

Phylogenetic relationships between proteins have been proven to be helpful for inferring protein functions [135, 136, 137]. Paralogues and orthologues, the two kinds of homologous proteins generated by gene duplication and speciation during evolution, respectively [138], may still share similar functions. Thus, the function of a protein may be inferred from that of its paralogues or orthologues, even though the level of their functional similarity may depend on their evolutionary distance and other factors. As most of phylogenetic-tree based methods assume orthologous proteins are more likely to share similar functions [137], they often generate a phylogenetic tree to elucidate the evolutionary relationships between a target protein and its homologous proteins at first, and then preferably use the functions of its orthologues to infer its function. SIFTER [139], Orthostrapper [140], RIO [141], and AFAWE [142] are typical methods in this category.

Apart from homologous relationships mentioned above, network-based methods exploit other relationships stored in protein networks. Assuming that neighboring proteins interacting in a protein-protein interaction (PPI) network may have similar protein function, some early network-based methods use the functions of the direct (radius-one) neighbors of a target protein in a PPI to infer its function. More advances in this direction include the consideration of statistically enriched functions within

neighbors [143, 33], the expansion of search from direct neighbors to radius-two and radius-three neighbors [143], and the development of more advanced function inference methods, such as Markov random field [144], random walk [145], and algorithms taking in account the global topology of a network [146, 147, 148]. In addition to PPI, Functional Linkage Networks (FLNs) [149] derived from protein interaction, gene expression data, phylogenetic profile, and genetic interaction [148, 150, 151, 152], have been used to predict protein function. More recently, Domain Co-occurrence Networks (DCN) has been used to predict protein function [33].

In an effort to directly link a protein with its function, machine learning methods, such as Support Vector Machines (SVM) and Artificial Neural Networks, have been developed to predict protein function from scratch. Machine learning methods usually generate features from protein sequence, secondary structure, hydrophobicity, subcellular location, and solvent accessibility, and then use these features as inputs to train a classifier to assign proteins to a number of predefined function categories. ProtFun [153] aims to classify a eukaryotic protein into 14 Gene Ontology (GO) categories and several Enzyme Commission classes. FFPred [154] uses features derived from protein disordered regions and protein sequence profiles with SVM to classify a protein into 300 Gene Ontology classes. With these approaches developed from a variety of perspectives, protein function prediction remains a challenging, largely unsolved problem, particularly when little information regarding homology and protein interaction is known about a target protein. Both the specificity and sensitivity of function prediction need to be improved in order to reliably make function predictions for most proteins. One consensus in the community is to combine multiple complementary methods and explore and integrate more diverse sources of information to enhance prediction accuracy and broaden the annotation scope [155, 156]. In this spirit, we developed a three-level method to cope with the complexity of function prediction at different levels, from high homology, to remote homology, to no homology, and

synergistically integrated them into a system that can make function prediction for almost all the target proteins in the 2011 Critical Assessment of Function Annotation (CAFA, <http://biofunctionprediction.org/>). At the first level, our method uses PSI-BLAST to search SwissProt [91] for significant homologues of a target protein; at the second level, it applies a sensitive profile-profile alignment tool HHSearch to search against Pfam [86] to gather remote homologues; and at the third level, it detects domains existent in the target protein, and then uses their neighboring domains found in a species-specific Domain Co-occurrence Networks (DCNs) to infer the functions of the target protein, even though there may be no homology between the target protein and its neighboring domains. Our method combining predictions generated at all three levels participated in the 2011 CAFA experiment. In comparison with three base-line methods, our method not only substantially expanded the sensitivity / recall of function prediction by adding profile search and domain network at the top of traditional PSI-BLAST search, but increased the semantic similarity between predicted function terms and true ones according to the Gene Ontology. Another advantage is that our method can readily predict the functions of multi-domain proteins by decomposing a protein into individual domains and aggregating the function predictions of each domain on a domain co-occurrence network.

6.3 Methods

6.3.1 Overview

We constructed three protein function predictors (predictor-1, predictor-2, predictor-3), whose predictions were submitted to CAFA as models 1, 2, and 3. The first two predictors combined function predictions derived from profile-sequence PSI-BLAST search, profile-profile HHSearch search, and domain co-occurrence networks using

different strategies. The third one used only predictions derived from PSI-BLAST search at the default threshold (i.e. 10).

6.3.2 Predictor-1

The threelevel method integrates profile-sequence alignment, profileprofile alignment, and Domain Co-occurrence Networks (DCNs) as shown in Figure 6.1. At the first level, PSI-BLAST was executed to search against Swiss-Prot [91].

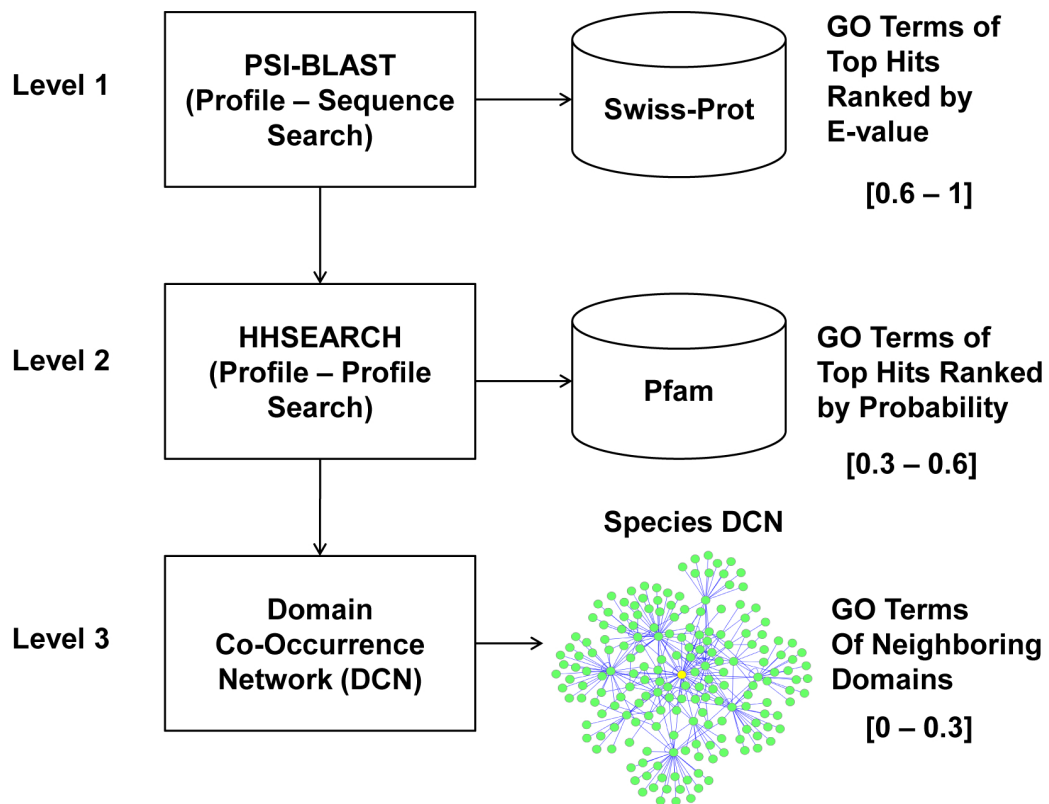


Figure 6.1: The architecture of the three-level prediction methodology used in our predictor-1.

The protein hits with $evalue \leq 0.01$ were chosen and ranked by $evalue$. Only the GO terms of the top one hit were included as predictions, whose confidence score S was calculated as:

$$S_{PSIBLAST} = 0.6 + 0.4 \times \frac{-\log_{10}(e)}{200}, S_{PSIBLAST} \in [0.6, 1.0]$$

, where e stands for evalue of the hit assigned by PSIBLAST. An upper limit of $S_{PSIBLAST}$ was set to 1.0. For example, when e-value equals to 0, $S_{PSIBLAST}$ is set to 1.0. So $S_{PSIBLAST}$ is in the range [0.6, 1]. The second level applied a profile-profile alignment tool HHSearch [157] to detect domains of a target protein. HHSearch generated a hidden Markov model (HMM) for a target protein, which was aligned with the HMM of each Pfam domain family, resulting in a probability score in the range of [0, 100] for each hit. Only the hits with probability score ≥ 80 were kept, and their GO terms were retrieved from the PfamA database as predictions. The confidence scores of the predicted GO terms were assigned as:

$$S_{HHSearch} = 0.3 + 0.3 \times \frac{P}{100}$$

, where p is the HHSearch probability score from 0 to 100. Thus, the confidence score of GO terms predicted from HHSearch hits is in the range [0.3, 0.6].

The target proteins without predictions made from the first two levels were considered hard cases. For hard cases, we used PSI-BLAST with the default threshold (i.e. 10) to search against both Swiss-Prot and the Gene Ontology database, and additionally applied the DCNbased “aggregated neighbor-counting” method [33] to make predictions. The DCN-based “aggregated neighbor-counting” method ran PSI-BLAST to search a target protein against the pre-built proteome databases to find its most closely related organism. Our database contains the wholegenome protein sequences of *H. sapiens*, *S. cerevisiae*, *C. elegans*, *D. melanogaster*, 15 plant species, and 398 single-chromosome prokaryotic organisms (detailed species names can be found at [33]). The organism whose protein was most similar to the target protein according to the PSI-BLAST search’s e-value was considered the most closely related

species for the target protein. The pre-constructed DCN of this species was used to make function predictions for the target. For the domains of the target protein detected by HHSearch at level two, the predictor gathered the GO terms of their radius-one neighboring domains in this DCN as predictions, whose confidence scores was proportional to their occurrence frequencies. If no GO terms could be found in radius-one neighboring domains, it extended to the search to radius-two neighboring domains and made predictions according to the same procedure. The confidence of a predicted GO term was calculated as

$$S_{DCN} = 0.3 \times f$$

, where f is the occurrence-frequency of the GO term - the number of the neighboring domains that have the GO term function divided by the total number of occurrences of all predicted GO terms. Thus, the confidence score of DCN-based predictions is in range $(0, 0.3]$. The ranges of confidence scores assigned to three levels were chosen according to a benchmarking on 100 proteins randomly selected from Gene Ontology before making predictions for the CAFA targets. We compared the performances of the predictions at each level and set their ranges of confidence scores based on their prediction accuracies on the benchmark proteins from high to low.

6.3.3 Predictor-2

Predictor-2 used PSI-BLAST to search a target protein against Swiss-Prot with e-value threshold 0.01, applied HHSearch to search the target against PfamA, and employed the DCN-based “aggregated neighbor-counting” method on radius-one neighbors, in order to gather GO terms at all the three levels. The same probability score threshold (≥ 80) of HHSearch was used as in Predictor-1. We assigned weights 4, 2,

and 1 to a GO term generated by PSI-BLAST, HHSearch, and DCN-based “aggregated neighbor-counting” method, respectively. The weighted frequency of each GO term was calculated and normalized. The normalized score was used as the confidence score of an individual GO term. For proteins without any predictions generated using these three methods, an additional PSI-BLAST search of the protein against Gene Ontology and Swiss-Prot with the default e-value threshold (i.e. 10) was executed in order to gather more hits if possible.

6.3.4 Predictor-3

Only a PSI-BLAST search against Swiss-Prot with the default threshold (i.e. 10) was performed in predictor 3. All of the PSI-BLAST hits were included to make prediction. The occurrence frequency of a GO term among all hits was used as its confidence score.

6.4 Results and Discussion

6.4.1 Overview

We blindly tested our method in the 2011 Critical Assessment of Function Annotations (CAFA) experiment. In total, CAFA released 48,298 protein targets whose functions were not known to predictors from all around the world to make prediction from Sept. to Jan., 2011. At the end, 436 of the targets whose functions were later known and deposited into the SwissProt database were used to evaluate the performances of the predictors. In order to gauge the advances in the field, CAFA released the predictions of three baseline methods. The Prior method used the frequency of GO terms in the SwissProt database to select 836 most frequent GO terms for each target as prediction. The BLAST method used BLAST [133] to search a target pro-

tein against groups of proteins, where proteins in each group shared a common GO term; and the maximum sequence identify of BLAST alignments with sequences in a group was used as confidence score to rank GO terms for the target protein. The GOTcha method [130] generated GOTcha I-Scores as the sum of the negative logarithm of the e-values resulted from the BLAST search and used them as confidence scores to select GO terms. During the CAFA experiment, our method submitted three predictions / models for each target, which were produced by three different ways of combining predictions of three levels. Model 1 mapped predictions of three levels into three intervals of confidence from high to low; model 2 weighed predictions of three levels differently from more to less in confidence score calculation; and model 3 simply used the frequency of GO terms of all PSI-BLAST hits as confidence scores to select GO terms. The details of these three models are described in the Method section. Table 1 lists the minimum, maximum, and average number of GO terms predicted by three baseline methods and our three prediction models. The following sub-sections report and discuss the performances of our predictors along with the three baseline predictors using complementary evaluation measures.

6.4.2 Precision and Recall of Top n Predictions

We evaluated these methods using precision and recall of GO terms on top n predictions ranked by prediction confidence scores, for each n in the range [1, 20]. One caveat is that multiple GO terms having the same confidence score received the same average rank. For example, if the confidence scores of top three GO term A, B, C are 0.9, 0.8, 0.8, respectively, the rank of them will be 1, 2.5 (i.e. $(2+3) / 2$), 2.5, respectively. The precision and recall of a protein i were calculated as:

$$Precision_i = \frac{Number_of_correctly_predicted_nodes}{number_of_predicted_nodes}$$

$$Precision_i = \frac{Number_of_correctly_predicted_nodes}{number_of_true_nodes}$$

Here, all of the top n predicted GO terms and the actual GO terms (determined by experimental methods) of protein i were propagated to the root of the Gene Ontology Directed Acyclic Graph (DAG). All the GO term nodes present in the paths of predicted GO terms toward the root were considered predicted GO terms; and all the GO term nodes existing in the paths of the actual GO terms toward the root were considered as true GO terms. The overlapping GO terms / nodes between predicted ones and true ones were considered correctly predicted nodes. For each n in $[1, 20]$, we calculated the precision and recall for each target protein and averaged them over 436 targets protein as the precision and recall on the data set. Thus, for each method, we got 20 precision-recall pairs to generate a precision-recall curve.

Figure 6.2 plots the precision-recall curves of six predictors. The plot shows that our methods tend to have higher recall and lower precision in comparison with the lower recall and higher precision of two baseline methods: Prior and GOtcha. For instance, our best performing model 1 can reach a recall value as high as ~ 0.55 , whereas the highest recall value of the baseline methods (i.e. Prior and GOtcha) is at ~ 0.27 . That the baseline methods, particularly the Prior method that selects most frequent GO terms in a general protein function database, can have a high precision but a low recall may be because these methods tend to predict more general GO terms in top n predictions that are closer to the root node of the Gene Ontology, but farther away from the most specific true GO terms. To illustrate this point using an example, we calculated the average depth (number of nodes from a GO term to the root node) of predicted GO terms of the target protein T07719 for the six predictors when recall is at ~ 0.1 . The average depth of our model 1, 2, and 3, and the Prior, BLAST, and GOtcha is 16.3, 16.3, 18.4, 13.9, 16.1, and 4.8, respectively, which shows that our models tried to predict deeper GO terms than the Prior and GOtcha. Although the

precision-recall curves of our methods and the three baseline methods largely occupy two different areas in the plot, when their recalls overlap within the range (~ 0.22 , ~ 0.27), our models have higher precisions than the Prior and GOtcha methods at the same recalls. The BLAST baseline method can only have a maximum recall at ~ 0.18 and a maximum precision at 0.31 , which clearly performed worse than our three methods including our least accurate model 3 based on PSI-BLAST search alone. This suggests that PSI-BLAST might work better than BLAST for protein function prediction, even though other factors such as how to rank GO terms based on alignments cannot be ruled out either.

It is also interesting to notice that the best recall value of our model 1 (~ 0.55) is higher than that of model 2 (~ 0.51), indicating that including radius-two domain neighbors in Domain Co-occurrence Network [33] may contribute to the increase of recall since model 1 used both radius-one and radius-two domain neighbors to make predictions whereas model 2 only used radius-one neighbors (see the Method section for details). That the recall of models 1 and 2 are much higher than that of model 3 (~ 0.41) demonstrates that profile-profile alignment (HHSearch [157]) and DCN can substantially increase the sensitivity of protein function prediction at the top of the profile-sequence search methods such as PSI-BLAST (e.g. model 3).

6.4.3 Precision and Recall Under a Sliding Threshold on Confidence Scores

We also calculated precisions and recalls according to a sliding threshold scheme, in which only the predictions with confidence scores higher than a threshold value t ($0 \leq t \leq 1$) were selected for evaluation. The predicted and actual GO terms were propagated to the root in the GO DAG as described in the sub-section above. At each threshold, we calculated precision and recall for each protein and used the average precision and recall on all 436 target proteins as the estimated prediction and recall

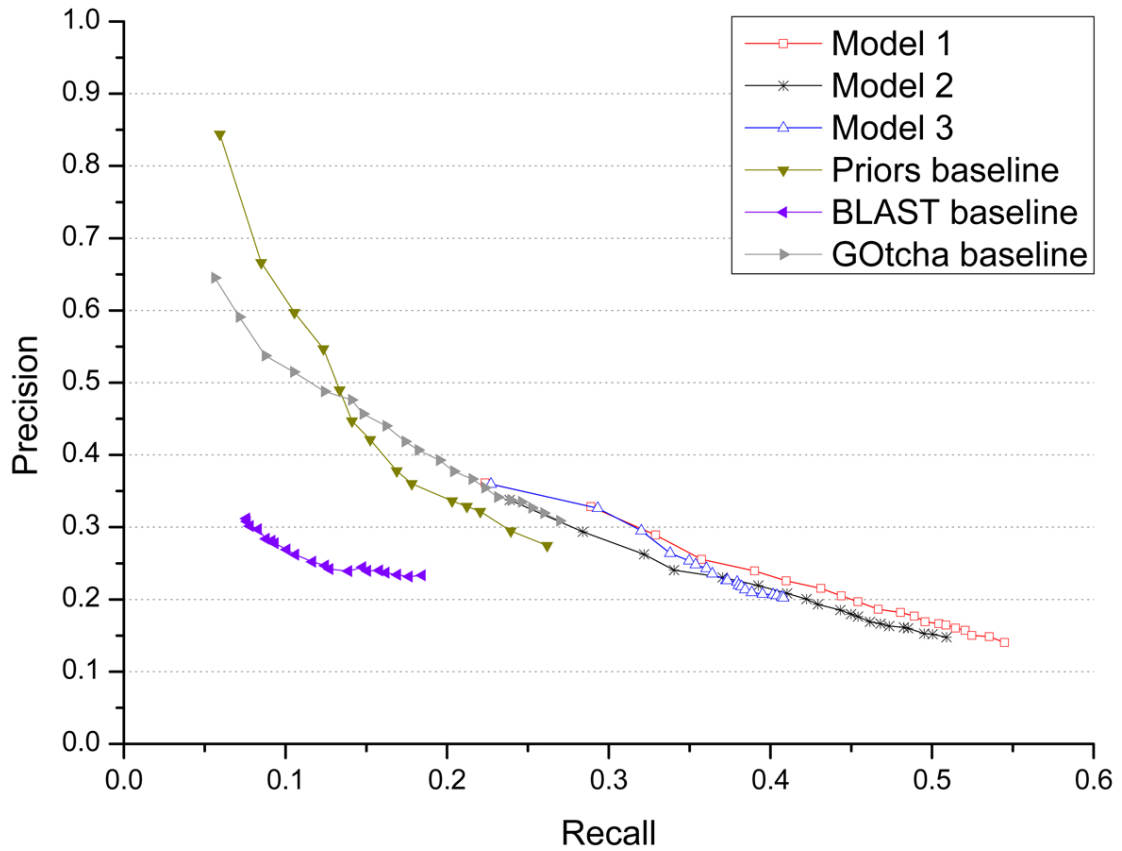


Figure 6.2: Precision and recall of our three models and three baseline methods when considering top n , $1 \leq n \leq 20$, predictions ranked by confidence scores.

for the threshold. By using the thresholds evenly distributed in the range $[0, 1]$ at step size 0.01, we calculated a series of precision-recall pairs for each of six predictors.

Figure 6.3 shows the precision-recall curves of our three models and three baseline methods. Because all the predictions with a confidence score above a threshold other than just top 20 predictions are selected for evaluation, all the predictors in Figure 6.3 can reach a higher recall than those in Figure 6.2. Particularly, the Prior method can yield the highest recall (0.82) among all predictors because it constantly predicts 836 GO terms for each target protein, which is several times more than all other predictors (Table 6.1). Thus, 0.82 could be considered an upper limit on recall that current methods can achieve. The other two baseline methods (BLAST and GOtcha) that predict more than twice as many GO terms as our methods have the maximum recall values 0.55 and 0.5 respectively, which are lower than the ones of our three models. It is worth noting that our model 1 that predicted ~ 73 GO terms on average delivered the second highest recall ~ 0.68 . When recall is higher than ~ 0.31 , our three models have higher precision than all three baseline methods at the same recall, while model 1 performed mostly better than or occasionally comparable to models 2 and 3. That model 1 mostly performed better than model 2 suggests that the sequential combination of three levels of predictions generated by PSI-BLAST, HHSearch and DCN is more effective than the weighted combination of these predictions. Another interesting observation is that model 3 that simply pooled all PSI-BLAST hits to generate GO term predictions performed better than the baseline BLAST and GOtcha methods throughout the entire recall range. It also worked better than the baseline Prior method when recall is $> \sim 0.15$, whereas the latter yielded better precision than all other methods when the recall is low ($< \sim 0.15$). The better performance of the Prior method in the low recall range may be explained by that its highly common GO term predictions were largely correct, but too far away from the specific GO function of target proteins. Thus its highest precision (~ 0.85) may serve as an upper

limit that current function prediction methods can aim to achieve. Table 6.2 shows the break-even values of our models and the three baseline methods when precision equals to recall. According to this criteria, our methods performed better than the baseline methods, while model 1 yielded the highest break-even value 0.306.

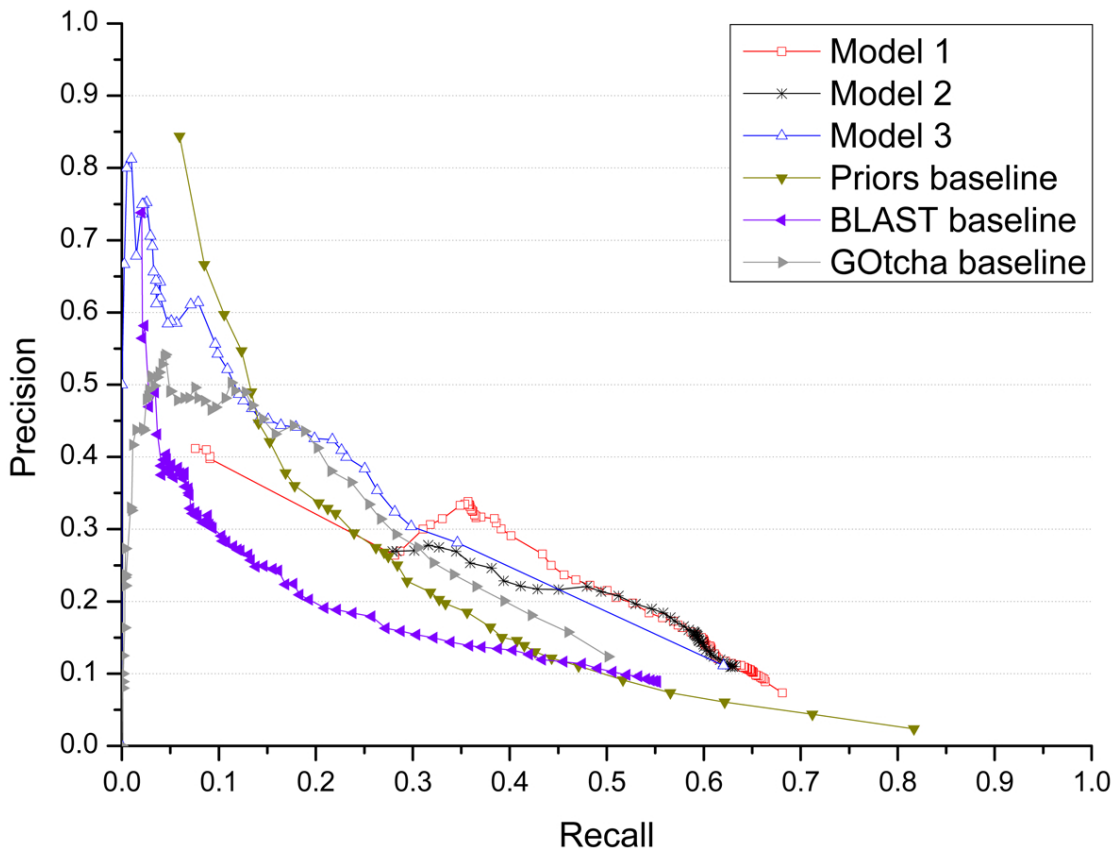


Figure 6.3: Precision and recall of our three models and three baseline methods when considering predictions with confidence score above a threshold t , $0 \leq t \leq 1$. A number of threshold values evenly distributed in the range $[0, 1]$ at step size 0.01 were used to calculate precisions and recalls.

In order to further analyze the amount of contributions made by profile-sequence alignment (PSI-BLAST), profile-profile alignment (HHSearch), and domain co-occurrence networks (DCN), we plotted a precision-recall curve of model 1 in Figure 6.4 to show how precision and recall changes, when progressively considering predictions resulted from PSI-BLAST search at level 1, from both PSI-BLAST and HHSearch searches at levels 1 and 2, and from all three methods at levels 1, 2 and 3. Figure 6.4 shows

Table 6.1: The minimum, maximum, and average number of predictions per target of our three models and the three baseline methods.

	Minimum	Maximum	Average
Model 1	1	100	73.3
Model 2	1	100	56.3
Model 3	1	100	57.8
Priors baseline	836	836	836.0
BLAST baseline	1	945	254.1
GOTcha baseline	2	519	135.7

Table 6.2: The break-even values between precision and recall (i.e., when precision = recall) of the six predictors. Average values are the averages of precisions and recalls at decision thresholds yielding the closest precision and recall values.

	Threshold	Precision	Recall	Average
Model 1	0.90	0.300	0.311	0.306
Model 2	0.81	0.269	0.279	0.274
Model 3	0.02	0.304	0.298	0.301
Priors baseline	0.20	0.268	0.270	0.269
BLAST baseline	0.46	0.202	0.193	0.198
GOTcha baseline	0.08	0.293	0.283	0.288

that the profile-profile alignment (HHSearch) extended the recall of profile-sequence alignment (PSI-BLAST) from 0.57 to 0.64, and DCN further increased the recall to 0.69. The results demonstrate that three levels of predictions are complementary and can be combined effectively to increase the sensitivity of protein function prediction. Particularly, the DCN method may contribute valuable function predictions when all homology search methods fail to find useful hits, even though the prediction precision in this ab initio situation may be low.

6.4.4 Evaluations by Semantic Similarity

In addition to precision and recall measures based on the exact match of GO terms, we evaluated these predictors in terms of semantic similarity between true GO terms and predicted GO terms. For two GO terms g_1 and g_2 , we obtained their paths r_1 and r_2 to the root of the GO DAG, and calculated the similarity score between g_1

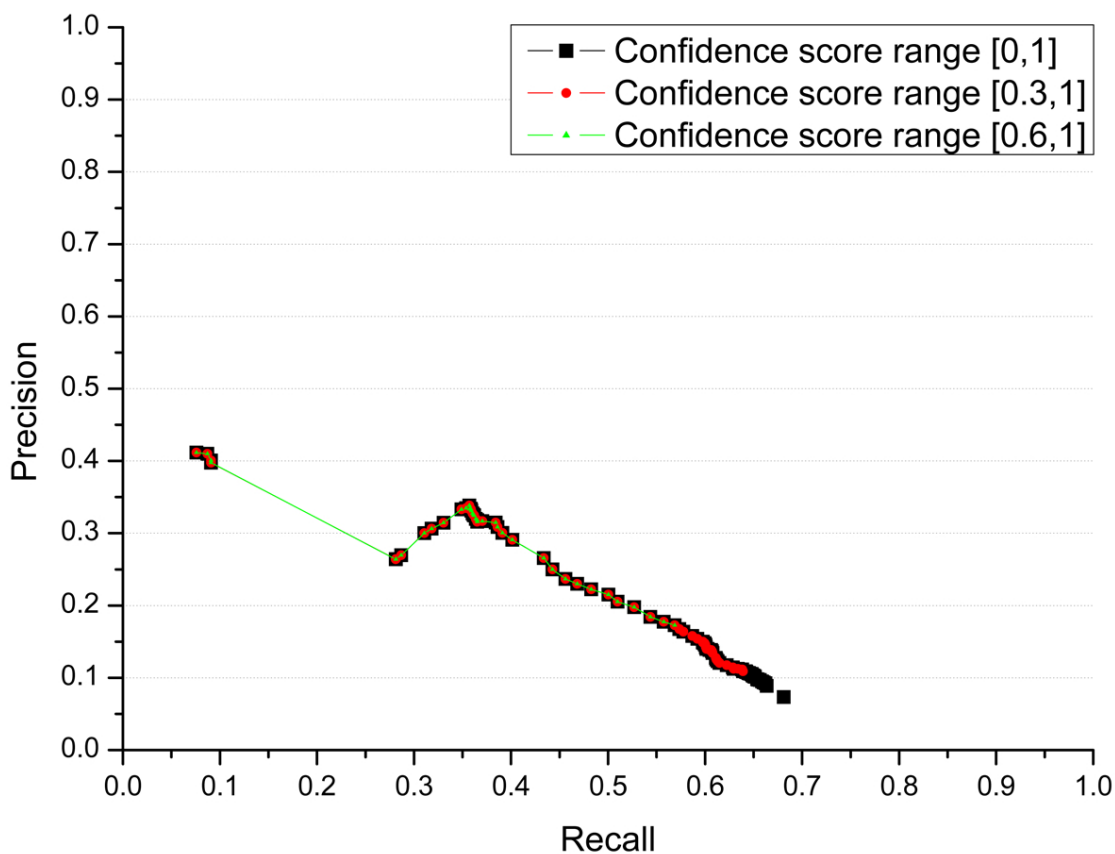


Figure 6.4: Precision and recall when progressively considering predictions with confidence score in ranges $[0, 1]$, $[0.3, 1]$, and $[0.6, 1]$ for our model 1 predictions. The predictions in the three ranges were predicted by three different methods: PSI-BLAST search, HHSearch search, and Domain Co-Occurrence Networks. Their precision and recall curves were drawn in three different colors, showing the higher level of prediction gradually increased recall at the expense of lower precision.

and g_2 as:

$$Sim(g_1, g_2) = \frac{\theta(r_1 \cap r_2)}{Max(\theta(r_1), \theta(r_2))}$$

,where $\theta(r_1)$ and $\theta(r_2)$ denotes the number of GO terms of paths r_1 and r_2 , respectively. The numerator is the number of common GO terms shared by paths r_1 and r_2 . For a target protein, every predicted GO term was compared with each of the true GO terms to calculate similarity scores; and the highest score was considered as the similarity score between a specific predicted GO term and the actual GO terms. Averaging the similarities over all predicted GO terms of a target generated the similarity score for the target. The average similarity scores of all the target protein was used as the prediction similarity score of a predictor. We computed the similarity scores of all predictors for top 1, 2, ..., 20 predicted GO terms respectively and plotted them in Figure 6.5. According to this measure, all our three models performed better than three baseline methods. In addition to the average similarity score for each target, we also calculated the best similarity score of a target - the highest similarity score among all GO term predictions for the target and averaged the best similarity scores over all the targets for each predictor. Figure 6.6 shows the best similarity scores of six predictors for top 1-20 predictions, which also demonstrates the better performances of our three models compared with the three baseline methods.

6.4.5 An Example Illustrating the Effectiveness of Domain Co-Occurrence Networks for Protein Function Prediction

We chose an example to illustrate the effectiveness of the DCN function prediction component when both profile-sequence alignment (PSI-BLAST [133]) and profile-profile alignment (HHSearch [157]) cannot make precise predictions. Figure 6.7 shows how functions were predicted by using Domain Co-occurrence Network (DCN) for tar-

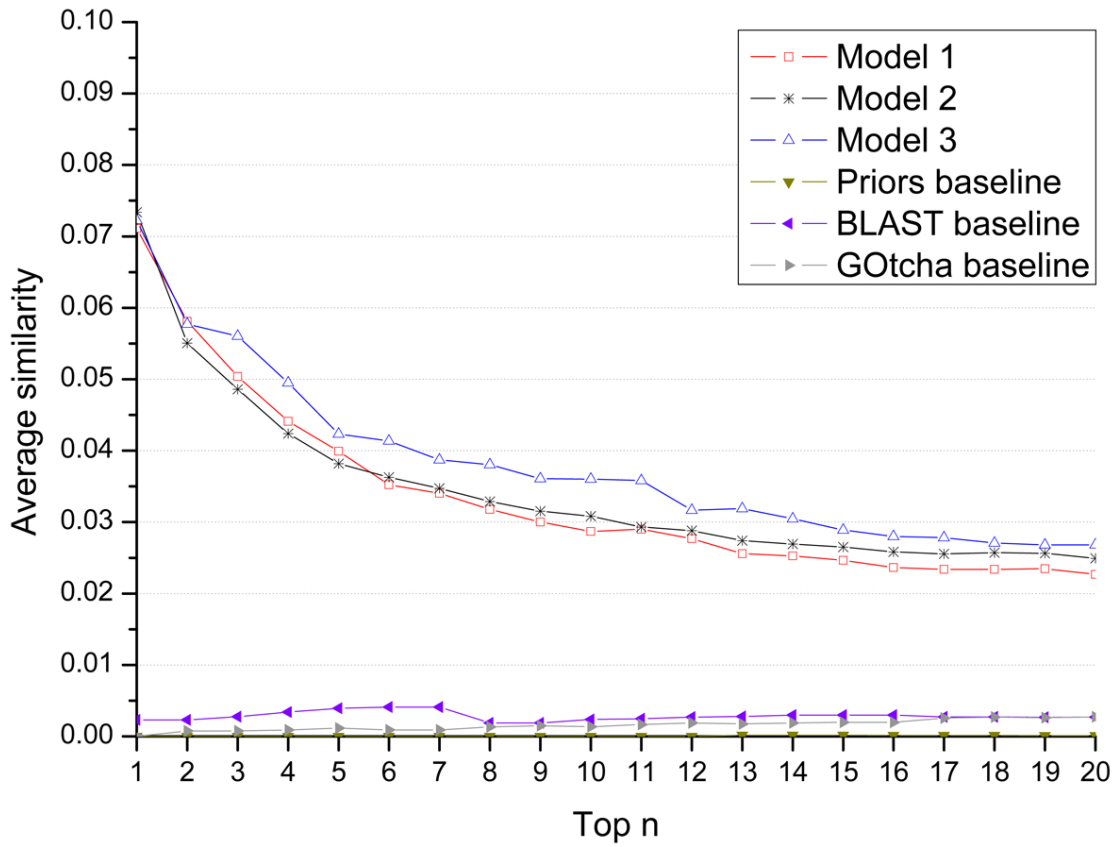


Figure 6.5: Average similarity scores of our three models and three baseline methods for top 1-20 predictions.

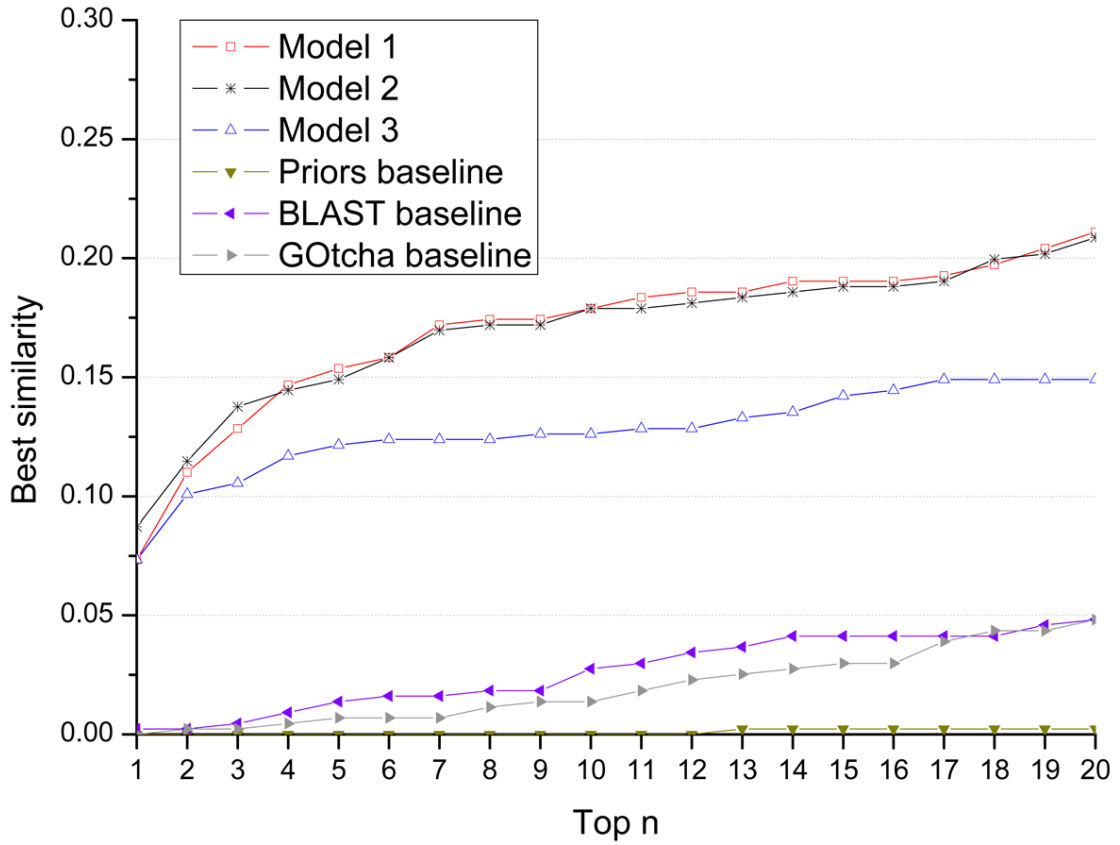


Figure 6.6: Best similarity scores of our three models and three baseline methods for top 1-20 predictions.

get T30248 - a multi-domain protein in *Mus musculus* (house mouse). Figure 6.7 (A) and (B) illustrate the tertiary structure of the protein predicted by MULTICOM [11] and electrostatic potentials (blue: positive charged; red: negative charged), calculated and visualized by DeepView (<http://spdbv.vital-it.ch/>). In order to make function prediction, the DCN method executed PSI-BLAST to search the target protein against our pre-built protein sequence database containing the proteome of *H. sapiens*, *S. cerevisiae*, *C. elegans*, *D. melanogaster*, 15 plants, and 398 bacteria species (detailed organism names can be found at [33]), and identified the most significant homologous hit - a *H. sapiens* protein (Swiss-Prot ID P68543) with PSI-BLAST e-value $3e-95$ and score 348. Then it utilized the DCN of *H. sapiens* (human) (Figure 6.7 (C)) to make prediction as follows. Firstly it used the profile-profile alignment tool HHSearch [157] to search the target protein against PfamA [86] database, which detected eight domain families with homologous probability ≤ 0.80 : SEP, UBX, Spt20, ubiquitin, UN_NPL4, Cobl, Rad60-SLD, and FERM_N. The four domains (SEP, UBX, ubiquitin, and FERM_N) that existed in the *H. sapiens* proteome were then used as central domains in the DCN of *H. sapiens* to identify domain neighbors. Because domains SEP, UBX, FERM_N do not have GO functional annotations in the Pfam database and the ubiquitin domain only has one general GO term annotation (GO:0006464), directly inferring precise function of the target from these domains was not possible. However, the DCN method was able to use the annotated GO terms of the neighboring domains of these four domains to make function prediction for the target as shown in Figure 6.7 (D), where red nodes denotes the central domains detected for the target and yellow nodes represents their radius-one neighboring domains. The GO terms of the neighboring domains were aggregated and ranked based on frequency. The top ranked GO terms were used as predictions. The frequencies were used as the confidence scores of the predicted GO terms. Similarly, radius-two neighboring domains (not illustrated) were applied to generate predictions in a similar way.

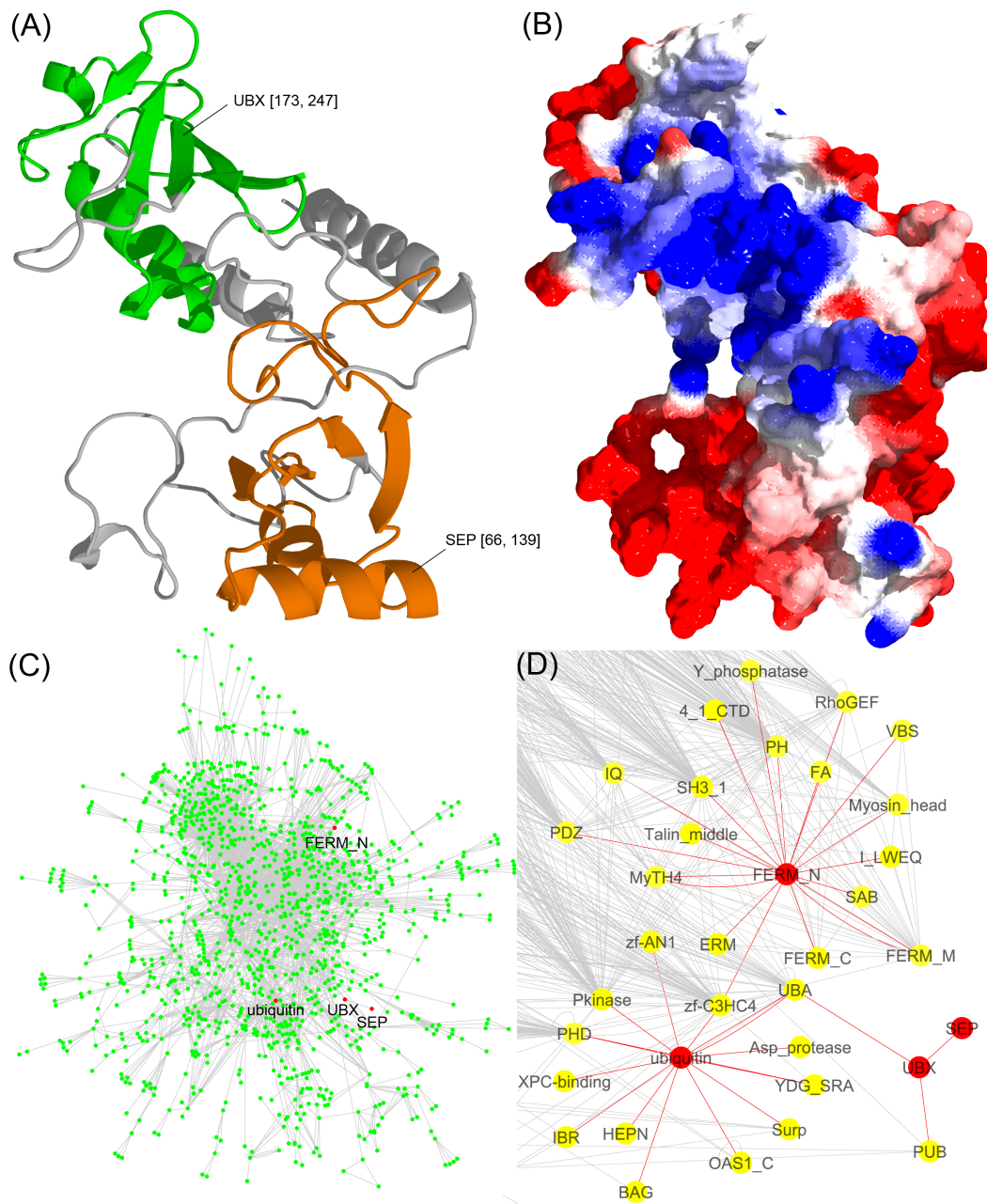


Figure 6.7: An example (CAFA target T30248) showing how DCN-based “aggregated neighbor-counting” method works. (A) Tertiary structure of target protein T30248 predicted by MULTICOM [11] (magenta: domain SEP; green: domain UBX). (B) Electrostatic of protein T30248, generated based on predicted structures in (A) (blue: positive; red: negative). (C) The main Domain Co-occurrence Networks (DCN) of *Homo sapiens* used to make functional prediction. (D) Radius-one neighbor domains of the four domains - ubiquitin, UBX, SEP, and FERM_N - of the target. The GO terms of the neighboring domains were used as predictions for the target. Detailed discussion can be found in “Results and Discussion” section.

The DNC method predicted a GO term GO:0006464 (protein modification process) using radius-one neighboring domains and predicted GO:0043687 (post-translational protein modification), GO: 0051246 (regulation of protein metabolic process) and GO: 0008152 (metabolic process) using radius-two neighboring domains, which were highly related to the two real GO terms of T30248 (Swiss-Prot ID Q99KJ0.1) - GO:0031396 (regulation of protein ubiquitination) and GO:0042176 (regulation of protein catabolic process). Moreover, the above-mentioned predicted GO terms all existed in the propagated paths from the actual GO terms to the root in the Gene Ontology Directed Acyclic Graph (DAG). Particularly, the true GO term GO:0042176 (regulation of protein catabolic process) and the predicted GO term GO:0051246 (regulation of protein metabolic process) had a high semantic similarity score of 0.730 calculated by the tool G-SESAME [158], where 1 indicates exactly the same and 0 completely different. Because the homology-based method did not produce any predictions for T30248, this example demonstrates that the DCN of a species, which may be different from the species of a target protein, can be used to make de novo function prediction for the target from scratch. It also shows that the DCN method can readily decompose a multi-domain protein into multiple domains and aggregate function predictions of individual domains as the prediction for the whole protein.

6.5 Conclusions

We designed and developed an automated three-level method to predict protein functions integrating profile-sequence homology search, profile-profile homology search and domain co-occurrence networks. We blindly tested different ways of combining predictions generated at the three levels on a large number of protein targets in the 2011 Critical Assessment of Function Annotation. The results showed that our methods integrating complementary predictions performed mostly better than three

standard baseline methods. Our experiments also clearly demonstrated that using profile-profile alignment (HHSearch) and domain co-occurrence networks not only increases the sensitivity of protein function prediction at top of the traditional BLAST- and PSI-BLAST-based homology search methods, but also make it possible to make ab initio predictions and handle multi-domain proteins readily.

Chapter 7

A Protein Domain Co-Occurrence Network Approach for Predicting Protein Function and Inferring Species Phylogeny

7.1 Abstract

Protein Domain Co-occurrence Network (DCN) is a biological network that has not been fully-studied. We analyzed the properties of the DCNs of *H. sapiens*, *S. cerevisiae*, *C. elegans*, *D. melanogaster*, and 15 plant genomes. These DCNs have the hallmark features of scale-free networks. We investigated the possibility of using DCNs to predict protein and domain functions. Based on our experiment conducted on 66 randomly selected proteins, the best of top 3 predictions made by our DCN-based aggregated neighbor-counting method achieved a semantic similarity score of 0.81 to the actual Gene Ontology terms of the proteins. Moreover, the top 3 predictions using neighbor-counting, χ^2 , and an SVM-based method achieved an accuracy of 66%, 59%, and 61%, respectively, when used to predict specific Gene Ontology

terms of human target domains. These predictions on average had a semantic similarity score of 0.82, 0.80, and 0.79 to the actual Gene Ontology terms, respectively. We also used DCNs to predict whether a domain is an enzyme domain, and our SVM-based and neighbor-inference method correctly classified 79% and 77% of the target domains, respectively. When using DCNs to classify a target domain into one of the six enzyme classes, we found that, as long as there is one EC number available in the neighboring domains, our SVM-based and neighboring-counting method correctly classified 92.4% and 91.9% of the target domains, respectively. Furthermore, we benchmarked the performance of using DCNs to infer species phylogenies on six different combinations of 398 single-chromosome prokaryotic genomes. The phylogenetic tree of 54 prokaryotic taxa generated by our DCNs-alignment-based method achieved a 93.45% similarity score compared to the Bergey’s taxonomy. In summary, our studies show that genome-wide DCNs contain rich information that can be effectively used to decipher protein function and reveal the evolutionary relationship among species.

7.2 Introduction

Biological systems, such as living cells, are composed of a large number of individual components (e.g., proteins, DNA, RNA, and small molecules). These molecules interact and form networks to carry out biological functions. Departing from the traditional reductionist approach of studying single targets, systems biology aims to identify the cellular molecular components and their interactions and analyze cellular responses at a large scale using high-throughput experimental techniques (e.g., genome sequencing, DNA microarrays, yeast two-hybrid experiments, proteomics, and metabolomics) and computational methods [159, 160, 161, 162, 163, 29, 164]. One promising approach to analyze the complex interactions among molecules is net-

work biology - studying the structure, dynamics, and function of biological networks at the system level [165].

Currently, network biology primarily focuses on metabolic, gene regulatory, and/or protein-protein interaction networks [166, 167, 168, 31, 169, 30, 170, 171, 172, 173, 174, 175]. Since proteins and their interactions play central roles in almost all biological processes, protein interaction networks have been a major target of network biology. Experimental techniques, such as yeast two-hybrid [176] and many computational methods have been developed to construct protein-protein interaction networks. The network approach to the study of protein interactions has shed light not only on the general principles that govern the evolution and functions of the proteins in a species (i.e. proteome) as a whole, but also the function and roles of a particular protein of interest. For instance, protein interaction networks can be used to identify hub proteins having critical biological functions, to predict biological pathways, and to infer the function of a protein according to its interactions with other proteins with known functions [177, 178, 179].

Despite many successful applications, the study of protein interaction networks is hindered by two serious problems [30, 180]. First, protein interaction networks constructed from high-throughput experimental techniques, such as yeast two-hybrid (Y2H), have a high level of false positives [30]. It was estimated that more than half of the protein interactions in an Y2H protein interaction network may be false positives [180]. Protein interactions predicted by computational methods are even noisier. Second, current protein interaction networks constructed for most species are far from complete. For instance, estimates suggest that only about 10% of the protein interactions in the human genome have been elucidated to date [30]. Thus, inferring protein function, interaction, and evolution from protein interaction networks might not be accurate and reliable.

Here, we propose to use protein domain co-occurrence networks (DCN) [181, 182]

to study the function and interaction of proteins at the proteome level. These networks make use of the co-occurrences of various protein domains in given proteins. A protein domain, usually a segment of continuous sequence within a protein, is considered the structural, functional, and evolutionary unit of proteins. One protein is generally composed of one or more domains (i.e. building blocks) that each might fold independently into a stable structure. Each domain often has a distinct biochemical function. Domains that are similar in sequence, structure, and function are grouped together in families/types. The proteome (i.e. the collection of all the proteins) of a species usually has representatives of thousands of domain types. These domains can exist as single-domain proteins or are combined together to form multi-domain proteins. Hence, in addition to sequential divergence, domain combination is another major mechanism of increasing the complexity of a proteome [183]. Nature tends to reuse and recombine existing building blocks to create new proteins, rather than to invent them de novo [184].

Domain combination represents a strong, permanent and definite interaction between domains, which can be captured by domain co-occurrence networks (DCN). A DCN is a graph consisting of all the protein domain types of a species as nodes. Two domain types (i.e. nodes) are connected by an edge if they co-exist in one protein [181]. Figure 7.1 shows the domain architecture (i.e., a series of domain types) of two multi-domain proteins in Arabidopsis and a DCN derived from the two proteins. Previous studies showed that DCNs of the yeast and human were scale-free and small-world networks [181], like other networks such as web hyper-link networks, social networks, and protein interaction networks. However, to date, the features and properties of DCNs have not been well explored, and they have not been used to study the functions and evolution of proteins. Compared to the well-studied protein-protein interaction networks, DCN has the following advantages, making it a very promising tool for studying proteomes at the system level: (1) Accurate and reliable. Domain

co-existence (or combination) relationship constructed from sequential analysis is almost 100% accurate, which is much more reliable than protein interactions predicted from experimental approaches (e.g., yeast two-hybrid). (2) Higher coverage. A DCN constructed from homologous sequence analysis usually can recall about 70% of domain co-existence relationships [185], compared to the very low coverage of protein interaction networks of most species. Thus, the inference based on DCNs is often more reliable. (3) Easy to construct. It is much easier to construct the DCN of a genome by comparing its protein sequences against known protein domain databases, such as Pfam [186] and ProDom [187], compared to building protein interaction networks through either experimental or other computational methods.

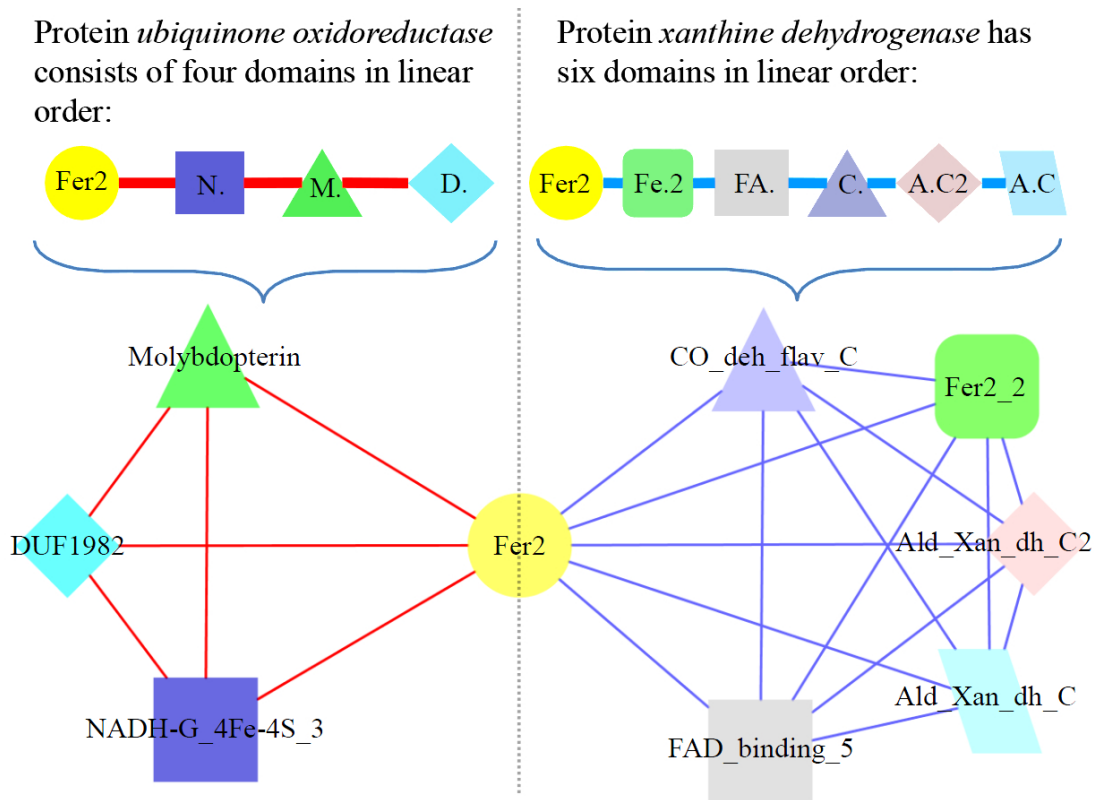


Figure 7.1: A small DCN consisting of two Arabidopsis proteins. Left: Protein ubiquinone oxidoreductase, which contains four Pfam domains. Right: Protein xanthine dehydrogenase, which has six Pfam domains. The two proteins share the same domain Fer2. An edge is drawn between domains co-occurring in the same protein. A fully connected sub-graph (clique) in the DCN corresponds to a protein.

In addition, DCNs also have two other distinct features. First, each node represents a domain type instead of an instance (i.e. sequence). Since DCNs of differing species share a large number of common domain types, DCNs can be more readily compared or aligned across species than sequence-based networks, such as protein-protein interaction networks. Second, an edge in a DCN represents a permanent combination relationship that is stronger than the relationship defined in protein-protein interaction networks, which is often intangible. Therefore, DCNs can provide richer information about the functional relationship between connected domains. Therefore, DCN is a very valuable target for biological network research and a useful tool for studying the evolution, function, and interaction of proteins.

In this study, we constructed the DCNs for *H. sapiens*, *S. cerevisiae*, *C. elegans*, *D. melanogaster*, and 15 plant genomes and performed statistical analysis and attack simulations. A “domain” used in our study is a Pfam entry, corresponding to a Pfam domain “family”, such as “Pkinase_Tyr (PF07714)” and “Helicase_C (PF00271)” [86]. The Pfam domains are often more like protein function units than structural ones. We utilized DCNs to predict domain functions, including GO terms or enzyme classes, and inferred prokaryotic species phylogenies for the first time. Our large-scale studies of DCNs on this diverse set of species demonstrate that DCNs can be readily and reliably constructed from a genome or a list of proteins of an organism, and, in conjunction with graph neighboring methods, can be used to effectively predict protein functions and accurately infer species phylogenies.

7.3 Materials and Methods

7.3.1 Construction of Domain Co-occurrence Networks

The only input data needed for constructing a domain co-occurrence network (DCN) is the whole genome protein sequences of the target organism. In order to have a broad coverage of various species with known genome sequences and cover several model species for our experiments, we construct DCNs for *Homo sapiens* (human, downloaded from NCBI: human genome resources ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/), *Saccharomyces cerevisiae* (yeast, downloaded from SGD [188]), *Caenorhabditis elegans* (downloaded from <http://www.uniprot.org/uniprot/?query=organism:6239+keyword:181>), *Drosophila melanogaster* (fruit fly, downloaded from <http://www.uniprot.org/uniprot/?query=organism:7227+keyword:181>), 15 plants species, and 398 single-chromosome prokaryotic species (downloaded from NCBI: microbial complete genomes taxonomy <ftp://ftp.ncbi.nih.gov/genomes/Bacteria/>). The 15 plant species include: *Chlamydomonas reinhardtii* (green alga), *Ostreococcus lucimarinus*, *Ostreococcus tauri*, *Ostreococcus RCC809*, *Chlorella vulgaris*, *Volvox carteri*, *Physcomitrella patens* (moss), *Selaginella moellendorffii* (gemmiferous spike-moss), *Oryza sativa* (rice), *Zea mays* (maize), *Sorghum bicolor* (sorghum), *Vitis vinifera* (grape), *Arabidopsis thaliana*, *Populus trichocarpa* (black cottonwood), and *Glycine max* (soybean). The protein sequences of *Arabidopsis thaliana* were downloaded from TAIR8 [92] (version 8), *Oryza sativa* (rice) from the “TIGR rice genome annotation resource” [189], *Zea mays* (maize) from the “MAGI website” (<http://magi.plantgenomics.iastate.edu/>), *Vitis vinifera* (grape) from “Genoscope” (<http://www.genoscope.cns.fr/externe/GenomeBrowser/Vitis/>), *Glycine max* (soybean) from “phytozome” website (<http://www.phytozome.net/soybean>), and the others from the Joint Genome Institute (JGI) website (<http://www.jgi.doe.gov/>). These 15 plants species represent virtually all major evolutionary stages of plants,

including alga, primitive land plants, and higher plants.

The program PfamScan is downloaded from the Pfam [86] FTP site (<ftp://ftp.sanger.ac.uk/pub/databases/Pfam/Tools/>), together with Pfam databases and hidden Markov model (HMM) libraries. This domain searching tool is locally installed, which incorporates HMMER (<http://hmmer.janelia.org/>) and BLAST to search against Pfam domain libraries. Each protein sequence of the target species was searched against Pfam using PfamScan. The domain hits with an e-value ≤ 0.01 were kept. When constructing a DCN, each domain is considered as a node, or vertices, in the undirected graph; and every two domains (i.e. nodes) are connected by an edge if they co-exist in one protein [181] as shown in Figure 7.1. Figure 7.2 illustrates the main DCN of *Arabidopsis thaliana* proteome visualized by Cytoscape [2].

7.3.2 Domain function prediction - GO terms

Because domains involved in the same biological process are more likely to co-occur in one protein, domains with similar functions tend to cluster in DCNs. Figure 7.3 shows a densely connected sub-graph identified in the DCN of *Arabidopsis*. This sub-graph consists of 10 domains, which form seven different proteins in *Arabidopsis*. The domains of each protein are circled by red dotted-line eclipse. Not surprisingly, all the proteins are identified to participate in RNA synthesis processes according to the Pfam annotations. Thus, the function of one protein (e.g. spb1_C+FtsJ) can be inferred from another protein (e.g. KOW+Supt5) that is connected through a path in the sub-graph and the central node, e.g. domain S4.

In order to demonstrate the potential of DCNs to infer protein function, we first focused on predicting the specific GO terms [12] of a domain from its neighboring nodes. Taking Figure 7.3 as an example, if we pretend to not know the GO terms of domain S4, we can use the GO terms of its radius one (immediate) neighboring domains: FtsJ, tRNA-synt_1b, Ribosomal_S4, PseudoU_synth_2, KOW, and Ribosomal_S4e to infer

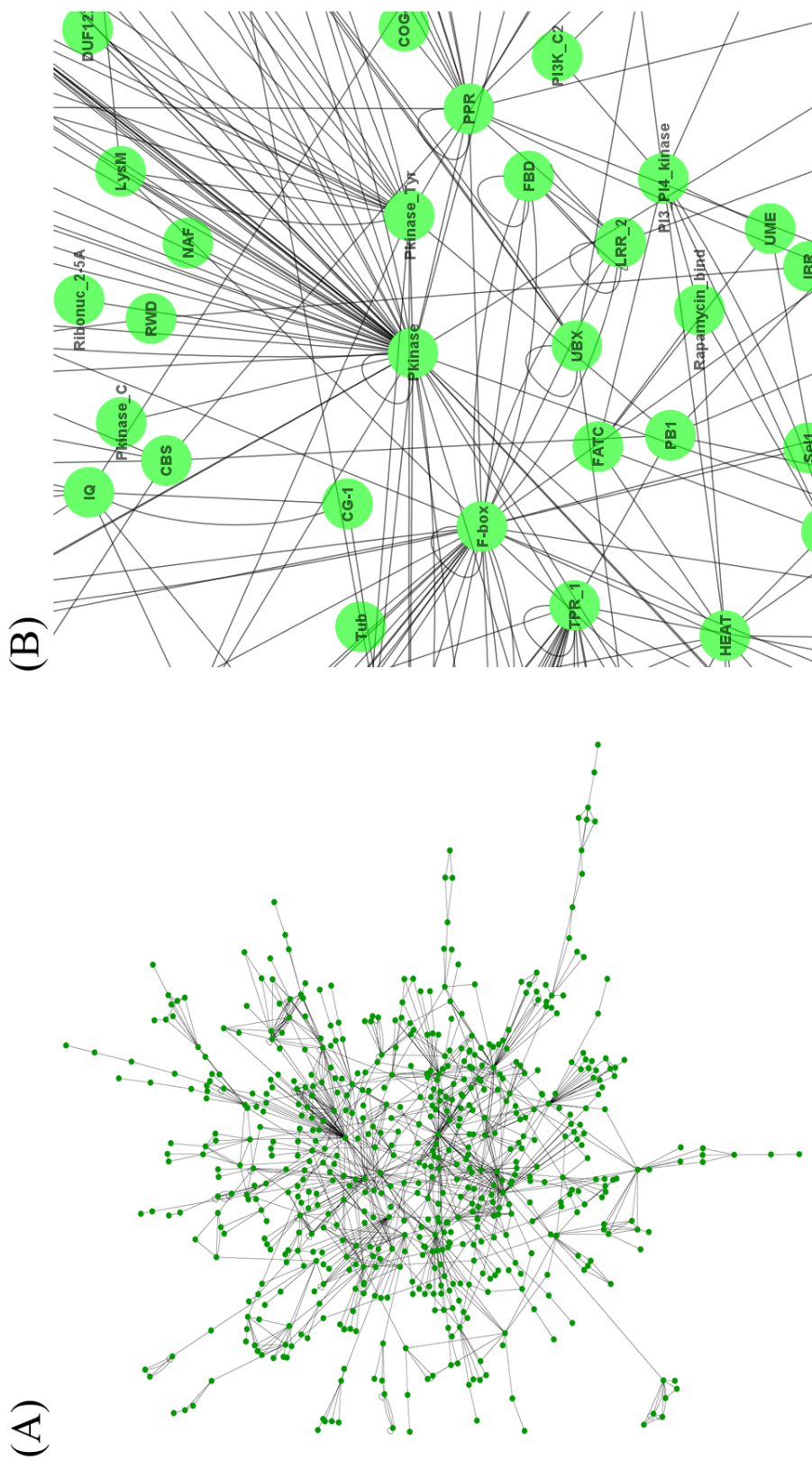


Figure 7.2: Domain co-occurrence network of Arabidopsis thaliana. In total, the DCN of Arabidopsis contains 141 disconnected sub-graphs (each one has no edges connecting to any other sub-graphs); and most of the sub-graphs have less than 10 nodes. (A) is the largest DCN sub-graph, or the main graph, of Arabidopsis DCNs that has 626 nodes and 1,304 edges. The graphs shown in Figure 7.1 and Figure 7.3 are two acutal examples of the small sub-graphs in Arabidopsis DCN. (B) is an enlarged partial view of Arabidopsis main DCN, in which domain Pkinase is a hub.

the GO terms of the central domain. We can also do that by incorporating radius two neighboring domains, including Spb1_C, Supt5, and RS4NT. In our experiments of predicting domain functions, we used *Arabidopsis thaliana*, *Saccharomyces cerevisiae*, and *Homo sapiens* as benchmarks. We included *Arabidopsis thaliana* because it has been well-studied as a model plant species and has high-quality genome annotation. We incorporated *Saccharomyces cerevisiae* and *Homo sapiens* because they are often used as model species for protein function prediction.

Three methods were used in our experiments to predict domain functions. The first is the most straightforward method, majority vote or neighbor-counting: count the appearance number of every GO term occurred in the neighboring nodes of the target domain; and then rank the GO terms based on this occurrence frequency; and the top ranked GO term(s) is (are) considered as the predicted GO term(s). This method was used by [32] to predict protein functions based on yeast protein-protein interaction networks.

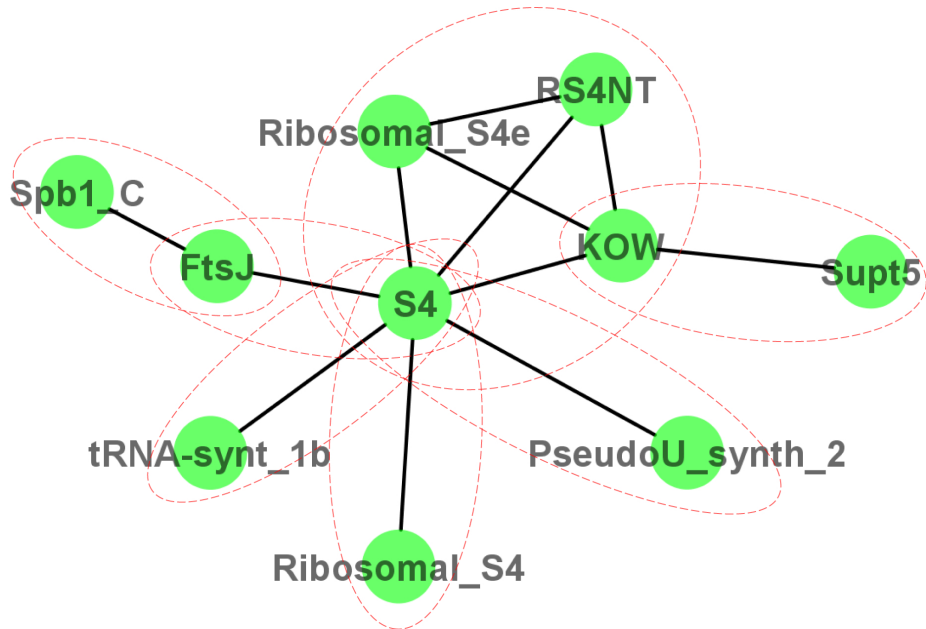


Figure 7.3: The relationship between sub-graphs and protein function. Each red eclipse encircles the domains co-occurred in one protein.

The second method in our experiment considers the distribution of every GO term in the entire DCN. This method was used by [143] to improve the performances of the neighbor-counting method mentioned above. This method ranks GO terms gathered from the neighboring domains by a χ^2 value:

$$\chi_i^2 = \frac{(n_i - e_i)^2}{e_i}$$

where i denotes a GO term, for example, GO:0090295, negative regulation of transcription by nitrogen catabolites; n_i denotes the observed number of GO term i in m -neighboring domains; and e_i denotes the expected number of GO term i appeared in m nodes.

Besides the above two methods, we also designed a Support Vector Machine (SVM)-based method. For each domain, we generated a feature-vector which included: (1) The occurrence frequency of each of the 31,398 GO terms defined by Gene Ontology [12], gathered from the neighboring nodes of the target domain. For example, if there are in total two GO terms occurring in the neighboring nodes, and each of them occurs once, the frequencies of both of these two GO terms are 0.5, and the frequencies of all the other GO terms are 0. (2) The occurrence frequency of each of the six enzyme families collected from the neighboring nodes of the target domain. (3) The occurrence frequency of each of the 20 amino acids of the target domain. From all the proteins of the target species, the ones that are found to have the target domain are gathered; and then the segments of the domain region are used to calculate the occurrence frequency of the 20 amino acids. (4) Secondary structure information: the occurrence frequency of helix (H), strand (E), and coil (C) of the target domain, which are calculated in a similar way as in (3). (5) Solvent accessibility information: the occurrence frequency of solvent exposed and buried amino acids of the target domain. The secondary structures and solvent accessibilities are predicted by SCRATCH [153]. Some of these features, such as secondary structure,

amino acid sequence, and solvent accessibilities, have been widely used in protein function prediction [190, 153, 191, 192]; therefore, it seemed reasonable to also test them in DCN-based predictions. However, our experiments showed that not all of these features made positive contributions to improvements in accuracy. Details are discussed in the "Results and Discussion" section.

According to TAIR8, Arabidopsis has 1,454 domains in total. 736 of these domains have at least one GO term available in Pfam (i.e. domains with known function) and have at least one GO term available in its radius one neighboring nodes. From these 736 nodes, we performed a leave-one-domain-out cross-validation. Figure 7.4 shows an example, in which we supposed there are in total only four domains existing in the DCN. Each time we left one domain out, which is domain *a* in Figure 7.4, and treated its GO terms unknown. From the remaining domains, which are domain *b*, *c*, and *d* in Figure 7.4, we generated a feature vector from each of these domains. A training dataset for each GO term was then constructed in which the domains having the function of this GO term were labeled positive and the ones not were labeled as negative examples. Figure 7.4 shows the training dataset for "GO term 2". If one domain contains several GO terms, which happens quite often, each of the GO terms is included as positive examples in its training dataset. As shown in Figure 7.4, "GO term 2" occurs in both domains *b* and *c*, so the training dataset of "GO term 2" contains two positive examples with feature vector *b* and *c*. The negative examples are randomly selected from all of the domains that do not have the GO term. If a feature vector of a domain has been included as a positive example, it will not be selected as a negative example.

A binary SVM model was trained for each of the GO terms occurring in the remaining domains using *SVM^{light}* [193], and then the target domain, domain *a* in Figure 7.4, was classified by each of these SVM models. Every SVM model generates a predicted value, based on which all the GO terms are ranked, and the top ranked

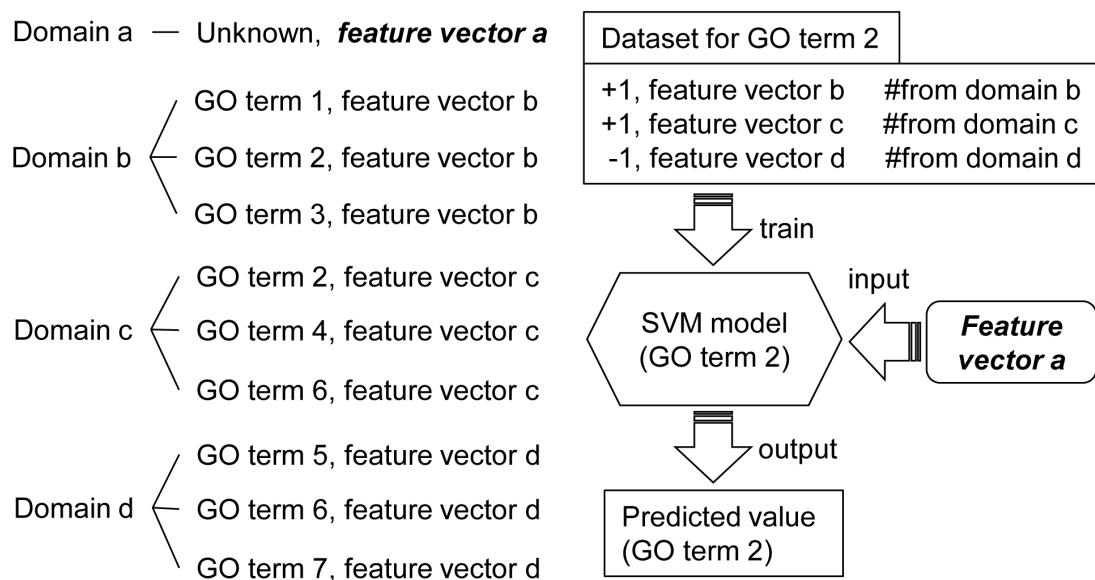


Figure 7.4: Predicting the GO terms of domain a using SVM classification. The DCN is supposed to have only four domains: *a*, *b*, *c*, and *d*, and the functions of domain a are treated unknown. It only shows the binary classification process for GO term 2. This same process should be applied on every GO term (GO term 1-7) that occurred in the training domains (domain *b*, *c*, and *d*). The top ranked GO term(s) is (are) treated as the final predicted functions.

GO terms are considered as the predicted GO terms. In Arabidopsis, 467 models on average were trained every time when classifying a target domain, and 403 for yeast, and 631 for human.

Many nodes in the DCN have a self-loop, i.e. an edge starting from and ending to the same node. If a domain with self-loop is selected as a target domain, it also exists in its own radius one neighboring domains, which makes the GO terms of the neighboring domains contain the real GO term(s) of the target domain. Therefore, the self-loop of every target domain, if exists, is removed.

7.3.3 Domain function prediction - enzyme family

We also used DCN to predict whether a domain is an enzyme domain; and if so, which of the six Enzyme Commission (EC) classes it belongs to. These six enzyme classes are: Oxidoreductases, Transferases, Hydrolases, Lyases, Isomerases, and Ligases, which have an EC number starting from 1 to 6, respectively. A mapping file between GO terms and Enzyme Commission (EC) numbers was downloaded from Gene Ontology website (<http://www.geneontology.org/external2go/ec2go>). If a domain contains at least one GO term that maps to an EC number, this domain is considered as an enzyme domain.

To predict whether a domain is an enzyme domain, we applied SVM along with all the features mentioned in the previous section. In Arabidopsis, there are a total of 307 enzyme domains that have at least one GO term or EC number available/known in the radius one neighboring domains, i.e. the GO term and EC classes occurrence frequency are not all 0. Each of these domains was considered as a positive example. The negative examples consisted of the domains with known functions that are not enzyme-related, i.e. none of its GO terms maps to an EC number. In this way, we eliminated the “Domain with Unknown Function” (DUF) domains, because these domains may not be non-enzyme domains. We gathered a total of 429 negative

domains/examples in Arabidopsis. A binary SVM model was built by *SVM^{light}* [193], and different kernels and combinations of features tested by several leave-one-out cross-validations. Besides SVM, we also tested a neighbor-inference method, by which if the neighboring nodes contain at least one EC number, we predicted the target (central) domain an enzyme domain; otherwise, non-enzyme domain.

Each of the 307 Arabidopsis enzyme domains having an enzyme class number starting from 1 to 6 based on the first digit of their EC numbers. *SVM^{multi-class}* [193] was used to build SVM models and classify a query domain into one of the six classes. A leave-one-out cross-validation was performed on these 307 enzyme domains, which contain 90 domains in EC class 1, 91 domains in EC class 2, 59 domains in EC class 3, 19 domains in EC class 4, 13 domains in EC class 5, and 37 domains in EC class 6.

7.3.4 Protein function prediction

To predict the functions of a query protein, we used a DCN-based aggregated neighbor-counting method. Given the amino acid sequence of a query protein, we run a profile-profile alignment tool HHsearch [157] against Pfam profile database (downloaded at <ftp://toolkit.lmb.uni-muenchen.de/HHsearch/databases/>) to detect Pfam domains. In order to determine the most relevant species of the query proteins, a PSI-BLAST search was performed against the whole genome protein sequences of *H. sapiens*, *S. cerevisiae*, *C. elegans*, *D. melanogaster*, 15 plants species, and 398 single-chromosome prokaryotic species, and the species having the most significant PSI-BLAST hit (with the minimum e-value) was considered the relevant species. The DCN of the relevant species was used to make functional prediction for the query protein. The aggregated neighbor-counting was then used to predict functions. From the HHsearch result report of each query protein, the detected Pfam domain(s) with an e-value ≤ 0.01 was (were) extracted. For each of these extracted domain(s), we

retrieved its (their) radius one neighboring domains, and the neighboring domains of all extracted domains were put together. The GO terms for these neighbor domains were retrieved from Pfam database and ranked by their occurrence frequencies. This list of ranked GO terms was our final prediction. Figure 7.5 shows an example of using the aggregated neighbor-counting method to predict protein functions.

7.3.5 Phylogenetic tree construction and its evaluation

To construct a phylogenetic tree of a group of species, we aligned their DCNs to identify conserved sub-networks, or common network topology, which can reveal the evolutionary significant patterns between species. This novel method is different from existent sequence-based methods, as when comparing the network topologies, since it uses the entire proteome of each species to infer the phylogenetic relationship.

To align the DCNs of two species a and b , we define the DCN of species a as a graph $G_a = (V_a, E_a)$, where V_a is the set containing all the vertices of graph G_a , and E_a is the set containing all the non-redundant edges in graph G_a . Similarly, we define the DCN of species b as a graph $G_b = (V_b, E_b)$. The mutual vertices between graph G_a and graph G_b (i.e., the same domains exist in both the DCNs of species a and b) are defined as set $V_{(a \cap b)} = V_a \cap V_b$. We identified the mutual vertices $V_{(a \cap b)}$, and calculate $NUM(E_a \cap E_b)$ - the number of mutual edges between graph G_a and graph G_b . Because we do not assign weights to edges, two edges from two graphs connecting the same vertices are considered equal. Then we count the total number of unique edges connecting only mutual vertices in both graph G_a and graph G_b , denoted as $NUM(E_a^{mutual-vertices} \cap E_b^{mutual-vertices})$. Then the similarity score between graph G_a and graph G_b is calculated as the number of mutual edges, divided by the total number of unique edges on the mutual vertices:

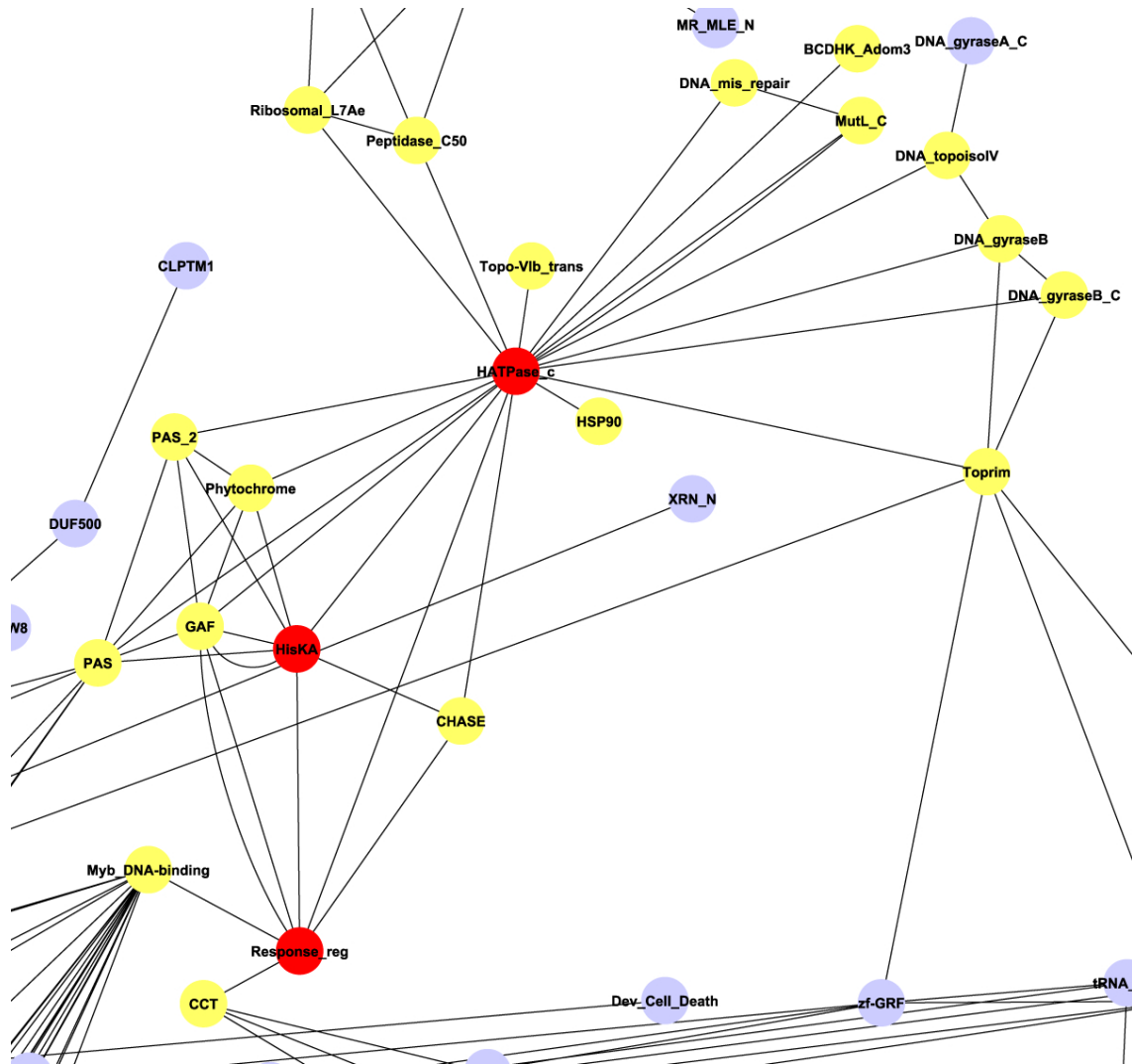


Figure 7.5: An example showing the DCN-based aggregated neighbor-counting method for protein function prediction. The query protein for this example contains 1,225 amino acids and has an ID “DDB_G0274191” in DictyBase (http://dictybase.org/gene/DDB_G0274191). Three unique Pfam domains (red vertices) were detected by HHsearch, each with an e-value ≤ 0.01 . It had a PSI-BLAST hit to a protein of *Vitis vinifera* (grape) with an e-value e-118. Therefore, the DCN of *Vitis vinifera* was used to make predictions. The vertices in yellow are the radius one neighboring vertices of the three Pfam domains. The GO terms of these yellow vertices were put together and ranked based on their occurrence frequencies. This query protein contains 14 real GO terms. Everyone of the top 3 predictions (GO:0005524, GO:0006355, and GO:0016020) ranked by DCN-based aggregated neighbor-counting method matches one of the real GO terms. Besides that, the 5th (GO:0007165), 7th (GO:0000155), 13th (GO:0000160), and 14th (GO:0000156) ranked GO terms have an exact match to the real GO terms.

$$S_{(a,b)} = \frac{NUM(E_a \cap E_b)}{NUM(E_a^{mutual-vertices} \cup E_b^{mutual-vertices})}$$

Figure 7.6 illustrates an example of the graph alignment algorithm, in which we align the two graphs as Figure 7.6 (A) and (B). We at first find the mutual vertices between (A) and (B), which are b , c , d , and e as shown in Figure 7.6 (C). Then three mutual edges between graph (A) and graph (B): $c-d$, $c-e$, and $e-d$ are picked as shown in Figure 7.6 (D). The seven unique edges on the mutual vertices between graph (A) and (B) are shown in Figure 7.6 (C). In this case, there are two edges occurred between node c and d , which is due to the fact that two proteins contain both domain c and d . These two redundant edges are considered as one edge. Thus the number of unique edges on the common vertices is 6. Therefore, the similarity score between graph Figure 7.6 (A) and (B) is equal to the number of edges in graph Figure 7.6 (D) divided by the number of unique edges in graph Figure 7.6 (C), which is $3/6 = 0.5$.

This alignment algorithm is straightforward, easy to implement, and has low computational complexity compared to complicated global alignment algorithms. If n is the total number of unique domains in the two DCNs, the computational complexity will be at most $O(n^2)$. In the future, we plan to try more advanced global graph-alignment algorithms, such as IsoRankN [194].

Given a group of species, we used our graph alignment algorithm for pair-wise comparisons and generated a distance matrix ($distancescore = 1 - similarityscore$). We generated a phylogenetic tree using the program “NEIGHBOR” in the phylogeny inference package PHYLIP [195], which implements the neighbor joining method [196].

Unlike the phylogenetic study of more advanced organisms, which has plenty of morphology and archeology evidence available, there are still uncertainties in bacteria taxonomy. However, the scientific community usually considers the classification

presented in the book Bergey’s Manual of Determinative Bacteriology [197] as the best approximation. Bergey defines a set of taxonomy codes to indicate the classification. For example, the organism *Lactobacillus casei* has a Bergey’s code B13.3.2.1.1 indicating it belongs to kingdom Bacteria, Division 13 (Firmicutes), Class 3 (Bacilli), Order 2 (Lactobacillales), Family 1 (Lactobacillaceae), and Genus 1 (Lactobacillus). In our work, we use Bergey’s classification as the reference, and compared our phylogenetic trees to this reference. The similarity between our phylogenetic tree and Bergey’s classification was calculated by ComPhy [198], which counts the number of agreed quartets between our phylogenetic tree and Bergey’s classification, and uses the percentage of the agreed quartets as the similarity measure, whereas a quartet is a sub-graph topology containing four taxa (tree nodes) [198].

In order to comprehensively evaluate the potential of DCNs to infer phylogeny, we conducted experiments on six datasets with different combinations of 398 single-

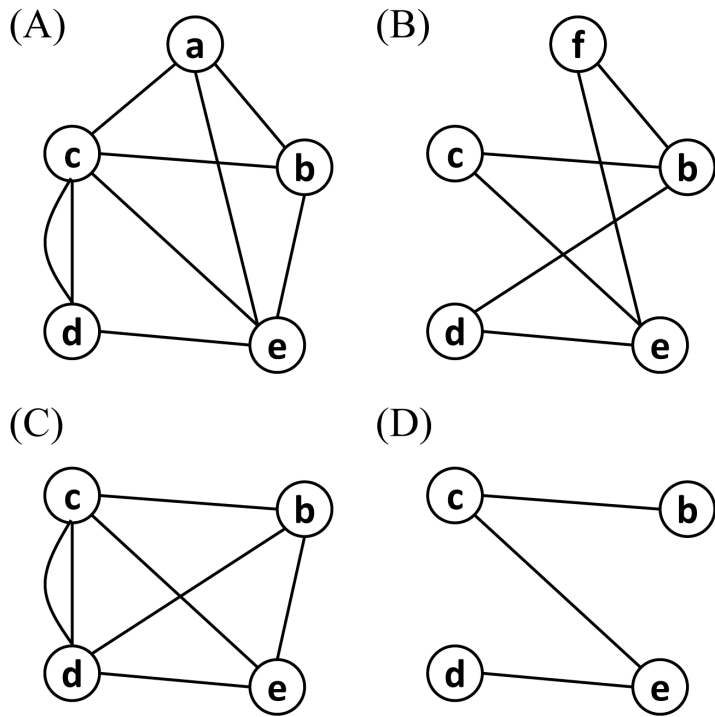


Figure 7.6: An example illustrating the graph alignment algorithm we utilized.

chromosome prokaryotic genomes (strains). These six datasets were previously used to compare several different phylogeny inference methods [198]. For direct comparison, we used the same version of the datasets: 432 prokaryotic genomes downloaded from NCBI in September 2007. After removing 34 multi-chromosome species, we included 398 species in our dataset, which contains 29 Archaea species and 369 Eubacteria species (Figure 7.7). Dataset 1 consists of 52 randomly selected species from Bergey’s taxonomy tree [198]. Dataset 2 contains 53 species, 28 of which are randomly selected from the Archaea species, and 25 of them are randomly selected from Eubacteria species. Dataset 3 contains all the 398 organisms. Dataset 4 is composed by Bacterial Division 12 (181 species). Division 12 is a large division containing approximately half of the 398 genomes. Dataset 5 is formed by Bacterial Division 12 (181 species) and Division 13 (96 species). These two are big clusters, and the phylogenetic tree generated on dataset 5 should contain two tiger clusters. Dataset 6 contains 54 organisms, which were obtained from Deeds [199], a phylogeny inference method using domain structures networks.

7.4 Results and Discussion

7.4.1 Statistical Properties of Domain Co-occurrence Networks

We analyzed the statistical properties of the DCNs of 15 plant species, yeast, and human, and found that they share several common features. Figure 7.7 (A) depicts the node-degree distribution of four example species, *Arabidopsis thaliana* (*Arabidopsis*), *Chlamydomonas reinhardtii* (green alga), *Zea mays* (maize), and *Physcomitrella patens* (moss). Appendix Figure B.1 shows the node-degree distributions of *H. sapiens*, *S. cerevisiae*, *C. elegans*, *D. melanogaster*, and 15 plant genomes. This log-log

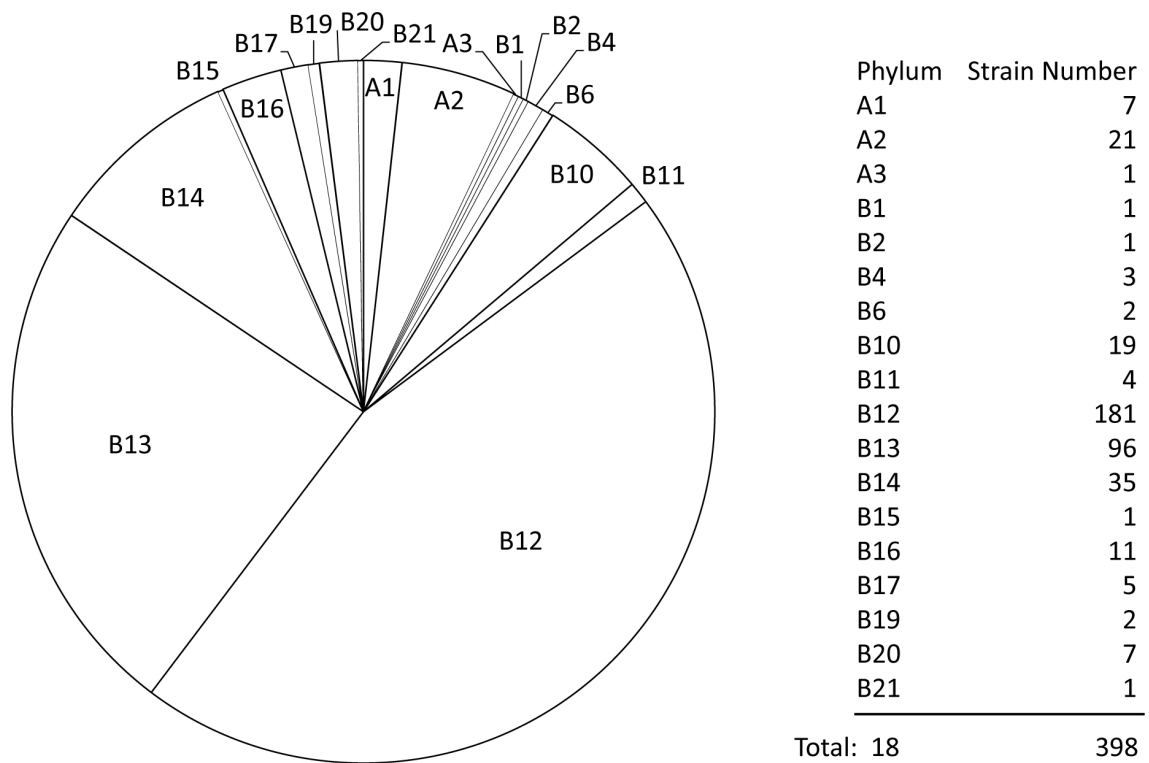


Figure 7.7: Composition details of the 398 single-chromosome prokaryotic genomes (strains)

plot shows that the number of nodes with a specific degree value (degree is the number of edges linked to a node) mathematically follows a power law distribution, because the logarithmic relationship between two variables approximates a linear relationship. This property shows that the DCNs are scale-free networks

$$P(k) \sim k^{-\gamma}$$

where $P(k)$ is the probability of having a node with k edges linking to it, and γ is a species-specific constant.

Figure 7.7 (B) plots another property of scale-free networks, the small-world phenomenon. Figure 7.7 (B) shows the frequency of node pairs whose shortest path has a specific value k ($k = 1, 2, 3\dots$). Appendix Figure B.2 shows the shortest path length distributions of *H. sapiens*, *S. cerevisiae*, *C. elegans*, *D. melanogaster*, and 15 plant genomes. The majority of the node pairs have a path shorter than 5, indicating that most of node pairs can be reached within five steps.

Average clustering coefficient is another important property of scale-free networks. For un-directed networks, the clustering coefficient of a node n is calculated by:

$$C_n = \frac{2e_n}{k_n(k_n - 1)}$$

where e_n is the number of connected node pairs among immediate (one edge away) neighboring nodes of the central node n , and k_n is the number of immediate neighboring nodes of the central node n [165, 200]. Clustering coefficient indicates the degree of the nodes to be clustered. The average clustering coefficient of the whole network is calculated by taking the average of all nodes' clustering coefficients. Figure 7.7 (C) plots the relationship between average clustering coefficient and node degrees. This log-log plot shows that the clustering coefficient distribution of DCNs also follows a power law. Appendix Figure B.3 shows the average clustering coefficient

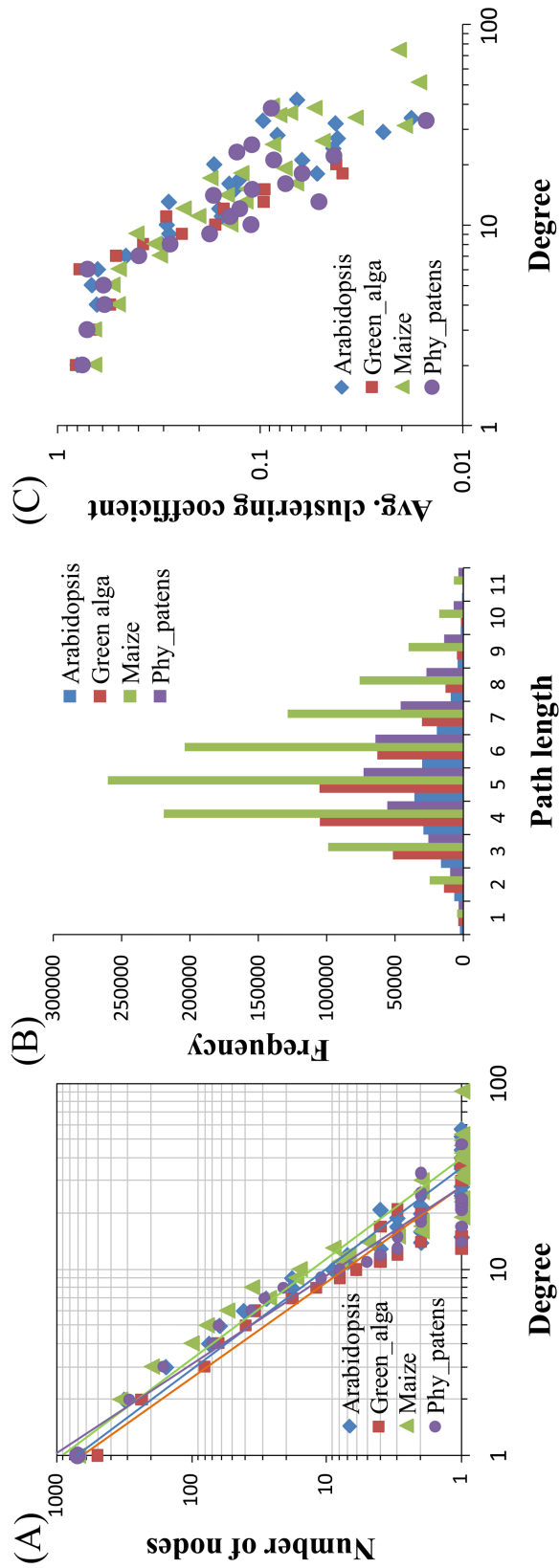


Figure 7.8: Statistical properties of DCNs of four representative species. (A) The degree distribution plots (scale-free); (B) the distributions of lengths of the shortest paths (small-world); (C) the log-log plots of clustering coefficients against degrees (hierarchical modularity).

distributions of *H. sapiens*, *S. cerevisiae*, *C. elegans*, *D. melanogaster*, and 15 plant genomes. Figure 7.7 (B) and (C) illustrate that in DCNs, the nodes with a small degree value, usually tend to form a complete graph (clique), or very dense sub-graph (i.e. with higher average clustering coefficient), and one dense sub-graph is connected with other dense sub-graphs by hub nodes, i.e. popular nodes with high degree, which are similar to politicians and celebrities in human social networks. This indicates that a DCN is a hierarchical network composed of densely connected modules [201].

Experiments conducted on yeast and human DCNs show that they follow the same properties (data not shown). Therefore, they share the same characteristics as other scale-free networks, such as social networks, the World-Wide Web, and some biological networks, including protein-protein interaction networks and metabolic networks.

7.4.2 Error and attack robustness of DCNs

Scale-free networks are usually very recalcitrant (remains un-fragmented, i.e., every node is connected, by some paths, to the others) to random removal of nodes, named as “failure”, but very vulnerable to “attack”, in which the nodes with highest degree are removed first [202, 28, 203]. We simulated both perturbations on DCNs.

The Average Shortest Path (ASP) of a network is the average of all the shortest paths between all pairs of nodes. A smaller ASP means the network has a better interconnectivity. If a network is fragmented, i.e., containing several independent sub-graph(s) whose nodes have no connections to the other sub-graphs, the path between these un-connected nodes become infinity, so does the ASP of the entire networks. As mentioned in the section “Materials and Methods”, the DCN for a given species usually contains several disconnected sub-graphs, with one main graph containing approximately >80% of all the domains. Therefore, it only makes sense to study the changes of ASP under the two perturbations (“failure” and “attack”) on the main network. According to our criteria, the simulations were terminated and

the network was considered fragmented even if one node became fragmented. This is probably a too stringent criteria compared with the study performed on yeast protein-protein interaction networks [202], where the shortest path between disconnected nodes were treated a large number instead of infinity. That criteria allows calculating the interconnectivity of a network after some nodes are fragmented.

We performed simulations on the DCNs of yeast, Arabidopsis, maize, soybean, and human. Taking the main network of Arabidopsis as an example, after removing the node with the highest degree, domain Helicase_C, the ASP of the network went to infinity, indicating that this attack behavior caused at least one node to be disconnected with the remaining network (i.e. network fragmented). Compared with “attack”, scale-free networks are usually less vulnerable to “failure”, in which nodes are removed randomly, because the probability of randomly selecting a node with high degree value is very low, according to the power law distribution. This trend is also found true in DCNs, but to a lesser extent. We performed 10 rounds of failure simulations on the DCNs of human, yeast, and Arabidopsis, and the average number of nodes deleted before the DCN had at least one nodes fragmented were 5.5, 4.0, and 7.3, with standard deviations 4.2, 3.5, and 5.9, respectively. These are the 0.4%, 1.6%, and 1.6% of all the vertices in their main networks, respectively.

The DCN was considered fragmented even if one node became fragmented. In order to study the interconnectivity of the networks after some of the nodes became fragmented and to make comparison with the study performed on yeast protein-protein interaction networks [202], instead of using infinity, we assigned a specific value as the shortest path when two nodes were disconnected. The value was equal to the maximum length of the shortest paths before the removal action. We observed that DCNs were very vulnerable to “attack”, i.e., the modified ASP had a sharp increase and reached the peak after removing 2.5% and 5% of the nodes of yeast and rice, respectively, and much more robust against “failure”, i.e., the modified ASP

had a much slower increase compared to “attack”. However, we found that DCNs were not as robust as yeast interaction networks under “failure”, as the ASP in yeast interaction networks remained almost unchanged after removing up to 50% of nodes based on study [202].

In summary, the results show that DCNs follow the general robustness characteristics of scale-free networks, but are more vulnerable to perturbations than protein interaction networks and other high-density scale-free networks. One possible reason is that the “domain” used in our DCNs sometime may be a coarse unit, which sometime may be further divided into smaller units. The other reason is probably because DCNs are less densely connected than protein interaction networks and the internet, or because of the much smaller size of DCNs. This may indicate that the interconnectivity of DCNs is not indispensably important, mainly because of the nature of DCNs. For the networks showing great robustness, their functions can only be carried out by interconnectivity, such as the Internet. On the other hand, for the domains that do not have a direct connection in the DCN, they may also be able to physically interact with the others and carry out functions. For example, two domains from two different proteins may be interactive and bind with each other, but they do not co-exist in one protein.

7.4.3 Domain function prediction - GO terms

To assess the capability of DCN to predict the GO terms of a target domain, we developed and evaluated three methods: neighbor-counting, χ^2 , and a SVM-based method, as shown in Table 7.1. The neighbor-counting method retrieves the GO terms of all neighboring vertices and ranks them by their occurrence frequencies. This ranked list of GO terms is its final predictions. The χ^2 method not only considers the occurrence frequencies of GO terms in the neighboring nodes, but also the overall distributions of the GO terms in the entire DCN. The SVM-base method uses

the known examples of the target domains to train a SVM model, and then uses this SVM model to make predictions. Details can be found in the “Materials and Methods” section. We evaluated the top 3 ranked GO terms by the criteria that if one of the top 3 GO terms matches one of the real GO terms of the target domains, we count it as a correct prediction. Similarly, the top 1 GO term evaluation criteria is that only if the top 1 GO term matches one of the real GO terms, it is considered as a correct prediction. For the neighbor-counting and χ^2 value methods, we calculated the percentage of correctly predicted domains. For the SVM-based method, we performed a leave-one-domain-out cross-validation, and report the average percentage of correctly predicted domains. These three methods were tested on the target domains whose “radius one” neighboring domains have at least one GO term available. The number of target domains for Arabidopsis, yeast, and human are 736, 518, and 953, respectively.

The evaluation criteria mentioned above is a yes-or-no binary criteria. Considering that two different GO terms may share functional similarity, we also calculated the average similarity score between the best prediction and the real GO terms (Table 7.2), using the tool G-SESAME [158], which defines the semantic similarity between two GO terms as the percentage of their common sub-graphs starting from the root of the GO “directed acyclic graph” (DAG) [158].

For the SVM-based method, we tested each of the five set features built on GO terms, EC numbers, amino acid sequences, secondary structures, and solvent accessibilities. We first used only GO term frequency and kept adding one of the other features. After comparing the performances of several leave-one-domain-out cross-validations, we found that the “GO terms frequency” feature, generated from the neighboring domains, had the largest positive influence on final prediction accuracy. Adding any one of the remaining features slightly decreased the accuracy by 1-3% on Arabidopsis. We also tested four kernel functions for the SVM model: linear,

Table 7.1: The prediction accuracy of using neighbor counting, χ^2 , and SVM when predicting GO terms. For SVM, it’s the average accuracy of a leave-one-domain-out cross-validation. “Top 1” (“Top 3”) indicates when “the top one ranked GO term” (“one of the top three ranked GO terms”) matches one of the actual GO terms of the target domain, it is considered a correct prediction. The values shown under “Top 1” are also the precision values for top 1 prediction. The precision values for ”Top 3” predictions can be found at Table 7.4.

Species	Top 3			Top 1		
	Neigh.-Count.	χ^2	SVM	Neigh.-Count.	χ^2	SVM
Arabidopsis	64.8%	59.6%	58.5%	50.4%	42.9%	47.6%
Yeast	67.0%	61.8%	61.1%	52.3%	41.9%	51.6%
Human	65.8%	58.3%	60.4%	47.4%	37.5%	45.1%

Table 7.2: The average semantic similarity scores of the best predictions, among “Top 1” or “Top 3” GO term(s). “Top 1” (“Top 3”) indicates that for each target domain, we calculate the pair-wise similarity scores between the top one (top three) ranked GO term(s) and the actual GO terms, and the highest score is considered as the similarity score of the best prediction. Semantic scores are calculated by the tool G-SESAME [158].

Species	Top 3			Top 1		
	Neigh.-Count.	χ^2	SVM	Neigh.-Count.	χ^2	SVM
Arabidopsis	0.814	0.790	0.767	0.680	0.652	0.620
Yeast	0.835	0.818	0.772	0.679	0.646	0.626
Human	0.826	0.790	0.791	0.665	0.619	0.630

polynomial, radial basis function (RBF), and sigmoid tanh, and found that the linear kernel function yielded the best accuracy. Therefore, the SVM method used in our evaluations (Table 7.1 and 7.2) only uses “GO term frequency” features consisting of 31,398 values and the linear kernel function. Moreover, we randomly selected different numbers of negative examples in each GO term’s training dataset (see Figure 7.4 for how to construct a training dataset), and found that the ratio 2:1 negative to positive examples generates the best accuracy, which is reported in Table 7.1 and 7.2.

Our results (Table 7.1 - 7.4) show that the SVM-based method has the better

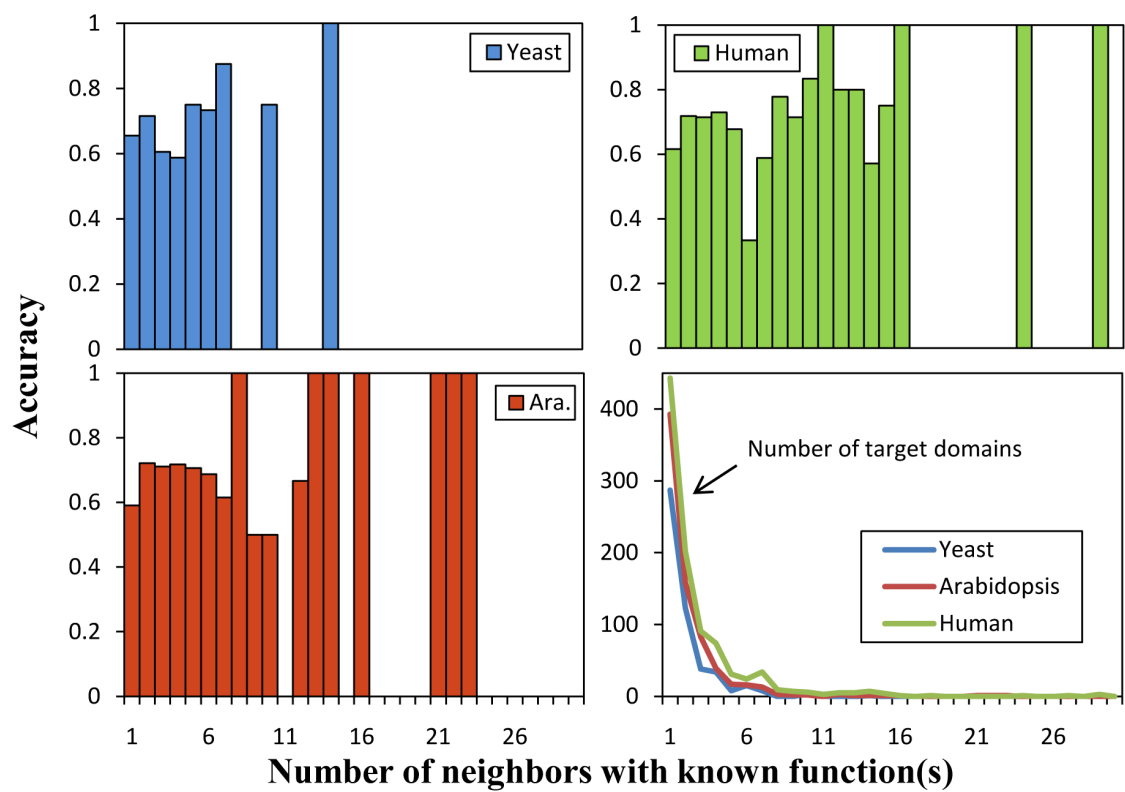


Figure 7.9: The number of neighbors with known functions versus the prediction accuracy of the neighbor-counting method. The more neighbors with known functions, the higher accuracies neighbor-counting method can achieve when predicting the functions of the central node. Most of the target domains have less than five neighbors with known functions.

performance compared to the χ^2 -based method when considering only top 1 predictions; and the simple neighbor-counting method, which relies solely on DCN topology, achieves the best performance, in terms of both the “yes-or-no” evaluation and the average similarity between the best predictions and real GO terms. The accuracy of top 3 predictions using neighbor counting on three species is more than 65% and top 1 predictions more than 47%. The average GO similarity of top 3 predictions is more than 0.81 and top 1 prediction more than 0.66. The good performance suggests that the connectivity of DCNs contains rather rich information for protein function prediction. To further investigate the neighbor counting method, we plotted Figure 7.9, the relationship between “the number of neighboring domains with known function” and “the prediction accuracy” (top 3, radius one neighboring domains) on three test species: Arabidopsis, yeast, and human. The results showed that, in general, more neighboring domains with known functions can generate better accuracy. Although most of the target domains have less than five neighboring domains with known function, as long as there is one neighboring domain with functional information available, neighbor-counting can correctly predict 60% of the domains (top 3). This shows that domain co-occurrence is a very useful indicator of protein function.

Table 7.3: The average recall value of using neighbor counting, χ^2 , and SVM when predicting GO terms. Recall was calculated as the correctly predicted GO terms divided by the total number of real GO terms. Average recall was the sum of recall value for each domain divided by the total number of test domains.

Species	Top 3			Top 1		
	Neigh.-Count.	χ^2	SVM	Neigh.-Count.	χ^2	SVM
Arabidopsis	47.0%	42.5%	40.4%	21.4%	18.6%	20.0%
Yeast	47.4%	44.3%	41.5%	22.0%	18.0%	20.9%
Human	45.8%	40.1%	40.4%	19.8%	16.5%	19.1%

We also conducted experiments on integrating both radius one and two neighboring domains. For neighbor-counting and SVM-based method, our results show that

Table 7.4: The average precision value of using neighbor counting, χ^2 , and SVM for top 3 prediction. Precision was calculated as the number of correctly predicted GO terms divided by the number of predictions a method made (in this case, three). Average precision was the sum of precision value for each domain divided by the total number of test domains. The precisions of “Top 1” can be found at Table 7.1.

Species	Top 3		
	Neigh.-Count.	χ^2	SVM
Arabidopsis	40.7%	36.8%	35.3%
Yeast	41.3%	38.0%	36.9%
Human	38.9%	34.1%	34.7%

the accuracy has a small decrease, by 2-3% on average, after including radius two neighbors (data not shown). For the χ^2 -based method, the accuracy drops by 10% on average (data not shown). This indicates that directly expanding the neighboring radius may not help improving accuracy, as more noise may be included. In the future, we plan to incorporate more advanced algorithms such as graph kernels [204] and graph random walk algorithm [205] to infer functions from DCNs.

We selected nine most promiscuous domains [206] (i.e., the hub domains or the vertices in DCNs with a high degree) and used the three methods: neighbor-counting, χ^2 , and SVM-based method, to predict their functions (Table 7.5-7.7). Our results show that in general, if considering top 3 hits, the neighbor-counting method performs better than χ^2 and SVM-based method. However, SVM-based method performs better on Arabidopsis and human if considering only top 1 prediction. Moreover, for some promiscuous domains, χ^2 and SVM-based method perform better than neighbor-counting. For example, the top 1 prediction of neighbor-counting method has a similarity score of 0.088 on human domain “Pkinase_Tyr”, whereas χ^2 method has a score of 1.000 on the same domain. Similarly, the best prediction of neighbor-counting has a similarity score of 0.080 on human domain “SET”, whereas the SVM-based method has a score of 1.000. This shows that each of these three methods

may have its own advantages when being applied to specific domains. Table 7.5 also shows that usually the higher degree value a domain has, the better performance neighbor-counting method can achieve if considering top 3 predictions.

7.4.4 Domain function prediction - Enzyme

For a target domain, we first used SVM and a neighbor-inference method to predict whether it is an enzyme domain. On enzyme domains, we tested two methods, neighbor-counting and SVM-based method, to classify it into one of the six enzyme classes.

For SVM-based enzyme “yes or no” predictions, we found that using “GO term frequency” feature with linear kernel function generated the best accuracy. Adding other features such as EC number, secondary structure, and solvent accessibilities did not significantly influence performance (data not shown). Table 7.8 and 7.9 compare the performance of using SVM and another method: neighbor-inference, by which if the neighboring domains have at least one EC number, it is predicted as an enzyme domain; otherwise, not. We calculated and report the sensitivity of positives (Q_p), the sensitivity of negatives (Q_n), and the Matthews correlation coefficient (MCC) [190]:

$$Q_p = \frac{TP}{TP + FN}$$

$$Q_n = \frac{TN}{TN + FP}$$

$$MCC = \frac{TP \times TN - FN \times FP}{\sqrt{(TP + FN) \times (TP + FP) \times (TN + FN) \times (TN + FP)}}$$

where TP is true positive; FN is false negative; TN is true negative; and FP is false positive. The Q_p and Q_n of SVM-based method are 0.66 and 0.87 on average, which indicates the predictions are more biased towards the negative side. However, the neighbor-reference method shows more balanced results, with the Q_p and Q_n of 0.79 and 0.78. The overall classification accuracy and the Matthews correlation coefficients of these two methods are very similar (i.e., about 0.54).

To classify an enzyme domain into one of six enzyme classes, we developed and evaluated two methods, neighbor-counting and a SVM-based method (Table 7.10). For the neighbor-counting method, the most frequent enzyme class in the neighboring domains is considered as the predicted enzyme class for the central node. For the SVM-based method, we tested several combinations of the five occurrence-frequency features built on GO terms, EC numbers, amino acid sequences, secondary structures, and solvent accessibilities, and found that using “GO term frequencies” plus “EC number frequencies” generated the best performance in the leave-one-domain-out cross-validation. Adding any one of the other features slightly decreased or did not influence accuracy. As shown in Table 7.10, when there are ≥ 2 EC numbers available in the neighboring domains, both SVM and neighbor-counting method (top 3) can achieve 100% classification accuracy. Neighbor-counting cannot work on the cases of 70 target domains whose neighboring nodes do not have an EC number, which results in the decrease of its accuracy to 0.711. This is a case on which the neighbor-counting method cannot work, but SVM can make predictions based on the “GO term frequencies” features. In general, as long as there is one EC number in the neighboring domains, the accuracy of both SVM-based and neighboring-counting can reach $>90\%$. This demonstrates the abilities of DCNs to infer enzyme classes. In the future, we will test the performances of inferring sub-classes of enzyme domains.

We also tested other kernel functions in the SVM model and tried to integrate radius two neighboring domains. Linear kernel function worked the best, and inte-

grating more neighboring domains slightly decreased the accuracy (data not shown).

7.4.5 Protein function prediction

In order to test the performance of applying the aggregated neighbor-counting method to predict protein functions in a real life scenario, we randomly selected 100 proteins from the Gene Ontology FASTA sequence database (http://archive.geneontology.org/lite/2010-12-11/go_20101211-seqdb-data.gz). The functions of these 100 proteins had been annotated, and their GO terms stored in the Gene Ontology database. Therefore, we treated these GO terms as their real GO terms. We ran HHsearch [157] to search each query sequence against the Pfam profile database to detect Pfam domains. Only the Pfam domains detected with an e-value ≤ 0.01 were kept. To determine the closest relevant species for a query sequence, we used PSI-BLAST [133] to search each of the 100 proteins against the whole genome protein sequences of *H. sapiens*, *S. cerevisiae*, *C. elegans*, *D. melanogaster*, 15 plants species, and 398 single-chromosome prokaryotic species. In total, 96 proteins had at least one PSI-BLAST hit with an e-value ≤ 0.01 , and the other four had an e-value: 1.8, 1.9, 2.7, and 3.7, respectively. The DCN of the species of the most significant hit of a protein is used to predict its function.

In total, DCN-based aggregated neighbor-counting method (details see “Materials and Methods” section) generated predictions for 66 out of 100 proteins. Among the 34 proteins on which DCNs failed to make predictions, 10 of them failed because no Pfam domain was detected by HHsearch with an e-value ≤ 0.01 ; 19 of them failed because the Pfam domain(s) detected by HHsearch could not be found in the DCN of the closest relevant species; and the others failed because none of their neighboring domains had GO terms available. Because we used PfamScan, a sequence-profile alignment tool, to detect Pfam domains when we constructed the DCNs (see the “Construction of Domain Co-occurrence Networks” sub-section in the “Materials and

Methods” section) but used HHsearch, a much more sensitive profile-profile alignment tool, to detect Pfam domains on these 100 proteins, some of the domains found by HHsearch could not be found by PfamScan. This also suggests that the coverage of our current DCNs can be improved by using more sensitive domain detection tools. Another reason why some of the detected domains cannot be found in the DCNs is that we used PSI-BLAST to identify relevant species, which may not be the right species for the query protein. Thus, adding the DCNs of more species into our system may increase the coverage of function prediction.

Table 7.11 reports the average precision, recall, and the semantic similarity score of the best predictions on the 66 proteins including both multi- and single-domain proteins. Table 7.12 shows the same measurements on the 9 single-domain proteins containing only one Pfam domain (i.e., only one domain with an HHsearch e-value ≤ 0.01). Precision was calculated as the number of correctly predicted GO terms (“exact match”) divided by the number of GO predictions. Recall was calculated as the correctly predicted GO terms (“exact match”) divided by the total number of real GO terms. Our results show that the top 1 prediction on single-domain proteins can achieve a similarity score of 0.636. Considering top 3 predictions on single-domain proteins, the best prediction has a similarity score of 0.834, and above 0.95 if considering more than top 4 predictions. Figure 7.5 shows a successful prediction example for one protein, in which all top 3 predictions match the real GO terms. For both domain function and protein function predictions, if more than two edges existed between two vertices, as the case shown in Figure 7.6 between node *c* and *d*, they were reduced to one edge.

7.4.6 Evaluation of DCN-Inferred Phylogeny

We evaluated our DCNs-alignment-based phylogeny inference method on six different combinations of 398 single-chromosome prokaryotic species (strains). We bench-

marked our DCNs-based method and compared its performances with that of five other state-of-the-art phylogeny inference methods (Table 7.13). These methods include (1) BPhyOG [207], a method based on overlapping genes (OG); (2) a method based on Composition Vector Tree (CVT) with $k=5$, where k is the length of strings [208]. The concept of “Composition Vector Tree” which used the string appearance frequencies to represent a genome, and the distance between two genomes was calculated as the Euclidean distance between the two “Composition Vector Trees”. (3) [209] extended (2) by using the appearance frequencies of all the strings with length $k \leq 5$, and named this vector “Complete Composition Vector”. (4) A method based on Structural Protein Domain Universe Graph (PDUG) [199]. This method incorporates a graph consisted of protein domains, which, to some extent, is similar to our DCNs-based method, but these two methods still have big differences. PDUG consists of all the protein domains and may be derived from several species, with known protein structures. Each of these domains is treated as a node in the PDUG. The protein structure and the structural similarity are based on the structural classification of protein “fold”, and the structural similarity is used to define the edge between two nodes. An organism is assigned nodes from the PDUG based on sequential similarities, and the distance matrix between graphs is generated based on the degree distribution. (5) ComPhy [198], a method based on gene Composite Distance (CD). Gene composite distance combines Gene Dispersion Distance, Genome Breakpoint Distance, and Gene Content Distance, and it achieves higher than 90% accuracy on all the six datasets. Detailed descriptions about ComPhy can be found at [198].

Table 7.13 reports the results of our novel DCNs-based method on the six datasets. It performs very well on dataset 6, with similarity score 93.45% to Bergey’s taxonomy, which contains 54 genomes that cover almost all the major clusters of the 398 genomes (Table 7.14). The phylogenetic tree generated on dataset 6 is shown in Figure 7.10, which largely complies with Bergey’s taxonomy. Our experiments also show

that DCNs-based method is robust as it has $>85\%$ accuracy on both datasets 1 and 2 which contain randomly selected species. Dataset 2 contains 52 species randomly selected from 398 chromosomes, whereas dataset 2 contains 53 species, half of which are randomly selected from Archaea and half randomly selected from Eubacteria. The accuracies on Datasets 4 and 5 are 82.75% and 80.42% , respectively, which shows DCN-based method has a decent sensitivity to distinguish and classify closely-related species, i.e. in the deeper level of the phylogenetic tree, as dataset 4 contains only genomes from Bacterial Division 12, and dataset 5 contains both Division 12 and 13. The overall accuracy on all the 398 chromosomes is 76% , which is slightly lower but still comparable with most of the other methods. In general, our DCNs-based method gains a comparable performance when compared to most of the other methods. ComPhy consistently showed a higher than 90% agreed percentage with Bergey's taxonomy. However, our DCNs-based method is based on organism-specific DCNs and graph alignment algorithm, which are completely different to the gene-based and sequence-based methods. It is likely that combining our DCNs-based method with the method in ComPhy can further improving the performance of ComPhy. Moreover, once the DCNs are constructed for candidate species, the graph alignment process is very fast and memory-efficient, because it does not need to consider all the genes in the whole chromosome, but only the nodes of DCNs, and on average one DCN of the 398 single-chromosome organisms contains 377 vertices. Using our PERL implementation of our graph alignment, with $O(n^2)$ complexity (n is the number of unique domains in the two DCNs), the average time of aligning two DCNs of the 398 species is around 1.6 seconds using a single 2.4GHz Intel(R) Xeon CPU at a 64-bit Linux machine.

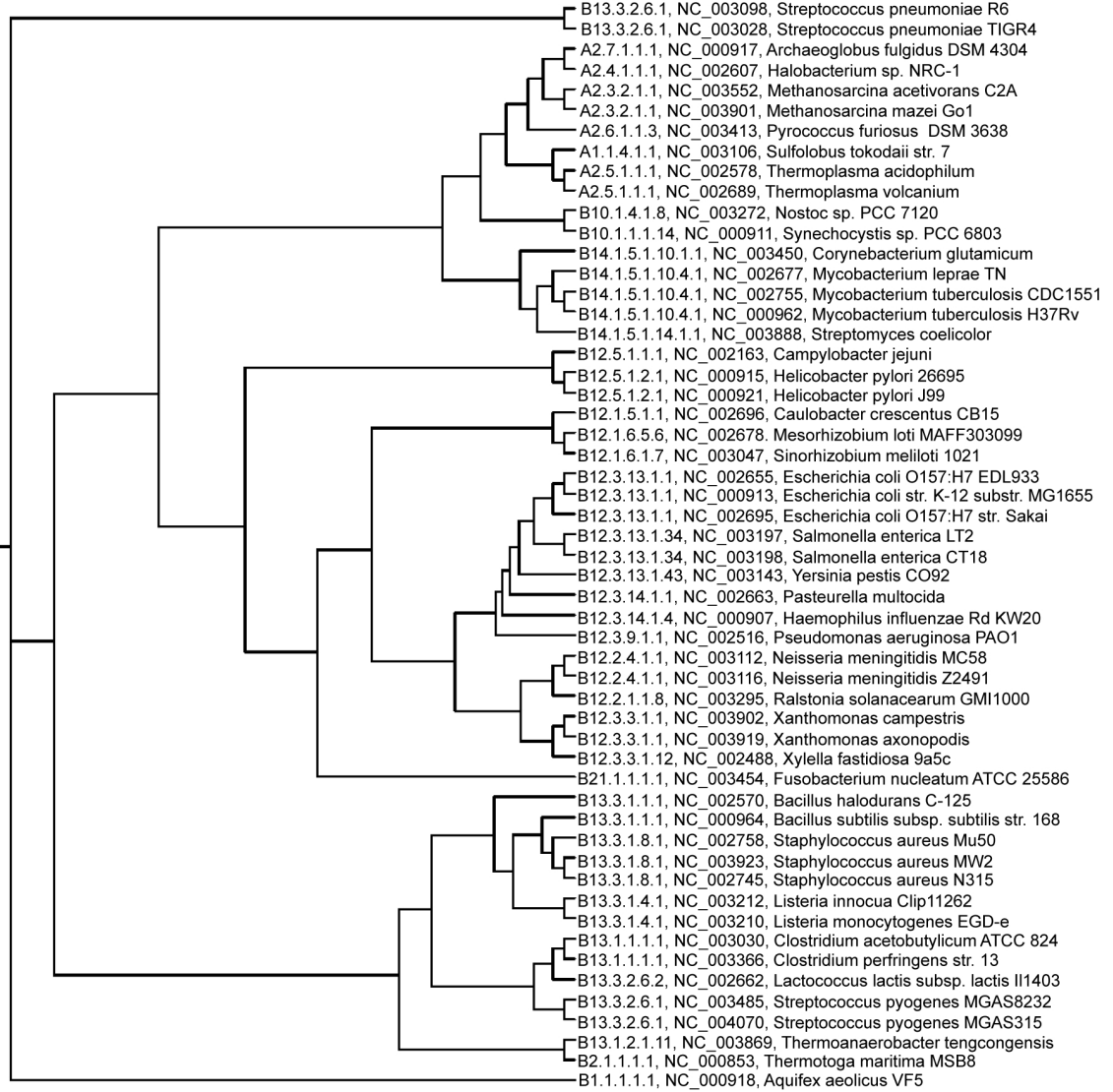


Figure 7.10: The phylogenetic tree generated on 54 single-chromosome prokaryotic taxa by our DCNs-alignment-based inference method. For each organism, the Bergey's code, NCBI ID, and the scientific name are shown. The percentage of the agreed quartets between this tree and the Bergey's taxonomy is 93.45%.

7.5 Conclusions and Future Work

We present a set of new DCN-based methods to predict protein function and to infer species phylogeny. We tested our methods on the genomes of several representative species and a large phylogeny benchmark. The results showed that DCN-based methods can predict protein function rather accurately, probably because DCNs constructed from all the proteins of a genome are rather complete and reliable in comparison with other protein networks. Our unique approach of constructing phylogeny by aligning species' DCNs also yields good performance that are comparable to other established methods, making it a complementary and valuable addition to the repository of phylogenetic analysis tools.

Despite initial promising results, there is still room to improve DCN-based function prediction and phylogeny inference. The “domains” used in our current version of DCNs are “domain families” defined by the Pfam database, which sometime could be further divided into sub-units. Using a more specific definition of domains might show different characteristics of DCNs. For example, we plan to use the structural domain definitions in the SCOP database [210] and ProDom [187] in our next study. However, since probably none of the current domain definitions is perfect and there is no unified ways of defining a domain, some noises are not completely avoidable.

In our current work, when predicting the GO terms of a target domain, we treated different GO terms in the neighboring domains independently. However, the semantic similarities between GO terms may be incorporated into the SVM and neighbor counting methods to improve protein function prediction. More advanced graph-kernels and random-walk graph kernels [145, 211, 212] can also be applied to DCNs to predict domain functions in order to take non-immediate neighboring domains into account. Similarly, advanced graph alignment algorithms, such as IsoRankN [194] may be used to align DCNs for phylogeny inference. Furthermore, the quality of the genome assembly and annotation is important to our DCNs-based phylogeny

inference method. It may be interesting to find how annotation quality influences the performance of DCN-based phylogeny inference method by using different versions of annotated genomes.

Table 7.5: Performances of neighbor-counting method on promiscuous domains. Definition of best prediction can be found at the caption of Table 7.2. “N/A” indicates this specific domain does not exist in the DCN of a species. “Average” indicates the average value of degree values and similarity scores. “Arab.” indicates “Arabidopsis”, and “Hum.” indicates “Human”. Some of these promiscuous domains were selected from [198].

Species	Degree		Top 3		Top 1				
	Arabidopsis	Yeast	Human	Arabidopsis	Yeast	Human	Arabidopsis	Yeast	Human
Helicase_C	44	24	55	1.000	1.000	1.000	0.387	1.000	1.000
PDZ	1	N/A	69	0.234	N/A	1.000	0.115	N/A	0.160
Pkinase	44	15	63	1.000	1.000	1.000	1.000	1.000	0.300
Pkinase_Tyr	17	N/A	41	1.000	N/A	1.000	0.246	N/A	0.088
Pkinase_C	1	4	14	1.000	1.000	1.000	1.000	0.171	0.300
PHD	35	11	43	1.000	0.557	1.000	0.096	0.557	0.557
AAA	24	21	21	1.000	1.000	1.000	1.000	1.000	1.000
SET	12	2	20	1.000	0.096	1.000	0.054	0.096	0.080
GATase	7	13	11	0.538	0.262	0.602	0.538	0.220	0.602
Average	20.6	12.9	37.4	0.864	0.702	0.956	0.493	0.578	0.454

Table 7.6: Performances of χ^2 method on promiscuous domains. Definition of the best prediction can be found at the caption of Table 7.2. The degree value of each promiscuous domain can be found at Table 7.5.

Species	Top 3			Top 1		
	Arabidopsis	Yeast	Human	Arabidopsis	Yeast	Human
Helicase_C	1.000	0.696	1.000	1.000	0.696	0.696
PDZ	0.234	N/A	0.482	0.234	N/A	0.284
Pkinase	1.000	1.000	1.000	1.000	1.000	0.573
Pkinase_Tyr	0.670	N/A	1.000	0.670	N/A	1.000
Pkinase_C	1.000	1.000	0.206	0.895	0.895	0.199
PHD	0.669	0.557	0.661	0.000	0.557	0.557
AAA	0.152	0.152	0.152	0.040	0.040	0.040
SET	0.058	0.096	0.488	0.058	0.096	0.058
GATase	0.302	0.602	0.602	0.538	0.602	0.602
Average	0.560	0.584	0.617	0.486	0.552	0.442

Table 7.7: Performances of SVM-based method on promiscuous domains. Definition of the best prediction can be found at the caption of Table 7.2. The degree value of each domain can be found at Table 7.5.

Species	Top 3			Top 1		
	Arabidopsis	Yeast	Human	Arabidopsis	Yeast	Human
Helicase_C	1.000	1.000	1.000	1.000	0.527	1.000
PDZ	0.000	N/A	0.226	0.120	N/A	0.226
Pkinase	0.202	0.171	0.181	1.000	0.037	0.173
Pkinase_Tyr	0.662	N/A	0.132	0.662	N/A	0.088
Pkinase_C	1.000	1.000	1.000	1.000	0.171	0.149
PHD	1.000	0.481	0.557	0.096	0.481	0.096
AAA	1.000	0.081	1.000	1.000	0.039	1.000
SET	1.000	0.485	1.000	1.000	0.000	1.000
GATase	0.337	0.527	0.602	0.079	0.527	0.602
Average	0.689	0.535	0.633	0.662	0.255	0.482

Table 7.8: The leave-one-out cross-validation results of the SVM-based enzyme “yes or no” predictions. “Ratio” standards for the percentage of correctly predicted domains in the cross-validation. The feature used are only GO term frequencies gained from radius one neighboring domains. “Pos.” and “Neg.” indicate the number of positive and negative examples.

Species	Pos.	Neg.	<i>TP</i>	<i>FN</i>	<i>TN</i>	<i>FP</i>	Q_p	Q_n	<i>MCC</i>	Ratio
Arabidopsis	309	432	205	104	374	104	0.66	0.87	0.54	0.78
Yeast	220	300	40	75	260	40	0.66	0.87	0.54	0.78
Human	312	646	185	127	593	53	0.59	0.91	0.55	0.81

Table 7.9: The accuracies of using “neighbor-inference” method for enzyme “yes or no” predictions. The “Ratio” indicates the percentage of correctly predicted domains. “Pos.” and “Neg.” indicate the number of positive and negative examples.

Species	Pos.	Neg.	<i>TP</i>	<i>FN</i>	<i>TN</i>	<i>FP</i>	Q_p	Q_n	<i>MCC</i>	Ratio
Arabidopsis	309	432	240	69	328	104	0.78	0.76	0.53	0.77
Yeast	220	300	177	43	231	69	0.80	0.77	0.56	0.78
Human	312	646	242	70	497	149	0.78	0.77	0.52	0.77

Table 7.10: Prediction accuracy of the neighbor-counting and SVM-base method when predicting EC families. For SVM, it reports the accuracy of a leave-one-domain-out cross-validation. Experiments were performed on Arabidopsis DCNs. “Known EC” indicates the number of known EC numbers occurrences in the radius one neighboring domains. “Target Number” indicates the number of target domains. The features used in SVM-based method are the GO and EC number occurrence-frequencies, with the linear kernel function.

Known EC	SVM	Neighbor-Counting		Target Domain Number
		Top 3	Top 1	
>= 0	0.803	0.711	0.674	307
>= 1	0.924	0.924	0.919	237
>= 2	1.000	1.000	0.989	89
>= 3	1.000	1.000	1.000	43
>= 4	1.000	1.000	1.000	21
>= 5	1.000	1.000	1.000	19
>= 6	1.000	1.000	1.000	17
>= 7	1.000	1.000	1.000	6

Table 7.11: The results of DCN-based aggregated neighbor-counting method on 66 randomly selected proteins. The ways of calculating precision and recall can be found at the captions of Table 7.4 and Table 7.3, respectively. Explanations of the best semantic similarity score can be found at the caption of Table 7.2. From the 100 proteins randomly selected from GO database, 66 proteins have predictions available by DCN-based aggregated neighbor-counting method.

Number of top predictions	Avg. Precision	Avg. Recall	Avg. similarity score of best predictions
1	36.4%	6.5%	0.600
2	33.3%	11.5%	0.724
3	30.3%	18.8%	0.805
4	26.9%	21.9%	0.855
5	23.9%	23.6%	0.874
6	23.2%	27.8%	0.897
7	21.2%	29.2%	0.913
8	19.9%	31.1%	0.913
9	18.6%	32.1%	0.913
10	17.1%	32.8%	0.913

Table 7.12: The results of DCN-based aggregated neighbor-counting method on 9 single-domain proteins.

Number of top predictions	Avg. Precision	Avg. Recall	Avg. similarity score of best predictions
1	33.3%	8.4%	0.636
2	27.8%	14.3%	0.799
3	22.2%	15.6%	0.834
4	27.8%	21.6%	0.960
5	22.2%	21.6%	0.960
6	20.4%	25.4%	0.960
7	17.5%	25.4%	0.960
8	16.7%	27.6%	0.960
9	16.0%	31.3%	0.960
10	15.6%	32.2%	0.960

Table 7.13: Accuracy comparisons between our DCNs-based method and other methods for phylogeny inference. OG = Overlapping Gene Distance [207]; CVT = Composition Vector Tree [208], and k is the length of string; CCV = Complete Composition Vector [209]; PDUG = Structural Protein Domain Universe Graph [199]; and CD = Composite Distance [198]. The accuracy of OG, CVT, SDD, and CD are directly retrieved from [198]. Accuracies are reflected by the percentage of the agreed quartets.

Dataset	Species Num.	OG	CVT ($k = 5$)	CCV ($k \leq 5$)	PDUG	CD	DCNs
1	52	83.93	88.29	87.82	N/A	90.29	85.71
2	53	85.49	87.92	86.27	N/A	90.74	87.00
3	398	85.52	78.86	79.03	N/A	90.07	76.71
4	181	80.34	87.19	87.19	N/A	98.30	82.75
5	277	81.89	83.19	83.28	N/A	90.71	80.42
6	54	88.27	91.47	91.39	81.57	96.55	93.45

Table 7.14: The composition of dataset 6 containing 54 single-chromosome prokaryotic organisms.

Division	Bergey's code	Number of species
Bacteria <i>Aquificae</i>	B1	1
Bacteria <i>Fusobacteria</i>	B21	1
Bacteria <i>Thermotogae</i>	B20	1
Bacteria <i>Cyanobacteria</i>	B10	2
Bacteria <i>Actinobacteria</i>	B14	5
Bacteria <i>Firmicutes</i>	B13	15
Bacteria <i>Proteobacteria</i>	B12	21
Archaea	A2	8

Appendix A

Supplementary Documents for Chapter 2

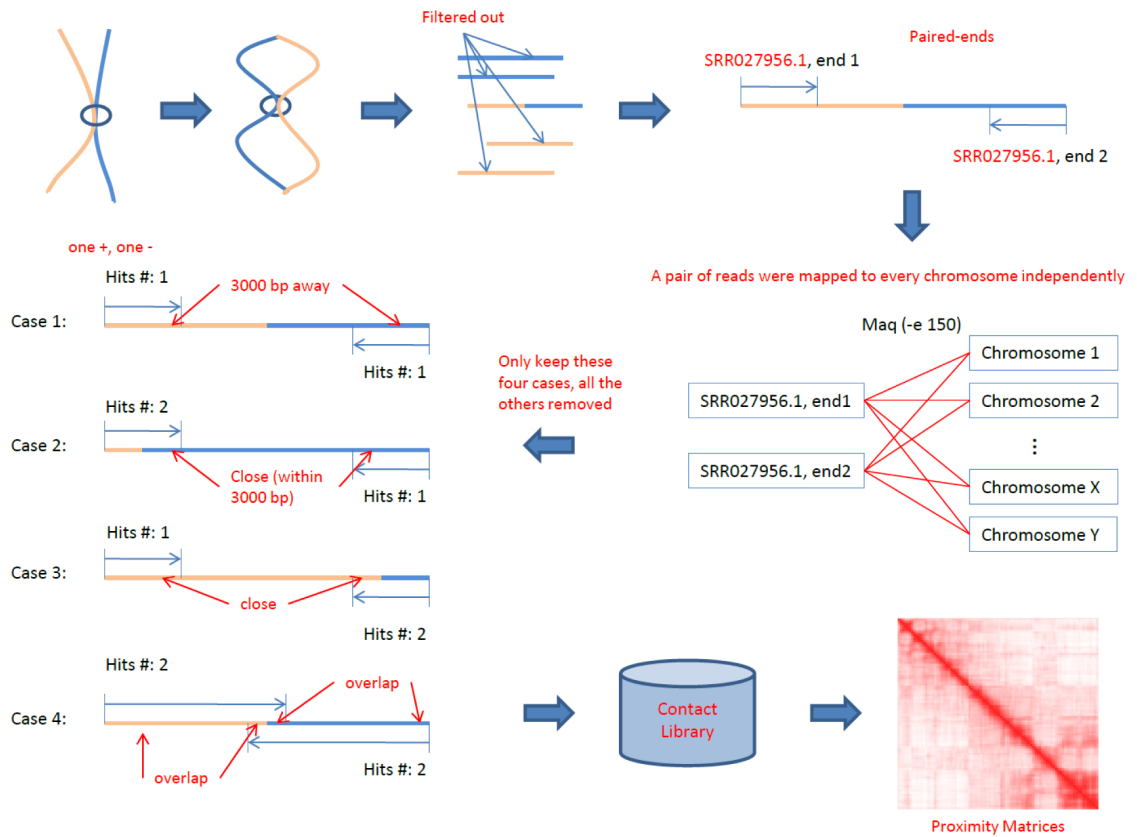


Figure A.1: An overview of the bioinformatics pipeline of analyzing Hi-C experimental reads data.

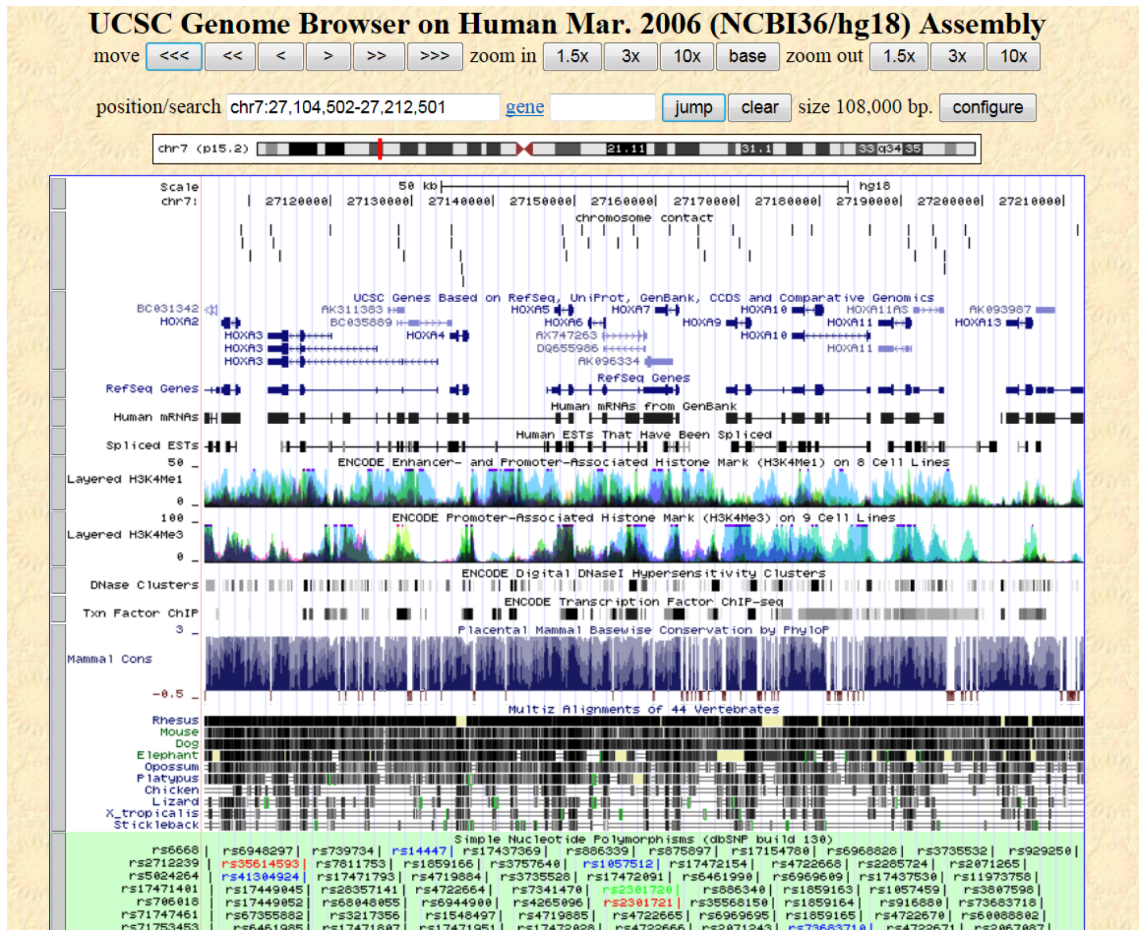


Figure A.2: The visualization of reads mapped to the HoxA gene region (27,104,502 - 27,212,501) on chromosome 7 of the human genome by the UCSC genome browser. The vertical line segments under the label “chromosome contact” denote the locations where the reads were mapped to. The reads data of the MHH-CALL-4 cell line was used.

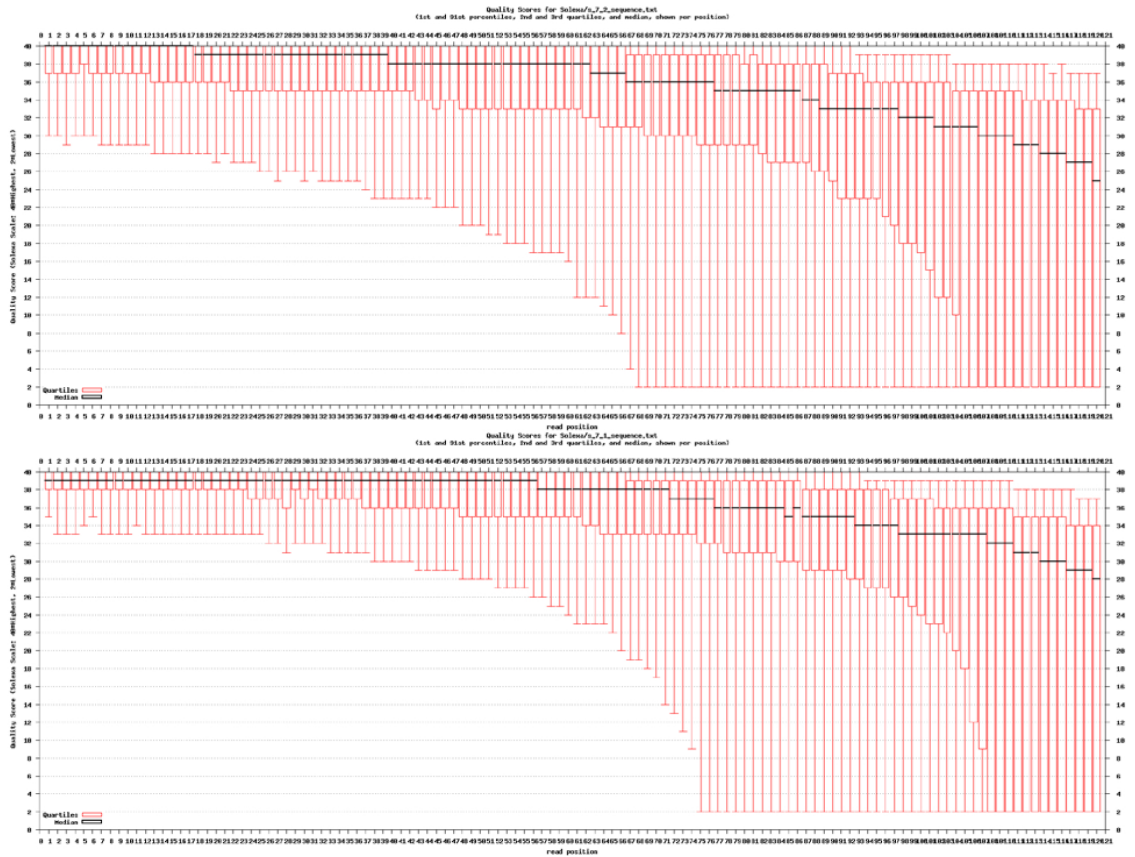


Figure A.3: The distribution of the sequencing qualities (Solexa-scale) of paired-end reads of the two malignant primary ALL B-cell data sets (i.e. quality scores versus nucleotide positions). The sequencing quality score at a position is calculated as $Q_{Solexa} = -10 \log_{10} \frac{p}{1-p}$, where p is the probability of a sequencing error at the position. A score 30 means the probability of a sequencing error at the position is ~ 0.001 . A score 20 or above may be considered acceptable. The plots show the median (the black curve), 1st and 91st percentiles, 2nd and 3rd quartiles from positions 1 to 120 in the reads data.

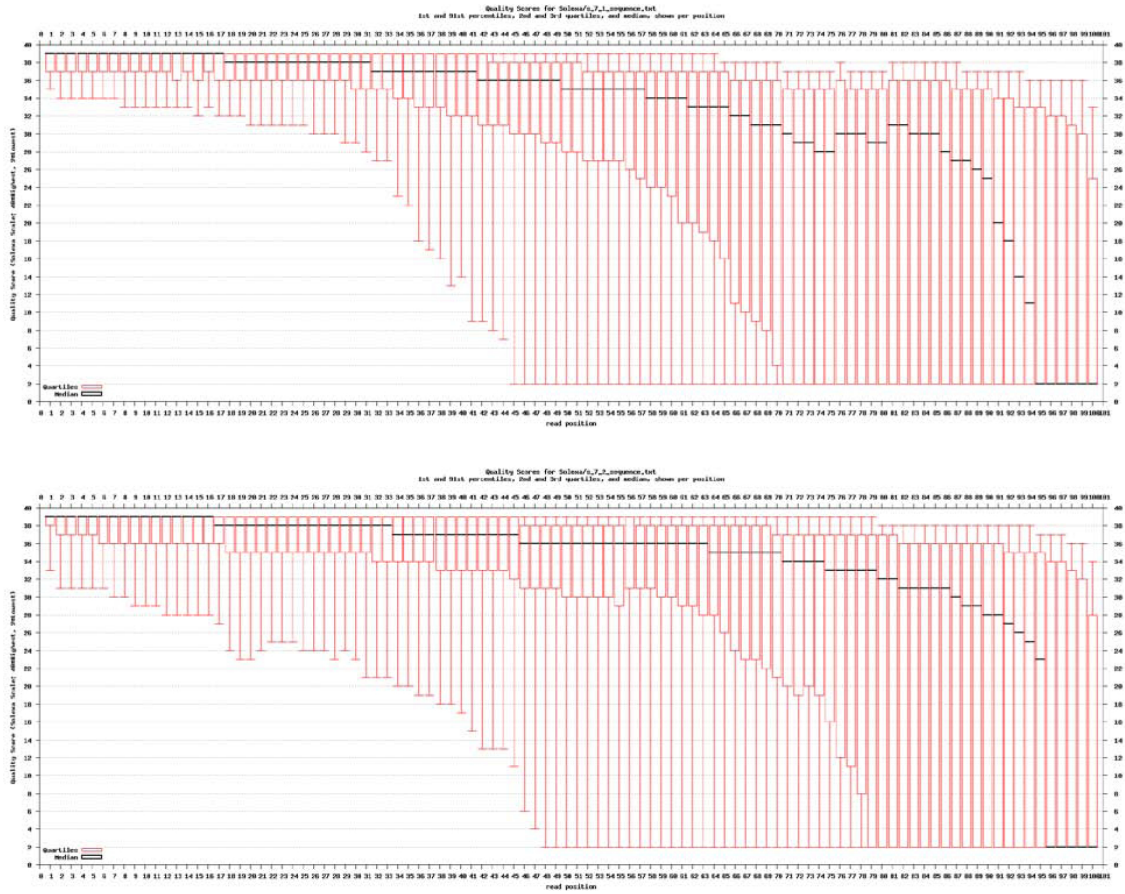


Figure A.4: The distribution of the sequencing qualities of paired-end reads of the two malignant MHH-CALL-4 cell line data sets.

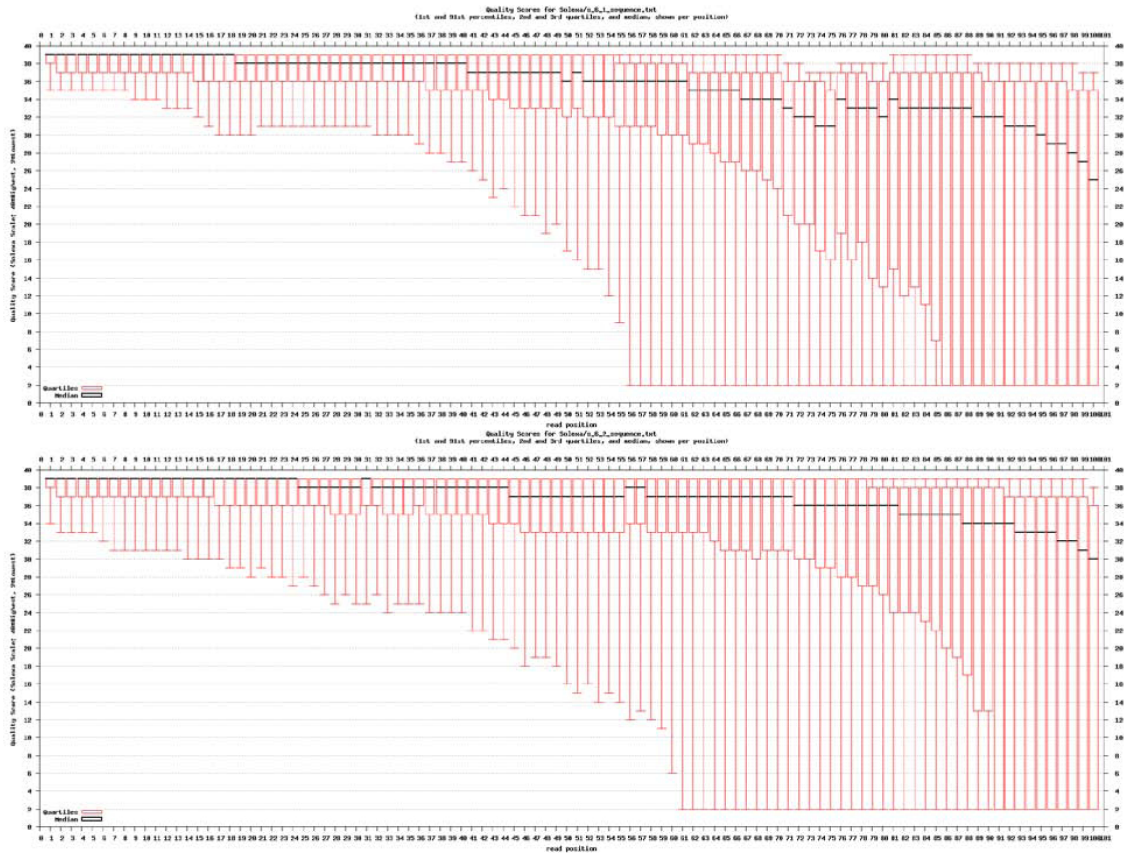


Figure A.5: The distribution of the sequencing qualities of paired-end reads of the two malignant lymphoma RL cell line data sets.

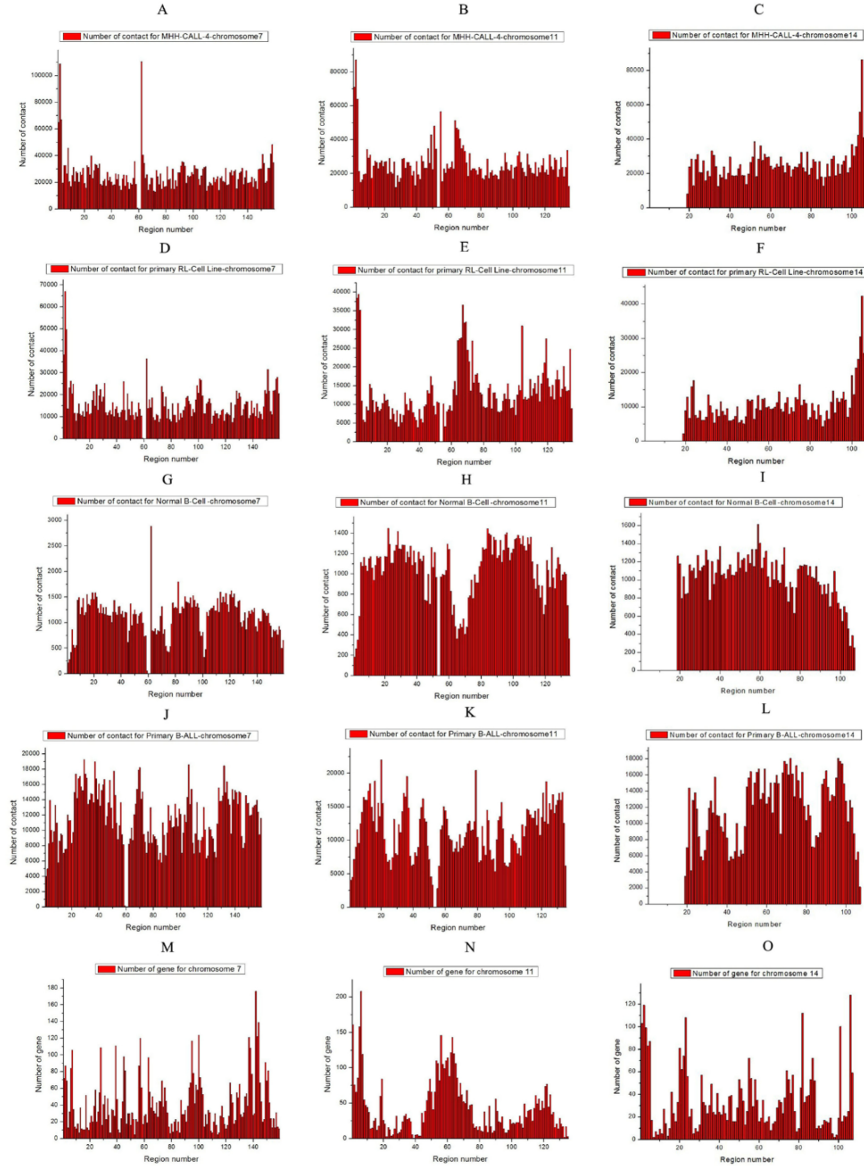


Figure A.6: The plots of contact numbers against regions of chromosome 7, 11 and 14 of four cell samples and the plots of gene numbers against regions of chromosome 7, 11 and 14. The X-axis in Plots A-L denotes chromosomal region index at resolution 1Mb and the Y-axis denotes the number of intra- and inter-chromosomal contacts in each region. An inter-chromosomal contact is a spatial contact between two different chromosomes, and an intra-chromosomal contact a contact within the same chromosome. A, B and C are the plots of chromosomes 7, 11, and 14 for the MHH-CALL-4 cell line respectively, D, E and F for the RL cell line, G, H and I for the normal B-Cell, and J, K and L for the Primary B-ALL cell. The plots show that the number of contacts generated from the sequence data is not evenly distributed along the chromosomes. The extra M, N and O plots show the number of genes in each region against the regions of chromosome 7, 11 and 14 separately.

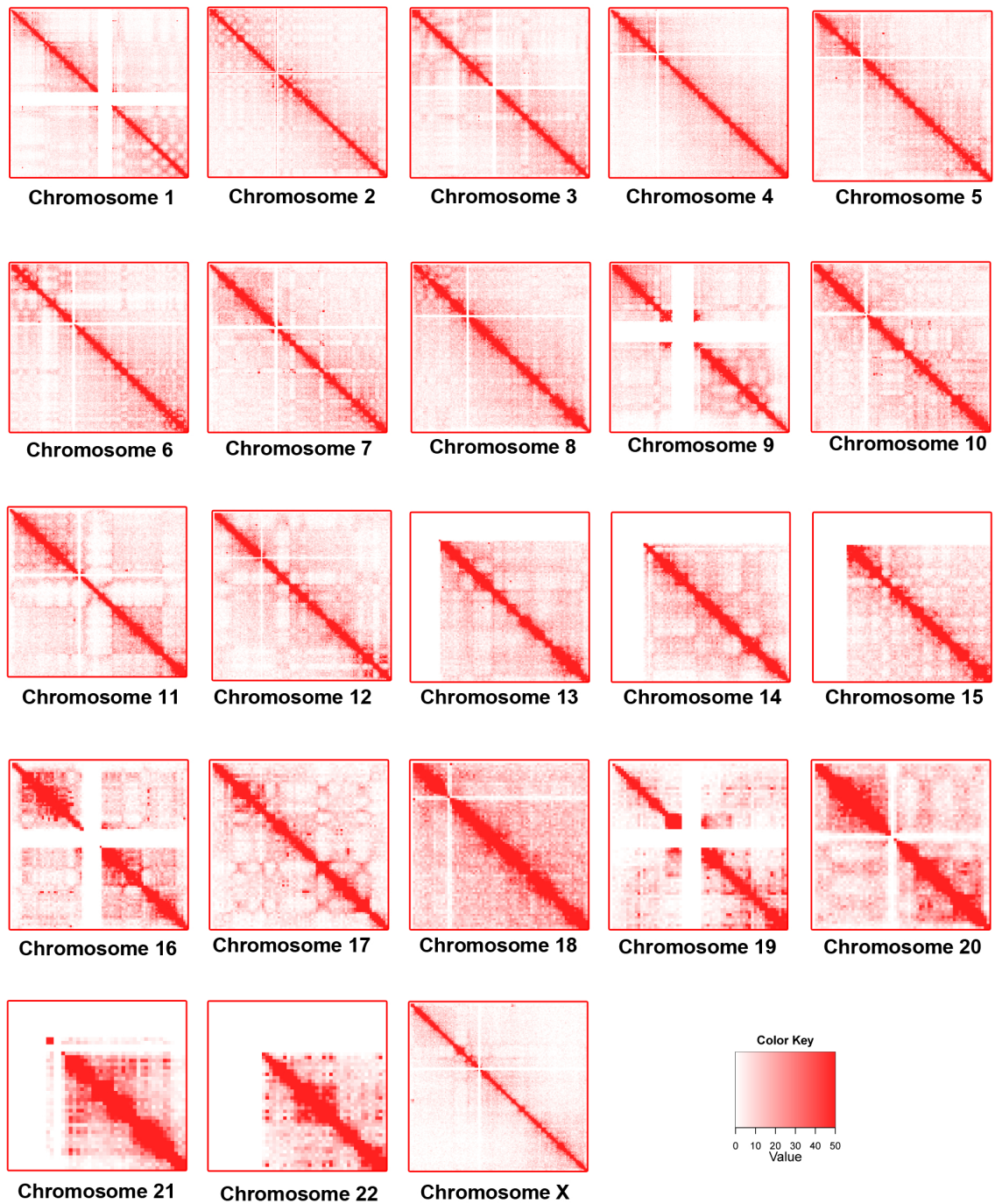


Figure A.7: The intra-chromosomal contact heat maps for all chromosomes of the primary ALL B-cell. Interested readers may contact us for images with higher resolution and for contact matrix data.

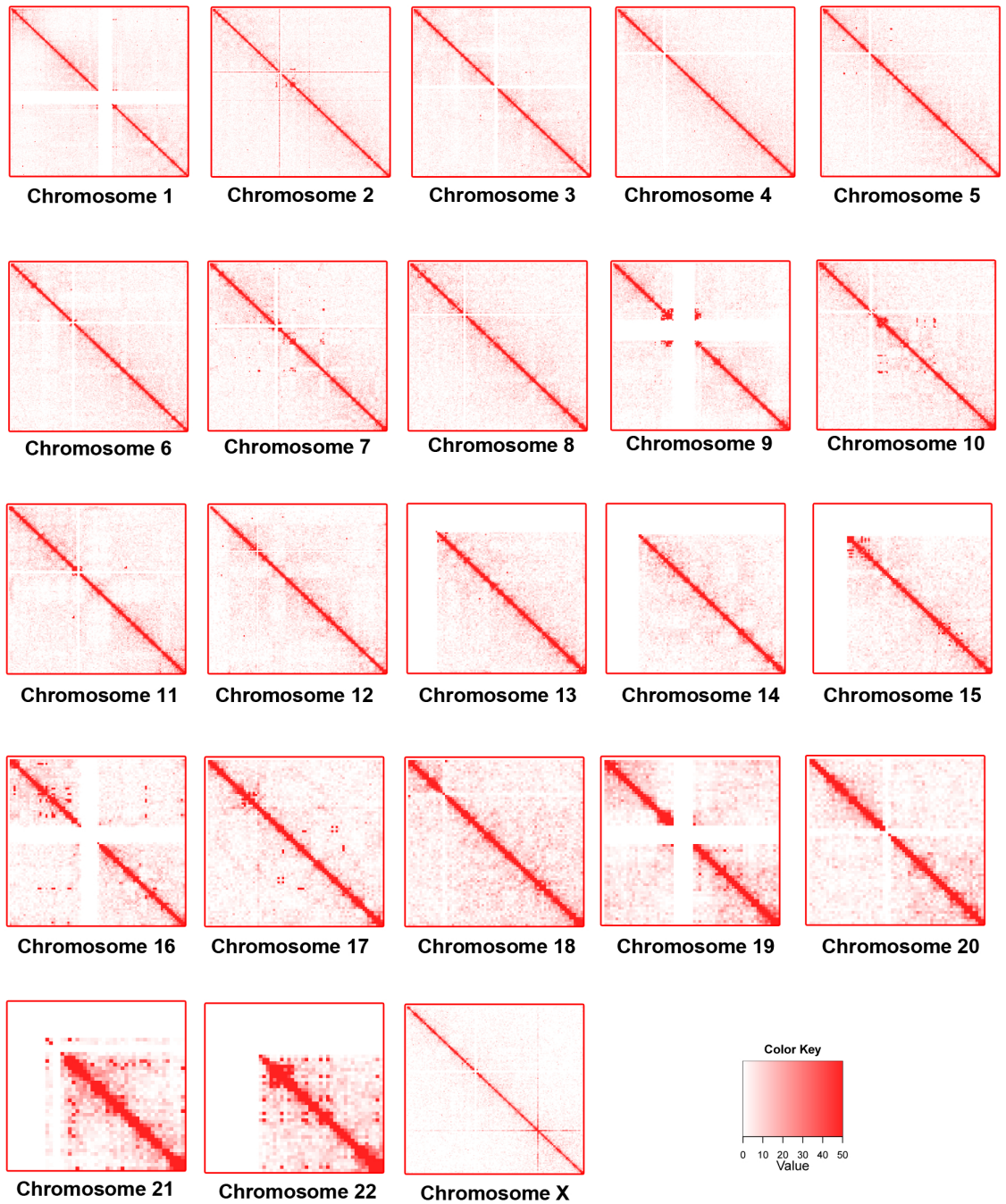


Figure A.8: The intra-chromosomal contact heat maps for all chromosomes of the MHH-CALL-4 cell line. Interested readers may contact us for images with higher resolution and for contact matrix data.

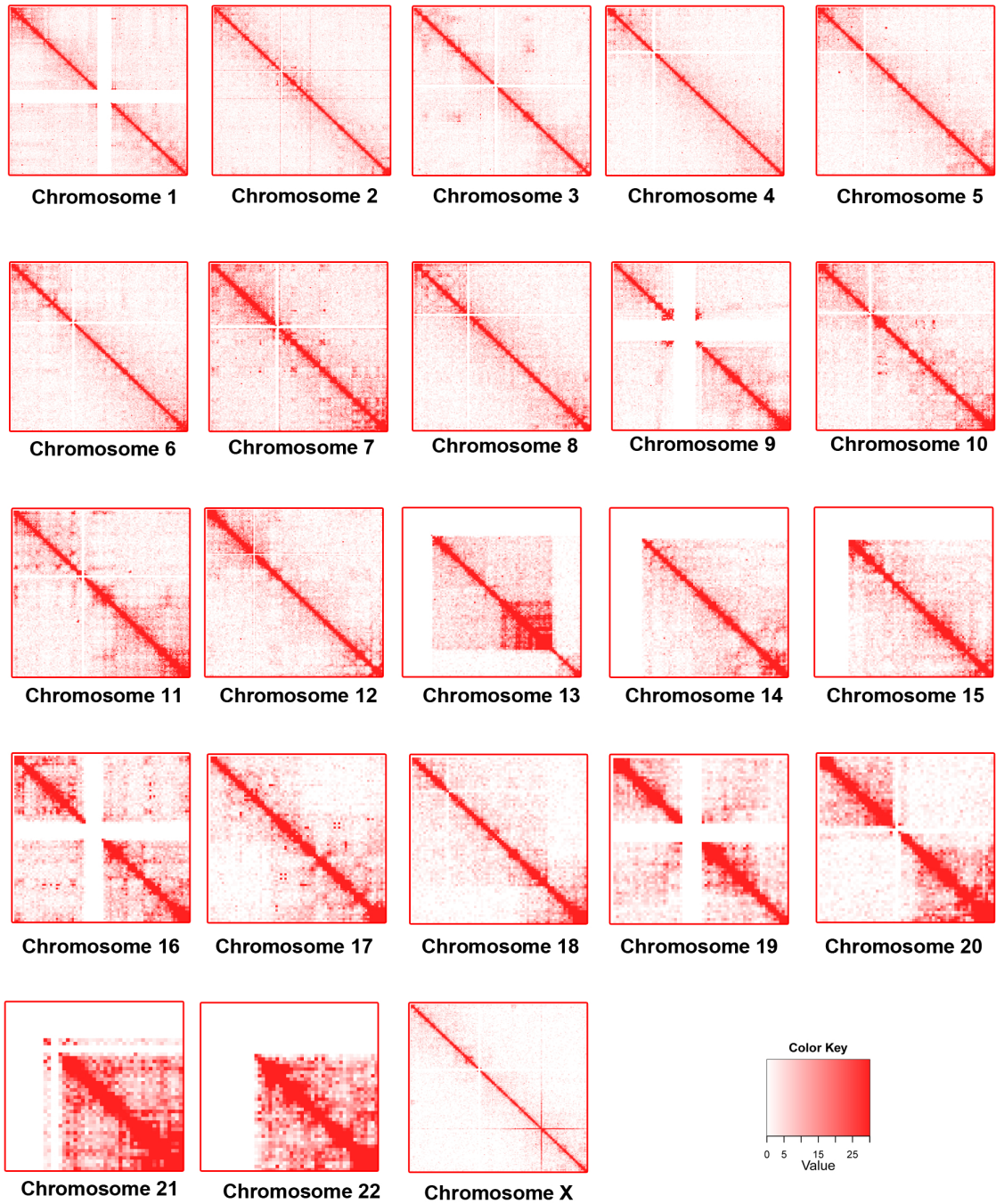


Figure A.9: The intra-chromosomal contact heat maps for all chromosomes of the RL cell line. Interested readers may contact us for images with higher resolution and for contact matrix data.

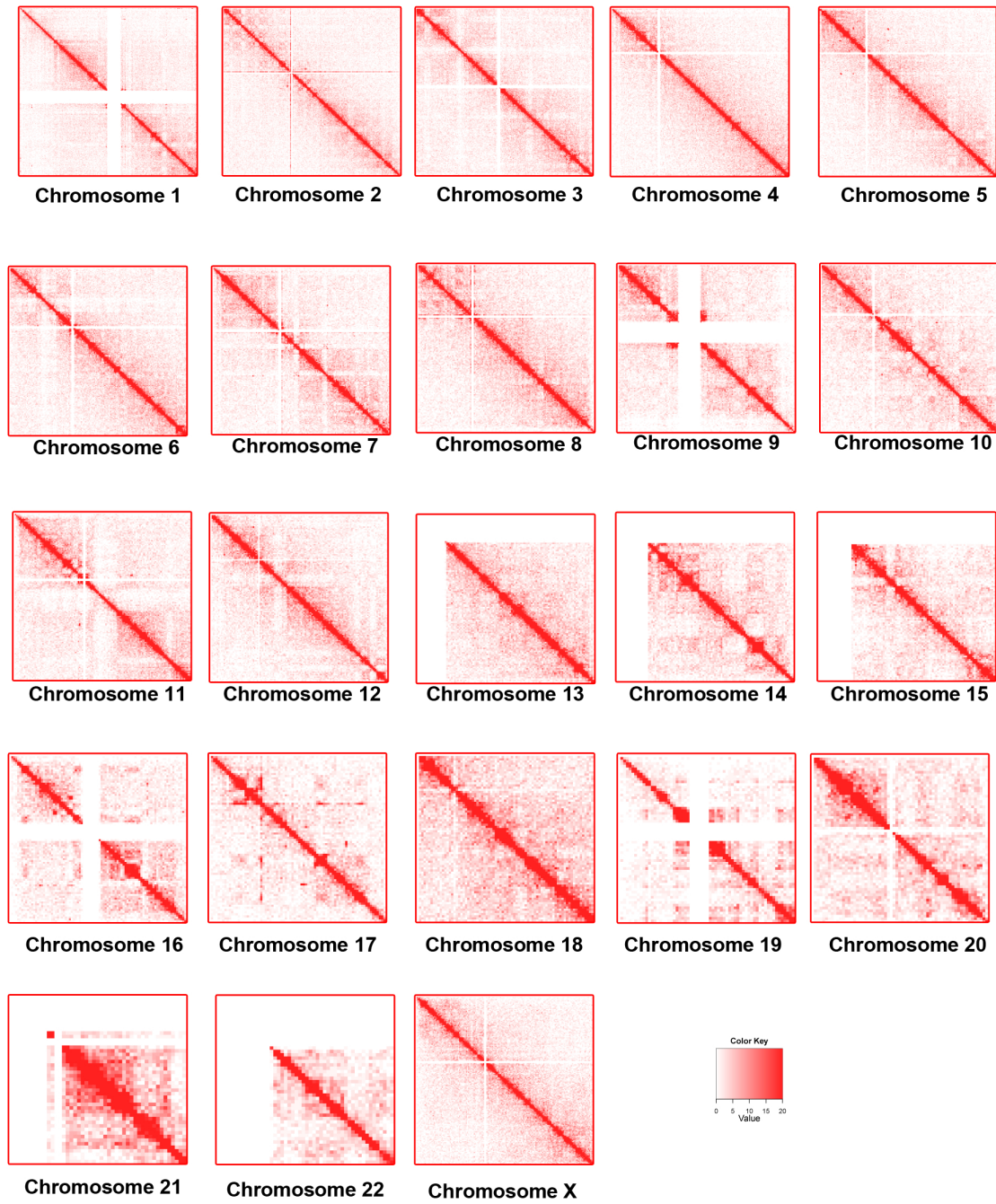


Figure A.10: The intra-chromosomal contact heat maps for all chromosomes for the normal B-cell line. Sequence reads data were downloaded from Lieberman-Aiden etc [1]. Mapping and construction of contact maps were carried out by our pipeline.

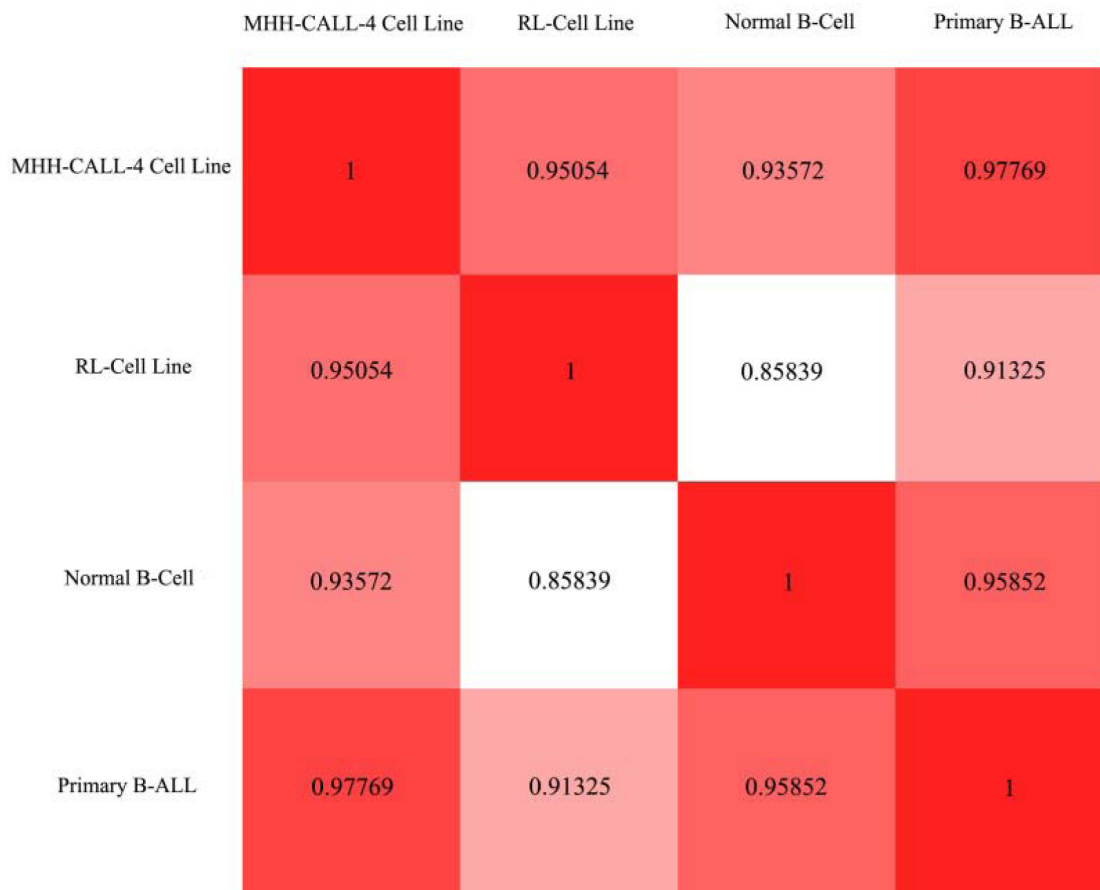


Figure A.11: The Pearson's correlation matrix for intra-chromosomal contact numbers between the normal B cell, primary ALL B-cell, MHH-CALL-4 cell line, and RL cell line. For each cell, the number of intra-chromosomal contacts for each of 23 pairs of chromosomes was calculated and was put into a vector. Thus, each cell sample has one intra-chromosomal contact vector. The matrix below shows the Pearson's correlation between each pairs of vectors of two cell samples.

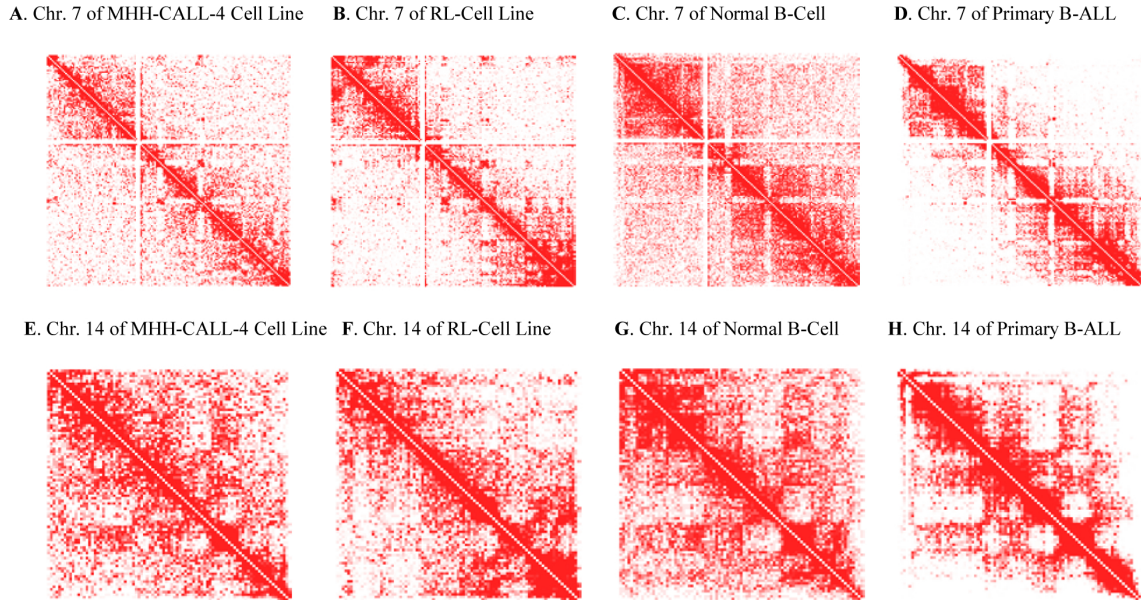


Figure A.12: Contact significance analysis of selected chromosomes. In order to check if the number of contacts between two specific chromosome regions is significantly large, we calculated the significance score (i.e. the probability of receiving this number of contacts or more) in each cell of an intra-chromosomal contact matrix at 1Mb resolution, assuming the background distribution of contact numbers follows the Poisson distribution. The parameter (λ : mean contact number) of the background distribution was set to the average of number of contacts in the matrix excluding contacts within the same region (i.e. diagonal line in a matrix). Sub-figures A, B, C and D illustrate the contact significant scores of the intra-chromosomal contact matrices of chromosome 7 of the MHH-CALL-4 cell line, RL cell line, normal B-cell and the Primary B-ALL cell, respectively. Sub-figures E, F, G and H depict the significance scores of intra-chromosomal contact matrices of chromosome 14 of the MHH-CALL-4 cell line, RL-cell line, normal B-cell line and the primary B-ALL cell, respectively. Darker red indicates higher significant score.

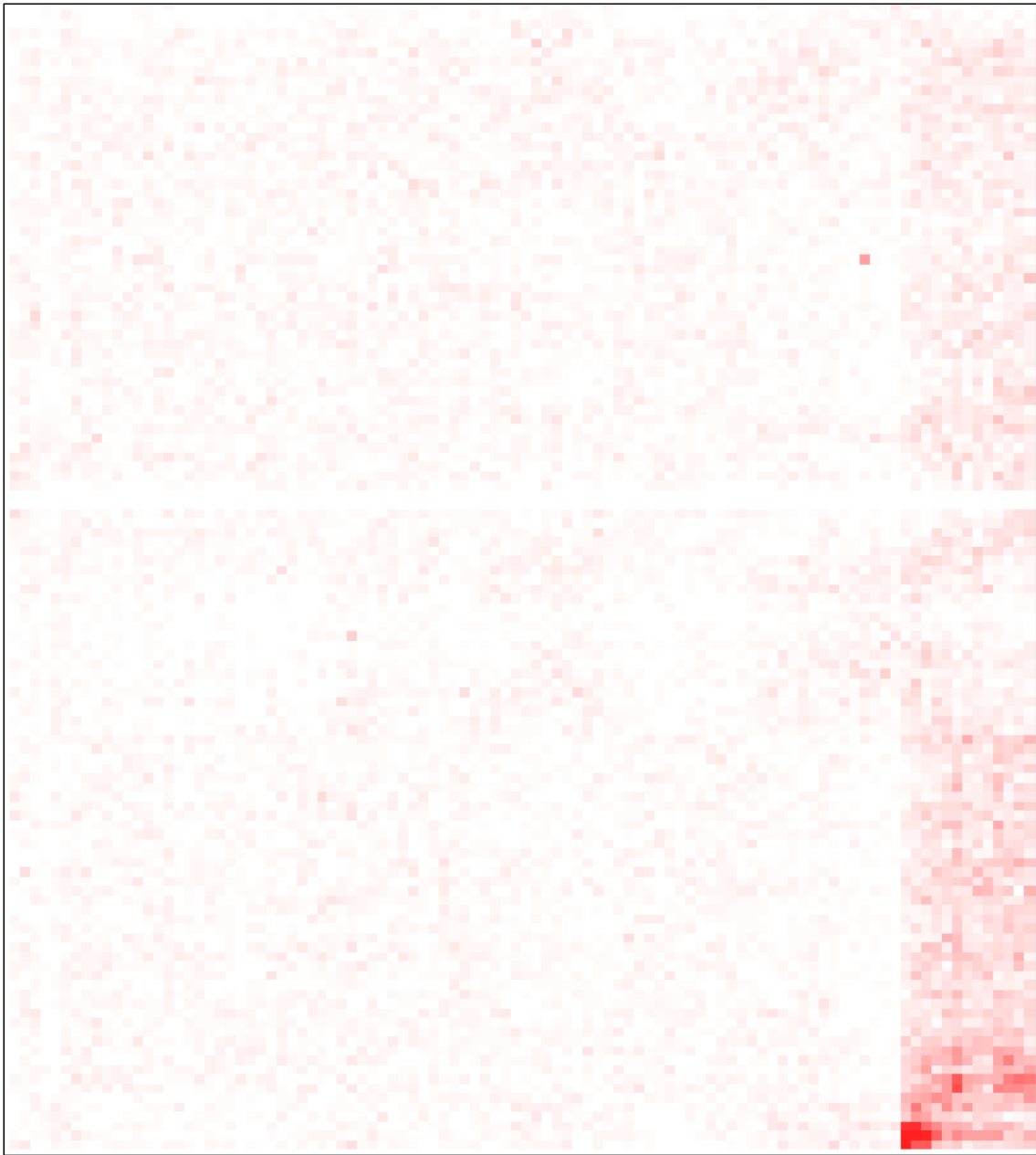
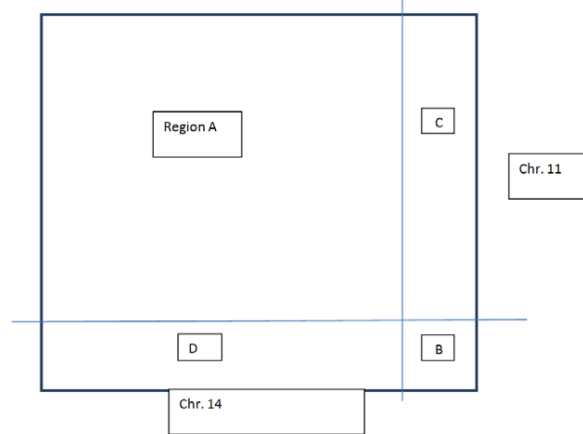


Figure A.13: The corrected inter-chromosomal contact map between translocated chromosomes 11 and 14 for the primary ALL B-cell. The method of calculating it can be found in Figure A.15.

A. Reconstructed contact matrix



B. Cartoon representation of translocation between chromosomes 11 and 14

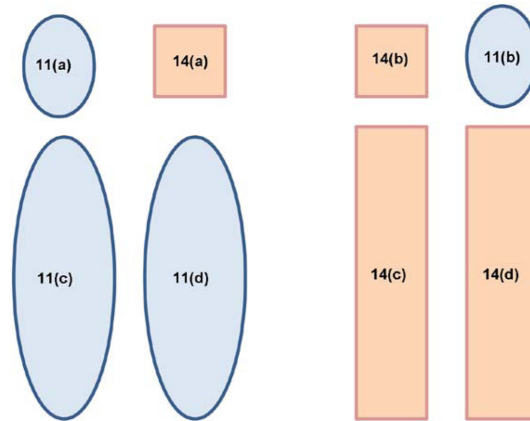


Figure A.14: The method of calculating the corrected inter-chromosomal contact matrix of translocated chromosomes. (A) Division of the corrected inter-chromosomal contact matrix between chromosomes 11 and 14 into three regions to be reconstructed separately. Region A contains the contacts between non-translocated segments in chromosome 11 (i.e. 11(c) and 11(d) in (B)) and non-translocated segments in chromosome 14 (i.e. 14(c) and 14(d) in (B)). Region B contains the contacts between translocated segments in chromosome 11 (i.e. 11(b) in (B)) and translocated segments in chromosome 14 (i.e. 14(a) in (B)). Region C contains the contacts between non-translocated segments in chromosome 11 (i.e. 11(c) and 11(d) in (B)) and translocated segments in chromosome 14 (i.e. 14(a) in (B)). Region D contains the contacts between non-translocated segments in chromosome 14 and translocation segment in chromosome 11. For the contacts in regions A and B, we divided the original contact numbers by 2 in order to estimate the inter-chromosome contacts. For region C, we normalized the value of each cell $C_{ij} = \max(0, C_{ij} - \text{average num of row } i \text{ in region A})$. For region D, we normalized the value of each cell $D_{ij} = \max(0, D_{ij} - \text{average num of column } j \text{ in region A})$.

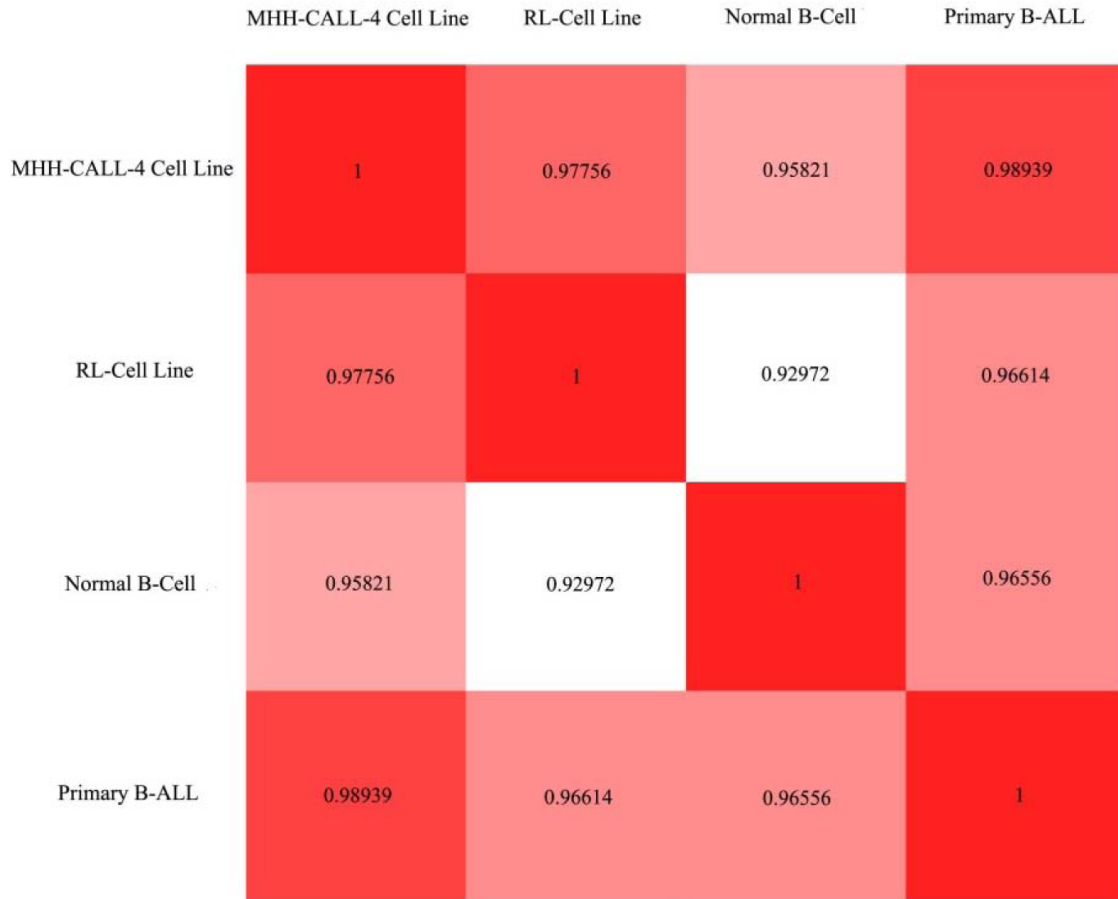


Figure A.15: The Pearson's correlation matrix for inter-chromosomal contact numbers between the normal B cell, primary ALL B-cell, MHH-CALL-4 cell line, and RL cell line. For each cell, the number of inter-chromosomal contacts between chromosomes were calculated and put into a vector. Thus, each cell has one vector to represent all its inter-chromosomal contact numbers. The matrix below shows the Pearson's correlation between each pairs of vectors of two cell samples.

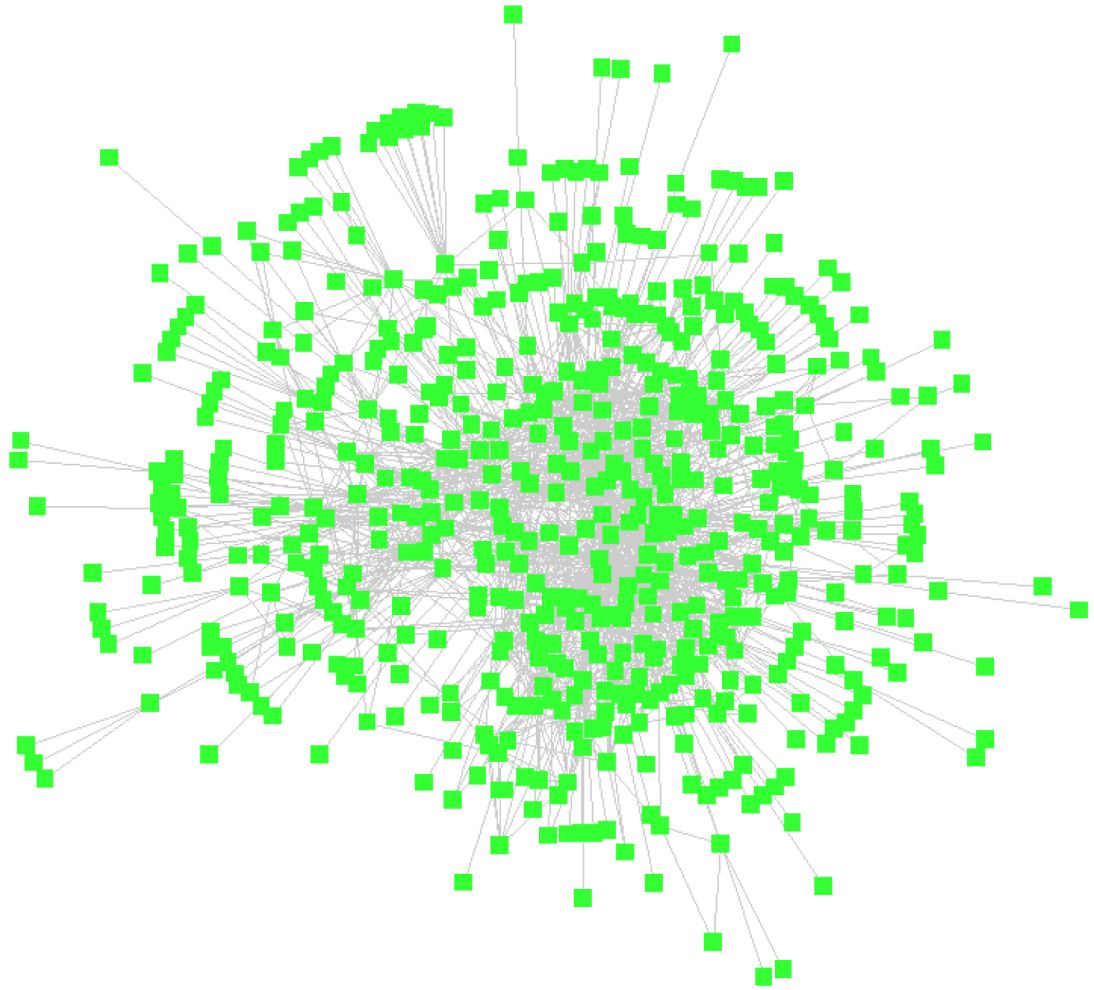


Figure A.16: The interaction network between transcription factor binding sites (TBSs) in the entire genome of the MHH-CALL-4 cell line.

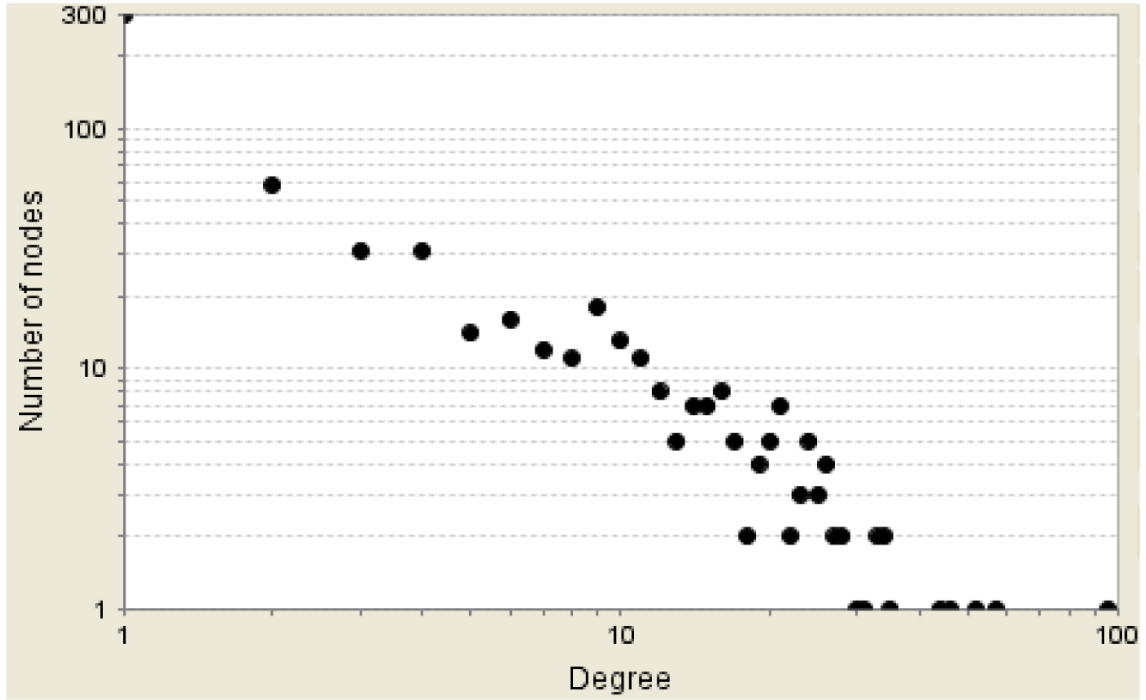


Figure A.17: The distribution of node degree of the TBS-TBS interaction network shown in Figure A.16.

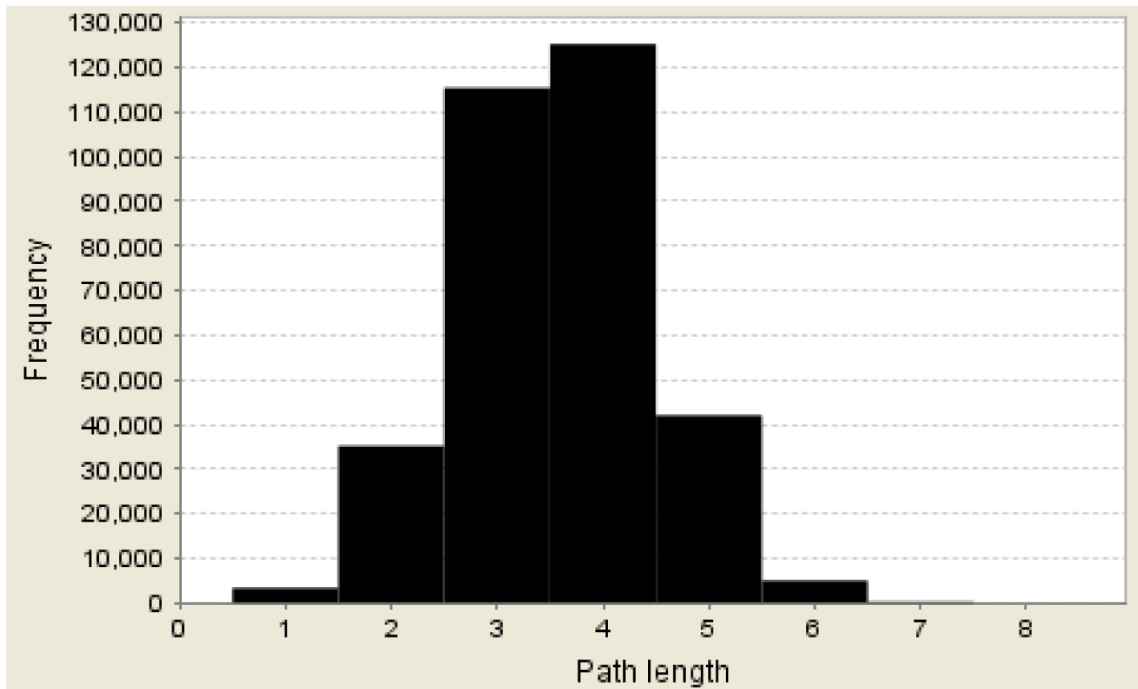


Figure A.18: The histogram of lengths of the shortest paths between any two nodes in the TBS-TBS interaction network shown in Figure A.16.

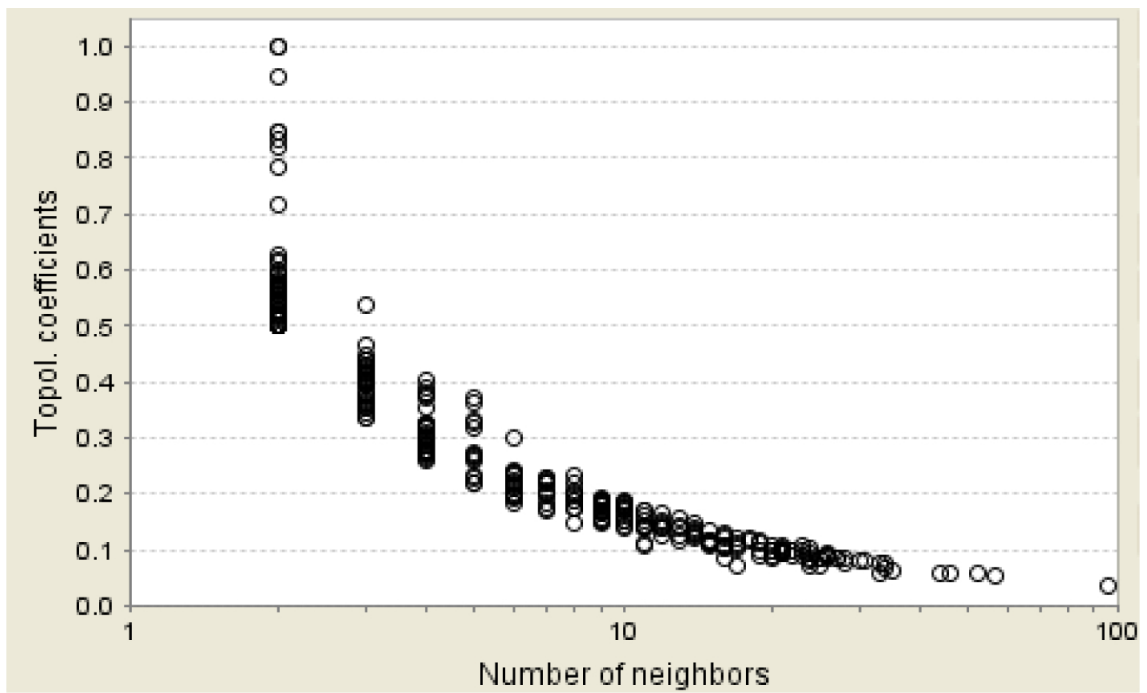


Figure A.19: The distribution of topological coefficients the TBS-TBS interaction network shown in Figure A.16.

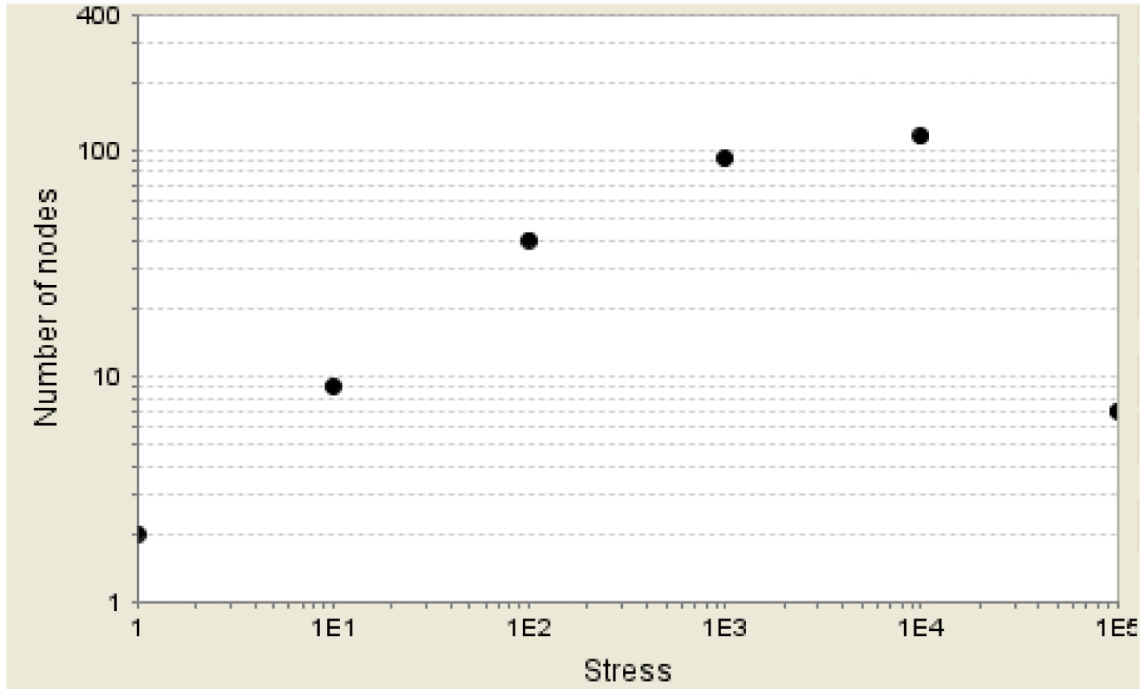


Figure A.20: The distribution of node stresses of the TBS-TBS interaction network shown in Figure A.16.

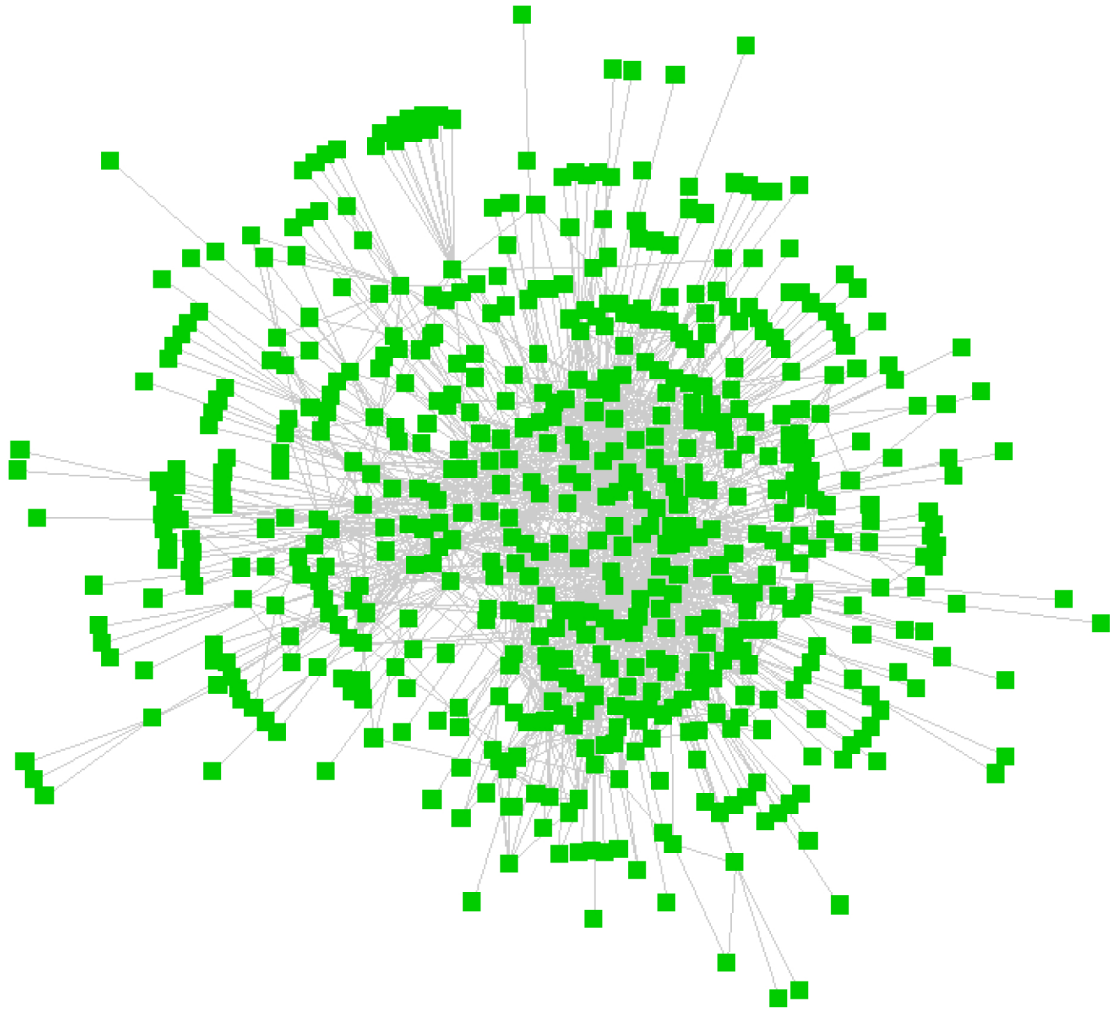


Figure A.21: The spatial interaction networks between genes and transcription factor binding sites (TFB) in chromosome 14 for the CALL-4 cell line. A node in the network denotes a gene or a TFB. Two nodes are connected by an edge if they are spatially contacted.

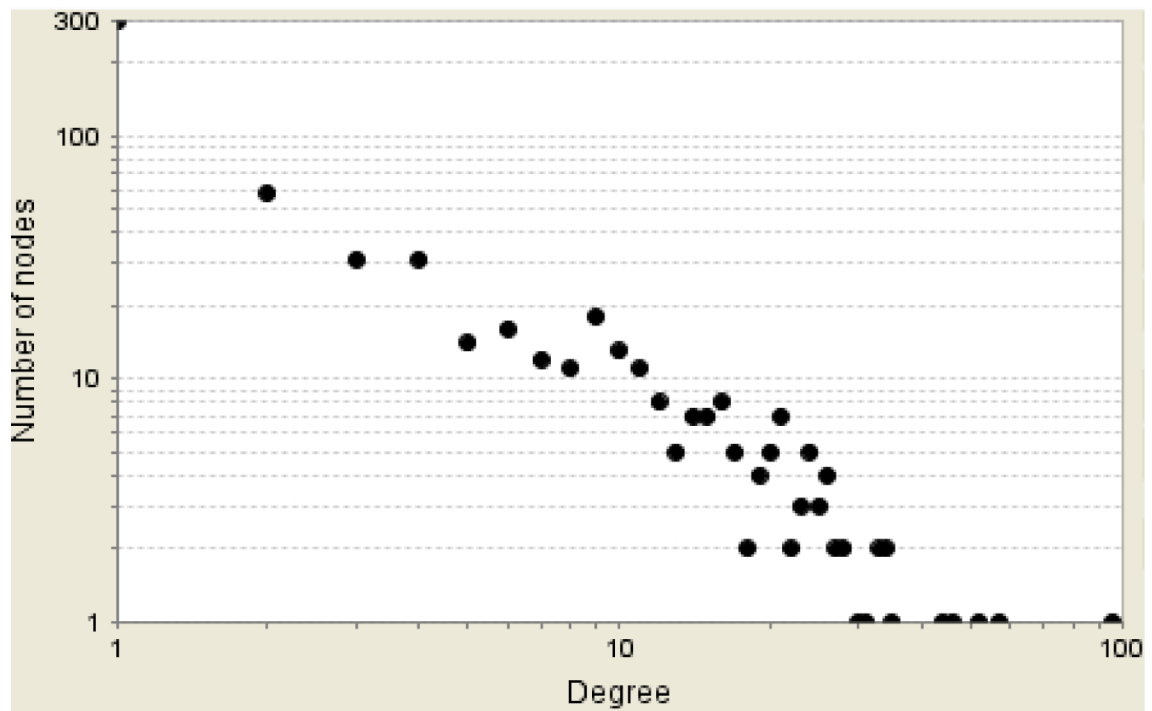


Figure A.22: The node degree distribution of the network shown in Figure A.21. It is shown that the frequency (number of nodes) is largely linear to the degree of the nodes on the log-log scale. This suggests that the network is likely a scale-free network.

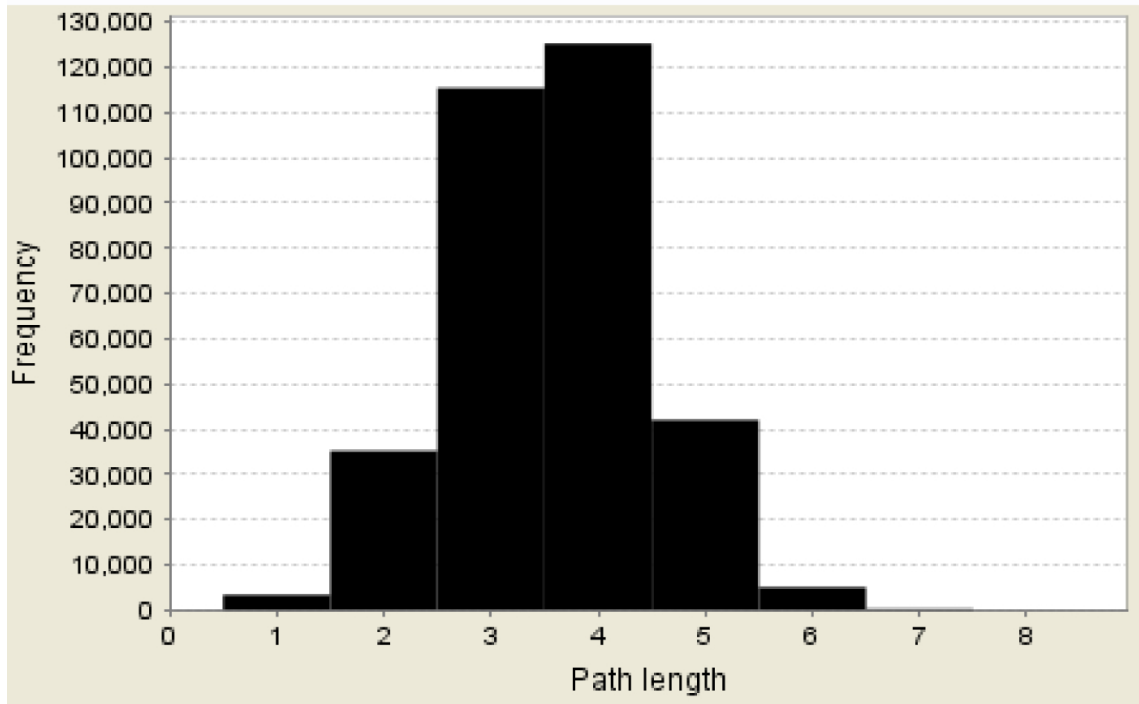


Figure A.23: The histogram of lengths of the shortest path between any two nodes in the network shown in Figure A.21.

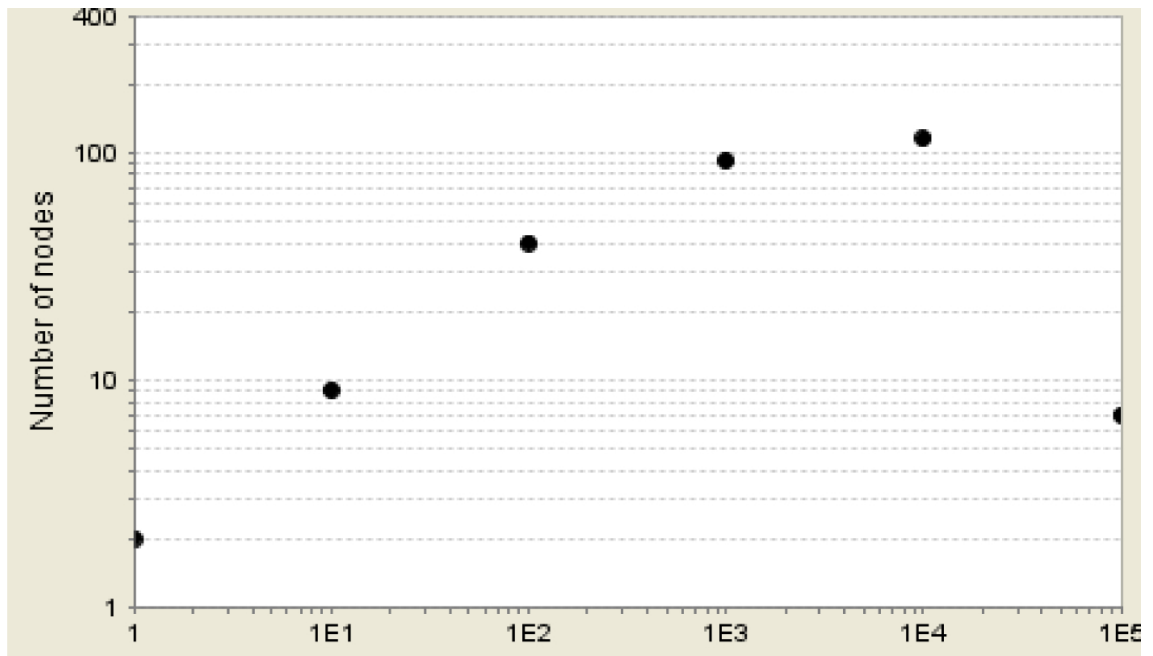


Figure A.24: The distribution of stress values of the network shown in Figure A.21.

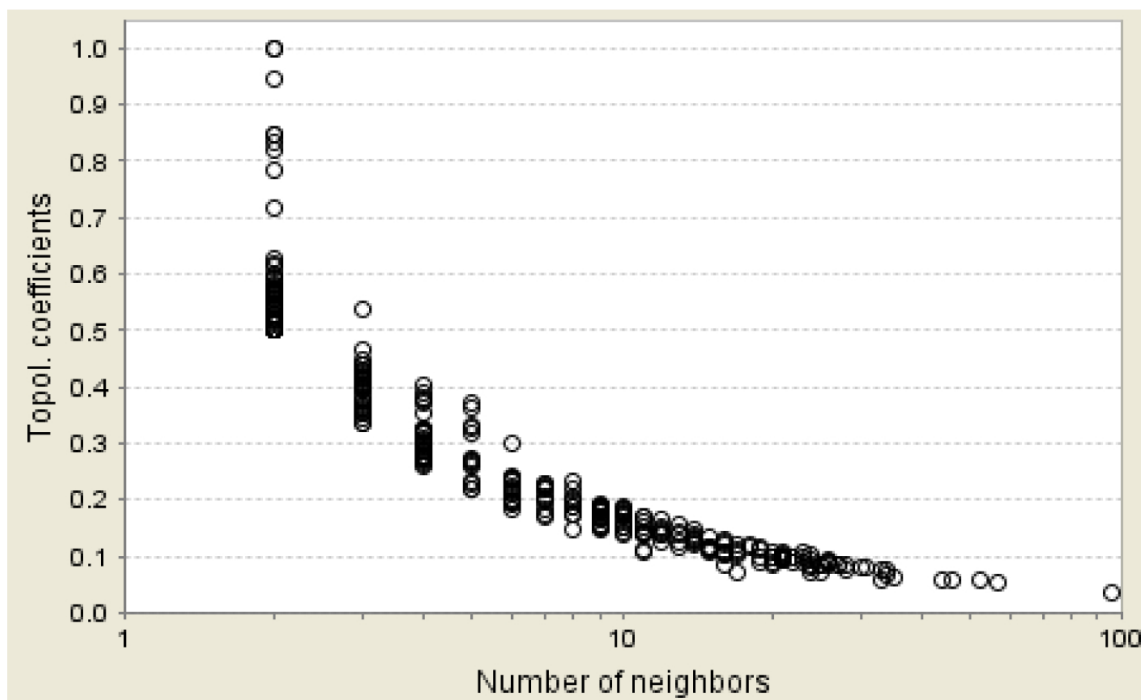
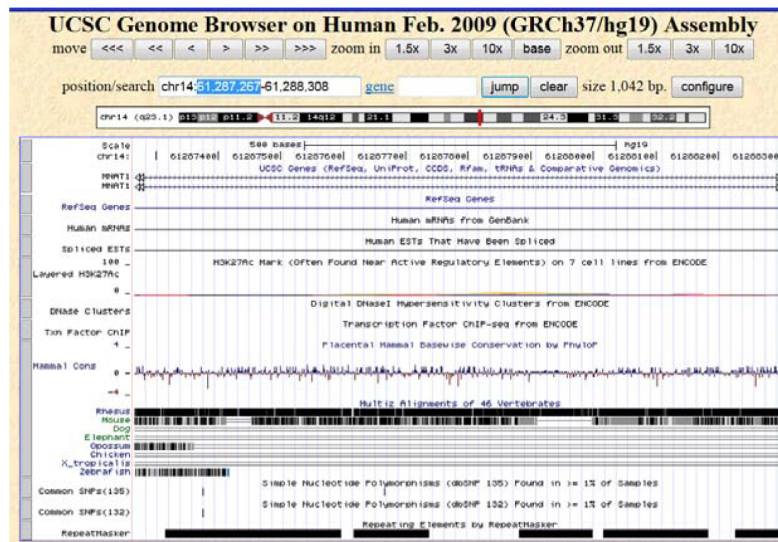


Figure A.25: The distribution of topological coefficients of the network shown in Figure A.21.

- A. The chromosomal region of the transcription factor binding site having most contacts chromosome 14 of the MHH-CALL-4 cell line



- B. Chromosomal region of the gene (GeneID:145508) on chromosome 14 of the MHH-CALL-4.

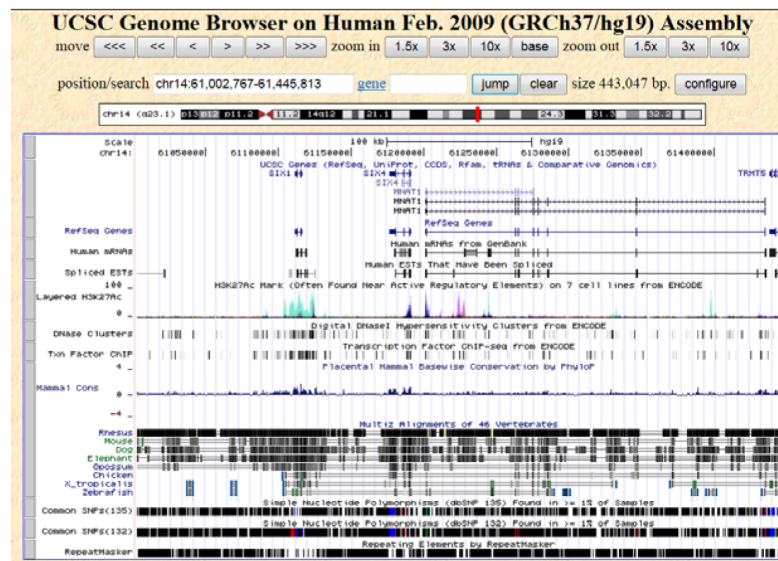


Figure A.26: (A) The chromosomal region of the transcription factor binding site on chromosome 14 of the MHH-CALL-4 cell line that has the highest contacts with other genes is visualized by the UCSC genome browser. This transcription factor binding site contacted 1460 times with GeneID:145508 (starting from the position 61002767 and ending at 61445813), 1 time with GeneID:7253, 2 times with GeneID:6710, 2 times with GeneID:56659, and 1 time with GeneID:9369. (B) The chromosomal region of the gene (GeneID:145508) that encodes a centrosomal protein (128kDa).

Table A.1: Reads coverage of the gene regions and non-gene regions. The reads coverage of gene region was calculated as the reads length multiply the number of contact in gene region / total length of gene region. The coverage of non-gene region was calculated as the read length * number of contact not in gene region / total length of non-gene region. Here the gene length was calculated according to the gene start and end information.

	Gene region	Non-gene region	Read length
Call4 cell line	2.81	2.36	100
RL cell line	1.47	1.15	100
Normal B-cell	0.19	0.17	76
ALL B-cell	1.79	1.49	120

Table A.2: Total number of reads for all samples mentioned in this work. The data of the normal B cell was downloaded from the publication [1]. The others were generated by us. One pair-end read pair contains two ends of reads. This table shows the total number of ends. For some cell / cell lines, we sequenced them more than one times and selected the one with the best quality to use in this work.

	Total number of reads	Used for analysis
Normal B cell	12,887,282	YES
RL1	60,272,006	NO
RL2	61,043,078	NO
RL3	65,579,872	NO
RL4	125,256,746	YES
Call4-1	62,741,712	NO
Call4-2	62,607,906	NO
Call4-3	133,542,778	YES
ALL B-Cell	77,888,742	YES

Appendix B

Statistical Properties of *H. sapiens*, *S. cerevisiae*, *C. elegans*, *D. melanogaster*, and 15 Plant Genomes

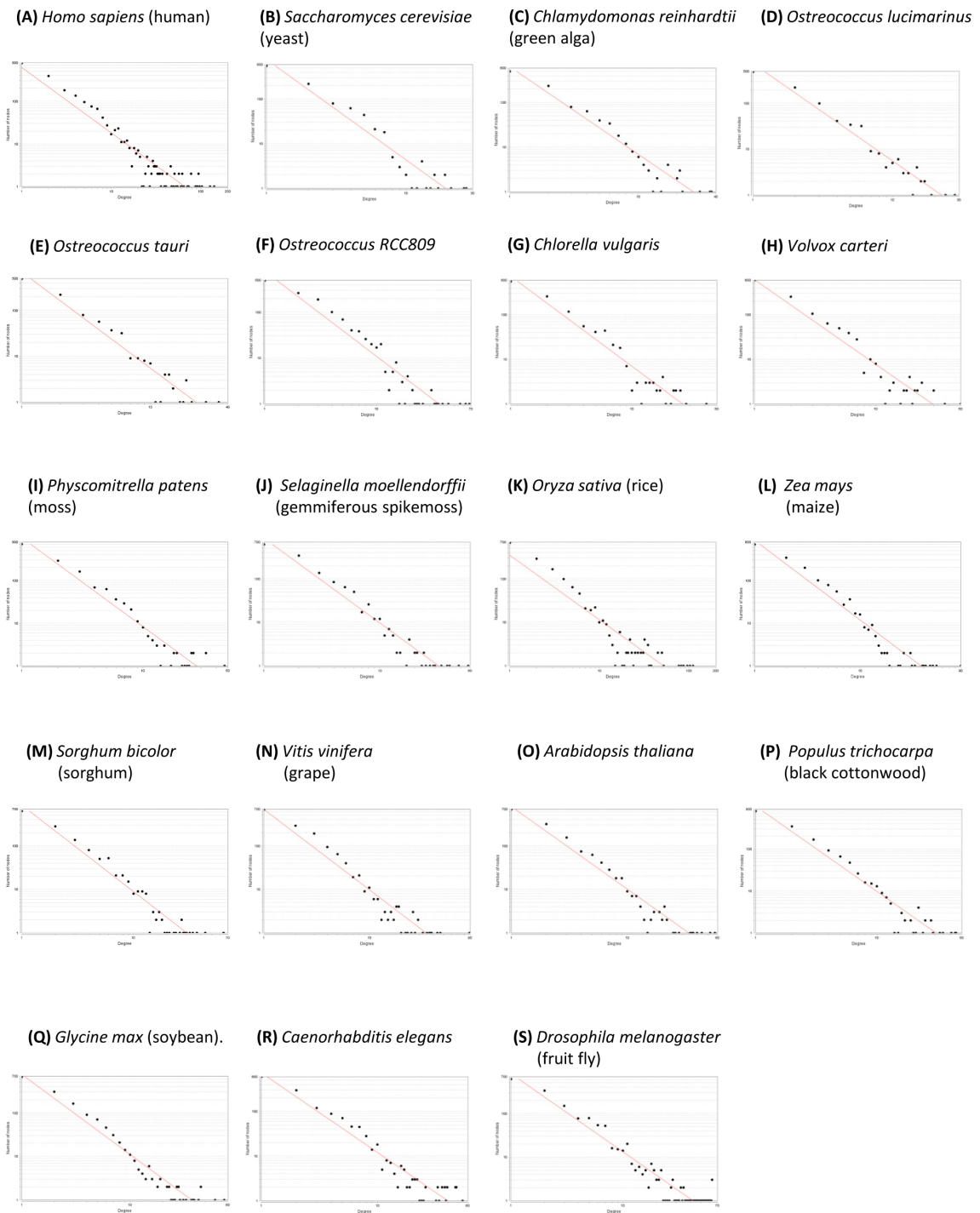


Figure B.1: The node degree distributions of *H. sapiens*, *S. cerevisiae*, *C. elegans*, *D. melanogaster*, and 15 plant genomes

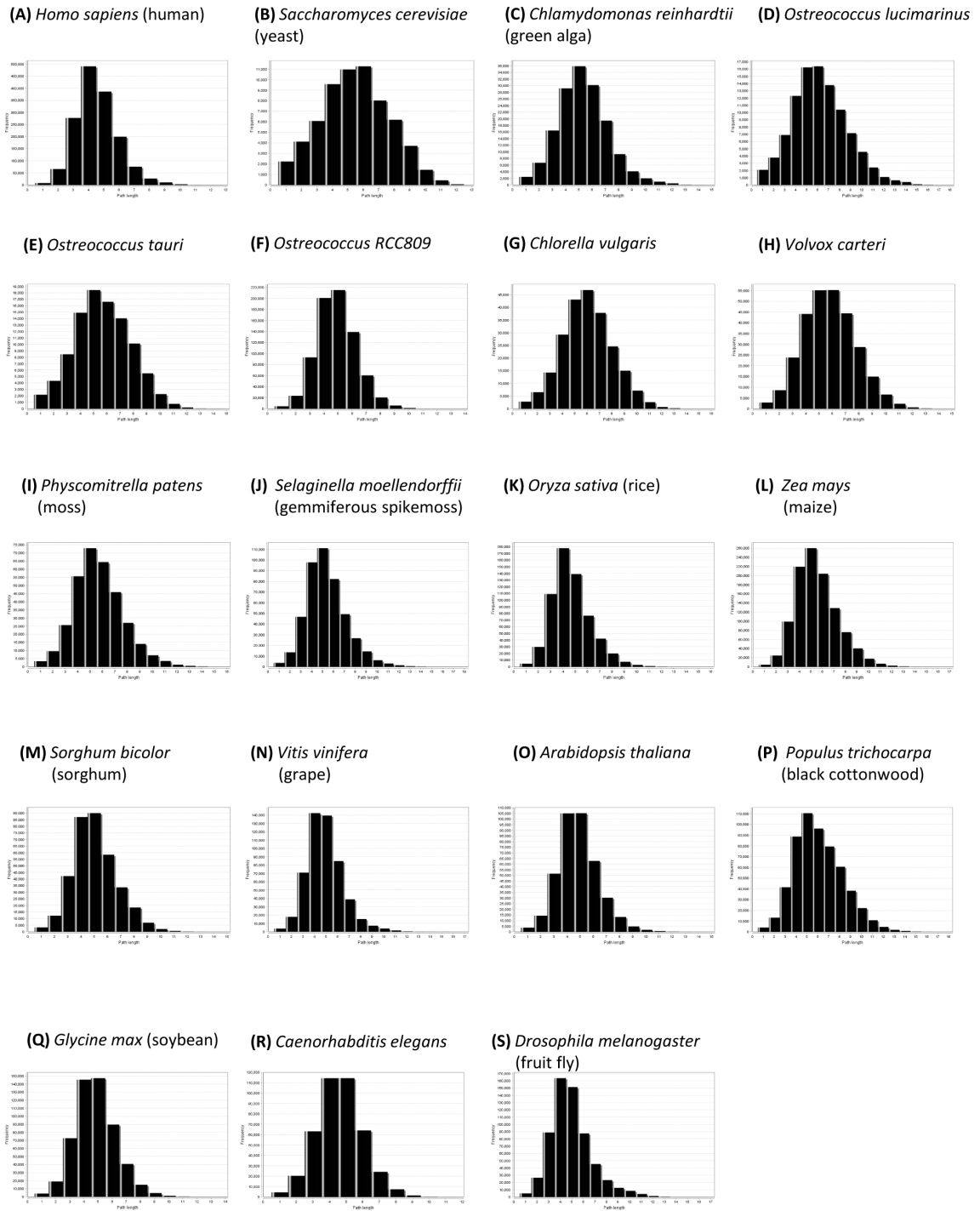


Figure B.2: The shortest path length distributions of *H. sapiens*, *S. cerevisiae*, *C. elegans*, *D. melanogaster*, and 15 plant genomes

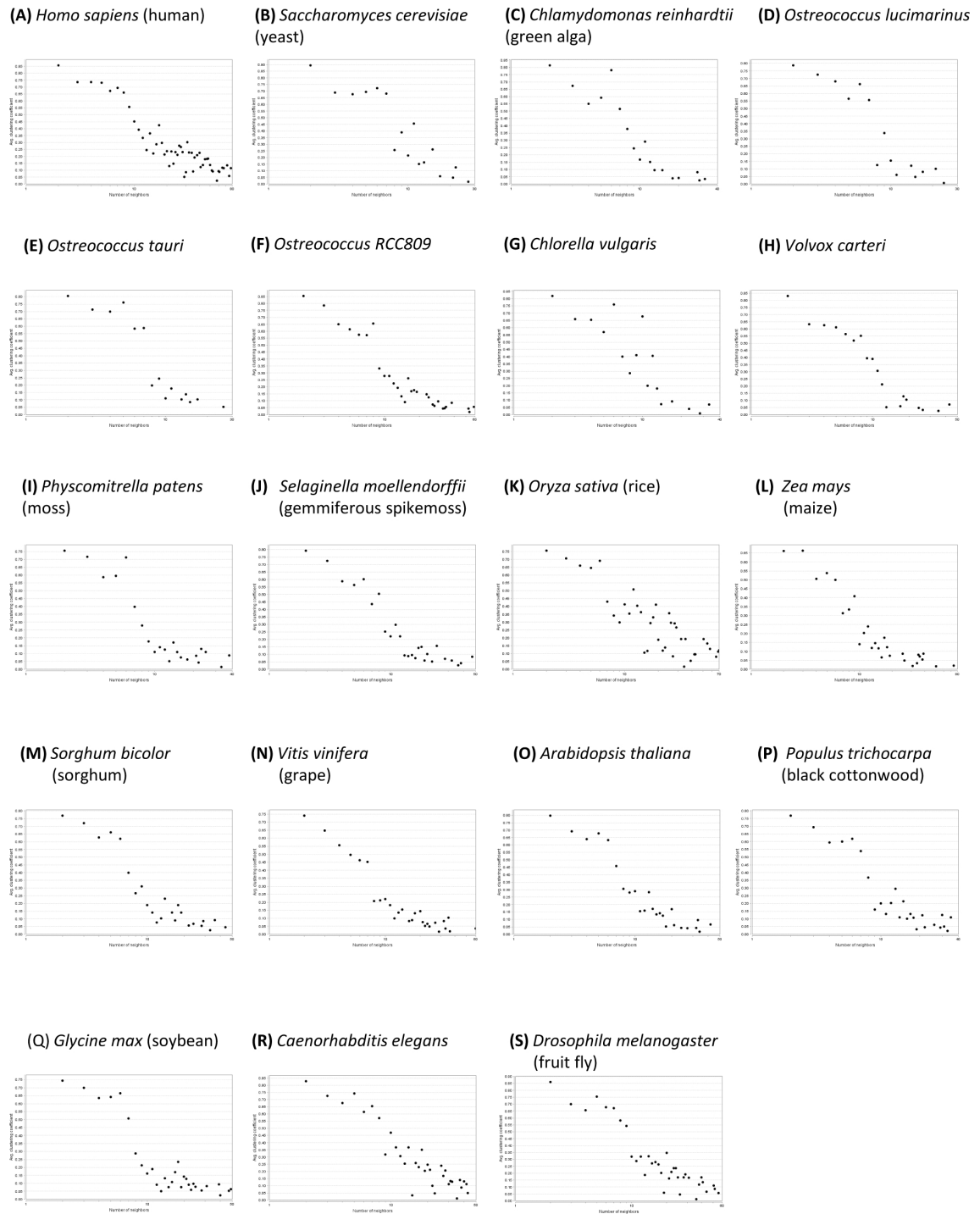


Figure B.3: The average clustering coefficient distributions of *H. sapiens*, *S. cerevisiae*, *C. elegans*, *D. melanogaster*, and 15 plant genomes

Appendix C

Web-based Bioinformatics Tools and Services

C.1 APOLLO: A Quality Assessment Service for Single and Multiple Protein Models

C.1.1 Overview

APOLLO can predict the absolute global quality of a single protein model, relative global qualities of multiple models, and residue-specific qualities of individual protein model.

C.1.2 URL

<http://sysbio.rnet.missouri.edu/apollo/>

C.1.3 Input

The input of APOLLO must be protein structural files in PDB format. APOLLO accepts one single protein model or a group of models (.zip for .tar.gz files) of the same target.

C.1.4 Output

APOLLO output absolute global quality scores generated by machine learning techniques and relative quality scores generated by pair-wise comparisons. It also outputs residue-specific quality scores for each model.

C.1.5 Software Architecture

PERL CGI was used in the server end of APOLLO. It executes algorithms in the background and output the results by sending emails or directly displaying them at the webpage.

C.2 Automated Assessment of CASP8 (2008) and CASP9 (2010) Predictions

C.2.1 Overview

An automated software pipeline was built to evaluate the CASP TS and QA predictions. CASP prediction groups were ranked based on prediction accuracy.

C.2.2 URL

CASP8: http://sysbio.rnet.missouri.edu/casp8_eva/ CASP9: http://sysbio.rnet.missouri.edu/casp9_assess/

C.2.3 Descriptions

The TS and QA predictions were downloaded from CASP website. The native structure for each target was downloaded from Protein Data Bank after released from the CASP website. TM-score (<http://zhanglab.ccmb.med.umich.edu/TM-score/>) was used to compare predicted structures with native.

Bibliography

- [1] E Lieberman-Aiden, NL Van Berkum, L Williams, M Imakaev, T Ragoczy, A Telling, I Amit, BR Lajoie, PJ Sabo, and MO Dorschner. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, 2009.
- [2] P Shannon, A Markiel, O Ozier, NS Baliga, JT Wang, D Ramage, N Amin, B Schwikowski, and T Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498–2504, 2003.
- [3] Lewis J et al. Alberts B, Bray D. *Molecular Biology of the Cell*. Chromosomal DNA and Its Packaging. New York: Garland Science, 3 edition, 1994.
- [4] T Cremer and C Cremer. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nature Reviews Genetics*, 2(4):292–301, 2001.
- [5] KJ Meaburn and T Misteli. Cell biology: chromosome territories. *Nature*, 445(7126):379–381, 2007.
- [6] PR Cook. Predicting three-dimensional genome structure from transcriptional activity. *Nature genetics*, 32(3):347–352, 2002.

- [7] S Cai, CC Lee, and T Kohwi-Shigematsu. SATB1 packages densely looped, transcriptionally active chromatin for coordinated expression of cytokine genes. *Nature genetics*, 38(11):1278–1288, 2006.
- [8] B Li, M Carey, and JL Workman. The role of chromatin during transcription. *Cell*, 128(4):707–719, 2007.
- [9] J Fraser, M Rousseau, S Shenker, MA Ferraiuolo, Y Hayashizaki, M Blanchette, and J Dostie. Chromatin conformation signatures of cellular differentiation. *Genome Biol*, 10(4):R37, 2009.
- [10] A Zemla. LGA: a method for finding 3d similarities in protein structures. *Nucleic Acids Research*, 31(13):3370–3374, 2003.
- [11] Z Wang, J Eickholt, and J Cheng. MULTICOM: a multi-level combination approach to protein structure prediction and its assessments in CASP8. *Bioinformatics*, 26(7):882–888, 2010.
- [12] M Ashburner, CA Ball, JA Blake, D Botstein, H Butler, JM Cherry, AP Davis, K Dolinski, SS Dwight, and JT Eppig. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.
- [13] JC Hansen. Conformational dynamics of the chromatin fiber in solution: determinants, mechanisms, and functions. *Annual review of biophysics and biomolecular structure*, 31(1):361–392, 2002.
- [14] PJ Horn and CL Peterson. Chromatin higher order folding–wrapping up transcription. *Science*, 297(5588):1824–1827, 2002.
- [15] MR Branco and A Pombo. Chromosome organization: new facts, new models. *Trends in cell biology*, 17(3):127–134, 2007.

- [16] J Mateos-Langerak, M Bohn, W De Leeuw, O Giromus, EMM Manders, PJ Verschure, MHG Indemans, HJ Gierman, DW Heermann, and R Van Driel. Spatially confined folding of chromatin in the interphase nucleus. *Proceedings of the National Academy of Sciences*, 106(10):3812–3817, 2009.
- [17] M Simonis, J Kooren, and W de Laat. An evaluation of 3C-based methods to capture DNA interactions. *Nature methods*, 4(11):895–901, 2007.
- [18] A Sanyal, D Bau, MA Marti-Renom, and J Dekker. Chromatin globules: a common motif of higher order chromosome structure? *Current Opinion in Cell Biology*, (23):325–331, 2011.
- [19] JG Gall and ML Pardue. Formation and detection of RNA-DNA hybrid molecules in cytological preparations. *Proceedings of the National Academy of Sciences*, 63(2):378–383, 1969.
- [20] J Dekker, K Rippe, M Dekker, and N Kleckner. Capturing chromosome conformation. *Science*, 295(5558):1306–1311, 2002.
- [21] Z Zhao, G Tavoosidana, M Sjolinder, A Gondor, P Mariano, S Wang, C Kanduri, M Lezcano, KS Sandhu, and U Singh. Circular chromosome conformation capture (4c) uncovers extensive networks of epigenetically regulated intra-and interchromosomal interactions. *Nature genetics*, 38(11):1341–1347, 2006.
- [22] J Dostie, TA Richmond, RA Arnaout, RR Selzer, WL Lee, TA Honan, ED Rubio, A Krumm, J Lamb, and C Nusbaum. Chromosome conformation capture carbon copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Research*, 16(10):1299–1309, 2006.
- [23] MA Ferraiuolo, M Rousseau, C Miyamoto, S Shenker, XQD Wang, M Nadler, M Blanchette, and J Dostie. The three-dimensional architecture of hox cluster silencing. *Nucleic Acids Research*, 38(21):7472–7484, 2010.

- [24] E Yaffe and A Tanay. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature genetics*, 43:10591065, 2011.
- [25] J Tomeczkowski, E Yakisan, B Wieland, A Reiter, K Welte, and KW Sykora. Absence of g-csf receptors and absent response to g-csf in childhood burkitt's lymphoma and b-all cells. *British journal of haematology*, 89(4):771–779, 1995.
- [26] GM Euskirchen, JS Rozowsky, CL Wei, WH Lee, ZD Zhang, S Hartman, O Emanuelsson, V Stolc, S Weissman, and MB Gerstein. Mapping of transcription factor binding regions in mammalian cells by chip: comparison of array-and sequencing-based technologies. *Genome Research*, 17(6):898–909, 2007.
- [27] D Karolchik, R Baertsch, M Diekhans, TS Furey, A Hinrichs, YT Lu, KM Roskin, M Schwartz, CW Sugnet, and DJ Thomas. The UCSC genome browser database. *Nucleic Acids Research*, 31(1):51–54, 2003.
- [28] R Albert, H Jeong, and AL Barabasi. Error and attack tolerance of complex networks. *Nature*, 406:378–382, 2000.
- [29] R Bonneau. Learning biological networks: from modules to dynamics. *Nature chemical biology*, 4(11):658–664, 2008.
- [30] L Hakes, JW Pinney, DL Robertson, and SC Lovell. Protein-protein interaction networks and biology - what's the connection? *Nature biotechnology*, 26(1):69–72, 2008.
- [31] O Rinner, LN Mueller, M Hubalek, M Muller, M Gstaiger, and R Aebersold. An integrated mass spectrometric and computational framework for the analysis of protein interaction networks. *Nature biotechnology*, 25(3):345–352, 2007.

- [32] B Schwikowski, P Uetz, and S Fields. A network of protein-protein interactions in yeast. *Nature biotechnology*, 18(12):1257–1261, 2000.
- [33] Z Wang, XC Zhang, MH Le, D Xu, G Stacey, and J Cheng. A protein domain co-occurrence network approach for predicting protein function and inferring species phylogeny. *PLoS ONE*, 6(3):e17906, 2011.
- [34] D Baker and A Sali. Protein structure prediction and structural genomics. *Science*, 294(5540):93–96, 2001.
- [35] J Cheng. A multi-template combination algorithm for protein comparative modeling. *BMC Structural Biology*, 8(1):18, 2008.
- [36] D Cozzetto, A Kryshchuk, and A Tramontano. Evaluation of CASP8 model quality predictions. *Proteins: Structure, Function, and Bioinformatics*, 77(S9):157–166, 2009.
- [37] Y Zhang and J Skolnick. Automated structure prediction of weakly homologous proteins on a genomic scale. *Proceedings of the National Academy of Sciences*, 101(20):7594–7599, 2004.
- [38] LJ McGuffin. Benchmarking consensus model quality assessment for protein fold recognition. *BMC Bioinformatics*, 8:345, 2007.
- [39] LJ McGuffin. The ModFOLD server for the quality assessment of protein structural models. *Bioinformatics*, 24(4):586, 2008.
- [40] M Paluszewski and K Karplus. Model quality assessment using distance constraints from alignments. *Proteins*, 75:540–549, 2008.
- [41] B Wallner and A Elofsson. Can correct protein models be identified? *Protein Science*, 12(5):1073–1086, 2003.

- [42] B Wallner and A Elofsson. Prediction of global and local model quality in casp7 using pcons and proq. *Proteins*, 69(8):184–93, 2007.
- [43] Y Zhang and J Skolnick. SPICKER: A clustering approach to identify near-native protein folds. *Journal of Computational Chemistry*, 25(6):865–871, 2004.
- [44] J Archie and K Karplus. Applying undertaker cost functions to model quality assessment. *Proteins*, 75:550–555, 2009.
- [45] P Benkert, SCE Tosatto, and D Schomburg. QMEAN: a comprehensive scoring function for model quality assessment. *Proteins*, 71(1), 2008.
- [46] M Cline, R Hughey, and K Karplus. Predicting reliable regions in protein sequence alignments, 2002.
- [47] J Qiu, W Sheffler, D Baker, and WS Noble. Ranking predicted protein structures with support vector regression. *Proteins*, 71(3):1175, 2008.
- [48] X Wang, Y Huang, and Y Xiao. Structural-symmetry-related sequence patterns of the proteins of beta-propeller family. *Journal of Molecular Graphics and Modelling*, 26(5):829–833, 2008.
- [49] J Cheng, Z Wang, AN Tegge, and J Eickholt. Prediction of global and local quality of casp8 models by multicom series. *Proteins*, 77(S9):181–184, 2009.
- [50] J Cheng, AZ Randall, MJ Sweredoski, and P Baldi. SCRATCH: a protein structure and structural feature prediction server. *Nucleic Acids Research*, 33(Web Server Issue):W72–W76, 2005.
- [51] P Larsson, MJ Skwark, B Wallner, and A Elofsson. Assessment of global and local model quality in CASP8 using Pcons and ProQ. *Proteins*, 77(S9):167–172, 2009.

- [52] Y Zhang and J Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4):702–710, 2004.
- [53] M Ben-David, O Noivirt-Birk, A Paz, J Prilusky, JL Sussman, Y Levy, and E Pearl. Assessment of CASP8 structure predictions for template free targets. *Proteins: Structure, Function, and Bioinformatics*, 77(Suppl 9):000–000, 2009.
- [54] LJ McGuffin and DB Roche. Rapid model quality assessment for protein structure predictions using the comparison of multiple models without structural alignments. *Bioinformatics*, 26(2):182–188, 2010.
- [55] Y Zhang. Progress and challenges in protein structure prediction. *Current Opinion in Structural Biology*, 18(3):342–348, 2008.
- [56] K Ginalski, A Elofsson, D Fischer, and L Rychlewski. 3d-jury: a simple approach to improve protein structure predictions, 2003.
- [57] LJ McGuffin. Prediction of global and local model quality in casp8 using the modfold server. *Proteins: Structure, Function, and Bioinformatics*, 77(S9):185–190, 2009.
- [58] Z Wang, A Tegge, and J Cheng. Evaluating the absolute quality of a single protein model using structural features and support vector machines. *Proteins*, 75:638–647, 2008.
- [59] M Chen J Li J Ma Y Wu, M Lu. Opus-ca: A knowledge-based potential function requiring only ca positions. *Protein Science*, 16:1449–1463, 2007.
- [60] A Randall and P Baldi. SELECTpro: effective protein model selection using a structure-based energy function resistant to blunders. *BMC Structural Biology*, 8:52, 2008.

- [61] D Cozzetto, A Kryshchak, M Ceriani, and A Tramontano. Assessment of predictions in the model quality assessment category. *Proteins*, 69(8):175–83, 2007.
- [62] Z Wang, J Eickholt, and J Cheng. APOLLO: A quality assessment service for single and multiple protein models. *Bioinformatics*, in press, 2011.
- [63] J Henkel. Soy: health claims for soy protein, questions about other components. *FDA consumer*, 34(3), 2000.
- [64] B Han, FM Rombouts, and MJR Nout. A chinese fermented soybean food. *International Journal of Food Microbiology*, 65(1-2):1–10, 2001. doi: DOI: 10.1016/S0168-1605(00)00523-7.
- [65] JE Carpenter and LP Gianessi. *Agricultural biotechnology: updated benefit estimates*. National Centre for Food and Agricultural Policy (NCFAP), Washington, USA, 2001.
- [66] J Schmutz, SB Cannon, J Schlueter, J Ma, D Hyten, Q Song, T Mitros, W Nelson, GD May, N Gill, M Peto, D Goodstein, JJ Thelen, J Cheng, T Sakurai, and T Umezawa. Genome sequence of the paleopolyploid soybean (*glycine max* (l.) merr.). *Nature*, page submitted, 2009.
- [67] M Jakoby, B Weisshaar, W Droge-Laser, J Vicente-Carbajosa, J Tiedemann, T Kroj, and F Parcy. bZIP transcription factors in arabidopsis. *Trends in Plant Science*, 7(3):106–111, 2002.
- [68] JC Reese. Basal transcription factors. *Current opinion in genetics & development*, 13(2):114–118, 2003.

- [69] ROJ Weinzierl. *Mechanisms of gene expression: structure, function and evolution of the basal transcriptional machinery*. Imperial College Press, London, UK, 1999.
- [70] C Corrinne. *Transcription factors and mammalian development*, volume 27, page 351. Academic Press, Inc., 1992.
- [71] Q Liu, M Kasuga, Y Sakuma, H Abe, S Miura, K Yamaguchi-Shinozaki, and K Shinozaki. Two transcription factors, DREB1 and DREB2, with an EREBP/AP2 DNA binding domain separate two cellular signal transduction pathways in drought-and low-temperature-responsive gene expression, respectively, in Arabidopsis. *The Plant Cell Online*, 10(8):1391–1406, 1998.
- [72] JL Riechmann, J Heard, G Martin, L Reuber, C Z, J Keddie, L Adam, O Pineda, OJ Ratcliffe, and RR Samaha. Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. *Science*, 290(5499):2105–2110, 2000.
- [73] Y Sakuma, K Maruyama, Y Osakabe, F Qin, M Seki, K Shinozaki, and K Yamaguchi-Shinozaki. Functional analysis of an arabidopsis transcription factor, DREB2A, involved in drought-responsive gene expression. *The Plant Cell Online*, 18(5):1292–1309, 2006.
- [74] CS Johnson, B Kolevski, and DR Smyth. TRANSPARENT TESTA GLABRA2, a trichome and seed coat development gene of arabidopsis, encodes a WRKY transcription factor. *The Plant Cell*, 14(6):1359–1375, 2002.
- [75] B Ulker and IE Somssich. Wrky transcription factors: from dna binding towards biological function. *Current Opinion in Plant Biology*, 7(5):491–498, 2004.
- [76] JL Shultz, D Kurunam, K Shopinski, MJ Iqbal, S Kazi, K Zobrist, R Bashir, S Yaegashi, N Lavu, AJ Afzai, CR Yesudas, MA Kassem, C Wu, H Zhang,

- CD Town, K Meksem, and DA Lightfoot. The Soybean Genome Database (SoyGD): a browser for display of duplicated, polyploid, regions and sequence tagged sites on the integrated physical and genetic maps of glycine max. *Nucleic Acids Research*, 34(Database Issue):D758–D765, 2006.
- [77] KCC Cheng and MV Stromvik. SoyXpress: a database for exploring the soybean transcriptome. *BMC genomics*, 9(1):368, 2008.
- [78] A Guo, X Chen, G Gao, H Zhang, Q Zhu, X Liu, Y Zhong, X Gu, K He, and J Luo. PlantTFDB: a comprehensive plant transcription factor database. *Nucleic Acids Research*, 36(Database Issue):D966–D969, 2008.
- [79] D Wilson, V Charoensawan, SK Kummerfeld, and SA Teichmann. DBD taxonomically broad transcription factor predictions: new content and functionality. *Nucleic Acids Research*, 36(Database Issue):D88–D92, 2007.
- [80] EM Zdobnov and R Apweiler. InterProScan—an integration platform for the signature-recognition methods in interpro. *Bioinformatics*, 17(9):847–848, 2001.
- [81] G Stoesser, MA Tuli, R Lopez, and P Sterk. The EMBL nucleotide sequence database. *Nucleic Acids Research*, 27(1):18–24, 1999.
- [82] R Apweiler, A Bairoch, CH Wu, WC Barker, B Boeckmann, S Ferro, E Gasteiger, H Huang, R Lopez, M Magrane, MJ Martin, DA Natale, C O’Donovan, N Redaschi, and LSL Yeh. UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, 32(Database Issue):D115–D119, 2004.
- [83] KD Pruitt, T Tatusova, and DR Maglott. NCBI reference sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, 00(Database Issue):D1–D5, 2006.

- [84] E Wingender, X Chen, R Hehl, H Karas, I Liebich, V Matys, T Meinhardt, M Prub, I Reuter, and F Schacherer. TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Research*, 28(1):316–319, 2000.
- [85] T Hubbard, D Barker, E Birney, G Cameron, Y Chen, L Clark, T Cox, J Cuff, V Curwen, and T Down. The ensembl genome database project. *Nucleic Acids Research*, 30(1):38–41, 2002.
- [86] A Bateman, L Coin, R Durbin, RD Finn, V Hollich, S Griffiths-Jones, A Khanna, M Marshall, S Moxon, and ELL Sonnhammer. The pfam protein families database. *Nucleic Acids Research*, 32(1):276–280, 2004.
- [87] M Madera, C Vogel, SK Kummerfeld, C Chothia, and J Gough. The SUPERFAMILY database in 2004: additions and improvements. *Nucleic Acids Research*, 32(Database Issue):D235–D239, 2004.
- [88] SK Burley, SC Almo, JB Bonanno, M Capel, MR Chance, T Gaasterland, D Lin, A Sali, FW Studier, and S Swaminathan. Structural genomics: beyond the human genome project. *Nature Genetics*, 23:151–158, 1999.
- [89] SK Burley. An overview of structural genomics. *Nature Structural & Molecular Biology*, 7:932–934, 2000.
- [90] HM Berman, J Westbrook, Z Feng, G Gilliland, T Bhat, H Weissig, IN Shindyalov, and PE Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 2000.
- [91] B Boeckmann, A Bairoch, R Apweiler, MC Blatter, A Estreicher, E Gasteiger, MJ Martin, K Michoud, C O’Donovan, and I Phan. The SWISS-PROT protein knowledgebase and its supplement trembl in 2003. *Nucleic Acids Research*, 31(1):365–370, 2003.

- [92] SY Rhee, W Beavis, TZ Berardini, G Chen, D Dixon, A Doyle, M Garcia-Hernandez, E Huala, G Lander, M Montoya, N Miller, LA Mueller, S Mundodi, L Reiser, J Tacklind, and DC Weems. The arabidopsis information resource (TAIR): a model organism database providing a centralized, curated gateway to arabidopsis biology, research materials and community. *Nucleic Acids Research*, (31):224–228, 2003.
- [93] I Letunic, RR Copley, B Pils, S Pinkert, J Schultz, and P Bork. SMART 5: domains in the context of genomes and networks. *Nucleic Acids Research*, 34(Database Issue):D257–D260, 2006.
- [94] M Kanehisa, M Araki, S Goto, M Hattori, M Hirakawa, M Itoh, T Katayama, S Kawashima, S Okuda, and T Tokimatsu. KEGG for linking genomes to life and the environment. *Nucleic Acids Research*, 36(Database Issue):D480–D484, 2008.
- [95] TK Attwood, MJ Blythe, DR Flower, A Gaulton, JE Mabey, N Maudling, L McGregor, AL Mitchell, G Moulton, and K Paine. PRINTS and PRINTS-S shed light on protein ancestry. *Nucleic Acids Research*, 30(1):239–241, 2002.
- [96] SV Angiuoli, A Gussman, W Klimke, G Cochrane, D Field, GM Garrity, CD Kodira, N Kyrpides, R Madupu, and V Markowitz. Toward an online repository of standard operating procedures (SOPs) for (meta) genomic annotation. *OMICS: A Journal of Integrative Biology*, 12(2):137–141, 2008.
- [97] NJ Mulder, R Apweiler, TK Attwood, A Bairoch, A Bateman, D Binns, M Biswas, P Bradley, P Bork, and P Bucher. InterPro: An integrated documentation resource for protein families, domains and functional sites. *Briefings in Bioinformatics*, 3(3):225–235, 2002.

- [98] N Hulo, A Bairoch, V Bulliard, L Cerutti, E De Castro, PS Langendijk-Genevaux, M Pagni, and CJA Sigrist. The PROSITE database. *Nucleic Acids Research*, 34(Database Issue):D227–D230, 2006.
- [99] RF Yeh, LP Lim, and CB Burge. Computational inference of homologous gene structures in the human genome. *Genome research*, 11(5):803–816, 2001.
- [100] BJ Haas, AL Delcher, SM Mount, JR Wortman, RK Smith Jr, LI Hannick, R Maiti, CM Ronning, DB Rusch, and CD Town. Improving the arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research*, 31(19):5654–5666, 2003.
- [101] TK Attwood, MDR Croning, DR Flower, AP Lewis, JE Mabey, P Scordis, JN Selley, and W Wright. PRINTS-S: the database formerly known as prints. *Nucleic Acids Research*, 28(1):225–227, 2000.
- [102] F Corpet, J Gouzy, and D Kahn. Recent improvements of the prodom database of protein domain families. *Nucleic Acids Research*, 27(1):263–267, 1999.
- [103] DH Haft, BJ Loftus, DL Richardson, F Yang, JA Eisen, IT Paulsen, and O White. TIGRFAMs: a protein family resource for the functional identification of proteins. *Nucleic Acids Research*, 29(1):41–43, 2001.
- [104] CH Wu, H Huang, LSL Yeh, and WC Barker. Protein family classification and functional annotation. *Computational Biology and Chemistry*, 27(1):37–47, 2003.
- [105] J Gough, K Karplus, R Hughey, and C Chothia. Assignment of homology to genome sequences using a library of hidden markov models that represent all proteins of known structure. *Journal of Molecular Biology*, 313(4):903–919, 2001.

- [106] DWA Buchan, AJ Shepherd, D Lee, FMG Pearl, SCG Rison, JM Thornton, and CA Orengo. Gene3D: structural assignment for whole genes and genomes using the cath domain structure database. *Genome Research*, 12(3):503–514, 2002.
- [107] H Mi, B Lazareva-Ulitsky, R Loo, A Kejariwal, J Vandergriff, S Rabkin, N Guo, A Muruganujan, O Doremiex, and MJ Campbell. The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Research*, 33(Database Issue):D284–D288, 2005.
- [108] T Lima, AH Auchincloss, E Coudert, G Keller, K Michoud, C Rivoire, V Buliard, E de Castro, C Lachaize, and D Baratin. HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot. *Nucleic Acids Research*, 37(Database issue):D471–D478, 2009.
- [109] DM Riano-Pachon, S Ruzicic, I Dreyer, and B Mueller-Roeber. PlnTFDB: an integrative plant transcription factor database. *BMC Bioinformatics*, 8(1):42, 2007.
- [110] K Kakar, M Wandrey, T Czechowski, T Gaertner, WR Scheible, M Stitt, I Torres-Jerez, Y Xiao, JC Redman, and HC Wu. A community resource for high-throughput quantitative RT-PCR analysis of transcription factor gene expression in medicago truncatula. *Plant Methods*, 4(1):18, 2008.
- [111] RC Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797, 2004.
- [112] R Hughey and A Krogh. *SAM: sequence alignment and modeling software system*. University of California at Santa Cruz, 1995.

- [113] C Yan, M Terribilini, F Wu, RL Jernigan, D Dobbs, and V Honavar. Predicting DNA-binding sites of proteins from amino acid sequence. *BMC Bioinformatics*, 7(1):262, 2006.
- [114] L Wang and SJ Brown. BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. *Nucleic Acids Research*, 34(Web Server issue):W243–W248, 2006.
- [115] S Hwang, Z Gou, and IB Kuznetsov. DP-Bind: a web server for sequence-based prediction of dna-binding residues in dna-binding proteins. *Bioinformatics*, 23(5):634–636, 2007.
- [116] K Yamasaki, T Kigawa, M Inoue, M Tateno, T Yamasaki, T Yabuki, M Aoki, E Seki, T Matsuda, and Y Tomo. Solution structure of the B3 DNA binding domain of the Arabidopsis cold-responsive transcription factor RAV1. *The Plant Cell Online*, 16(12):3448–3459, 2004.
- [117] W Kaplan and TG Littlejohn. Swiss-PDB viewer (deep view). *Briefings in Bioinformatics*, 2(2):195–197, 2001.
- [118] S Jones, HP Shanahan, HM Berman, and JM Thornton. Using electrostatic potentials to predict dna-binding sites on dna-binding proteins. *Nucleic Acids Research*, 31(24):7189–7198, 2003.
- [119] N Alexandrov and I Shindyalov. PDP: protein domain parser, 2003.
- [120] J Cheng. DOMAC: an accurate, hybrid protein domain prediction server. *Nucleic Acids Research*, 35(Web Server Issue):W354–W356, 2007.
- [121] GE Crooks, G Hon, JM Chandonia, and SE Brenner. WebLogo: a sequence logo generator. *Genome Research*, 14(6):1188–1190, 2004.

- [122] DA Benson, MS Boguski, DJ Lipman, J Ostell, BF Ouellette, BA Rapp, and DL Wheeler. Genbank. *Nucleic Acids Research*, 27(1):12–17, 1999.
- [123] M Libault, T Joshi, K Takahashi, A Hurley-Sommer, K Puricelli, S Blake, D Xu, HT Nguyen, and G Stacey. Large-scale analysis of putative soybean regulatory gene expression identifies a Myb gene involved in soybean nodule development. *Plant Physiology*, 151:1207–1220, 2009.
- [124] JA Schlueter, JY Lin, SD Schlueter, IF Vasylenko-Sanders, S Deshpande, J Yi, M O’Bleness, BA Roe, RT Nelson, and BE Scheffler. Gene duplication and paleopolyploidy in soybean and the implications for whole genome sequencing. *BMC genomics*, 8(1):330, 2007.
- [125] JA Schlueter, P Dixon, C Granger, D Grant, L Clark, JJ Doyle, and RC Shoemaker. Mining EST databases to resolve evolutionary events in major crop species. *Genome*, 47(5):868–876, 2004.
- [126] SS Merchant, SE Prochnik, O Vallon, EH Harris, SJ Karpowicz, GB Witman, A Terry, A Salamov, LK Fritz-Laylin, and L Marechal-Drouard. The chlamydomonas genome reveals the evolution of key animal and plant functions. *Science*, 318(5848):245–250, 2007.
- [127] SA Rensing, D Lang, AD Zimmer, A Terry, A Salamov, H Shapiro, T Nishiyama, PF Perroud, EA Lindquist, and Y Kamisugi. The physcomitrella genome reveals evolutionary insights into the conquest of land by plants. *Science*, 319(5859):64–69, 2008.
- [128] S Scofield and JAH Murray. KNOX gene function in plant stem cell niches. *Plant Molecular Biology*, 60(6):929–946, 2006.

- [129] M Libault, T Joshi, VA Benedito, D Xu, MK Udvardi, and G Stacey. Legume transcription factor genes; what makes legumes so special? *Plant Physiology*, 151:991–1001, 2009.
- [130] D Martin, M Berriman, and G Barton. GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes. *BMC Bioinformatics*, 5(1):178, 2004.
- [131] G Zehetner. OntoBlast function: From sequence similarities directly to potential functional annotations by ontology terms. *Nucleic Acids Research*, 31(13):3799–3803, 2003.
- [132] S Hennig, D Groth, and H Lehrach. Automated gene ontology annotation for anonymous sequence data. *Nucleic Acids Research*, 31(13):3712–3715, 2003.
- [133] 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, 1997.
- [134] T Hawkins, M Chitale, S Luban, and D Kihara. PFP: Automated prediction of gene ontology functional annotations with confidence scores using protein sequence data. *Proteins: Structure, Function, and Bioinformatics*, 74(3):566–582, 2009.
- [135] JA Eisen. A phylogenomic study of the MutS family of proteins. *Nucleic Acids Research*, 26(18):4291–4300, 1998.
- [136] M Goodman, J Czelusniak, GW Moore, AE Romero-Herrera, and G Matsuda. Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. *Systematic Biology*, 28(2):132–163, 1979.

- [137] K Sjolander. Phylogenomic inference of protein molecular function: advances and challenges. *Bioinformatics*, 20(2):170–179, 2004.
- [138] ELL Sonnhammer and EV Koonin. Orthology, paralogy and proposed classification for paralog subtypes. *Trends in Genetics*, 18(12):619–620, 2002.
- [139] BE Engelhardt, MI Jordan, KE Muratore, and SE Brenner. Protein molecular function prediction by Bayesian phylogenomics. *PLoS computational biology*, 1(5):e45, 2005.
- [140] CEV Storm and ELL Sonnhammer. Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. *Bioinformatics*, 18(1):92–99, 2002.
- [141] C Zmasek and S Eddy. RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics*, 3(1):14, 2002.
- [142] A. Jocker, F. Hoffmann, A. Groscurth, and H. Schoof. Protein function prediction and annotation in an integrated environment powered by web services (afawe). *Bioinformatics*, 24(20):2393–2394, 2008.
- [143] H Hishigaki, K Nakai, T Ono, A Tanigami, and T Takagi. Assessment of prediction accuracy of protein function from protein-protein interaction data. *Yeast*, 18(6):523–531, 2001.
- [144] M. Deng, K. Zhang, S. Mehta, T. Chen, and F. Sun. Prediction of protein function using protein-protein interaction data. *Journal of Computational Biology*, 10(6):947–960, 2003.
- [145] KM Borgwardt, CS Ong, S Schonauer, SVN Vishwanathan, AJ Smola, and HP Kriegel. Protein function prediction via graph kernels. *Bioinformatics*, 21(Suppl 1):i47–i56, 2005.

- [146] R Sharan, I Ulitsky, and R Shamir. Network-based prediction of protein function. *Molecular Systems Biology*, 3(1), 2007.
- [147] A Vazquez, A Flammini, A Maritan, and A Vespignani. Global protein function prediction in protein-protein interaction networks. *Nature Biotechnology*, 21:697700, 2003.
- [148] U. Karaoz, TM Murali, S. Letovsky, Y. Zheng, C. Ding, C.R. Cantor, and S. Kasif. Whole-genome annotation by using evidence integration in functional-linkage networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(9):2888–2893, 2004.
- [149] EM Marcotte, M Pellegrini, MJ Thompson, TO Yeates, and D Eisenberg. A combined algorithm for genome-wide prediction of protein function. *Nature*, 402(6757):83–86, 1999.
- [150] B Linghu, E Snitkin, D Holloway, A Gustafson, Y Xia, and C DeLisi. High-precision high-coverage functional inference from integrated data sources. *BMC Bioinformatics*, 9(1):119, 2008.
- [151] XM Zhao, L Chen, and K Aihara. Protein function prediction with the shortest path in functional linkage graph and boosting. *International journal of bioinformatics research and applications*, 4(4):375–384, 2008.
- [152] N Massjouni, CG Rivera, and TM Murali. VIRGO: computational prediction of gene functions. *Nucleic Acids Research*, 34(suppl 2):W340–W344, 2006.
- [153] LJ Jensen, R Gupta, HH Staerfeldt, and S Brunak. Prediction of human protein function according to gene ontology categories. *Bioinformatics*, 19(5):635–642, 2003.

- [154] AE Lobley, T. Nugent, CA Orengo, and DT Jones. Ffpred: an integrated feature-based function prediction server for vertebrate proteomes. *Nucleic Acids Research*, 36(suppl 2):W297–W302, 2008.
- [155] T Hawkins, M Chitale, and D Kihara. New paradigm in protein function prediction for large scale omics analysis. *Molecular BioSystems*, 4(3):223–231, 2008.
- [156] R Rentzsch and CA Orengo. Protein function prediction-the power of multiplicity. *Trends in biotechnology*, 27(4):210–219, 2009.
- [157] J Soding, A Biegert, and AN Lupas. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Research*, 33(Web Server Issue):W244–W248, 2005.
- [158] Z Du, L Li, CF Chen, PS Yu, and JZ Wang. G-SESAME: web tools for go-term-based gene similarity analysis and knowledge discovery. *Nucleic Acids Research*, 37(Web Server issue):W345, 2009.
- [159] LH Hartwell, JJ Hopfield, S Leibler, and AW Murray. From molecular to modular cell biology. *Nature*, 402(6761):C47–C52, 1999.
- [160] T Ideker, T Galitski, and L Hood. A new approach to decoding life: Systems biology. *Annual Review of Genomics and Human Genetics*, 2:343–372, 2001.
- [161] H Kitano. Computational systems biology. *Nature*, 420(6912):206–210, 2002.
- [162] M Hucka, A Finney, HM Sauro, H Bolouri, JC Doyle, and H Kitano. The systems biology markup language (sbml): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531, 2003.
- [163] J Cheng, L Scharenbroich, P Baldi, and E Mjolsness. Sigmoid: towards an intelligent, scalable, software infrastructure for pathway bioinformatics and systems biology. *IEEE Intelligent Systems*, 20(3):1–10, 2005.

- [164] A Zhang. *Protein interaction networks: computational analysis*. Cambridge University Press, 2009.
- [165] AL Barabasi and ZN Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113, 2004.
- [166] MB Elowitz and S Leibler. A synthetic oscillatory network of transcriptional regulators. *Nature*, 403(6767):335–338, 2000.
- [167] D Segre, D Vitkup, and GM Church. Analysis of optimality in natural and perturbed metabolic networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(23):15112–15117, 2002.
- [168] P Uetz, L Giot, G Cagney, TA Mansfield, RS Judson, JR Knight, D Lockshon, V Narayan, M Srinivasan, and P Pochart. A comprehensive analysis of protein-protein interactions in *saccharomyces cerevisiae*. *Nature*, 403(6770):623–627, 2000.
- [169] R Singh, J Xu, and B Berger. Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proceedings of the National Academy of Sciences*, 105(35):12763–12768, 2008.
- [170] F Ramirez, A Schlicker, Y Assenov, T Lengauer, and M Albrecht. Computational analysis of human protein interaction networks. *Proteomics*, 7(15):2541–2552, 2007.
- [171] ACF Lewis, NS Jones, MA Porter, and DM Charlotte. The function of communities in protein interaction networks at multiple scales. *BMC Systems Biology*, 4(1):100, 2010.

- [172] F Li, P Li, W Xu, Y Peng, X Bo, and S Wang. Perturbation analyzer: a tool for investigating the effects of concentration perturbation on protein interaction networks. *Bioinformatics*, 26(2):275–277, 2010.
- [173] S Agarwal, CM Deane, MA Porter, and NS Jones. Revisiting date and party hubs: Novel approaches to role assignment in protein interaction networks. *PLoS Comput Biol*, 6(6):e1000817, 2010.
- [174] TP Nguyen and F Jordan. A quantitative approach to study indirect effects among disease proteins in the human protein interaction network. *BMC Systems Biology*, 4(1):103, 2010.
- [175] G Wu, X Feng, and L Stein. A human functional protein interaction network and its application to cancer data analysis. *Genome Biology*, 11(5):R53, 2010.
- [176] T Ito, T Chiba, R Ozawa, M Yoshida, M Hattori, and Y Sakaki. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci.*, 98(8):4569–4574, 2001.
- [177] J Scott, T Ideker, RM Karp, and R Sharan. Efficient algorithms for detecting signaling pathways in protein interaction networks. *Journal of Computational Biology*, 13(2):133–144, 2006.
- [178] X Chen, M Liu, and R Ward. Protein function assignment through mining cross-species protein-protein interactions. *PLoS ONE*, 3(2):e1562, 2008.
- [179] C Zhang, T Joshi, GN Lin, and D Xu. An integrated probabilistic approach for gene function prediction using multiple sources of high-throughput data. *Int. J. of Computational Biology and Drug Design*, page in press, 2009.

- [180] P Bork, LJ Jensen, C von Mering, AK Ramani, I Lee, and EM Marcotte. Protein interaction networks from yeast to human. *Current Opinion in Structural Biology*, 14(3):292–299, 2004.
- [181] S Wuchty and E Almaas. Evolutionary cores of domain co-occurrence networks. *BMC Evolutionary Biology*, 5(1):24, 2005.
- [182] S Wuchty. Scale-free behavior in protein domain networks. *Molecular biology and evolution*, 18(9):1694–1702, 2001.
- [183] JH Fong, LY Geer, AR Panchenko, and SH Bryant. Modeling the evolution of protein domain architectures using maximum parsimony. *Journal of Molecular Biology*, 366(1):307–315, 2007.
- [184] K Sarah and T Sarah. Protein domain organisation: adding order. *BMC Bioinformatics*, 10(1):39, 2009.
- [185] D Ekman, AK Bjirklund, J Frey-Skott, and A Elofsson. Multi-domain proteins in the three kingdoms of life: orphan domains and other unassigned regions. *Journal of Molecular Biology*, 348(1):231–243, 2005.
- [186] EL Sonnhammer, SR Eddy, E Birney, A Bateman, and R Durbin. Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Research*, 26(1):320, 1998.
- [187] F Servant, C Bru, S Carrere, E Courcelle, J Gouzy, D Peyruc, and D Kahn. ProDom: automated clustering of homologous domains. *Briefings in Bioinformatics*, 3(3):246–251, 2002.
- [188] L Issel-Tarver, KR Christie, K Dolinski, R Andrada, R Balakrishnan, CA Ball, G Binkley, S Dong, SS Dwight, and DG Fisk. Saccharomyces genome database. *Methods in enzymology*, 350:329–346, 2002.

- [189] S Ouyang, W Zhu, J Hamilton, H Lin, M Campbell, K Childs, F Thibaud-Nissen, RL Malek, Y Lee, and L Zheng. The TIGR rice genome annotation resource: improvements and new features. *Nucleic Acids Research*, 35(suppl 1):D883–D887, 2006.
- [190] CZ Cai, LY Han, ZL Ji, and YZ Chen. Enzyme family classification by support vector machines. *Proteins: Structure, Function, and Bioinformatics*, 55(1):66–76, 2004.
- [191] JC Whisstock and AM Lesk. Prediction of protein function from protein sequence and structure. *Quarterly reviews of biophysics*, 36(3):307–340, 2004.
- [192] KM Borgwardt and HP Kriegel. Kernel methods for protein function prediction, 2005.
- [193] T Joachims. *Making large scale SVM learning practical. Advances in Kernel Methods - Support Vector Learning*. MIT Press, Cambridge, USA, 1999.
- [194] CS Liao, K Lu, M Baym, R Singh, and B Berger. IsoRankN: spectral methods for global alignment of multiple protein networks. *Bioinformatics*, 25(12):i253–i258, 2009.
- [195] J Felsenstein. PHYLIP-phylogeny inference package (version 3.2). *Cladistics*, 5(1):164–166, 1989.
- [196] N Saitou and M Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 4(4):406–425, 1987.
- [197] DH Bergey and JG Holt. *Bergey’s manual of determinative bacteriology*. Lippincott Williams & Wilkins, 1994.

- [198] G Lin, Z Cai, S Chakraborty, and D Xu. ComPhy: prokaryotic composite distance phylogenies inferred from whole-genome gene sets. *BMC Bioinformatics*, 10(Suppl 1):S5, 2009.
- [199] EJ Deeds, H Hennessey, and EI Shakhnovich. Prokaryotic phylogenies inferred from protein structural domains. *Genome Research*, 15(3):393–402, 2005.
- [200] DJ Watts and SH Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.
- [201] E Ravasz, AL Somera, DA Mongru, ZN Oltvai, and AL Barabasi. Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551–1555, 2002.
- [202] D Li, J Li, S Ouyang, J Wang, S Wu, P Wan, Y Zhu, X Xu, and F He. Protein interaction networks of *saccharomyces cerevisiae*, *caenorhabditis elegans* and *drosophila melanogaster*: large-scale organization and robustness. *Proteomics*, 6(2):456–461, 2006.
- [203] H Jeong, B Tombor, R Albert, ZN Oltvai, and AL Barabasi. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, 2000.
- [204] H Saigo, M Hattori, H Kashima, and K Tsuda. Reaction graph kernels predict ec numbers of unknown enzymatic reactions in plant secondary metabolism. *BMC Bioinformatics*, 11(Suppl 1):S31, 2010.
- [205] K Komurov, MA White, and PT Ram. Use of data-biased random walks on graphs for the retrieval of context-specific networks from genomic data. *PLoS Comput Biol*, 6(8):e1000889, 2010.
- [206] MK Basu, L Carmel, IB Rogozin, and EV Koonin. Evolution of protein domain promiscuity in eukaryotes. *Genome Research*, 18(3):449–461, 2008.

- [207] Y Luo, C Fu, DY Zhang, and K Lin. BPhyOG: an interactive server for genome-wide inference of bacterial phylogenies based on overlapping genes. *BMC Bioinformatics*, 8(1):266, 2007.
- [208] L Gao, J Qi, JD Sun, and BL Hao. Prokaryote phylogeny meets taxonomy: An exhaustive comparison of composition vector trees with systematic bacteriology. *Science in China Series C: Life Sciences*, 50(5):587–599, 2007.
- [209] X Wu, Z Cai, XF Wan, T Hoang, R Goebel, and G Lin. Nucleotide composition string selection in HIV-1 subtyping using whole genomes. *Bioinformatics*, 23(14):1744–1752, 2007.
- [210] AG Murzin, SE Brenner, T Hubbard, and C Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247(4):536–540, 1995.
- [211] GRG Lanckriet, M Deng, N Cristianini, MI Jordan, and WS Noble. Kernel-based data fusion and its application to protein function prediction in yeast. volume 9, pages 300–311. Proceedings of the Pacific Symposium on Biocomputing.
- [212] S Hiroto, H Masahiro, K Hisashi, and T Koji. Reaction graph kernels predict ec numbers of unknown enzymatic reactions in plant secondary metabolism. *BMC Bioinformatics*, 11(Suppl 1):S31, 2010.

VITA

Zheng Wang was born in Jinan, Shandong, China. He received a Bachelor's degree from Shandong University of Finance and Economics, China, in 2004, and a Master of Computer Science degree from the University of New Brunswick, Canada, in 2007. He started his Ph.D. studies in the Computer Science Department at the University of Missouri in Spring 2008. He received a Shumaker Bioinformatics Fellowship during his PhD studies and a "Mizzou Advantage Preparing Future Faculty" Postdoctoral Fellowship (four candidates university-wide received this fellowship in 2012) after he earned his Ph.D. degree at the University of Missouri.

Protein structural prediction is an area that he has been continuously working in since 2005. He started from protein secondary structure prediction and worked on tertiary structure prediction during his PhD studies. He is also interested in other bioinformatics problems including protein functional prediction, human chromosomal conformation, species phylogeny inference, and biological networks.