

**BIOINFORMATICS TOOL AND MODEL DEVELOPMENT
FOR STUDYING BIOLOGICAL NETWORKS AND
PROTEIN-PROTEIN INTERACTIONS**

ZHANG PENG

(B.Sc. (Hons.), National University of Singapore)

A THESIS SUBMITTED
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
DEPARTMENT OF PHARMACY
NATIONAL UNIVERSITY OF SINGAPORE

2016

Supervisor:
Professor Chen Yu Zong

Examiners:
Professor Sung Wing Kin
Associate Professor Zhang Louxin
Dr Verma Chandra, A*STAR

Declaration

*I hereby declare that the thesis is my original work and
it has been written by me in its entirety.*

*I have duly acknowledged all the sources of information
which have been used in the thesis.*

*This thesis has also not been submitted for any degree
in any university previously.*



ZHANG PENG

08 August 2016

Acknowledgements

Foremost, I would like to sincerely express my heartfelt gratitude to my PhD supervisor, Professor Chen Yu Zong, whose academic advice, insightful vision, and spiritual encouragement have sparked my scientific inspiration and rigorous thinking along my four years of PhD study. Prof Chen's elaborate guidance and extraordinary perspectives has enabled my enjoyable explorations in the field of computational biology and biocheminformatics. He has enlightened my philosophy of research and life, and all of which I learnt from him are truly my life-long spirit and backbone. My wholehearted appreciation, admiration and respect to Professor Chen Yu Zong is much more than that I can express. Also, I would like to give my best wishes to his beloved family.

Secondly, I am very grateful for the helps from our BIDD (BioInformatics and Drug Design) group members, particularly Dr. Zhu Feng and Dr. Liu Xin who mentored my FYP and UOPS projects when I was an undergraduate student majoring computational biology in NUS, and they are the ones who brought me into this BIDD family. Furthermore, I would like to thank the BIDD members, including Dr. Tao Lin, Dr. Zhang Chen, Dr. Qin Chu, Ms. Chen Shangying and Mr. Zeng Xian. The friendship, discussions and collaborations with them had greatly promoted my academic knowledge and enriched my research experience.

Moreover, I shall thank Professor Roger Kamm from MIT, who granted me a student research fellowship in SMART BioSyM during the summer 2012, which well bridged me from the undergraduate study to the graduate research.

This opportunity had offered me the experience in bio-imaging processing and triggered my senses of doing research. I should also give my gratefulness to Associate Professor Go Mei Lin, Associate Professor Brain Dymock and Professor Greg Tucker-Kellogg for giving me the chances as a teaching assistant in (PR3101 Medicinal Chemistry) in Department of Pharmacy, and (LSM2241 Bioinformatics) in YLL School of Medicine. These experience had largely broaden my scientific vision and solidified my knowledge foundations.

Finally, my utmost gratefulness goes to my beloved parents and wife, for their everlasting love and support. I would thank my parents for giving birth to me in the delicate city of Yang Zhou in Jiangsu China, and raising me up as a decent and upright young man. I also heartily thank my lovely wife, Ms. Yang Zhou (the same name as my hometown), for being extraordinarily supportive and considerate to me all the time. The warmth from the family is always the strongest power in my heart.

It is my greatest pleasure to have all of them in my PhD journey. Thank you all!



ZHANG PENG

08 August 2016

Table of Contents

| | |
|--|-------|
| Declaration | I |
| Acknowledgements | II |
| Table of Contents | IV |
| Summary | VIII |
| List of Tables | XI |
| List of Figures | XIII |
| List of Abbreviations | XV |
| List of Publications | XVIII |
| CHAPTER 1 Introduction | 1 |
| 1.1 Introduction of Network Descriptors in Systems Biology | 1 |
| 1.1.1 Origins of Network Descriptors | 1 |
| 1.1.2 Applications of Network Descriptors in Systems Biology | 5 |
| 1.1.3 Publicly Accessible Tools for Computing Network Descriptors.. | 7 |
| 1.2 Introduction of Tissue-Specific Protein Interaction Networks..... | 10 |
| 1.2.1 Studies on Tissue-Specific Protein Interaction Networks | 10 |
| 1.2.2 Databases for Tissue-Specific Protein Interaction Networks | 13 |
| 1.3 Introduction to the Prediction of Protein-Protein Interaction Kinetic Constants | 17 |
| 1.3.1 Inhibition of Protein-Protein Interaction in Drug Discovery | 17 |
| 1.3.2 Knowledge of Kinetic Constants in Molecular Interactions | 21 |
| 1.3.3 Prediction of Protein-Protein Interaction and its Kinetic Constants | 24 |
| 1.4 Objectives and Outline of the Research Described in this Thesis..... | 27 |
| 1.4.1 Overall Objectives | 27 |
| 1.4.2 Overall Outline..... | 30 |

| | |
|---|-----------|
| CHAPTER 2 PROFEAT Webserver Development for Computing Biological Network Descriptors | 32 |
| 2.1 Background and Motivations | 32 |
| 2.2 Materials and Methods | 43 |
| 2.2.1 Network Descriptor Computational Methods | 43 |
| 2.2.2 Network File Format | 47 |
| 2.2.3 Performance Evaluation Methods | 49 |
| 2.3 Results | 53 |
| 2.3.1 PROFEAT Network Module Structure and Access..... | 53 |
| 2.3.2 Input and Output in PROFEAT Network Module | 57 |
| 2.3.3 Comparative Performance Evaluations..... | 66 |
| 2.4 Discussion | 72 |
| 2.4.1 Applications of network descriptors in genome-derived networks | 72 |
| 2.4.2 Applications of network descriptors in interactome-derived networks..... | 73 |
| 2.4.3 Applications of network descriptors in transcriptome-derived networks..... | 75 |
| 2.4.4 Applications of network descriptors in metabolome-derived networks..... | 77 |
| 2.4.5 Applications of network descriptors in diseasome-derived networks..... | 78 |
| 2.4.6 Perspectives..... | 79 |
| CHAPTER 3 TISPIN Database Development for Human Tissue-Specific Protein Interaction Networks | 81 |
| 3.1 Background and Motivations | 81 |
| 3.2 Materials and Methods | 86 |
| 3.2.1 Data Source | 86 |

| | |
|--|------------|
| 3.2.2 Generation of Network Information | 89 |
| 3.3 Results | 92 |
| 3.3.1 TISPIN Database Structure and Access..... | 92 |
| CHAPTER 4 Quantitative Sequence-Kinetic Constants Relationship for Predicting Protein-Protein Interaction Kinetic Constants..... | 99 |
| 4.1 Background and Motivations | 99 |
| 4.2 Materials and Methods | 102 |
| 4.2.1 Data Collection | 102 |
| 4.2.2 Calculation of Protein-Protein Interaction Features..... | 105 |
| 4.2.3 Machine Learning Method (Support Vector Regression)..... | 107 |
| 4.2.4 Machine Learning Method (Random Forests) | 109 |
| 4.2.5 Performance Evaluation | 110 |
| 4.2.6 Workflow of QSKR Study..... | 112 |
| 4.3 Results | 113 |
| 4.3.1 QSKR Prediction Performance on K_d Dataset..... | 113 |
| 4.3.2 QSKR Prediction Performance on kon and koff Datasets | 116 |
| 4.4 Conclusion and Discussion | 120 |
| CHAPTER 5 Concluding Remarks..... | 122 |
| 5.1 Major Finding and Contributions..... | 123 |
| 5.1.1 Merits of Upgrading PROFEAT Webserver for Computing Biological Network Descriptors..... | 123 |
| 5.1.2 Merits of Developing TISPIN Database for Providing Human Tissue-Specific Protein Interaction Networks | 125 |
| 5.1.3 Merits of Studying the Quantitative Sequence-Kinetic Constants Relationship to Predict Protein-Protein Interaction Kinetic Constants | 126 |
| 5.2 Limitations and Suggestions for Further Studies | 127 |
| 5.2.1 Limitations and Suggestions for PROFEAT Webserver | 127 |

| | |
|--|------------|
| 5.2.2 Limitations and Suggestions for TISPIN Database | 129 |
| 5.2.3 Limitations and Suggestions for QSKR Study | 130 |
| BIBLIOGRAPHY | 131 |
| APPENDICES | 147 |
| Section A: Network Descriptors in PROFEAT Webserver..... | 147 |
| Section B: Definitions and Algorithms of Network Descriptors | 156 |
| B.1 Node-Level Descriptors | 157 |
| B.2 Network-Level Descriptors | 169 |
| B.3 Edge-Level Descriptors..... | 191 |
| Section C: Protein-Protein Interaction Dataset for Studying Equilibrium Dissociation Constant (K_d) | 192 |
| Section D: Protein-Protein Interaction Dataset for Association Rate Constant (k_{on}) and Dissociation Rate Constant (k_{off})..... | 210 |

Summary

This thesis described my research in the development of bioinformatics tools (PROFEAT webserver, TISPIN database) to facilitate the study of complex biological networks, and the construction of machine-learning models for predicting the protein-protein interaction (PPI) kinetic constants, to facilitate the quantitative investigation of the drug competitive binding to the PPIs.

To cater to the extensive needs of quantitative analysis of biological, disease, and pharmacological networks, PROFEAT webserver was upgraded with a new module (<http://bidd2.nus.edu.sg/cgi-bin/profeat2016/main.cgi>) for computing the biological network descriptors. This tool offered some distinguished advantages over the other publically accessible tools by: 1) providing the most comprehensive and diverse (up-to 379 vs 3~100 in other tools) network descriptors at node/edge-level (local properties), and network-level (global properties); 2) covering different network types (undirected/directed, unweighted/weighted edges or nodes) for representing different kinds of biological networks (binary/oriented, constant/varying binding constants or molecular levels); 3) offering user-friendly access; 4) supporting different network formats to be compatible with the major network software; and 5) enabling the automatic detection, split, and computation of multiple disconnected networks from a single input. PROFEAT would considerably facilitate the functional biological investigations by providing the systematic properties of molecular interaction networks, offering the expanded understandings of biological complex systems, and revealing the higher-level clues of the mechanisms.

As another aspect of biological networks, tissue-specific networks have transcended the global interaction network with more precise focus and improved capability at the tissue/cell-level for studying functional biology, disease/drug-response mechanism, and discovery of the biomarkers/targets for the diagnostics/therapeutics of the diseases. Therefore, a new database TISPIN (<http://bidd2.nus.edu.sg/TISPIN/home.php>) was developed to provide the comprehensive information on human TIssue-Specific Protein Interaction Networks. Currently, TISPIN prototype 1.0 is fully functional, with well-designed web interface and architecture. It has several improved features over the existing resources, by delivering: 1) network files in various formats compatible with the major network software; 2) network visualizations; 3) network descriptors at node-level (each protein) and network-level (the entire network); 4) extensive annotations of protein name, gene symbol, UniProt ID, NCBI ID, biological process/cellular component/molecular function, and therapeutic targets; and 5) downloadable links.

Studies on PPI networks have greatly facilitated the understandings of functional biology. Specifically, some certain PPIs have been explored as the potential therapeutic targeting space in drug discovery, where the drug potency is highly depended on the competitive advantages over the substrates of targets, but the researchers are sometimes less clear about the potency needed to ensure the drug to be competitive against the protein partners (the binding affinity of a drug to the target must be stronger than that of the substrate to the target). Therefore, there is a need to predict PPI kinetics (K_d , k_{on} , and k_{off}), so as to determine the needed potencies for drugs to inhibit PPIs. A study on Quantitative-Sequence-Kinetic-Constants-Relationship (QSKR) was conducted

by applying regression algorithms on the newly expended PPI datasets, to predict the kinetic constants by using the features from primary sequences. the best models achieved $R_{test}^2 = 0.63, 0.55, 0.67$ for K_d , k_{on} , k_{off} datasets respectively in cross-validation.

List of Tables

CHAPTER 1

| | |
|--|----|
| Table 1-1 Examples of some typical representations of networks | 4 |
| Table 1-2 List of the network descriptors provided by the publically accessible tools that do not require programing skill..... | 9 |
| Table 1-3 Inhibitors of p53-MDM2 interaction | 18 |

CHAPTER 2

| | |
|--|----|
| Table 2-1 The supported network types with different biological representations, and the number of network descriptors (both the full-set and the slim-set) computed in PROFEAT | 35 |
| Table 2-2 Typical interpretations and biological implications of the slim set of network descriptors | 36 |
| Table 2-3 The number of network descriptors, the list of network types, and visualization features of PROFEAT and other publically accessible tools..... | 39 |
| Table 2-4 List of the network descriptors (node-level & network-level) in different categories and their selected applications in systems biology | 40 |
| Table 2-5 Ten tissue-specific PPI networks for CPU running time evaluation..... | 50 |
| Table 2-6 Job execution procedures needed by PROFEAT and other public tools to compute the selected descriptors of an undirected unweighted network | 52 |
| Table 2-7 The required file(s) for each input network type | 58 |
| Table 2-8 Sample input and output of an undirected un-weighted network | 60 |
| Table 2-9 Sample input and output of an undirected edge-weighted network..... | 61 |
| Table 2-10 Sample input and output of an undirected node-weighted network..... | 62 |
| Table 2-11 Sample input and output of an undirected edge-node-weighted network | 63 |
| Table 2-12 Sample input and output of a directed un-weighted network | 64 |
| Table 2-13 Sample input and output of a single file with disconnected networks | 65 |

| | |
|---|----|
| Table 2-14 CPU time in computing the slim set of PROFEAT network descriptors for 10 human tissue-specific PPI networks of 5 network types..... | 66 |
|---|----|

| | |
|--|----|
| Table 2-15 Comparison of the computed network descriptor value and the job execution time for three human tissue-specific PPI networks (A. hippocampus, B. muscle, and C. ovary) by PROFEAT and other public tools NetworkX, Cytoscape and Gephi..... | 69 |
|--|----|

CHAPTER 3

| | |
|--|----|
| Table 3-1 Comparison between TISPIN 1.0 and the other relevant databases (TissueNet, SPECTRA, and IID) that provide tissue-specific PPI networks..... | 85 |
|--|----|

| | |
|--|----|
| Table 3-2 Human tissues/cells covered in TISPIN database..... | 88 |
|--|----|

CHAPTER 4

| | |
|---|-----|
| Table 4-1 PROFEAT protein feature categories, descriptions, and dimensions | 106 |
|---|-----|

| | |
|--|-----|
| Table 4-2 The best QSKR model performance in predicting the K_d value | 113 |
|--|-----|

| | |
|---|-----|
| Table 4-3 The best QSKR performance in predicting the k_{on} and k_{off} value | 116 |
|---|-----|

APPENDICES

| | |
|--|-----|
| Table S-1 List of the node-level descriptors covered in PROFEAT | 148 |
|--|-----|

| | |
|--|-----|
| Table S-2 List of the network-level descriptors covered in PROFEAT..... | 150 |
|--|-----|

| | |
|--|-----|
| Table S-3 List of the edge-level descriptors covered in PROFEAT | 155 |
|--|-----|

| | |
|---|-----|
| Table S-4 Protein-protein interaction dataset for studying equilibrium dissociation constants (K_d)..... | 192 |
|---|-----|

| | |
|--|-----|
| Table S-5 Protein-protein interaction dataset for studying association rate constant (k_{on}) and dissociation rate constants (k_{off}) | 210 |
|--|-----|

List of Figures

CHAPTER 1

| | |
|---|----|
| Figure 1-1 Schematic regulation of p53 by MDM2 | 18 |
| Figure 1-2 Scheme of bimolecular kinetics and thermodynamics | 22 |

CHAPTER 2

| | |
|---|----|
| Figure 2-1 Graphic illustration of the network descriptors (degree, selfloop, triangle, clustering coefficient, closeness centrality, betweenness centrality, and eccentricity) in a hypothetic network..... | 44 |
| Figure 2-2 Homepage of PROFEAT webserver 2016 | 54 |
| Figure 2-3 Biological network descriptor module in PROFEAT webserver | 55 |
| Figure 2-4 Computational flowchart for PROFEAT network descriptors | 56 |
| Figure 2-5 CPU time (mins) in computing the slim set of PROFEAT network descriptors for the networks in Table 2-14 with respect to the number of nodes (left) and the number of edges (right) | 67 |

CHAPTER 3

| | |
|--|----|
| Figure 3-1 Distribution of TISPIN covered tissues/cells in human systems..... | 87 |
| Figure 3-2 Schematic of the data sources and the incorporated information in TISPIN | 90 |
| Figure 3-3 Network visualization example for “T cell”: (A) the global protein interaction network and (B) the largest connected protein interaction network | 91 |
| Figure 3-4 Home page of TISPIN database | 93 |
| Figure 3-5 Search result page by an example quick search for “Immune” | 94 |
| Figure 3-6 Network detail page for example “T Cell” | 96 |

CHAPTER 4

| | |
|---|-----|
| Figure 4-1 Distribution of kinetic constants value in dataset (K_d , k_{on} , k_{off})..... | 104 |
| Figure 4-2 Schematic of support vector regression..... | 108 |
| Figure 4-3 Schematic of regressive random forests | 109 |
| Figure 4-4 Workflow of QSKR study to predict PPI kinetic constants | 112 |
| Figure 4-5 The best QSKR model performance in predicting the K_d value..... | 114 |
| Figure 4-6 Heat plot of K_d prediction performance in SVR parameter optimization | 115 |
| Figure 4-7 Plot of the predicted K_d value versus the actual K_d value by using the best QSKR model in K_d dataset..... | 115 |
| Figure 4-8 The best QSKR model performance in predicting k_{on} and k_{off} value | 117 |
| Figure 4-9 Heat plot of k_{on} prediction performance in SVR parameter optimization | 118 |
| Figure 4-10 Plot of the predicted k_{on} value versus the actual k_{on} value by using the best QSKR model in k_{on} dataset | 118 |
| Figure 4-11 Heat plot of k_{off} prediction performance in SVR parameter optimization | 119 |
| Figure 4-12 Plot of the predicted k_{off} value versus the actual k_{off} value by using the best QSKR model in k_{off} dataset..... | 119 |

List of Abbreviations

| | |
|------------------------|--|
| ArrayExpress | Repository for Microarray Gene Expression Data at EBI |
| BIDD | BioInformatics and Drug Design group, NUS |
| BioGRID | Biological General Repository for Interaction Datasets |
| CYJS | Cytoscape-defined Network File Format |
| Cytoscape | Platform for Complex Network Analysis Visualization |
| DIP | Database of Interacting Proteins |
| EBI | European Bioinformatics Institute |
| FDA | Food and Drug Administration |
| GEO | Gene Expression Omnibus at NCBI |
| Gephi | Graph Visualization Platform |
| GO | Gene Ontology |
| GraphWeb | Webserver for Graph Analysis of Biological Networks |
| GUI | Graphical User Interface |
| HIPPIE | Human Integrated Protein-Protein Interaction Reference |
| HPRD | Human Protein Reference Database |
| Hubba | Hub Objects Analyzer |
| igraph | R Package for Network Analysis |
| IID | Integrated Interactions Database |
| InnateDB | Innate Immunity Interactions & Pathways Database |
| IntAct | IntAct Molecular Interaction Database |
| K_d | Equilibrium Dissociation Constant |
| k_{off} | Rate of Dissociation |
| k_{on} | Rate of Association |
| KDBI | Kinetic Data of Biomolecular Interactions Database |

| | |
|----------------------|---|
| LOOCV | Leave-One-Out Cross Validation |
| MINT | Molecular Interaction database |
| MS/MS | Tandem mass spectrometry |
| MSE | Mean Squared Error |
| MySQL | Open-Source Relational Database Management System |
| NAViGaTOR | Software for Network Analysis, Visualization, Graphing |
| NCBI | National Center for Biotechnology Information |
| NCI | National Cancer Institute |
| NET | Pajek NET Network File Format |
| NetworkX | Python Package for Complex Networks |
| OMIM | Online Mendelian Inheritance in Man |
| Pajek | Program for Large Network Analysis |
| PDB | Protein Data Bank |
| PDBBind | Binding Affinity Database for Protein-Ligand Complex |
| PINA | Protein Interaction Network Analysis Platform |
| PLS | Partial Least Squares |
| PPI | Protein-Protein Interaction |
| PROFEAT | Protein Feature Server |
| Q² | Coefficient of Determination for Testing in Validation |
| QSKR | Quantitative Sequence-Kinetics Relationship |
| QuACN | R Package for Quantitative Analysis of Networks |
| R² | Coefficient of Determination for Training in Validation |
| RBF | Gaussian Radial-basis Kernel Function |
| RF | Random Forest |
| RNAseq | RNA Sequencing |
| SIF | Simple Interaction File Format |

| | |
|--------------------|--|
| SPECTRA | Specific Tissue/Tumor Related PPI Networks Analyzer |
| SpectralNET | Application for Spectral Graph Analysis, Visualization |
| STRING | Search Tool for Retrieval of Interacting Gene/Protein |
| SVG | Scalable Vector Graphics |
| SVM | Support Vector Machine |
| SVR | Support Vector Regression |
| TCGA | The Cancer Genome Atlas |
| TISPIN | Tissue-Specific Protein Interaction Network Database |
| TissueNet | Tissue Protein-Protein Interaction Network Database |
| TTD | Therapeutic Target Database |
| tYNA | The Yale Network Analyzer |
| UniProt | Universal Protein Resource |
| VANESA | Software for Visualization & Analysis of Systems Biology Networks |
| VisANT | Visualization and Analysis Tool for Biological Interaction Data |
| XML | Extensible Markup Language Network File Format |

List of Publications

Publications:

1. **P. Zhang**, L. Tao, X. Zeng, C. Qin, S.Y. Chen, F. Zhu, Z.R. Li, Y.Y. Jiang, W.P. Chen, and Y.Z. Chen. A Protein Network Descriptor Server and Investigation of Protein, Disease, Metabolic and Drug Targeted Networks. *Briefings in Bioinformatics* pii: bbw071 (2016) **IF: 8.399**

2. **P. Zhang**, L. Tao, X. Zeng, C. Qin, S.Y. Chen, F. Zhu, Z.R. Li, and Y.Z. Chen. PROFEAT Update: A Protein Features Web-Server with Added Facility to Compute Network Descriptors for Studying Omics-Derived Networks. *Journal of Molecular Biology (Special Issue for Computation Resources)* pii: S0022-2836(16)30428-4 (2016) **IF: 4.333**

3. Y.H. Li, J.Y. Xu, L. Tao, X. Zeng, S.Y. Chen, **P. Zhang**, C. Qin, C. Zhang, F. Zhu, and Y.Z. Chen. SVM-Prot 2016: A Web-Server for Machine Learning Prediction of Protein Functional Families from Sequence Irrespective of Similarity. *PLOS ONE* 11(8):e0155290 (2016) **IF: 3.234**

4. S.Y. Chen, C. Qin, J.E. Sin, X. Yang, L. Tao, X. Zeng, **P. Zhang**, C.M. Gao, Y.Y. Jiang, C. Zhang, Y.Z. Chen, and W.K. Chui. Discovery of Novel Dual VEGFR2 and Src Inhibitors using a Multi-Step Virtual Screening Approach. *Future Medicinal Chemistry* 9(1):7-24 (2016) **IF: 3.345**

5. S.Y. Chen, **P. Zhang**, X. Liu, Q. Chu, L. Tao, C. Zhang, S.Y. Yang, Y.Z. Chen, and W.K. Chui. Towards Cheminformatics-based Estimation of Drug Therapeutic Index: Predicting the Protective Index of Anticonvulsants using a new Quantitative Structure-Index Relationship Approach. *Journal of Molecular Graphics and Modelling* 67:102-110 (2016) **IF: 1.674**

6. L. Tao, **P. Zhang**, C. Qin, S.Y. Chen, C. Zhang, Z. Chen, F. Zhu, S.Y. Yang, Y.Q. Wei, and Y.Z. Chen. Recent Progresses in the Exploration of Machine Learning Methods as in-silico ADME Prediction Tools. *Advanced Drug Delivery Reviews* 86:83-100 (2015) **IF: 15.04**

7. L. Tao, F. Zhu, C. Qin, C. Zhang, S.Y. Chen, **P. Zhang**, C.L. Zhang, **IF:**
C.Y. Tan, C.M. Gao, Y.Y. Jiang, and Y.Z. Chen. Clustered Distribution **5.228**
of Natural Product Drug Leads in the Chemspace Influenced by the
Privileged Target-Sites. *Scientific Reports* 5:9325 (2015)

8. C. Zhang, L. Tao, C. Qin, **P. Zhang**, S.Y. Chen, X. Zeng, F. Xu, Z. Chen, **IF:**
S.Y. Yang, and Y.Z. Chen. CFam: A Chemical Families Database Based **9.112**
on Iterative Selection of Functional Seeds and Seed-Directed Compound
Clustering. *Nucleic Acids Research* 43:D558-65 (2015)

9. C. Qin, L. Tao, Y.H. Pang, C. Zhang, S.Y. Chen, **P. Zhang**, Y.Y. Jiang, **IF:**
and Y.Z. Chen. The Assessment of the Readiness of Molecular **5.228**
Biomarker-Based Mobile Health Technologies for Healthcare
Applications. *Scientific Reports* 5:17854 (2015)

10. C. Zhang, C. Qin, L. Tao, F. Zhu, S.Y. Chen, **P. Zhang**, S.Y. Yang, Y. **IF:**
Q. Wei, and Y.Z. Chen. A Resource for Facilitating the Tools in **7.903**
Education and Implementation of Genomics-Informed Personalized
Medicine. *Clinical Pharmacology & Therapeutic* 95:590-591 (2014)

11. C. Qin, C. Zhang, F. Zhu, F. Xu, S.Y. Chen, **P. Zhang**, Y.H. Li, S.Y. **IF:**
Yang, Y.Q. Wei, L. Tao, and Y.Z. Chen Therapeutic Target Database **9.112**
Update 2014: Resource for Facilitating Bench-to-Clinic Discovery,
Investigation, Application, and Management of Targeted Therapeutics.
Nucleic Acids Research 42(1):D1118-23 (2014)

12. F. Zhu, Z. Shi, C. Qin, L. Tao, X. Liu, F. Xu, L. Zhang, X.H. Liu, J.X. **IF:**
Zhang, B.C. Han, **P. Zhang**, and Y.Z. Chen. Therapeutic Target **9.112**
Database Update 2012: Resource for Facilitating Target-Oriented Drug
Discovery. *Nucleic Acids Research* 40(D1):D1128-D1136 (2012)

Manuscripts:

13. L. Tao, Y.F. Zhong, S.H. Paw, X. Zeng, C. Qin, **P. Zhang**, S.Y. Chen, W.D. He, Y. Tan, H.X. Liu, Y.Y. Jiang, W.P. Chen, and Y.Z. Chen. The Database and Bioinformatics Studies of Probiotics. *Briefings in Bioinformatics* (2017) *[UNDER REIVEW]* **IF: 8.399**

14. X. Zeng, L. Tao, **P. Zhang**, Q. Chu, S.Y. Chen, W.D. He, H.X. Liu, S.Y. Yang, Y.Y. Jiang, and Y.Z. Chen. HEROD: A Human Ethnic and Regional Specific Omics Database. *Bioinformatics* (2016) *[UNDER REIVEW]* **IF: 5.766**

15. C. Zhang, Y.M. Shao, X.H. Ma, S. Cheong, C. Qin, L. Tao, **P. Zhang**, S.Y. Chen, X. Zeng, H.X. Liu, G. Pastorin, Y.Y. Jiang, and Y.Z. Chen. Pharmacological Relationships and Ligand Discovery of G Protein-Coupled Receptors Revealed by Simultaneous Ligand and Receptor Clustering. *Journal of Chemical Information and Modeling* (2016) *[UNDER REIVEW]* **IF: 3.738**

CHAPTER 1 Introduction

1.1 Introduction of Network Descriptors in Systems Biology

1.1.1 Origins of Network Descriptors

A network/graph is basically a collection of points (nodes/vertices in network theory, proteins in protein-protein networks) connected in pairs by lines (edges/links in network theory, protein-protein interactions in protein-protein networks). The concept of network was first introduced and best developed in the area of sociology, by using the network to represent a group of people with interpersonal relationships, where each node represents a person and each edge represents the relationship between two individuals¹.

For a small network, the visual inspection might be good enough to identify the important nodes and the connection patterns of a network. However, when the network size grows, visual inspection will be no longer effective in capturing the complex network features and patterns. To address this problem, the network/graph theory has been well developed and extended to use a number of network descriptors (or properties, features, measures) for representing various patterns and characteristics of a network, and these mathematical and statistical network descriptors are mostly originated directly or indirectly from sociology¹. These network descriptors are generally grouped into node-level, edge-level, and network-level, where the node-level descriptors are to measure the local properties (e.g. degree, centrality, etc.) for each node or member in the network, the edge-level descriptors are to measure the properties for each edge or link (e.g. edge weight, edge betweenness centrality, etc.), and the network-level

descriptors are to measure the global properties (e.g. complexity, efficiency, etc.) of the entire network or system. Network descriptors have been shown to be very helpful and powerful to understand the structure and patterns of a network quantitatively, which lead to useful and deep insights of a system¹.

As early as 1950s-1990s, the ideas of network descriptors have been gradually emerged. Two parameters (compactness and stress) were defined to characterize the internal network structure in the communication networks². The centrality indices, representing the degree of importance for each node in the network, were then introduced and applied in the sociological and psychological studies^{3,4}. The clustering coefficient, which indicates the tendency of forming clusters or groups in a network, was created in studying the interpersonal relationships⁵ and then further developed in studying psychology⁶. The sociologist *Linton Freeman* introduced the betweenness centrality as a new measure of individual's importance in a network³. After some time, several network descriptors (e.g. eccentricity, variation, deviation, unipolarity, etc.) were proposed by mathematicians for studying chemistry⁷.

Some of the above mentioned network properties (degree, clustering coefficient, and betweenness centrality) are described here: **Degree** " deg_i " of node i is the number of edges linked to it. **Clustering Coefficient**^{8,9} of node i is defined as " $cluster_i = 2e_i / (deg_i (deg_i - 1))$ ", where e_i is the number of connected pairs between all neighbours of node i . **Betweenness Centrality**^{10,11} quantifies the number of times a node serving as a linking bridge along the shortest path between two other nodes, by " $centralityBtw_i = \sum_{s \neq i \neq t} \sigma_{st}(i) / \sigma_{st}$ ", where s and t are two different nodes from node i , σ_{st} is the number of shortest paths from s

to t , and $\sigma_{st}(i)$ is the number of shortest paths from s to t via i . It reflects the extent of control of that node exerting over the interactions with other nodes in the network.

Since 2000, there have been some more network descriptors evolved in various areas. Network efficiency was introduced to measure how efficient the information is exchanged within a network, and well applied in analyzing the communication and transportation networks¹². A group of topological indices (e.g. alpha index, beta index, Pi index, eta index, hierarchy, etc.) were applied as the geographical measures for the transport systems¹³. PageRank centrality was developed by website search engine Google to determine the importance of each website¹⁴. Node-weighted clustering coefficient was proposed to analyze the networks with heterogeneous node weights, in the study of Earth's spatial network and international trade network¹⁵.

Some of the above mentioned network properties (efficiency, hierarchy, and PageRank centrality) are described here: **Efficiency**¹² measures information exchange efficiency across the network to determine the cost-effectiveness of the connection structure, by the equation " $efficiency_G = (\sum_{i \neq j}^N \frac{1}{D_{ij}}) / N(N-1)$ ". **Hierarchy**¹³ is the gradient of the linear power-law regression, by fitting the \log_{10} (node frequency) over the \log_{10} (degree distribution). Hierarchy is calculated by " $y = a \cdot x^{hierarchy}$ ", where x is the degree distribution and y is the node frequency of that degree. **PageRank Centrality**^{16,17,18,19,20,21} is an algorithm implemented in Google search engine to rank the websites, according to the webpage connections in the World Wide Web. It initializes the PageRank

centralities to an equal probability value $1/N$ for all nodes. The equation

" $pageRank_i = \frac{1-d}{N} + d \cdot \sum_{j=1}^N A_{ij} \cdot \frac{pageRank_j}{deg_j}$ " iteratively updates the centrality

value by using a constant damping factor d , its neighbors' PageRank centrality value, and its degree value. The algorithm stops running, when the PageRank centrality converges.

With the more and more successful applications of the network descriptors, a new realm has been opened – network biology, which will be discussed in the next **Section 1.1.2**.

Table 1-1 Examples of some typical representations of networks

| Network | | Node | Edge |
|----------------|-----------------------------|--------------------|------------------------|
| Non-Biological | Friendship Network | Person | Friendship |
| | World Wide Web | Web Page | Hyperlink |
| | International Trade Network | Country | Trade Relation |
| | Land Transportation Network | Station & Terminal | Road & Railway |
| Biological | Protein Interaction Network | Protein | Protein Interaction |
| | Gene Co-Expression Network | Gene | Expression Correlation |
| | Metabolic Network | Metabolite | Metabolic Reaction |
| | Neural Network | Neuron | Synapse |

1.1.2 Applications of Network Descriptors in Systems Biology

Recently, networks have been widely used in many biological investigations, as a representation of the connections (e.g. physical interactions, regulatory relationships, co-expressions) between the appropriate biological elements (e.g. proteins, genes, metabolites, drugs)^{1,22}. A variety of network descriptors initially developed for capturing the structural patterns in the areas of sociology, mathematics, physics, economics, have been applied for the quantitative analysis of biological networks, which are increasingly required for more extensive investigations of biological^{8,22,23,24,25}, disease^{26,27,28,29,30} and pharmacological^{31,32,33,34,35} processes. These analyses have been largely facilitated by the knowledge of the network descriptors that characterize the connectivity, topology and complexity properties of the relevant protein-protein interaction networks, gene regulatory networks, gene co-expression networks, metabolic networks, and drug-target networks.

Some of the network descriptors have been used for studying the biological networks, which to a large extent share the same architectural features with other complex networks⁸. So far, the well-established graph theory, from the fields of mathematics and computer science, has revealed the enrichment patterns, systematic understandings, high-level relationships, and network-based clues in various biological networks^{22,29,36}. For instance, the betweenness centrality has been employed for the assessment of protein druggability based on the profiles of the drug targets in the protein network³⁷, and the modularity analysis of interaction information in a liver metabolism network¹¹. The clustering coefficient and topological coefficient have been used for analyzing the organizational and structural properties of human protein network³⁸. The

neighborhood connectivity has been applied for measuring the specificity and stability of topology in protein networks³⁹.

Some of the above mentioned network features (neighbourhood connectivity, and topological coefficient) are described here: **Neighborhood Connectivity**³⁹

of a node is the number of its neighbours, $neighbourConnect_i = \frac{\sum_{j=1}^N A_{ij} \cdot deg_j}{deg_i}$.

Topological Coefficient³⁸ estimates the tendency of the nodes to share neighbours by " $topology_i = avg \left\{ \frac{J(i,j)}{deg_i} \right\}$ ", using j represents all the nodes sharing at least one neighbour with i , and $J(i, j)$ is the number of shared neighbours between node i and node j .

Nonetheless, a substantial number of the network descriptors have not yet been used but are potentially useful for the analysis of a larger variety of features in biological networks. For instance, the geographical indices from the study of transport systems¹³ are potentially applicable for describing the spatial and structural properties of the biological networks, and the topological robustness from the study of social networks⁴⁰ can be potentially employed for measuring the overall robustness or the alternative signaling capability of biological networks. More example applications of network descriptors in systems biology were briefly summarized in **Table 2-4** in **Chapter 2** (PROFEAT Webserver Development for Computing Biological Network Descriptors).

1.1.3 Publicly Accessible Tools for Computing Network Descriptors

A number of resources are been publically available for computing the network descriptors, particularly the user-interface-based software or webservers:

1. Cytoscape⁴¹ (<http://www.cytoscape.org>)
2. NAViGaTOR⁴² (<http://ophid.utoronto.ca/navigator>)
3. Gephi⁴³ (<https://gephi.org>)
4. VANESA⁴⁴ (<http://vanesa.sf.net>)
5. Pajek⁴⁵ (<http://vlado.fmf.uni-lj.si/pub/networks/pajek>)
6. SpectralNET⁴⁶ (<https://www.broadinstitute.org/software/spectralnet>)
7. PINA⁴⁷ (<http://cbg.garvan.unsw.edu.au/pina>)
8. Hubba⁴⁸ (<http://hub.iis.sinica.edu.tw/cytoHubba>)
9. GraphWeb⁴⁹ (<http://biit.cs.ut.ee/graphweb/>)
10. tYNA⁵⁰ (<http://tyna.gersteinlab.org/tyna>)
11. VisANT⁵¹ (<http://visant.bu.edu/>)

These eleven software and webservers have enabled the computation of approximately 23, 13, 10, 10, 9, 9, 8, 6, 4, 4 and 3 network descriptors respectively (**Table 1-2**).

Moreover, users knowledgeable of the programming languages can use:

1. Python library NetworkX⁵²:
<https://networkx.github.io>
2. R package igraph⁵³:
<https://cran.r-project.org/web/packages/igraph/index.html>
3. R package QuACN⁵⁴:
<https://cran.r-project.org/web/packages/QuACN/index.html>

These three programming tools are able to compute about 100 network properties. However, they are hardly applicable for the users without computation expertise, especially the biologist⁵⁵. Furthermore, compared to the literature-reported network descriptors (**Table 2-4**, **Appendix Section A**, **Table S-1**, **S-2**, and **S-3**), these resources have only covered a limited number of network descriptors, while some of the uncovered network descriptors have already been successfully applied in systems biology studies. For instance, the PageRank centrality from Google search algorithm has been used for analyzing the metabolic networks and gene regulatory networks^{19,21,56}; the interconnectivity has been applied to prioritize the disease-associated genes^{57,58,59}, and the edge-weighted clustering coefficient has been utilized to predict the significant gene modules in the gene co-expression network^{60,61}.

Table 1-2 List of the network descriptors provided by the publically accessible tools that do not require programing skill

| Tool Name (No. of Descriptors) | List of Provided Network Descriptors | |
|--|---|--|
| | Node-Level | Network-Level |
| Cytoscape (23) | degree, in/out-degree, self-loops, clustering coefficient, topological coefficient, neighbourhood connectivity, avg shortest path length, eccentricity, radiality, stress, closeness centrality, betweenness centrality | number of nodes/edges/self-loops, density, diameter, radius, centralization, heterogeneity, avg neighbourhood, avg path length |
| NAViGaTOR (13) | clustering coefficient, degree centrality, betweenness centrality | number of nodes/edges, density, min/avg/max degree, diameter, avg clustering coefficient, avg path length |
| Gephi (10) | degree, clustering coefficient, betweenness centrality, closeness centrality, eigenvector centrality, HITS | diameter, avg clustering coefficient, density, avg shortest path length |
| VANESA (10) | degree, avg/max shortest path length | min/avg/max degree, density, characteristic shortest path length, centralization, clustering coefficient |
| Pajek (9) | degree, avg shortest path length, degree centrality, closeness centrality, betweenness centrality | diameter, degree centralization, closeness centralization, betweenness centralization |
| SpectralNET (9) | degree, clustering coefficient, min/avg/max shortest path length | number of nodes, diameter, avg clustering coefficient, characteristic shortest path length |
| PINA (8) | degree, shortest path length, clustering coefficient, closeness centrality, betweenness centrality, degree centrality, eigenvector centrality | diameter |
| Hubba (6) | degree, bottleneck, subgraph centrality, edge percolation component, max neighbourhood component, density of max neighbourhood | n.a. |
| GraphWeb (4) | betweenness centrality | number of nodes/edges, density |
| tYNA (4) | degree, clustering coefficient, eccentricity, betweenness centrality | n.a. |
| VisANT (3) | degree, shortest path length, clustering coefficient | n.a. |

1.2 Introduction of Tissue-Specific Protein Interaction Networks

1.2.1 Studies on Tissue-Specific Protein Interaction Networks

Tissue-specificity is an important aspect in the functional study of systems biology, and in the investigation of certain disease mechanisms and drug responses, as it reflects the different genetic types (e.g. protein subtypes), different expression levels, different interactions of the participating molecules, and different functional roles in diversely different tissues and cells^{62,63}. In the last few years, the increasing efforts have been directed at studying the human tissue-specific networks, where most of such papers were published after 2011. The research interests have been mainly focused on the comparison between the tissue-specific networks and the global networks^{64,65,66,67,68}, and the disease-related applications of tissue-specific networks^{62,63,69,70,71,72}.

Bossi et al. introduced the human tissue-specific protein-protein interaction networks⁶⁴ by combining the physical protein interactions and the tissue-specific gene expression from microarray experiments⁷³. *Bossi* defined those universally expressed proteins in all tissues as the house-keeping proteins, and those expressed only in a few tissues as the tissue-specific proteins⁶⁴. *Lin et al.* analyzed the topological and organizational properties of the house-keeping proteins and the tissue-specific proteins in 19 human tissue-specific PPI networks⁶⁵, by mapping the tissue-specific gene expression data from HuGE (Human Gene Expression) Index database⁷⁴ to the PPI database HPRD (Human Protein Reference Database)⁷⁵. *Lin* used three network descriptors (degree, betweenness centrality, and closeness centrality) for comparing the house-keeping proteins and the tissue-specific proteins against the expected mean of

randomly selected proteins in the network, which have shown certain house-keeping and tissue-specific behaviours. *Lin* found that the house-keeping proteins favor to occupy central positions in the network, while the tissue-specific proteins tend to be more peripheral⁶⁵. *Lopes et al.* also constructed the tissue-specific PPI networks by using physical PPI data and Affymetrix microarray gene expression data. Through the functional comparison between the tissue-specific networks and the global networks, *Lopes* observed the substantial enrichment of specific proteins and pathways in the tissue-specific networks, while in contrast, the global network had no significant functional enrichment, making the analysis difficult to produce any critical findings⁶⁶. As evident from these studies, topological properties and functional enrichment of the tissue-specific networks have been shown markedly different from the global network, and the analysis of global network instead of tissue-specific network would lead to the loss of biological information, and result in the misinterpretation of biological functions.

Tissue-specific interactome has also been intensively applied for studying the disease mechanisms. *Greene et al.* deemed that the understanding of tissue-specific networks is important for identifying the varying functional roles of genes and proteins across different tissues, and helpful in developing the improved diagnostics and therapeutics⁶³. *Dezso et al.* found that tissue-specific proteins were more likely to be drug targets and biomarkers, by analyzing the network topology and ontology enrichment of the tissue-specific networks⁶⁹. *Guan et al.* utilized the tissue-specific networks to predict the gene/protein candidates associated with certain phenotype or disease⁶². For instance, testis-specific network enabled the prediction of gene candidates related with male

fertility and spermatogenesis, which was also experimentally validated⁶². *Shahin et al.* proposed the method of using the tissue-specific interactome to study disease mechanism, by demonstrating a case study on brain-specific interactome for Alzheimer's and Parkinson's diseases, which implicated the disease-related pathways and the potential therapeutic targets⁷⁰. In some other studies, tissue-specific PPI networks have also improved the prioritization of disease-causing genes^{71,76}, and enhanced the understandings of the molecular mechanisms underlying the hereditary diseases⁷².

Based on these literature-reported applications of tissue-specific networks, we observed that the tissue-specific approaches surpass the traditional global methods with the remarkably improved capability in addressing the tissue/cell-level questions for functional biology, disease mechanism, and target/biomarker identification.

However, the tissue-protein relationships in these studies were always determined by the microarray gene expression data. Interestingly, *Emig et al.* re-evaluated the tissue specificity by using both microarray and RNAseq gene expression data to infer the tissue-specific networks. *Emig* observed many interactions, classified as highly tissue-specific by microarray data, were considerably found in all tissues from RNAseq data⁶⁷. This work concluded that microarray data is not sensitive enough for the low expressed genes, thus suggesting that microarray-based tissue distribution was less reliable^{67,77}. The low sensitivity of microarray technology in detecting gene expression has also been reported in many other studies, in comparison with sequencing technology^{78,79,80,81}. The advance of sequencing technology has offered us better

resolution in detecting the mRNA abundance, especially the low expressed ones. Nevertheless, the tissue distribution of proteins was still determined by transcriptional-level data, rather than protein-level expression evidence. While there is a recent study showing that the squared Pearson correlation coefficient (R^2) was ~ 0.4 between the transcriptional mRNA expression level and the protein abundance, implying $\sim 40\%$ of the variations in protein abundance could be explained by the mRNA expression⁸². The remaining $\sim 60\%$ of the protein abundance variations would require more post-transcriptional explanations⁸². Therefore, the tissue-specific protein networks, in which the tissue-protein associations derived from protein-level evidence, will be of great interest, due to its higher reliability.

1.2.2 Databases for Tissue-Specific Protein Interaction Networks

There have been a rapid growth of the publically accessible protein-protein interaction databases, famously BioGRID⁸³, DIP⁸⁴, HIPPIE⁸⁵, HPRD⁸⁶, InnateDB⁸⁷, IntAct⁸⁸, MINT⁸⁹, STRING⁹⁰, and so on. However, there are still a limited number of databases providing the tissue-specific protein-protein interaction networks (particularly TissueNet⁹¹, SPECTRA⁹², and IID⁹³), although the tissue-specific networks have been proofed to be very useful in investigating the functional biology and disease mechanism in tissue- or cell-level.

TissueNet⁹¹ database (<http://netbio.bgu.ac.il/tissuenet>), published in 2012, integrated 67,439 PPIs (from BioGRID⁸³, DIP⁸⁴, IntAct⁸⁸, and MINT⁸⁹) between 11,225 human proteins, and 16 human tissues profiled by microarray

and RNAseq. Users could only query one protein identifier and then select a tissue of interest in the homepage. TissueNet will return all the interacting partners (immediate neighbours) with tissue annotations in a single graphical network map. This map distinguishes the proteins expressed in ≥ 14 tissues in blue, proteins expressed in ≤ 3 tissues in orange, and the rest proteins in grey. Each node can be clicked for showing more protein information (name, Entrez, and tissue). However, “service error” frequently occurred when querying the TissueNet database, and unfortunately the download option is not available for users to conduct further study.

SPECTRA⁹² (<http://alpha.dmi.unict.it/spectra>), published in 2015, is an integrated knowledge base of human tissue-specific PPI networks, which combined 175,841 interactions (from BioGRID⁸³, DIP⁸⁴, HPRD⁸⁶, IntAct⁸⁸, and MINT⁸⁹) between 16,435 proteins, and gene expression data of 107 human normal tissues (from ArrayExpress, GEO, and TCGA). To search this database, users should firstly choose to search for all genes or some selected genes, secondly choose the tissues and specify the gene expression datasets, and thirdly choose the PPI datasets of interest. The returned output is a table of PPIs that matched with the user-defined settings, and these PPIs can also be visualized in a network panel. SPECTRA provides a link to download this table, which gives the names of Gene1 and Gene2, the expressions of Gene1 and Gene2, and the distributed tissues of each listed PPI.

IID⁹³ (Integrated Interactions Database) (<http://dcv.uhnres.utoronto.ca/iid>), published in 2016, provides the tissue-specific PPIs for 6 species (yeast, worm, fly, rat, mouse, and human) and up to 30 tissues. IID collected 1,566,043 protein

interaction data including both the experimentally detected PPIs (from BioGRID⁸³, DIP⁸⁴, HPRD⁸⁶, IntAct⁸⁸, InnateDB⁸⁷, and MINT⁸⁹) and the predicted PPIs from four published datasets^{94,95,96,97}. Its tissue distribution information was calculated based on eight microarray gene expression datasets downloaded from NCBI GEO database. To submit a query, users need to enter a protein or gene identifier, select the species and the tissues. Three output options (“View Results”, “Download Results”, and “Graphical Summary”) are provided, where the output result is a list of the interacting partners of the query. This table is downloadable, and it shows the query ID, query UniProt, partner UniProt, query symbol, partner symbol, species, and evidence type (experimental / predicted). However, the tissue distribution is neither provided in this table, nor found in the “Graphical Summary”. Although there is an option labeled as “Tissue” in the “Graphical Summary”, it is always empty or irresponsive in our testing.

All these findings and conclusions on TissueNet, SPECTRA, and IID databases were based on our observations till 08 Aug 2016.

These databases have made the groundbreaking contributions in providing the tissue-specific PPI networks, however there are still some major limitations and drawbacks:

- 1) Prior knowledge is required to search these databases, where users should be able to at least provide the names/symbols/IDs of genes or proteins.

- 2) The output is always a list of PPIs that the queried gene or protein is involved. TissueNet and SPECTRA provide the visualization of these PPIs as a graphical network, while IID only gives these PPIs in a table.
- 3) SPECTRA and IID have a link for downloading the output table in plain text format, but TissueNet does not have any download options.
- 4) The output tables downloaded from SPECTRA and IID are not compatible with any network analysis software, particularly Cytoscape, for further study.
- 5) None of the database provides the quantitative network properties / descriptors of the tissue-specific PPI networks, despite the network descriptors have already been widely used for analyzing the biological networks.

1.3 Introduction to the Prediction of Protein-Protein Interaction Kinetic Constants

1.3.1 Inhibition of Protein-Protein Interaction in Drug Discovery

In drug discovery research, efforts have been primarily directed at the searching of small molecules targeting to a specific protein, to inhibit the downstream functions of that protein. More recently, a new approach for drug discovery has gained increasing attentions with highly-expected promising potentials in principle, which is to find small molecules for targeting macromolecular complexes and disrupting protein-protein interactions⁹⁸. Protein-protein interactions are playing key roles in most biological processes, such as growth, differentiation, communication, and termination of the programmed cell death, hence blocking these PPIs would provide means to modulate the signalling activities. It represents an important target space for therapeutic intervention, suggesting a new avenue for the design of next generation of therapeutics^{98,99}.

As an example of therapeutic disruption in PPI, inhibition of *p53-MDM2* interaction has demonstrated its potentials in cancer treatment, which activates *p53* induction and its biological apoptotic responses for cancer cells¹⁰⁰ (**Figure 1-1**). The tumour suppressor *p53* induces cell death by apoptosis, in response to various stress conditions (e.g. DNA damage, activated oncogenes, hypoxia, etc.). With the loss of *p53* tumour-suppression activity, the cell will favour the development of cancer in a high proliferation rate. *MDM2*, a *p53*-specific *E3* ubiquitin ligase, is the cellular antagonist of *p53* acting to limit the *p53*-growth-suppressive function in cells. As a negative regulator of *p53*, *MDM2* binds to *p53* and inhibits the transcriptional activity which is driven by *p53*, resulting in

uncontrolled cell proliferation. Therefore, the inhibition of *p53-MDM2* interaction has become an emerging strategy to activate the apoptosis of *p53* for tumour treatment^{101,102}.

Figure 1-1 Schematic regulation of *p53* by *MDM2*

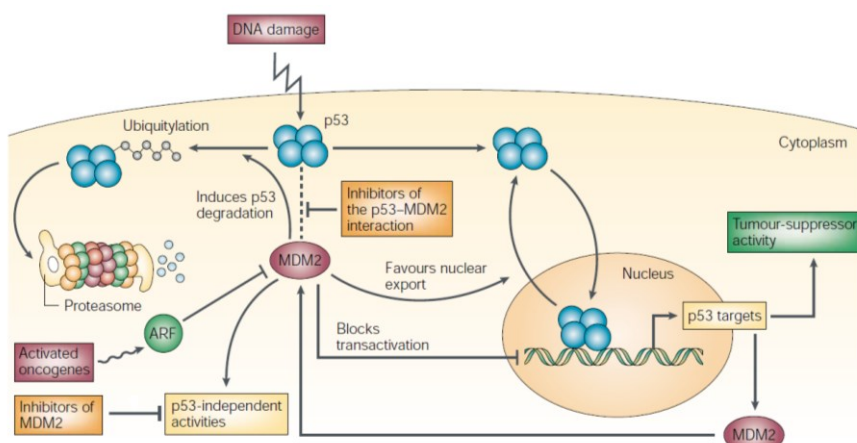
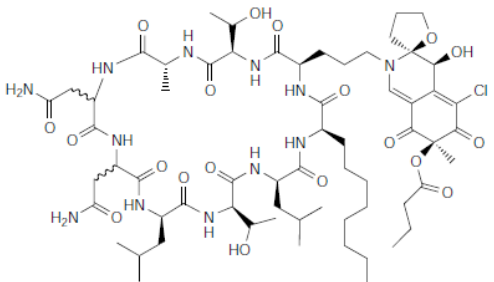


Table 1-3 Inhibitors of *p53-MDM2* interaction

| Peptide / Compound | Sequence / Structure | IC ₅₀ (μM) |
|--|--|-----------------------|
| Wild-type <i>p53</i> | Ac-Gln-Glu-Thr-Phe-Ser-Asp-Leu-Trp-Lys-Leu-Leu-Pro-NH ₂ | 8.7 |
| Phage-derived peptide | Ac-Met-Pro-Arg-Phe-Met-Asp-Tyr-Trp-Glu-Gly-Leu-Asn-NH ₂ | 0.3 |
| Truncated phage-derived peptide | Ac-Phe-Met-Asp-Tyr-Trp-Glu-Gly-Leu-Asn-NH ₂ | 8.9 |
| Constrained wide-type peptide | Ac-Glu-Thr-Phe-Aib-Asp-Aib-Trp-Lys-Aib-Leu-Aib-Glu-NH ₂ | 5.2 |
| Constrained peptide 3 | Ac-Phe-Met-Aib-Tyr-Trp-Glu-Ac-Leu-Asn-NH ₂ | 2.2 |
| Peptide 5 with a PMP at position 22 | Ac-Phe-Met-Aib-Pmp-Trp-Glu-Ac-Leu-Asn-NH ₂ | 0.3 |
| Peptide 6 with a 6ClTrp at position 22 | Ac-Phe-Met-Aib-Pmp-6ClTrp-Glu-Ac-Leu-Asn-NH ₂ | 0.005 |
| Chlorofusin |  | 4.6 |

Patrick Chene, from Novartis, published an article “Inhibiting the *p53-MDM2* interaction: an important target for cancer therapy” in *Nature Review Cancer*, in which he summarized some inhibitors for *p53-MDM2* interactions, including peptides, synthetic compounds, and natural products (**Table 1-3**). These inhibitors bind to *MDM2* binding site, such that repelling *p53* from interacting with *MDM2*¹⁰⁰. To achieve the PPI inhibition, *Patrick* deemed that the initial efforts is to obtain peptides or compounds having higher binding potency than *p53* wild-type peptide¹⁰⁰. It's observed that most of the inhibitors in **Table 1-3** demonstrating up-to thousand-fold increase in binding potency, compared with the wild-type *p53* peptide¹⁰⁰.

Therefore, the strength of the interaction between *p53* and *MDM2* plays as an important criterion in discovering the potent and effective inhibitors. Being determined by several methods, the experimental K_d (equilibrium dissociation constant) for *p53-MDM2* protein complex ranges from 60-700 nM, such that a good inhibitor for *p53-MDM2* interaction should have a lower K_d value with either *MDM2* or *p53*¹⁰⁰. *Grasberger*, from Johnson & Johnson laboratory, demonstrated their optimized benzodiazepine binding to *MDM2* with $K_d=80$ nM.

Besides *p53-MDM2* interaction, some more protein-protein interaction inhibitions having been attempted for therapeutic purposes, including β -catenin-*TCF* interaction, *JNK-JIP* interaction, *BAD-BAK* interaction, interaction of interleukin *IL-2* with its receptor, and so on^{98,101,103}.

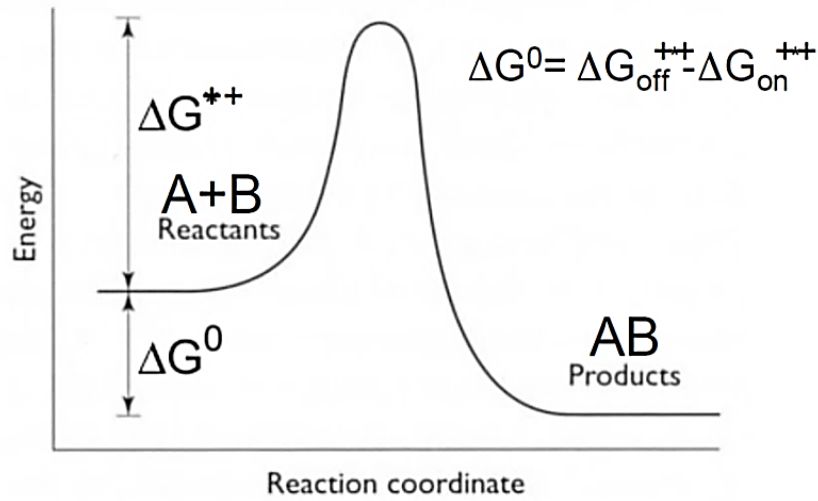
Based on these above facts, we learnt that the drug potency is highly depended on the competitive advantage over the substrates of drug targets, especially in

discovering new therapeutic agents in disrupting the protein-protein interactions. However, sometimes the researchers are less clear about what potency is necessary to ensure the drug competitive over its target substrate. In other words, for such a competitive drug binding, the affinity of the protein-drug interaction on its own gives no evidence of the effective inhibition outcome. Rather, the protein-drug binding affinity becomes promising only if being stronger than the affinity of the wide-type protein interaction partner, which the drug is competing with. Therefore, there is a need to determine the potency needed for drugs against specific PPIs, through the prediction of PPI kinetic constants. With this goal, a detailed knowledge on molecular interaction kinetics would provide us more understandings about this study.

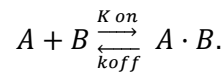
1.3.2 Knowledge of Kinetic Constants in Molecular Interactions

The increasing call for the understanding of the molecular interactions in biological systems has raised the need for interpretative and predictive models for biochemical kinetics studies. In the past decade, physiochemical and structural knowledge have been applied to study the molecular interactions, and some computational methods have allowed the assessment of thermodynamics and kinetics for the association-dissociation processes in protein-protein interactions^{104,105}. Protein-protein interactions demonstrate a very broad spectrum in structural and energetic variability. The binding affinities between a protein pair may range more than 10-order difference in magnitude, corresponding to vast Gibbs free energy changes^{104,106}. Conceptually, the study of PPI kinetics is primarily associated with the transition of the energy states in an energy landscape paradigm, where the states and the kinetics of transition can be described by the depth of energy wells and the heights of energy barriers¹⁰⁷.

Generally in a bimolecular interaction, the association of a protein complex ($A \cdot B$) from the unbound proteins ($A+B$) can be described in **Figure 1-2**, with respect to the Gibbs free energy changes. The energy variation in the scheme shows the stabilization of the transition state through protein-protein binding, where ΔG^0 is the free-energy difference between the reactants ($A+B$) and the product ($A \cdot B$), determining whether the reaction or interaction will occur spontaneously¹⁰⁸.

Figure 1-2 Scheme of bimolecular kinetics and thermodynamics

To illustrate a dimeric protein-protein interaction kinetics mathematically, consider the following reaction for the reversible interaction between protein A and protein B. The binding affinity (K_d , equilibrium dissociation constant, unit: M) of the interaction between protein A and protein B is not only determined by how easy for the two protein partners to get associated to reach the bound state (k_{on} , rate of association constant, unit: $M^{-1}s^{-1}$), but also how easy for them to get separated into their individual conformations (k_{off} , rate of dissociation constant, unit: s^{-1}).



At bimolecular interaction equilibrium, the mass action law applies, which states that: the chemical equilibrium is a dynamic process where the rates of the forward reaction and the backward reaction must be equal at the state of chemical equilibrium. Therefore the following equation is generated.

$$\frac{d[A \cdot B]}{dt} = [A][B]k_{on} - [A \cdot B]k_{off} = 0.$$

Such that,

$$[A][B]k_{on} = [A \cdot B]k_{off}.$$

Which defines the equilibrium dissociation constant K_d (unit: M) as:

$$k_d = \frac{k_{off}}{k_{on}} = \frac{[A][B]}{[A \cdot B]}.$$

In terms of free energy, the equation above can be converted into:

$$k_d = \frac{k_{off}}{k_{on}} = \frac{A * e^{(-\Delta G_{off}^{++}/RT)}}{A * e^{(-\Delta G_{on}^{++}/RT)}} = e^{(-(\Delta G_{off}^{++} - \Delta G_{on}^{++})/RT)} = e^{(-\Delta G^0/RT)}.$$

Such that,

$$\Delta G^0 = -R \cdot T \cdot \ln(k_d).$$

Since ΔG^0 is the standard free-energy changes, R is the gas constant, and T is the absolute temperate, it is obviously observed that ΔG^0 has linear relationship with logarithm of K_d . Hence, the elucidation and prediction of equilibrium dissociation constant (K_d) in protein-protein interaction will help to reveal the free energy difference from the reactants to the product¹⁰⁸.

Recently, there have been made many efforts in the study of protein-protein interaction kinetics. From a computational perspective, *Brownian* dynamics simulation has been widely used to study the enzyme-substrate/inhibitor association, through the effects of mutation, pH, viscosity, and ionic strength^{109,110}. *Selzer et al.* developed an enhancement program to predict the protein association rate by mutating the electrostatic interactions between proteins¹⁰⁶. *Zhou et al.* developed a theory for the protein-protein association rates based on the transition state¹¹¹. *Xie et al.* modelled sequence-kinetics relationship for the antigen-antibody interaction¹¹². *Bai et al.* predicted the kinetic constants for 62 PPIs by using the structure-based properties¹¹³. *Iain et al.* carried out the protein-protein binding affinity prediction on 137 protein complexes with known PDB structures^{114,115}.

1.3.3 Prediction of Protein-Protein Interaction and its Kinetic Constants

In molecular biology, there is a sequence hypothesis states that, the protein function (including molecular recognition) is determined by its three-dimensional structure, which is determined or coded in the protein primary sequence¹¹⁶. This hypothesis is the basis for the sequence-based prediction of protein function, structure, interaction, and so on.

In the last decade, researchers had attempted different computational approaches to predict the protein-protein interactions. One approach is to predict PPIs by solely using the information from the amino acid sequences. *Joel et al.* carried out a study on PPI prediction from primary sequences, by collecting 2,664 PPIs from DIP database¹¹⁷. *Joel* announced their SVM prediction model accuracy reached 80% on average, by using three properties (charge, hydrophobicity, surface tension) to represent the proteins¹¹⁸. To improve this study, *Siaw et al.* assessed the sequence-based PPI prediction by using the artificial shuffled sequences as the negative set, and obtained a higher accuracy at 94%¹¹⁹. *Jiankuan et al.* attempted a larger dataset with 11,855 interactions, and they represented the protein features in binary vectors (1/0) to indicate the existence/inexistence of a certain domain. *Jiankuan* achieved average 70% for both sensitivity and specificity¹²⁰. *Guo et al.* calculated seven physiochemical properties (hydrophobicity, hydrophilicity, volumes, polarity, polarizability, solvent-accessible surface area, and net charge index) of amino acids to reflect the PPI interactions. They combined autocovariance with SVM for predicting 11,474 PPIs, and achieved the accuracy at 88%¹²¹.

With the increasing interests in PPI prediction, PPI kinetics prediction has also attracted many attentions in these years. Particularly, *Xie et al.* constructed the quantitative sequence-kinetics relationship (QSKR) between recombinant *Fab67P* and the coat protein of tobacco mosaic virus¹¹². *Xie* converted 25 mutated peptides into three molecular descriptors (Van Der Waals volume, net charge index, and hydrophobic parameter). PLS (partial least squares) method was used for the modeling, and a good result was obtained ($R^2=0.823$ for training dataset and $Q^2=0.9$ for testing dataset)¹¹². *Bai et al.* studied the linear modelling to predict the PPI kinetics by using 37 structure-based properties. Their dataset had 62 PPIs, and the models gave the performance R^2 at 0.770, 0.732, and 0.801 for K_d , k_{on} , and k_{off} and datasets respectively in leave-one-out cross-validation¹¹³. *Iain et al.* constructed the protein-protein binding affinity prediction on a set of 137 protein complexes with known PDB structures¹¹⁵, by using 200 interacting descriptors calculated from various standalone programs¹¹⁴. *Iain* used four regression algorithms (Random Forest, Regression Tree, Multivariate Adaptive Regression Splines, and Radial Basis Function Interpolation), and showed that all algorithms achieved correlation coefficient between 0.69 and 0.75 against the experimentally measured value in leave-one-out cross-validation. *Iain* also further concluded that the conformational selection mechanism of protein binding was supported from their kinetic rate constant predictions¹²². *Ma et al.* conducted the PPI equilibrium dissociation constant (K_d) prediction of 133 protein complexes, by applying 432 physiochemical and structural features to represent the proteins. Random forests modelling gave the performance R^2 at 0.708 in leave-one-out cross-validation¹²³.

After years of efforts in predicting the protein-protein interaction kinetic constants, these studies have provided many valuable experience and significant findings. However, there are still some limitations:

- 1) The PPI kinetics datasets were not diverse enough to represent a larger protein feature space, where the largest PPI dataset used in the previous studies only covered 137 protein complexes¹¹⁵.
- 2) The availability of PDB structures for the interacting protein complexes was highly relied on, where the structure of the protein pair is only a “frozen” view of the complex, but ignoring the kinetics nature of protein-protein association and dissociation^{105,124}.
- 3) Some proteins do not have the 3D structures available yet, and various software are needed to compute the PPI features based the protein structures, which might not be applicable for many biological labs.

PPI inhibition is an extremely complex issue when going to in vitro and in vivo studies, as there are tremendous proteins, physiochemical factors, and metabolic processes in the interaction environment. Nevertheless, overcoming the wild-type PPI kinetic constants is one of the key rules in designing the competitive PPI inhibitor. In this thesis, we mainly focused on the modeling of the protein-protein interactions for the interaction kinetic constants prediction.

1.4 Objectives and Outline of the Research Described in this Thesis

1.4.1 Overall Objectives

The main goals of the research described in this thesis are the development of bioinformatics tools (PROFEAT webserver and TISPIN database) to facilitate the study of complex biological networks, and the construction of the QSKR machine-learning models for predicting the kinetic constants of protein-protein interactions. Essentially, in this thesis, PROFEAT webserver is to be upgraded to compute the biological network descriptors, TISPIN database is to be constructed to provide human tissue-specific protein interaction networks, and QSKR study is to build the regression machine learning models to predict the PPI kinetic constants.

The first objective is to update PROFEAT webserver by adding the new module for computing the biological network descriptors, with the key features: 1) collect and integrate the definitions and algorithms of network descriptors from various fields of study (e.g. sociology, physics, mathematics, economics, transportation, biology, etc.), such that provide the most comprehensive and diverse network descriptors at both node-level (local properties), edge-level (local properties), and network-level (global properties); 2) support different network types (undirected/directed, unweighted/weighted edge or nodes) for representing different kinds of biological networks (binary/oriented, constant/varying binding constants or molecular levels); 3) offer the user-friendly access modes for simple input/output, requiring easy operation with minimal manual interventions; 4) support different network file formats, which

are compatible with the major network analysis software; and 5) enable the automatic detection, split, and computation of multiple disconnected networks from a single input file. Additionally, systems biology applications, case studies, and detailed definitions/algorithms of the network descriptors are to be comprehensively documented and provided.

The second target is to construct a new database (TISPIN) for providing the Tissue-Specific Protein Interaction Networks, by delivering: 1) network files in various network formats that are compatible with the major network software; 2) network visualization; 3) computed network descriptors in node-level (for each protein) and network-level (for the entire network); 4) protein annotations in terms of protein name, gene symbol, UniProt ID/ACC, NCBI protein ID, GO biological process / cellular component / molecular function, and therapeutic targets; and 5) comprehensive download links for all information. In the current stage of developing the database prototype TISPIN 1.0, we mainly focus on building up the database architecture and interface, and making it fully functional based on the primary data source collected from HPRD database, which is an expert-curated reliable source of protein-protein interactions^{125,126}. HPRD provides not only the PPIs, but also the tissue-protein associations, based on the literature text mining for tissue distribution of proteins⁷⁵. Unlike the other databases that use the microarray and RNAseq gene expression data to infer the tissue-protein associations at the transcription-level, the tissue distribution in TISPIN is on the basis of the protein-level evidence.

The third goal is to study Quantitative Sequence-Kinetic Constants Relationship (QSKR), by hypothesizing that “for a pair of interacting proteins, there exists a

general quantitative relationship between the information from protein primary sequences and the interaction kinetic constants.” The approach by only using the protein primary sequences will be more universally applicable than those using the protein structures. In this proof of concept study, we are to expand the PPI library with known kinetics, and extend the applications of support vector regression and random forests algorithms onto this highly diverse PPI dataset, to predict the kinetic constants (K_d , k_{on} and k_{off}), solely based on the features generated from amino acid sequences.

1.4.2 Overall Outline

In **Chapter 1** (Introduction), an overview of the background knowledge is given, essentially the bioinformatics tools (webserver and database) in studying the complex biological networks and the machine-learning prediction for protein-protein interaction kinetic constants. **Section 1.1** describes the historical origins and the modern biological applications of the network descriptors, and then introduces the publicly accessible tools/software for computing the network descriptors. **Section 1.2** discusses the recent studies and the relevant databases on tissue-specific protein networks. In **Section 1.3**, the emerging trend in discovering PPI inhibitor is introduced, by emphasizing the importance of PPI kinetic constants in drug discovery, and summarizing the research in predicting the PPI kinetic constants.

Chapter 2 (PROFEAT Webserver Development for Computing Biological Network Descriptors) provides the detailed motivations, sources, and methods in upgrading the PROFEAT webserver for computing the biological network descriptors. The input, output, file formats, illustrative examples, and comparative performance evaluations are provided for a better understanding of this new function in PROFEAT webserver. Moreover, we will also summarize and discuss some typical applications of network descriptors in studying the genome / interactome / transcriptome / metabolome / diseasome-derived biological networks.

Chapter 3 (TISPIN Database Development for Human Tissue-Specific Protein Interaction Networks) presents the data sources (PPIs and tissue-protein associations) and the workflow in constructing the TISPIN database prototype

1.0. The database interfaces and architectures are introduced, as well as the informative contents provided in TISPIN. As a new database providing tissue-specific protein interaction networks, the comparison with other relevant databases is presented, and the outstanding features and the current limitations are also discussed.

Chapter 4 (Quantitative Sequence-Kinetic Constants Relationship for Predicting Protein-Protein Interaction Kinetic Constants) provides a proof-of-concept study in predicting the PPI kinetic constants (K_d , k_{on} and k_{off}) by only using the information derived from protein primary sequences. Data collection, PPI feature generation, and regression machine learning methods are introduced, and the predictive performance are evaluated.

In **Chapter 5** (Concluding Remarks), we discuss the major contributions of this thesis, as well as the limitations of the current work. Suggestions for further studies and improvable aspects are also proposed.

Additionally, the supplementary information are delivered in **Appendices**, where **Section A** and **Section B** provides the full list and the detailed definitions/algorithms of all the network descriptors implemented in PROFEAT webserver. **Section C** and **Section D** provides the newly expanded PPI dataset used in QSKR study for predicting the PPI kinetic constants (K_d , k_{on} and k_{off}).

CHAPTER 2 PROFEAT Webserver Development for Computing Biological Network Descriptors

2.1 Background and Motivations

Quantitative analysis of biological networks are needed for more extensive investigations of biological^{8,22,23,24,25}, disease^{26,27,28,29,30} and pharmacological^{31,32,33,34,35} processes. These analyses can be facilitated by the knowledge of the network descriptors that characterize the connectivity, topology, organizational, robustness, and stability properties of the relevant protein-protein interaction, gene regulatory, metabolic and drug-target networks. A number of network descriptors (e.g. centrality indices, clustering coefficient, topological coefficient, neighborhood connectivity, etc.) initially developed as graph/network theory in such areas as sociology, mathematics and physics, have been successfully applied for studying the biological networks. These network descriptors have facilitated to reveal enrichment patterns, systematic understandings, and network-based clues in biological networks^{29,36}. Nonetheless, a substantial number of the network descriptors (e.g. geographical indices, topological robustness, etc.) have not yet been used but are potentially useful for analyzing more diverse features of biological networks.

Currently, a number of public GUI-based computational resources are available for calculating network descriptors, particularly Cytoscape⁴¹, NAViGaTOR⁴², Gephi⁴³, VANESA⁴⁴, Pajek⁴⁵, SpectralNET⁴⁶, PINA⁴⁷, Hubba⁴⁸, GraphWeb⁴⁹, tYNA⁵⁰ and VisANT⁵¹ (**Table 1-2**). For users with programming skills can use Python library NetworkX⁵², R package igraph⁵³, and R package QuACN⁵⁴ to

compute the network properties, however these programming tools are hardly applicable for the users without computation expertise⁵⁵. Compared to the literature-reported network descriptors (**Appendix Section A**, **Section B**, and **Table S-1**, **S-2**, and **S-3**), these computational resources have covered a limited number of network descriptors, while some of the uncovered network descriptors (e.g. PageRank centrality^{19,21,56}, interconnectivity^{57,58,59}, weighted clustering coefficient^{60,61}, etc.) have already been shown their usefulness in systems biology.

Therefore, there is a need for the relevant web-servers to provide more comprehensive coverage and more user-friendly means in computing the network descriptors for studying biological networks. Hence, we introduced a new network descriptor module in PROFEAT webserver at (<http://bidd2.nus.edu.sg/cgi-bin/profeat2016/main.cgi>), which was previously introduced¹²⁷ and updated¹²⁸ as a webserver for computing the structural and physicochemical descriptors of proteins, peptides and protein-protein interaction pairs.

This new module supports the computation of 227 descriptors (31 node-level, 195 network-level, and 1 edge-level) for an undirected un-weighted network (e.g. un-oriented network with uniform binding constants and molecular levels), 367 descriptors (85 node-level, 277 network-level, and 5 edge-level) for an undirected edge-weighted network (e.g. un-oriented network with varying binding constants and uniform molecular levels), 239 descriptors (39 node-level, 199 network-level, and 1 edge-level) for an undirected node-weighted network (e.g. un-oriented network with uniform binding constants and varying molecular

levels), 379 descriptors (93 node-level, 281 network-level, and 5 edge-level) for an undirected edge-node-weighted network (e.g. un-oriented network with varying binding constants and varying molecular levels), and 23 descriptors (11 node-level and 12 network-level) for a directed un-weighted network (e.g. oriented process with uniform binding constants and molecular levels).

Apart from the full-set of network descriptors, a sub-group of the network descriptors, which have been extensively used in studying biological networks^{38,39} or applied for probing specific biological or therapeutic questions¹²⁹, were selected into a slim-set of network descriptors. The numbers of the network descriptors in both the full-set and the slim-set for the different network types, and their biological representations were tabulated **Table 2-1**. The typical interpretations and biological implications of the slim-set of the network descriptors were summarized in **Table 2-2**.

Table 2-1 The supported network types with different biological representations, and the number of network descriptors (both the full-set and the slim-set) computed in PROFEAT

| Network Type | Biological Representations | Full-Set of Network Descriptors | | | | Slim-Set of Network Descriptors | | | |
|--------------------------------------|--|---------------------------------|---------------|------------|------------|---------------------------------|---------------|------------|-----------|
| | | Node Level | Network Level | Edge Level | Total | Node Level | Network Level | Edge Level | Total |
| Undirected Un-Weighted Network | un-oriented network with uniform binding constants, uniform molecular levels | 31 | 195 | 1 | 227 | 19 | 28 | 1 | 48 |
| Undirected Edge-Weighted Network | un-oriented network with varying binding constants, uniform molecular levels | 85 | 277 | 5 | 367 | 41 | 44 | 5 | 90 |
| Undirected Node-Weighted Network | un-oriented network with uniform binding constants, varying molecular levels | 39 | 199 | 1 | 239 | 23 | 28 | 1 | 52 |
| Undirected EdgeNode-Weighted Network | un-oriented network with varying binding constants, varying molecular levels | 93 | 281 | 5 | 379 | 45 | 44 | 5 | 94 |
| Directed Un-Weighted Network | oriented network with uniform binding constants, uniform molecular levels | 11 | 12 | 0 | 23 | 5 | 11 | 0 | 16 |

Table 2-2 Typical interpretations and biological implications of the slim set of network descriptors

| Network Descriptor | Level | Typical Interpretation and Biological Implication |
|--|---------|--|
| Connectivity/Adjacency-based Properties | | |
| Degree | Node | Number of interacting partners |
| Number of Selfloops | Node | Number of homodimers formed by two identical molecules |
| Number of Triangles | Node | Number of the smallest unit of molecular interaction clusters |
| Clustering Coefficient | Node | Tendency of each molecule to form groups in the network |
| Neighborhood Connectivity | Node | Indicate if a molecule is near the high-degree hubs of the network |
| Topological Coefficient | Node | Extent of a molecule in sharing its partners in the network |
| Interconnectivity | Node | How close of a molecule is connected with its neighbours, reflecting the alternative signaling capacity |
| Bridging Coefficient | Node | How well the molecule is linked between high-degree hubs |
| Degree Centrality | Node | Prioritize the molecules by their number of interactions |
| Number of Nodes and Edges | Network | Number of molecules and interactions in the biological network |
| Number of Selfloops | Network | Total number of homodimers formed in the network |
| Maximum / Minimum Connectivity | Network | The highest / lowest number of interactions for a molecule |
| Average Number of Neighbours | Network | The average number of interactions for all molecules |
| Network Density | Network | Efficiency of information transmitting in the biological network |
| Average Clustering Coefficient | Network | Overall tendency of all molecules to form groups in the network |
| Transitivity | Network | Another measure of tendency of forming groups in the network |
| Heterogeneity | Network | Reflect the tendency of a network to have molecular hubs |
| Degree Centralization | Network | Indicate the network is highly connected or decentralized |
| Shortest Path Length-based Properties | | |
| Average Shortest Path Length | Node | A measure of signal transmission distances or reaction steps from one molecule to all other molecules in the network |
| Eccentricity | Node | Identify the peripheral or marginal molecules in the network |
| Radiality | Node | Another indicator for peripheral molecules in the network |
| Closeness Centrality | Node | A measure of how fast the signaling information or reaction spreads from one molecule to all other molecules |
| Eccentricity Centrality | Node | A similar measure as closeness centrality |
| Load Centrality (Stress) | Node | The extent of a molecule involved in efficient signal transmission |
| Betweenness Centrality | Node | The importance of a molecule in efficient alternative signaling |
| Bridging Centrality | Node | How much information flowing through the molecule |
| Network Diameter | Network | The longest signal transmission or reaction distance |
| Network Radius | Network | The shortest signal transmission or reaction distance |
| Characteristic Path Length | Network | The average signal transmission or reaction distance |
| Average Eccentricity | Network | The overall peripherality of all molecules in the network |
| Global Efficiency | Network | The efficiency of information exchange, signaling transmission, or chemical reaction in the biological network |

| Topological Indices | | |
|--|---------|---|
| Hierarchy | Network | Index for power-law distribution of molecular interactions |
| Robustness | Network | Stability of a biological network for studying diseases and variations |
| Wiener, BalabanJ , Randic Connectivity Index | Network | Well-known topological properties for molecular characterization |
| Eigenvector-based Complexity Indices | | |
| Eigenvector Centrality | Node | The iteratively converged importance of a molecule by considering the importance of its interaction partners |
| Page Rank Centrality | Node | The iteratively converged importance of a molecule by considering the importance of its interaction partners and its number of partners |
| Graph Energy, Laplacian Energy | Network | Well-known eigenvalue-derived properties in mathematical chemistry |
| Entropy-based Complexity Indices | | |
| Information Content on Degree Equality | Network | Entropy of probability distribution of the molecular interactions |
| Radial Centric Information Index | Network | Entropy of probability distribution of the peripheral molecules |
| Bonchev Information Index | Network | Entropy of probability distribution of the efficient signaling transmission distances |
| Edge-Weighted Properties | | |
| Strength | Node | Indicate if a molecule having strong interactions with its partners |
| Assortativity | Node | Indicate if a molecule having strong interactions with its partners and also near the high-degree hubs in the network |
| Edge-Weighted Interconnectivity | Node | A complexity measure of how close and how strong a molecule is interacting with its partners |
| Edge-Weighted Transitivity | Network | A measure of tendency of forming groups in the weighted network |
| Edge Weight | Edge | Interaction kinetic constants, binding affinity, correlation coefficient between molecular levels, interaction score, etc. |
| Edge-Betweenness | Edge | Prioritize the important interactions in the biological network, and facilitate the identification of key modules or clusters |
| Node-Weighted Properties | | |
| Node Weight | Node | Molecular level, expression level, expression fold change, etc. |
| Node-Weighted Neighbourhood Score | Node | Identify the regions with high molecular abundance if the node weight is molecular level, or the regions with high differentially expressed genes if the node weight is expression fold change. |
| Directed Properties | | |
| In-Degree | Node | The number of molecules that control or regulate a specific molecule |
| Out-Degree | Node | The number of molecules that are controlled or regulated by a specific molecule |
| Directed Local Clustering Coefficient | Node | Tendency of each molecule to form circulated groups in the network |
| In-Degree (Avg, Max, Min) | Network | The average / highest / lowest number of molecules that control or regulate other molecules in the network |
| Out-Degree (Avg, Max, Min) | Network | The average / highest / lowest number of molecules that are controlled or regulated by other molecules in the network |
| Directed Global Clustering Coefficient | Network | Overall tendency of all molecules to form circulated groups in the network |

Table 2-3 summarized the number of computed network descriptors, the supported network types, the capability to automatically detect-and-split the multiple networks from a single input network file, the requirement for programming expertise, and the network visualization function of PROFEAT in comparison with the other 14 publically accessible tools.

Table 2-4 summarized the selected typical applications of network descriptors in systems biology studies, where the names of network descriptors were *italicized* in this table, and the network descriptors were categorized by the matrices used for the calculation based on their definitions and algorithms (**Appendix Section B**): 1) adjacency-based properties, 2) shortest path length-based properties, 3) topological indices, 4) entropy-based complexity indices, 5) eigenvalue-based complexity indices, 6) edge-weighted properties, 7) node-weighted properties, and 8) directed properties.

Table 2-3 The number of network descriptors, the list of network types, and visualization features of PROFEAT and other publically accessible tools

| Tool Name | Number of Descriptors | Network Types | | | | | Auto-Detect-&-Split Multiple | Program Skills Required? | Network Visualization |
|-------------|-----------------------|---------------|---------------|---------------|-------------------|----------------------|------------------------------|--------------------------|-----------------------|
| | | Un-Weighted | Edge-Weighted | Node-Weighted | EdgeNode-Weighted | Directed Un-Weighted | | | |
| PROFEAT | up to 317 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✗ | ✗ |
| NetworkX | ~ 100 | ✓ | ✓ | ✗ | ✗ | ✓ | ✗ | ✓ | ✗ |
| Igraph | ~ 100 | ✓ | ✓ | ✗ | ✗ | ✓ | ✗ | ✓ | ✗ |
| QuACN | ~ 100 | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ | ✗ |
| Cytoscape | ~ 23 | ✓ | ✗ | ✗ | ✗ | ✓ | ✗ | ✗ | ✓ |
| NAViGaTOR | ~ 13 | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ |
| Gephi | ~ 10 | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ |
| VANESA | ~ 10 | ✓ | ✓ | ✗ | ✗ | ✓ | ✗ | ✗ | ✓ |
| Pajek | ~ 9 | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ |
| SpectralNET | ~ 9 | ✓ | ✓ | ✓ | ✓ | ✗ | ✗ | ✗ | ✓ |
| PINA | ~ 8 | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ |
| Hubba | ~ 6 | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ |
| GraphWeb | ~ 4 | ✓ | ✓ | ✗ | ✗ | ✓ | ✗ | ✗ | ✓ |
| tYNA | ~ 4 | ✓ | ✗ | ✗ | ✗ | ✓ | ✗ | ✗ | ✓ |
| VisANT | ~ 3 | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ |

Table 2-4 List of the network descriptors (node-level & network-level) in different categories and their selected applications in systems biology

| Network Descriptors | Applications in Systems Biology |
|--|---|
| Adjacency-based Properties | |
| Node-Level: Degree, Scaled Connectivity, Number of Selfloops / Triangles, Zscore, Clustering Coefficient, Topological Coefficient, Neighborhood Connectivity, Interconnectivity, Degree Centrality, Bridging Coefficient | <p><i>Degree, average neighbours</i> and <i>density</i> implicated the genes in disease network¹³⁰.</p> <p><i>Degree</i> and <i>clustering coefficient</i> validated if the drugs are highly associated with proteins in the drug-target network³², and predicted candidate genes in coronary artery disease¹³¹.</p> <p><i>Topological coefficient</i> and <i>clustering coefficient</i> identified high-confidence interactions in a large-scale PPI network³⁸.</p> <p><i>Clustering coefficient</i> provided molecular characterization of gene co-expression network¹³², illustrated the hierarchical architecture of metabolism^{8,133}, identified the functional modules from genomic associations¹³⁴, and predicted the protein functions by network-based methods¹³⁵.</p> <p><i>Neighbourhood connectivity</i> measured the specificity and stability of protein networks and gene regulatory networks¹³⁶.</p> <p><i>Interconnectivity</i> prioritized the disease genes in drug-target network⁵⁷.</p> <p><i>Density, heterogeneity</i>, and <i>degree centralization</i> used to compare the PPI networks between drosophila and yeast¹³⁷.</p> |
| Shortest Path Length-based Properties | |
| Node-Level: Average Shortest Path Length, Eccentric, Eccentricity, Radiality, Distance Sum, Deviation, Distance Deviation, Closeness Centrality, Eccentricity Centrality, Harmonic Centrality, Residual Centrality, Stress Centrality, Betweenness Centrality, Bridging Centrality | <p><i>Eccentricity</i> and <i>distance deviation</i> identified the metabolic biomarkers¹³⁸.</p> <p><i>Radiality</i> used to analyze gene regulatory networks¹³⁹.</p> <p><i>Centrality</i> and <i>peripherality (eccentricity, radiality)</i> implicated genes in disease network¹³⁰.</p> <p><i>Shortest path length, betweenness centrality, closeness centrality, radiality</i> and <i>integration</i> explored protein-drug interactome for lung cancer¹⁴⁰, identified the hubs and bridging nodes in drug addiction mechanisms¹⁴¹.</p> <p><i>Betweenness centrality, degree centrality, bridging centrality</i> and other centrality measures exposed the relations between network topology and system function of proteins^{31,131,142,143,144}, and classified the important nodes in drug discovery¹⁴⁵.</p> <p><i>Characteristic path length</i> and <i>global efficiency</i> used to describe the brain neuro-connectivity network¹⁴⁶.</p> |
| Network-Level: Total Distance, Shape Coefficient, Diameter, Radius, Characteristic Path Length, Network Eccentricity, Average Eccentricity, Network Eccentric, Unipolarity Eccentric Connectivity, Integration, Variation, Average Distance, Mean Distance Deviation, Centralization, Global Efficiency | |

Table 2-4 (continued) List of the network descriptors (node-level & network-level) in different categories and their selected applications in systems biology

| Topological Indices | |
|---|--|
| <p>Node-Level: <i>N.A.</i></p> <p>Network-Level: Edge Complexity Index, ABC Index, Randic Connectivity Index, Zagreb Indices, Narumi Indices, Alpha Index, Beta Index, Pi Index, Eta Index, Hierarchy, Robustness, Medium Articulation, Complexity Indices, Wiener Index, Hyper-Wiener, Harary Indices, Compactness, Superpendentic Index, Hyper-Distance-Path Index, BalabanJ, BalabanJ-like Indices, Geometric Arithmetic Indices, Product of Row Sums, Topological Indices, Szeged Index, Efficiency Complexity</p> | <p>Exponent of power-law degree distribution (<i>hierarchy index</i>), provided molecular characterization of cellular state in gene co-expression network¹³², characterized the yeast genetic interaction network¹⁴⁷, measured the robustness of protein interaction networks and genetic regulatory networks¹³⁶.</p> <p><i>Wiener index</i>, <i>BalabanJ index</i>, <i>Randic connectivity index</i>, <i>Zagreb indices</i>, and <i>graph complexity index</i> applied to access the complexity in chemistry and biology^{148,149}.</p> <p><i>Medium articulation</i> and <i>efficiency complexity</i> evaluated for measuring the graph features of PPI, genetic interaction, and metabolic networks¹⁵⁰.</p> <p><i>Complexity indices</i> and <i>BalabanJ index</i> classified the metabolic networks from 3 domains of life¹⁵¹.</p> |
| Entropy-based Complexity Indices | |
| <p>Node-Level: <i>N.A.</i></p> <p>Network-Level: Entropy on (degree equality / edge equality / edge magnitude / distance degree / distance degree equality), Radial Centric Information Index, Distance Degree Compactness, Distance Degree Centric Index, Graph Distance Complexity, Information Layer Index, Bonchev Information Indices, Balaban-like Information Indices</p> | <p><i>Information-theoretic entropy measures</i> identified and ranked the highly discriminating metabolic biomarker candidates for obesity¹³⁸.</p> <p><i>Radial centric</i> and <i>degree equality-information index</i> classified the metabolic networks of 43 organisms from 3 domains of life¹⁵¹.</p> <p><i>Bonchev indices</i> and some other entropy measures were evaluated for potential use in biology and chemistry^{152,153}.</p> |
| Eigenvalue-based Complexity Indices | |
| <p>Node-Level: Eigenvector Centrality, PageRank Centrality</p> <p>Network-Level: Graph Energy, Laplacian Energy, Spectral Radius, Estrada Index, Laplacian Estrada Index, Quasi-Weiner Index, Mohar Indices, Graph Index Complexity, 50 Dehmer-defined Entropy by Matrices of (adjacency / laplacian / distance / distance path / augmented vertex degree / extended adjacency / vertex connectivity / random walk markov / weighted structure function 1 / weighted structure function 2)</p> | <p><i>PageRank centrality</i> identified prognostic marker genes of pancreatic cancer⁵⁶, and identified protein target in metabolic networks¹⁹. The <i>PageRank centrality/degree quotient</i> scored and found the non-hub important nodes in microbial networks from 3 distinct organisms¹⁹.</p> <p><i>Eigenvector centrality</i>, together with other centralities, were applied to predict the synthetic genetic interactions^{154,155}.</p> <p><i>Graph index complexity</i> measured the features of real-world systems, including PPI network, genetic interaction network, and metabolic network¹⁵⁰.</p> <p><i>Dehmer proposed 50 eigenvalue descriptors</i>, possessing high discriminative power to capture structural information, to predict biological and pharmacological properties¹⁵⁶.</p> |

Table 2-4 (continued) List of the network descriptors (node-level & network-level) in different categories and their selected applications in systems biology

| Edge-Weighted Properties | |
|---|--|
| Node-Level: Strength, Assortativity, Disparity, Geometric Mean of Triangles, Edge-Weighted Local Clustering Coefficient, Edge-Weighted Interconnectivity | <i>Edge-weighted clustering coefficient</i> identified the significant gene modules in co-expression network ⁶⁰ . <i>Edge-weighted interconnectivity</i> ranked the candidate disease genes in biological networks ⁵⁸ . <i>Edge-weighted transitivity</i> used to describe the brain neuro-connectivity network ¹⁴⁶ . |
| Network-Level: Weighted Transitivity, Edge-Weighted Global Clustering Coefficient | |
| Node-Weighted Properties | |
| Node-Level: Node Weight, Node-Weighted Cross Degree, Node-Weighted Local Clustering Coefficient, Node-Weighted Neighbourhood Score | <i>Node-weighted neighbourhood score</i> prioritized the novel disease genes for the prediction of drug targets for a given disease ⁵⁷ . |
| Network-Level: Total Node Weight, Node-Weighted Global Clustering Coefficient | |
| Directed Properties | |
| Node-Level: In-Degree, Out-Degree, Directed Local Clustering Coefficient, Neighbourhood Connectivity (only in, only out, in-and-out), Average Directed Neighbour Degree | <i>In/out-degree</i> and <i>directed clustering coefficient</i> analyzed the gene regulatory networks under different conditions ¹⁵⁷ , and applied to identify and rank the regulators in the directed biological networks ¹⁵⁸ . <i>Directed clustering coefficient</i> and <i>average directed neighbour degree</i> studied the neuro-connectivity networks ¹⁴⁶ . |
| Network-Level: In-Degree (max, avg, min), Out-Degree (max, avg, min), Directed Global Clustering Coefficient | |

2.2 Materials and Methods

2.2.1 Network Descriptor Computational Methods

The PROFEAT computed network descriptors were broadly grouped into two local property and global property. Some popular ones were selected and introduced here, briefly.

The first group (**Appendix Table S-1** and **Section B.1**) consisted of the node-level descriptors that are calculated based on the connectivity/adjacency matrix (e.g. degree, selfloop, triangle, clustering coefficient), and based on shortest-path-length matrix (e.g. closeness centrality, betweenness centrality, eccentricity). These descriptors were illustrated in **Figure 2-1**. **Degree** deg_i is the number of edges or interactions directly linked to the studied node⁸. Number of **Selfloops** is the number of edges linking to itself. Number of **Triangles** " $tri_i = \frac{1}{2} \sum_{j=1}^N \sum_{k=1}^N A_{ij} A_{ik} A_{jk}$ " implies the level of segregation at each node, and it's the basis for measuring the global transitivity¹⁴⁶. **Clustering Coefficient** is locally defined as " $cluster_i = \frac{2e_i}{deg_i(deg_i-1)}$ " and globally defined as " $cluster_G = \frac{1}{N} \sum_{i=1}^N cluster_i$ ", where N is number of nodes, e_i is the number of links among all neighbours of node i , $e_i = 0$ if node i has less than 2 neighbours⁹. Global clustering coefficient characterizes the overall tendency of the nodes to form groups or clusters in the network⁸. **Closeness Centrality** is defined as the reciprocal of the average shortest path length, a measure of information spreading speed from a given node to the other reachable nodes in the network¹⁵⁹:

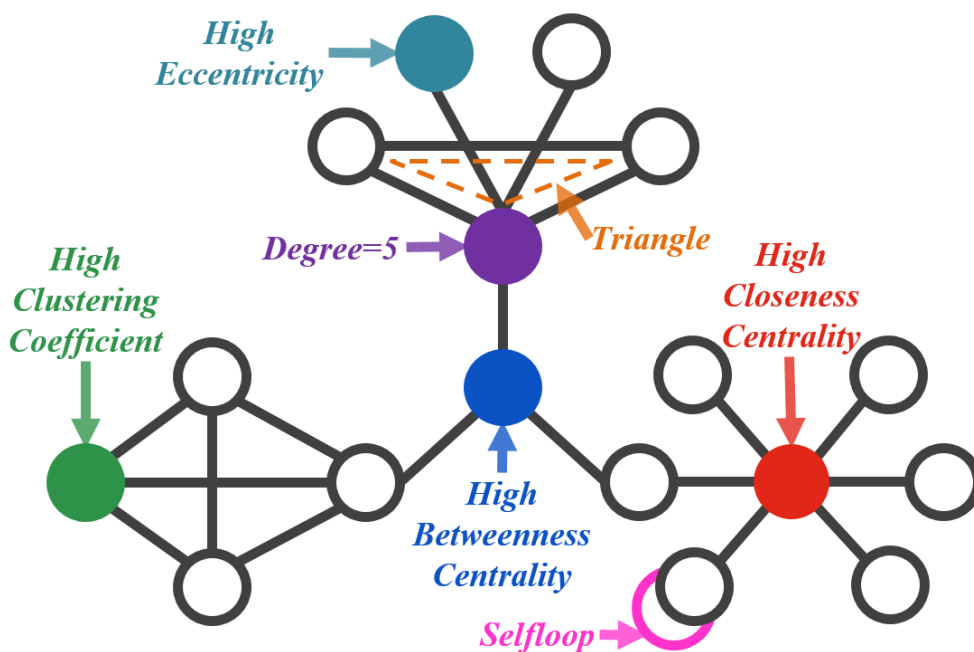
" $closeness_i = \left(\frac{1}{N} \sum_{j=1}^N D_{ij} \right)^{-1}$ ", where D_{ij} represents the shortest path length

between i and j ¹⁶⁰. **Betweenness Centrality** " $betweenness_i = \frac{\sum_{s \neq i \neq t} \sigma_{st}(i)}{\sigma_{st}}$ "

indicates the number of times a node serving as a bridge along the shortest path between any other two nodes in the network, where node s and node t are different from node i , $\sigma_{st}(i)$ is the number of shortest paths from node s to node t passing through node i , and σ_{st} is the number of all shortest paths from node s to node t ¹⁰. Betweenness centrality reflects the degree of control of that specific node exerting over the interactions with other nodes in the network, and implies the importance of a molecule acting in the efficient alternative signaling.

Eccentricity " $eccentricity_i = \max \{D_{ij}\}$ " is the largest shortest path length between node i and all the others, identifying the peripheral nodes in the network.

Figure 2-1 Graphic illustration of the network descriptors (degree, selfloop, triangle, clustering coefficient, closeness centrality, betweenness centrality, and eccentricity) in a hypothetic network



The second group of network descriptors (**Appendix Table S-2** and **Section B.2**) were the network-level features, including the descriptors that are calculated based on the adjacency matrix (e.g. degree centralization, and heterogeneity), the eigenvalue-based complexity indices (e.g. graph energy), and the entropy-based complexity indices (e.g. information content of degree equality). **Degree Centralization** (or Connectivity Centralization) is useful to differentiate the highly connected (like star-shaped) and the decentralized networks¹⁶¹, for studying the structural difference between networks. **Heterogeneity** measures the variation of degree distribution, implying the tendency of a network to have hubs. This descriptor is biological meaningful, as biological networks usually have some central nodes highly connected and the rest nodes very little connected, thus giving a high heterogeneity value. These two descriptors are computed by firstly calculating the network density " $density_G = 2 \cdot E / N(N - 1)$ " where E is the number of edges. The degree centralization is " $centralization_G = \frac{N}{N-2} \left(\frac{\max(deg_G)}{N-1} - density_G \right)$ " and the heterogeneity is " $heterogeneity_G = \sqrt{N \cdot \sum_{i=1}^N (deg_i^2) / (\sum_{i=1}^N deg_i)^2 - 1}$ "¹³⁷. **Graph Energy** of a network is the summation of all its non-zero eigenvalues $\{\lambda_1, \lambda_2 \dots \lambda_k\}$ based on the adjacency matrix, " $Energy_G = \sum_{i=1}^k |\lambda_i|$ "¹⁶². **Information Content of Degree Equality** is defined by the equation: " $I_{vertexDegree} = - \sum_{i=1}^{k^d} \frac{N_i^d}{N} \cdot \log_2 \left(\frac{N_i^d}{N} \right)$ ", which uses the Shannon's entropy formula to measure the probability distribution of vertex degree in the network, where N_i^d is the number of nodes having the same degree, and k^d is the maximum of degree¹⁶³.

To facilitate the studies of biological networks of varying molecular levels and/or binding constants, PROFEAT also provided the edge/node-weighted descriptors. For instances, **Edge-Weighted Clustering Coefficient** has been recently applied to the prediction of the significant gene modules in gene co-

expression network^{60,61}, which is defined " $cluster^{EW}_i = \frac{\sum_{j=1}^N \sum_{k=1}^N \widehat{W}_{ij} \widehat{W}_{ik} \widehat{W}_{jk}}{(\sum_{k=1}^N \widehat{W}_{ij})^2 - \sum_{k=1}^N \widehat{W}_{ij}^2}$ ".

Node-Weighted Cross Degree and **Node-Weighted Local Clustering Coefficient** have been used to analyze the networks with heterogeneous node weights, in the study of Earth's spatial network and international trade network¹⁵.

These descriptors are calculated by firstly generating the extended adjacency matrix by " $ExtA_{ij} = A_{ij} + \delta_{ij}$ ", where A_{ij} is the adjacency matrix and δ_{ij} is the

Kronecker's delta constant. **Node-Weighted Cross Degree** is defined by

" $crossdeg^{NW}_i = \sum_{j=1}^N ExtA_{ij} \cdot NW_i$ ", where NW_i is the node weight of node i .

Node-Weighted Local Clustering Coefficient is then calculated:

$$cluster^{NW}_i = \frac{1}{crossdeg^{NW}_i^2} \sum_{j=1}^N \sum_{k=1}^N ExtA_{ij} \cdot NW_j \cdot ExtA_{ik} \cdot NW_k \cdot ExtA_{jk} \quad ,$$

which is assumed to be zero if the node-weighted cross degree value is zero.

Directed Local Clustering Coefficient was introduced to measure the brain connectivity, as the neuro-connections is considered as directed edges¹⁴⁶. It is

defined " $cluster^D_i = \frac{\frac{1}{2} \sum_{j,h \in N} (A_{ij} + A_{ji})(A_{ih} + A_{hi})(A_{jh} + A_{hj})}{(deg_i^+ + deg_i^-)(deg_i^+ + deg_i^- - 1) - 2 \sum_{j \in N} A_{ij} \cdot A_{ji}}$ ", where deg_i^+

and deg_i^- are the in/out-degree of node i .

2.2.2 Network File Format

SIF Network File Format

SIF, namely Simple Interaction File, is compatible with the majority of the network software (including Cytoscape, Gephi, GraphWeb, Hubba, NAViGaTOR, PINA, SpectralNET, tYNA), and have been used for storing biological interaction data in databases such as Pathway Commons¹⁶⁴. SIF is tab-delimited, specifying the two linked nodes in each line, with the relationship type in between. The following example illustrated the unweighted SIF format, to represent the biological binary interaction networks (e.g. protein-protein interaction network, gene regulatory network, gene co-expression network, drug-target network, etc.).

[Node A] tab [Relationship] tab [Node B]

Edge-weighted SIF is defined by extending the fourth column for the numerical edge weight between the two connected nodes. In biological networks, the edge weight could be PPI kinetic constants, PPI binding affinity, gene co-expression association, interaction confidence level, or some other measures of the strength between the interacting molecules.

[Node A] tab [Relationship] tab [Node B] tab [Edge Weight]

Directed SIF format is the same as the original SIF format, with the added direction information. For the two nodes in each line, the earlier node is meant to point to the latter node. Here, the previous unweighted SIF format meant that *Node A* points to *Node B* ($A \rightarrow B$). Biological directed network usually represents the oriented process (e.g. signalling pathway, metabolic reaction, etc.).

NET Network File Format

NET format, developed by software Pajek, mainly includes three sections (**vertices*, **edges*, and **arcs*) in its file structure, where (**vertices*) section lists all the nodes; (**edges*) section lists all the undirected interactions between two nodes, with an optional edge weight in the third column; and (**arcs*) section lists all the directed interactions, pointing from the earlier node to the later node.

**vertices*

[Node A]

[Node B]

[Node C]

**edges*

[Node A] tab [Node C] tab [Edge Weight]

**arcs*

[Node B] tab [Node C] tab [Edge Weight]

The above example meant there are three nodes (**vertices*) A, B, C in the network, where there are one undirected interaction (**edge*) between *Node A* and *Node C*, and one directed interaction (**arcs*) from *Node B* to *Node C*.

TXT Node Weight File Format

The node weight file (in tab-delimited text format) is separated from the network file. It specifies the node label in the first column and its numerical node weight in the second column, while the node label must be exactly matched with the network file. Biologically, node weight may represent the molecular level (e.g. gene expression, RNAseq count, protein abundance, etc.).

[Node Label] tab [Node Weight]

2.2.3 Performance Evaluation Methods

Performance evaluation of CPU running time was carried out by testing 10 different-scaled human tissue-specific PPI networks of 5 different network types. These networks were constructed based on 38,131 protein-protein interactions and 111,152 tissue-protein associations collected from HPRD database^{86,165,166}. By grouping the PPIs according to their distributed tissues, the tissue-specific lists of PPIs were obtained and their largest connected components were extracted as the human tissue-specific PPI networks for this performance evaluation.

10 tissue-specific PPI networks (**Table 2-5**) were selected with the number of nodes varying from 63 to 2,317 and the number of edges varying from 91 to 4,942. Each network was constructed into five different types. The first four types were undirected unweighted, undirected edge-weighted, undirected node-weighted, and undirected edge-node-weighted networks respectively with the edge-weights or node-weights randomly generated. The fifth type was the directed unweighted network with the direction of each edge tentatively assigned from the left-node to the right-node in the input SIF file. The CPU running time for computing the slim-set of PROFEAT network descriptors were evaluated on a Dell OptiPlex9010 desktop computer with Intel Core i7-3770 3.4GHz CPU and 20GB RAM.

Table 2-5 Ten tissue-specific PPI networks for CPU running time evaluation

| Tissue | Human Systems | Network Size | |
|-----------------|-----------------|--------------|--------------|
| | | No. of Nodes | No. of Edges |
| Lymph Node | Immune | 63 | 91 |
| Hippocampus | Nervous | 107 | 146 |
| Bone Marrow | Immune | 189 | 348 |
| Muscle | Musculoskeletal | 315 | 632 |
| Small Intestine | Digestive | 616 | 980 |
| Colon | Digestive | 988 | 1951 |
| Ovary | Reproductive | 1165 | 2230 |
| Spleen | Immune | 1292 | 2543 |
| Pancreas | Endocrine | 1625 | 3336 |
| Lung | Respiratory | 2317 | 4942 |

PROFEAT computed network descriptor values and the job execution times were also evaluated against those computed from the three popular tools NetworkX, Cytoscape and Gephi. As different software calculate different sets of network properties and some software only allows the computation of a fixed set of properties, it is hard to ensure all tools to compute the same amount of information in running time comparison.

Therefore, we selected and evaluated 8 descriptors (including degree, number of triangles, local clustering coefficient, global clustering coefficient, closeness centrality, betweenness centrality, connectivity centralization, and heterogeneity), which are covered by all these tools (PROFEAT, NetworkX, Cytoscape, and Gephi).

These 8 network properties were computed for 3 undirected unweighted human tissue-specific PPI networks, which were hippocampus, muscle and ovary with

(107 / 315 / 1165 nodes and 146 / 632 / 2230 edges respectively). As the CPU running times on the public tools cannot be directly obtained, we used the job execution times (from the time of input file to the time of output file, roughly the CPU time plus 5 seconds on PROFEAT) instead for measuring the CPU time cost in calculating and obtaining these 8 descriptors.

Table 2-6 described the job execution procedures needed by PROFEAT, NetworkX, Cytoscape, and Gephi to compute these network properties for an undirected unweighted network.

Table 2-6 Job execution procedures needed by PROFEAT and other public tools to compute the selected descriptors of an undirected unweighted network

| Tool Name | | PROFEAT | NetworkX | Cytoscape | Gephi |
|--------------------------|-------------------------|--|--|--|---|
| Platform | | Web-Server | Python Programming | User-Interface Software | User-Interface Software |
| Network Auto-Split | | ✓ | ✗ | ✗ | ✗ |
| Program Expertise | | ✗ | ✓ | ✗ | ✗ |
| Network Visual | | ✗ | ✗ | ✓ | ✓ |
| Job Execution Procedures | Input | Upload the SIF/NET formatted network file into the input field of undirected unweighted network. | Import ‘ <i>networkx</i> ’ in python, and code an extra program to read network file by repetitively calling function ‘ <i>add_edge</i> ’ to create a graph <i>G</i> . | Import the SIF network file. | Import the network spreadsheet with a header indicating ‘ <i>source-target-type</i> ’. Select: ‘ <i>Separator: tab</i> ’ and ‘ <i>Tables: edge table</i> ’. |
| | Computational Operation | No operation is needed. Tick the ‘ <i>slim set</i> ’ option, and click ‘ <i>submit</i> ’. | Each function should be called to calculate the descriptor, e.g. <i>G.degree.values</i> , <i>triangles</i> , <i>clustering</i> , <i>average_clustering</i> , <i>closeness centrality</i> , <i>betweenness centrality</i> | Go to: <i>Tools</i> → <i>NetworkAnalyzer</i> → <i>Network Analysis</i> → <i>Analyze Network</i> . Select ‘ <i>treat network as undirected</i> ’. | Under ‘ <i>Overview</i> ’, and in ‘ <i>Stats</i> ’ panel, click each to run: <i>Average Degree</i> , <i>Avg. Clustering Coefficient</i> , <i>Avg. Path Length</i> |
| | Output | The descriptors are printed on the output page, and a text file is given for download. | The descriptors are stored as python variables. Extra coding is needed to save the descriptor values into a file. | Network-level descriptors are popped out in a result panel, and can be exported by ‘ <i>Save Statistics</i> ’ into a ‘ <i>.netstats</i> ’ file. Node-level descriptors are given in a table panel, and can be copy-&-paste into a text file. | Network-level descriptors are given in <i>Statistics</i> panel, without export or download option. Node-level descriptors are given in ‘ <i>Data Laboratory</i> ’, and an ‘ <i>Export Table</i> ’ button is also given. |
| Computed Descriptors | | The above operations compute the slim-set of network descriptors. | The above operations compute the selected descriptors by calling the specific functions. | The above operations compute all (~23) descriptors provided in Cytoscape. | The above operations compute the selected descriptors by running the certain statistic modules. |

2.3 Results

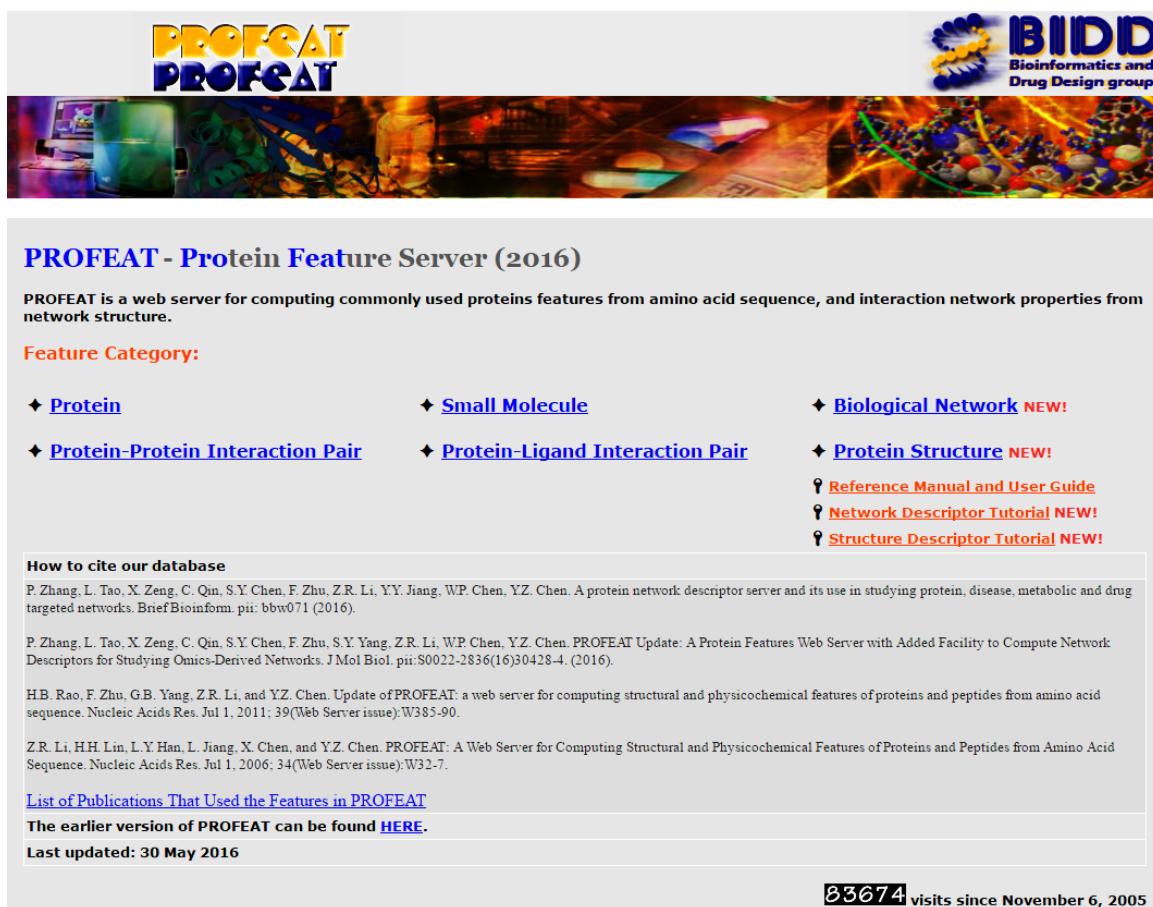
2.3.1 PROFEAT Network Module Structure and Access

To facilitate more extensive use of network descriptors in systems biology, we upgraded PROFEAT webserver by adding the biological network descriptor module at (<http://bidd2.nus.edu.sg/cgi-bin/profeat2016/network/profnew.cgi>). The homepage of the PROFEAT webserver 2016 was given in **Figure 2-2**, the module for biological network descriptor computation (**Figure 2-3**) was composed of five data input fields (undirected un-weighted, undirected edge-weighted, undirected node-weighted, undirected edge-node-weighted, and directed un-weighted networks) with a radio button to choose to compute the full-set or the slim-set of network properties. The flowchart for computing the biological network descriptors was illustrated in **Figure 2-4**.

Given an input network file, each type of network descriptors can be computed by uploading the file in a particular input field followed by the click of the “Submit” button at the bottom of the input fields. Once the job is submitted, the network is read and the adjacency matrix is stored in hashable dictionary data type for faster data access. The deep-first-search is then carried out to check and split the disconnected networks if any. The adjacency-based shortest path lengths and the edge-weighted shortest path lengths are computed and also stored in hashable data matrices, followed by the calculation of each descriptor according to its definition and algorithm (see **Appendix Section B**). The output file for each input network is then stored and printed at such a URL (<http://bidd2.nus.edu.sg/cgi-bin/profeat2016/network/profeat-result.cgi?uid=net-X>), where the numerical ‘X’ is a uniquely assigned 5-digit network id for each

individual job. For a small-sized network input, the output will be immediately displayed at the result window. For a large-sized network input, users could access the URL later to retrieve and download the results, as it may take longer to process large networks.

Figure 2-2 Homepage of PROFEAT webserver 2016



PROFEAT

BIDD
Bioinformatics and
Drug Design group

PROFEAT - Protein Feature Server (2016)

PROFEAT is a web server for computing commonly used proteins features from amino acid sequence, and interaction network properties from network structure.

Feature Category:

- ◆ [Protein](#)
- ◆ [Small Molecule](#)
- ◆ [Biological Network](#) **NEW!**
- ◆ [Protein-Protein Interaction Pair](#)
- ◆ [Protein-Ligand Interaction Pair](#)
- ◆ [Protein Structure](#) **NEW!**
- 🔑 [Reference Manual and User Guide](#)
- 🔑 [Network Descriptor Tutorial](#) **NEW!**
- 🔑 [Structure Descriptor Tutorial](#) **NEW!**

How to cite our database

P. Zhang, L. Tao, X. Zeng, C. Qin, S.Y. Chen, F. Zhu, Z.R. Li, Y.Y. Jiang, W.P. Chen, Y.Z. Chen. A protein network descriptor server and its use in studying protein, disease, metabolic and drug targeted networks. *BriefBioinform.* pii: bbw071 (2016).

P. Zhang, L. Tao, X. Zeng, C. Qin, S.Y. Chen, F. Zhu, S.Y. Yang, Z.R. Li, W.P. Chen, Y.Z. Chen. PROFEAT Update: A Protein Features Web Server with Added Facility to Compute Network Descriptors for Studying Omics-Derived Networks. *J Mol Biol.* pii: S0022-2836(16)30428-4. (2016).

H.B. Rao, F. Zhu, G.B. Yang, Z.R. Li, and Y.Z. Chen. Update of PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Res.* Jul 1, 2011; 39(Web Server issue):W385-90.

Z.R. Li, H.H. Lin, L.Y. Han, L. Jiang, X. Chen, and Y.Z. Chen. PROFEAT: A Web Server for Computing Structural and Physicochemical Features of Proteins and Peptides from Amino Acid Sequence. *Nucleic Acids Res.* Jul 1, 2006; 34(Web Server issue):W32-7.

[List of Publications That Used the Features in PROFEAT](#)

The earlier version of PROFEAT can be found [HERE](#).

Last updated: 30 May 2016

83674 visits since November 6, 2005

Figure 2-3 Biological network descriptor module in PROFEAT webserver

Biological Network Descriptors

Un-Directed Network:

Un-Weighted Network • upload your network, or ☐ upload sample network

Please upload the network file in [SIF format](#) or [NET format](#):

→ **Upload Network File** Choose File No file chosen

Sample Un-Weighted Network File [HERE](#)

Edge-Weighted Network • upload your network, or ☐ upload sample network

Please upload the network file in [Edge-Weighted SIF format](#) or [Edge-Weighted NET format](#):

→ **Upload Network File** Choose File No file chosen

Sample Edge-Weighted Network File [HERE](#)

Edge Weight: 1 2 3

Node-Weighted Network • upload your network, or ☐ upload sample network

Please upload the network file in [SIF format](#) or [NET format](#):

→ **Upload Network File** Choose File No file chosen

Please upload the node weight file in [Node-Weighted TXT format](#):

→ **Upload Node Weight File** Choose File No file chosen

Sample Network File [HERE](#)

Sample Node Weight File [HERE](#)

Node Weight: 1 2 3

Edge & Node-Weighted Network • upload your network, or ☐ upload sample network

Please upload the network file in [Edge-Weighted SIF format](#) or [Edge-Weighted NET format](#):

→ **Upload Network File** Choose File No file chosen

Please upload the node weight file in [Node-Weighted TXT format](#):

→ **Upload Node Weight File** Choose File No file chosen

Sample Edge-Weighted Network File [HERE](#)

Sample Node Weight File [HERE](#)

Edge Weight: 1 2 3

Node Weight: 1 2 3

Directed Network:

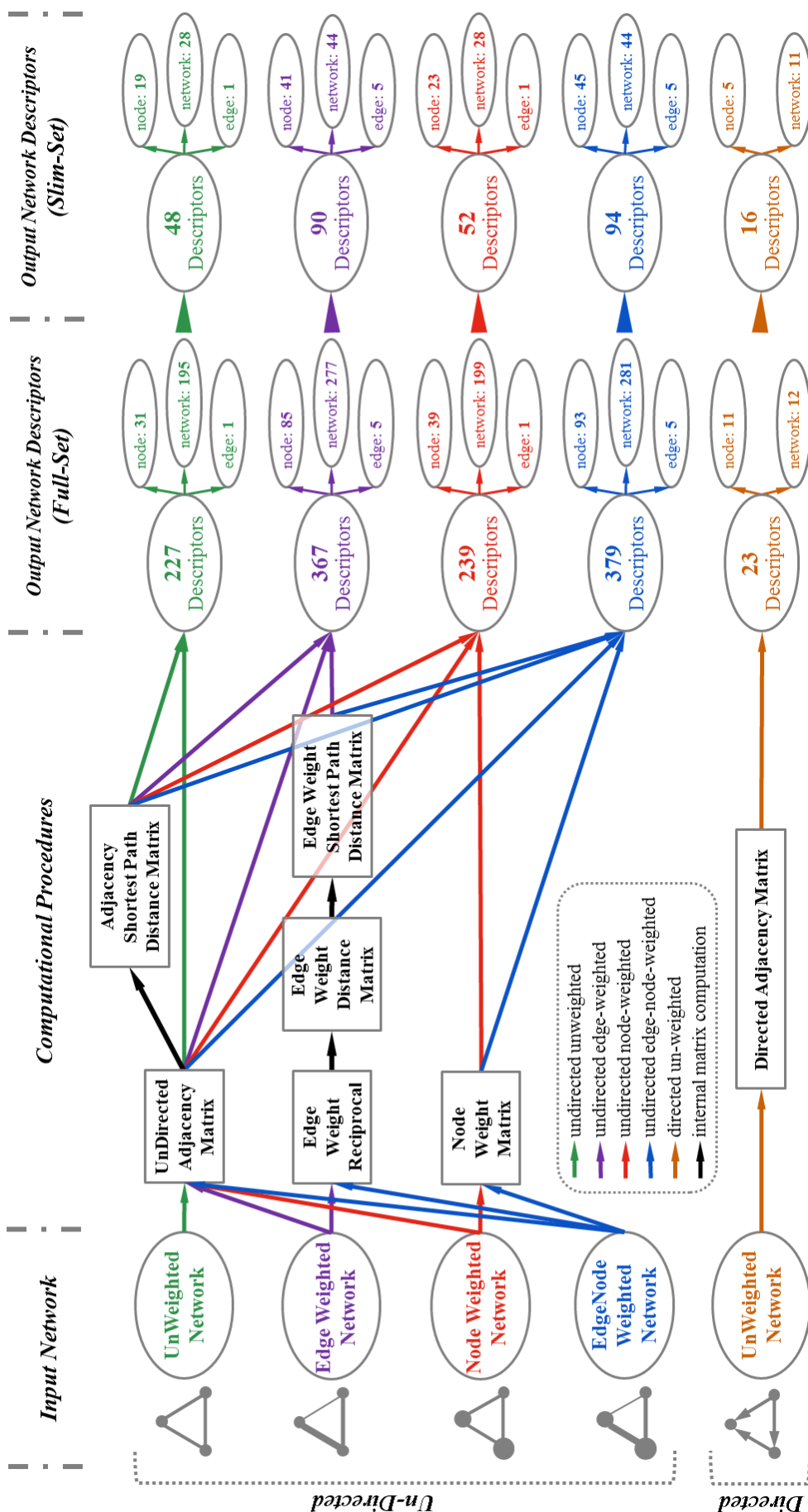
Un-Weighted Network • upload your network, or ☐ upload sample network

Please upload the network file in [Directed SIF format](#) or [Directed NET format](#):

→ **Upload Network File** Choose File No file chosen

Sample Directed Un-Weighted Network File [HERE](#)

• Full • Slim
Submit
Reset

Figure 2-4 Computational flowchart for PROFEAT network descriptors

2.3.2 Input and Output in PROFEAT Network Module

PROFEAT supports either simple interaction file (SIF) or nested network file (NET) as the input network file format. For computing the network descriptors, the following information is required for different network types (**Table 2-7**).

An undirected un-weighted network only needs the binary interaction information, and such a case study is provided (**Table 2-8**). To compute an undirected edge-weighted network, the edge weight is required in the input network file (**Table 2-9**). Note that the edge length is inversely related to the edge weight, as the higher edge weight is typically representing the stronger interaction or the closer relation¹⁴⁶, such that the edge-weighted-distance descriptors are calculated based on the reciprocal of the edge weights. The undirected node-weighted network needs an additional node weight file, where the node label should be correctly matched to the network file (**Table 2-10**). The undirected edge-node-weighted network requires both the edge-weighted network file and the node weight file together for computing the descriptors (**Table 2-11**). For all weighted networks, the weight normalization is carried out, such that weighted properties will be calculated based on both the original and the normalized weight. Lastly, for a directed unweighted network (**Table 2-12**), the SIF format defines that the earlier node points to the latter node, and the NET format defines the directed links in the **arc* section.

The output file of PROFEAT network descriptors is well organized by delivering (1) a header information, starting with “!” to indicate the input network file name, total number of networks, total number of nodes, and total

number of edges, which are based on the original input network given by users; (2) the node-level descriptors in a matrix, where the row represents a network descriptor for all nodes, and the column represents all the network descriptors for one node; and (3) the network-level descriptors, where each row shows one network descriptor.

Table 2-7 The required file(s) for each input network type

| Input Network Type | Required File(s) | | | |
|--------------------------------------|-------------------------|----------------------------|-----------------------|-----------------------|
| | Unweighted Network File | Edge-Weighted Network File | Node-Weight Text File | Directed Network File |
| Undirected Un-Weighted Network | ✓ | | | |
| Undirected Edge-Weighted Network | | ✓ | | |
| Undirected Node-Weighted Network | ✓ | | ✓ | |
| Undirected EdgeNode-Weighted Network | | ✓ | ✓ | |
| Directed Un-Weighted Network | | | | ✓ |

Case studies of different network types were provided in **Table 2-8, 2-9, 2-10, 2-11, and 2-12**. Moreover, quantitative network analysis may get trouble with the mixed networks in the data collection. The available tools have not yet provided the function to detect and split the multiple disconnected networks from a single input file.

To provide a solution for such a case, we implemented an additional function in PROFEAT, and illustrated in **Table 2-13**. A network file “*sample_network_multiple.sif*” containing 3 separated networks was inputted, and the global adjacency was checked if there were multiple separated networks included in a single input file. PROFEAT enabled the automatic detecting of each connected network, renaming of them by adding suffix, ranking by their number of nodes, and computing the network descriptors for each one sub-network respectively.

Table 2-8 Sample input and output of an undirected un-weighted network

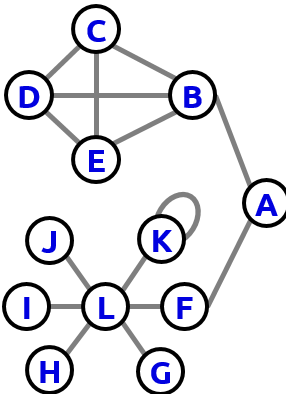
| Sample Input | | |
|---|---|--|
| Network Graphics | Network in SIF | Network in NET |
|  | <pre> A interact B B interact C B interact D B interact E C interact D D interact E C interact E A interact F L interact F L interact K K interact K L interact J L interact I L interact H L interact G </pre> | <pre> *vertices A B C D E F G H I J K L *edges A B B C B D B E C D D E C E A F L F L K K K L J L I L H L G *arcs </pre> |
| Sample Output | | |
| <pre> ! Input Network File Name: sample_network.sif ! Total Number of Networks: 1 ! Total Number of Nodes: 12 ! Total Number of Edges: 15 # Network File: sample_network.sif {12 Nodes; 15 Edges} # # Node-Level Descriptors [G10.0.0] Node Label: A B ... L [G10.1] Un-Weighted Features [G10.1.1] Degree: 2 4 ... 6 ... # # Network-Level Descriptors [G11.1] Un-Weighted Features [G11.1.1] Number of Nodes: 12 [G11.1.2] Number of Edges: 15 ... </pre> | | |

Table 2-9 Sample input and output of an undirected edge-weighted network

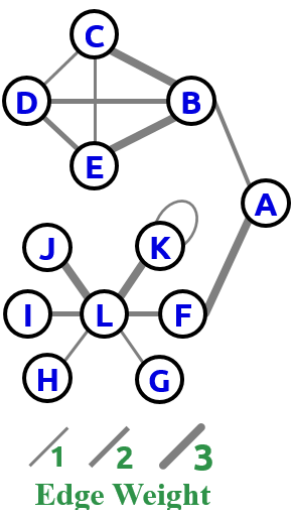
| Sample Input | | | |
|--|--|---|--|
| Network Graphics | Network in SIF | Network in NET | |
|  | <pre>A interact B 2 B interact C 3 B interact D 2 B interact E 3 C interact D 1 D interact E 2 C interact E 1 A interact F 3 L interact F 2 L interact K 3 K interact K 1 L interact J 3 L interact I 2 L interact H 1 L interact G 1</pre> | <pre>*vertices A B C D E F G H I J K L *edges A B 2 B C 3 B D 2 B E 3 C D 1 D E 2 C E 1 A F 3 L F 2 L K 3 K K 1 L J 3 L I 2 L H 1 L G 1 *arcs</pre> | |
| Sample Output | | | |
| <pre>! Input Network File Name: sample_network_edge_weighted.sif ! Total Number of Networks: 1 ! Total Number of Nodes: 12 ! Total Number of Edges: 15 # Network File: sample_network_edge_weighted.sif {12 Nodes; 15 Edges} # # Node-Level Descriptors [G10.0.0] Node Label: A B ... L [G10.1] Un-Weighted Features [G10.1.1] Degree: 2 4 ... 6 ... [G10.2] Original Edge-Weighted Features [G10.2.11] Edge-Weight Avg Shortest Path Length: 1.06 1.24 ... 1.0 ... [G10.2N] Normalized Edge-Weighted Features [G10.2N.11] N. Edge-Weight Avg Shortest Path Length: 0.46 0.56 ... 0.46 ... # # Network-Level Descriptors [G11.1] Un-Weighted Features [G11.1.1] Number of Nodes: 12 [G11.1.2] Number of Edges: 15 ... [G11.2] Original Edge-Weighted Features [G11.2.14] Edge-Weight Total Distance: 93.0 ... [G11.2N] Normalized Edge-Weighted Features [G11.2N.14] N. Edge-Weight Total Distance: 45.2 ...</pre> | | | |

Table 2-10 Sample input and output of an undirected node-weighted network

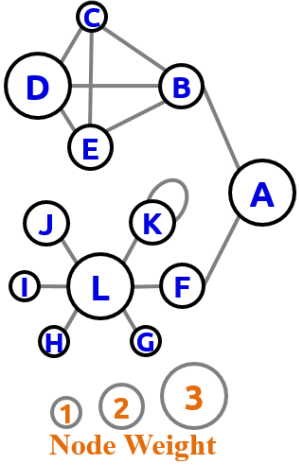
| Sample Input | | | |
|--|---|--|--|
| Network Graphics | Network in SIF | Network in NET | Node Weight |
|  <p>Node Weight</p> | <pre> A interact B B interact C B interact D B interact E C interact D D interact E C interact E A interact F L interact F L interact K K interact K L interact J L interact I L interact H L interact G </pre> | <pre> *vertices A B C D E F G H I J K L *edges A B B C B D B E C D D E C E A F L F L K K K L J L I L H L G *arcs </pre> | <pre> A 3 B 2 C 1 D 3 E 2 F 2 G 1 H 1 I 1 J 2 K 2 L 3 </pre> |
| Sample Output | | | |
| <pre> ! Input Network File Name: sample_network.sif ! Input Node Weight File Name: sample_network_node_weighted.txt ! Total Number of Networks: 1 ! Total Number of Nodes: 12 ! Total Number of Edges: 15 # Network File: sample_network.sif {12 Nodes; 15 Edges} # # Node-Level Descriptors [G10.0.0] Node Label: A B ... L [G10.1] Un-Weighted Features [G10.1.1] Degree: 2 4 ... 6 ... [G10.3] Original Node-Weighted Features [G10.3.38] Node Weight: 3 2 ... 3 ... [G10.3N] Normalized Node-Weighted Features [G10.3N.38] N. Node Weight: 1 0.5 ... 1 ... # # Network-Level Descriptors [G11.1] Un-Weighted Features [G11.1.1] Number of Nodes: 12 [G11.1.2] Number of Edges: 15 ... [G11.3] Original Node-Weighted Features [G11.3.150] Total Node Weight: 23 ... [G11.3N] Normalized Node-Weighted Features [G11.3N.150] N. Total Node Weight: 5.53 ... </pre> | | | |

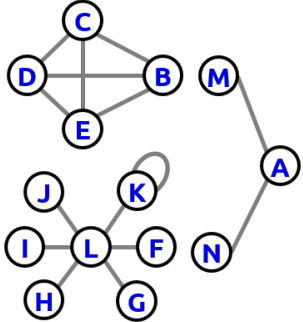
Table 2-11 Sample input and output of an undirected edge-node-weighted network

| Sample Input | | | |
|--|--|---|--|
| Network Graphics | Network in SIF | Network in NET | Node Weight |
| <p>Edge Weight</p> <p>Node Weight</p> | <pre> A interact B 2 B interact C 3 B interact D 2 B interact E 3 C interact D 1 D interact E 2 C interact E 1 A interact F 3 L interact F 2 L interact K 3 K interact K 1 L interact J 3 L interact I 2 L interact H 1 L interact G 1 </pre> | <pre> *vertices A B C D E F G H I J K L *edges A B 2 B C 3 B D 2 B E 3 C D 1 D E 2 C E 1 A F 3 L F 2 L K 3 K K 1 L J 3 L I 2 L H 1 L G 1 *arcs </pre> | <pre> A 3 B 2 C 1 D 3 E 2 F 2 G 1 H 1 I 1 J 2 K 2 L 3 </pre> |
| Sample Output | | | |
| <pre> ! Input Network File Name: sample_network_edge_weighted.sif ! Input Node Weight File Name: sample_network_node_weighted.txt ! Total Number of Networks: 1 ! Total Number of Nodes: 12 ! Total Number of Edges: 15 # Network File: sample_network_edge_weighted.sif {12 Nodes; 15 Edges} # # Node-Level Descriptors [G10.0.0] Node Label: A B ... L [G10.1] Un-Weighted Features [G10.1.1] Degree: 2 4 ... 6 ... [G10.2] Original Edge-Weighted Features ... [G10.3] Original Node-Weighted Features ... [G10.2N] Normalized Edge-Weighted Features ... [G10.3N] Normalized Node-Weighted Features ... # # Network-Level Descriptors [G11.1] Un-Weighted Features [G11.1.1] Number of Nodes: 12 [G11.1.2] Number of Edges: 15 ... [G11.2] Original Edge-Weighted Features ... [G11.3] Original Node-Weighted Features ... [G11.2N] Normalized Edge-Weighted Features ... [G11.3N] Normalized Node-Weighted Features ... </pre> | | | |

Table 2-12 Sample input and output of a directed un-weighted network

| Sample Input | | |
|--|--|--|
| Network Graphics | Network in SIF | Network in NET |
| | <pre> A direct-to B C direct-to B D direct-to B E direct-to B D direct-to C E direct-to D E direct-to C F direct-to A F direct-to L K direct-to L K direct-to K J direct-to L L direct-to I L direct-to H L direct-to G </pre> | <pre> *vertices A B C D E F G H I J K L *edges *arcs A B C B D B E B D C E D E C F A F L K L K K J L L I L H L G </pre> |
| Sample Output | | |
| <pre> ! Input Network File Name: sample_network_directed.sif ! Total Number of Networks: 1 ! Total Number of Nodes: 12 ! Total Number of Edges: 15 # Network File: sample_network_directed.sif {12 Nodes; 15 Edges} # # Node-Level Descriptors [G10.0.0] Node Label: A B ... L [G10.4] Directed Features [G10.4.41] In-Degree: 1 4 ... 3 [G10.4.42] Out-Degree: 1 0 ... 3 ... # # Network-Level Descriptors [G11.4] Directed Features [G11.4.1] Number of Nodes: 12 [G11.4.2] Number of Edges: 15 ... </pre> | | |

Table 2-13 Sample input and output of a single file with multiple networks

| Sample Input | | |
|---|---|--|
| Network Graphics | Network in SIF | Network in NET |
|  | <pre> A interact M B interact C B interact D B interact E C interact D D interact E C interact E A interact N L interact F L interact K K interact K L interact J L interact I L interact H L interact G </pre> | <pre> *vertices A B C D E F G H I J K L M N *edges A M B C B D B E C D D E C E A N L F L K K K L J L I L H L G *arcs </pre> |
| Sample Output | | |
| <pre> ! Input Network File Name: sample_network_multiple.sif ! Total Number of Networks: 3 ! Total Number of Nodes: 14 ! Total Number of Edges: 15 # Network File: sample_network_multiple_sub_1.sif {7 Nodes; 7 Edges} # # Node-Level Descriptors [G10.0.0] Node Label: F G ... L [G10.1] Un-Weighted Features [G10.1.1] Degree: 1 1 ... 6 ... # # Network-Level Descriptors [G11.1] Un-Weighted Features # Network File: sample_network_multiple_sub_2.sif {4 Nodes; 6 Edges} # # Node-Level Descriptors [G10.0.0] Node Label: B C D E [G10.1] Un-Weighted Features ... # # Network-Level Descriptors [G11.1] Un-Weighted Features # Network File: sample_network_multiple_sub_3.sif {3 Nodes; 2 Edges} # # Node-Level Descriptors [G10.0.0] Node Label: A M N [G10.1] Un-Weighted Features ... # # Network-Level Descriptors [G11.1] Un-Weighted Features </pre> | | |

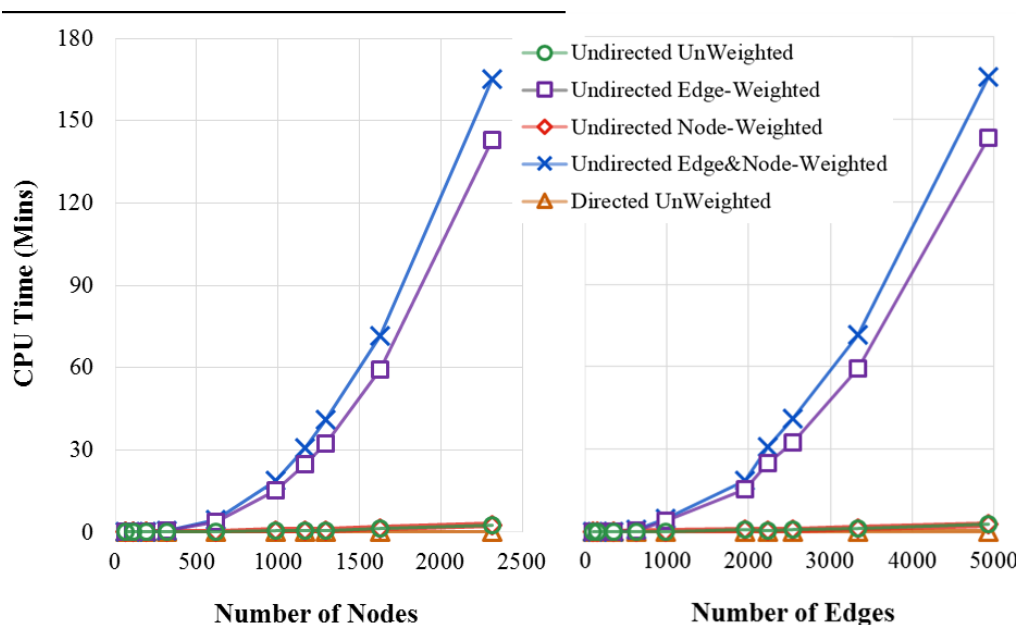
2.3.3 Comparative Performance Evaluations

The evaluation of CPU running time in computing the slim set of PROFEAT network descriptors on the 10 testing networks was summarized in **Table 2-14** and **Figure 2-5**. Approximately, the CPU time for the un-weighted network was within 30 seconds for a network having no more than 1000 nodes or edges, the CPU time was less than 1 minute if the network size is less than 1500 nodes or 3000 edges, while the CPU time increased to over 3 minutes if the network gets larger than 2,300 nodes or 5,000 edges. On the other hand, the CPU time for the edge-weighted network was about 15 minutes if the network has around 1,000 nodes or 2,000 edges. The edge-node-weighted network required the highest computational time, by costing about 30 minutes for the network with 1,200 nodes or 2,300 edges. For the directed unweighted network descriptors, it cost no more than 4 seconds for all the testing networks.

Table 2-14 CPU time in computing the slim set of PROFEAT network descriptors for 10 human tissue-specific PPI networks of 5 network types

| Tissue | Network Size | | CPU Time (Mins) for Different Network Types | | | | |
|-----------------|--------------|--------------|---|---------------|---------------|-------------------|----------|
| | No. of Nodes | No. of Edges | Un-Weighted | Edge-Weighted | Node-Weighted | EdgeNode-Weighted | Directed |
| Lymph Node | 63 | 91 | 0.008 | 0.014 | 0.009 | 0.018 | 0.007 |
| Hippocampus | 107 | 146 | 0.010 | 0.033 | 0.011 | 0.040 | 0.007 |
| Bone Marrow | 189 | 348 | 0.018 | 0.134 | 0.018 | 0.155 | 0.008 |
| Muscle | 315 | 632 | 0.044 | 0.551 | 0.045 | 0.660 | 0.009 |
| Small Intestine | 616 | 980 | 0.189 | 3.83 | 0.178 | 4.66 | 0.013 |
| Colon | 988 | 1951 | 0.672 | 15.42 | 0.653 | 18.67 | 0.021 |
| Ovary | 1165 | 2230 | 0.609 | 24.87 | 0.612 | 30.71 | 0.026 |
| Spleen | 1292 | 2543 | 0.765 | 32.34 | 0.761 | 40.98 | 0.029 |
| Pancreas | 1625 | 3336 | 1.38 | 59.26 | 1.39 | 71.45 | 0.043 |
| Lung | 2317 | 4942 | 2.68 | 143.10 | 2.72 | 165.33 | 0.067 |

Figure 2-5 CPU time (mins) in computing the slim set of PROFEAT network descriptors for the networks in Table 2-14 with respect to the number of nodes (left) and the number of edges (right)



In comparison of the eight computed network descriptor value and the job execution time by PROFEAT and other public tools (NetworkX, Cytoscape and Gephi), three human tissue-specific PPI networks (hippocampus (107 nodes, 146 edges), muscle (315 nodes, 632 edges) and ovary (1165 nodes, 2230 edges)) were tested. The comparative results were summarized in **Table 2-15**, where the job execution time counted the time cost from the time of input file to the time of output file (**Table 2-6**).

The eight evaluated network descriptors included three network-level descriptors and five node-level descriptors. The maximum values and the corresponding node's gene symbols were given for the node-level descriptors. PROFEAT computed values of all the evaluated descriptors in the three networks were in good agreement with those computed from other tools, while there were some minor variations might be caused by rounding precision.

However, we observed that Gephi software calculated significantly different values for local/global clustering coefficient for all the three networks. The algorithm for computing clustering coefficient in Gephi was not found anywhere (its website or its publication), such that we assumed Gephi might apply different algorithms or definitions in this case.

The job execution times of PROFEAT slim-set for the first two networks were faster than those of the public tools (5 seconds *vs* 10-15 seconds, and 8 seconds *vs* 15-20 seconds), and PROFEAT takes higher time cost than the other tools for the third network (45 seconds *vs* 30 seconds). The longer job execution times of PROFEAT arose from its computation of a larger number of network descriptors in contrast to the computation of a smaller set of user-selected descriptors by the other tools. However, users may not have the prior knowledge in the selection of network descriptors, especially the biologist.

Table 2-15 Comparison of the computed network descriptor value and the job execution time for three human tissue-specific PPI networks (A. hippocampus, B. muscle, and C. ovary) by PROFEAT and other public tools NetworkX, Cytoscape and Gephi

(A) Hippocampus-specific PPI network (107 nodes, 146 edges)

| Tool Name | PROFEAT | NetworkX | Cytoscape | Gephi |
|-------------------------------|-----------------------------------|--------------------|--------------------|------------------------------|
| Network Descriptor | Computed Network Descriptor Value | | | |
| Degree | 16 (SRC) | 16 (SRC) | 16 (SRC) | 16 (SRC) |
| Number of Triangle | 3 (DLG2) | 3 (DLG2) | n.a. | 4 (ACTB) |
| Closeness Centrality | 0.234 (GRIN2B) | 0.234 (GRIN2B) | 0.233 (GRIN2B) | 0.231 (GRIN2B) |
| Betweenness Centrality | 0.443 (GRIN2B) | 0.443 (GRIN2B) | 0.443 (GRIN2B) | 0.443 (GRIN2B) |
| Local Clustering Coefficient | 1 (LIN7A, PDCD6IP) | 1 (LIN7A, PDCD6IP) | 1 (LIN7A, PDCD6IP) | 1 (7 Proteins) ^{A4} |
| Global Clustering Coefficient | 0.026 | 0.026 | 0.026 | 0.227 |
| Connectivity Centralization | 0.118 | n.a. | 0.115 | n.a. |
| Heterogeneity | 0.910 | n.a. | 0.930 | n.a. |
| | Job Execution Time | | | |
| | ~ 5 seconds | ~ 10 seconds | ~ 30 seconds | ~ 30 seconds |

^{A4} List of 7 proteins that have local clustering coefficient = 1, by Gephi: [APOE, DYRK1A, LIN7A, NEK9, RABAC1, RELN, SYNE1]

Table 2-15 (*continue*) Comparison of the computed network descriptor value and the job execution time for three human tissue-specific PPI networks (A. hippocampus, B. muscle, and C. ovary) by PROFEAT and other public tools NetworkX, Cytoscape and Gephi

(B) Muscle-specific PPI network (315 nodes, 632 edges)

| Tool Name | PROFEAT | NetworkX | Cytoscape | Gephi |
|-------------------------------|-----------------------------------|----------------------------------|----------------------------------|----------------------------------|
| Network Descriptor | Computed Network Descriptor Value | | | |
| Degree | 20 (AR) | 20 (AR) | 20 (AR) | 20 (AR) |
| Number of Triangle | 12 (INSR) | 12 (INSR) | n.a. | 12 (INSR, PTK2, DMD) |
| Closeness Centrality | 0.310 (AKT1) | 0.310 (AKT1) | 0.309 (AKT1) | 0.309 (AKT1) |
| Betweenness Centrality | 0.213 (AKT1) | 0.213 (AKT1) | 0.213 (AKT1) | 0.213 (AKT1) |
| Local Clustering Coefficient | 1 (18 Proteins) ^{B1} | 1 (18 Proteins) ^{B2} | 1 (18 Proteins) ^{B3} | 1 (34 Proteins) ^{B4} |
| Global Clustering Coefficient | 0.104 | 0.104 | 0.104 | 0.315 |
| Connectivity Centralization | 0.048 | n.a. | 0.047 | n.a. |
| Heterogeneity | 0.916 | n.a. | 0.924 | n.a. |
| | Job Execution Time | | | |
| | ~ 25 seconds | ~ 15 seconds | ~ 40 seconds | ~ 40 seconds |

^{B1} List of 18 proteins that have local clustering coefficient = 1, by PROFEAT: [AVEN, BCL2L10, BCL6, CD36, CFL2, DLL1, DVL1, DVL3, EPB49, FLT4, FOXO3, IKZF2, IKZF5, IRAK2, IRAK3, NLRP1, PFKM, VEGFB]

^{B2} List of 18 proteins that have local clustering coefficient = 1, by NetworkX: [AVEN, BCL2L10, BCL6, CD36, CFL2, DLL1, DVL1, DVL3, EPB49, FLT4, FOXO3, IKZF2, IKZF5, IRAK2, IRAK3, NLRP1, PFKM, VEGFB]

^{B3} List of 18 proteins that have local clustering coefficient = 1, by Cytoscape: [AVEN, BCL2L10, BCL6, CD36, CFL2, DLL1, DVL1, DVL3, EPB49, FLT4, FOXO3, IKZF2, IKZF5, IRAK2, IRAK3, NLRP1, PFKM, VEGFB]

^{B4} List of 34 proteins that have local clustering coefficient = 1, by Gephi: [FOXO3, MAP3K13, AVEN, CFL2, DLL1, EPB49, CD36, ARMCX2, NFATC2, PFKM, PIK3CD, ERBB2IP, VEGFB, IRAK2, IRAK3, UBE2G1, IRF5, PAK3, MBD2, MSN, ALDH2, DVL1, DVL3, BCL6, MEF2C, PKM2, PYGM, IKZF2, IKZF5, CDC42BPA, RPS19, LDHA, STAU1, MSTN]

Table 2-15 (*continue*) Comparison of the computed network descriptor value and the job execution time for three human tissue-specific PPI networks (A. hippocampus, B. muscle, and C. ovary) by PROFEAT and other public tools NetworkX, Cytoscape and Gephi

(C) Ovary-specific PPI network (1165 nodes & 2230 edges)

| Tool Name | PROFEAT | NetworkX | Cytoscape | Gephi |
|-------------------------------|-----------------------------------|----------------------------------|----------------------------------|----------------------------------|
| Network Descriptor | Computed Network Descriptor Value | | | |
| Degree | 48 (AR) | 48 (AR) | 48 (AR) | 48 (AR) |
| Number of Triangle | 31 (SRC) | 31 (SRC) | n.a. | 31 (SRC) |
| Closeness Centrality | 0.307 (AR) | 0.307 (AR) | 0.307 (AR) | 0.307 (AR) |
| Betweenness Centrality | 0.208 (AR) | 0.208 (AR) | 0.208 (AR) | 0.208 (AR) |
| Local Clustering Coefficient | 1 (28 Proteins) ^{C1} | 1 (28 Proteins) ^{C2} | 1 (28 Proteins) ^{C3} | 1 (78 Proteins) ^{C4} |
| Global Clustering Coefficient | 0.061 | 0.061 | 0.061 | 0.248 |
| Connectivity Centralization | 0.037 | n.a. | 0.037 | n.a. |
| Heterogeneity | 1.205 | n.a. | 1.228 | n.a. |
| | Job Execution Time | | | |
| | ~ 6 minutes | ~ 30 seconds | ~ 1 minute | ~ 1 minute |

^{C1} List of 28 proteins that have local clustering coefficient = 1, by PROFEAT: [BMP15, CYP11A1, DCLRE1C, DIS3, EXO1, FSHR, HR, IRAK3, LPHN1, MSH3, NLRP1, NUP50, ORC4L, PHC2, PTPRE, RAD52, RBP1, SAP30, SFMBT1, TBL1X, TBL1XR1, TBPL1, TNFRSF10D, TNKS1BP1, ZNF652]

^{C2} List of 28 proteins that have local clustering coefficient = 1, by NetworkX: [BMP15, CYP11A1, DCLRE1C, DIS3, EXO1, FSHR, HR, IRAK3, LPHN1, MSH3, NLRP1, NUP50, ORC4L, PHC2, PTPRE, RAD52, RBP1, SAP30, SFMBT1, TBL1X, TBL1XR1, TBPL1, TNFRSF10D, TNKS1BP1, ZNF652]

^{C3} List of 28 proteins that have local clustering coefficient = 1, by Cytoscape: [BMP15, CYP11A1, DCLRE1C, DIS3, EXO1, FSHR, HR, IRAK3, LPHN1, MSH3, NLRP1, NUP50, ORC4L, PHC2, PTPRE, RAD52, RBP1, SAP30, SFMBT1, TBL1X, TBL1XR1, TBPL1, TNFRSF10D, TNKS1BP1, ZNF652]

^{C4} List of 78 proteins that have local clustering coefficient = 1, by Gephi: [CYP11A1, DCLRE1C, NUP50, XPO4, LPHN1, NLRP1, ARNT2, IRAK3, RORA, SAP30, RBP1, FOXP1, BMP15, LYPLA1, ERP29, FSHR, MERTK, PDZRN3, FLNB, PPP1R9A, TBL1X, TBL1XR1, TNFRSF10D, MAML1, HR, PCBD1, ORC4L, STX3, PAK4, BICD2, GC, PHC2, SFMBT1, LTBP1, ZNF652, TPI1, NFAT5, FIGF, VEGFC, FBXW11, TRIM25, LATS1, DIS3, TBPL1, NBR1, TERF2, EMILIN1, PIK3CD, EXO1, MSH3, CLDN1, LDHA, OXTR, RPS6KA2, SYNJ2, CADM1, TOB2, CNOT6, MBTPS1, STON2, DGKH, TESC, MTUS1, CTSC, AKR7A3, EPHA8, EPHA3, P4HA2, BCL2L10, ANGPT2, ANGPTL1, TNKS1BP1, RNF216, ASCL3, ADCY5, RPS6KC1, SETD2, FGG]

2.4 Discussion

The usefulness of the network descriptors in characterizing the connectivity, topology and complexity properties of the biological networks are illustrated in the following cases of literature-reported studies of the biological networks built from the genome (e.g. genetic interaction network¹⁶⁷), interactome (e.g. protein-protein interactions¹³⁶ and drug-target interactions¹⁴⁰), transcriptome (e.g. gene co-expression network based on the pairwise profile-similarity comparison¹³², and gene regulatory network derived from regulatory interactions between transcription factors and target genes¹⁵⁷), metabolome (e.g. metabolomics correlation network constructed based on the correlations among metabolite levels¹⁶⁸), and diseasome (e.g. human disease-gene network generated from OMIM disorder-disease gene associations³²) profiles respectively.

2.4.1 Applications of network descriptors in genome-derived networks

A yeast genetic interaction network of ~4,000 cooperative gene-pairs among ~1,000 genes has been constructed by the systematic analysis of functionally cooperative double mutants, which has been subsequently analyzed by using the network descriptor degree (the number of mutant genes cooperative with a mutant gene) to show that the network follows a power-law degree distribution containing many genes with few interactions and a few genes with many interactions, and these few genes are more important for fitness than less connected genes¹⁴⁷.

2.4.2 Applications of network descriptors in interactome-derived networks

The extensive studies of protein-protein interaction have generated rich knowledge and data for investigating the network behavior of proteins. For instance, a protein-protein interaction map has been constructed as a resource for annotating the proteome, which has been used for probing the topological properties of the human protein-protein network that connects 1,705 human proteins via 3,186 interactions³⁸. Based on the analysis of the network descriptors of this network, it was found that the average clustering coefficient, a measure of the tendency of the proteins to form groups, diminishes when the number of interactions per protein increases, indicating a hierarchical organization of the network. The topological coefficient, a measure of the extent to which a protein shares interaction partners with other proteins, decreases with the number of connections, suggesting that hubs do not have more common neighbors than proteins with fewer connections.

In another study of the yeast protein-protein interaction network of 4,549 physical interactions between 3,278 proteins, based on the analysis of the network descriptor degree (the number of proteins interacting with a protein), it has been found that the links between high-degree proteins are systematically suppressed whereas those between a high and a low degree protein are favored, which decreases the likelihood of crosstalk between different functional modules of the cell and increases the overall robustness of a network by localizing effects of harmful perturbations¹³⁶.

The targets of approved drugs possess such specific target-like characteristics as the appropriate druggable structures, substantial dissimilarity to human

proteins, and distinguished systems and tissue distribution profiles^{32,37,169,170}. In particular, these targets are distinguished in the human protein-protein networks such that specific network descriptors may be used as the quantitative determinants of drug targets in these networks^{32,37}. In a study of the global relationships between drug targets in the human interactome network³², a target protein network was constructed by using 394 targets of 890 approved drugs wherein these targets are connected by their commonly targeted drug(s). In this network, 788 drugs share targets and 305 targets are connected to one another. This network was then overlaid onto the human protein-protein network¹⁷¹ composed of 7,533 proteins and 22,052 non-self-interacting and non-redundant interactions. Overall, 260 targets were mapped onto the human protein-protein network, which on average have a higher degree (with 42% more interacting proteins) than that of the non-target proteins in the same network.

In the study of the druggability properties of 304 targets of approved drugs in the human protein-protein interaction network³⁷, a protein-protein network model of 7,764 proteins and 28,149 interactions was derived from the Human Protein Reference Database⁸⁶. The drug targets were found to have the increased average betweenness centrality, suggesting their tendency to bridge two or more clusters of relatively closely interacting proteins.

In the analysis of a drug-target network derived by docking 1,000 FDA-approved drugs to 2,500 protein pockets of the human genome¹⁴⁰, three network descriptors degree (the number of drugs sharing the same target with a drug), betweenness centrality (the number of times a drug serves as a linking bridge along the shortest path between two drugs) and clustering coefficient (the

tendency of a drug to form clusters with other drugs in the network) have been used for comparative analysis of this network with respect to a compound-protein network derived by docking 1,592 compounds from the NCI diversity set to 1918 protein pockets, which showed that the drug-target network has a significantly lower degree, comparable betweenness and slightly lower clustering coefficient, suggesting that the drugs share less number of targets and are more loosely connected than the NCI active compounds. In particular, anticancer drugs are among the drugs with the highest degree and betweenness, and most anticancer compounds are also the most selective compounds in the network.

2.4.3 Applications of network descriptors in transcriptome-derived networks

Based on the exhaustive pairwise gene expression profile similarity comparison, a yeast gene co-expression network has been constructed and analyzed by using two network descriptors, degree (the number of genes co-expressed with a gene) and clustering coefficient (the level of the clustering of co-expressed genes)¹³². The analysis indicated that the network follows a clear power-law degree distribution not correlated with the mean expression levels, and the average clustering coefficient of the network is several orders of magnitude greater than that predicted by a pure scale-free growth model, indicative of an underlying hierarchical organization of modularity in the network. The degree descriptor has also been used to derive a co-expressed protein-protein interaction degree measure as a robust predictor of protein evolutionary rate irrespective of experimental method¹⁴¹.

The yeast gene regulatory network has been built from 7,074 regulatory interactions between 142 transcription factors and 3,420 target genes (interactions between two transcription factors, or a transcription factor and a non-transcription factor target)¹⁵⁷. The topological property of that network was studied by using four network descriptors in-degree (the number of transcription factors regulating a target), out-degree (the number of target genes for each transcription factor), path length (the number of intermediate regulators between a transcription factor and a terminating target gene), and clustering coefficient (the level of inter-regulation of the transcription factor). The small in-degrees indicate that transcription factors regulate in simpler combinations, and the large out-degrees imply that each transcription factor has greater regulatory influence by targeting more genes simultaneously. The short paths signify faster propagation of the regulatory signal, while long paths suggest slower action arising from the formation of regulatory chains to control intermediate phases. High clustering coefficients indicate greater inter-regulation between transcription factors. The analysis of two sub-networks in the endogenous processes (cell cycle and sporulation) and three sub-networks of the exogenous states (diauxic shift, DNA damage and stress response) has suggested that these networks have been evolved to produce large-scale rapid responses in the exogenous states, and carefully coordinated processes in the endogenous conditions.

2.4.4 Applications of network descriptors in metabolome-derived networks

Network descriptors have been applied to identify the highly predictive biomarker candidates for obesity¹³⁸, where blood samples were collected and 59 metabolites were measured by quantitative MS/MS, from 52 persons (22 obesity and 30 control). To infer the metabolic networks, metabolites were represented by nodes, and edges were defined by two statistical methods to generate one correlation network and one ratio network. The correlation network was constructed based on the correlation (adjusted p-value) of metabolites concentrations between obesity and control, and the ratio network was constructed based on the statistical test for difference between the ratios of metabolites in obesity and control. Once the networks inferred, only the largest connected components were extracted by discarding the disconnected parts, for computing the network descriptors (degree, clustering coefficient, eccentricity, distance deviation, and information-theoretic measures). These descriptors were ranked by feature selection, and applied for classification to evaluate the predictive power. Finally, this study obtained the highly discriminating metabolic biomarker candidates for obesity.

A metabolomic correlation network in *Arabidopsis* has been constructed based on the significant correlations among the metabolite levels in the root tissues and the aerial parts obtained by the gas chromatography-time-of-flight/mass spectrometry and published information respectively¹⁶⁸. Six network descriptors have been used to assess the threshold-dependent changes in the network topology: degree (the number of metabolites significantly correlated to a metabolite), clustering coefficient (the level of the clustering of significantly correlated metabolites), network density (existing metabolite correlations

divided by the number of possible correlations), average path length (the extent of correlation between each metabolite and all the rest metabolites), number of connected components (number of metabolites correlated with another metabolite), and the number of edges (number of metabolite correlations). This work revealed the networks that contain tissue- and/or genotype- dependent metabolomics clusters, and some of these clusters are related to the respective biochemical pathways¹⁶⁸.

2.4.5 Applications of network descriptors in diseasome-derived networks

A human disease-gene network of 1,284 distinct disorders and 1,777 disease-related genes has been generated from the OMIM-based disorder-disease gene associations such that a link is established between a disorder and a disease gene if a mutation in that gene leads to the disorder¹³⁰. The distribution behavior of the drug targets in this network has also been studied³² by using the network descriptor degree (the number of genes connected to a disorder or the number of disorders connected to a gene), which showed that, for both the disorder nodes connected to a drug target and the disease gene nodes encoding a drug target, their average degrees are higher than random cases. Moreover, the distribution of the drug targets in this network exhibits a clustered pattern with the targets primarily enriched in some regions of the network. Specifically, starting from a node in the network, the ratio of drug targets with respect to the distance from the node was measured, which showed a strong enrichment in the first and the second neighbors and thus a bias toward clustering of drug targets in the network.

2.4.6 Perspectives

The systems biology studies frequently require the use of multiple approaches from the perspectives of genetic sequences, protein sequences, protein structures, molecular interactions and biological networks/pathways. Biological functional studies particularly at the cellular level or systems level can be greatly enhanced by the exploration of the network/graph theories, descriptors and models developed in other fields^{2,3,4,5,6,12,13,40} and also in the study of systems biology^{8,26,27,31,32,172}, which offers much more expanded perspectives and avenues to the understanding of biological systems and cellular internal organization, evolution and dynamic behavior than the studies based on the concept of individual molecule or independent group of molecules⁸.

However, the progress towards more extensive and more reliable network-based studies of the biological, disease and therapeutically relevant processes may be constrained by the insufficient information about biological networks, the limited capability of the available network analysis and modeling methods, and the inadequate computational resources for facilitating the analysis and modeling of biological networks.

By providing the facility of the computation of diverse network descriptors useful for studying biological systems, PROFEAT complements the other resources in the information^{164,173}, modeling tools¹⁷⁴, parameters¹⁷⁵ of biological networks. These plus more enhanced ability in generating and analyzing various biological networks from the genome¹⁶⁷, interactome^{136,140}, transcriptome^{132,157}, metabolome¹⁶⁸, and diseasome³² profiles will enable more comprehensive and in-depth investigations of the functional roles and the dynamics of the biological

networks in regulating biological and cellular systems⁸, disease processes²⁶ and therapeutic actions³¹.

CHAPTER 3 TISPIN Database Development for Human Tissue-Specific Protein Interaction Networks

3.1 Background and Motivations

In the last few years, increasing efforts have been observed in studying the human tissue-specific networks, which reflect the different roles of proteins, genes, and pathways in diverse complex tissues and cells^{62,63}. The human tissue-specific networks have offered more precise focus and improved capability at the tissue- or cell- level for studying functional biology, disease / drug-response mechanism, and discovery of the biomarkers / targets for the diagnostics / therapeutics of the diseases.

One main research interests is to understand and compare the functional capability between the tissue-specific network and the global network^{64,65,66,67,68}. *Bossi et al.* introduced the human tissue-specific PPI networks⁶⁴ by combining the physical protein interactions and the tissue-specific gene expression⁷³, and further defined house-keeping proteins and tissue-specific proteins in PPI networks⁶⁴. *Lin et al.* analyzed the topological and organizational properties of house-keeping proteins and tissue-specific proteins in 19 human tissue-specific PPI networks⁶⁵, and found that the house-keeping proteins favor to occupy central positions, while the tissue-specific proteins were more peripheral⁶⁵. *Lopes et al.* observed the substantial enrichment of specific proteins and pathways in the tissue-specific networks, while in contrast, the global networks had no significant enrichment identified⁶⁶. As evident from these studies, topological properties and functional enrichment of the tissue-specific PPI

networks have been shown significantly different from the global PPI networks, and the analysis of the global network instead of the tissue-specific network would lead the loss of biological information, and result in the misinterpretation of biological function.

Another main research interest is focused on the study of disease mechanisms through the application of tissue-specific networks^{62,63,69,70,71,72}. *Greene et al.* deemed that the understanding of tissue-specific networks was important in identifying the different functional roles of genes and proteins across different tissues, thus facilitating the development of the improved diagnostics and therapeutics⁶³. *Dezso et al.* analyzed the network topology and ontology enrichment of tissue-specific networks, and found that tissue-specific genes/proteins were more likely to be biomarkers and drug targets⁶⁹. *Guan et al.* utilized the tissue-specific networks to predict the gene/protein candidates associated with certain phenotype or disease⁶². *Shahin et al.* demonstrated a case study on brain-specific interactome for Alzheimer's and Parkinson's diseases, and implicated the disease-related pathways and the potential therapeutic targets⁷⁰. Tissue-specific PPI networks have also improved the prioritization of disease-causing genes^{71,76}, and enhanced the understandings of the molecular mechanisms underlying the hereditary diseases⁷².

Based on these literature reviews, it is observed that the tissue-specific PPI networks transcend the global PPI network with the considerably improved capability in addressing the tissue/cell-level questions for functional biology, disease mechanism, and target/biomarker identification. However, in these studies, the tissue distributions of proteins were determined by microarray gene

expression data. *Emig et al.* re-evaluated the tissue-specific networks by using both microarray and RNAseq gene expression, and observed that many interactions, classified as highly tissue-specific by microarray data, were substantially found in all tissues by using RNAseq data⁶⁷. *Emig et al.* concluded that microarray was not sensitive enough for the low expressed genes, thus the tissue distribution derived from microarray data was less reliable^{67,77}. This finding was also reported in many other studies that compared microarray technology against sequencing technology^{78,79,80,81}. Nevertheless, a recent study showed that the squared correlation coefficient (R^2) between the transcriptional mRNA expression level and the protein abundance was only ~ 0.4 , implying about 40% of the variations in protein abundance can be explained by mRNA expression, and the remaining 60% would require more post-transcriptional measurements⁸². Therefore, the tissue-protein associations derived based on the protein-level evidence should be more reliable and more real than the ones derived from the transcriptional gene expression data, especially when constructing the tissue-specific protein-protein interaction networks.

To investigate such biological networks, a relevant network database would offer an advantage in starting and expediting the study, however there are very limited resources (particularly TissueNet⁹¹, SPECTRA⁹², and IID⁹³) providing the tissue-specific PPI networks, which have been introduced in detail previously in **Section 1.2.2**. These databases have made the groundbreaking contributions, however there are still some major drawbacks. Briefly, 1) prior knowledge is required to search these databases, as they only accept the input of protein name or gene name; 2) the output is only a list of PPIs that contain the queried gene or protein, not a PPI network; 3) some databases have no

download option, or have no human tissue information; 4) the downloaded table (from IID and SPECTRA) is not compatible with any network analysis software, particularly Cytoscape; and 5) none of the databases provide the network properties or descriptors for quantitative analysis.

Therefore, TISPIN database (TIssue-Specific Protein Interaction Networks) was constructed for delivering the following information: 1) network files in various formats (SIF, XML, and CYJS) that are compatible with the major network software; 2) network visualization (the global network, and the largest connected network); 3) computed network descriptors in node-level (local properties for each protein) and network-level (global properties for the entire network); 4) protein annotations in terms of protein name, gene symbol, UniProt ID/ACC, NCBI protein reference ID, biological process / cellular component / molecular function, and therapeutic targets; and 5) comprehensive download links for all the information in TISPIN. In the current stage of developing TISPIN prototype 1.0, we mainly focus on building up the database interface and architecture, and the data source is primarily collected from HPRD, which have been evaluated as a reliable source of protein-protein interactions, as it was manually curated by expert biologist to reduce the errors^{125,126}. HPRD provides not only the human PPIs, but also the human tissue-protein associations, which were based on the literatures of tissue distributions of the expressed proteins⁷⁵. Unlike the other databases (TissueNet, SPECTRA, and IID) that use the microarray and/or RNAseq data to infer the tissue-protein associations in transcription-level, the tissue distribution in TISPIN is on the basis of the protein-level evidence. The detailed comparison between TISPIN 1.0 and the other relevant databases was presented in **Table 3-1**.

Table 3-1 Comparison between TISPIN 1.0 and the other relevant databases (TissueNet, SPECTRA, and IID) that provide tissue-specific PPI networks

| Database | TISPIN 1.0 | TissueNet | SPECTRA | IID |
|-------------------------------------|--|---|---|--|
| Year | 2016 | 2012 | 2015 | 2016 |
| Species | human | human | human | yeast, worm, fly, rat, mouse, human |
| # Tissues | 87 | 16 | 107 | 30 |
| # Proteins | 9,616 | 11,225 | 16,435 | 68,831 |
| # PPIs | 39,240 | 67,439 | 175,841 | 1,566,043 |
| PPI Type | experimental | experimental | experimental | experimental & predicted |
| PPI Source | HPRD | BioGRID, DIP, IntAct, MINT | BioGRID, DIP, HPRD, IntAct, MINT | BioGRID, DIP, HPRD, IntAct, InnateDB, MINT |
| Tissue Source | protein expression | microarray & sequencing | microarray & sequencing | microarray |
| | protein level | transcription level | transcription level | transcription level |
| Network Files | ✓ | ✗ | ✗ | ✗ |
| Network Visualizations | ✓ | ✓ | ✓ | ✗ |
| Network Properties | ✓ | ✗ | ✗ | ✗ |
| Download Option | ✓ | ✗ | ✓ | ✓ |
| Compatibility with Network Software | ✓ | ✗ | ✗ | ✗ |
| Query Method | quick search, search by system, search by tissue | provide gene name, and select tissue name | select gene data, select tissue, select expression data, and select interaction data | provide gene IDs, select species, and select tissues |
| Output Method | a new page with visualizations, properties, annotations, and download links | a network map | a table of PPIs, and a network visualization | a table of PPIs |
| Other Comments | This prototype is fully functional with outstanding features, but limited by its relatively small data source. Further improvement will be made. | “Service Error” occurred frequently when querying the database. | There was no trouble met in querying this database, except its relatively slow responses. | Tissue information was not found anywhere in the result. |

3.2 Materials and Methods

3.2.1 Data Source

TISPIN database prototype 1.0 (Tissue-Specific Protein Interaction Network) was primarily derived from the data source in HPRD database^{86,165,166}, which was developed for providing the curated proteomic information pertaining to human proteins. Initially, the raw data included 39,240 protein-protein interactions among 9,616 human proteins and 112,158 protein-tissue associations. By removing the entries with special characters (e.g. '-', '\$', '@') and the duplicates, we obtained 38,131 protein-protein interactions among 9,084 human proteins and 111,152 tissue-protein associations.

Each interacting protein pair was checked against the tissue-protein association information, to find whether these two proteins exist in the same tissue or cell. If yes, this protein pair was then added to the tissue-specific PPI list. After scanning through all the protein pairs, each tissue had a group of PPIs belonging to itself. We called this group of PPIs as the human tissue-specific global protein interaction network, and its largest connected interaction sub-network was also extracted. In HPRD database, the tissue distribution of proteins was collected from searching of literature databases⁷⁵, based on the protein-expression evidences.

Finally, we generated 87 human tissue/cell-specific protein interaction networks, which were categorized into 11 different human systems (**Figure 3-1**), including: cardiovascular system (n=8), digestive system (n=10), endocrine system (n=4), excretory system (n=4), immune system (n=15), integumentary system (n=6),

musculoskeletal system (n=7), nervous system (n=21), reproductive system (n=9), respiratory system (n=2), and fetus (n=1). The names of all these human tissues and cells were given in **Table 3-2**.

Figure 3-1 Distribution of TISPIN covered tissues/cells in human systems

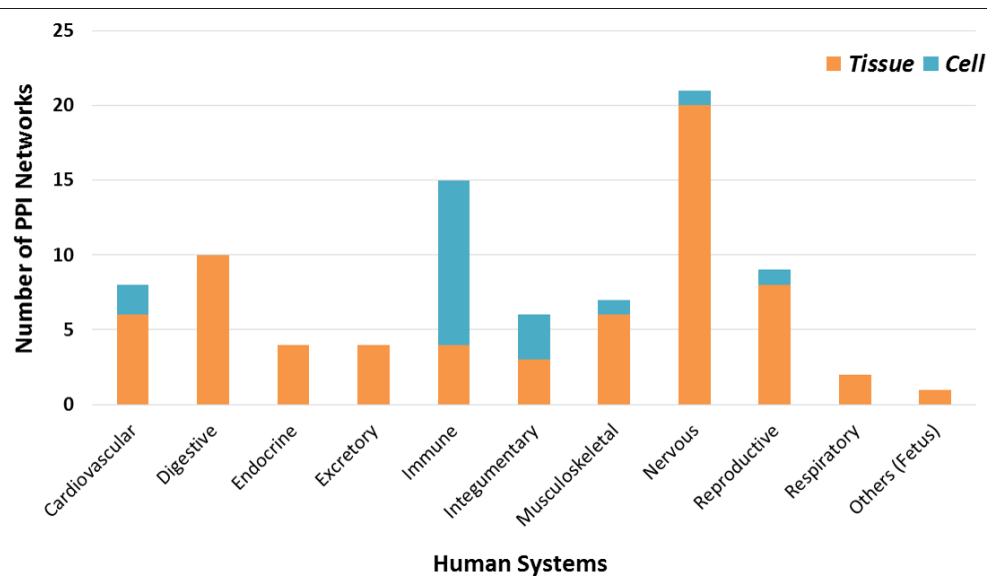


Table 3-2 Human tissues/cells covered in TISPIN database

| Human Systems | Tissue | Cells |
|----------------------|---|--|
| Cardiovascular | aorta, blood, blood vessel, heart, plasma, serum | platelet, red blood cell |
| Digestive | colon, duodenum, intestine, islets of Langerhans, liver, saliva, salivary gland, small intestine, stomach, vermiform appendix | - |
| Endocrine | adrenal gland, pancreas, pituitary gland, thyroid gland | - |
| Excretory | kidney, lacrimal gland, tear, urinary bladder | - |
| Immune | bone marrow, lymph node, spleen, tonsil | B cell, eosinophil, granulocyte, hematopoietic stem cell, leukocyte, lymphocyte, macrophage, monocyte, natural killer cell, neutrophil, T cell |
| Integumentary | epidermis, mammary epithelium, skin | keratinocyte, skin fibroblast, umbilical vein endothelial cell |
| Musculoskeletal | adipose tissue, bone, cartilage, muscle, skeletal muscle, smooth muscle | chondrocyte |
| Nervous | amygdala, brain, caudate nucleus, cerebellum, cerebral cortex, cornea, corpus callosum, eye, frontal cortex, frontal lobe, hippocampus, medulla oblongata, nervous system, putamen, retina, spinal cord, substantia nigra, subthalamic nucleus, temporal lobe, thalamus | dendritic cell |
| Reproductive | endometrium, mammary gland, ovary, placenta, prostate, semen, testis, uterus | spermatozoa |
| Respiratory | lung, trachea | - |
| Others | fetus | - |

3.2.2 Generation of Network Information

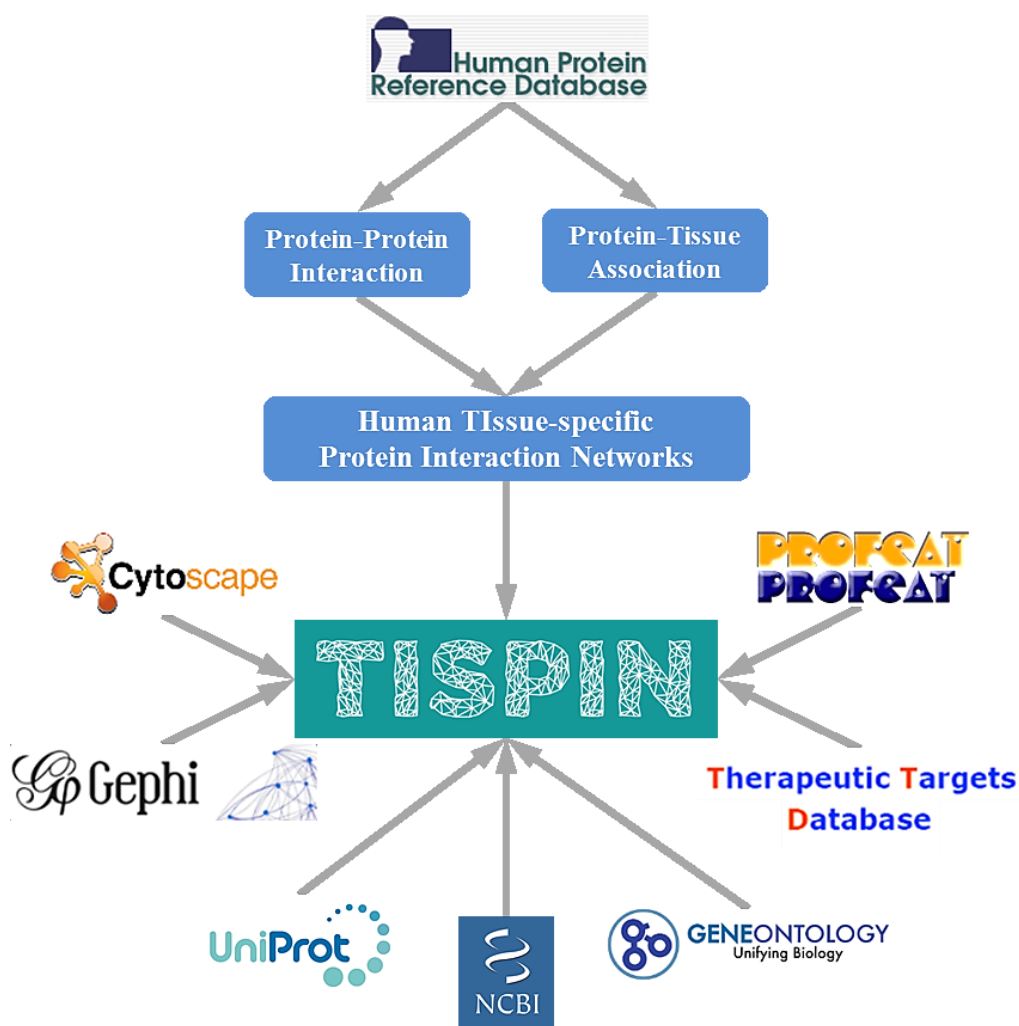
As described previously, we obtained a set of all PPIs belonging to each human tissue or cell, which was considered as the global PPI network in that particular tissue or cell. We had also coded the additional program to count the disconnected subnetworks that have more than 2 interacting proteins, and extract the largest connected PPI network for each tissue or cell. In TISPIN database, a board spectrum of network information have been generated (network files, network visualization, network descriptors, and protein annotations) for each PPI network, and the informative tables were stored by MySQL in the BIDD server. **Figure 3-2** showed the schematic of the data sources and the incorporated information in TISPIN database 1.0.

For the entire tissue-specific protein interaction networks, TISPIN database provides: 1) networks files in SIF format (simple interaction file), XML format (schema-based network format), and CYJS format (Cytoscape-defined network format)⁴¹; 2) network visualization in PNG and SVG format (scalable vector graphics), generated in the Force-Directed Layout by Cytoscape software⁴¹ (an example was given in **Figure 3-3 A**); 3) protein list in the network by giving the gene name, Uniprot ID, Uniprot ACC, NCBI protein reference ID, full protein name, and a Boolean value for each protein if it is a successful therapeutic target by checking against Therapeutic Target Database^{176,177}; 4) annotated protein list in terms of the biological process, cellular component, and molecular function¹⁷⁸.

For the largest connected protein interaction network for each tissue, TISPIN delivers: (1) networks files in SIF format; (2) network visualization in PNG and SVG format, with the varying node color representing the proteins' degrees

(number of its interaction partners) and the varying node size representing the proteins' betweenness centralities (a measure of its capability in bridging important modules in the network), generated in the Force Atlas Layout by Gephi software⁴³ (an example was given in **Figure 3-3 B**); (3) network descriptors in network-level and node-level, computed by PROFEAT webserver.

Figure 3-2 Schematic of the data sources and the incorporated information in TISPIN



3.3 Results

3.3.1 TISPIN Database Structure and Access

TISPIN (Tissue-Specific Protein Interaction Network) database prototype 1.0 was accessible at (<http://bidd2.nus.edu.sg/TISPIN/home.php>), enabling the users to search, visualize, analyze, and download the global and the largest connected PPI networks for 87 different human tissues and cells. TISPIN home page was shown in **Figure 3-4**, and there were 4 tabs in the top menu bar: “HOME” for the homepage, “OVERVIEW” for the brief introduction of the database; “BROWSE NETWORKS” for all the 87 tissue-specific protein interaction networks ordered by alphabet in default; “STATISTICS” for the network distribution in human systems and the names of all the tissues and cells (**Figure 3-1** and **Table 3-2**); “DOWNLOAD” for the bulky zipped files for all protein interaction network data.

In the centre of TISPIN home page, there were 3 searching moods given: “Quick Search” for searching any keyword typed in; “Search by System” for choosing one or more human systems in the check box from a drop down menu; “Search by Tissue” for choosing one or more tissues in the check box from a drop down menu. **Figure 3-5** was the search result page by a quick search for “Immune”. In the search result page, networks were ordered by tissue names alphabetically in the first column, the following columns were human systems, type (tissue or cell), number of proteins, number of interactions, number of subnetworks that having more than 2 interacting proteins (the lone pair of PPIs were excluded), and the thumbnail for the entire protein interaction networks.

Figure 3-4 Home page of TISPIN database


TISPIN HOME OVERVIEW BROWSE NETWORKS STATISTICS DOWNLOAD

TISPIN: Tissue-Specific Protein Interaction Networks

TISPIN

A database providing the network files, visualization, descriptors and protein information for the human tissue-specific protein interaction networks

Quick Search Search by System Search by Tissue

Keyword (e.g. Kidney, Pancreas, T Cell, Cardiovascular, Nervous) 

ABOUT US

BiInformatics & Drug Design Group (BIDD) is a research group led by Professor Chen Yu Zong, based in Department of Pharmacy, National University of Singapore.

Our research interests mainly focus on bio-chemo-informatics, computational biology, and computational drug discovery.

CONTACT

Prof. Chen Yu Zong
Email: phacyz@nus.edu.sg

Mr. Zhang Peng
Email: zhangpeng@u.nus.edu

LOCATION

#05-03, Block MD1,
12 Science Drive 2,
Singapore 117549

LAST UPDATE: 03 July 2016

Figure 3-5 Search result page by an example quick search for “Immune”
















TISPIN

HOMEOVERVIEWBROWSE NETWORKSSTATISTICSDOWNLOAD

SEARCH RESULTS » IMMUNE

Show20entries

Search:

| Tissue | Human System | Type | # Proteins | # Interactions | # Subnetworks | Network Image |
|-------------------------|--------------|--------|------------|----------------|---------------|---|
| B Cell | Immune | Cell | 164 | 323 | 4 |  |
| Bone marrow | Immune | Tissue | 324 | 472 | 7 |  |
| Eosinophil | Immune | Cell | 40 | 51 | 3 |  |
| Granulocyte | Immune | Cell | 28 | 31 | 2 |  |
| Hematopoietic stem cell | Immune | Cell | 119 | 232 | 2 |  |
| Leukocyte | Immune | Cell | 1578 | 3956 | 10 |  |
| Lymph node | Immune | Tissue | 219 | 244 | 10 |  |
| Lymphocyte | Immune | Cell | 204 | 285 | 5 |  |
| Macrophage | Immune | Cell | 128 | 183 | 4 |  |
| Monocyte | Immune | Cell | 383 | 677 | 4 |  |
| Natural killer cell | Immune | Cell | 42 | 46 | 4 |  |
| Neutrophil | Immune | Cell | 229 | 382 | 4 |  |
| Spleen | Immune | Tissue | 1540 | 2767 | 18 |  |
| T Cell | Immune | Cell | 227 | 423 | 6 |  |
| Tonsil | Immune | Tissue | 62 | 78 | 5 |  |

Showing 1 to 15 of 15 entries

Previous1Next

Figure 3-6 showed the network detail page for the protein interaction network of T cell. The page began with the basic network information (tissue, human system, type, number of proteins, number of interactions, and number of subnetworks), and followed by the network visualizations for the entire PPI network and the largest connected PPI network in T cell (also in **Figure 3-3**). The PROFEAT-computed network descriptors were provided for its largest subnetwork, including 10 selected network-level descriptors (number of nodes, number of edges, maximum connectivity, network density, network diameter, network radius, average clustering coefficient, characteristic path length, heterogeneity, and global efficiency), and 10 selected node-level descriptors (degree, number of selfloops, number of triangles, clustering coefficient, neighborhood connectivity, topological coefficient, closeness centrality, betweenness centrality, and PageRank centrality) for the 5 highest-degree proteins. The full set of network descriptors could be obtained via the links given in the download section. The therapeutic targets in this network were listed by matching against TTD database. Finally, the comprehensive download links were provided for users to get the information on the human T cell-specific protein interaction network.

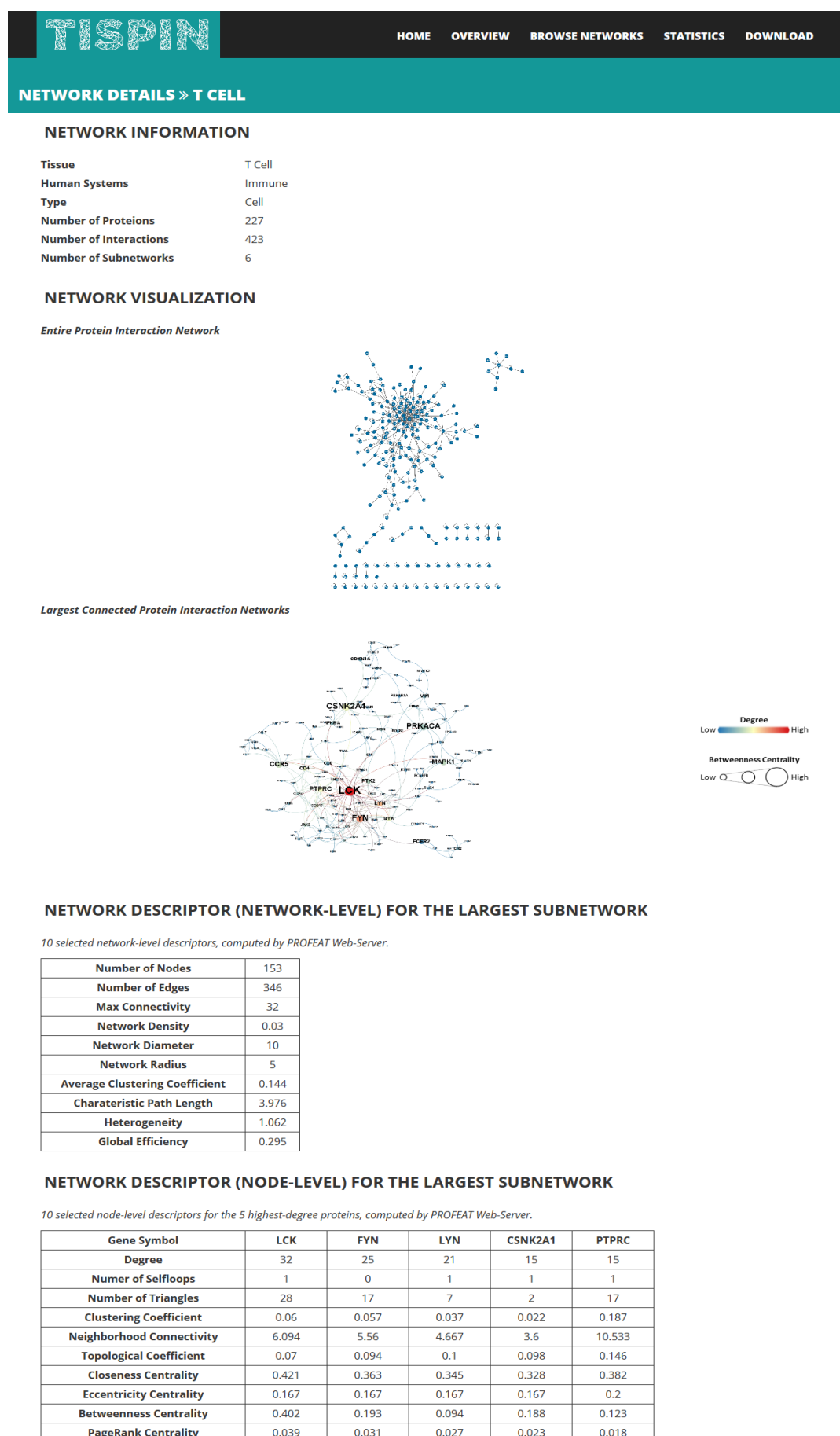
Figure 3-6 Network detail page for example “T Cell”










Figure 3-6 (continued) Network detail page for example “T Cell”**THERAPEUTIC TARGETS**

Matched against Therapeutic Target Database (TTD).






Gene Symbol: ADORA2A , CCR5 , CD28 , CD86 , CSF3R , CTLA4 , CXCR4 , FYN , IL1R1 , IL2RA , ITGAL , JAK3 , LCK , MS4A1 , MTOR , PDE4A ,

DOWNLOAD LINKS

Entire Protein Interaction Networks for T Cell-specific PPI network.

| File Description | Format | Download |
|---|--------|---|
| Network Files | | |
| Entire tissue-specific PPI network in Simple interaction format | SIF |  |
| Entire tissue-specific PPI network in Schema-based network format | XML |  |
| Entire tissue-specific PPI network in Cytoscape network format | CYJS |  |
| Network Visualization | | |
| Image of the entire tissue-specific PPI network | PNG |  |
| Scalable vector graphics of the entire tissue-specific PPI network | SVG |  |
| Protein Information in Network | | |
| List of gene name, UniProt ID/ACC, NCBI ID, protein name, Boolean value for therapeutic target, and GO annotations (Biological Process / Cellular Component / Molecular Function) | TXT |  |
| List of proteins under GO Biological Process Annotation | TXT |  |
| List of proteins under GO Cellular Component Annotation | TXT |  |
| List of proteins under GO Molecular Function Annotation | TXT |  |

Largest Connected Protein Interaction Networks for T Cell-specific PPI network.

| File Description | Format | Download |
|---|--------|---|
| Network Files | | |
| Largest tissue-specific PPI sub-network in Simple interaction format | SIF |  |
| Network Visualization | | |
| Image of the largest tissue-specific PPI sub-network | PNG |  |
| Scalable vector graphics of the largest tissue-specific PPI sub-network | SVG |  |
| Network Descriptors | | |
| Network-level descriptors for characterizing the PPI network globally | TXT |  |
| Node-level descriptors for characterizing each protein locally | TXT |  |

Currently, TISPIN database prototype 1.0 has been fully functional, with the well-designed interface and architecture. So far, TISPIN has demonstrated its significant advantages over the other databases, while there is still big room for improvement, especially in enlarging the PPI data size from more relevant databases, and integrating more tissue-protein associations based on protein-expression evidence from more reliable resources. The detailed perspectives will be discussed later in **Section 5.2**.

CHAPTER 4 Quantitative Sequence-Kinetic Constants Relationship for Predicting Protein-Protein Interaction Kinetic Constants

4.1 Background and Motivations

In molecular and systems biological studies, protein-protein interactions play key roles in cellular signalling processes, and kinetic information in protein complex is even more important in understanding the systematic molecular interactions and the biochemical events¹¹³. Elucidating the kinetic information of the interacting protein complex will reveal the mechanism in the protein-protein bindings, facilitate the simulation of the dynamics of biological pathways, and promote the systems biology investigations¹²².

In medicinal chemistry, the exploration of new therapeutic agent is not only limited by the chemical synthesis imaginations and the natural product resources, but also limited by the understanding of protein-substrate interaction kinetics, which provides the critical information in potent and effective drug design^{98,99}. This is because the drug potency is highly depended on the competitive advantages over the substrates of drug targets, but the researchers are sometimes less clear about the potency needed to ensure the competitiveness of the drug against the target substrate. In other words, for such a competitive drug binding, the affinity of the protein-drug interaction on its own offers no proof to the effective inhibition outcome. Rather, the protein-drug binding affinity becomes applicable only if being stronger than the affinity of the wide-type protein partner, which the drug is competing against.

As small molecule inhibitors for protein-protein interactions have been increasingly identified as potential therapeutic agents (e.g. drug discovery for inhibiting *p53-MDM2* interaction^{100,102}), it is therefore of significant interest and importance to know the binding affinity between a pair of interacting proteins¹⁷⁹. An accurate prediction of PPI kinetic constants will greatly elucidate the quantitative relationship between the protein information and their equilibrium constant (K_d) / association rate constant (k_{on}) / dissociation rate constant (k_{off}), and such that enable the further facilitation of the basic research on studying the protein-protein interaction kinetics and the applied research on discovering the potent and competitive protein-protein interaction inhibitors.

As previously introduced in **Section 1.3.3**, some computational approaches have been attempted to predict the PPI kinetic constants. Particularly, *Bai et al.* built the linear models to predict kinetic constants of 62 PPIs by using 37 structure-based properties, and gave the performance R^2 at 0.801, 0.732 and 0.770 for k_{off} , k_{on} and K_d data respectively in leave-one-out cross-validation¹¹³. *Iain et al.* predicted the protein-protein binding affinity on 137 protein complexes with known PDB structures¹¹⁵, by using 200 descriptors calculated from various protein structure-based software¹¹⁴. *Iain* applied four machine learning algorithms, and achieved R^2 between 0.69 and 0.75 in leave-one-out cross-validation. *Ma et al.* conducted the PPI equilibrium constant (K_d) prediction of 133 protein complexes, by 432 physiochemical and structural features. *Ma* built the regression model by random forest algorithm, and obtained R^2 at 0.708 in leave-one-out cross-validation¹²³.

In the current process of discovering PPI inhibitors, the main efforts have been suggested to extend the prediction programs for modeling the PPI binding affinity, and to expand the libraries of PPIs with known wide-type PPI kinetics information¹⁸⁰. The previously reviewed studies on PPI kinetics prediction have provided the valuable experience and the significant findings. However, the PPI kinetics datasets were not diverse enough, where at most 137 PPIs were used for building the prediction models. Moreover, the protein 3D structures for the interacting protein complexes were highly relied on for calculating the PPI features, while the fact is that some proteins do not have PDB structures available yet. Therefore, a larger PPI kinetics dataset is needed for representing larger protein feature space, and the prediction method by only using the information from amino acid sequences should be attempted, as the sequence-based approach is more universally applicable than those based on the structural and functional information of proteins.

In this proof of concept study of Quantitative Sequence-Kinetic Constants Relationship (QSKR), we expanded the PPI kinetics library to 820 entries, and investigated the applications of support vector regression and random forest algorithms on this highly diverse protein-protein interaction datasets, to predict their kinetic constants (K_d , k_{on} and k_{off}), by solely using the features generated from protein primary sequences, without the protein 3D structures.

4.2 Materials and Methods

4.2.1 Data Collection

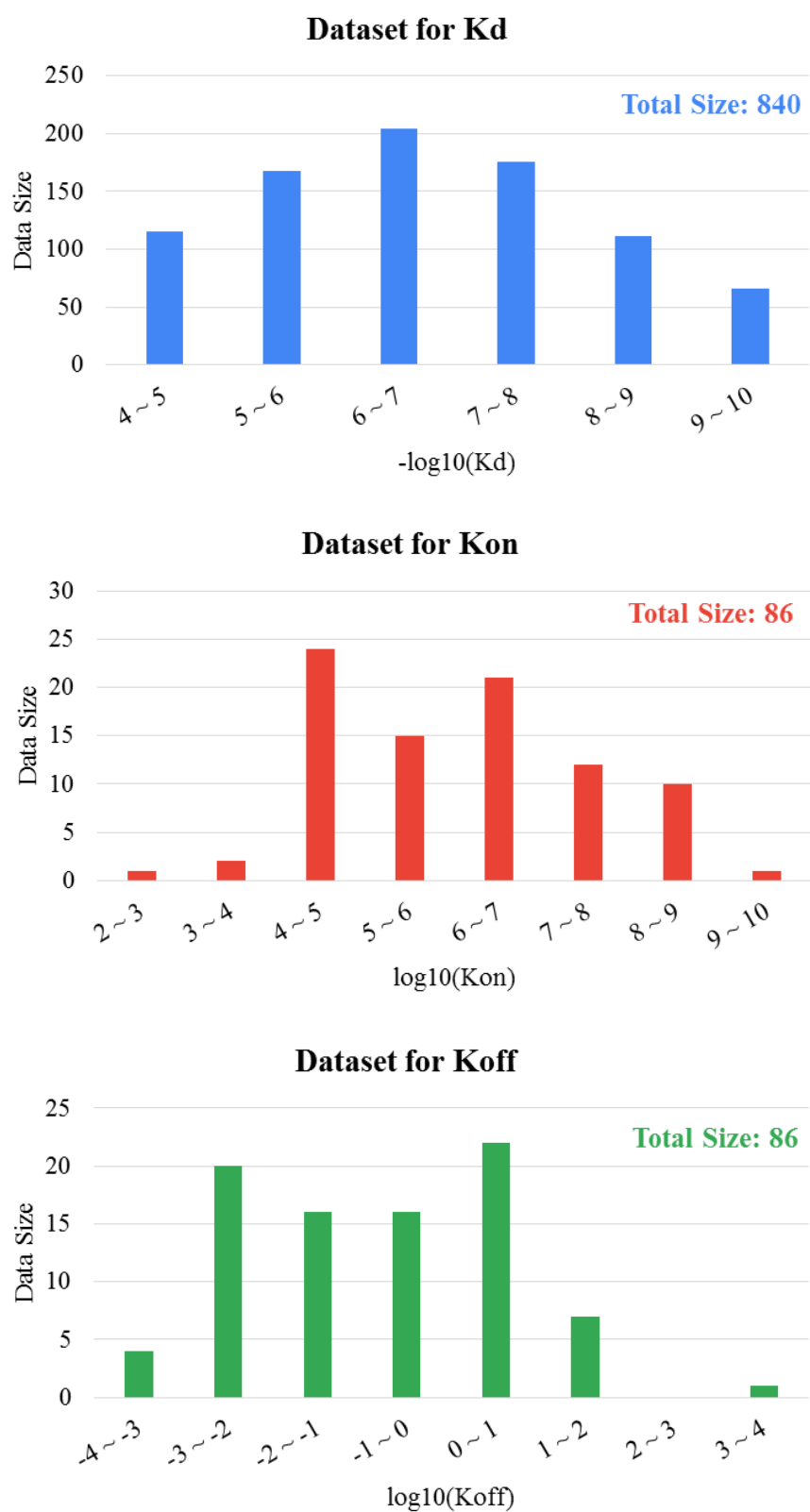
To obtain a non-redundant and highly diverse dataset for protein-protein interaction kinetics, we approached PDBbind database¹⁸¹, KDBI database¹³⁸, a structure-based benchmark for protein-protein binding affinity¹¹⁵, and extensive literature reviews (PMIDs were provided for each collected PPI in **Appendix Section C** and **Section D**). The inclusion and exclusion criteria were applied in the data collection: 1) it should have at least one kinetic constant information (K_d , k_{on} and k_{off}) for the protein complex; 2) it should not have more than 2 interacting proteins in the complex; 3) protein-peptide complex was not included; 4) small ligands, DNAs, RNAs should not be involved in the interaction; 5) for the same pair of interacting proteins, the kinetic data published later was adopted. In addition, if there were any two kinetic constants collected for one specific PPI, then the third kinetic constant would be calculated by $K_d = k_{off} / k_{on}$.

Therefore, a diverse and comprehensive PPI dataset with kinetics information was collected, and then further refined by removing the very few PPI entries having $K_d > 10^{-4}M$ (100 μM , very weak binding affinity), because it is meaningless to find drugs even weaker; and the very few PPIs entries having $K_d < 10^{-10}M$ (0.1 nM, very strong binding affinity), although there are many drugs at picomolar affinities, there are very few PPIs having picomolar affinities. In addition, due to the small number of PPIs having $K_d > 10^{-4}M$ or $K_d < 10^{-10}M$, we have severely insufficient representations in the protein feature space for the very strong and the very weak binding affinity. The insufficient PPI

representations would make the developed models statistically unreliable and would adversely affect the prediction power in the major domain of PPIs which have their kinetic constants $10^{-10}\text{M} < K_d < 10^{-4}\text{M}$.

Hence, we were more interested in the PPIs having equilibrium dissociation constant $-\log(K_d)$ between 4 and 10. The final refined PPI dataset included 840 protein complexes with K_d data, and 86 protein complexes with k_{on} and k_{off} data. All these protein complexes were searched against Uniprot database to obtain their sequences, and Protein Data Bank database to check if there exist any protein 3D structures for the interacting protein complexes or the single proteins. Among these collected PPIs, there were 50 protein complexes having no PDB structures available yet. The detailed information of this PPI library was provided in **Appendix Section C** and **Section D**.

In this PPI library, there were 1,283 unique proteins, and their kinetic constants vary in magnitude largely, corresponding to a vast Gibbs free energy difference^{104,108}. $-\log_{10}(K_d)$ value ranged from 4 to 10, $\log_{10}(k_{on})$ value ranged from 2 to 9, and $\log_{10}(k_{off})$ value ranged from -4 to 4. The distribution of kinetic constants value in each dataset were illustrated in **Figure 4-1**.

Figure 4-1 Distribution of kinetic constants value in dataset (K_d , k_{on} , k_{off})

4.2.2 Calculation of Protein-Protein Interaction Features

Protein features were calculated by PROFEAT, a web server developed for computing the structural and physiochemical features of proteins from the primary sequences^{127,182} (<http://bidd2.nus.edu.sg/cgi-bin/profeat2016/main.cgi>). By giving the amino acid sequence of a single protein, PROFEAT enabled the computation of 356 protein features from 17 feature categories (**Table 4-1**), including: 1) amino acid composition; 2) hydrophobicity; 3) Van Der Waals volume; 4) polarity; 5) polarizability; 6) charge; 7) secondary structure information; 8) solvent accessibility; 9) surface tension; 10) molecular weight; 11) solubility in water; 12) number of hydrogen bond donor in the side chain; 13) number of hydrogen bond acceptor in the side chain; 14) ClogP; 15) amino acid flexibility index; 16) Bogen's protein-protein interface hotspot propensity; and 17) Ma's protein-protein interface propensity¹²⁷. It is noted that PROFEAT calculated the protein features based on the input amino acid sequence, but it was limited by not taking account of the physiochemical property changes in the dynamic environment (e.g. the change of charge in the phosphorylation).

These protein features described the informative patterns of composition, transition, and distribution based on the amino acid sequences. So far, the PROFEAT-generated protein features have been successfully applied to answer some biological questions, by receiving >260 citations in the last 10 years. Particularly, PROFEAT has been used for predicting protein folding and structural classes, functional classes, and subcellular locations, with accuracy at 72-95%, 83-97%, and 79-91% respectively^{183,184,185}.

Table 4-1 PROFEAT protein feature categories, descriptions, and dimensions

| Category | Feature Description | Dimension |
|--------------|--|------------|
| 1 | Amino acid composition | 20 |
| 2 | Hydrophobicity | 21 |
| 3 | Van Der Waals volumes | 21 |
| 4 | Polarity | 21 |
| 5 | Polarizability | 21 |
| 6 | Charge | 21 |
| 7 | Secondary structure | 21 |
| 8 | Solvent accessibility | 21 |
| 9 | Surface tension | 21 |
| 10 | Molecular weight | 21 |
| 11 | Solubility in water | 21 |
| 12 | No. of hydrogen bond donor in the side chain | 21 |
| 13 | No. of hydrogen bond acceptor in the side chain | 21 |
| 14 | CLogP | 21 |
| 15 | Amino acid flexibility index | 21 |
| 16 | Bogan's Protein-protein interface hotspot propensity | 21 |
| 17 | Ma's Protein-protein interface propensity | 21 |
| Total | | 356 |

All of the 1,283 proteins in the dataset were carefully identified and curated. Their sequences were collected from Uniprot database, and then submitted to PROFEAT webserver for computing their protein features. For an interacting protein complex (protein A and protein B), we concatenated their protein features to create the PPI features for the protein complex (A·B), where this vectorization method has been successfully applied to represent PPIs^{128,186}. Therefore, $V_{AB}=\{V_A(i)\oplus V_B(i), i=1\dots356\}$ and $V_{BA}=\{V_B(i)\oplus V_A(i), i=1\dots356\}$ were used to get a set of 1680 PPI feature vectors with 712-dimensional features for the K_d dataset, and two sets of 172 PPI feature vectors with 712 dimensional-features for the k_{on} and k_{off} datasets.

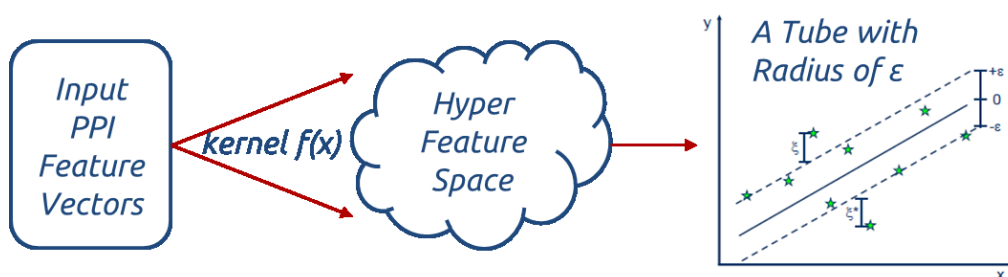
4.2.3 Machine Learning Method (Support Vector Regression)

To build the QSKR models, ε -SVR (support vector regression) algorithm was applied, which is a supervised statistical regression algorithm based on the classification algorithm SVM (support vector machine)^{187,188}. SVM defines a mapping or a kernel (e.g. polynomial, Gaussian radial basis function) from the input feature vector space to the class label space. To achieve this task, the input data is projected into a much higher dimensional feature space through the kernel function, and then a hyperplane or a set of hyperplanes is/are constructed in this high dimensional space to classify the input data by finding a maximum margin, which represents the largest separation between the two classes¹⁸⁹.

In ε -SVR modeling, suppose that the given training data is $[(x_1, y_1), (x_2, y_2) \dots (x_n, y_n)] \subset \chi$, where x_i represents the PPI feature vector, y_i is the logarithmic PPI kinetic constants $[-\log_{10}(K_d), \log_{10}(k_{on}), \text{ or } \log_{10}(k_{off})]$, n is the number of the vectors, and χ denotes the modeling PPI feature space. The goal of ε -SVR is to find a function $f(x)$ that has at most ε deviation from the actual value y_i for all the training data. In other words, ε -SVR constructs a “tube” with the radius of ε to involve as many training points as possible (**Figure 4-2**).

In this study, libSVM¹⁹⁰ was adopted to build the ε -SVR QSKR model, and Gaussian radial basis kernel function (RBF) (shown below) was selected to project the initial PPI feature vectors into a highly dimensional feature space, as RBF has been extensively and consistently showing better performance than the other kernel functions^{191,192,193}.

$$K(x_i, x_j) = e^{-|x_i - x_j|^2 / 2\sigma^2}.$$

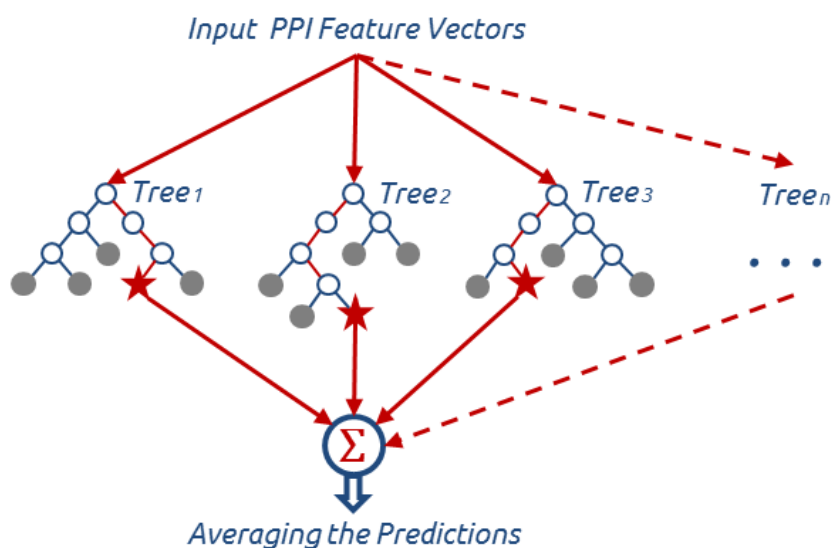
Figure 4-2 Schematic of support vector regression

To optimize the ϵ -SVR model, three parameters should be considered, which are *cost* (penalty factor), ϵ (noise tolerance), and *gamma*. Parameter *cost* determines the trade-off between the flatness of the kernel function and the tolerance for deviations larger than ϵ , and *epsilon* (ϵ) is proportional the noise variance of the dataset. In this study, a hard-margin SVR was used by constantly setting parameter *cost* to 10,000, as the soft-margin SVR allows too many errors in fitting the model. Parameters *gamma* and *epsilon* were then fine-tuned in a grid-screening, by setting *epsilon* from 0.1 to 10 with step-size of 0.1, and *gamma* from 1 to 100 with step-size of 0.5. Therefore, a total of 20,000 combinations of SVR parameters were evaluated to optimize each QSKR model.

4.2.4 Machine Learning Method (Random Forests)

Random Forests (RF) is basically an ensemble of decision tree predictors that each tree depends on the random vector values sampled independently¹⁹⁴. When the number of trees in the forest grows large enough, the generalization error of RF converges. As a non-linear machine learning algorithm, Random Forests is applicable for both classification (by majority voting based on the predicted classes from all the trees) and regression (by averaging the predicted values from all the trees) (Figure 4-3). Each decision tree is grown by a bootstrap sampling of the training dataset. Each node is split by a subset of features randomly chosen at that node, where the best node split is selected based on the criterion to minimize the variance within the branches. Each leaf in each decision tree is an entry in the training data¹⁹⁵.

Figure 4-3 Schematic of regressive random forests



In this study, a Matlab tool of regressive Random Forests by *Breiman*¹⁹⁴ was applied. Primarily, there are two key parameters *ntree* (number of trees to grow) and *mtry* (number of features randomly sampled at each node). Published

advices on parameter optimization were followed, which suggested to set *mtry* at approximately 1/6, 1/3 and 2/3 of the feature size¹⁹⁵, such that *mtry* will be 118, 237, and 474 in my case, as each PPI has a vector of 712-dimensional features. Random Forests has been shown to possess a broad range of optimal values, and also demonstrated the robustness of prediction performances against parameter changes, which concluded that fine-tuning was not necessary in optimizing the Random Forests. Therefore, instead of a large-scale parameter searching, a number of selected different parameter combinations were evaluated, by setting *ntree* to {100, 1000, 5000, 10000} and *mtry* to {100 (\approx 118), 250 (\approx 237), 500 (\approx 474)}. In this Matlab tool, there are some extra Boolean parameters (e.g. *importance*, *locallmp*, *proximity*, *oob_prox*, *keep_inbag*, *corr_bias*, etc.), which were kept as default in this study.

4.2.5 Performance Evaluation

As this study is to investigate the protein sequence-based modeling and prediction of PPI kinetic constants, and there were 50 PPIs that have no available 3D structures yet in the K_d dataset. Therefore, the K_d dataset, containing 840 PPIs in total, was split into a set of the non-3D-structural 50 PPIs for external validation, and a set of the remaining 790 PPIs for internal training purpose. The 10-fold internal cross-validation was applied to train and optimize the QSKR model for predicting the kinetic equilibrium constant (K_d). To carry out the 10-fold cross-validation, 790 PPIs were randomly and exclusively split into 10 pieces of equal-sized sub-datasets. And then, each of the 10 sub-datasets was selected as the testing dataset, while the rest 9 sub-datasets were merged as the training datasets to build the model.

On the other hand, due to the small-size (86 PPIs) of k_{on} and k_{off} datasets, their QSKR models were trained and optimized by leave-one-out cross-validation, and no external validation was applied. In leave-one-out cross-validation (LOOCV), one single PPI was excluded each time for the testing purpose, while the model was built based on the remaining 85 PPIs.

To evaluate the regressive prediction performance, the squared Pearson correlation coefficient (R^2), which implies the level of explained variability in the statistical model, was used. R^2 ranges between 0 and 1, where the closer to 1 indicating the higher degree of fitness between the prediction and the observation. Conventionally, the squared Pearson correlation coefficient in fitting the training data in cross-validation is notated by R^2 , and that in predicting the testing data in cross-validation is notated by Q^2 . The optimized combination of parameters was chosen when both of R^2 and Q^2 mutually achieve the relatively maximum. Additionally, the Root Mean Squared Error ($RMSE$) was also used for evaluating the performance, which measures the difference between the estimated value and the actual value.

Squared Pearson Correlation Coefficient (R^2):

$$R^2 = 1 - \frac{SS_{err}}{SS_{tot}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}.$$

Root Mean Squared Error ($RMSE$):

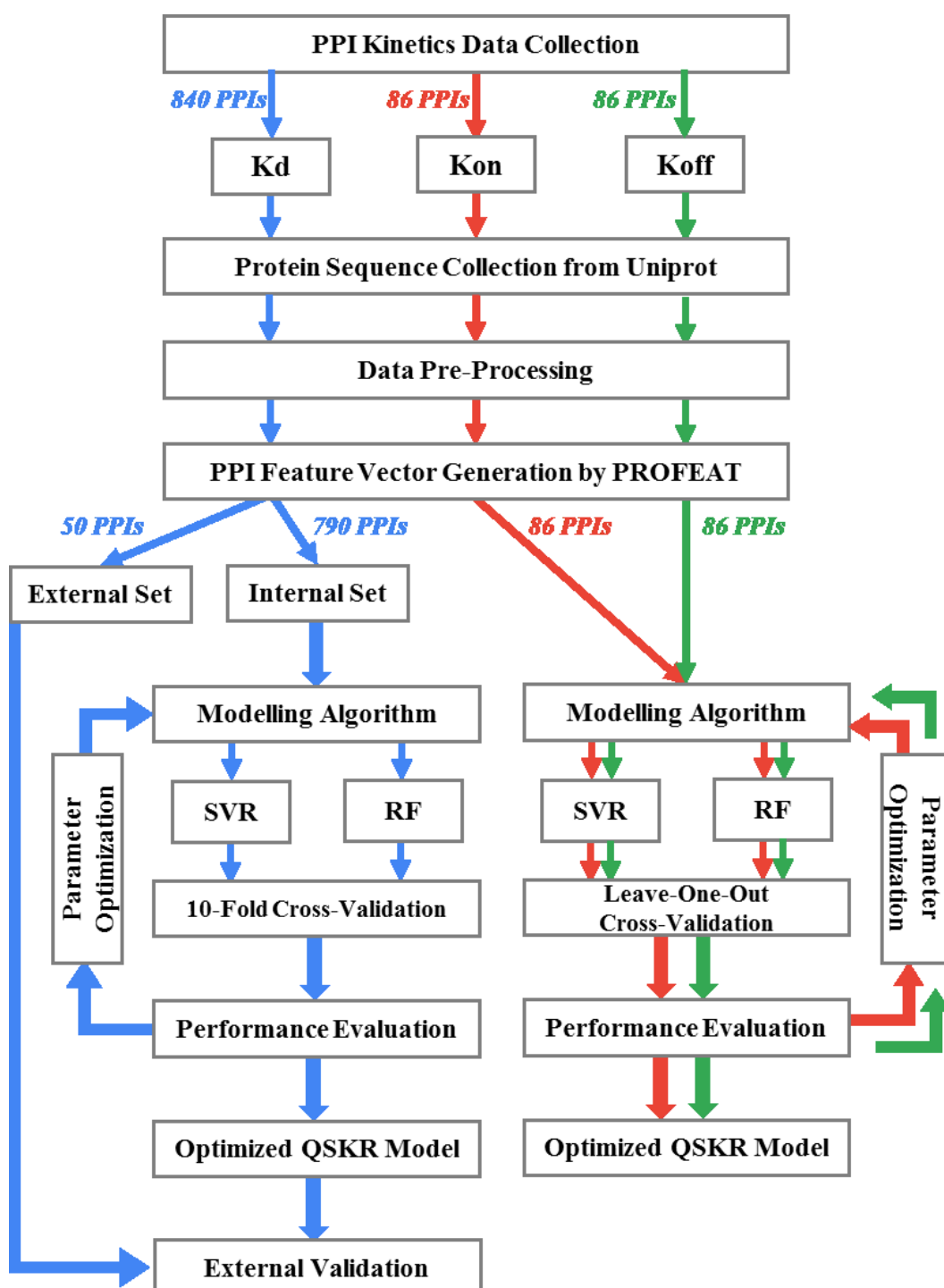
$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2}.$$

Where y_i is each actual value, \bar{y} is the mean of the actual value, and \hat{y}_i denotes each predicted value.

4.2.6 Workflow of QSKR Study

The schematic QSKR workflow to predict the PPI kinetic constants was illustrated in **Figure 4-4**, where the blue-, red- and green-colored directed lines represented the workflows for K_d , k_{on} and k_{off} datasets respectively.

Figure 4-4 Workflow of QSKR study to predict PPI kinetic constants



4.3 Results

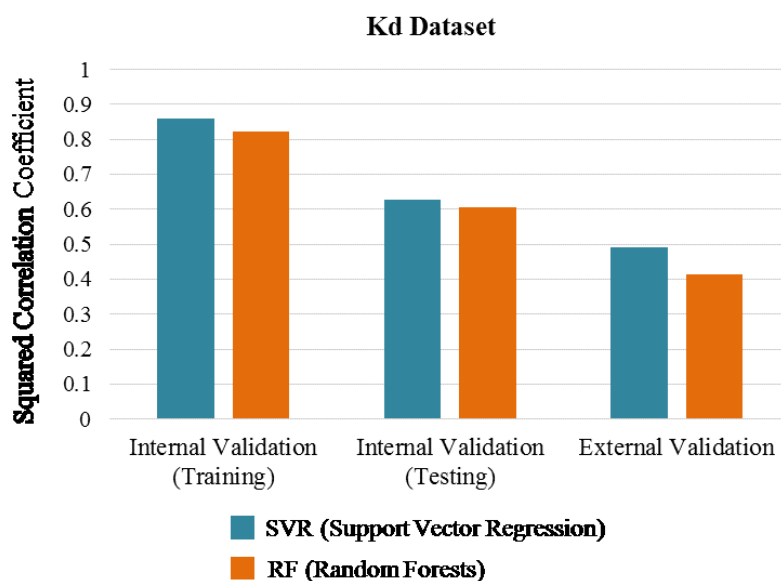
4.3.1 QSKR Prediction Performance on K_d Dataset

In the modeling of protein-protein interactions for predicting their equilibrium dissociation constant (K_d), there were totally 840 protein complexes collected, in which 790 PPIs were split into internal validation set for constructing and optimizing the QSKR model by SVR and RF machine-learning methods, as well as a set of 50 PPIs, that have no PDB structures, for external validation.

From **Table 4-2** and **Figure 4-5**, we observed that the best QSKR model in predicting PPI kinetic constant K_d was constructed by SVR algorithm, slightly outperformed RF algorithm. The best QSKR model achieved ($R^2 = 0.859$, $RMSE_{train} = 0.523$, $Q^2 = 0.628$, $RMSE_{test} = 0.901$) in 10-fold cross-validation, and ($R^2_{external} = 0.491$, $RMSE_{external} = 1.173$) in external validation.

Table 4-2 The best QSKR model performance in predicting the K_d value

| Algorithm | Internal 10-fold Cross-Validation | | | | External Validation | |
|------------|-----------------------------------|----------------|-------|---------------|---------------------|-------------------|
| | R^2 | $RMSE_{train}$ | Q^2 | $RMSE_{test}$ | $R^2_{external}$ | $RMSE_{external}$ |
| SVR | 0.859 | 0.523 | 0.628 | 0.901 | 0.491 | 1.173 |
| RF | 0.823 | 0.597 | 0.607 | 0.973 | 0.415 | 1.381 |

Figure 4-5 The best QSKR model performance in predicting the K_d value

The best QSKR model was obtained through optimizing the parameters in SVR and RF respectively. Below, we presented the SVR parameter optimization to find the best model (**Figure 4-6**), and the performance plot of the best model found (**Figure 4-7**). **Figure 4-6** illustrated the heat plot of prediction results in grid-searching of 20,000 SVR parameter combinations. In this heat plot, each grid represented one combination of *epsilon* and *gamma*, and the hotter color implied the higher R^2 . The left plot (A) was the average R^2 of training dataset, and the right plot (B) was the average R^2 of testing dataset, in internal 10-fold cross-validation. The region with highest performance was circled in the plot, due to the printing may not differentiate the red and the dark orange easily.

Therefore, the best QSKR model by SVR was achieved at $\gamma \approx 88$ and $\epsilon \approx 1.6$. **Figure 4-7** laid out the plot of the predicted K_d value versus the actual K_d value in the best QSKR model. The blue dots represented the internal testing data ($Q^2 = 0.628$, $RMSE_{test} = 0.901$), and the orange dots represented the external validating data ($R^2_{external} = 0.491$, $RMSE_{external} = 1.173$).

Figure 4-6 Heat plot of K_d prediction performance in SVR parameter optimization

(A) the average R^2 of training dataset, and (B) the average Q^2 of testing dataset, in internal 10-fold cross-validation

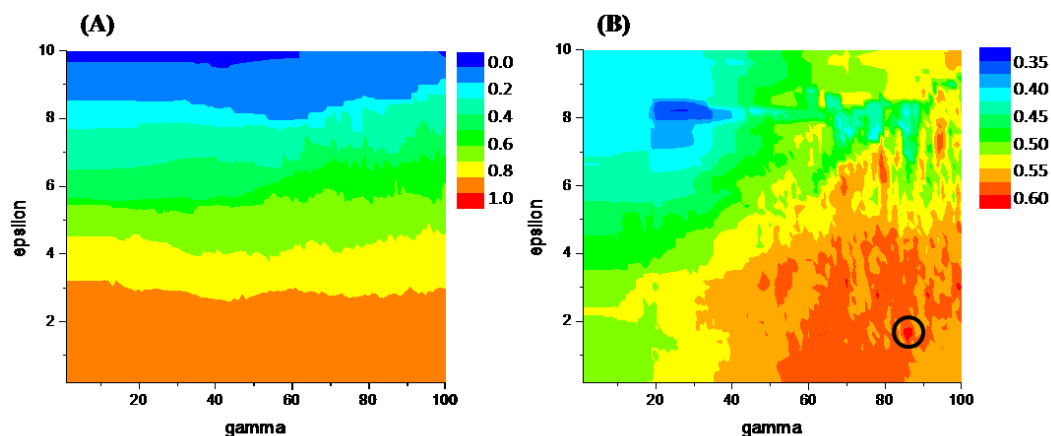
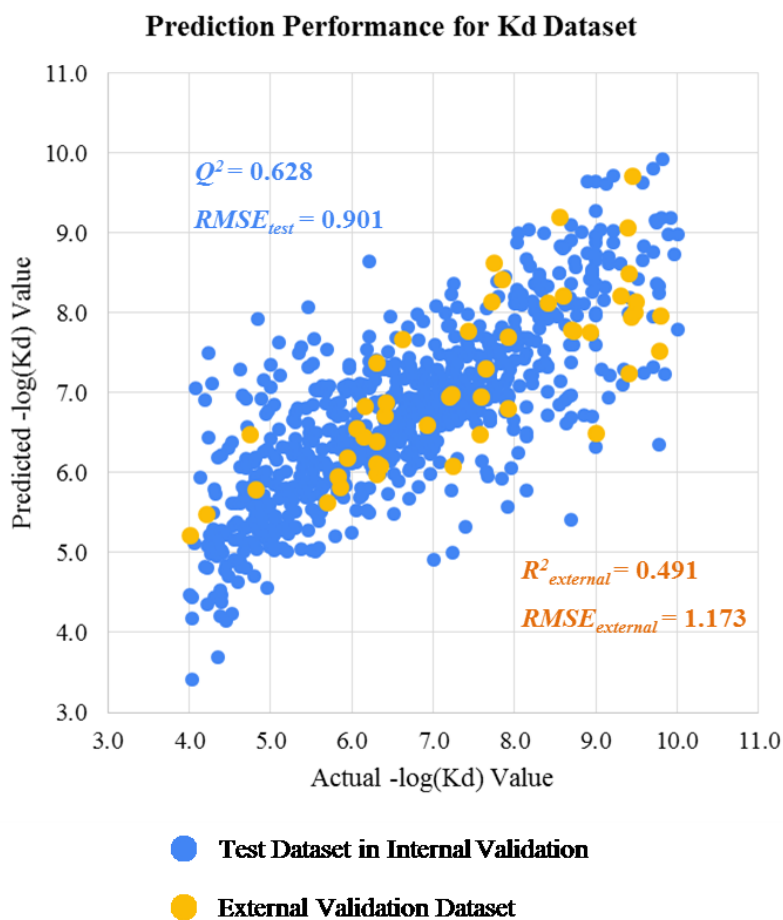


Figure 4-7 Plot of the predicted K_d value versus the actual K_d value by using the best QSKR model in K_d dataset

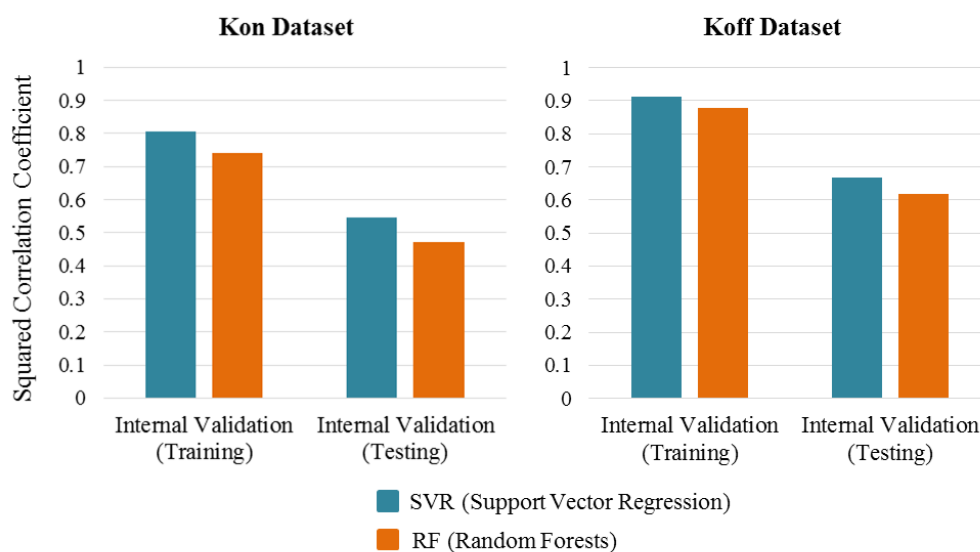


4.3.2 QSKR Prediction Performance on k_{on} and k_{off} Datasets

In the prediction of the association rate constant (k_{on}) and the dissociation rate constant (k_{off}), 86 protein complexes were modelled, and leave-one-out cross-validation (LOOCV) was applied to optimize the QSKR models. **Table 4-3** and **Figure 4-8** presented the highest modeling performance by SVR and RF algorithms, in k_{on} and k_{off} datasets respectively. The best QSKR model in predicting the association rate constant k_{on} was constructed by SVR, reaching ($R^2 = 0.807$, $RMSE_{train} = 0.612$, $Q^2 = 0.545$, $RMSE_{test} = 0.980$) in LOOCV. The best performance in predicting the dissociation rate constant k_{off} was also delivered by SVR, and the LOOCV gave ($R^2 = 0.807$, $RMSE_{train} = 0.612$) for internal training, and ($Q^2 = 0.545$, $RMSE_{test} = 0.980$) for internal testing.

Table 4-3 The best QSKR performance in predicting the k_{on} and k_{off} value

| Algorithm | Leave-One-Out Cross-Validation | | | | | | | |
|------------|--------------------------------|----------------|-------|---------------|-------------------|----------------|-------|---------------|
| | k_{on} Dataset | | | | k_{off} Dataset | | | |
| | R^2 | $RMSE_{train}$ | Q^2 | $RMSE_{test}$ | R^2 | $RMSE_{train}$ | Q^2 | $RMSE_{test}$ |
| SVR | 0.807 | 0.612 | 0.545 | 0.980 | 0.912 | 0.472 | 0.667 | 0.961 |
| RF | 0.741 | 0.725 | 0.472 | 1.314 | 0.877 | 0.539 | 0.619 | 0.977 |

Figure 4-8 The best QSKR model performance in predicting k_{on} and k_{off} value

As we observed, the best QSKR models for predicting k_{on} and k_{off} value were again constructed by SVR method, transcended RF method in this study. The parameter optimizations (heat plots) in searching for the best SVR models were illustrated in **Figure 4-9** for k_{on} dataset, and **Figure 4-11** for k_{off} dataset. The best QSKR models by SVR were built by setting ($gamma \approx 72$, $epsilon \approx 2$) for k_{on} dataset, and ($gamma \approx 70$, $epsilon \approx 0.8$) for k_{off} dataset. **Figure 4-10** and **Figure 4-12** showed the plots of the predicted value versus the actual value by using the best QSKR models for k_{on} and k_{off} datasets respectively. The red dots in **Figure 4-10** denoted the internal testing data in k_{on} dataset ($Q^2 = 0.545$, $RMSE_{test} = 0.980$), and the green dots in **Figure 4-12** denoted the internal testing data in k_{off} dataset ($Q^2 = 0.667$, $RMSE_{test} = 0.961$).

Figure 4-9 Heat plot of k_{on} prediction performance in SVR parameter optimization

(A) the average R^2 of training dataset, and (B) the average Q^2 of testing dataset, in internal leave-one-out cross-validation

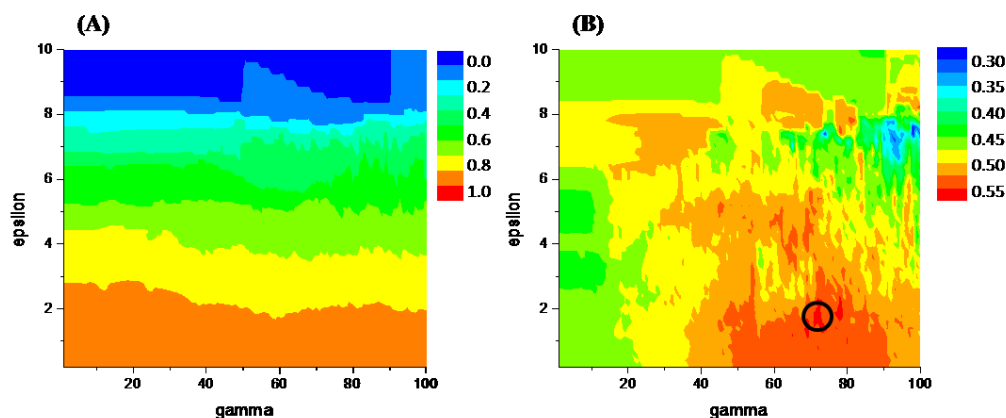


Figure 4-10 Plot of the predicted k_{on} value versus the actual k_{on} value by using the best QSKR model in k_{on} dataset

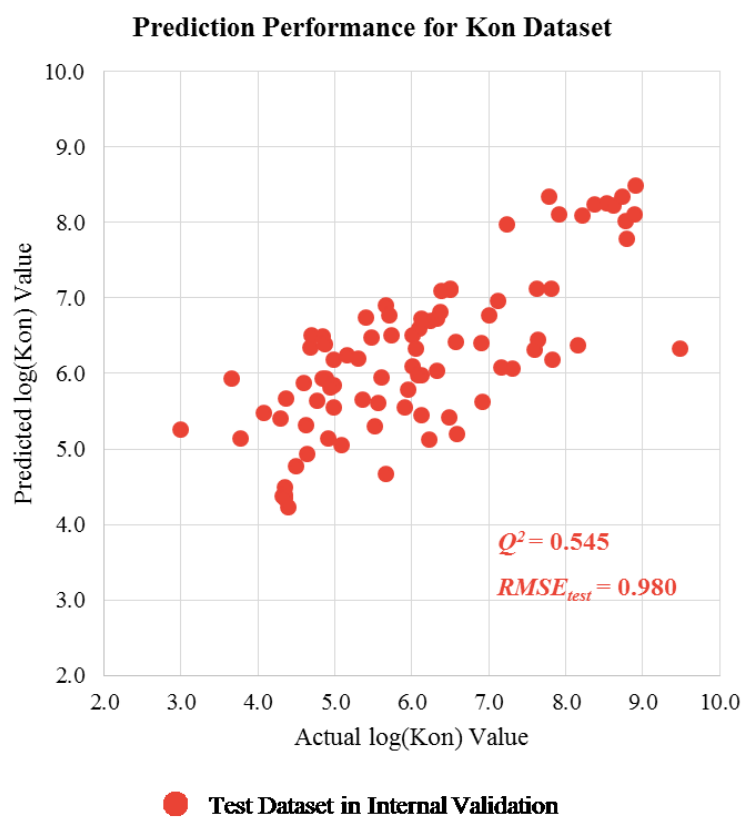


Figure 4-11 Heat plot of k_{off} prediction performance in SVR parameter optimization

(A) the average R^2 of training dataset, and (B) the average Q^2 of testing dataset, in internal leave-one-out cross-validation

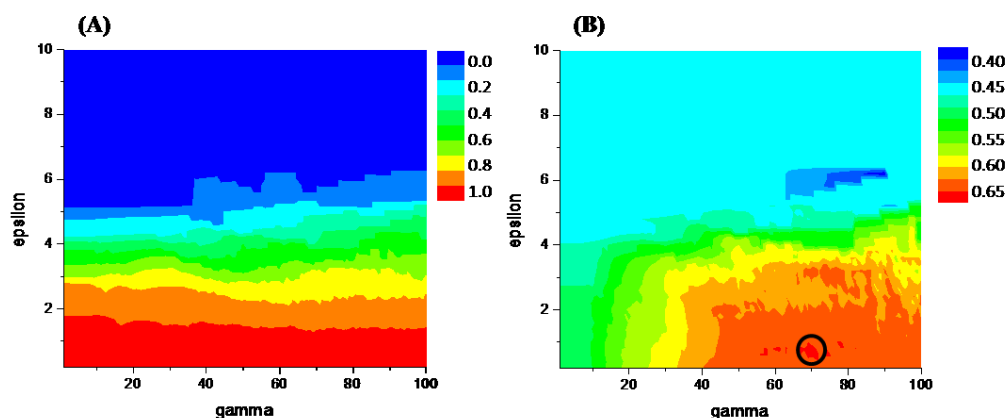
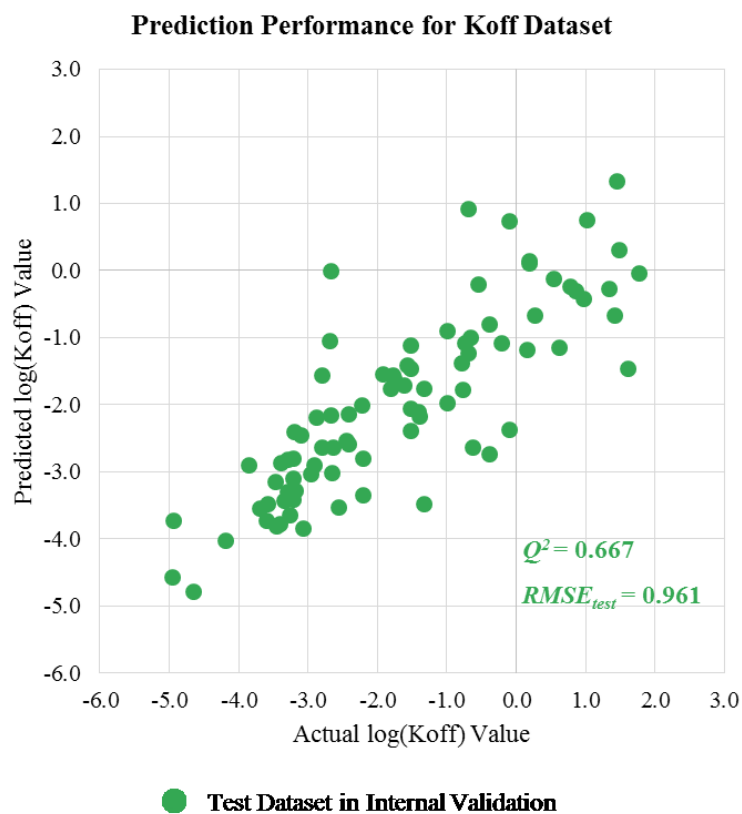


Figure 4-12 Plot of the predicted k_{off} value versus the actual k_{off} value by using the best QSKR model in k_{off} dataset



4.4 Conclusion and Discussion

As a proof of concept, we collected a far more diverse (~ 6 -times larger) PPI library than the largest previous study, and our QSKR study has evaluated the feasibility of quantitative prediction of protein-protein interaction kinetic constants (K_d , k_{on} and k_{off}) by only using the protein features computed from protein primary sequence.

It was also observed that, in this study, Support Vector Regression outperformed Random Forests in modeling and predicting all the three PPI kinetic constants, which might because we carried out more refined grid search for optimizing the SVR parameters, while RF demonstrated quite stable performance within a range of parameter value, so we did not apply refined grid search for optimizing RF parameters.

Our best QSKR models achieved ($R^2 = 0.86$, $Q^2 = 0.63$, $RMSE_{test} = 0.90$) for K_d dataset in the internal 10-fold cross-validations, while the external validation reached ($R^2_{external} = 0.49$, $RMSE_{external} = 1.17$) in predicting the 50 PPIs that have no available 3D protein structures, which would be an impossible task for those structure-based prediction methods. Compared with some reported work that gave K_d prediction $Q^2 = 0.69\sim 0.77$ in leave-one-out cross-validation, our result was a bit lower (0.63 vs $0.69\sim 0.77$), however we modelled the 6-times larger protein feature space (840 vs up-to 137) and applied 10-fold cross-validation, rather than leave-one-out cross-validation that is supposed to give a higher Q^2 value.

On the other hand, the best QSKR models achieved ($R^2 = 0.81$, $Q^2 = 0.55$, $RMSE_{test} = 0.98$) for the k_{on} dataset, and ($R^2 = 0.91$, $Q^2 = 0.67$, $RMSE_{test} = 0.96$) for the k_{off} dataset, in the internal leave-one-out cross-validations. The relatively lower performance for k_{on} prediction and the relatively higher performance for k_{off} prediction in this study, were consistent with the previous findings based on the structure-based study of 62 PPIs, which gave $Q^2 = 0.732$ and 0.801 in predicting k_{on} and k_{off} respectively. We speculated that the worse prediction of k_{on} is very likely because of the water molecules have to be displaced before the protein binding, and the on rates have strong dependence on ionic strength whereas the off rates are relatively insensitive. Currently, there are no good ways of accounting for this, even for the structure-based methods.

Our work had demonstrated an evidence of general quantitative relationships between protein primary sequence information and protein-protein interaction kinetic constants, thus offering an opportunity to predict the binding affinities between protein pairs having unknown structures. This QSKR study may help in efficient selection of bioactive compounds with sufficient potencies to compete with the protein substrates, so as to further facilitate the drug discovery for inhibiting the protein-protein interactions.

However, this work is limited by not fully considering the post-translational modifications, experimental conditions, physicochemical factors, which will be discussed in section 5.2.

CHAPTER 5 Concluding Remarks

In this thesis, my primary interests were focused on the development of bioinformatics tools (PROFEAT webserver and TISPIN database) to facilitate the study of complex biological networks, and the construction of machine-learning models for predicting the protein-protein interaction kinetic constants.

The merits of each study were delivered in the **Section 5.1**. The limitations of the current work and the suggestions for further studies were discusses in the **Section 5.2**.

5.1 Major Finding and Contributions

5.1.1 Merits of Upgrading PROFEAT Webserver for Computing Biological Network Descriptors

The major update of PROFEAT for computing the biological network descriptors could greatly cater to the extensive needs of quantitative analysis of biological, disease, and pharmacological networks.

PROFEAT webserver has a number of remarkable distinguished advantages over the other publically accessible tools:

- 1) it provided the most comprehensive and diverse (up-to 379 *vs* 3~100 in other tools) network descriptors at the node-level (local properties), the edge-level (local properties) and the network-level (global properties);
- 2) it broadly covered different network types (undirected/directed, unweighted/weighted edges or nodes) for representing different kinds of biological networks (binary/oriented, constant/varying binding constants or molecular levels);
- 3) it was very user-friendly in simple input/output, and it required easy operation with minimal manual interventions;
- 4) it supported different network file formats to be compatible with the major network analysis software;
- 5) it enabled the automatic detection, split, and computation of multiple disconnected networks from a single input file.

In perspective, PROFEAT would considerably facilitate the functional biological investigations by providing the systematic properties of molecular interaction networks, offering the expanded understandings of biological complex systems, and revealing the higher-level clues of what the mechanisms could be.

This work will enable the further applications of machine learning methods, especially deep learning, into the study of systems biology networks, because machine learning methods require large quantity of features to represent the systems and to train the model, which was impossible before the our work. Therefore, we opened up a new door between biological science and computer science, and produced many more research possibilities and opportunities.

5.1.2 Merits of Developing TISPIN Database for Providing Human Tissue-Specific Protein Interaction Networks

The birth of TISPIN database (<http://bidd2.nus.edu.sg/TISPIN/home.php>), which provided the comprehensive information on human tissue-specific protein interaction networks, would offer an advantage in starting and expediting the studies in biological networks, tissue-specificity, disease mechanism, and biomarker/target identifications.

As a prototype, TISPIN 1.0 has already demonstrated its strength against other relevant databases. Differently from other databases that used transcriptional data to infer the tissue-protein associations, TISPIN was on the basis of protein-expression evidence, which is more reliable and meaningful to derive the protein interaction networks. So far, the database interface and architecture have been accomplished, and TISPIN 1.0 was fully functional in delivering:

- 1) network files in various formats compatible with the major network software;
- 2) network visualizations;
- 3) computed network descriptors in node-level (for each protein) and network-level (for the entire network);
- 4) protein annotations in terms of protein name, gene symbol, UniProt ID, NCBI ID, biological process / cellular component / molecular function, and therapeutic targets;
- 5) comprehensive downloadable links for all of the information in TISPIN.

5.1.3 Merits of Studying the Quantitative Sequence-Kinetic Constants Relationship to Predict Protein-Protein Interaction Kinetic Constants

The first highlight of this QSKR study was the collection of a much more diverse (~ 6 -times larger) PPI library than the largest previous study (840 PPIs vs 137 PPIs), which could be a benchmarked resource for further studies on protein-protein interaction kinetics.

As for the core objective, this study evaluated and confirmed the feasibility of quantitative prediction of protein-protein interaction kinetic constants (K_d , k_{on} and k_{off}) by only using the information from protein primary sequence. The best QSKR performance achieved ($R^2 = 0.86$, $Q^2 = 0.63$) in modeling the K_d dataset, ($R^2 = 0.81$, $Q^2 = 0.55$) in modeling the k_{on} dataset, and ($R^2 = 0.91$, $Q^2 = 0.67$) in modeling the k_{off} dataset, in internal cross-validations. Moreover, the external validation in K_d dataset gave ($R^2_{external} = 0.49$) in predicting the 50 PPIs that do not have protein 3D structures, which would be not applicable for those structure-based prediction methods.

This work may complement with other experimental and computational approaches for more efficient selection of bioactive compounds with sufficient potencies to compete with the protein substrates, so as to further facilitate the drug discovery for PPI inhibitors.

5.2 Limitations and Suggestions for Further Studies

5.2.1 Limitations and Suggestions for PROFEAT Webserver

Besides the successful development of PROFEAT webserver for computing the comprehensive biological network descriptors, the current main limitation is the speed of computation. For calculating the full-set of network properties of an undirected unweighted network, PROFEAT would finish the task within 30 seconds for a network having no more than 300 nodes or 600 edges, and within 5 minutes for a network having no more than 1,000 nodes or 2,000 edges. However, due to the large number of the network descriptors and the high complexity in computing some specific properties, the time cost could get substantially higher when the network size grows larger, especially for the weighted networks.

By the recently added option for the slim-set of network descriptors, the running time for an undirected unweighted network (1625 nodes, 3336 edges) was reduced from 26 minutes (full-set) to 1.4 minutes (slim-set), and the running time for an undirected edge-weighted network (315 nodes, 632 edges) was reduced from 100 seconds (full-set) to 30 seconds (slim-set). Nevertheless, to improve the functionality and the efficiency of PROFEAT, more efficient algorithms and programming structures should be implemented for further improvement.

For a more comprehensively functional PROFEAT webserver, more network descriptors could be implemented, particularly expanding the features for directed networks. An interactive network map would be incorporated as well,

by providing an attractive function for users to manipulate the network components, focus on some specific molecules of interest in the network, and exam the network features of those molecules in a more user-friendly interface.

In addition, PROFEAT could be further improved by enabling the powerful function for the variation analysis of network properties, which is to allow the users to input different sets of networks or different sets of edge/node weight (e.g. gene expression, RNAseq count) in different health status, by different treatments, or at different sampling time points.

Furthermore, we would eagerly apply these diverse PROFEAT network descriptors into the study of various biological networks (protein-protein interaction networks, gene co-expression networks, metabolic networks) for further investigations: evaluating the important network descriptors for different biological systems, discovering more characteristics of drug targets and biomarkers in biological networks, and so on.

5.2.2 Limitations and Suggestions for TISPIN Database

TISPIN database prototype 1.0 has shown its significant advantages in well-designed web interface and broad spectrum of the provided network-related information. However, the current version of TISPIN 1.0 only adopted the data from HPRD database, which is far less than enough. To enlarge the PPI data size, more sources should be considered, particularly BioGRID, DIP, HIPPIE, HPRD, InnateDB, IntAct, MINT, and STRING databases. To enrich the tissue-protein associations based on the protein-level expression evidence, Human Protein Atlas^{196,197} and Protein Abundance Database¹⁹⁸ should be integrated.

Moreover, an extra search function for “Search by Proteins/Genes” should be added, more network file format (e.g. NET) should be supported, and regular (quarterly or semi-annual) update should be carried out to keep the pace with the other databases.

We would build up the networks based on transcriptional data as well, and provide users the options to choose the network of interest. The tissue-specific network could be derived from protein-level data, transcriptional-level data, or both. If selecting the network combining both the protein-level and transcriptional-level data, we could use different color to represent the source of the interaction (e.g. red for physical interaction, green for high correlation).

Thereafter, TISPIN database would be much more reliable, functional, informative, and applicable in biological network studies.

5.2.3 Limitations and Suggestions for QSKR Study

QSKR study has provided a proof of concept in predicting the PPI kinetic constants (K_d , k_{on} and k_{off}) from the protein primary sequences. It achieved acceptably good performance, but not as robust as the other published predictions that based on the protein 3D structures. We deem that our sequence-based modeling method would offer a great opportunity to predict the PPI kinetic constants without knowing the protein complex structures. In further perspective, a rational classification of proteins according to the functional families or the binding mechanisms may be suggested, rather than training the general PPI models. Therefore, different QSKR prediction models will be built for different protein functional classes or different interaction mechanisms, such that making the PPI kinetic constants predictions more specific, more accurate, and more interpretative.

In addition, effects of post-translational modifications should be considered, for example the phosphorylation can increase k_{off} value leading to the stabilization of one of the PPI partners that is needed for biological activity. More attentions should be made in post-translational modifications, by incorporating the PTM prediction tools provided in ExPASy resource portal¹⁹⁹.

For further deciphering the PPI kinetics, we should not only consider the nature of the interacting protein partners, but also the effects of temperature, pH, buffer, viscosity, etc. in the interaction environment.

BIBLIOGRAPHY

- 1 Newman, M. E. J. *Networks: An Introduction*. (Oxford University Press, 2010).
- 2 Shimbel, A. Structural parameters of communication networks. *The bulletin of mathematical biophysics*, 15, 501-507, (1953).
- 3 Freeman, L. C. A set of measures of centrality based on betweenness. *Sociometry*, 40, 35-41, (1977).
- 4 Sabidussi, G. The centrality index of a graph. *Psychometrika*, 31, 581-603, (1966).
- 5 Davis, J. A. Clustering and hierarchy in interpersonal relations: testing two graph theoretical models on 742 sociomatrices. *American Sociological Review*, 35, 843-851, (1970).
- 6 Holland, P. W. & Leinhardt, S. Transitivity in structural models of small groups. *Small Group Res.*, 2, 107-124, (1971).
- 7 Skorobogatov, V. A. & Dobrynin, A. A. Metrical analysis of graphs. *MATCH Commun Math Comp Chem*, 23, 105-155, (1988).
- 8 Barabasi, A. L. & Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nat Rev Genet*, 5, 101-113, (2004).
- 9 Watts, D. J. & Strogatz, S. H. Collective dynamics of 'small-world' networks. *Nature*, 393, 440-442, (1998).
- 10 Brandes, U. A faster algorithm for betweenness centrality. *J Math Sociol*, 25, (2001).
- 11 Yoon, J., Blumer, A. & Lee, K. An algorithm for modularity analysis of directed and weighted biological networks based on edge-betweenness centrality. *Bioinformatics*, 22, 3106-3108, (2006).
- 12 Latora, V. & Marchiori, M. Efficient behavior of small-world networks. *Physical Review Letters*, 87, 198701, (2001).
- 13 Rodrigue, J. P. *The Geography of Transport Systems*. Third edn, (Routledge, 2013).
- 14 Langville, A. N. & Meyer, C. D. A survey of eigenvector methods of web information retrieval. *SIAM Rev.*, 47, 135-161, (2005).
- 15 Wiedermann, M., Donges, J. F., Heitzig, J. & Kurths, J. Node-weighted interacting network measures improve the representation of real-world complex systems. *EPL (Europhysics Letters)*, 102, 28007, (2013).
- 16 Langville, A. N. & Meyer, C. D. *Google's PageRank and Beyond: The Science of Search Engine Rankings*. (Princeton University Press, 2012).
- 17 Michael, W. B. & Murray, B. *Understanding Search Engines: Mathematical Modeling and Text Retrieval*. Vol. 8 (Society for Industrial and Applied Mathematics, 1999).

- 18 Brin, S. & Page, L. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *7th International World-Wide Web Conference*, (1998).
- 19 Banky, D., Ivan, G. & Grolmusz, V. Equal opportunity for low-degree network nodes: a PageRank-based method for protein target identification in metabolic graphs. *PLoS One*, 8, e54204, (2013).
- 20 Gleich, D. F. PageRank beyond the Web. *SIAM Rev*, 57, 321-363, (2014).
- 21 Ivan, G. & Grolmusz, V. When the Web meets the cell: using personalized PageRank for analyzing protein interaction networks. *Bioinformatics*, 27, 405-407, (2011).
- 22 Ma'ayan, A. Introduction to network analysis in systems biology. *Sci Signal*, 4, tr5, (2011).
- 23 Yook, S. H., Oltvai, Z. N. & Barabasi, A. L. Functional and topological characterization of protein interaction networks. *Proteomics*, 4, 928-942, (2004).
- 24 Winterbach, W., Van Mieghem, P., Reinders, M., Wang, H. & de Ridder, D. Topology of molecular interaction networks. *BMC Syst Biol*, 7, 90, (2013).
- 25 Koyuturk, M. Algorithmic and analytical methods in network biology. *Wiley Interdiscip Rev Syst Biol Med*, 2, 277-292, (2010).
- 26 Barabasi, A. L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *Nat Rev Genet*, 12, 56-68, (2011).
- 27 Cho, D. Y., Kim, Y. A. & Przytycka, T. M. Network biology approach to complex diseases. *PLoS Comput Biol*, 8, e1002820, (2012).
- 28 Furlong, L. I. Human diseases through the lens of network biology. *Trends Genet*, 29, 150-159, (2013).
- 29 Zhang, B., Tian, Y. & Zhang, Z. Network biology in medicine and beyond. *Circ Cardiovasc Genet*, 7, 536-547, (2014).
- 30 Schadt, E. E. & Björkegren, J. L. NEW: Network-Enabled Wisdom in biology, medicine and healthcare. *Sci Transl Med*, 4, 115rv111, (2012).
- 31 Hopkins, A. L. Network pharmacology: the next paradigm in drug discovery. *Nat Chem Biol*, 4, 682-690, (2008).
- 32 Yildirim, M. A., Goh, K. I., Cusick, M. E., Barabasi, A. L. & Vidal, M. Drug-target network. *Nat Biotechnol*, 25, 1119-1126, (2007).
- 33 Pujol, A., Mosca, R., Farres, J. & Aloy, P. Unveiling the role of network and systems biology in drug discovery. *Trends Pharmacol Sci*, 31, 115-123, (2010).
- 34 Guney, E., Menche, J., Vidal, M. & Barabasi, A. L. Network-based in silico drug efficacy screening. *Nat Commun*, 7, 10331, (2016).
- 35 Chandra, N. & Padiadpu, J. Network approaches to drug discovery. *Expert Opin Drug Discov*, 8, 7-20, (2013).
- 36 Marx, V. Cancer: smoother journeys for molecular data. *Nat Methods*, 12, 299-302, (2015).

- 37 Yao, L. & Rzhetsky, A. Quantitative systems-level determinants of human genes targeted by successful drugs. *Genome Res*, 18, 206-213, (2008).
- 38 Stelzl, U. *et al.* A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, 122, 957-968, (2005).
- 39 Maslov, S. & Sneppen, K. Specificity and stability in topology of protein networks. *Science*, 296, 910-913, (2002).
- 40 Piraveenan, M., Uddin, S. & Chung, K. S. K. in *Advances in Social Networks Analysis and Mining, IEEE* 38-45 (IEEE, Istanbul, 2012).
- 41 Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 13, 2498-2504, (2003).
- 42 Djebbari, A. *et al.* NAViGaTOR: large scalable and interactive navigation and analysis of large graphs. *Internet Mathematics*, 7, 314-347, (2011).
- 43 Bastian, M., Heymann, S. & Jacomy, M. Gephi: an open source software for exploring and manipulating networks. *3rd International AAAI Conference on Weblogs and Social Media*, (2009).
- 44 Brinkrolf, C. *et al.* VANESA - a software application for the visualization and analysis of networks in system biology applications. *J Integr Bioinform*, 11, 239, (2014).
- 45 Batagelj, V. & Mrvar, A. Pajek – Program for Large Network Analysis. *Connections*, 21, 47-57, (1998).
- 46 Forman, J. J., Clemons, P. A., Schreiber, S. L. & Haggarty, S. J. SpectralNET--an application for spectral graph analysis and visualization. *BMC Bioinformatics*, 6, 260, (2005).
- 47 Wu, J. *et al.* Integrated network analysis platform for protein-protein interactions. *Nat Methods*, 6, 75-77, (2009).
- 48 Lin, C. Y. *et al.* Hubba: hub objects analyzer--a framework of interactome hubs identification for network biology. *Nucleic Acids Res*, 36, W438-443, (2008).
- 49 Reimand, J., Tooming, L., Peterson, H., Adler, P. & Vilo, J. GraphWeb: mining heterogeneous biological networks for gene modules with functional significance. *Nucleic Acids Res*, 36, W452-459, (2008).
- 50 Yip, K. Y., Yu, H., Kim, P. M., Schultz, M. & Gerstein, M. The tYNA platform for comparative interactomics: a web tool for managing, comparing and mining multiple networks. *Bioinformatics*, 22, 2968-2970, (2006).
- 51 Hu, Z. *et al.* VisANT: data-integrating visual framework for biological networks and modules. *Nucleic Acids Res*, 33, W352-357, (2005).
- 52 Hagberg, A. A., Schult, D. A. & Swart, P. J. Exploring network structure, dynamics, and function using networkX. *Proceedings of 7th Python in Science Conference* 11-15, (2008).

- 53 Gabor Csardi , T. N. The igraph software package for complex network research. *InterJ. Complex Syst.*, (2006).
- 54 Mueller, L. A., Kugler, K. G., Dander, A., Graber, A. & Dehmer, M. QuACN: an R package for analyzing complex biological networks quantitatively. *Bioinformatics*, 27, 140-141, (2011).
- 55 Doncheva, N. T., Assenov, Y., Domingues, F. S. & Albrecht, M. Topological analysis and interactive visualization of biological networks and protein structures. *Nat Protoc*, 7, 670-685, (2012).
- 56 Winter, C. *et al.* Google goes cancer: improving outcome prediction for cancer patients by network-based ranking of marker genes. *PLoS Comput Biol*, 8, e1002511, (2012).
- 57 Emig, D. *et al.* Drug target prediction and repositioning using an integrated network-based approach. *PLoS One*, 8, e60618, (2013).
- 58 Hsu, C. L., Huang, Y. H., Hsu, C. T. & Yang, U. C. Prioritizing disease candidate genes by a gene interconnectedness-based approach. *BMC Genomics*, 12 Suppl 3, S25, (2011).
- 59 Zhu, C., Kushwaha, A., Berman, K. & Jegga, A. G. A vertex similarity-based framework to discover and rank orphan disease-related genes. *BMC Syst Biol*, 6 Suppl 3, S8, (2012).
- 60 Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol*, 4, Article17, (2005).
- 61 Saramäki, J., Kivelä, M., Onnela, J. P., Kaski, K. & Kertész, J. Generalizations of the clustering coefficient to weighted complex networks. *Physical Review E*, 75, 027105, (2007).
- 62 Guan, Y. *et al.* Tissue-specific functional networks for prioritizing phenotype and disease genes. *PLoS Comput Biol*, 8, e1002694, (2012).
- 63 Greene, C. S. *et al.* Understanding multicellular function and disease with human tissue-specific networks. *Nat Genet*, 47, 569-576, (2015).
- 64 Bossi, A. & Lehner, B. Tissue specificity and the human protein interaction network. *Mol Syst Biol*, 5, 260, (2009).
- 65 Lin, W. H., Liu, W. C. & Hwang, M. J. Topological and organizational properties of the products of house-keeping and tissue-specific genes in protein-protein interaction networks. *BMC Syst Biol*, 3, 32, (2009).
- 66 Lopes, T. J. *et al.* Tissue-specific subnetworks and characteristics of publicly available human protein interaction databases. *Bioinformatics*, 27, 2414-2421, (2011).
- 67 Emig, D., Kacprowski, T. & Albrecht, M. Measuring and analyzing tissue specificity of human genes and protein complexes. *EURASIP J Bioinform Syst Biol*, 2011, 5, (2011).
- 68 Liu, W., Wang, J., Wang, T. & Xie, H. Construction and analyses of human large-scale tissue specific networks. *PLoS One*, 9, e115074, (2014).

- 69 Dezso, Z. *et al.* A comprehensive functional analysis of tissue specificity of human gene expression. *BMC Biol*, 6, 49, (2008).
- 70 Mohammadi, S. & Grama, A. A convex optimization approach for identification of human tissue-specific interactomes. *Bioinformatics*, 32, i243-i252, (2016).
- 71 Magger, O., Waldman, Y. Y., Rupp, E. & Sharan, R. Enhancing the prioritization of disease-causing genes through tissue specific protein interaction networks. *PLoS Comput Biol*, 8, e1002690, (2012).
- 72 Barshir, R., Shwartz, O., Smoly, I. Y. & Yeager-Lotem, E. Comparative analysis of human tissue interactomes reveals factors leading to tissue-specific manifestation of hereditary diseases. *PLoS Comput Biol*, 10, e1003632, (2014).
- 73 Su, A. I. *et al.* A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A*, 101, 6062-6067, (2004).
- 74 Hsiao LL *et al.* A compendium of gene expression in normal human tissues. *Physiol Genomics*, 7, 97-104, (2001).
- 75 Peri, S. *et al.* Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res*, 13, 2363-2371, (2003).
- 76 Yeager-Lotem, E. & Sharan, R. Human protein interaction networks across tissues and diseases. *Front Genet*, 6, 257, (2015).
- 77 Emig, D. & Albrecht, M. Tissue-specific proteins and functional implications. *J Proteome Res*, 10, 1893-1903, (2011).
- 78 Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M. & Gilad, Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*, 18, 1509-1517, (2008).
- 79 Fu, X. *et al.* Estimating accuracy of RNA-Seq and microarrays with proteomics. *BMC Genomics*, 10, 161, (2009).
- 80 Zhao, S., Fung-Leung, W. P., Bittner, A., Ngo, K. & Liu, X. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PLoS One*, 9, e78644, (2014).
- 81 Sirbu, A., Kerr, G., Crane, M. & Ruskin, H. J. RNA-Seq vs dual- and single-channel microarray data: sensitivity analysis for differential expression and clustering. *PLoS One*, 7, e50986, (2012).
- 82 Vogel, C. & Marcotte, E. M. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat Rev Genet*, 13, 227-232, (2012).
- 83 Chatr-Aryamontri, A. *et al.* The BioGRID interaction database: 2015 update. *Nucleic Acids Res*, 43, D470-478, (2015).
- 84 Salwinski, L. *et al.* The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res*, 32, D449-451, (2004).
- 85 Schaefer, M. H. *et al.* HIPPIE: Integrating protein interaction networks with experiment based quality scores. *PLoS One*, 7, e31826, (2012).

- 86 Keshava Prasad, T. S. *et al.* Human Protein Reference Database--2009 update. *Nucleic Acids Res*, 37, D767-772, (2009).
- 87 Breuer, K. *et al.* InnateDB: systems biology of innate immunity and beyond--recent updates and continuing curation. *Nucleic Acids Res*, 41, D1228-1233, (2013).
- 88 Kerrien, S. *et al.* The IntAct molecular interaction database in 2012. *Nucleic Acids Res*, 40, D841-846, (2012).
- 89 Licata, L. *et al.* MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res*, 40, D857-861, (2012).
- 90 Szklarczyk, D. *et al.* STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res*, 43, D447-452, (2015).
- 91 Barshir, R. *et al.* The TissueNet database of human tissue protein-protein interactions. *Nucleic Acids Res*, 41, D841-844, (2013).
- 92 Micale, G., Ferro, A., Pulvirenti, A. & Giugno, R. SPECTRA: An Integrated Knowledge Base for Comparing Tissue and Tumor-Specific PPI Networks in Human. *Front Bioeng Biotechnol*, 3, 58, (2015).
- 93 Kotlyar, M., Pastrello, C., Sheahan, N. & Jurisica, I. Integrated interactions database: tissue-specific view of the human and model organism interactomes. *Nucleic Acids Res*, 44, D536-541, (2016).
- 94 Elefsinioti, A. *et al.* Large-scale de novo prediction of physical protein-protein association. *Mol Cell Proteomics*, 10, M111 010629, (2011).
- 95 Kotlyar, M. *et al.* In silico prediction of physical protein interactions and characterization of interactome orphans. *Nat Methods*, 12, 79-84, (2015).
- 96 Zhang, Q. C. *et al.* Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature*, 490, 556-560, (2012).
- 97 Rhodes, D. R. *et al.* Probabilistic model of the human protein-protein interaction network. *Nat Biotechnol*, 23, 951-959, (2005).
- 98 Steven Fletcher, A. D. H. Protein-protein interaction inhibitors: small molecules from screening technique. *current topics in medicinal chemistry*, 7, 922-927, (2007).
- 99 Arkin, M. R. & Whitty, A. The road less traveled: modulating signal transduction enzymes by inhibiting their protein-protein interactions. *Curr Opin Chem Biol*, 13, 284-290, (2009).
- 100 Chene, P. Inhibiting the p53-MDM2 interaction: an important target for cancer therapy. *Nature reviews. Cancer*, 3, 102-109, (2003).
- 101 Casey, F. P., Pihan, E. & Shields, D. C. Discovery of small molecule inhibitors of protein-protein interactions using combined ligand and target score normalization. *J Chem Inf Model*, 49, 2708-2717, (2009).
- 102 Moll, U. M. & Petrenko, O. The MDM2-p53 interaction. *Mol Cancer Res*, 1, 1001-1008, (2003).
- 103 Basse, M. J. *et al.* 2P2Idb: a structural database dedicated to orthosteric modulation of protein-protein interactions. *Nucleic Acids Res*, 41, D824-827, (2013).

- 104 Dell'Orco, D. Fast predictions of thermodynamics and kinetics of protein-protein recognition from structures: from molecular design to systems biology. *Molecular bioSystems*, 5, 323-334, (2009).
- 105 Schreiber, G. Kinetic studies of protein-protein interactions. *Current Opinion in Structural Biology*, 12, 41-47, (2002).
- 106 Tzvia Selzer, S. A. a. G. S. Rational design of faster associating binding protein complexes. *Nature*, 7, 537-541, (2000).
- 107 HANS FRAUENFELDER, S. G. S., PETER G. WOLYNES. The energy landscapes and motions of proteins. *Science*, 254, 1598-1603, (1991).
- 108 Haynie, D. T. *Biological Thermodynamics*. (Cambridge University Press, 2008).
- 109 Haddadian, E. J. & Gross, E. L. A Brownian dynamics study of the interactions of the luminal domains of the cytochrome b6f complex with plastocyanin and cytochrome c6: the effects of the Rieske FeS protein on the interactions. *Biophysical journal*, 91, 2589-2600, (2006).
- 110 Scott H. NORTHRUP, a. H. P. E. Kinetics of protein-protein association explained by Brownian. *Proc. Natl. Acad. Sci. USA*, 89, 3338-3342, (1992).
- 111 Alsallaq, R. & Zhou, H. X. Prediction of protein-protein association rates from a transition-state theory. *Structure*, 15, 215-224, (2007).
- 112 Rong-Kai, X. *et al.* Quantitative sequence-kinetics relationship in antigen-antibody interaction kinetics based on a set of descriptors [corrected]. *Chemical biology & drug design*, 76, 345-349, (2010).
- 113 Bai, H. *et al.* Predicting kinetic constants of protein-protein interactions based on structural properties. *Proteins*, 79, 720-734, (2011).
- 114 Iain H. Moal, R. A. a. P. A. B. Protein-protein binding affinity prediction on a diverse set of structures. *Bioinformatics*, (2011).
- 115 Kastritis, P. L. *et al.* A structure-based benchmark for protein-protein binding affinity. *Protein science : a publication of the Protein Society*, 20, 482-491, (2011).
- 116 Brenner, S. Frederick Sanger (1918–2013). *Science*, 343, 262, (2014).
- 117 Xenarios, I. *et al.* DIP: the database of interacting proteins. *Nucleic Acids Res*, 28, 289-291, (2000).
- 118 Gough, J. R. B. a. D. A. Predicting protein-protein interactions from primary structure. *Bioinformatics*, 17, 455-460, (2001).
- 119 Lo, S. L., Cai, C. Z., Chen, Y. Z. & Chung, M. C. Effect of training datasets on support vector machine prediction of protein-protein interactions. *Proteomics*, 5, 876-884, (2005).
- 120 Jiankuan Ye, C. K., and Ilya Muchnik. Sequence-based Protein-Protein Interaction Prediction Optimized for Target Selection in Biological Experiments in *Proceedings of the 2005 IEEE, Engineering in Medicine and Biology 27th Annual Conference*.

- 121 Guo, Y., Yu, L., Wen, Z. & Li, M. Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences. *Nucleic Acids Res*, 36, 3025-3030, (2008).
- 122 Moal, I. H. & Bates, P. A. Kinetic rate constant prediction supports the conformational selection mechanism of protein binding. *PLoS computational biology*, 8, e1002351, (2012).
- 123 Ma, D., Guo, Y., Luo, J., Pu, X. & Li, M. Prediction of protein-protein binding affinity using diverse protein-protein interface features. *Chemometrics and Intelligent Laboratory Systems*, 138, 7-13, (2014).
- 124 Schreiber, G., Haran, G. & Zhou, H. X. Fundamental aspects of protein-protein association kinetics. *Chemical reviews*, 109, 839-860, (2009).
- 125 Lu, L. J. *et al.* Comparing classical pathways and modern networks: towards the development of an edge ontology. *Trends Biochem Sci*, 32, 320-331, (2007).
- 126 Mathivanan, S. *et al.* An evaluation of human protein-protein interaction data in the public domain. *BMC Bioinformatics*, 7 Suppl 5, S19, (2006).
- 127 Li, Z. R. *et al.* PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Res*, 34, W32-37, (2006).
- 128 Rao, H. B., Zhu, F., Yang, G. B., Li, Z. R. & Chen, Y. Z. Update of PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Research*, 39, W385-W390, (2011).
- 129 Yao, L. & Rzhetsky, A. Quantitative systems-level determinants of human genes targeted by successful drugs. *Genome Res*, 18, 206-213, (2008).
- 130 Goh, K. I. *et al.* The human disease network. *Proc Natl Acad Sci U S A*, 104, 8685-8690, (2007).
- 131 Zhang, L., Li, X., Tai, J., Li, W. & Chen, L. Predicting candidate genes based on combined network topological features: a case study in coronary artery disease. *PLoS One*, 7, e39542, (2012).
- 132 Carter, S. L., Brechbuhler, C. M., Griffin, M. & Bond, A. T. Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics*, 20, 2242-2250, (2004).
- 133 Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. & Barabási, A. L. Hierarchical organization of modularity in metabolic networks. *Science*, 297, 1551-1555, (2002).
- 134 Snel, B., Bork, P. & Huynen, M. A. The identification of functional modules from the genomic association of genes. *Proc Natl Acad Sci U S A*, 99, 5890-5895, (2002).
- 135 Sharan, R., Ulitsky, I. & Shamir, R. Network-based prediction of protein function. *Mol Syst Biol*, 3, 88, (2007).
- 136 Maslov, S. & Sneppen, K. Specificity and stability in topology of protein networks. *Science*, 296, 910-913, (2002).

- 137 Dong, J. & Horvath, S. Understanding network concepts in modules. *BMC Syst Biol*, 1, 24, (2007).
- 138 Netzer, M. *et al.* A network-based feature selection approach to identify metabolic signatures in disease. *J Theor Biol*, 310, 216-222, (2012).
- 139 Dirk Koschützki , F. S. Centrality analysis methods for biological networks and their application to gene regulatory networks. *Gene Regul Syst Bio*, 2, 193-201, (2008).
- 140 Peng, X. *et al.* Exploring a structural protein-drug interactome for new therapeutics in lung cancer. *Mol Biosyst*, 10, 581-591, (2014).
- 141 Pang, K., Cheng, C., Xuan, Z., Sheng, H. & Ma, X. Understanding protein evolutionary rate by integrating gene co-expression with protein interactions. *BMC Syst Biol*, 4, 179, (2010).
- 142 Jacunski, A. & Tatonetti, N. P. Connecting the dots: applications of network medicine in pharmacology and disease. *Clin Pharmacol Ther*, 94, 659-669, (2013).
- 143 Joy, M. P., Brock, A., Ingber, D. E. & Huang, S. High-betweenness proteins in the yeast protein interaction network. *Journal of Biomedicine and Biotechnology*, 96-103, (2005).
- 144 Yu, H., Kim, P. M., Sprecher, E., Trifonov, V. & Gerstein, M. The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Comput Biol*, 3, e59, (2007).
- 145 Harrold, J. M., Ramanathan, M. & Mager, D. E. Network-based approaches in drug discovery and early development. *Clin Pharmacol Ther*, 94, 651-658, (2013).
- 146 Rubinov, M. & Sporns, O. Complex network measures of brain connectivity: uses and interpretations. *Neuroimage*, 52, 1059-1069, (2010).
- 147 Tong, A. H. *et al.* Global mapping of the yeast genetic interaction network. *Science*, 303, 808-813, (2004).
- 148 Emmert-Streib, F. & Dehmer, M. Networks for systems biology: conceptual connection of data and function. *IET Syst Biol*, 5, 185-207, (2011).
- 149 Bonchev, D. *Complexity in Chemistry, Biology, and Ecology*. (Springer US, 2007).
- 150 Kim, J. & Wilhelm, T. What is a complex graph? *Physica A: Statistical Mechanics and its Applications*, 387, 2637-2652, (2008).
- 151 Mueller, L. A., Kugler, K. G., Netzer, M., Graber, A. & Dehmer, M. A network-based approach to classify the three domains of life. *Biol Direct*, 6, 53, (2011).
- 152 Dehmer, M. & Mowshowitz, A. A history of graph entropy measures. *Information Sciences*, 181, 57-78, (2011).
- 153 Dehmer, M., Grabner, M. & Varmuza, K. Information indices with high discriminative power for graphs. *PLoS One*, 7, (2012).

- 154 Paladugu, S. R., Zhao, S., Ray, A. & Raval, A. Mining protein networks for synthetic genetic interactions. *BMC Bioinformatics*, 9, 426, (2008).
- 155 You, Z. H., Yin, Z., Han, K., Huang, D. S. & Zhou, X. A semi-supervised learning approach to predict synthetic genetic interactions by combining functional and topological properties of functional gene network. *BMC Bioinformatics*, 11, 343, (2010).
- 156 Dehmer, M. Uniquely discriminating molecular structures using novel eigenvalue-based descriptors. *MATCH Commun. Math. Comput. Chem.*, 67, 147-172, (2012).
- 157 Luscombe, N. M. *et al.* Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, 431, 308-312, (2004).
- 158 Pei Wang, J. L., Xinghuo Yu. Identification of important nodes in directed biological networks: a network motif approach. *PLoS One*, 9, e106132, (2014).
- 159 Newman, M. E. J. A measure of betweenness centrality based on random walks. *Social Networks*, 27, (2003).
- 160 Dijkstra, E. W. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1, 269-271, (1959).
- 161 Ma, H. W. & Zeng, A. P. The connectivity structure, giant strong component and centrality of metabolic networks. *Bioinformatics*, 19, 1423-1430, (2003).
- 162 Gutman, I. & Zhou, B. Laplacian energy of a graph. *Linear Algebra and its Applications*, 414, 29-37, (2006).
- 163 Leroy, G. *Information Theoretic Indices for Characterization of Chemical Structures*. Vol. 27 (Wiley, 1985).
- 164 Cerami, E. G. *et al.* Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res*, 39, D685-690, (2011).
- 165 Mishra, G. R. *et al.* Human protein reference database--2006 update. *Nucleic Acids Res*, 34, D411-414, (2006).
- 166 Peri, S. *et al.* Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res*, 32, D497-501, (2004).
- 167 Tong, A. H. *et al.* Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science*, 294, 2364-2368, (2001).
- 168 Fukushima, A., Kusano, M., Redestig, H., Arita, M. & Saito, K. Metabolomic correlation-network modules in Arabidopsis based on a graph-clustering approach. *BMC Syst Biol*, 5, 1, (2011).
- 169 Zheng, C. J. *et al.* Therapeutic targets: progress of their exploration and investigation of their characteristics. *Pharmacol Rev*, 58, 259-279, (2006).
- 170 Hopkins AL & CR, G. The druggable genome. *Nat Rev Drug Discov*, 1, 727-730, (2002).
- 171 Rual, J. F. *et al.* Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437, 1173-1178, (2005).

- 172 Zhang, J. *et al.* Analysis of bypass signaling in EGFR pathway and profiling of bypass genes for predicting response to anticancer EGFR tyrosine kinase inhibitors. *Mol Biosyst*, 8, 2645-2656, (2012).
- 173 Kanehisa, M. *et al.* Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res*, 42, D199-205, (2014).
- 174 Chelliah, V. *et al.* BioModels: ten-year anniversary. *Nucleic Acids Res*, 43, D542-548, (2015).
- 175 Kumar, P. *et al.* Update of KDBI: kinetic data of bio-molecular interaction database. *Nucleic Acids Research*, 37, D636-D641, (2009).
- 176 Zhu, F. *et al.* Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery. *Nucleic Acids Res*, 40, D1128-1136, (2012).
- 177 Yang, H. *et al.* Therapeutic target database update 2016: enriched resource for bench to clinical drug target and targeted pathway information. *Nucleic Acids Res*, 44, D1069-1074, (2016).
- 178 Gene Ontology, C. Gene Ontology Consortium: going forward. *Nucleic Acids Res*, 43, D1049-1056, (2015).
- 179 Chen, J., Sawyer, N. & Regan, L. Protein-protein interactions: general trends in the relationship between binding affinity and interfacial buried surface area. *Protein science : a publication of the Protein Society*, 22, 510-515, (2013).
- 180 Mullard, A. Protein-protein interaction inhibitors get into the groove. *Nat Rev Drug Discov*, 11, 173-175, (2012).
- 181 Renxiao Wang, X. F., Yipin Lu, Chao-Yie Yang, and Shaomeng Wang. The PDBbind Database: Methodologies and Updates. *J. Med. Chem*, 48, 4111-4119, (2004).
- 182 Rao, H. B., Zhu, F., Yang, G. B., Li, Z. R. & Chen, Y. Z. Update of PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic acids research*, 39, W385-390, (2011).
- 183 Chou, K. C. & Cai, Y. D. Using functional domain composition and support vector machines for prediction of protein subcellular location. *The Journal of biological chemistry*, 277, 45765-45769, (2002).
- 184 Sun, S. H. a. Z. Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, 17, 721-728, (2001).
- 185 Bhasin, M. & Raghava, G. P. Classification of nuclear receptors based on amino acid composition and dipeptide composition. *The Journal of biological chemistry*, 279, 23262-23266, (2004).
- 186 Shen, J. *et al.* Predicting protein-protein interactions based only on sequences information. *Proc Natl Acad Sci U S A*, 104, 4337-4341, (2007).
- 187 Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A. J. & Vapnik, V. Support Vector Regression Machines in *Advances in Neural Information Processing Systems*. 155-161 (MIT Press).

- 188 Cortes, C. & Vapnik, V. Support-Vector Networks. *Machine Learning*, 20, 273-297, (1995).
- 189 Debasish Basak, S. P. a. D. C. P. Support Vector Regression. *Neural Information Processing*, 11, 203-224, (2007).
- 190 Lin, C.-C. C. a. C.-J. LIBSVM - A Library for Support Vector Machines. (2001).
- 191 R. Burbidge, M. T., B. Buxton, S. Holden. Drug design by machine learning - SVM for pharmaceutical data analysis. *Computers and Chemistry*, 26, 5-14, (2001).
- 192 Trotter, M. W. B., Buxton, B. F. & Holden, S. B. Support Vector Machines in Combinatorial Chemistry. *Measurement and Control*, 34, 235-239, (2001).
- 193 Ryszard Czerminski, A. Y. a. D. H. Use of SVM in Pattern Classification - Application to QSAR Studies. *Quant. Struct.-Act. Relat*, 20, 227-240, (2001).
- 194 Breiman, L. Random Forests. *Machine Learning*, 45, 5-32, (2001).
- 195 Liaw, A. & Wiener, M. in *R News* Vol. Vol. 2/3 18-23 (Wien : R Foundation for Statistical Computing, 2002).
- 196 Uhlen, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science*, 347, 1260419, (2015).
- 197 Uhlen, M. *et al.* Towards a knowledge-based Human Protein Atlas. *Nat Biotechnol*, 28, 1248-1250, (2010).
- 198 Wang, M. *et al.* PaxDb, a database of protein abundance averages across all three domains of life. *Mol Cell Proteomics*, 11, 492-500, (2012).
- 199 Artimo, P. *et al.* ExPASy: SIB bioinformatics resource portal. *Nucleic Acids Res*, 40, W597-603, (2012).
- 200 Milo, R. *et al.* Network Motifs Simple Building Blocks of Complex Networks. *Science*, 298, 824-827, (2002).
- 201 Hwang, W., Cho, Y., Zhang, A. & Ramanathan, M. Bridging Centrality: Identifying Bridging Nodes In Scale-free Networks. *12th ACM International Conference on Knowledge Discovery and Data Mining*, (2006).
- 202 Freeman, L. C. Centrality in Social Networks Conceptual Clarification. *Social Networks*, 1, 215-239, (1978).
- 203 Rochat, Y. Closeness Centrality Extended to Unconnected Graphs: The Harmonic Centrality Index. *ASNA*, (2009).
- 204 Dangalchev, C. Residual Closeness in Networks. *Physica A: Statistical Mechanics and its Applications*, 365, 556-564, (2006).
- 205 K. Stephenson, M. Z. Rethinking centrality: methods and examples. *Social Networks*, 11, 1-37, (1989).
- 206 Newman, M. E. J. Betweenness centrality based on random walks. *Social Networks*, 27, 39-54, (2005).

- 207 Borgatti, S. P. Centrality and network flow. *Social Networks*, 27, 55-71, (2005).
- 208 Straffin, P. D. Linear Algebra in Geography: Eigenvectors of Networks. *Mathematics Magazine*, 53, 269-276, (1980).
- 209 Newman, M. E. J. *Mathematics of Networks*. 2nd edn, (Palgrave Macmillan, 2008).
- 210 Barrat, A., Barthélemy, M., Pastor-Satorras, R. & Vespignani, A. The architecture of complex weighted networks. *Proc Natl Acad Sci U S A*, 101, 3747-3752, (2004).
- 211 Leung, C. C. & Chau, H. F. Weighted Assortative And Disassortative Networks Model. *Physica A: Statistical Mechanics and its Applications*, 378, 591-602, (2008).
- 212 Barthélemy, M. Architecture of Complex Weighted Networks. *Talk Series on Networks and Complex Systems*, (2005).
- 213 Onnela, J. P., Saramäki, J., Kertész, J. & Kaski, K. Intensity and Coherence of Motifs in Weighted Complex Networks. *Physical Review E*, 71, (2005).
- 214 Holme, P., Park, S. M., Kim, B. J. & Edling, C. R. Korean University Life in a Network Perspective: Dynamics of a Large Affiliation Network. *Physica A: Statistical Mechanics and its Applications*, 373, 821-830, (2007).
- 215 Mangioni, G. *Complex Networks: Results of the 1st International Workshop on Complex Networks*. 1st edn, Vol. 207 (Springer-Verlag Berlin Heidelberg, 2009).
- 216 Newman, M. E. J. Mixing patterns in networks. *Physical Review E*, 67, (2003).
- 217 Petitjean, M. Applications of the Radius-Diameter Diagram to the Classification of Topological and Geometrical Shapes of Chemical Compounds. *J. Chem. Inf. Comput. Sci.*, 32, 331-337, (1992).
- 218 Sharma, V., Goswami, R. & Madan, A. K. Eccentric Connectivity Index: A Novel Highly Discriminating Topological Descriptor for Structure-Property and Structure-Activity Studies. *J. Chem. Inf. Comput. Sci.*, 37, 273-282, (1997).
- 219 Li, X. & Gutman, I. *Mathematical Aspects of Randić-Type Molecular Structure Descriptors*. (University of Kragujevac, 2006).
- 220 Furtula, B., Graovac, A. & Vukičević, D. Atom-Bond Connectivity Index of Trees. *Discrete Applied Mathematics*, 157, 2828-2835, (2009).
- 221 Diudea, M. V., Gutman, I. & Lorentz, J. *Molecular Topology*. (Nova Publishing, 2001).
- 222 Vukičević, D. & Furtula, B. Topological Index Based on The Ratios of Geometrical and Arithmetical Means of End-Vertex Degrees of Ddges. *Journal of Mathematical Chemistry*, 46, 1369-1376, (2009).

- 223 Khalifeh, M. H., Yousefi-Azari, H. & Ashrafi, A. R. The First and Second Zagreb Indices of Some Graph Operations. *Discrete Appl. Math.*, 157, 804-811, (2009).
- 224 Gutman, I. Graph theory and molecular orbitals. XII. Acyclic polyenes. *The Journal of Chemical Physics*, 62, 3399, (1975).
- 225 Narumi, H. & Katayama, M. *Simple Topological Index. A Newly Devised Index Characterizing the Topological Nature of Structural Isomers of Saturated Hydrocarbons*, Hokkaido Univ. Japan, (1984).
- 226 Narumi, H. New Topological Indices for Finite and Infinite Systems. *MATCH Commun. Math. Chem.*, 22, 195-207, (1987).
- 227 Dehmer, M., Grabner, M. & Furtula, B. Structural discrimination of networks by using distance, degree and eigenvalue-based measures. *PLoS One*, 7, e38564, (2012).
- 228 Wiener, H. Structural Determination of Paraffin Boiling Points. *Journal of the American Chemical Society*, 69, 17-20, (1947).
- 229 Klein, D. J., Lukovits, I. & Gutman, I. On the Definition of the Hyper-Wiener Index for Cycle-Containing Structures. *J. Chem. Inf. Comput. Sci.*, 35, 50-62, (1995).
- 230 Balaban, A. *Topological Indices and Related Descriptors in QSAR and QSPAR*. (CRC Press, 2000).
- 231 Doyle, J. K. & Graver, J. E. Mean Distance in a Graph. *Discrete Mathematics*, 17, (1977).
- 232 Gupta, S. Superpendentic Index: A Novel Topological Descriptor for Predicting Biological Activity. *J. Chem. Inf. Comput. Sci.*, 39, 272-277, (1999).
- 233 Diudea, M. V. Walk Numbers eWM: Wiener-Type Numbers of Higher Rank. *J. Chem. Inf. Comput. Sci.*, 36, 535-540, (1996).
- 234 Todeschini, R. & Consonni, V. *Handbook of Molecular Descriptors*. (Wiley VCH, 2008).
- 235 Balaban, A. Highly Discriminating Distance-Based Topological Index. *Chemical Physics Letters*, 89, (1982).
- 236 Bono Lucic *et al.* *On the Novel Balaban-Like and Balaban-Detour-Like Molecular Descriptors*. Vol. 5 (Nova Science Publishers, 2013).
- 237 Zhou, B., Gutman, I., Furtula, B. & Du, Z. B. On Two Types of Geometric-Arithmetic Index. *Chemical Physics Letters*, 482, 153-155, (2009).
- 238 Khadika, P. V. The Szeged Index and an Analogy with the Wiener Index. *J. Chem. Inf. Comput. Sci.*, 35, 547-550, (1995).
- 239 Schultz, H. P., Schultz, E. B. & Schultz, T. P. Topological Organic Chemistry. 4. Graph theory, Matrix Permanents, and Topological Indices of Alkanes. *J. Chem. Inf. Comput. Sci.*, 32, 69-72, (1992).
- 240 Schultz, H. P. Topological Organic Chemistry. 1. Graph Theory and Topological Indices of Alkanes. *J. Chem. Inf. Comput. Sci.*, 29, 227-228, (1989).

- 241 Gutman, I. Selected properties of the Schultz molecular topological index. *J. Chem. Inf. Comput. Sci.*, 34, 1087-1089, (1994).
- 242 Mowshowitz, A. & Dehmer, M. Entropy and the Complexity of Graphs Revisited. *Entropy*, 14, 559-570, (2012).
- 243 Bonchev, D., Mekenyan, O. V. & Trinajstii, N. Isomer Discrimination by Topological Information Approach. *Journal of Computational Chemistry*, 2, 127-148, (1981).
- 244 Balaban, A. T., Bertelsen, S. & Basak, S. C. New centric topological indexes for acyclic molecules (trees) and substituents (rooted trees) and coding of rooted trees. *MATCH Communications in Mathematical and in Computer Chemistry*, 30, 55-72, (1994).
- 245 Dehmer, M., Sivakumar, L. & Varmuza, K. *On Distance-Based Entropy Measures*. Vol. 12 (University of Kragujevac, 2012).
- 246 Raychaudhury, C., Ray, S. K., Ghosh, J. J., Roy, A. B. & Basak, S. C. Discrimination of Isomeric Structures Using Information Theoretic Topological Indices. *Journal of Computational Chemistry*, 5, 581-588, (1984).
- 247 Konstantinova, E. V. & Paleev, A. A. Sensitivity of Topological Indices of Polycyclic Graphs. *Vychisl Sistemy*, 136, 38-48, (2006).
- 248 Bertz, S. H. The First General Index of Molecular Complexity. *J. Am. Chem. Soc.*, 103, 3599-3601, (1981).
- 249 Bonchev, D. & Trinajstic, N. Information theory, distance matrix, and molecular branching. *The Journal of Chemical Physics*, 67, (1977).
- 250 Balaban, A. T. New Vertex Invariant and Topological Indices of Chemical Graphs Based on Information on Distances. *Journal of Mathematical Chemistry*, 8, 383-397, (1991).
- 251 Gradshteyn, I. S. & Ryzhik, I. M. *Tables of Integrals, Series, and Products*. 6th edn, 1115-1116 (Academic Press, 2000).
- 252 Estrada, E. Characterization of 3D Molecular Structure. *Chemical Physics Letters*, 319, 713-718, (2000).
- 253 Fath-Tabar, G. H., Ashra, A. R. & Gutman, I. Note on Estrada and L-Estrada Indices of Graphs. *Classe des Sciences Mathématiques et Naturelles, Sciences mathématiques naturelles*, CXXXIX, 1-16, (2009).
- 254 Karacali, B. Hierarchical Motif Vectors for Prediction of Functional Sites in Amino Acid Sequences Using Quasi-Supervised Learning. *Ieee-Acm Transactions on Computational Biology and Bioinformatics*, 9, 1432-1441, (2012).
- 255 Mohar B., Babic D. & N., T. A Novel Definition of the Wiener Index for Trees. *J. Chem. Inf. Comput. Sci.*, 33, 153-154, (1993).
- 256 Dehmer, M., Emmert-Streib, F., Tsoy, Y. R. & Varmuza, K. *Quantifying Structural Complexity of Graphs: Information Measures in Mathematical Chemistry*. (Nova Science Publishers, 2011).

- 257 Luo, J. & Magee, C. L. Detecting evolving patterns of self-organizing networks by flow hierarchy measurement. *Complexity*, 16, 53-61, (2011).
- 258 Brandes, U. On variants of shortest-path betweenness centrality and their generic computation. *Social Networks*, 30, 136-145, (2008).

APPENDICES

Section A: Network Descriptors in PROFEAT Webserver

Based on the indexing rule in PROFEAT webserver, each network descriptor was indexed as (X, Y, Z) , where node-level descriptors were indexed by $X=G10$, network-level descriptors were indexed by $X=G11$, and edge-level descriptors were indexed by $X=G12$. Secondly, descriptors were labelled as un-weighted, edge-weighted, node-weighted, or directed by $Y=1, 2, 3, 4$ respectively. Properties based on the normalized weight was labelled by an extra “N” in Y . Thirdly, Z represented the descriptor ID in **Table S-1, S-2, S-3**.

For some examples, $(G10, 1, 7)$ is the node-level un-weighted neighbourhood connectivity, $(G10, 2, 25)$ is the node-level edge-weighted betweenness centrality, $(G10, 4, 49)$ is the node-level directed local clustering coefficient, $(G11, 2N, 196)$ is the network-level normalized edge-weighted transitivity, $(G11, 3, 202)$ is the network-level node-weighted global clustering coefficient, and $(G12, 2N, 2)$ is the edge-level normalized edge-weighted edge betweenness.

In the tables below, all descriptors were grouped into different categories according to their definitions and algorithms, and each column listed the computed descriptors for each network type. Therefore, some notations were given: “O” ($Y = 1$) represents the features calculated based on un-weighted network adjacency information, “—” ($Y = 2$) represents the features calculated based on edge weight, “●” ($Y = 3$) represents the features calculated based on

node weight, and “↗” ($Y = 4$) represents the features calculated based on the directed information.

Additionally, a slim set of network descriptors were selected, which is a cut-down version of the PROFEAT network descriptors that have been particularly applied in studying systems biology or probing specific therapeutic questions. The descriptors in the slim-set were marked by “★” in the ID column.

Table S-1 List of the node-level descriptors covered in PROFEAT

| ID | (G10) Node-Level Network Descriptor | Network Type | | | | |
|---|---|--------------|---------------|---------------|--------------------|-------------|
| | | Un-Directed | | | | Directed |
| | | Un-Weighted | Edge Weighted | Node Weighted | Edge-Node Weighted | Un-Weighted |
| Connectivity/Adjacency-based Properties | | | | | | |
| ★1 | Degree | ○ | ○ | ○ | ○ | |
| 2 | Scaled Connectivity | ○ | ○ | ○ | ○ | |
| ★3 | Number of Selfloops | ○ | ○ | ○ | ○ | ↗ |
| ★4 | Number of Triangles | ○ | ○ | ○ | ○ | ↗ |
| 5 | Z Score | ○ | ○ | ○ | ○ | |
| ★6 | Clustering Coefficient | ○ | ○ | ○ | ○ | |
| ★7 | Neighborhood Connectivity | ○ | ○ | ○ | ○ | |
| ★8 | Topological Coefficient | ○ | ○ | ○ | ○ | |
| ★9 | Interconnectivity | ○ | ○ | ○ | ○ | |
| ★10 | Bridging Coefficient | ○ | ○ | ○ | ○ | |
| ★11 | Degree Centrality | ○ | ○ | ○ | ○ | |
| Shortest Path Length-based Properties | | | | | | |
| ★12 | Average Shortest Path Length | ○ | ○ — | ○ | ○ — | |
| 13 | Distance Sum | ○ | ○ — | ○ | ○ — | |
| ★14 | Eccentricity | ○ | ○ — | ○ | ○ — | |
| 15 | Eccentric | ○ | ○ — | ○ | ○ — | |
| 16 | Deviation | ○ | ○ — | ○ | ○ — | |
| 17 | Distance Deviation | ○ | ○ — | ○ | ○ — | |
| ★18 | Radiality | ○ | ○ — | ○ | ○ — | |
| ★19 | Closeness Centrality (avg) | ○ | ○ — | ○ | ○ — | |
| 20 | Closeness Centrality (sum) | ○ | ○ — | ○ | ○ — | |
| ★21 | Eccentricity Centrality | ○ | ○ — | ○ | ○ — | |
| 22 | Harmonic Closeness Centrality | ○ | ○ — | ○ | ○ — | |
| 23 | Residual Closeness Centrality | ○ | ○ — | ○ | ○ — | |

| | | | | | | |
|---|---------------------------------------|---|-----|---|-----|---|
| ★24 | Load Centrality (Stress) | ○ | ○ — | ○ | ○ — | |
| ★25 | Betweenness Centrality | ○ | ○ — | ○ | ○ — | |
| 26 | Normalized Betweenness | ○ | ○ — | ○ | ○ — | |
| ★27 | Bridging Centrality | ○ | ○ — | ○ | ○ — | |
| 28 | CurrentFlow Betweenness | ○ | ○ — | ○ | ○ — | |
| 29 | CurrentFlow Closeness | ○ | ○ — | ○ | ○ — | |
| Eigenvector-based Complexity Indices | | | | | | |
| ★30 | Eigenvector Centrality | ○ | ○ | ○ | ○ | |
| ★31 | Page Rank Centrality | ○ | ○ | ○ | ○ | |
| Edge-Weighted Properties | | | | | | |
| ★32 | Strength | | — | | — | |
| ★33 | Assortativity | | — | | — | |
| 34 | Disparity | | — | | — | |
| 35 | Geometric Mean of Triangles | | — | | — | |
| 36 | Barrat's Local Clustering Coefficient | | — | | — | |
| 37 | Onnela's Local Clustering Coefficient | | — | | — | |
| 38 | Zhang's Local Clustering Coefficient | | — | | — | |
| 39 | Holme's Local Clustering Coefficient | | — | | — | |
| ★40 | Edge-Weighted Interconnectivity | | — | | — | |
| Node-Weighted Properties | | | | | | |
| ★41 | Node Weight | | | ● | ● | |
| 42 | Node Weighted Cross Degree | | | ● | ● | |
| 43 | Node Weighted Local Clustering Coeff. | | | ● | ● | |
| ★44 | Node-Weighted Neighbourhood Score | | | ● | ● | |
| Directed Properties | | | | | | |
| ★45 | In-Degree | | | | | ↗ |
| 46 | In-Degree Centrality | | | | | ↗ |
| ★47 | Out-Degree | | | | | ↗ |
| 48 | Out-Degree Centrality | | | | | ↗ |
| ★49 | Directed Local Clustering Coefficient | | | | | ↗ |
| 50 | Neighbourhood Connectivity (only in) | | | | | ↗ |
| 51 | Neighbourhood Connectivity (only out) | | | | | ↗ |
| 52 | Neighbourhood Connectivity (in & out) | | | | | ↗ |
| 53 | Average Directed Neighbour Degree | | | | | ↗ |

Table S-2 List of the network-level descriptors covered in PROFEAT

| ID | (G11) Network-Level Network Descriptor | Network Type | | | | |
|---|--|-----------------|------------------|------------------|---------------------------|-----------------|
| | | Un-Directed | | | | Directed |
| | | Un- Weighted | Edge Weighted | Node Weighted | Edge- Node Weighted | Un- Weighted |
| Connectivity/Adjacency-based Properties | | | | | | |
| ★1 | Number of Nodes | ○ | ○ | ○ | ○ | ○ |
| ★2 | Number of Edges | ○ | ○ | ○ | ○ | ○ |
| ★3 | Number of Selfloops | ○ | ○ | ○ | ○ | ↗ |
| ★4 | Maximum Connectivity | ○ | ○ | ○ | ○ | |
| ★5 | Minimum Connectivity | ○ | ○ | ○ | ○ | |
| ★6 | Average Number of Neighbours | ○ | ○ | ○ | ○ | |
| 7 | Total Adjacency | ○ | ○ | ○ | ○ | |
| ★8 | Network Density | ○ | ○ | ○ | ○ | ↗ |
| ★9 | Average Clustering Coefficient | ○ | ○ | ○ | ○ | |
| ★10 | Transitivity | ○ | ○ | ○ | ○ | |
| ★11 | Heterogeneity | ○ | ○ | ○ | ○ | |
| ★12 | Degree Centralization | ○ | ○ | ○ | ○ | |
| 13 | Central Point Dominance | ○ | ○ | ○ | ○ | |
| 14 | Degree Assortativity Coefficient | ○ | ○ | ○ | ○ | |
| Shortest Path Length-based Properties | | | | | | |
| 15 | Total Distance | ○ | ○ — | ○ | ○ — | |
| ★16 | Network Diameter | ○ | ○ — | ○ | ○ — | |
| ★17 | Network Radius | ○ | ○ — | ○ | ○ — | |
| 18 | Shape Coefficient | ○ | ○ — | ○ | ○ — | |
| ★19 | Characterisite Path Length | ○ | ○ — | ○ | ○ — | |
| 20 | Network Eccentricity | ○ | ○ — | ○ | ○ — | |
| ★21 | Average Eccentricity | ○ | ○ — | ○ | ○ — | |
| 22 | Network Eccentric | ○ | ○ — | ○ | ○ — | |
| 23 | Eccentric Connectivity | ○ | ○ — | ○ | ○ — | |
| 24 | Unipolarity | ○ | ○ — | ○ | ○ — | |
| 25 | Integration | ○ | ○ — | ○ | ○ — | |
| 26 | Variation | ○ | ○ — | ○ | ○ — | |
| 27 | Average Distance | ○ | ○ — | ○ | ○ — | |
| 28 | Mean Distance Deviation | ○ | ○ — | ○ | ○ — | |
| 29 | Centralization | ○ | ○ — | ○ | ○ — | |
| ★30 | Global Efficiency | ○ | ○ — | ○ | ○ — | |
| Topological Indices | | | | | | |
| 31 | Edge Complexity Index | ○ | ○ | ○ | ○ | |
| ★32 | Randic Connectivity Index | ○ | ○ | ○ | ○ | |
| 33 | Atom-Bond Connectivity Index | ○ | ○ | ○ | ○ | |
| 34 | Zagreb Index 1 | ○ | ○ | ○ | ○ | |
| 35 | Zagreb Index 2 | ○ | ○ | ○ | ○ | |
| 36 | Zagreb Index Modified | ○ | ○ | ○ | ○ | |

| | | | | | | |
|---|---|---|-----|---|-----|--|
| 37 | Zagreb Index Augmented | ○ | ○ | ○ | ○ | |
| 38 | Zagreb Index Variable | ○ | ○ | ○ | ○ | |
| 39 | Narumi-Katayama Index | ○ | ○ | ○ | ○ | |
| 40 | Narumi-Katayama Index (log) | ○ | ○ | ○ | ○ | |
| 41 | Narumi Geometric Index | ○ | ○ | ○ | ○ | |
| 42 | Narumi Harmonic Index | ○ | ○ | ○ | ○ | |
| 43 | Alpha Index | ○ | ○ | ○ | ○ | |
| 44 | Beta Index | ○ | ○ | ○ | ○ | |
| 45 | Pi Index | ○ | ○ | ○ | ○ | |
| 46 | Eta Index | ○ | ○ | ○ | ○ | |
| ★47 | Hierarchy | ○ | ○ | ○ | ○ | |
| ★48 | Robustness | ○ | ○ | ○ | ○ | |
| 49 | Medium Articulation | ○ | ○ | ○ | ○ | |
| 50 | Complexity Index A | ○ | ○ — | ○ | ○ — | |
| 51 | Complexity Index B | ○ | ○ — | ○ | ○ — | |
| ★52 | Wiener Index | ○ | ○ — | ○ | ○ — | |
| 53 | Hyper-Wiener | ○ | ○ — | ○ | ○ — | |
| 54 | Harary Index 1 | ○ | ○ — | ○ | ○ — | |
| 55 | Harary Index 2 | ○ | ○ — | ○ | ○ — | |
| 56 | Compactness Index | ○ | ○ — | ○ | ○ — | |
| 57 | Superpendentic Index | ○ | ○ — | ○ | ○ — | |
| 58 | Hyper-Distance-Path Index | ○ | ○ | ○ | ○ | |
| ★59 | BalabanJ Index | ○ | ○ — | ○ | ○ — | |
| 60 | BalabanJ-like 1 Index | ○ | ○ — | ○ | ○ — | |
| 61 | BalabanJ-like 2 Index | ○ | ○ — | ○ | ○ — | |
| 62 | BalabanJ-like 3 Index | ○ | ○ — | ○ | ○ — | |
| 63 | Geometric Arithmetic Index 1 | ○ | ○ | ○ | ○ | |
| 64 | Geometric Arithmetic Index 2 | ○ | ○ — | ○ | ○ — | |
| 65 | Geometric Arithmetic Index 3 | ○ | ○ — | ○ | ○ — | |
| 66 | Szeged Index | ○ | ○ — | ○ | ○ — | |
| 67 | Product Of Row Sums | ○ | ○ — | ○ | ○ — | |
| 68 | Product Of Row Sums (log) | ○ | ○ — | ○ | ○ — | |
| 69 | Schultz Topological Index | ○ | ○ — | ○ | ○ — | |
| 70 | Gutman Topological Index | ○ | ○ — | ○ | ○ — | |
| 71 | Efficiency Complexity | ○ | ○ — | ○ | ○ — | |
| Entropy-based Complexity Indices | | | | | | |
| ★72 | Information Content (Degree Equality) | ○ | ○ | ○ | ○ | |
| 73 | Information Content (Edge Equality) | ○ | ○ | ○ | ○ | |
| 74 | Information Content (Edge Magnitude) | ○ | ○ | ○ | ○ | |
| 75 | Information Content (Distance Degree) | ○ | ○ | ○ | ○ | |
| 76 | Information Content (Dist Deg Equality) | ○ | ○ | ○ | ○ | |
| ★77 | Radial Centric Information Index | ○ | ○ | ○ | ○ | |
| 78 | Distance Degree Compactness | ○ | ○ | ○ | ○ | |
| 79 | Distance Degree Centric Index | ○ | ○ | ○ | ○ | |

| | | | | | | |
|--|-------------------------------------|---|---|---|---|--|
| 80 | Graph Distance Complexity | ○ | ○ | ○ | ○ | |
| 81 | Information Layer Index | ○ | ○ | ○ | ○ | |
| ★82 | Bonchev Information Index 1 | ○ | ○ | ○ | ○ | |
| ★83 | Bonchev Information Index 2 | ○ | ○ | ○ | ○ | |
| ★84 | Bonchev Information Index 3 | ○ | ○ | ○ | ○ | |
| 85 | Balaban-like Information Index 1 | ○ | ○ | ○ | ○ | |
| 86 | Balaban-like Information Index 2 | ○ | ○ | ○ | ○ | |
| Eigenvalue-based Complexity Indices | | | | | | |
| ★87 | Graph Energy | ○ | ○ | ○ | ○ | |
| ★88 | Laplacian Energy | ○ | ○ | ○ | ○ | |
| 89 | Spectral Radius | ○ | ○ | ○ | ○ | |
| 90 | Estrada Index | ○ | ○ | ○ | ○ | |
| 91 | Laplacian Estrada Index | ○ | ○ | ○ | ○ | |
| 92 | Quasi-Weiner Index | ○ | ○ | ○ | ○ | |
| 93 | Mohar Index 1 | ○ | ○ | ○ | ○ | |
| 94 | Mohar Index 2 | ○ | ○ | ○ | ○ | |
| 95 | Graph Index Complexity | ○ | ○ | ○ | ○ | |
| 96 | Adjacency Matrix HM (S=1) | ○ | ○ | ○ | ○ | |
| 97 | Adjacency Matrix SM (S=1) | ○ | ○ | ○ | ○ | |
| 98 | Adjacency Matrix ISM (S=1) | ○ | ○ | ○ | ○ | |
| 99 | Adjacency Matrix PM (S=1) | ○ | ○ | ○ | ○ | |
| 100 | Adjacency Matrix IPM (S=1) | ○ | ○ | ○ | ○ | |
| 101 | Laplacian Matrix HM (S=1) | ○ | ○ | ○ | ○ | |
| 102 | Laplacian Matrix SM (S=1) | ○ | ○ | ○ | ○ | |
| 103 | Laplacian Matrix ISM (S=1) | ○ | ○ | ○ | ○ | |
| 104 | Laplacian Matrix PM (S=1) | ○ | ○ | ○ | ○ | |
| 105 | Laplacian Matrix IPM (S=1) | ○ | ○ | ○ | ○ | |
| 106 | Distance Matrix HM (S=1) | ○ | ○ | ○ | ○ | |
| 107 | Distance Matrix SM (S=1) | ○ | ○ | ○ | ○ | |
| 108 | Distance Matrix ISM (S=1) | ○ | ○ | ○ | ○ | |
| 109 | Distance Matrix PM (S=1) | ○ | ○ | ○ | ○ | |
| 110 | Distance Matrix IPM (S=1) | ○ | ○ | ○ | ○ | |
| 111 | Distance Path Matrix HM (S=1) | ○ | ○ | ○ | ○ | |
| 112 | Distance Path Matrix SM (S=1) | ○ | ○ | ○ | ○ | |
| 113 | Distance Path Matrix ISM (S=1) | ○ | ○ | ○ | ○ | |
| 114 | Distance Path Matrix PM (S=1) | ○ | ○ | ○ | ○ | |
| 115 | Distance Path Matrix IPM (S=1) | ○ | ○ | ○ | ○ | |
| 116 | Aug. Vertex Degree Matrix HM (S=1) | ○ | ○ | ○ | ○ | |
| 117 | Aug. Vertex Degree Matrix SM (S=1) | ○ | ○ | ○ | ○ | |
| 118 | Aug. Vertex Degree Matrix ISM (S=1) | ○ | ○ | ○ | ○ | |
| 119 | Aug. Vertex Degree Matrix PM (S=1) | ○ | ○ | ○ | ○ | |
| 120 | Aug. Vertex Degree Matrix IPM (S=1) | ○ | ○ | ○ | ○ | |
| 121 | Extended Adjacency Matrix HM (S=1) | ○ | ○ | ○ | ○ | |
| 122 | Extended Adjacency Matrix SM (S=1) | ○ | ○ | ○ | ○ | |

Appendices

| | | | | | | |
|-----|--------------------------------------|---|---|---|---|--|
| 123 | Extended Adjacency Matrix ISM (S=1) | ○ | ○ | ○ | ○ | |
| 124 | Extended Adjacency Matrix PM (S=1) | ○ | ○ | ○ | ○ | |
| 125 | Extended Adjacency Matrix IPM (S=1) | ○ | ○ | ○ | ○ | |
| 126 | Vertex Connectivity Matrix HM (S=1) | ○ | ○ | ○ | ○ | |
| 127 | Vertex Connectivity Matrix SM (S=1) | ○ | ○ | ○ | ○ | |
| 128 | Vertex Connectivity Matrix ISM (S=1) | ○ | ○ | ○ | ○ | |
| 129 | Vertex Connectivity Matrix PM (S=1) | ○ | ○ | ○ | ○ | |
| 130 | Vertex Connectivity Matrix IPM (S=1) | ○ | ○ | ○ | ○ | |
| 131 | Random Walk Markov HM (S=1) | ○ | ○ | ○ | ○ | |
| 132 | Random Walk Markov SM (S=1) | ○ | ○ | ○ | ○ | |
| 133 | Random Walk Markov ISM (S=1) | ○ | ○ | ○ | ○ | |
| 134 | Random Walk Markov PM (S=1) | ○ | ○ | ○ | ○ | |
| 135 | Random Walk Markov IPM (S=1) | ○ | ○ | ○ | ○ | |
| 136 | Weighted Struct. Func. IM1 HM (S=1) | ○ | ○ | ○ | ○ | |
| 137 | Weighted Struct. Func. IM1 SM (S=1) | ○ | ○ | ○ | ○ | |
| 138 | Weighted Struct. Func. IM1 ISM (S=1) | ○ | ○ | ○ | ○ | |
| 139 | Weighted Struct. Func. IM1 PM (S=1) | ○ | ○ | ○ | ○ | |
| 140 | Weighted Struct. Func. IM1 IPM (S=1) | ○ | ○ | ○ | ○ | |
| 141 | Weighted Struct. Func. IM2 HM (S=1) | ○ | ○ | ○ | ○ | |
| 142 | Weighted Struct. Func. IM2 SM (S=1) | ○ | ○ | ○ | ○ | |
| 143 | Weighted Struct. Func. IM2 ISM (S=1) | ○ | ○ | ○ | ○ | |
| 144 | Weighted Struct. Func. IM2 PM (S=1) | ○ | ○ | ○ | ○ | |
| 145 | Weighted Struct. Func. IM2 IPM (S=1) | ○ | ○ | ○ | ○ | |
| 146 | Adjacency Matrix HM (S=2) | ○ | ○ | ○ | ○ | |
| 147 | Adjacency Matrix SM (S=2) | ○ | ○ | ○ | ○ | |
| 148 | Adjacency Matrix ISM (S=2) | ○ | ○ | ○ | ○ | |
| 149 | Adjacency Matrix PM (S=2) | ○ | ○ | ○ | ○ | |
| 150 | Adjacency Matrix IPM (S=2) | ○ | ○ | ○ | ○ | |
| 151 | Laplacian Matrix HM (S=2) | ○ | ○ | ○ | ○ | |
| 152 | Laplacian Matrix SM (S=2) | ○ | ○ | ○ | ○ | |
| 153 | Laplacian Matrix ISM (S=2) | ○ | ○ | ○ | ○ | |
| 154 | Laplacian Matrix PM (S=2) | ○ | ○ | ○ | ○ | |
| 155 | Laplacian Matrix IPM (S=2) | ○ | ○ | ○ | ○ | |
| 156 | Distance Matrix HM (S=2) | ○ | ○ | ○ | ○ | |
| 157 | Distance Matrix SM (S=2) | ○ | ○ | ○ | ○ | |
| 158 | Distance Matrix ISM (S=2) | ○ | ○ | ○ | ○ | |
| 159 | Distance Matrix PM (S=2) | ○ | ○ | ○ | ○ | |
| 160 | Distance Matrix IPM (S=2) | ○ | ○ | ○ | ○ | |
| 161 | Distance Path Matrix HM (S=2) | ○ | ○ | ○ | ○ | |
| 162 | Distance Path Matrix SM (S=2) | ○ | ○ | ○ | ○ | |
| 163 | Distance Path Matrix ISM (S=2) | ○ | ○ | ○ | ○ | |
| 164 | Distance Path Matrix PM (S=2) | ○ | ○ | ○ | ○ | |
| 165 | Distance Path Matrix IPM (S=2) | ○ | ○ | ○ | ○ | |
| 166 | Aug. Vertex Degree Matrix HM (S=2) | ○ | ○ | ○ | ○ | |

| | | | | | | |
|---------------------------------|--|---|---|---|---|---|
| 167 | Aug. Vertex Degree Matrix SM (S=2) | ○ | ○ | ○ | ○ | |
| 168 | Aug. Vertex Degree Matrix ISM (S=2) | ○ | ○ | ○ | ○ | |
| 169 | Aug. Vertex Degree Matrix PM (S=2) | ○ | ○ | ○ | ○ | |
| 170 | Aug. Vertex Degree Matrix IPM (S=2) | ○ | ○ | ○ | ○ | |
| 171 | Extended Adjacency Matrix HM (S=2) | ○ | ○ | ○ | ○ | |
| 172 | Extended Adjacency Matrix SM (S=2) | ○ | ○ | ○ | ○ | |
| 173 | Extended Adjacency Matrix ISM (S=2) | ○ | ○ | ○ | ○ | |
| 174 | Extended Adjacency Matrix PM (S=2) | ○ | ○ | ○ | ○ | |
| 175 | Extended Adjacency Matrix IPM (S=2) | ○ | ○ | ○ | ○ | |
| 176 | Vertex Connectivity Matrix HM (S=2) | ○ | ○ | ○ | ○ | |
| 177 | Vertex Connectivity Matrix SM (S=2) | ○ | ○ | ○ | ○ | |
| 178 | Vertex Connectivity Matrix ISM (S=2) | ○ | ○ | ○ | ○ | |
| 179 | Vertex Connectivity Matrix PM (S=2) | ○ | ○ | ○ | ○ | |
| 180 | Vertex Connectivity Matrix IPM (S=2) | ○ | ○ | ○ | ○ | |
| 181 | Random Walk Markov HM (S=2) | ○ | ○ | ○ | ○ | |
| 182 | Random Walk Markov SM (S=2) | ○ | ○ | ○ | ○ | |
| 183 | Random Walk Markov ISM (S=2) | ○ | ○ | ○ | ○ | |
| 184 | Random Walk Markov PM (S=2) | ○ | ○ | ○ | ○ | |
| 185 | Random Walk Markov IPM (S=2) | ○ | ○ | ○ | ○ | |
| 186 | Weighted Struct. Func. IM1 HM (S=2) | ○ | ○ | ○ | ○ | |
| 187 | Weighted Struct. Func. IM1 SM (S=2) | ○ | ○ | ○ | ○ | |
| 188 | Weighted Struct. Func. IM1 ISM (S=2) | ○ | ○ | ○ | ○ | |
| 189 | Weighted Struct. Func. IM1 PM (S=2) | ○ | ○ | ○ | ○ | |
| 190 | Weighted Struct. Func. IM1 IPM (S=2) | ○ | ○ | ○ | ○ | |
| 191 | Weighted Struct. Func. IM2 HM (S=2) | ○ | ○ | ○ | ○ | |
| 192 | Weighted Struct. Func. IM2 SM (S=2) | ○ | ○ | ○ | ○ | |
| 193 | Weighted Struct. Func. IM2 ISM (S=2) | ○ | ○ | ○ | ○ | |
| 194 | Weighted Struct. Func. IM2 PM (S=2) | ○ | ○ | ○ | ○ | |
| 195 | Weighted Struct. Func. IM2 IPM (S=2) | ○ | ○ | ○ | ○ | |
| Edge-Weighted Properties | | | | | | |
| ★196 | Weighted Transitivity | | — | | — | |
| 197 | Barrat's Global Clustering Coefficient | | — | | — | |
| 198 | Onnela's Global Clustering Coefficient | | — | | — | |
| 199 | Zhang's Global Clustering Coefficient | | — | | — | |
| 200 | Holme's Global Clustering Coefficient | | — | | — | |
| Node-Weighted Properties | | | | | | |
| 201 | Total Node Weight | | | ● | ● | |
| 202 | Node Weighted Global Clustering Coeff | | | ● | ● | |
| Directed Properties | | | | | | |
| ★203 | Average In-Degree | | | | | ↗ |
| ★204 | Maximum In-Degree | | | | | ↗ |
| ★205 | Minimum In-Degree | | | | | ↗ |
| ★206 | Average Out-Degree | | | | | ↗ |
| ★207 | Maximum Out-Degree | | | | | ↗ |

| | | | | | | |
|------|--|--|--|--|--|---|
| ★208 | Minimum Out-Degree | | | | | ↗ |
| ★209 | Directed Global Clustering Coefficient | | | | | ↗ |
| 210 | Directed Flow Hierarchy | | | | | ↗ |

Table S-3 List of the edge-level descriptors covered in PROFEAT

| ID | (G12) Edge-Level Network Descriptor | Network Type | | | | |
|----|---|--------------|---------------|---------------|--------------------|-------------|
| | | Un-Directed | | | | Directed |
| | | Un-Weighted | Edge Weighted | Node Weighted | Edge-Node Weighted | Un-Weighted |
| ★1 | Edge Weight | | ○ — | | ○ — | |
| ★2 | Edge-Betweenness | ○ | ○ — | ○ | ○ — | |

Section B: Definitions and Algorithms of Network Descriptors

For a connected and undirected network, some basic matrices will be generated:

❖ Un-weighted matrix

- ♦ Adjacency matrix “ A ”, with $A_{ij}=A_{ji}=1$, if exists an edge linking node i and j . Otherwise, $A_{ij}=A_{ji}=0$.

❖ Edge-weight matrix

- ♦ Edge weight matrix “ EW ”: $EW_{ij}=EW_{ji}$ =weight between node i and j .
- ♦ Normalized edge weight matrix “ $NorEW$ ”, by the following definition.
Here, the constant factor 0.99 in the denominator is to slightly enlarge the domain from minimum value to maximum value, such that ensure the normalized minimum edge weight not to be zero.

$$NorEW_{ij} = \frac{EW_{ij} - \min\{EW\}}{\max\{EW\} - 0.99 * \min\{EW\}}.$$

❖ Node-weighted matrix

- ♦ Node weight “ NW ”: NW_i =node weight of node i , based on the input.
- ♦ Normalized node weight “ $NorNW$ ”. Again, the denominator is slightly enlarged to ensure the normalized minimum node weight not to be zero.

$$NorNW_i = \frac{NW_i - \min\{NW\}}{\max\{NW\} - 0.99 * \min\{NW\}}.$$

For a connected and directed network, directed adjacency matrix will be generated:

❖ Un-weighted matrix

- ♦ Directed adjacency matrix “ a ”, where $a_{ij}=1$, if exists a directed link from node i pointing to node j . $a_{ji}=1$ only if exists another directed link from node j pointing to node i .

The network descriptors were introduced by the IDs in **Table S-1**, **S-2**, and **S-3**. As some descriptors can be derived from either un-weighted adjacency matrix or weighted matrix, we mainly introduced the un-weighted ones, and the weighted ones can be easily obtained by substituting the algorithm with the weighted matrix.

B.1 Node-Level Descriptors

❖ Feature Category: Adjacency-based Properties

1. Degree

Degree of a node i “ deg_i ” is the number of edges linked to it.

2. Scaled Connectivity

$$scaledConnect_i = \frac{deg_i}{\max\{deg_G\}}.$$

3. Number of Selfloops

Selfloops of a node i “ $selfloop_i$ ” is the number of edges linking to itself.

4. Number of Triangles¹⁴⁶

$$tri_i = \frac{1}{2} \sum_{j=1}^N \sum_{k=1}^N A_{ij} A_{ik} A_{jk}.$$

5. Z Score^{13,200}

Z score is a connectivity index of a node, based on the degree distribution of a network. It has been applied in discovering network motifs in some studies.

$$zscore_i = \frac{deg_i - \text{avg}\{deg_G\}}{\text{dev}\{deg_G\}}.$$

6. Clustering Coefficient ^{8,9}

Clustering coefficient of a node i is defined as below, where e_i is the number of connected pairs between all neighbours of node i . It is assumed to 0, if $deg_i < 2$.

$$cluster_i = \frac{2e_i}{deg_i(deg_i - 1)}.$$

7. Neighborhood Connectivity ³⁹

The connectivity of a node is the number of its neighbours. The neighbourhood connectivity of a node i is defined as its average connectivity of all neighbours.

$$neighbourConnect_i = \frac{\sum_{j=1}^N A_{ij} \cdot deg_j}{deg_i}.$$

8. Topological Coefficient ³⁸

In calculating topological coefficient, j represents all the nodes sharing at least one neighbour with i , and $J(i, j)$ is the number of shared neighbours between i and j . If there is a direct edge between i and j , plus an additional 1 to $J(i, j)$. It is a measure to estimate the tendency of the nodes to share neighbours.

$$topology_i = avg \left\{ \frac{J(i, j)}{deg_i} \right\}.$$

9. Interconnectivity ^{57,58,59}

Firstly, the interconnectivity score is generated for each edge in the network. $N(i)$ is the neighbours of node i , such that $|N(i) \cap N(j)|$ is the number of shared neighbours between node i and node j .

$$ICN_edge_{ij} = A_{ij} \cdot \left(\frac{2 + |N(i) \cap N(j)|}{\sqrt{deg_i \cdot deg_j}} \right).$$

Next, the node's interconnectivity is calculated based on the ICN_edge scores.

$$ICN_node_i = \frac{1}{deg_i} \sum_{j=1}^N ICN_edge_{ij}.$$

10. Bridging Coefficient ²⁰¹

Bridging coefficient describes how well the node is linked between high-degree nodes.

$$bridge_i = \frac{deg_i^{-1}}{\sum_{j=1}^N A_{ij} \cdot \frac{1}{deg_j}}.$$

11. Degree Centrality ²⁰²

$$centralityDeg_i = \frac{deg_i}{N - 1}.$$

❖ Feature Category: Shortest Path Length-based Properties

12. Average Shortest Path Length ¹⁶⁰

Shortest path lengths are computed to generate an $N \times N$ matrix for storing the pairwise shortest path lengths, such that D_{ij} is the shortest path length between node i and j .

For an unweighted network, the shortest path length is basically the minimum number of edges linking between any two nodes. For an edge-weighted network, the weighted shortest path length could be generated based on the edge weight matrix. Here, $avgSPL_i$ is the average length of shorest paths between node i and all other nodes.

$$avgSPL_i = \frac{1}{N - 1} \sum_{j=1}^N D_{ij}.$$

13. Distance Sum ⁷

Distance sum is obtained by adding up all the shortest paths from node i .

$$distSum_i = \sum_{j=1}^N D_{ij}.$$

14. Eccentricity ⁷

Eccentricity is the maximum shortest path length between node i and all the other nodes.

$$eccentricity_i = \max \{D_{ij}\}.$$

15. Eccentric ⁷

Different from eccentricity measure, eccentric index is the absolute difference between the nodes' eccentricities and the graph's average eccentricity.

$$eccentric_i = |eccentricity_i - avg \{eccentricity_G\}|.$$

16. Deviation ⁷

Node's deviation measures the difference between the node's distance sum and the graph's unipolarity, which is the minimum of distance sums in the graph.

$$deviation_i = distSum_i - unipolarity_G.$$

17. Distance Deviation ⁷

$$distDev_i = |distSum_i - distAvg_G|.$$

18. Radiality ¹⁰

Radiality is computed by subtracting the average shortest path length of node i from the diameter plus 1, and the result is then divided by the network diameter. High value of radiality implies the node is generally nearer to other nodes, while a low radiality indicates the node is peripheral in the network.

$$radiality_i = \frac{diameter_G - avgSPL_i + 1}{diameter_G}.$$

19. Closeness Centrality (avg) ^{4,10,159}

The closeness centrality of a node is defined as the reciprocal of the average shortest path length. It measures how fast information spreads from a given node to other reachable nodes in the network.

$$centralityCloseAvg_i = \frac{1}{\frac{1}{N} \sum_{j=1}^N D_{ij}}.$$

20. Closeness Centrality (sum)

$$centralityCloseSum_i = \frac{1}{\sum_{j=1}^N D_{ij}}.$$

21. Eccentricity Centrality

$$centralityEccentricity_i = \frac{1}{\max\{D_{ij}\}}.$$

22. Harmonic Centrality ²⁰³

Harmonic closeness is the sum of reciprocals of the shortest path lengths for each node.

$$centralityHar_i = \sum_{j=1}^N \frac{1}{D_{ij}}.$$

23. Residual Centrality ²⁰⁴

$$centralityRes_i = \sum_{j=1}^N \frac{1}{2^{D_{ij}}}.$$

24. Load Centrality ^{2,10}

The load centrality (stress centrality) of node i is the fraction of all shortest paths that passing through it. A node has a high load centrality if it is involved in a high number of shortest paths.

25. Betweenness Centrality ^{10,11}

The betweenness centrality quantifies the number of times a node serving as a linking bridge along the shortest path between two other nodes. It is computed by the following equation, where s , t , $\sigma_{st}(v)$ are defined as same as the previous stress centrality, and σ_{st} is the number of shortest paths from s to t . The betweenness centrality reflects the extent of control of that node exerting over the interactions with other nodes in the network.

$$centralityBtw_i = \frac{\sum_{s \neq i \neq t} \sigma_{st}(i)}{\sigma_{st}}.$$

26. Normalized Betweenness Centrality

$$centralityBtwNor_i = \frac{centralityBtw_i - \min\{centralityBtw_G\}}{\max\{centralityBtw_G\} - \min\{centralityBtw_G\}}.$$

27. Bridging Centrality ²⁰¹

The bridging centrality of a node is the product of the bridging coefficient and the betweenness centrality. A higher bridging centrality means more information flowing through that node.

$$centralityBridge_i = bridge_i \cdot centralityBtw_i.$$

28. Current Flow Betweenness ^{205,206,207}

Previously, the betweenness centrality is based on the shortest path length in the network. Here, the current flow betweenness centrality is assumed that information efficiently spreading in the network like an electrical current, as a current flow analog.

Firstly, the resistance R of an edge is defined, where $r(e)=1/w(e)$ and $w(e)$ is the weight of an edge e . For unweighted networks, $w(e)=1$ for all edges.

Secondarily, a vector b , namely supply, is defined where current enters and leaves the network. Since there should be as much current entering as leaving the network, $\sum b(v)=0$.

$$b_{st}(v) = \begin{cases} I, & v=s \\ -I, & v=t \\ 0, & \text{otherwise} \end{cases}$$

Thirdly, the electrical current c is defined and it should follow the law below.

Kirchhoff's Current Law (for every $v \in V$):

$$\sum_{(v,w) \in E} c(v,w) - \sum_{(u,v) \in E} c(u,v) = b(v).$$

Kirchhoff's Potential Law (for every current cycle $e_1 \dots e_k$ in the network):

$$\sum_{i=1}^k c(e_i) = 0.$$

Lastly, the potential difference p is defined by *Ohm's Law*, where $p(e)=c(e)/r(e)$. To calculate the current flow betweenness, throughput $\tau(v)$ of a node v , and throughput $\tau(e)$ of an edge e are defined:

$$\tau(v) = \frac{1}{2} \left(-|b(v)| + \sum_e |c(e)| \right).$$

$$\tau(e) = |c(e)|.$$

Therefore, current flow betweenness (sometimes also called random-walk betweenness) is then defined, where τ_{st} denotes the throughput of a s - t current, and $N_b = (N-1)(N-2)$.

$$CFbetween_i = \frac{1}{N_b} \sum_{s,t \in V} \tau_{st}(i).$$

29. Current Flow Closeness^{205,206,207}

The current flow closeness centrality is a variant of the current flow betweenness centrality, by using the analog of shortest path length in electrical networks.

$$CFclose_i = \frac{N_c}{\sum_{s \neq t} p_{st}(s) - p_{st}(t)}.$$

Where, $N_c = (N-1)$, and $p_{st}(s)-p_{st}(t)$ denotes the effective resistance of s - t current, interpreted as an alternative measure of distance between node s and node t .

❖ **Feature Category: Eigenvector-based Centrality Indices**

30. Eigenvector Centrality ^{208,209}

Eigenvector centrality is the eigenvalue-based methods to approximate the importance of each node in a network. It assumes that each node's centrality is the sum of its neighbors' centrality values, which is saying that an important node should be linking to important neighbors.

In its definition algorithm, the eigenvector centralities for all nodes are initialized to 1 at the beginning, and then an eigenvalue-based function is applied to iteratively converge the centrality to a fixed value, by considering the neighbourhood relationships and the neighbors' centrality values. Let $\{\lambda_1, \lambda_2 \dots \lambda_k\}$ be the non-zero eigenvalues of adjacency matrix of the network, and λ_{max} is the maximum eigenvalue.

$$centralityEigen_i = \frac{1}{\lambda_{max}} \sum_{j=1}^N A_{ij} \cdot centralityEigen_j.$$

31. PageRank Centrality ^{16,17,18,19,20,21}

PageRank is an algorithm implemented in Google search engine to rank the websites, according to the webpage connections in the World Wide Web. It is a variant of eigenvector centrality, by initializing the PageRank centralities to an equal probability value $1/N$ for all nodes.

]The equation below will iteratively update the node centrality value by using a constant damping factor d , its neighbors' PageRank centrality value, and its degree. The algorithm stops running, when the PageRank centrality converges, and the constant damping factor d is generally assumed to 0.85.

$$pageRank_i = \frac{1-d}{N} + d \cdot \sum_{j=1}^N A_{ij} \cdot \frac{pageRank_j}{deg_j}.$$

❖ **Feature Category: Edge-Weighted Properties**

32. Strength ²¹⁰

The node's strength is the sum of all the edge weights connected to that node.

$$strength_i = \sum_{j=1}^N A_{ij} \cdot W_{ij}.$$

33. Assortativity ^{210,211}

In an unweighted graph, assortativity is as the same as the neighbourhood connectivity. For a weighted graph, it is defined as below.

$$assortativity_i = \frac{1}{strength_i} \sum_{j=1}^N W_{ij} \cdot deg_j.$$

34. Disparity ²¹²

$$disparity_i = \sum_{j=1}^N \left(\frac{A_{ij} \cdot W_{ij}}{strength_i} \right)^2.$$

35. Geometric Mean of Triangles ¹⁴⁶

$$geo_tri_i = \frac{1}{2} \sum_{j=1}^N \sum_{k=1}^N \sqrt[3]{W_{ij} W_{ik} W_{jk}}.$$

36. Barrat's Local Clustering Coefficients ⁶¹

$$clusterBarrat_i = \frac{1}{strength_i(deg_i - 1)} \sum_{j=1}^N \sum_{k=1}^N \left(A_{ij} A_{ik} A_{jk} \cdot \frac{W_{ij} + W_{ik}}{2} \right).$$

37. Onnela's Local Clustering Coefficients ^{61,213}

$$clusterOnnela_i = \frac{1}{deg_i \cdot (deg_i - 1)} \sum_{j=1}^N \sum_{k=1}^N (\widehat{W}_{ij} \widehat{W}_{ik} \widehat{W}_{jk})^{1/3}.$$

$$\widehat{W}_{ij} = \frac{W_{ij}}{\max\{W\}}.$$

38. Zhang's Local Clustering Coefficients ^{60,61}

$$clusterZhang_i = \frac{\sum_{j=1}^N \sum_{k=1}^N \widehat{W}_{ij} \widehat{W}_{ik} \widehat{W}_{jk}}{(\sum_{k=1}^N \widehat{W}_{ij})^2 - \sum_{k=1}^N \widehat{W}_{ij}^2}.$$

39. Holme's Local Clustering Coefficients ^{61,214}

$$clusterHolme_i = \frac{\sum_{j=1}^N \sum_{k=1}^N \widehat{W}_{ij} \widehat{W}_{ik} \widehat{W}_{jk}}{\max\{W\} \cdot \sum_{j=1}^N \sum_{k=1}^N \widehat{W}_{ij} \widehat{W}_{ik}}.$$

40. Edge-Weighted Interconnectivity ⁵⁸

The edge-weighted interconnectivity is defined similarly with the unweighted interconnectivity. Firstly, the interconnectivity score for each edge is calculated. Where, W_{ij} is the weight of the edge linking node i and node j , and the previously defined $strength_i$ is the sum of weights of connected edges to node i .

$$EW_{ICN_{edge_{ij}}} = \frac{2W_{ij} + \sum_{u \in N(i) \cap N(j)} W_{iu} W_{ju}}{\sqrt{strength_i \cdot strength_j}}.$$

Next, the edge-weighted interconnectivity for each node is calculated based on EW_ICN_edge scores.

$$EW_{ICN_{node_i}} = \frac{1}{deg_i} \sum_{j=1}^N EW_{ICN_{edge_{ij}}}.$$

❖ Feature Category: Node-Weighted Properties

41. Node Weight

Node weight NW_i is directly extracted from the node weight matrix generated.

42. Node Weighted Cross Degree ¹⁵

For analyzing networks with heterogeneous node weights, the next two node-weighted measures were derived from the recent economic trading network study, where $ExtA$ is the extended adjacency matrix " $ExtA_{ij}=A_{ij}+\delta_{ij}$ " and δ_{ij} is Kronecker's delta constant.

$$\delta_{ij} = \begin{cases} 0, & \text{if } i \neq j \\ 1, & \text{if } i = j \end{cases}$$

$$NW_{crossdeg_i} = \sum_{j=1}^N ExtA_{ij} \cdot NW_j.$$

43. Node Weighted Local Clustering Coefficient ¹⁵

This node-weighted local clustering coefficient works, only if the node-weighted cross degree is not zero, otherwise the local clustering coefficient will be assumed as zero.

$$NWcluster_i = \frac{1}{NWcrossdeg_i^2} \sum_{j=1}^N \sum_{k=1}^N ExtA_{ij} \cdot NW_j \cdot ExtA_{ik} \cdot NW_k \cdot ExtA_{jk}.$$

44. Node-Weighted Neighbourhood Score ⁵⁷

This score was defined in a disease-gene network study, by assigning the fold change of gene expression as the node weight. $neighbour(i)$ denotes all neighbours of node i .

$$NWneighbourhood_i = \frac{1}{2}NW_i + \frac{1}{2} \cdot \frac{\sum_{j \in neighbour(i)} NW_j}{|neighbour(i)|}.$$

❖ Feature Category: Directed Properties

45. In-Degree ^{41,146}

As previously mentioned, “ A ” represents the undirected adjacency matrix and “ a ” represents the directed adjacency matrix. $a_{ij}=1$ means a directed edge pointing from node i to j . In-degree of a node counts the number of directed edges pointing to itself.

$$deg_i^+ = \sum_{j \in N} a_{ji}.$$

46. In-Degree Centrality

The in-degree centrality for a node is the fraction of nodes its incoming edges are connected to.

47. Out-Degree ^{41,146}

Out-degree of a node counts the number of directed edges pointing out of itself.

$$deg_i^- = \sum_{j \in N} a_{ij}.$$

48. Out-Degree Centrality

The out-degree centrality for a node is the fraction of nodes its outgoing edges are connected to.

49. Directed Local Clustering Coefficient ⁴¹

In directed networks, local clustering coefficient is defined as below.

$$cluster_i^\pm = \frac{e_i}{(deg_i^+ + deg_i^-)(deg_i^+ + deg_i^- - 1)}.$$

50. Neighbourhood Connectivity (only in) ⁴¹

It is the average out-connectivity of all in-neighbours of node i.

$$neighbourConnectivity_i^+ = \frac{\sum_{j \in N} a_{ji} \cdot deg_j^-}{\sum_{j \in N} a_{ji}}.$$

51. Neighbourhood Connectivity (only out) ⁴¹

It is the average in-connectivity of all out-neighbours of node i.

$$neighbourConnectivity_i^- = \frac{\sum_{j \in N} a_{ij} \cdot deg_j^+}{\sum_{j \in N} a_{ij}}.$$

52. Neighbourhood Connectivity (in & out) ⁴¹

It is the average connectivity of all neighbours for each node.

$$neighbourConnectivity_i^\pm = \frac{\sum_{j \in N} a_{ij} \cdot (deg_j^+ + deg_j^-) + \sum_{j \in N} a_{ji} \cdot (deg_j^+ + deg_j^-)}{\sum_{j \in N} a_{ji} + \sum_{j \in N} a_{ij}}.$$

53. Average Directed Neighbour Degree ¹⁴⁶

$$avgDirectedNeighbourDeg_i^\pm = \frac{\sum_{j \in N} [(a_{ij} + a_{ji}) \cdot (deg_j^+ + deg_j^-)]}{2 \cdot (deg_j^+ + deg_j^-)}.$$

B.2 Network-Level Descriptors

❖ Feature Category: Connectivity/Adjacency-based Properties

1. Number of Nodes

The number of the nodes (or vertices) in the network, noted as N .

2. Number of Edges

The number of edges (or links) in the network, noted as E .

3. Number of Selfloops

$$selfloops_G = \sum_{i=1}^N selfloop_i .$$

4. Maximum Connectivity

$$connectivityMax_G = \max\{deg_G\} .$$

5. Minimum Connectivity

$$connectivityMin_G = \min\{deg_G\} .$$

6. Average Number of Neighbors

The average of the number of neighbours (or degree, connectivity) for all nodes.

$$neighbourAvg_G = \frac{1}{N} \sum_{i=1}^N deg_i .$$

7. Total Adjacency ¹⁴⁹

The total adjacency is the half of the sum of the adjacency matrix entries.

$$totalAdjacency_G = \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N A_{ij} .$$

8. Network Density ¹⁴⁹

The network density measures the efficiency of the information progression in a network in time. The denominator $N*(N-1)/2$ is the maximum number of links if the network is completely connected. For a directed network, the denominator is $N*(N-1)$.

$$density_G = \frac{E}{N(N-1)/2}.$$

9. Global Clustering Coefficient ^{8,9}

Network clustering coefficient is the average of all the node-level clustering coefficients.

$$cluster_G = \frac{1}{N} \sum_{i=1}^N cluster_i.$$

10. Transitivity ¹⁴⁶

Transitivity is calculated based on the number of triangles for each node in the network.

$$transitivity_G = \frac{2 * \sum_{i=1}^N tri_i}{\sum_{i=1}^N deg_i(deg_i - 1)}.$$

11. Heterogeneity ¹³⁷

Heterogeneity measures the variation of degree distribution, reflecting the tendency of a network to have hubs. This index is biologically meaningful, as biological networks are usually heterogeneous with some central nodes highly connected and the rest nodes having few connections in the network.

$$heterogeneity_G = \sqrt{\frac{N \cdot \sum_{i=1}^N (deg_i^2)}{(\sum_{i=1}^N deg_i)^2} - 1}.$$

12. Degree Centralization ¹³⁷

Degree centralization (or sometimes called as, connectivity centralization) is useful for distinguishing such characteristics as highly connected networks (e.g. star-shaped) or decentralized networks, which have been used for studying the structural differences of metabolic networks.

$$centralizationDeg_G = \frac{N}{N-2} \left(\frac{connectivityMax_G}{N-1} - density_G \right).$$

13. Central Point Dominance²¹⁵

Central point dominance is defined based on the measure of betweenness centrality.

$$centralDominance_G = \frac{1}{N-1} \sum_{i=1}^N (\max\{centralityBtw_i\} - centralityBtw_i).$$

14. Degree Assortativity Coefficient²¹⁶

It measures the similarity of degree with respect to each edge in the network, by calculating the standard Pearson correlation coefficient between the degrees of the two connecting vertices of each edge. Its value lies in between -1 and 1, where 1 represents perfect assortativity and -1 indicates perfect dissortativity.

❖ Feature Category: Shortest Path Length-based Properties

15. Total Distance¹⁴⁹

It is the sum of all the non-redundant pairwise shortest path distances in the network.

$$totalDistance_G = \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N D_{ij}.$$

16. Network Diameter⁴¹

The network diameter is the largest distance in shortest path length matrix.

$$diameter_G = \max\{D_{ij}\}.$$

17. Network Radius⁴¹

The network radius is the smallest distance in shortest path length matrix.

$$radius_G = \min\{D_{ij}\}.$$

18. Shape Coefficient²¹⁷

The shape coefficient of a network is defined by its radius and its diameter.

$$shapeCoef_G = \frac{diameter_G - radius_G}{radius_G}.$$

19. Characteristic Path Length ⁴¹

The characteristic path length is the average distance in shortest path length matrix.

$$CPL_G = \frac{\sum_{i=1}^N avgSPL_i}{N}.$$

20. Network Eccentricity ⁷

$$eccentricity_G = \sum_{i=1}^N eccentricity_i.$$

21. Average Eccentricity ⁷

$$eccentricityAvg_G = \frac{eccentricity_G}{N}.$$

22. Network Eccentric ⁷

$$eccentric_G = \frac{1}{N} \sum_{i=1}^N eccentric_i.$$

23. Eccentric Connectivity ²¹⁸

This index is defined as the sum of the product of eccentricity and degree of each node, it has been shown the high correlation with regard to physical properties of diverse nature in various datasets.

$$eccentricConnect_G = \sum_{i=1}^N eccentric_i \cdot deg_i.$$

24. Unipolarity ⁷

It measures the minimal distance sum (sum of shortest path lengths) value.

$$unipolarity_G = \min\{distSum_i\}.$$

25. Integration ⁷

Network integration is the sum of all nodes' distance sum.

$$integration_G = \frac{1}{2} \sum_{i=1}^N distSum_i.$$

26. Variation ⁷

The network variation is defined as the maximum variance in the node-level measures.

$$variation_G = \max\{deviation_i\}.$$

27. Average Distance ⁷

This measures the mean shorest path length by dividing the integration by the number of nodes.

$$distAvg_G = \frac{2 \cdot integration_G}{N}.$$

28. Mean Distance Deviation ⁷

This mean distance deviation is to average the node-level distance deviation values.

$$distDevMean_G = \frac{1}{N} \sum_{i=1}^N distDev_i.$$

29. Centralization ⁷

This centralization sums the variance value for all nodes in the network.

$$centralization_G = \sum_{i=1}^N deviation_i.$$

30. Global Efficiency ¹²

The global efficiency is a measure of the information exchange efficiency across the entire network. It can be used to determine the cost-effectiveness of the network structure.

$$efficiency_G = \frac{1}{N(N-1)} \sum_{i \neq j}^N \frac{1}{D_{ij}}.$$

❖ **Feature Category: Topological Indices**

31. Edge Complexity Index ¹⁴⁹

The global edge complexity is defined by dividing the total adjacency by N^2 .

$$edgeComplexity_G = \frac{totalAdjacency_G}{N^2}.$$

32. Randic Connectivity Index ²¹⁹

The randic index is a function of the connectivity of edges.

$$randic_G = \sum_{E_{i,j} \in G} (deg_i \cdot deg_j)^{-\frac{1}{2}}.$$

33. Atom-Bond Connectivity Index ²²⁰

The ABC index is a graph-invariant measure, which has been applied to study the stability of chemical structure. Here, it is used to describe the stability of a network structure.

$$ABC_G = \sum_{E_{i,j} \in G} \left(\frac{deg_i + deg_j - 2}{deg_i \cdot deg_j} \right)^{\frac{1}{2}}.$$

34. Zagreb Index 1 ^{221,222,223,224}

There are five Zagreb indices variants are defined based on the nodes' degree.

$$zagreb1_G = \sum_{i=1}^N deg_i^2.$$

35. Zagreb Index 2

$$zagreb2_G = \sum_{E_{i,j} \in G} deg_i \cdot deg_j.$$

36. Modified Zagreb Index

$$zagrebModified_G = \sum_{E_{i,j} \in G} \frac{1}{deg_i \cdot deg_j}.$$

37. Augmented Zagreb Index

$$zagrebAugmented_G = \sum_{E_{i,j} \in G} \left(\frac{deg_i \cdot deg_j}{deg_i + deg_j - 2} \right)^3.$$

38. Variable Zagreb Index

$$zagrebVariable_G = \sum_{E_{i,j} \in G} \frac{deg_i + deg_j - 2}{deg_i \cdot deg_j}.$$

39. Narumi-Katayama Index ²²⁵

The NK index is the product of degrees of all nodes. It has been shown the relationships with thermodynamics properties. Additionally, its logged index, geometric index, and harmonic Index are provided as follows. In our program, if Narumi index goes beyond *sys.maxsize*, then Narumi Index and Narumi Geometric Index will be assigned as zero.

$$narumi_G = \prod_{i=1}^N deg_i.$$

40. Narumi-Katayama Index (log)

$$narumiLog_G = \log_2 \left(\prod_{i=1}^N deg_i \right).$$

41. Narumi Geometric Index ²²⁶

$$narumiGeo_G = \left(\prod_{i=1}^N deg_i \right)^{\frac{1}{N}}.$$

42. Narumi Harmonic Index ²²⁶

$$narumiHar_G = \frac{N}{\sum_{i=1}^N (deg_i)^{-1}}.$$

43. Alpha Index ¹³

Alpha index (Meshedness Coefficient) is a connectivity to evaluate the number of cycles in a network in comparison with maximum number of cycles, such that the higher alpha index, the more connected nodes. Trees and simple graphs have alpha index=0, and a completely connected network have alpha index=1.

$$alpha_G = \frac{E - N}{\frac{N(N-1)}{2} - (N-1)}.$$

44. Beta Index ¹³

It measures the network connectivity, by the ratio of the number of edges over the number of nodes. Simple networks have beta value less than 1, and more complex networks have higher beta index.

$$beta_G = \frac{E}{N}.$$

45. Pi Index ¹³

Pi index is the relationship between the total length and the diameter, having a similar meaning with the definition of π , indicating of the shape of the network.

$$pi_G = \frac{\sum_{i=1}^N \sum_{j=1}^N A_{ij}}{diameter_G}.$$

46. Eta Index ¹³

Eta index is the average adjacency per edge. Adding nodes reduce the eta index.

$$eta_G = \frac{\sum_{i=1}^N \sum_{j=1}^N A_{ij}}{E}.$$

47. Hierarchy ¹³

Hierarchy index is the gradient of the linear power-law regression, by fitting $\log_{10}(\text{node frequency})$ over $\log_{10}(\text{degree distribution})$. It usually has the value between 1 and 2, where the low hierarchy indicates the weak hierarchical relationship. Hierarchy is notated as h in the fitted regression equation $y=ax^h$, where x is the degree distribution and y is the node frequency of that degree.

$$y = a \cdot x^{hierarchy}.$$

48. Robustness ⁴⁰

Robustness is to measure the stability of a network under node-removal attacks. By removing each node, the size of the largest fragmented component S is used to define the robustness.

$$robustness_G = \frac{\sum_{k=1}^N S_k}{N(N-1)}.$$

49. Medium Articulation ^{150,153,227}

Medium articulation MA is a complexity measure of a network, reaching its maximum with medium number of edges. It is defined based on the redundancy (MA_R) and the mutual information (MA_I).

$$MA_G = MA_R \cdot MA_I .$$

Redundancy MA_R is defined as:

$$MA_R = 4 \left(\frac{R - R_{path}}{R_{clique} - R_{path}} \right) \left(1 - \frac{R - R_{path}}{R_{clique} - R_{path}} \right) .$$

$$R = \frac{1}{E} \sum_{i=1}^N \sum_{j>i}^N \log_{10}(deg_i \cdot deg_j) .$$

$$R_{clique} = 2 \cdot \log_{10}(N - 1) .$$

$$R_{path} = 2 \cdot \frac{N - 2}{N - 1} \log_{10} 2 .$$

Mutual information MA_I is defined as:

$$MA_I = 4 \left(\frac{I - I_{clique}}{I_{path} - I_{clique}} \right) \left(1 - \frac{I - I_{clique}}{I_{path} - I_{clique}} \right) .$$

$$I = \frac{1}{E} \sum_{i=1}^N \sum_{j>i}^N \log_{10} \frac{2 E}{deg_i \cdot deg_j} .$$

$$I_{clique} = \log_{10} \left(\frac{N}{N - 1} \right) .$$

$$I_{path} = \log_{10}(N - 1) - \frac{N - 3}{N - 1} \log_{10} 2 .$$

50. Complexity Index A ¹⁴⁹

It is the ratio of total adjacency and the total distance of a network.

$$complexityA_G = \frac{totalAdjacency_G}{totalDistance_G} .$$

51. Complexity Index B ¹⁴⁹

It is defined by the ratio of vertex degree and its distance sum for each vertex.

$$complexityB_G = \sum_{i=1}^N \frac{deg_i}{distSum_i} .$$

52. Wiener Index ²²⁸

Wiener index measures the sum of the shortest path lengths between all pairs of vertices.

$$wiener_G = \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N D_{ij}.$$

53. Hyper-Wiener Index ²²⁹

$$hyperWiener_G = \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N (D_{ij}^2 + D_{ij}).$$

54. Harary Index 1 ²³⁰

$$harary1_G = \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N D_{ij}^{-1}.$$

55. Harary Index 2 ²³⁰

$$harary2_G = \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N D_{ij}^{-2}.$$

56. Compactness ²³¹

This is based on Wiener index, by dividing the Wiener index by $N(N-1)$.

$$compactness_G = \frac{4 \cdot wiener_G}{N(N-1)}.$$

57. Superpendentic Index ²³²

$$superpendentic_G = \left(\sum_{i=1}^N \sum_{j=1}^N D_{ij} \right)^{\frac{1}{2}}.$$

58. Hyper-Distance-Path Index ^{233,234}

This index is consist of two parts: the exactly Wiener index, and the delta number.

$$hyper_path_G = \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N D_{ij} + \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \binom{D_{ij}}{2}.$$

59. BalabanJ Index ²³⁵

This BalabanJ index counts into the distance sum of the two end-vertex for each edge. BalabanJ index has been proven to be relevant to the network branching. There are another three differently defined variants of BalabanJ indices are given in the followings.

$$Jm_G = \frac{E}{\mu + 1} \sum_{E_{i,j} \in G} (disSum_i \cdot disSum_j)^{-\frac{1}{2}}.$$

Where, $\mu = E + I - N$, which denotes the cyclomatic number of a graph.

60. BalabanJ-Like Index 1 ²³⁶

$$Jm1_G = \frac{E}{\mu + 1} \sum_{E_{i,j} \in G} (disSum_i \cdot disSum_j)^{\frac{1}{2}}.$$

61. BalabanJ-Like Index 2 ²³⁶

$$Jm2_G = \frac{E}{\mu + 1} \sum_{E_{i,j} \in G} (disSum_i + disSum_j)^{\frac{1}{2}}.$$

62. BalabanJ-Like Index 3 ²³⁶

$$Jm3_G = \frac{E}{\mu + 1} \sum_{E_{i,j} \in G} \left(\frac{disSum_i \cdot disSum_j}{disSum_i + disSum_j} \right)^{\frac{1}{2}}.$$

63. Geometric Arithmetic Index 1 ^{222,237}

GA index consists of the geometrical and the arithmetic means of the end-to-end degree of an edge.

$$GA1_G = \sum_{E_{i,j} \in G} \frac{2\sqrt{deg_i \cdot deg_j}}{deg_i + deg_j}.$$

64. Geometric Arithmetic Index 2 ^{222,237}

There are 2 extended geometric-arithmetic indices, which make use of the information of the shortest path lengths. In some studies, the geometric-arithmetic indices have shown its power in characterizing the network structure features.

$$GA2_G = \sum_{E_{i,j} \in G} \frac{2\sqrt{n_i \cdot n_j}}{(n_i + n_j)}.$$

$$n_i := |x \in node(G), D_{xi} < D_{xj}|.$$

$$n_j := |x \in node(G), D_{xj} < D_{xi}|.$$

In the definition of geometric arithmetic index 2 (GA2), x is a node, n_i is the number of nodes closer to node i , and n_j is the number of nodes closer to node j , while the nodes with same distance to node i and node j are ignored.

65. Geometric Arithmetic Index 3 ^{222,237}

$$GA3_G = \sum_{E_{i,j} \in G} \frac{2\sqrt{m_i \cdot m_j}}{(m_i + m_j)}.$$

$$m_i := |y \in edge(G), D_{yi} < D_{yj}|.$$

$$m_j := |y \in edge(G), D_{yi} < D_{yj}|.$$

In the definition of geometric arithmetic index 3 (GA3), y is an edge in the graph, the distance between edge y to node i is defined as $D_{yi} = \min \{D_{pi}, D_{qi}\}$, where p and q are the two ends of edge y . In the context above, m_i is number of edges closer to node i and m_j is the number of edges closer to node j , while the edges with same distance to node i and node j are not counted.

66. Szeged Index ²³⁸

$$szeged_G = \sum_{E_{i,j} \in G} n_i \cdot n_j.$$

Where n_i and n_j are as defined as the previous geometric-arithmetic index 2.

67. Product of Row Sums ²³⁹

If PRS is greater than `sys.maxsize`, it will be assigned as zero in the program.

$$PRS_G = \prod_{i=1}^N distSum_i.$$

68. Product of Row Sums (log)

$$PRSLog_G = \log_2 \left(\prod_{i=1}^N distSum_i \right).$$

69. Schultz Topological Index ²⁴⁰

By using adjacency matrix A , shorest path distance matrix D , and the vertex degree vector v , Schultz defined a topological index to described the network structure. In the equation below, $(D+A)$ forms an addictive $N \times N$ matrix, and this matrix is then multiplied by a $1 \times N$ vector v , such that obtaining another $1 \times N$ vector. The sum of all the elements in the resultant vector is called the Schultz topological index.

$$schultz_G = \sum_{i=1}^N [v(D + A)]_i .$$

70. Gutman Topological Index ²⁴¹

Gutman topological index is a further defined, where ADA is matrix multiplication.

$$gutman_G = \sum_{i=1}^N \sum_{j=1}^N [ADA]_{ij} .$$

71. Efficiency Complexity ^{150,153,227}

Efficiency complexity is motivated in analyzing the weighted networks, as it suggests to measure not only the shortest path lengths but also the cost (number of links).

$$EC_G = 4 \left(\frac{E - E_{path}}{1 - E_{path}} \right) \left(1 - \frac{E - E_{path}}{1 - E_{path}} \right) .$$

$$E = \frac{2}{N(N-1)} \sum_{i=1}^N \sum_{j>i}^N \frac{1}{D(i,j)} .$$

$$E_{path} = \frac{2}{N(N-1)} \sum_{i=1}^N \left(N - \frac{N-i}{i} \right) .$$

❖ **Feature Category: Entropy-Based Complexity Indices**

72. Information Content (Degree Equality) ^{163,242}

This information content measures the probability distribution of vertex degree, where N_i^d is the number of nodes having the same degree, and k^d is the maximum of degree.

$$I_{vertexDegree} = - \sum_{i=1}^{k^d} \frac{N_i^d}{N} \cdot \log_2 \left(\frac{N_i^d}{N} \right).$$

73. Information Content (Edge Equality) ²⁴³

This measure is based on the probability distribution of edge connectivity, where each edge has an end-to-end connectivity value. Let (a, b) and $a \leq b$ be the edge's end-to-end connectivity, such that the edges having the same edge connectivity will be grouped into the same subset.

$$I_{edgeEquality} = - \sum_{i=1}^{k^{edge}} \frac{E_i}{E} \cdot \log_2 \left(\frac{E_i}{E} \right).$$

Where, E_i is the number of edges having the same end-to-end connectivity, and k^{edge} is the number of different edge subsets.

74. Information Content (Edge Magnitude) ²⁴³

As another measure based on the edge information, it is defined by the connectivity magnitude of each edge, and $randic_G$ is the network-level randic connectivity index introduced previously.

$$I_{edgeMagnitude} = - \sum_{E_{i,j} \in G} \frac{(deg_i \cdot deg_j)^{-1/2}}{randic_G} \cdot \log_2 \left(\frac{(deg_i \cdot deg_j)^{-1/2}}{randic_G} \right).$$

75. Information Content (Distance Degree) ¹⁶³

The distance degree of a node i is equivalent to the $distSum_i$ defined previously.

$$I_{distanceDegree} = - \sum_{i=1}^N \frac{distSum_i}{2 \cdot Wiener_G} \cdot \log_2 \left(\frac{distSum_i}{2 \cdot Wiener_G} \right).$$

76. Information Content (Distance Degree Equality) ¹⁶³

The probability distribution regarding on the nodes' distance degree value gives the definition of the mean information content on distance degree equality. In the equation below, k^{dd} is the number of node groups in the distribution of distance degree, N^{dd}_i is the number of nodes having the same distance degree.

$$I_{distanceDegreeEquality} = - \sum_{i=1}^{k^{dd}} \frac{N^{dd}_i}{N} \cdot \log_2 \left(\frac{N^{dd}_i}{N} \right).$$

77. Radial Centric Information Index ^{163,242}

Radial centric information measures the probability distribution of vertex eccentricity, where N^e_i is the number of nodes having the equal eccentricity value i , and k^e is the maximum of eccentricity.

$$I_{radialCentric} = - \sum_{i=1}^{k^e} \frac{N^e_i}{N} \cdot \log_2 \left(\frac{N^e_i}{N} \right).$$

78. Distance Degree Compactness ²⁴⁴

This measure is defined based on the distribution of nodes' locations from the centre of a network, where the centre is determined by the closeness centrality score in this case. Here, Q_k is the sum of distance degree of all nodes that located at the same topological distance k from the centre.

$$I_{compactness} = 2Weiner_G \cdot \log_2(2Weiner_G) - \sum_k Q_k \cdot \log_2(Q_k).$$

79. Distance Degree Centric Index ²⁴⁵

$$I_{distanceDegreeCentric} = - \sum_{i=1}^{K^c} \frac{N_i}{N} \log_2 \frac{N_i}{N}.$$

Where N_i is the number of nodes having the same eccentricity and the same degree, K^c is the number of equivalent classes of N_i .

80. Graph Distance Complexity ²⁴⁶

Similar as $I_{infoLayer}$, this distance complexity includes the nodes' distance sums.

$$I_{distanceComplexity} = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^{ecc_i} N_j^i \cdot \frac{j}{distSum_i} \cdot \log_2 \left(\frac{j}{distSum_i} \right).$$

81. Information Layer Index ²⁴⁷

$$I_{infoLayer} = -\sum_{i=1}^N \sum_{j=1}^{ecc_i} \frac{N_j^i}{N} \cdot \log_2 \left(\frac{N_j^i}{N} \right).$$

In the equation, ecc_i is the eccentricity value of node i , and N_j^i is the number of nodes in the j^{th} sphere of node i . In other words, N_j^i is the number of nodes in shorest distance j away from node i .

82. Bochev Information Index 1 ^{248,249}

Bochev indices applies the probability distribution of the shortest path lengths to the Shannon's entropy formula, and it has three different variants.

$$I_{bochev1} = -\frac{1}{N} \cdot \log_2 \left(\frac{1}{N} \right) - \sum_{i=1}^{diameter_G} \frac{2k_i}{N^2} \cdot \log_2 \left(\frac{2k_i}{N^2} \right).$$

Where $diameter_G$ is the maximum distance between two nodes in the network, and k_i is the occurrence of distance i in the shortest path length matrix D_{ij} .

83. Bochev Information Index 2 ^{248,249}

$$I_{bochev2} = -Weiner_G \cdot \log_2(Weiner_G) - \sum_{i=1}^{diameter_G} i \cdot k_i \cdot \log_2(i).$$

84. Bochev Information Index 3 ^{248,249}

$$I_{bochev3} = -\sum_{i=1}^{diameter_G} \frac{2k_i}{N(N-1)} \cdot \log_2 \left(\frac{2k_i}{N(N-1)} \right).$$

85. Balaban-like Information Index 1 ^{152,250}

BalabanJ indices are defined by the distance degree of each node. Balaban-like information index 1 & 2 are defined based on the distribution of distance degree.

$$I_{balaban1} = \frac{E}{\mu + 1} \sum_{E_{i,j} \in G} [u_i \cdot u_j]^{-1/2}.$$

$$u_i = - \sum_{k=1}^{dimeter} \frac{k \cdot g_k}{distSum_k} \cdot \log_2 \left(\frac{k}{distSum_k} \right).$$

$$\mu = E + 1 - N.$$

Where g_k is the number of nodes at distance k from node i .

86. Balaban-like Information Index 2 ^{152,250}

$$I_{balaban2} = \frac{E}{\mu + 1} \sum_{E_{i,j} \in G} [v_i \cdot v_j]^{-1/2}.$$

$$v_i = distSum_i \cdot \log_2(distSum_i) - u_i.$$

❖ Feature Category: Eigenvalue-Based Complexity Indices

87. Graph Energy ¹⁶²

Given a network, let $\{\lambda_1, \lambda_2 \dots \lambda_k\}$ be the non-zero eigenvalues of its adjacency matrix, such that k is the number of eigenvalues and λ_{max} is the maximum of the eigenvalues.

$$Energy_G = \sum_{i=1}^k |\lambda_i|.$$

88. Laplacian Energy ¹⁶²

Laplacian matrix L_{ij} is generated based on the degree and the adjacency relationships as below, producing $\mu_i : \{\mu_1, \mu_2 \dots, \mu_k\}$ as the Laplacian eigenvalues of the network.

$$L_{ij} = \begin{cases} -1 & \text{if } A_{ij} = 1 \\ deg_i & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

$$LaplacianEnergy_G = \sum_{i=1}^k \left| \mu_i - \frac{2E}{N} \right|.$$

89. Spectral Radius ²⁵¹

$$SpRadius_G = \max\{|\lambda_i|\}.$$

90. Estrada Index ²⁵²

$$Estrada_G = \sum_{i=1}^k e^{\lambda_i}.$$

91. Laplacian Estrada Index ²⁵³

$$LaplacianEstrada_G = \sum_{i=1}^k e^{\mu_i}.$$

92. Quasi-Wiener Index ²⁵⁴

Quasi-Wiener is defined by Laplacian eigenvalues. The last eigenvalue μ_k is excluded as always being zero.

$$quasiWeiner_G = N \sum_{i=1}^{k-1} \frac{1}{\mu_i}.$$

93. Mohar Index 1 ^{234,255}

$$mohar1_G = \frac{1}{N} \cdot quasiWeiner_G \cdot \log_2 \left(\sum_{i=1}^{k-1} \mu_i \right).$$

94. Mohar Index 2 ^{234,255}

$$mohar2_G = \frac{4}{N \cdot \mu_{k-1}}.$$

95. Graph Index Complexity ¹⁵⁰

$$Cr_G = 4 \cdot cr \cdot (1 - cr).$$

$$cr = \frac{\lambda_{max} - 2 \cos \frac{\pi}{N+1}}{N - 1 - 2 \cos \frac{\pi}{N+1}}.$$

96 - 195. A Set of Eigenvalue-Based Descriptors ^{156,256}

There are 5 novel eigenvalue-based descriptors recently introduced, namely HMG, SMG, ISMG, PMG, and IPMG. Let M be a re-defined matrix based on the given graph G, and $\{\lambda_1, \lambda_2 \dots \lambda_k\}$ be its non-zero eigenvalues.

As the factor “s” may have different discrimination power for different networks, we thus provide these eigenvalue-based descriptors at both $s = 1$ and $s = 2$.

$$HM_G = - \sum_{i=1}^k \left[\frac{|\lambda_i|^{\frac{1}{s}}}{\sum_{j=1}^k |\lambda_j|^{\frac{1}{s}}} \log_2 \left(\frac{|\lambda_i|^{\frac{1}{s}}}{\sum_{j=1}^k |\lambda_j|^{\frac{1}{s}}} \right) \right].$$

$$SM_G = \sum_{i=1}^k |\lambda_i|^{\frac{1}{s}}.$$

$$ISM_G = \frac{1}{\sum_{i=1}^k |\lambda_i|^{\frac{1}{s}}}.$$

$$PM_G = \prod_{i=1}^k |\lambda_i|^{\frac{1}{s}}.$$

$$IPM_G = \frac{1}{\prod_{i=1}^k |\lambda_i|^{\frac{1}{s}}}.$$

These 5 eigenvalue-based descriptors could be applied to the following 10 differently re-defined matrices, including (1) adjacency matrix, (2) laplacian matrix, (3) distance matrix, (4) distance path matrix, (5) augmented vertex degree matrix, (6) extended adjacency matrix, (7) vertex connectivity matrix, (8) random walk Markov matrix, (9) weighted structure function matrix 1, and (10) weighted structure function matrix 2, which are defined as follows.

Therefore, totally 50 eigenvalue-based descriptors are calculated in this set.

- (1) Adjacency matrix A_{ij} , is initially generated based on the network connections.
- (2) Laplacian matrix L_{ij} , is introduced previously in defining the Laplacian energy.
- (3) Distance matrix D_{ij} , is the shortest distance between all the nodes.
- (4) Distance path matrix DP_{ij} , is derived from the distance matrix, by counting all the internal paths between a pair of nodes, including their shortest paths.

$$DP_{ij} = \binom{D_{ij} + 1}{2}.$$

- (5) Augmented vertex degree matrix AVD_{ij} , is defined by the nodes' degree and distance matrix.

$$AVD_{ij} = \frac{deg_j}{2^{D_{ij}}}.$$

- (6) Extended adjacency matrix EA_{ij} , is a symmetric matrix based on nodes' degree.

$$EA_{ij} = \begin{cases} \frac{1}{2} \left(\frac{deg_i}{deg_j} + \frac{deg_j}{deg_i} \right) & \text{if } A_{ij} = 1 \\ 0 & \text{otherwise} \end{cases}$$

- (7) Vertex connectivity matrix VC_{ij} , is another symmetric matrix based on nodes' degree.

$$VC_{ij} = \begin{cases} \frac{1}{\sqrt{deg_i \cdot deg_j}} & \text{if } A_{ij} = 1 \\ 0 & \text{otherwise} \end{cases}$$

- (8) Radom walk Markov matrix RWM_{ij} , is a non-symmetric matrix based on the nodes' degree. It is based on the assumption that each neighbour node can be reached from a given node with the same probability, such that the probability of reaching the neighbor of node i is $1/deg_i$. The generated distribution of walks is called the simple random walks.

$$RWM_{ij} = \begin{cases} \frac{1}{deg_i} & \text{if } A_{ij} = 1 \\ 0 & \text{otherwise} \end{cases}$$

- (9) Weighted structure function matrix 1 IM1_{ij} , is a more complexly defined matrix. In the following definitions, $radius_G$ is the maximum shortest path length in the network, and $|S_d(i)|$ is the number of nodes that are at the shortest distance d away from the node i .

$$f1(i) = \sum_{d=1}^{radius_G} (radius_G + 1 - d) \cdot |S_d(i)|.$$

$$pf1(i) = \frac{f1(i)}{\sum_{j=1}^N f1(j)}.$$

$$IM1_{ij} = 1 - \frac{|pf1(i) - pf1(j)|}{2^{D_{ij}}}.$$

(10) Weighted structure function matrix 2 $IM2_{ij}$, is defined slightly different.

$$f2(i) = \sum_{d=1}^{radius_G} (radius_G \cdot e^{1-d}) \cdot |S_d(i)|.$$

$$pf2(i) = \frac{f2(i)}{\sum_{j=1}^N f2(j)}.$$

$$IM2_{ij} = 1 - \frac{|pf2(i) - pf2(j)|}{2^{D_{ij}}}.$$

❖ Feature Category: Edge-Weighted Properties

196. Weighted Transitivity ¹⁴⁶

$$weighted_transitivity_G = \frac{\sum_{i=1}^N geo_tri_i}{\sum_{i=1}^N deg_i(deg_i - 1)}.$$

197. Barrat's Global Clustering Coefficients ⁶¹

$$clusterBarrat_G = \frac{1}{N} \sum_{i=1}^N clusterBarrat_i.$$

198. Onnela's Global Clustering Coefficients ^{61,213}

$$clusterOnnela_G = \frac{1}{N} \sum_{i=1}^N clusterOnnela_i.$$

199. Zhang's Global Clustering Coefficients ^{60,61}

$$clusterZhang_G = \frac{1}{N} \sum_{i=1}^N clusterZhang_i.$$

200. Holme's Global Clustering Coefficients ^{61,214}

$$clusterHolme_G = \frac{1}{N} \sum_{i=1}^N clusterHolme_i.$$

❖ **Feature Category: Node-Weighted Properties**

201. Total Node Weight

$$total_NW_G = \sum_{i=1}^N NW_i .$$

202. Node Weighted Global Clustering Coefficient ¹⁵

$$NWcluster_G = \frac{1}{N} \sum_{i=1}^N NWcluster_i .$$

❖ **Feature Category: Directed Properties**

203. Average In-Degree

$$avg_deg_G^+ = \frac{1}{N} \sum_{i \in N} deg_i^+ .$$

204. Maximum In-Degree

$$max_deg_G^+ = \max\{deg_i^+\} .$$

205. Minimum In-Degree

$$min_deg_G^+ = \min\{deg_i^+\} .$$

206. Average Out-Degree

$$avg_deg_G^- = \frac{1}{N} \sum_{i \in N} deg_i^- .$$

207. Maximum Out-Degree

$$max_deg_G^- = \max\{deg_i^-\} .$$

208. Minimum Out-Degree

$$min_deg_G^- = \min\{deg_i^-\} .$$

209. Directed Global Clustering Coefficient ⁴¹

$$cluster_G^\pm = \frac{1}{N} \sum_{i \in N} cluster_i^\pm .$$

210. Directed Flow Hierarchy ²⁵⁷

Flow hierarchy is a measurement of the percentage of edges that not involved in any directed cycles in the directed network.

B.3 Edge-Level Descriptors

1. Edge Weight

Edge weight EW_i is directly extracted from the user-provided edge weight list.

2. Edge Betweenness^{10,258}

Similarly with the definition of the node-level betweenness centrality. The edge betweenness quantifies the number of times an edge serving as a linking bridge along the shortest path between two nodes. In the following equation, node s and node t are two different nodes in the network, $\sigma_{st}(e)$ is the number of shorest paths from s to t that passing through the edge e , and σ_{st} is the number of shorest paths from node s to node t .

$$edgeBetweenness_e = \frac{\sum_{s \neq t} \sigma_{st}(e)}{\sigma_{st}}.$$

Section C: Protein-Protein Interaction Dataset for Studying Equilibrium Dissociation Constant (K_d)

Table S-4 Protein-protein interaction dataset for studying equilibrium dissociation constants (K_d)

| NO. | Protein A Uniprot | Protein B Uniprot | K_d (nM) | $\log(K_d)$ | PDB ID | References (PMID) |
|-----|----------------------|----------------------|------------|-------------|--------|----------------------|
| 1 | P0CG48 | Q15819 | 9.80E-05 | 4.01 | 1ZGU | 15943484 |
| 2 | O00499 | P01106 | 9.33E-05 | 4.03 | 1MV0 | 15943484 |
| 3 | P49792 | P12497 | 9.40E-05 | 4.03 | 4LQW | 25301850 |
| 4 | P0CH08 | P39940 | 9.06E-05 | 4.04 | 3OLM | 15943484 |
| 5 | P63279 | Q9UBT2 | 8.74E-05 | 4.06 | 2PX9 | 15943484 |
| 6 | Q9Y4K3 | P25942 | 8.40E-05 | 4.08 | 1LB6 | 12140561 |
| 7 | P55036 | P0CG48 | 7.30E-05 | 4.14 | 1YX6 | 15943484 |
| 8 | P07800 | P0A2D5 | 6.47E-05 | 4.19 | 2PMC | 15943484 |
| 9 | O43791 | Q6P8B3 | 6.36E-05 | 4.20 | 3HQH | 15943484 |
| 10 | P26675 | P62993 | 6.00E-05 | 4.22 | 1AZE | 15943484 |
| 11 | Q13191 | P0CH28 | 6.00E-05 | 4.22 | 2O0B | 21213247 |
| 12 | O95071 | P0CH28 | 6.00E-05 | 4.22 | 2QHO | 15943484 |
| 13 | Q8N3F8 | Q9H4M9 | 5.70E-05 | 4.24 | 2KSP | 15943484 |
| 14 | I3UIB4 | B2ZUN0 | 5.75E-05 | 4.24 | 4IKA | 25301850 |
| 15 | P20645 | Q9NZ52 | 5.60E-05 | 4.25 | 1JUQ | 15943484 |
| 16 | P0AE67 | P06974 | 5.50E-05 | 4.26 | 1U8T | 15943484 |
| 17 | P37090 | Q62940 | 5.30E-05 | 4.28 | 1ISH | 15943484 |
| 18 | P10412 | Q13185 | 5.20E-05 | 4.28 | 3TZD | 15943484 |
| 19 | P62987 | Q6R3M4 | 5.10E-05 | 4.29 | 2KWU | 15943484 |
| 20 | P04637 | Q92793 | 5.00E-05 | 4.30 | 1JSP | 15943484 |
| 21 | Q96RT1 | P04626 | 5.00E-05 | 4.30 | 1MFG | 12444095 |
| 22 | P16144 | Q15149 | 4.90E-05 | 4.31 | 3F7P | 15943484 |
| 23 | P62826 | P49791 | 4.90E-05 | 4.31 | 3GJ3 | 19505478 |
| 24 | P25823 | O76922 | 4.80E-05 | 4.32 | 3NTH | 20713507 |
| 25 | P49791 | P62826 | 4.70E-05 | 4.33 | 3GJ5 | 15943484 |
| 26 | P63010 | P42566 | 4.60E-05 | 4.34 | 2IV9 | 15943484 |
| 27 | O08398 | Q9WZU0 | 4.50E-05 | 4.35 | 4A2A | 15943484 |
| 28 | P53112 | P80667 | 4.40E-05 | 4.36 | 1N5Z | 15943484 |
| 29 | P63086 | Q64346 | 4.40E-05 | 4.36 | 2FYS | 15943484 |
| 30 | Q8WUM4 | Q9BY43 | 4.40E-05 | 4.36 | 3C3O | 15943484 |
| 31 | Q07817 | Q9Y5Z4 | 4.13E-05 | 4.38 | 3R85 | 15943484 |
| 32 | P22681 | P62837 | 4.20E-05 | 4.38 | 4A49 | 15943484 |
| 33 | Q8WUM4 | Q96CF2 | 4.10E-05 | 4.39 | 3C3R | 18511562 |
| 34 | O15105 | Q9HAU4 | 4.00E-05 | 4.40 | 2DJY | 15943484 |
| 35 | P32790 | P0CG63 | 4.00E-05 | 4.40 | 2JT4 | 15943484 |
| 36 | O60716 | P12830 | 4.00E-05 | 4.40 | 3L6X | 20371349 |
| 37 | O96017 | Q5PSV9 | 4.00E-05 | 4.40 | 3VA4 | 15943484 |
| 38 | P62940 | P12497 | 3.85E-05 | 4.41 | 4DGE | 15943484 |

| | | | | | | |
|----|--------|--------|----------|------|------|----------|
| 39 | P05556 | Q71LX4 | 3.60E-05 | 4.44 | 3G9W | 15943484 |
| 40 | P00698 | Q8ZYM8 | 3.53E-05 | 4.45 | 4GLA | 25301850 |
| 41 | P61964 | Q6P823 | 3.50E-05 | 4.46 | 2H9M | 15943484 |
| 42 | P62158 | P37840 | 3.50E-05 | 4.46 | 2M55 | 25301850 |
| 43 | P10275 | Q13772 | 3.30E-05 | 4.48 | 1T5Z | 15943484 |
| 44 | Q9HD42 | Q9UN37 | 3.34E-05 | 4.48 | 2JQ9 | 15943484 |
| 45 | Q99814 | P27540 | 3.00E-05 | 4.52 | 2A24 | 15943484 |
| 46 | P00523 | P41240 | 3.00E-05 | 4.52 | 3D7U | 15943484 |
| 47 | O00189 | P05067 | 2.96E-05 | 4.53 | 3L81 | 20230749 |
| 48 | P10636 | Q13526 | 2.90E-05 | 4.54 | 1I8H | 15943484 |
| 49 | P0AA04 | P69797 | 2.84E-05 | 4.55 | 1VRC | 15943484 |
| 50 | Q9J0X9 | Q9JJW6 | 2.80E-05 | 4.55 | 2KT5 | 15943484 |
| 51 | P36108 | P52917 | 2.80E-05 | 4.55 | 2V6X | 15943484 |
| 52 | P03536 | Q04637 | 2.70E-05 | 4.57 | 1LJ2 | 15943484 |
| 53 | P00533 | P22681 | 2.58E-05 | 4.59 | 3OB2 | 15943484 |
| 54 | Q9H251 | Q9Y6N9 | 2.54E-05 | 4.60 | 2KBR | 15943484 |
| 55 | Q02843 | Q8WUM4 | 2.45E-05 | 4.61 | 2XS8 | 15943484 |
| 56 | Q13114 | Q92844 | 2.39E-05 | 4.62 | 1L0A | 15943484 |
| 57 | Q8IUQ4 | Q9HB71 | 2.40E-05 | 4.62 | 2A25 | 15943484 |
| 58 | P54939 | Q8C351 | 2.40E-05 | 4.62 | 2K00 | 15943484 |
| 59 | P07766 | Q8TE68 | 2.40E-05 | 4.62 | 2ROL | 15943484 |
| 60 | P25054 | Q9ULH1 | 2.36E-05 | 4.63 | 2RQU | 15943484 |
| 61 | P68431 | Q92794 | 2.33E-05 | 4.63 | 3V43 | 15943484 |
| 62 | O75496 | P31274 | 2.22E-05 | 4.65 | 2LP0 | 15943484 |
| 63 | P0CG48 | Q9UKV5 | 2.26E-05 | 4.65 | 2LVQ | 15943484 |
| 64 | Q14677 | Q9UEU0 | 2.20E-05 | 4.66 | 2V8S | 18033301 |
| 65 | P62944 | O60331 | 2.20E-05 | 4.66 | 3H1Z | 19903820 |
| 66 | Q89VT8 | Q89VT6 | 2.15E-05 | 4.67 | 4H2S | 25301850 |
| 67 | P08839 | P0AA04 | 2.10E-05 | 4.68 | 2XDF | 15943484 |
| 68 | P01850 | Q8NKX2 | 2.00E-05 | 4.70 | 1KTK | 15943484 |
| 69 | P11884 | Q62760 | 2.00E-05 | 4.70 | 1OM2 | 15943484 |
| 70 | P02638 | Q15208 | 2.00E-05 | 4.70 | 1PSB | 15943484 |
| 71 | P0CG48 | Q9UMX0 | 2.00E-05 | 4.70 | 2JY6 | 15943484 |
| 72 | P20701 | Q9UMF0 | 2.00E-05 | 4.70 | 3BN3 | 18691975 |
| 73 | O43157 | P63000 | 1.89E-05 | 4.72 | 3SUA | 15943484 |
| 74 | P13232 | P16871 | 1.80E-05 | 4.74 | 3DI2 | 15943484 |
| 75 | P0CG48 | Q9WUB0 | 1.72E-05 | 4.76 | 3B08 | 15943484 |
| 76 | P62993 | P27870 | 1.70E-05 | 4.77 | 1GCQ | 21213247 |
| 77 | P20936 | P01112 | 1.70E-05 | 4.77 | 1WQ1 | 21213247 |
| 78 | P0CG48 | Q96S82 | 1.70E-05 | 4.77 | 2DEN | 15943484 |
| 79 | P62937 | P12497 | 1.60E-05 | 4.80 | 1AK4 | 21213247 |
| 80 | P03406 | P06241 | 1.60E-05 | 4.80 | 1AVZ | 21213247 |
| 81 | P06241 | P27986 | 1.60E-05 | 4.80 | 1AZG | 15943484 |
| 82 | P06876 | P45481 | 1.50E-05 | 4.82 | 1SB0 | 15943484 |
| 83 | P10275 | Q15596 | 1.50E-05 | 4.82 | 1T63 | 15943484 |
| 84 | P22216 | P34217 | 1.50E-05 | 4.82 | 2A0T | 15943484 |
| 85 | P04050 | Q05543 | 1.50E-05 | 4.82 | 2L0I | 15943484 |
| 86 | P06492 | P53999 | 1.50E-05 | 4.82 | 2PHE | 15943484 |

Appendices

| | | | | | | |
|-----|--------|--------|----------|------|------|--------------------|
| 87 | P62331 | Q02241 | 1.50E-05 | 4.82 | 3VHX | 15943484 |
| 88 | P48510 | P0CG63 | 1.48E-05 | 4.83 | 1WR1 | 15943484 |
| 89 | Q96AE4 | Q9UHX1 | 1.40E-05 | 4.85 | 2KXH | 15943484 |
| 90 | P00974 | P35030 | 1.40E-05 | 4.85 | 2R9P | 15943484 |
| 91 | O75376 | P04150 | 1.40E-05 | 4.85 | 3H52 | 15943484 |
| 92 | P06731 | Q57254 | 1.31E-05 | 4.88 | 2VER | 15943484 |
| 93 | Q60603 | Q60603 | 1.32E-05 | 4.88 | 4LLO | 25301850 |
| 94 | P54725 | P55036 | 1.29E-05 | 4.89 | 1P9D | 15943484 |
| 95 | P00533 | Q9UJM3 | 1.30E-05 | 4.89 | 2RF9 | 15943484 |
| 96 | P18206 | Q8IY67 | 1.26E-05 | 4.90 | 3H2U | 15943484 |
| 97 | O95727 | Q9BY67 | 1.25E-05 | 4.90 | 4H5S | 25301850 |
| 98 | P04213 | P0A0L5 | 1.20E-05 | 4.92 | 2AQ3 | 21213247 |
| 99 | P0CG48 | Q9JLQ0 | 1.20E-05 | 4.92 | 2LZ6 | 25301850 |
| 100 | Q7LBR1 | Q9UBP0 | 1.20E-05 | 4.92 | 3EAB | 15943484 |
| 101 | P47160 | Q04338 | 1.20E-05 | 4.92 | 3ONL | 15943484 |
| 102 | P41182 | Q9Y618 | 1.14E-05 | 4.94 | 1R2B | 15943484 |
| 103 | O95166 | P27797 | 1.15E-05 | 4.94 | 3DOW | 15943484 |
| 104 | P55072 | Q9UNN5 | 1.12E-05 | 4.95 | 3QC8 | 15943484 |
| 105 | P69783 | P0A6F3 | 1.10E-05 | 4.96 | 1GLA | 21213247 |
| 106 | O35179 | Q9JK66 | 1.10E-05 | 4.96 | 2KNB | 15943484 |
| 107 | P0CG53 | P46934 | 1.10E-05 | 4.96 | 2XBB | 15943484 |
| 108 | P02309 | Q9WYW0 | 1.05E-05 | 4.98 | 2H2H | 15943484 |
| 109 | P02679 | Q2G015 | 1.05E-05 | 4.98 | 2VR3 | 15943484 |
| 110 | P14737 | P22216 | 1.02E-05 | 4.99 | 1J4P | 15943484 |
| 111 | P9WJ71 | P9WGG9 | 1.03E-05 | 4.99 | 3VEP | 25301850 |
| 112 | P04610 | Q92831 | 1.00E-05 | 5.00 | 1JM4 | 15943484 |
| 113 | P11362 | Q8WU20 | 1.00E-05 | 5.00 | 1XR0 | 15943484 |
| 114 | P03069 | P19659 | 1.01E-05 | 5.00 | 2LPB | 15943484 |
| 115 | P09038 | P04271 | 1.00E-05 | 5.00 | 2M49 | 25301850 |
| 116 | Q16611 | P55957 | 1.00E-05 | 5.00 | 2M5B | 25301850 |
| 117 | P09803 | P0DJM0 | 1.00E-05 | 5.00 | 2OMW | 15943484 |
| 118 | P00431 | P00004 | 1.00E-05 | 5.00 | 2PCB | 21213247 |
| 119 | P07948 | P22575 | 9.58E-06 | 5.02 | 1WA7 | 15943484 |
| 120 | Q15843 | Q63KH5 | 9.40E-06 | 5.03 | 4HCP | 15943484 |
| 121 | P61830 | Q9WYW0 | 9.20E-06 | 5.04 | 2H2G | 15943484 |
| 122 | O75381 | P40855 | 9.17E-06 | 5.04 | 2W85 | 15943484 |
| 123 | P06729 | P19256 | 9.00E-06 | 5.05 | 1QA9 | 21213247 |
| 124 | P55036 | P62972 | 8.90E-06 | 5.05 | 2KDE | 15943484 |
| 125 | P56524 | P61981 | 9.00E-06 | 5.05 | 3UZD | 15943484 |
| 126 | P0CG48 | P21580 | 9.00E-06 | 5.05 | 3VUX | 25301850 |
| 127 | P62993 | Q9UQC2 | 8.70E-06 | 5.06 | 2VWF | 19523899 |
| 128 | O15164 | P68431 | 8.80E-06 | 5.06 | 3O34 | 15943484 |
| 129 | P10515 | Q64536 | 8.30E-06 | 5.08 | 3CRK | 15943484 |
| 130 | Q8WUM4 | Q9WC62 | 8.00E-06 | 5.10 | 2R02 | 15943484 |
| 131 | Q02242 | Q9NZQ7 | 8.00E-06 | 5.10 | 3BIK | 15943484 |
| 132 | Q12933 | Q15628 | 7.80E-06 | 5.11 | 1F3V | 15943484 |
| 133 | P20338 | Q9H1K0 | 7.70E-06 | 5.11 | 1Z0K | 21213247 |
| 134 | O97428 | P68135 | 7.58E-06 | 5.12 | 1SQK | 23055910, 15163400 |

Appendices

| | | | | | | |
|-----|--------|--------|----------|------|------|----------|
| 135 | Q13153 | Q14155 | 7.50E-06 | 5.12 | 1ZSG | 15943484 |
| 136 | Q06455 | Q99081 | 7.00E-06 | 5.15 | 2KNH | 15943484 |
| 137 | P35962 | Q8WUM4 | 7.00E-06 | 5.15 | 2R05 | 15943484 |
| 138 | Q9NZL4 | P08107 | 6.50E-06 | 5.19 | 1XQS | 21213247 |
| 139 | P0CH28 | Q9UJ41 | 6.40E-06 | 5.19 | 2C7M | 15943484 |
| 140 | Q9UJ41 | P0CH28 | 6.40E-06 | 5.19 | 2C7N | 15943484 |
| 141 | O14980 | O95149 | 6.50E-06 | 5.19 | 3GB8 | 15943484 |
| 142 | P12493 | P62937 | 6.10E-06 | 5.21 | 2X2D | 15943484 |
| 143 | P24394 | P35568 | 6.00E-06 | 5.22 | 1IRS | 15943484 |
| 144 | P01851 | P0DJY8 | 6.00E-06 | 5.22 | 1L0X | 15943484 |
| 145 | P55284 | Q99NH2 | 6.00E-06 | 5.22 | 2KOH | 15943484 |
| 146 | Q96FZ7 | Q9UN37 | 5.80E-06 | 5.24 | 2K3W | 15943484 |
| 147 | Q99523 | Q9UJY5 | 5.40E-06 | 5.27 | 3G2U | 15943484 |
| 148 | Q05195 | Q60520 | 5.20E-06 | 5.28 | 1S5Q | 15943484 |
| 149 | P63000 | Q00722 | 5.30E-06 | 5.28 | 2FJU | 21213247 |
| 150 | P38398 | Q13085 | 5.20E-06 | 5.28 | 3COJ | 15943484 |
| 151 | Q9NZ52 | Q15276 | 5.00E-06 | 5.30 | 1P4U | 12858162 |
| 152 | O55164 | Q8N3R9 | 5.00E-06 | 5.30 | 1Y76 | 15943484 |
| 153 | P35236 | P63086 | 5.00E-06 | 5.30 | 2GPH | 15943484 |
| 154 | Q9HTK8 | Q9HTK9 | 5.00E-06 | 5.30 | 2V3B | 15943484 |
| 155 | P10275 | P52293 | 5.00E-06 | 5.30 | 3BTR | 15943484 |
| 156 | P17427 | Q9Z0R4 | 5.00E-06 | 5.30 | 3HS8 | 20160082 |
| 157 | Q80UL9 | P97792 | 5.00E-06 | 5.30 | 3MJ7 | 20813955 |
| 158 | Q5SJ82 | Q5SJ83 | 5.00E-06 | 5.30 | 3T1Q | 15943484 |
| 159 | Q13330 | Q13547 | 5.00E-06 | 5.30 | 4BKX | 25301850 |
| 160 | P30801 | Q9CXW3 | 4.90E-06 | 5.31 | 2JTT | 15943484 |
| 161 | P0A8Q6 | P0ABT2 | 4.80E-06 | 5.32 | 2W9R | 19373253 |
| 162 | Q18PE0 | Q61006 | 4.70E-06 | 5.33 | 3ML4 | 15943484 |
| 163 | P63017 | Q91YN9 | 4.50E-06 | 5.35 | 3CQX | 15943484 |
| 164 | P98158 | Q9H9E1 | 4.50E-06 | 5.35 | 3V2X | 15943484 |
| 165 | O80297 | Q8X965 | 4.40E-06 | 5.36 | 2X9A | 15943484 |
| 166 | P04637 | Q9WYW0 | 4.30E-06 | 5.37 | 2H2D | 15943484 |
| 167 | P98170 | Q15750 | 4.20E-06 | 5.38 | 2POP | 15943484 |
| 168 | P0CG47 | Q9UK80 | 4.20E-06 | 5.38 | 2Y5B | 15943484 |
| 169 | Q3S4A7 | Q9ZNV9 | 4.10E-06 | 5.39 | 4EUK | 25301850 |
| 170 | P07900 | Q16543 | 4.00E-06 | 5.40 | 2K5B | 15943484 |
| 171 | P18181 | Q07763 | 4.00E-06 | 5.40 | 2PTT | 15943484 |
| 172 | P32797 | P13382 | 3.80E-06 | 5.42 | 3OIQ | 20877309 |
| 173 | P21890 | P0A3C8 | 3.60E-06 | 5.44 | 1EWY | 21213247 |
| 174 | P01112 | Q07889 | 3.60E-06 | 5.44 | 1NVU | 21213247 |
| 175 | P26449 | P41695 | 3.60E-06 | 5.44 | 2I3S | 15943484 |
| 176 | Q14449 | P01112 | 3.60E-06 | 5.44 | 4K81 | 25301850 |
| 177 | P11717 | Q9NZ52 | 3.50E-06 | 5.46 | 1LF8 | 15943484 |
| 178 | Q3YE50 | Q775D6 | 3.47E-06 | 5.46 | 2IOU | 15943484 |
| 179 | P21549 | P50542 | 3.50E-06 | 5.46 | 3R9A | 15943484 |
| 180 | Q8RSY1 | Q94F62 | 3.50E-06 | 5.46 | 3TL8 | 15943484 |
| 181 | P54727 | P55036 | 3.40E-06 | 5.47 | 1UEL | 15943484 |
| 182 | P02829 | Q16543 | 3.37E-06 | 5.47 | 1US7 | 14718169 |

| | | | | | | |
|-----|--------|--------|----------|------|------|--------------------|
| 183 | P9WHN4 | P9WQN4 | 3.40E-06 | 5.47 | 3M9D | 15943484 |
| 184 | Q9LT31 | Q9SN68 | 3.40E-06 | 5.47 | 4G01 | 25301850 |
| 185 | P02829 | Q12449 | 3.30E-06 | 5.48 | 1USV | 15943484 |
| 186 | P48736 | P01112 | 3.20E-06 | 5.49 | 1HE8 | 21213247 |
| 187 | P35546 | P97465 | 3.20E-06 | 5.49 | 1UEF | 15943484 |
| 188 | P25084 | Q9I494 | 3.20E-06 | 5.49 | 4NG2 | 25301850 |
| 189 | P01236 | P05710 | 3.15E-06 | 5.50 | 3NPZ | 15943484 |
| 190 | P04637 | P45481 | 3.10E-06 | 5.51 | 2L14 | 15943484 |
| 191 | P53365 | P63000 | 3.00E-06 | 5.52 | 1I4D | 21213247 |
| 192 | P05362 | P20701 | 3.00E-06 | 5.52 | 1MQ8 | 21213247 |
| 193 | Q9NP10 | P22692 | 3.00E-06 | 5.52 | 2DSP | 15943484 |
| 194 | Q24139 | Q9VVI3 | 3.00E-06 | 5.52 | 2EZ5 | 15943484 |
| 195 | P10408 | Q8CVI4 | 3.00E-06 | 5.52 | 2VDA | 15943484 |
| 196 | C4M622 | C4M4W4 | 3.00E-06 | 5.52 | 4DVG | 15943484 |
| 197 | P62158 | Q13936 | 2.90E-06 | 5.54 | 2LQC | 15943484 |
| 198 | P30533 | Q07954 | 2.80E-06 | 5.55 | 2FYL | 15943484 |
| 199 | P45974 | P0CG48 | 2.82E-06 | 5.55 | 2G45 | 15943484 |
| 200 | O75604 | P0CH28 | 2.80E-06 | 5.55 | 2HD5 | 15943484 |
| 201 | P0A9P4 | P0AA25 | 2.70E-06 | 5.57 | 1F6M | 21213247 |
| 202 | P26449 | P47074 | 2.70E-06 | 5.57 | 2I3T | 15943484 |
| 203 | P62155 | Q99LM3 | 2.70E-06 | 5.57 | 2K3S | 15943484 |
| 204 | P04637 | Q09472 | 2.70E-06 | 5.57 | 2K8F | 15943484 |
| 205 | Q4KMG0 | Q62226 | 2.70E-06 | 5.57 | 3D1M | 15943484 |
| 206 | P02299 | P05205 | 2.50E-06 | 5.60 | 1KNE | 15943484 |
| 207 | Q13191 | Q96B97 | 2.50E-06 | 5.60 | 2BZ8 | 15943484 |
| 208 | P06701 | P11938 | 2.49E-06 | 5.60 | 3OWT | 15943484 |
| 209 | O89100 | P70218 | 2.40E-06 | 5.62 | 1UTI | 15100220 |
| 210 | P04637 | Q7ZUW7 | 2.40E-06 | 5.62 | 2Z5T | 15943484 |
| 211 | P20339 | Q15075 | 2.40E-06 | 5.62 | 3MJH | 15943484 |
| 212 | Q99PF4 | Q99PJ1 | 2.40E-06 | 5.62 | 4AQE | 15943484 |
| 213 | P08581 | P22681 | 2.36E-06 | 5.63 | 3BUX | 15943484 |
| 214 | P60712 | P02584 | 2.30E-06 | 5.64 | 2BTF | 21213247 |
| 215 | Q9V444 | Q9V452 | 2.30E-06 | 5.64 | 2BYK | 15943484 |
| 216 | P54764 | O43921 | 2.30E-06 | 5.64 | 2WO3 | 19836338 |
| 217 | P30622 | Q14203 | 2.30E-06 | 5.64 | 3E2U | 15943484 |
| 218 | P70280 | Q96NW4 | 2.30E-06 | 5.64 | 4B93 | 15943484 |
| 219 | P40189 | Q98823 | 2.20E-06 | 5.66 | 1I1R | 21287608 |
| 220 | P20701 | P32942 | 2.20E-06 | 5.66 | 1T0P | 21287608 |
| 221 | Q60520 | Q96QT6 | 2.20E-06 | 5.66 | 2L9S | 15943484 |
| 222 | P22059 | Q9P0L0 | 2.10E-06 | 5.68 | 2RR3 | 15943484 |
| 223 | P17870 | P49951 | 2.10E-06 | 5.68 | 3GC3 | 19710023 |
| 224 | P17785 | P60903 | 2.00E-06 | 5.70 | 1BT6 | 15943484 |
| 225 | Q15819 | P61088 | 2.00E-06 | 5.70 | 1J7D | 11473255 |
| 226 | P28482 | Q15418 | 2.00E-06 | 5.70 | 4H3P | 25301850 |
| 227 | O43447 | O43172 | 1.97E-06 | 5.71 | 1MZW | 12875835 |
| 228 | P68433 | P83917 | 1.90E-06 | 5.72 | 1GUW | 15943484 |
| 229 | Q03386 | P01112 | 1.90E-06 | 5.72 | 1LFD | 21213247 |
| 230 | O00560 | Q01344 | 1.90E-06 | 5.72 | 1OBX | 12842047, 12679023 |

Appendices

| | | | | | | |
|-----|--------|--------|----------|------|------|----------|
| 231 | P68135 | Q8K4J6 | 1.90E-06 | 5.72 | 2V52 | 19008859 |
| 232 | Q9UJ41 | Q9UL25 | 1.80E-06 | 5.74 | 2OT3 | 17450153 |
| 233 | Q96FZ7 | Q9BRG1 | 1.80E-06 | 5.74 | 3HTU | 15943484 |
| 234 | C4ZQ55 | P09883 | 1.80E-06 | 5.74 | 3O0E | 15943484 |
| 235 | Q89VT8 | A9CHM9 | 1.73E-06 | 5.76 | 4H2W | 25301850 |
| 236 | P01857 | O75015 | 1.70E-06 | 5.77 | 1E4K | 21213247 |
| 237 | P01112 | P04049 | 1.70E-06 | 5.77 | 3KUD | 15943484 |
| 238 | Q7XJ80 | Q9SUR9 | 1.64E-06 | 5.79 | 2JKI | 15943484 |
| 239 | O88522 | P0CG48 | 1.60E-06 | 5.80 | 2ZVO | 15943484 |
| 240 | O88597 | P89884 | 1.58E-06 | 5.80 | 3BL2 | 15943484 |
| 241 | P56524 | Q9H9E1 | 1.60E-06 | 5.80 | 3V31 | 15943484 |
| 242 | P58340 | P62258 | 1.55E-06 | 5.81 | 3UAL | 15943484 |
| 243 | P11940 | Q9H074 | 1.50E-06 | 5.82 | 1JH4 | 15943484 |
| 244 | Q53EZ4 | Q8WUM4 | 1.50E-06 | 5.82 | 3E1R | 15943484 |
| 245 | P01033 | P50281 | 1.53E-06 | 5.82 | 3MA2 | 15943484 |
| 246 | P10826 | Q15788 | 1.50E-06 | 5.82 | 4DM8 | 15943484 |
| 247 | P22681 | P43405 | 1.49E-06 | 5.83 | 3BUW | 15943484 |
| 248 | P61088 | Q9Y4K3 | 1.48E-06 | 5.83 | 3HCT | 15943484 |
| 249 | P02710 | P60615 | 1.40E-06 | 5.85 | 1ABT | 15943484 |
| 250 | Q8BSL7 | Q9UJY5 | 1.40E-06 | 5.85 | 1J2J | 12679809 |
| 251 | Q07817 | Q14457 | 1.40E-06 | 5.85 | 2PON | 15943484 |
| 252 | P41182 | Q6W2J9 | 1.32E-06 | 5.88 | 3BIM | 15943484 |
| 253 | P12295 | P14739 | 1.30E-06 | 5.89 | 1UUG | 15943484 |
| 254 | P32324 | P11439 | 1.30E-06 | 5.89 | 1ZM4 | 21213247 |
| 255 | P54198 | Q9Y294 | 1.30E-06 | 5.89 | 2I32 | 15943484 |
| 256 | P06213 | P81122 | 1.30E-06 | 5.89 | 3BU6 | 15943484 |
| 257 | P70207 | O35464 | 1.30E-06 | 5.89 | 3OKY | 20877282 |
| 258 | Q3J179 | Q53119 | 1.30E-06 | 5.89 | 4HH3 | 25301850 |
| 259 | P08754 | P41220 | 1.25E-06 | 5.90 | 2V4Z | 15943484 |
| 260 | O08808 | P60766 | 1.25E-06 | 5.90 | 3EG5 | 15943484 |
| 261 | O14713 | O00522 | 1.24E-06 | 5.91 | 4DX8 | 25301850 |
| 262 | P0A988 | Q47155 | 1.20E-06 | 5.92 | 1OK7 | 15943484 |
| 263 | P54787 | P0CH28 | 1.20E-06 | 5.92 | 1P3Q | 15943484 |
| 264 | P78324 | Q08722 | 1.20E-06 | 5.92 | 2JJS | 15943484 |
| 265 | P06400 | P62136 | 1.20E-06 | 5.92 | 3N5U | 15943484 |
| 266 | Q8H1R0 | Q9CAJ0 | 1.20E-06 | 5.92 | 3RT0 | 15943484 |
| 267 | P10515 | Q15120 | 1.17E-06 | 5.93 | 1Y8N | 15943484 |
| 268 | A8MT69 | Q8N2Z9 | 1.17E-06 | 5.93 | 4DRA | 15943484 |
| 269 | O75533 | Q96I25 | 1.10E-06 | 5.96 | 2PEH | 15943484 |
| 270 | Q7DB61 | Q7DB62 | 1.10E-06 | 5.96 | 4KT5 | 25301850 |
| 271 | O22265 | P37107 | 1.06E-06 | 5.97 | 2HUG | 15943484 |
| 272 | P03081 | P30153 | 1.06E-06 | 5.97 | 2PKG | 15943484 |
| 273 | Q9Y6D9 | Q13257 | 1.04E-06 | 5.98 | 1GO4 | 15943484 |
| 274 | Q9VK33 | Q8ST83 | 1.03E-06 | 5.99 | 4C5G | 25301850 |
| 275 | P13051 | P15927 | 1.00E-06 | 6.00 | 1DPU | 15943484 |
| 276 | P02751 | Q53971 | 1.00E-06 | 6.00 | 1O9A | 15943484 |
| 277 | P02743 | P12318 | 1.00E-06 | 6.00 | 3D5O | 15943484 |
| 278 | Q9UPP1 | Q6NXT2 | 1.00E-06 | 6.00 | 3KV4 | 20023638 |

| | | | | | | |
|-----|--------|--------|----------|------|------|--------------------|
| 279 | O43791 | Q13618 | 1.00E-06 | 6.00 | 4EOZ | 15943484 |
| 280 | Q8WZ42 | O75147 | 9.40E-07 | 6.03 | 3KNB | 20489725 |
| 281 | P09883 | P0A855 | 9.20E-07 | 6.04 | 2IVZ | 15943484 |
| 282 | P01732 | P06239 | 9.00E-07 | 6.05 | 1Q69 | 15943484 |
| 283 | P38398 | Q9BX63 | 9.00E-07 | 6.05 | 1T29 | 15943484 |
| 284 | Q9X005 | Q9X006 | 9.00E-07 | 6.05 | 2F9Z | 15943484 |
| 285 | P03255 | P06400 | 9.00E-07 | 6.05 | 2R7G | 15943484 |
| 286 | O74515 | P87314 | 8.90E-07 | 6.05 | 2Z34 | 15943484 |
| 287 | P62805 | Q16576 | 9.00E-07 | 6.05 | 3CFV | 15943484 |
| 288 | P45448 | Q61066 | 9.00E-07 | 6.05 | 3F5C | 15943484 |
| 289 | P61006 | Q01968 | 9.00E-07 | 6.05 | 3QBT | 15943484 |
| 290 | P22692 | P05019 | 8.65E-07 | 6.06 | 1WQJ | 15642270 |
| 291 | O25675 | O25119 | 8.70E-07 | 6.06 | 4FQ0 | 25301850 |
| 292 | P08165 | P00257 | 8.56E-07 | 6.07 | 1E6E | 21287608 |
| 293 | Q93009 | P03211 | 8.60E-07 | 6.07 | 1YY6 | 15808506 |
| 294 | P70365 | P42226 | 8.00E-07 | 6.10 | 1OJ5 | 14757047 |
| 295 | P02766 | P02753 | 8.00E-07 | 6.10 | 1RLB | 21213247 |
| 296 | P56817 | Q9UJY5 | 8.00E-07 | 6.10 | 1UJJ | 15943484 |
| 297 | P63165 | Q9P0U3 | 7.87E-07 | 6.10 | 2IY1 | 15943484 |
| 298 | P62328 | P68135 | 8.00E-07 | 6.10 | 4PL8 | 23055910, 18327913 |
| 299 | P07992 | P23025 | 7.80E-07 | 6.11 | 2JNW | 15943484 |
| 300 | P01112 | Q5EBH1 | 7.70E-07 | 6.11 | 3DDC | 15943484 |
| 301 | Q8NB78 | P84243 | 7.41E-07 | 6.13 | 4GU0 | 25301850 |
| 302 | P16410 | Q96CW1 | 7.00E-07 | 6.15 | 1H6E | 15943484 |
| 303 | Q47038 | Q57254 | 7.00E-07 | 6.15 | 2IXQ | 15943484 |
| 304 | P78423 | Q7TDW8 | 6.80E-07 | 6.17 | 3ONA | 15943484 |
| 305 | Q86YC2 | P51587 | 6.60E-07 | 6.18 | 3EU7 | 19609323 |
| 306 | O60880 | Q13291 | 6.50E-07 | 6.19 | 1D4T | 10549287 |
| 307 | P13498 | P14598 | 6.40E-07 | 6.19 | 1WLP | 15943484 |
| 308 | Q13618 | Q9NVR0 | 6.50E-07 | 6.19 | 4APF | 15943484 |
| 309 | P55210 | P98170 | 6.30E-07 | 6.20 | 1I51 | 15943484 |
| 310 | P22681 | Q9C004 | 6.10E-07 | 6.21 | 3BUN | 15943484 |
| 311 | P23827 | P48740 | 6.10E-07 | 6.21 | 4IW4 | 25301850 |
| 312 | P26718 | Q29983 | 6.00E-07 | 6.22 | 1HYR | 21287608 |
| 313 | O60880 | Q13291 | 6.00E-07 | 6.22 | 1KA7 | 15943484 |
| 314 | P11474 | Q9UBK2 | 6.00E-07 | 6.22 | 1XB7 | 15337744 |
| 315 | P00431 | P00044 | 6.00E-07 | 6.22 | 2B10 | 21287608 |
| 316 | P62158 | Q13557 | 6.00E-07 | 6.22 | 2WEL | 15943484 |
| 317 | Q92673 | Q9UJY5 | 6.00E-07 | 6.22 | 3G2S | 15943484 |
| 318 | P40742 | Q01960 | 6.00E-07 | 6.22 | 3SYN | 15943484 |
| 319 | Q9ES57 | O54901 | 6.00E-07 | 6.22 | 4BFI | 25301850 |
| 320 | P00044 | P00431 | 5.88E-07 | 6.23 | 2JTI | 15943484 |
| 321 | O75531 | P50402 | 5.90E-07 | 6.23 | 2ODG | 15943484 |
| 322 | Q62120 | Q91ZM2 | 5.50E-07 | 6.26 | 2HDX | 15943484 |
| 323 | P19793 | Q15596 | 5.50E-07 | 6.26 | 3OAP | 15943484 |
| 324 | O14936 | O75334 | 5.50E-07 | 6.26 | 3TAC | 15943484 |
| 325 | P08362 | Q9GJT3 | 5.20E-07 | 6.28 | 3ALZ | 15943484 |
| 326 | P43146 | Q9HD67 | 5.30E-07 | 6.28 | 3AU4 | 15943484 |

| | | | | | | |
|-----|--------|--------|----------|------|------|--------------------|
| 327 | Q96EP0 | Q9BYM8 | 5.20E-07 | 6.28 | 4DBG | 15943484 |
| 328 | P01024 | P20023 | 5.00E-07 | 6.30 | 3OED | 15943484 |
| 329 | C4X1R9 | B5XTS6 | 5.00E-07 | 6.30 | 4AWX | 25301850 |
| 330 | P09581 | Q8R1R4 | 5.00E-07 | 6.30 | 4EXP | 15943484 |
| 331 | O08603 | O54709 | 4.86E-07 | 6.31 | 1JSK | 15943484 |
| 332 | Q9WVE9 | Q9Z0R4 | 4.90E-07 | 6.31 | 3HS9 | 15943484 |
| 333 | P60953 | Q15811 | 4.78E-07 | 6.32 | 3QBV | 15943484 |
| 334 | Q91YR1 | P68135 | 4.70E-07 | 6.33 | 3DAW | 23055910, 12429826 |
| 335 | P09581 | P07141 | 4.55E-07 | 6.34 | 3EJJ | 15943484 |
| 336 | P62834 | P04049 | 4.42E-07 | 6.35 | 3KUC | 20361980 |
| 337 | P08476 | P21674 | 4.30E-07 | 6.37 | 2ARP | 15943484 |
| 338 | A5IFX1 | P62820 | 4.30E-07 | 6.37 | 3TKL | 15943484 |
| 339 | Q13164 | Q13163 | 4.30E-07 | 6.37 | 4IC7 | 25301850 |
| 340 | P45983 | Q9WVI9 | 4.20E-07 | 6.38 | 1UKH | 15943484 |
| 341 | O60271 | P62330 | 4.20E-07 | 6.38 | 2W83 | 15943484 |
| 342 | O08604 | Q83156 | 4.20E-07 | 6.38 | 4G59 | 15943484 |
| 343 | P01730 | P06239 | 4.00E-07 | 6.40 | 1Q68 | 15943484 |
| 344 | P02994 | P32471 | 4.00E-07 | 6.40 | 2B7C | 15943484 |
| 345 | P0DJM0 | P12830 | 4.00E-07 | 6.40 | 2OMZ | 17715295 |
| 346 | P02549 | P11277 | 4.00E-07 | 6.40 | 3LBX | 15943484 |
| 347 | P09936 | P0CG48 | 3.85E-07 | 6.41 | 3KW5 | 15943484 |
| 348 | O74774 | Q9USL5 | 3.90E-07 | 6.41 | 3MCA | 20890290 |
| 349 | P60953 | Q9UQB8 | 3.91E-07 | 6.41 | 4JS0 | 19293156 |
| 350 | Q61188 | Q921E6 | 3.80E-07 | 6.42 | 2QXV | 15943484 |
| 351 | P67775 | Q9UIC8 | 3.80E-07 | 6.42 | 3P71 | 15943484 |
| 352 | P17119 | Q12045 | 3.70E-07 | 6.43 | 4ETP | 15943484 |
| 353 | P04486 | P32776 | 3.60E-07 | 6.44 | 2K2U | 15943484 |
| 354 | P00644 | Q1WCB7 | 3.60E-07 | 6.44 | 2KHS | 15943484 |
| 355 | P01024 | C8LN82 | 3.60E-07 | 6.44 | 2WY8 | 21055811 |
| 356 | O00522 | P61224 | 3.60E-07 | 6.44 | 4HDO | 25301850 |
| 357 | P11940 | Q9BPZ3 | 3.50E-07 | 6.46 | 1JGN | 15943484 |
| 358 | P04637 | P09429 | 3.46E-07 | 6.46 | 2LY4 | 15943484 |
| 359 | P63104 | P04049 | 3.46E-07 | 6.46 | 4IHL | 25301850 |
| 360 | P06400 | Q01094 | 3.40E-07 | 6.47 | 1O9K | 15943484 |
| 361 | P27601 | Q9ES67 | 3.40E-07 | 6.47 | 3CX8 | 18940608 |
| 362 | P0ABH9 | P0A8Q6 | 3.30E-07 | 6.48 | 1R6Q | 21213247 |
| 363 | P07560 | P39958 | 3.30E-07 | 6.48 | 3CPH | 21213247 |
| 364 | O00213 | P05067 | 3.30E-07 | 6.48 | 3DXE | 15943484 |
| 365 | P05067 | Q02410 | 3.20E-07 | 6.49 | 1X11 | 15943484 |
| 366 | P55957 | Q07820 | 3.20E-07 | 6.49 | 2KBW | 15943484 |
| 367 | P12023 | Q9DBR4 | 3.20E-07 | 6.49 | 2ROZ | 15943484 |
| 368 | Q2M3X8 | P68135 | 3.20E-07 | 6.49 | 4B1V | 15943484 |
| 369 | O96013 | O96013 | 3.20E-07 | 6.49 | 4L67 | 25301850 |
| 370 | P0AE67 | P07363 | 3.00E-07 | 6.52 | 1A0O | 21287608 |
| 371 | Q05195 | Q62141 | 3.00E-07 | 6.52 | 1PD7 | 15943484 |
| 372 | P15374 | P0CG48 | 3.00E-07 | 6.52 | 1XD3 | 21213247 |
| 373 | P22216 | P39009 | 3.00E-07 | 6.52 | 2JQL | 15943484 |
| 374 | P21333 | Q8WUP2 | 3.00E-07 | 6.52 | 2W0P | 15943484 |

Appendices

| | | | | | | |
|-----|--------|--------|----------|------|------|----------|
| 375 | P62158 | Q08209 | 3.00E-07 | 6.52 | 2W73 | 15943484 |
| 376 | P50750 | O60563 | 3.00E-07 | 6.52 | 3BLH | 18566585 |
| 377 | P52003 | Q9HXS2 | 3.00E-07 | 6.52 | 3ECH | 15943484 |
| 378 | P78318 | P67775 | 3.00E-07 | 6.52 | 4IYP | 25301850 |
| 379 | Q13485 | Q15796 | 2.96E-07 | 6.53 | 1U7V | 15943484 |
| 380 | Q96F46 | Q96PD4 | 2.92E-07 | 6.53 | 3JVF | 15943484 |
| 381 | P10600 | P37173 | 2.90E-07 | 6.54 | 1KTZ | 21213247 |
| 382 | O60486 | O75326 | 2.90E-07 | 6.54 | 3NVQ | 15943484 |
| 383 | E8X8J1 | E8XBD7 | 2.90E-07 | 6.54 | 3ZET | 25301850 |
| 384 | P62942 | P36897 | 2.80E-07 | 6.55 | 1B6C | 21213247 |
| 385 | P15529 | Q2KS96 | 2.84E-07 | 6.55 | 3L89 | 15943484 |
| 386 | Q811U3 | Q9JIR4 | 2.70E-07 | 6.57 | 1ZUB | 15943484 |
| 387 | P04271 | P52907 | 2.60E-07 | 6.59 | 1MQ1 | 15943484 |
| 388 | P03372 | Q15596 | 2.50E-07 | 6.60 | 1GWR | 15943484 |
| 389 | P49642 | P49643 | 2.50E-07 | 6.60 | 4BPX | 25301850 |
| 390 | P62491 | Q7L804 | 2.50E-07 | 6.60 | 4C4P | 25301850 |
| 391 | P60953 | Q07960 | 2.40E-07 | 6.62 | 1GRN | 21213247 |
| 392 | P05230 | P22607 | 2.30E-07 | 6.64 | 1RY7 | 21287608 |
| 393 | Q56312 | Q56310 | 2.30E-07 | 6.64 | 1U0S | 15289606 |
| 394 | P03126 | Q96QZ7 | 2.30E-07 | 6.64 | 2KPL | 15943484 |
| 395 | Q12018 | Q12395 | 2.27E-07 | 6.64 | 3O6B | 15943484 |
| 396 | O35235 | O35305 | 2.30E-07 | 6.64 | 3QBQ | 15943484 |
| 397 | O89100 | Q13094 | 2.20E-07 | 6.66 | 1H3H | 15943484 |
| 398 | Q15465 | Q96QV1 | 2.20E-07 | 6.66 | 3HO5 | 15943484 |
| 399 | Q07440 | Q91ZE9 | 2.10E-07 | 6.68 | 2VOG | 15943484 |
| 400 | O15151 | P04637 | 2.10E-07 | 6.68 | 3DAB | 15943484 |
| 401 | P52799 | P54764 | 2.03E-07 | 6.69 | 3GXU | 15943484 |
| 402 | P16410 | P33681 | 2.00E-07 | 6.70 | 1I8L | 15943484 |
| 403 | P04631 | P52907 | 2.00E-07 | 6.70 | 1MWN | 15943484 |
| 404 | P04637 | Q8WTS6 | 2.00E-07 | 6.70 | 1XQH | 15943484 |
| 405 | Q15843 | Q96LD8 | 2.00E-07 | 6.70 | 1XT9 | 15943484 |
| 406 | P54784 | P21691 | 2.00E-07 | 6.70 | 1ZHI | 21213247 |
| 407 | P26043 | Q62170 | 2.01E-07 | 6.70 | 2EMT | 15943484 |
| 408 | P48061 | P61073 | 2.00E-07 | 6.70 | 2K05 | 15943484 |
| 409 | P0C077 | P0C079 | 2.00E-07 | 6.70 | 2KC8 | 15943484 |
| 410 | O43559 | Q9UM73 | 2.00E-07 | 6.70 | 2KUP | 15943484 |
| 411 | Q92900 | Q9HAU5 | 2.00E-07 | 6.70 | 2WJV | 15943484 |
| 412 | P09883 | P04482 | 2.00E-07 | 6.70 | 2WPT | 16109424 |
| 413 | P89884 | Q14457 | 2.00E-07 | 6.70 | 3DVU | 15943484 |
| 414 | P08362 | P15529 | 2.00E-07 | 6.70 | 3INB | 15943484 |
| 415 | P21279 | Q01970 | 2.00E-07 | 6.70 | 3OHM | 20966218 |
| 416 | P00760 | P84781 | 2.02E-07 | 6.70 | 4AOQ | 25301850 |
| 417 | P39517 | P39998 | 2.00E-07 | 6.70 | 4BRU | 25301850 |
| 418 | P00747 | P00779 | 1.97E-07 | 6.71 | 1L4D | 21287608 |
| 419 | B8H5L1 | B8H5L0 | 1.93E-07 | 6.71 | 3T0Y | 15943484 |
| 420 | Q96BN8 | P0CG47 | 1.96E-07 | 6.71 | 3ZNZ | 25301850 |
| 421 | P07276 | P32776 | 1.90E-07 | 6.72 | 2LOX | 15943484 |
| 422 | P35222 | Q9DBG9 | 1.90E-07 | 6.72 | 3DIW | 15943484 |

Appendices

| | | | | | | |
|-----|--------|--------|----------|------|------|----------|
| 423 | Q10103 | P09988 | 1.90E-07 | 6.72 | 3G7L | 19362535 |
| 424 | P01138 | P07174 | 1.86E-07 | 6.73 | 1SGI | 15943484 |
| 425 | P06400 | P52293 | 1.80E-07 | 6.74 | 1PJM | 15943484 |
| 426 | P11234 | Q15311 | 1.84E-07 | 6.74 | 2KWI | 15943484 |
| 427 | P01024 | P08603 | 1.80E-07 | 6.74 | 2XQW | 15943484 |
| 428 | P10844 | P46097 | 1.80E-07 | 6.74 | 4KBB | 25301850 |
| 429 | O70161 | P26039 | 1.70E-07 | 6.77 | 1Y19 | 15943484 |
| 430 | P63000 | Q16512 | 1.70E-07 | 6.77 | 2RMK | 15943484 |
| 431 | O89100 | Q6PB44 | 1.70E-07 | 6.77 | 2W10 | 15943484 |
| 432 | P23615 | Q06505 | 1.70E-07 | 6.77 | 3OAK | 15943484 |
| 433 | P50489 | Q8IKV6 | 1.65E-07 | 6.78 | 3SRI | 15943484 |
| 434 | Q9BUL8 | O00506 | 1.64E-07 | 6.79 | 3W8H | 25301850 |
| 435 | P06103 | P40217 | 1.60E-07 | 6.80 | 3ZWL | 15943484 |
| 436 | P55075 | P21802 | 1.55E-07 | 6.81 | 2FDB | 21287608 |
| 437 | P61972 | P62825 | 1.50E-07 | 6.82 | 1A2K | 21213247 |
| 438 | Q96BD6 | Q96IZ0 | 1.50E-07 | 6.82 | 2JK9 | 20561531 |
| 439 | P29083 | P32780 | 1.50E-07 | 6.82 | 2RNR | 15943484 |
| 440 | P03182 | Q16611 | 1.50E-07 | 6.82 | 2XPX | 15943484 |
| 441 | P01834 | P05067 | 1.51E-07 | 6.82 | 4HIX | 25301850 |
| 442 | O53512 | P9WIC1 | 1.40E-07 | 6.85 | 2W19 | 15943484 |
| 443 | P61964 | Q03164 | 1.40E-07 | 6.85 | 3EMH | 15943484 |
| 444 | Q5JII0 | Q5JII1 | 1.40E-07 | 6.85 | 3VYR | 15943484 |
| 445 | O54921 | P11233 | 1.37E-07 | 6.86 | 1UAD | 15943484 |
| 446 | P05230 | P11362 | 1.36E-07 | 6.87 | 1EVT | 21287608 |
| 447 | Q1EHW4 | Q60520 | 1.34E-07 | 6.87 | 2RMS | 15943484 |
| 448 | Q13291 | O35324 | 1.31E-07 | 6.88 | 1I3Z | 15943484 |
| 449 | Q63373 | Q8N0W4 | 1.32E-07 | 6.88 | 2WQZ | 15943484 |
| 450 | P05230 | P21802 | 1.30E-07 | 6.89 | 3OJM | 15943484 |
| 451 | P00760 | P01062 | 1.20E-07 | 6.92 | 1G9I | 15943484 |
| 452 | P08160 | Q14790 | 1.20E-07 | 6.92 | 1I4E | 15943484 |
| 453 | P0ABB0 | P0ABA4 | 1.20E-07 | 6.92 | 2A7U | 15943484 |
| 454 | P12830 | A4GWL5 | 1.20E-07 | 6.92 | 2OMX | 15943484 |
| 455 | P53741 | Q01477 | 1.19E-07 | 6.92 | 2QIY | 15943484 |
| 456 | Q15554 | Q9BSI4 | 1.20E-07 | 6.92 | 3BU8 | 15943484 |
| 457 | O89100 | Q60787 | 1.18E-07 | 6.93 | 1OEB | 15943484 |
| 458 | P09803 | P14923 | 1.16E-07 | 6.94 | 3IFQ | 15943484 |
| 459 | Q5SSZ7 | Q8BFU0 | 1.14E-07 | 6.94 | 4C99 | 25301850 |
| 460 | O95630 | Q9Y3E7 | 1.13E-07 | 6.95 | 2XZE | 15943484 |
| 461 | Q96RJ3 | Q9Y275 | 1.09E-07 | 6.96 | 1OQE | 15943484 |
| 462 | P50542 | P22307 | 1.09E-07 | 6.96 | 2C0L | 17157249 |
| 463 | P60842 | Q61823 | 1.10E-07 | 6.96 | 3EIQ | 15943484 |
| 464 | P06766 | P18887 | 1.10E-07 | 6.96 | 3K75 | 15943484 |
| 465 | O15085 | P61586 | 1.10E-07 | 6.96 | 3KZ1 | 15943484 |
| 466 | O08808 | Q3US76 | 1.02E-07 | 6.99 | 2F31 | 15943484 |
| 467 | Q62768 | Q9JIS1 | 1.00E-07 | 7.00 | 2CJS | 15943484 |
| 468 | P39104 | Q06389 | 1.00E-07 | 7.00 | 2JU0 | 15943484 |
| 469 | P12830 | A4GWM6 | 1.00E-07 | 7.00 | 2OMT | 15943484 |
| 470 | P61964 | P68431 | 1.00E-07 | 7.00 | 4A7J | 15943484 |

| | | | | | | |
|-----|--------|--------|----------|------|------|----------|
| 471 | Q12118 | Q12285 | 1.00E-07 | 7.00 | 4ASW | 25301850 |
| 472 | Q9Y376 | Q9P289 | 9.91E-08 | 7.00 | 4FZA | 25301850 |
| 473 | P08254 | P01033 | 9.50E-08 | 7.02 | 1UEA | 21287608 |
| 474 | O75695 | Q9WUL7 | 9.50E-08 | 7.02 | 3BH6 | 15943484 |
| 475 | P03407 | P08631 | 9.60E-08 | 7.02 | 3REB | 15943484 |
| 476 | Q8ML92 | Q8WXI2 | 9.25E-08 | 7.03 | 3BS5 | 18287031 |
| 477 | P69905 | Q9NZD4 | 9.30E-08 | 7.03 | 3IA3 | 15943484 |
| 478 | P00573 | P00806 | 9.20E-08 | 7.04 | 1ARO | 21287608 |
| 479 | P08581 | P14210 | 9.00E-08 | 7.05 | 1SHY | 15943484 |
| 480 | Q03555 | P20781 | 9.00E-08 | 7.05 | 2FTS | 16511563 |
| 481 | O60609 | Q5T4W7 | 9.00E-08 | 7.05 | 2GH0 | 15943484 |
| 482 | Q8N488 | Q99496 | 9.00E-08 | 7.05 | 3IXS | 15943484 |
| 483 | Q15046 | Q13155 | 9.00E-08 | 7.05 | 4DPG | 25301850 |
| 484 | P54253 | Q96RK0 | 8.85E-08 | 7.05 | 4J2L | 25301850 |
| 485 | Q15637 | P26368 | 8.40E-08 | 7.08 | 2M0G | 25301850 |
| 486 | Q14974 | O95149 | 8.30E-08 | 7.08 | 2QNA | 18028944 |
| 487 | P63280 | P63165 | 8.20E-08 | 7.09 | 2UYZ | 17491593 |
| 488 | P40189 | P15018 | 8.00E-08 | 7.10 | 1PVH | 21213247 |
| 489 | P62820 | Q5ZSQ3 | 7.90E-08 | 7.10 | 2WWX | 19942850 |
| 490 | P24941 | P61024 | 7.70E-08 | 7.11 | 1BUH | 21213247 |
| 491 | O43323 | Q96QV1 | 7.36E-08 | 7.13 | 2WG3 | 15943484 |
| 492 | Q62226 | Q96QV1 | 7.39E-08 | 7.13 | 2WG4 | 15943484 |
| 493 | Q1RGE4 | Q2RHX9 | 7.30E-08 | 7.14 | 2PV1 | 17825319 |
| 494 | P08253 | P16035 | 7.10E-08 | 7.15 | 1GXD | 21287608 |
| 495 | O35718 | Q00560 | 7.00E-08 | 7.15 | 2HMH | 16905102 |
| 496 | P13861 | Q06455 | 6.70E-08 | 7.17 | 2KYG | 15943484 |
| 497 | P16757 | Q29980 | 6.60E-08 | 7.18 | 2WY3 | 15943484 |
| 498 | P01391 | P02710 | 6.50E-08 | 7.19 | 1LXH | 15943484 |
| 499 | Q9JI78 | P54728 | 6.50E-08 | 7.19 | 2F4M | 16500903 |
| 500 | P09038 | P11362 | 6.19E-08 | 7.21 | 1CVS | 21287608 |
| 501 | P12004 | P39748 | 6.00E-08 | 7.22 | 1UL1 | 15943484 |
| 502 | P11233 | P15879 | 6.00E-08 | 7.22 | 2A9K | 21213247 |
| 503 | B7UM99 | P16333 | 6.00E-08 | 7.22 | 2CI9 | 15943484 |
| 504 | P00760 | P84781 | 6.09E-08 | 7.22 | 4AOR | 25301850 |
| 505 | O83922 | O83923 | 6.00E-08 | 7.22 | 4DI3 | 15943484 |
| 506 | P21645 | G7RM21 | 5.90E-08 | 7.23 | 4IHH | 25301850 |
| 507 | P84022 | Q13485 | 5.80E-08 | 7.24 | 1U7F | 15943484 |
| 508 | P52630 | P45481 | 5.80E-08 | 7.24 | 2KA4 | 15943484 |
| 509 | P45481 | Q04207 | 5.70E-08 | 7.24 | 2LWW | 25301850 |
| 510 | P9WJK2 | P71590 | 5.80E-08 | 7.24 | 3OUN | 15943484 |
| 511 | O34208 | O66100 | 5.70E-08 | 7.24 | 3TU3 | 15943484 |
| 512 | E8SYK9 | E8SYK8 | 5.60E-08 | 7.25 | 3W6J | 25301850 |
| 513 | Q9BXB1 | Q2MKA7 | 5.65E-08 | 7.25 | 4KT1 | 25301850 |
| 514 | P84078 | Q5T5U3 | 5.50E-08 | 7.26 | 2J59 | 15943484 |
| 515 | Q01567 | Q01565 | 5.50E-08 | 7.26 | 3UYM | 15943484 |
| 516 | P39769 | Q9VHA0 | 5.40E-08 | 7.27 | 1PK1 | 15943484 |
| 517 | P04629 | P29353 | 5.30E-08 | 7.28 | 1SHC | 15943484 |
| 518 | O00834 | Q9XYH7 | 5.30E-08 | 7.28 | 2K2S | 15943484 |

Appendices

| | | | | | | |
|-----|--------|--------|----------|------|------|--------------------|
| 519 | P42224 | P45481 | 5.20E-08 | 7.28 | 2KA6 | 15943484 |
| 520 | P53192 | Q12154 | 5.10E-08 | 7.29 | 3ZS8 | 15943484 |
| 521 | Q9BG57 | P63073 | 5.00E-08 | 7.30 | 1EJ4 | 15943484 |
| 522 | P60953 | Q9JK83 | 5.00E-08 | 7.30 | 1NF3 | 15943484 |
| 523 | P26368 | Q15637 | 5.00E-08 | 7.30 | 1OPI | 15943484 |
| 524 | P05434 | P11927 | 5.00E-08 | 7.30 | 1OQP | 15943484 |
| 525 | Q96SB4 | Q07955 | 5.00E-08 | 7.30 | 3BEG | 18342604, 14555757 |
| 526 | O75581 | O94907 | 5.00E-08 | 7.30 | 3S8V | 15943484 |
| 527 | P39517 | P25644 | 5.00E-08 | 7.30 | 4BRW | 25301850 |
| 528 | O15162 | P52293 | 4.58E-08 | 7.34 | 1Y2A | 15943484 |
| 529 | Q14980 | Q8VDU0 | 4.60E-08 | 7.34 | 3RO2 | 15943484 |
| 530 | Q9RE09 | P08877 | 4.50E-08 | 7.35 | 1KKL | 21213247 |
| 531 | P35557 | Q07071 | 4.50E-08 | 7.35 | 4LC9 | 18809676 |
| 532 | P22681 | Q9Z200 | 4.30E-08 | 7.37 | 1YVH | 15943484 |
| 533 | P84092 | P18508 | 4.22E-08 | 7.37 | 2PR9 | 18305175 |
| 534 | P00747 | P49054 | 4.20E-08 | 7.38 | 1I5K | 15943484 |
| 535 | O54924 | P11233 | 4.20E-08 | 7.38 | 1ZC4 | 15943484 |
| 536 | P09651 | Q92973 | 4.20E-08 | 7.38 | 2H4M | 15943484 |
| 537 | P39748 | P52293 | 4.20E-08 | 7.38 | 3UVU | 15943484 |
| 538 | P56589 | P40855 | 4.08E-08 | 7.39 | 3AJB | 21102411 |
| 539 | Q96L35 | P52799 | 4.00E-08 | 7.40 | 2HLE | 21213247 |
| 540 | P09052 | A1Z6E0 | 4.00E-08 | 7.40 | 2IHS | 15943484 |
| 541 | P49841 | Q92837 | 3.90E-08 | 7.41 | 1GNG | 15943484 |
| 542 | P18206 | P54939 | 3.90E-08 | 7.41 | 1RKC | 15943484 |
| 543 | P12003 | P26039 | 3.90E-08 | 7.41 | 1T01 | 15943484 |
| 544 | P55211 | Q8XAL7 | 3.87E-08 | 7.41 | 3V3K | 25301850 |
| 545 | Q47112 | P13479 | 3.80E-08 | 7.42 | 3GJN | 15019791 |
| 546 | P60604 | Q9UKV5 | 3.80E-08 | 7.42 | 4LAD | 25301850 |
| 547 | P05019 | P24593 | 3.70E-08 | 7.43 | 1H59 | 15943484 |
| 548 | P10415 | Q07813 | 3.53E-08 | 7.45 | 2XA0 | 15943484 |
| 549 | P09372 | P0A6Y8 | 3.50E-08 | 7.46 | 1DKG | 21287608 |
| 550 | P78310 | P36711 | 3.50E-08 | 7.46 | 1P6A | 15943484 |
| 551 | P42768 | Q8X2U1 | 3.50E-08 | 7.46 | 2K42 | 15943484 |
| 552 | Q9IH62 | P52799 | 3.50E-08 | 7.46 | 2VSM | 18488039 |
| 553 | P45481 | Q9Y6Q9 | 3.40E-08 | 7.47 | 1KBH | 15943484 |
| 554 | P10844 | P29101 | 3.40E-08 | 7.47 | 2NM1 | 17167421 |
| 555 | P11362 | P10686 | 3.30E-08 | 7.48 | 3GQI | 19665973 |
| 556 | P50984 | Q8WSF8 | 3.26E-08 | 7.49 | 2BR8 | 15943484 |
| 557 | O49908 | Q43866 | 3.10E-08 | 7.51 | 2XQR | 15943484 |
| 558 | P40056 | Q12154 | 3.10E-08 | 7.51 | 3SJD | 15943484 |
| 559 | P04637 | Q13625 | 3.00E-08 | 7.52 | 1YCS | 15943484 |
| 560 | P60568 | P01589 | 3.01E-08 | 7.52 | 1Z92 | 21287608 |
| 561 | P56945 | Q8N5H7 | 3.00E-08 | 7.52 | 3T6G | 15943484 |
| 562 | Q05195 | Q60520 | 2.90E-08 | 7.54 | 1G1E | 15943484 |
| 563 | Q9HYC5 | Q9HYC4 | 2.81E-08 | 7.55 | 3WA5 | 25301850 |
| 564 | P38919 | Q9HCG8 | 2.84E-08 | 7.55 | 4C9B | 25301850 |
| 565 | P10845 | Q496J9 | 2.80E-08 | 7.55 | 4JRA | 25301850 |
| 566 | P03180 | Q13651 | 2.70E-08 | 7.57 | 1Y6N | 15943484 |

Appendices

| | | | | | | |
|-----|--------|--------|----------|------|------|--------------------|
| 567 | P0A855 | P0A912 | 2.70E-08 | 7.57 | 2HQS | 21213247 |
| 568 | A6QG59 | P01024 | 2.60E-08 | 7.59 | 3D5S | 15943484 |
| 569 | P68135 | P06396 | 2.50E-08 | 7.60 | 1H1V | 21213247 |
| 570 | P52800 | P54763 | 2.50E-08 | 7.60 | 1KGY | 15943484 |
| 571 | Q7Z6A9 | Q92956 | 2.50E-08 | 7.60 | 2AW2 | 15943484 |
| 572 | P38507 | P01857 | 2.30E-08 | 7.64 | 1FC2 | 21213247 |
| 573 | P27782 | Q02248 | 2.30E-08 | 7.64 | 3OUW | 15943484 |
| 574 | P06180 | P52293 | 2.20E-08 | 7.66 | 1PJN | 15943484 |
| 575 | Q9CU62 | Q9CW03 | 2.20E-08 | 7.66 | 2WD5 | 15943484 |
| 576 | P12643 | Q5D734 | 2.20E-08 | 7.66 | 3BK3 | 15943484 |
| 577 | P21279 | P41220 | 2.20E-08 | 7.66 | 4EKC | 25301850 |
| 578 | P35658 | Q9UMR2 | 2.16E-08 | 7.67 | 3FHC | 15943484 |
| 579 | P31809 | Q9J3E7 | 2.14E-08 | 7.67 | 3R4D | 15943484 |
| 580 | P00533 | P01133 | 2.10E-08 | 7.68 | 1IVO | 10840042 |
| 581 | P60604 | Q9UKV5 | 2.10E-08 | 7.68 | 3H8K | 19560420 |
| 582 | H9T8H3 | I2KQ03 | 2.11E-08 | 7.68 | 4G6V | 15943484 |
| 583 | A0AEF6 | A0AEF5 | 2.08E-08 | 7.68 | 4IU3 | 25301850 |
| 584 | P60953 | Q61036 | 2.00E-08 | 7.70 | 1EES | 15943484 |
| 585 | P06396 | P68135 | 2.00E-08 | 7.70 | 1EQY | 8987989 |
| 586 | P63104 | Q29495 | 2.00E-08 | 7.70 | 1IB1 | 21213247 |
| 587 | P62158 | P40136 | 2.00E-08 | 7.70 | 1K93 | 15943484 |
| 588 | P22002 | P54287 | 2.00E-08 | 7.70 | 1VYT | 15943484 |
| 589 | Q14145 | Q16236 | 2.00E-08 | 7.70 | 2FLU | 16888629 |
| 590 | Q80S15 | P78310 | 2.00E-08 | 7.70 | 2J12 | 16923808 |
| 591 | P52272 | Q92973 | 2.00E-08 | 7.70 | 2OT8 | 15943484 |
| 592 | P36404 | Q9Y2Y0 | 2.00E-08 | 7.70 | 3DOE | 19368893, 10488091 |
| 593 | A5IHF0 | Q5ZYC9 | 2.00E-08 | 7.70 | 3FXD | 15943484 |
| 594 | P9WJ66 | P9WGH4 | 2.00E-08 | 7.70 | 3HUG | 15943484 |
| 595 | Q96FW1 | P0CG48 | 2.00E-08 | 7.70 | 4I6L | 25301850 |
| 596 | P00533 | P01135 | 1.90E-08 | 7.72 | 1MOX | 10840042 |
| 597 | O43521 | P0C6Z1 | 1.80E-08 | 7.74 | 2WH6 | 15943484 |
| 598 | P62136 | Q12972 | 1.78E-08 | 7.75 | 3V4Y | 15943484 |
| 599 | F2WK69 | F2WK70 | 1.78E-08 | 7.75 | 4G6U | 15943484 |
| 600 | P62554 | P0AES4 | 1.75E-08 | 7.76 | 1X75 | 15943484 |
| 601 | Q15223 | Q69091 | 1.71E-08 | 7.77 | 3U82 | 15943484 |
| 602 | Q15554 | Q9NYB0 | 1.65E-08 | 7.78 | 3K6G | 15943484 |
| 603 | P27487 | K0BRG7 | 1.67E-08 | 7.78 | 4KR0 | 25301850 |
| 604 | P26043 | P35330 | 1.64E-08 | 7.79 | 1J19 | 12554651 |
| 605 | Q9BYF1 | P59594 | 1.60E-08 | 7.80 | 2AJF | 21213247 |
| 606 | Q2RSB2 | P0C188 | 1.60E-08 | 7.80 | 2OOR | 21213247 |
| 607 | P36711 | P78310 | 1.50E-08 | 7.82 | 1KAC | 21287608 |
| 608 | P11717 | Q59EZ3 | 1.53E-08 | 7.82 | 2L29 | 15943484 |
| 609 | P18206 | Q9Y490 | 1.47E-08 | 7.83 | 1SYQ | 15943484 |
| 610 | P55407 | P55408 | 1.49E-08 | 7.83 | 2Q0O | 15943484 |
| 611 | Q06124 | P02751 | 1.40E-08 | 7.85 | 4JE4 | 25301850 |
| 612 | Q7Z6M4 | Q96CB9 | 1.33E-08 | 7.88 | 4FZV | 15943484 |
| 613 | P68400 | P67870 | 1.30E-08 | 7.89 | 1JWH | 21213247 |
| 614 | P45481 | Q99967 | 1.30E-08 | 7.89 | 1R8U | 15943484 |

Appendices

| | | | | | | |
|-----|--------|--------|----------|------|------|------------------|
| 615 | O88653 | Q9JHS3 | 1.28E-08 | 7.89 | 1VET | 15263099 |
| 616 | P01024 | Q6GA60 | 1.30E-08 | 7.89 | 2NOJ | 15943484 |
| 617 | P16474 | Q08199 | 1.30E-08 | 7.89 | 3QML | 15943484 |
| 618 | O08689 | P19883 | 1.23E-08 | 7.91 | 3HH2 | 15943484 |
| 619 | P03252 | P03274 | 1.20E-08 | 7.92 | 1AVP | 15943484 |
| 620 | P07897 | Q05546 | 1.20E-08 | 7.92 | 1TDQ | 15943484 |
| 621 | Q96QT6 | Q9UBU8 | 1.20E-08 | 7.92 | 2LKM | 15943484 |
| 622 | Q7RTN6 | Q9Y376 | 1.20E-08 | 7.92 | 3GNI | 15943484 |
| 623 | Q06124 | P02751 | 1.20E-08 | 7.92 | 4JEG | 25301850 |
| 624 | Q9BUL8 | Q9P289 | 1.17E-08 | 7.93 | 4GEH | 25301850 |
| 625 | Q01842 | Q7K119 | 1.11E-08 | 7.95 | 1SV0 | 15943484 |
| 626 | P9WNK4 | P9WNK6 | 1.10E-08 | 7.96 | 1WA8 | 15943484 |
| 627 | P12497 | O75475 | 1.10E-08 | 7.96 | 2B4J | 21213247 |
| 628 | P42260 | Q63273 | 1.10E-08 | 7.96 | 3QLU | 15943484 |
| 629 | P02788 | Q54972 | 1.03E-08 | 7.99 | 2PMS | 15943484 |
| 630 | B6KAM0 | B6KV60 | 1.02E-08 | 7.99 | 2Y8T | 15943484 |
| 631 | P08476 | P38444 | 1.00E-08 | 8.00 | 1NYS | 15943484 |
| 632 | P08476 | P38445 | 1.00E-08 | 8.00 | 1NYU | 15943484 |
| 633 | P25054 | Q02248 | 1.00E-08 | 8.00 | 1V18 | 15943484 |
| 634 | P00588 | Q99075 | 1.00E-08 | 8.00 | 1XDT | 9659904, 7961874 |
| 635 | O07347 | P83749 | 1.00E-08 | 8.00 | 2J7P | 15943484 |
| 636 | P07463 | P04775 | 1.00E-08 | 8.00 | 2KXW | 15943484 |
| 637 | Q07817 | Q9BXH1 | 1.00E-08 | 8.00 | 2M04 | 25301850 |
| 638 | Q6CUS2 | Q12745 | 1.00E-08 | 8.00 | 3K8P | 20005805 |
| 639 | O14763 | P50591 | 9.74E-09 | 8.01 | 1DU3 | 21287608 |
| 640 | P00698 | Q8JGG7 | 9.40E-09 | 8.03 | 2I26 | 15943484 |
| 641 | P11277 | P16157 | 9.29E-09 | 8.03 | 3KBT | 15943484 |
| 642 | Q8JL80 | O60486 | 9.40E-09 | 8.03 | 3NVN | 20727575 |
| 643 | O88574 | Q60520 | 9.20E-09 | 8.04 | 2LD7 | 15943484 |
| 644 | O95931 | Q99496 | 9.20E-09 | 8.04 | 3GS2 | 15943484 |
| 645 | P46655 | P46672 | 9.00E-09 | 8.05 | 2HRK | 21213247 |
| 646 | P26645 | P62158 | 8.80E-09 | 8.06 | 1IWQ | 15943484 |
| 647 | O35274 | P62136 | 8.70E-09 | 8.06 | 3EGG | 15943484 |
| 648 | Q9Y6N7 | O94813 | 8.20E-09 | 8.09 | 2V9T | 17848514 |
| 649 | P61765 | P32851 | 8.10E-09 | 8.09 | 4JEH | 25301850 |
| 650 | P62158 | Q13698 | 7.90E-09 | 8.10 | 2VAY | 15943484 |
| 651 | Q9HCJ2 | Q9Y2I2 | 7.90E-09 | 8.10 | 3ZYJ | 15943484 |
| 652 | Q00805 | Q01083 | 7.70E-09 | 8.11 | 3C9A | 15943484 |
| 653 | Q9LT31 | Q9SN68 | 7.30E-09 | 8.14 | 2EFD | 15943484 |
| 654 | Q96CW9 | Q9HBW1 | 7.30E-09 | 8.14 | 3ZYI | 15943484 |
| 655 | P45481 | Q16665 | 7.00E-09 | 8.15 | 1L8C | 15943484 |
| 656 | P09787 | P09788 | 7.00E-09 | 8.15 | 1WVE | 15943484 |
| 657 | P61925 | O14980 | 7.00E-09 | 8.15 | 2L1L | 15943484 |
| 658 | P24863 | P49336 | 7.05E-09 | 8.15 | 3RGF | 15943484 |
| 659 | P08476 | P27040 | 6.87E-09 | 8.16 | 1S4Y | 15943484 |
| 660 | Q6XVZ2 | P18206 | 6.61E-09 | 8.18 | 2HSQ | 15943484 |
| 661 | O61667 | P41958 | 6.40E-09 | 8.19 | 1TY4 | 15943484 |
| 662 | Q9D777 | O14836 | 6.40E-09 | 8.19 | 1XU1 | 21213247 |

Appendices

| | | | | | | |
|-----|--------|--------|----------|------|------|--------------------|
| 663 | Q9N6S8 | P01038 | 6.50E-09 | 8.19 | 1YVB | 21213247 |
| 664 | P16471 | P01236 | 6.50E-09 | 8.19 | 3D48 | 21889455, 17785459 |
| 665 | P10493 | Q05793 | 6.00E-09 | 8.22 | 1GL4 | 21287608 |
| 666 | P49137 | Q16539 | 6.00E-09 | 8.22 | 2ONL | 15943484 |
| 667 | Q6RJQ3 | O14763 | 5.96E-09 | 8.22 | 4I9X | 25301850 |
| 668 | P04275 | P07359 | 5.80E-09 | 8.24 | 1M10 | 21213247 |
| 669 | P02751 | P80188 | 5.77E-09 | 8.24 | 4GH7 | 15943484 |
| 670 | P00760 | Q1EG59 | 5.60E-09 | 8.25 | 2UUY | 21213247 |
| 671 | P01241 | P10912 | 5.50E-09 | 8.26 | 1HWH | 21287608 |
| 672 | Q9X0C6 | Q9X0C8 | 5.00E-09 | 8.30 | 1GPW | 21213247 |
| 673 | P01009 | P00760 | 5.00E-09 | 8.30 | 1OPH | 21213247 |
| 674 | P09527 | P37727 | 5.00E-09 | 8.30 | 1VG0 | 15943484 |
| 675 | P14280 | P83326 | 5.00E-09 | 8.30 | 1XG2 | 15722470 |
| 676 | Q05489 | Q05490 | 5.00E-09 | 8.30 | 2ES4 | 15943484 |
| 677 | Q9W2R4 | Q9VB22 | 5.00E-09 | 8.30 | 4A1S | 15943484 |
| 678 | P00730 | P01075 | 5.00E-09 | 8.30 | 4CPA | 21213247 |
| 679 | Q6X1E6 | Q7KQK5 | 4.80E-09 | 8.32 | 2Z8V | 15943484 |
| 680 | P01241 | P16471 | 4.70E-09 | 8.33 | 1BP3 | 21287608 |
| 681 | Q6X1E6 | Q7KQK5 | 4.70E-09 | 8.33 | 2Z8W | 15943484 |
| 682 | P07596 | P29600 | 4.50E-09 | 8.35 | 3BX1 | 15943484 |
| 683 | O88513 | Q8R4E9 | 4.40E-09 | 8.36 | 2ZXX | 15943484 |
| 684 | P27884 | P62158 | 4.32E-09 | 8.36 | 3DVM | 15943484 |
| 685 | P62826 | P49792 | 4.30E-09 | 8.37 | 1RRP | 21287608 |
| 686 | P50456 | P69786 | 4.10E-09 | 8.39 | 3BP8 | 21213247 |
| 687 | P17150 | Q13651 | 4.00E-09 | 8.40 | 1LQS | 15943484 |
| 688 | P19878 | Q15080 | 4.00E-09 | 8.40 | 1OEY | 15943484 |
| 689 | P26447 | P35579 | 4.00E-09 | 8.40 | 2LNK | 15943484 |
| 690 | P08603 | Q9JXV4 | 4.00E-09 | 8.40 | 2W81 | 15943484 |
| 691 | P57740 | Q8WUM0 | 4.00E-09 | 8.40 | 3CQC | PMC2446439 |
| 692 | A9UTG5 | A9V0L3 | 3.90E-09 | 8.41 | 2XHE | 15943484 |
| 693 | O14893 | Q16637 | 3.30E-09 | 8.48 | 2LEH | 15943484 |
| 694 | Q07817 | Q9BXH1 | 3.00E-09 | 8.52 | 4HNJ | 25301850 |
| 695 | P62593 | P35804 | 2.80E-09 | 8.55 | 1JTG | 21287608 |
| 696 | P00730 | P81511 | 2.80E-09 | 8.55 | 2ABZ | 21213247 |
| 697 | P03528 | Q9Y624 | 2.80E-09 | 8.55 | 3EOY | 15943484 |
| 698 | P00760 | Q9S9F3 | 2.69E-09 | 8.57 | 3RDZ | 15943484 |
| 699 | P05221 | P52293 | 2.70E-09 | 8.57 | 3UL1 | 15943484 |
| 700 | Q8U094 | Q8U093 | 2.50E-09 | 8.60 | 1WDW | 21213247 |
| 701 | P49137 | P47811 | 2.50E-09 | 8.60 | 2OZA | 21213247 |
| 702 | P36894 | P12643 | 2.40E-09 | 8.62 | 2QJ9 | 15943484 |
| 703 | P81274 | Q1MX18 | 2.40E-09 | 8.62 | 3SF4 | 15943484 |
| 704 | P52293 | Q3UYV9 | 2.40E-09 | 8.62 | 3UKZ | 15943484 |
| 705 | Q9I2Q0 | Q9I2Q1 | 2.42E-09 | 8.62 | 4EQA | 15943484 |
| 706 | O43566 | P63096 | 2.30E-09 | 8.64 | 3ONW | 15943484 |
| 707 | Q9BUL8 | Q9P289 | 2.15E-09 | 8.67 | 3W8I | 25301850 |
| 708 | P12272 | Q14974 | 2.10E-09 | 8.68 | 1M5N | 15943484 |
| 709 | P25054 | P35222 | 2.10E-09 | 8.68 | 1TH1 | 15943484 |
| 710 | P68135 | P00639 | 2.00E-09 | 8.70 | 1ATN | 21287608 |

Appendices

| | | | | | | |
|-----|--------|--------|----------|------|------|--------------------|
| 711 | P32499 | Q02821 | 2.00E-09 | 8.70 | 2C1T | 15943484 |
| 712 | P62161 | Q05586 | 2.00E-09 | 8.70 | 2HQQ | 15943484 |
| 713 | P29340 | P39718 | 2.00E-09 | 8.70 | 2Y9M | 15943484 |
| 714 | Q07820 | O43521 | 2.00E-09 | 8.70 | 3KJ0 | 20066663 |
| 715 | P08603 | Q19KF7 | 2.00E-09 | 8.70 | 4AYI | 15943484 |
| 716 | O75473 | Q2MKA7 | 2.00E-09 | 8.70 | 4BSR | 25301850 |
| 717 | P05113 | Q01344 | 1.90E-09 | 8.72 | 3QT2 | 15943484 |
| 718 | Q16552 | Q96F46 | 1.90E-09 | 8.72 | 4HSA | 25301850 |
| 719 | Q2KIG3 | Q5EPH2 | 1.80E-09 | 8.74 | 3OSL | 15943484 |
| 720 | O43707 | P18206 | 1.78E-09 | 8.75 | 1YDI | 15943484 |
| 721 | O14745 | P26043 | 1.70E-09 | 8.77 | 2D10 | 15943484 |
| 722 | Q8I6U4 | Q966X9 | 1.70E-09 | 8.77 | 2OUL | 21213247 |
| 723 | P62993 | Q07889 | 1.48E-09 | 8.83 | 1GBQ | 7566970 |
| 724 | P01024 | P68799 | 1.40E-09 | 8.85 | 2GOX | 21213247 |
| 725 | P61765 | P32851 | 1.40E-09 | 8.85 | 4JEU | 25301850 |
| 726 | P15086 | Q5EPH2 | 1.30E-09 | 8.89 | 1ZLI | 21213247 |
| 727 | Q93IS4 | Q8ZRL5 | 1.28E-09 | 8.89 | 4J32 | 25301850 |
| 728 | P10912 | P01241 | 1.20E-09 | 8.92 | 1A22 | 21889455, 15147191 |
| 729 | P0CE48 | P0A6P1 | 1.10E-09 | 8.96 | 1EFU | 21287608 |
| 730 | Q8H0K8 | P18429 | 1.10E-09 | 8.96 | 2B42 | 21213247 |
| 731 | P52293 | Q9JIH2 | 1.10E-09 | 8.96 | 2C1M | 15943484 |
| 732 | P78310 | Q65914 | 1.10E-09 | 8.96 | 2JIK | 15943484 |
| 733 | P12643 | P36894 | 1.10E-09 | 8.96 | 2QJA | 15943484 |
| 734 | P70444 | Q07440 | 1.10E-09 | 8.96 | 2VOI | 15943484 |
| 735 | P0AD64 | P35804 | 1.10E-09 | 8.96 | 3N4I | 15943484 |
| 736 | P56634 | P80403 | 1.00E-09 | 9.00 | 1CLV | 15943484 |
| 737 | P00772 | P19957 | 1.00E-09 | 9.00 | 1FLE | 21213247 |
| 738 | P02774 | P68135 | 1.00E-09 | 9.00 | 1KXP | 23055910, 2910852 |
| 739 | P03079 | P06239 | 1.00E-09 | 9.00 | 1LCJ | 15943484 |
| 740 | Q02248 | Q9NSA3 | 1.00E-09 | 9.00 | 1MIE | 15943484 |
| 741 | P05121 | P04004 | 1.00E-09 | 9.00 | 1OC0 | 12808446, 9065424 |
| 742 | P35804 | P62593 | 1.00E-09 | 9.00 | 1S0W | 9890878 |
| 743 | P13393 | P51123 | 1.00E-09 | 9.00 | 1TBA | 15943484 |
| 744 | P22301 | Q13651 | 1.00E-09 | 9.00 | 1Y6K | 15837194 |
| 745 | P0DKX7 | P62158 | 1.00E-09 | 9.00 | 1YRT | 15943484 |
| 746 | P0DKX7 | P62158 | 1.00E-09 | 9.00 | 1YRU | 15943484 |
| 747 | P09060 | P09061 | 1.00E-09 | 9.00 | 2BP7 | 15943484 |
| 748 | P00698 | Q8AXH5 | 1.00E-09 | 9.00 | 2I25 | 21213247 |
| 749 | Q9Y6N9 | Q495M9 | 1.00E-09 | 9.00 | 3K1R | 20142502 |
| 750 | P50983 | Q8WSF8 | 8.80E-10 | 9.06 | 2BYP | 15943484 |
| 751 | P93343 | Q40409 | 8.50E-10 | 9.07 | 2O98 | 15943484 |
| 752 | P21802 | P05230 | 7.91E-10 | 9.10 | 1DJS | 21287608 |
| 753 | O00330 | P09622 | 7.80E-10 | 9.11 | 2F5Z | 15943484 |
| 754 | Q1PIV4 | P23371 | 7.55E-10 | 9.12 | 2GAF | 21287608 |
| 755 | P61326 | Q9Y5S9 | 7.00E-10 | 9.15 | 1P27 | 15943484 |
| 756 | P47224 | P55258 | 7.00E-10 | 9.15 | 2FU5 | 15943484 |
| 757 | Q07440 | Q99ML1 | 7.00E-10 | 9.15 | 2VOF | 15943484 |
| 758 | Q6UVW9 | D3W0D1 | 6.70E-10 | 9.17 | 4IOP | 25301850 |

Appendices

| | | | | | | |
|-----|--------|--------|----------|-------|------|----------|
| 759 | Q07817 | Q92934 | 6.00E-10 | 9.22 | 1G5J | 15943484 |
| 760 | P12830 | Q4EQX8 | 6.00E-10 | 9.22 | 2OMU | 15943484 |
| 761 | P05132 | P12369 | 6.00E-10 | 9.22 | 3IDC | 15943484 |
| 762 | Q53176 | Q53177 | 5.00E-10 | 9.30 | 1OGY | 15943484 |
| 763 | O41925 | P47992 | 5.00E-10 | 9.30 | 2NYZ | 21213247 |
| 764 | P38526 | Q9WZF8 | 4.70E-10 | 9.33 | 3GTY | 15943484 |
| 765 | O34800 | O34853 | 4.60E-10 | 9.34 | 3O6Q | 15943484 |
| 766 | P09038 | P21802 | 3.81E-10 | 9.42 | 1IIL | 21287608 |
| 767 | Q5PY49 | Q03405 | 3.30E-10 | 9.48 | 2I9B | 21213247 |
| 768 | P0C1S6 | Q9EYW6 | 3.10E-10 | 9.51 | 1PXV | 21213247 |
| 769 | Q93IS4 | Q8ZRL5 | 2.69E-10 | 9.57 | 4HFF | 25301850 |
| 770 | D5C6F6 | D5C6F7 | 2.69E-10 | 9.57 | 4HFK | 25301850 |
| 771 | P34130 | Q16620 | 2.60E-10 | 9.59 | 1HCF | 21213247 |
| 772 | Q4KC90 | Q4KC91 | 2.60E-10 | 9.59 | 4KT3 | 25301850 |
| 773 | P00766 | P01051 | 2.00E-10 | 9.70 | 1ACB | 21213247 |
| 774 | P05798 | P11540 | 2.00E-10 | 9.70 | 1AY7 | 21213247 |
| 775 | P00766 | P80060 | 2.00E-10 | 9.70 | 1GL1 | 21213247 |
| 776 | Q14116 | Q9DHU8 | 2.00E-10 | 9.70 | 4EEE | 15943484 |
| 777 | O00206 | Q9Y6Y9 | 1.86E-10 | 9.73 | 4G8A | 15943484 |
| 778 | P06886 | A0A5B4 | 1.80E-10 | 9.74 | 2IJ0 | 15943484 |
| 779 | C7B6Y3 | P00698 | 1.80E-10 | 9.74 | 3M18 | 15943484 |
| 780 | P49763 | P17948 | 1.70E-10 | 9.77 | 1RV6 | 21213247 |
| 781 | P13423 | P58335 | 1.70E-10 | 9.77 | 1T6B | 21213247 |
| 782 | P06869 | P35456 | 1.70E-10 | 9.77 | 3LAQ | 15943484 |
| 783 | P05112 | P24394 | 1.60E-10 | 9.80 | 1IAR | 21287608 |
| 784 | P02768 | Q51911 | 1.50E-10 | 9.82 | 2VDB | 21213247 |
| 785 | P62826 | Q14974 | 1.40E-10 | 9.85 | 1IBR | 21287608 |
| 786 | P00760 | P01055 | 1.30E-10 | 9.89 | 1D6R | 15943484 |
| 787 | P0C8E7 | P10147 | 1.20E-10 | 9.92 | 3FPU | 20041127 |
| 788 | P18010 | P18206 | 1.10E-10 | 9.96 | 2GWW | 15943484 |
| 789 | P00791 | P19400 | 1.00E-10 | 10.00 | 1F34 | 15943484 |
| 790 | P01579 | P15260 | 1.00E-10 | 10.00 | 1FG9 | 15943484 |
| 791 | P09883 | Q03708 | 1.00E-04 | 4.00 | - | 15019791 |
| 792 | Q16539 | P15336 | 6.20E-05 | 4.21 | - | 16156785 |
| 793 | P0ADV1 | P0ADV9 | 1.50E-05 | 4.82 | - | 24123237 |
| 794 | Q96RD9 | P01860 | 1.06E-05 | 4.97 | - | 23616577 |
| 795 | P28482 | P14921 | 7.30E-06 | 5.14 | - | 20361728 |
| 796 | P62161 | Q13158 | 2.00E-06 | 5.70 | - | 23760276 |
| 797 | Q96RD9 | P01857 | 1.75E-06 | 5.76 | - | 23616577 |
| 798 | Q96RD9 | P01859 | 1.37E-06 | 5.86 | - | 23616577 |
| 799 | P50454 | P02452 | 1.14E-06 | 5.94 | - | 7983065 |
| 800 | P50454 | P20908 | 8.73E-07 | 6.06 | - | 7983065 |
| 801 | P50454 | P02458 | 7.17E-07 | 6.14 | - | 7983065 |
| 802 | P50454 | P02461 | 7.12E-07 | 6.15 | - | 7983065 |
| 803 | P52193 | P19137 | 5.00E-07 | 6.30 | - | 8626465 |
| 804 | Q02388 | P02452 | 5.00E-07 | 6.30 | - | 16563355 |
| 805 | Q02388 | P08123 | 5.00E-07 | 6.30 | - | 16563355 |
| 806 | P09883 | P09881 | 5.00E-07 | 6.30 | - | 15019791 |

Appendices

| | | | | | | |
|-----|--------|--------|----------|------|---|----------|
| 807 | P60709 | P21333 | 4.62E-07 | 6.34 | - | 8282102 |
| 808 | P60709 | P12814 | 4.00E-07 | 6.40 | - | 8282102 |
| 809 | P50454 | P02462 | 3.83E-07 | 6.42 | - | 7983065 |
| 810 | P02766 | P02795 | 2.44E-07 | 6.61 | - | 18237193 |
| 811 | Q07817 | P00004 | 1.20E-07 | 6.92 | - | 17905676 |
| 812 | P09882 | P13479 | 6.40E-08 | 7.19 | - | 15019791 |
| 813 | Q02388 | O15230 | 6.00E-08 | 7.22 | - | 16563355 |
| 814 | O14745 | P15311 | 5.80E-08 | 7.24 | - | 19857202 |
| 815 | P04419 | P09881 | 3.80E-08 | 7.42 | - | 15019791 |
| 816 | P60953 | O00401 | 2.70E-08 | 7.57 | - | 19293156 |
| 817 | Q2TAM5 | B2VQE1 | 2.60E-08 | 7.59 | - | 9023117 |
| 818 | Q2TAM5 | Q00403 | 2.30E-08 | 7.64 | - | 9023117 |
| 819 | P01034 | A5HIII | 1.90E-08 | 7.72 | - | 8718861 |
| 820 | P09882 | P04482 | 1.80E-08 | 7.74 | - | 15019791 |
| 821 | P04419 | Q03708 | 1.40E-08 | 7.85 | - | 15019791 |
| 822 | P56464 | O25928 | 1.20E-08 | 7.92 | - | 17049879 |
| 823 | P04419 | P13479 | 1.20E-08 | 7.92 | - | 15019791 |
| 824 | Q02388 | Q02388 | 4.00E-09 | 8.40 | - | 16563355 |
| 825 | P02774 | P63258 | 2.80E-09 | 8.55 | - | 2910852 |
| 826 | P02774 | P60712 | 2.60E-09 | 8.59 | - | 2910852 |
| 827 | Q02388 | P53420 | 2.00E-09 | 8.70 | - | 16563355 |
| 828 | P80416 | P07858 | 1.90E-09 | 8.72 | - | 7875311 |
| 829 | P00748 | Q9S879 | 1.20E-09 | 8.92 | - | 24336918 |
| 830 | P09882 | Q03708 | 1.00E-09 | 9.00 | - | 15019791 |
| 831 | P63241 | P49366 | 5.00E-10 | 9.30 | - | 10229683 |
| 832 | P01034 | P09668 | 4.20E-10 | 9.38 | - | 8718861 |
| 833 | P80416 | P09668 | 4.00E-10 | 9.40 | - | 7875311 |
| 834 | P04275 | P00451 | 4.00E-10 | 9.40 | - | 8885147 |
| 835 | Q47112 | P09881 | 3.70E-10 | 9.43 | - | 15019791 |
| 836 | Q47112 | P04482 | 3.60E-10 | 9.44 | - | 15019791 |
| 837 | Q2TAM5 | P20226 | 3.40E-10 | 9.47 | - | 9023117 |
| 838 | P01034 | P07858 | 3.20E-10 | 9.49 | - | 8718861 |
| 839 | P08246 | P19957 | 1.67E-10 | 9.78 | - | 8439544 |

Section D: Protein-Protein Interaction Dataset for Association

Rate Constant (k_{on}) and Dissociation Rate Constant (k_{off})

Table S-5 Protein-protein interaction dataset for studying association rate constant (k_{on}) and dissociation rate constants (k_{off})

| NO. | Protein A Uniprot | Protein B Uniprot | k_{on} | $\log(k_{on})$ | k_{off} | $\log(k_{off})$ | PDB ID | References (PMID) |
|-----|----------------------|----------------------|----------|----------------|-----------|-----------------|--------|----------------------|
| 1 | P40189 | Q98823 | 9.80E+02 | 2.99 | 2.20E-03 | -2.66 | 1IIR | 21287608 |
| 2 | P08165 | P00257 | 4.43E+03 | 3.65 | 3.80E-03 | -2.42 | 1E6E | 21287608 |
| 3 | P08254 | P01033 | 5.70E+03 | 3.76 | 5.00E-04 | -3.30 | 1UEA | 21287608 |
| 4 | P00747 | P00779 | 1.17E+04 | 4.07 | 2.30E-03 | -2.64 | 1L4D | 21287608 |
| 5 | P68135 | P02774 | 2.20E+04 | 4.34 | 1.10E-05 | -4.96 | 1KXP | 21287608 |
| 6 | P02774 | P68135 | 2.20E+04 | 4.34 | 2.20E-05 | -4.66 | 1KXP | 23055910, 2910852 |
| 7 | P08253 | P16035 | 2.25E+04 | 4.35 | 1.60E-03 | -2.80 | 1GXD | 21287608 |
| 8 | P55075 | P21802 | 3.76E+04 | 4.58 | 5.84E-03 | -2.23 | 2FDB | 21287608 |
| 9 | P01241 | P16471 | 4.23E+04 | 4.63 | 2.00E-04 | -3.70 | 1BP3 | 21287608 |
| 10 | O88653 | Q9JHS3 | 4.68E+04 | 4.67 | 5.97E-04 | -3.22 | 1VET | 21287608 |
| 11 | P62826 | P49792 | 5.80E+04 | 4.76 | 2.50E-04 | -3.60 | 1RRP | 21287608 |
| 12 | P26718 | Q29983 | 6.75E+04 | 4.83 | 3.90E-02 | -1.41 | 1HYR | 21287608 |
| 13 | P36711 | P78310 | 7.31E+04 | 4.86 | 1.10E-03 | -2.96 | 1KAC | 21287608 |
| 14 | P20701 | P32942 | 7.38E+04 | 4.87 | 1.62E-01 | -0.79 | 1T0P | 21287608 |
| 15 | P62826 | Q14974 | 8.50E+04 | 4.93 | 1.16E-05 | -4.94 | 1IBR | 21287608 |
| 16 | P01241 | P10912 | 9.30E+04 | 4.97 | 5.50E-04 | -3.26 | 1HWH | 21287608 |
| 17 | P09038 | P11362 | 9.64E+04 | 4.98 | 5.96E-03 | -2.22 | 1CVS | 21287608 |
| 18 | P62593 | P35804 | 1.19E+05 | 5.08 | 3.30E-04 | -3.48 | 1JTG | 21287608 |
| 19 | P05230 | P11362 | 2.24E+05 | 5.35 | 3.05E-02 | -1.52 | 1EVT | 21287608 |
| 20 | P10493 | Q05793 | 3.60E+05 | 5.56 | 2.10E-03 | -2.68 | 1GL4 | 21287608 |
| 21 | O14763 | P50591 | 3.91E+05 | 5.59 | 3.81E-03 | -2.42 | 1DU3 | 21287608 |
| 22 | P00747 | P00779 | 4.52E+05 | 5.66 | 3.56E-03 | -2.45 | 1BML | 21287608 |
| 23 | Q1PIV4 | P23371 | 5.30E+05 | 5.72 | 4.00E-04 | -3.40 | 2GAF | 21287608 |
| 24 | P21802 | P05230 | 8.02E+05 | 5.90 | 6.35E-04 | -3.20 | 1DJS | 21287608 |
| 25 | P05230 | P22607 | 8.79E+05 | 5.94 | 2.02E-01 | -0.69 | 1RY7 | 21287608 |
| 26 | P68135 | P00639 | 1.00E+06 | 6.00 | 2.00E-03 | -2.70 | 1ATN | 21287608 |
| 27 | P09038 | P21802 | 1.18E+06 | 6.07 | 4.51E-04 | -3.35 | 1IIL | 21287608 |
| 28 | O97428 | P68135 | 1.20E+06 | 6.08 | 9.10E+00 | 0.96 | 1SQK | 23055910, 15163400 |
| 29 | P09372 | P0A6Y8 | 1.30E+06 | 6.11 | 4.60E-02 | -1.34 | 1DKG | 21287608 |
| 30 | P09038 | P21802 | 1.33E+06 | 6.12 | 6.52E-04 | -3.19 | 1EV2 | 21287608 |
| 31 | Q91YR1 | P68135 | 3.83E+06 | 6.58 | 1.80E+00 | 0.26 | 3DAW | 23055910, 12429826 |
| 32 | P60568 | P01589 | 7.80E+06 | 6.89 | 2.35E-01 | -0.63 | 1Z92 | 21287608 |
| 33 | P0CE48 | P0A6P1 | 1.00E+07 | 7.00 | 3.00E-02 | -1.52 | 1EFU | 21287608 |
| 34 | P05112 | P24394 | 1.30E+07 | 7.11 | 2.10E-03 | -2.68 | 1IAR | 21287608 |
| 35 | P02584 | P60712 | 1.40E+07 | 7.15 | 1.82E-01 | -0.74 | 2BTF | 23055910, 11052670 |
| 36 | P13479 | P09883 | 1.70E+07 | 7.23 | 6.11E-03 | -2.21 | 2GYK | 21287608 |

Appendices

| | | | | | | | | |
|----|--------|--------|----------|------|----------|-------|------|--------------------|
| 37 | P00573 | P00806 | 3.80E+07 | 7.58 | 3.50E+00 | 0.54 | 1ARO | 21287608 |
| 38 | P00639 | P68135 | 4.30E+07 | 7.63 | 2.58E+01 | 1.41 | 2A3Z | 23055910, 11146629 |
| 39 | P0AE67 | P07363 | 6.30E+07 | 7.80 | 2.20E+01 | 1.34 | 1A0O | 21287608 |
| 40 | P49137 | P47811 | 6.60E+07 | 7.82 | 2.20E-01 | -0.66 | 2OZA | 21287608 |
| 41 | P00431 | P00044 | 3.00E+09 | 9.48 | 1.90E+03 | 3.28 | 2B10 | 21287608 |
| 42 | P50454 | P20908 | 1.89E+04 | 4.28 | 1.65E-02 | -1.78 | - | 7983065 |
| 43 | P50454 | P02452 | 2.08E+04 | 4.32 | 2.36E-02 | -1.63 | - | 7983065 |
| 44 | P50454 | P02461 | 2.18E+04 | 4.34 | 1.55E-02 | -1.81 | - | 7983065 |
| 45 | P50454 | P02458 | 2.38E+04 | 4.38 | 1.71E-02 | -1.77 | - | 7983065 |
| 46 | P50454 | P02462 | 3.06E+04 | 4.49 | 1.17E-02 | -1.93 | - | 7983065 |
| 47 | P35804 | Q6SJ61 | 4.20E+04 | 4.62 | 6.00E-04 | -3.22 | - | 9890878 |
| 48 | Q2TAM5 | B2VQE1 | 4.90E+04 | 4.69 | 1.30E-03 | -2.89 | - | 9023117 |
| 49 | Q2TAM5 | Q00403 | 6.80E+04 | 4.83 | 1.60E-03 | -2.80 | - | 9023117 |
| 50 | P16471 | P01236 | 8.00E+04 | 4.90 | 5.00E-04 | -3.30 | - | 21889455, 17785459 |
| 51 | P62993 | Q07889 | 9.45E+04 | 4.98 | 1.38E-04 | -3.86 | - | 7566970 |
| 52 | P80416 | P07858 | 1.40E+05 | 5.15 | 2.66E-04 | -3.58 | - | 7875311 |
| 53 | P52193 | P19137 | 2.00E+05 | 5.30 | 1.00E-01 | -1.00 | - | 8626465 |
| 54 | Q07817 | P00004 | 2.50E+05 | 5.40 | 3.00E-02 | -1.52 | - | 17905676 |
| 55 | P23827 | P03952 | 2.90E+05 | 5.46 | 6.30E-05 | -4.20 | - | 7781771 |
| 56 | P10912 | P01241 | 3.20E+05 | 5.51 | 3.90E-04 | -3.41 | - | 21889455, 15147191 |
| 57 | Q02388 | O15230 | 4.50E+05 | 5.65 | 2.70E-02 | -1.57 | - | 16563355 |
| 58 | P00748 | Q9S879 | 5.00E+05 | 5.70 | 6.00E-04 | -3.22 | - | 24336918 |
| 59 | P60709 | P12814 | 1.00E+06 | 6.00 | 4.00E-01 | -0.40 | - | 8282102 |
| 60 | P01034 | P07858 | 1.10E+06 | 6.04 | 3.50E-04 | -3.46 | - | 8718861 |
| 61 | P60709 | P21333 | 1.30E+06 | 6.11 | 6.00E-01 | -0.22 | - | 8282102 |
| 62 | P00533 | P01133 | 1.63E+06 | 6.21 | 3.00E-02 | -1.52 | - | 10840042 |
| 63 | P62328 | P68135 | 1.70E+06 | 6.23 | 1.40E+00 | 0.15 | - | 23055910, 18327913 |
| 64 | P00533 | P01135 | 2.08E+06 | 6.32 | 4.00E-02 | -1.40 | - | 10840042 |
| 65 | P80416 | P09668 | 2.10E+06 | 6.32 | 8.40E-04 | -3.08 | - | 7875311 |
| 66 | Q2TAM5 | P20226 | 2.30E+06 | 6.36 | 7.90E-04 | -3.10 | - | 9023117 |
| 67 | P01034 | A5HIII | 2.40E+06 | 6.38 | 4.60E-02 | -1.34 | - | 8718861 |
| 68 | P04275 | P00451 | 3.00E+06 | 6.48 | 1.20E-03 | -2.92 | - | 8885147 |
| 69 | Q02388 | P02452 | 3.10E+06 | 6.49 | 1.50E+00 | 0.18 | - | 16563355 |
| 70 | Q02388 | P08123 | 3.10E+06 | 6.49 | 1.50E+00 | 0.18 | - | 16563355 |
| 71 | P08246 | P19957 | 3.60E+06 | 6.56 | 6.00E-04 | -3.22 | - | 8439544 |
| 72 | P28482 | P14921 | 8.08E+06 | 6.91 | 5.90E+01 | 1.77 | - | 20361728 |
| 73 | P06396 | P68135 | 2.00E+07 | 7.30 | 4.00E-01 | -0.40 | - | 8987989 |
| 74 | Q02388 | Q02388 | 4.20E+07 | 7.62 | 1.70E-01 | -0.77 | - | 16563355 |
| 75 | P09883 | P09881 | 6.00E+07 | 7.78 | 2.82E+01 | 1.45 | - | 15019791 |
| 76 | P09883 | P04482 | 8.00E+07 | 7.90 | 8.00E-01 | -0.10 | - | 15019791 |
| 77 | Q02388 | P53420 | 1.40E+08 | 8.15 | 2.80E-01 | -0.55 | - | 16563355 |
| 78 | Q47112 | P13479 | 1.60E+08 | 8.20 | 5.90E+00 | 0.77 | - | 15019791 |
| 79 | Q47112 | P04482 | 2.30E+08 | 8.36 | 1.00E-01 | -1.00 | - | 15019791 |
| 80 | P04419 | P13479 | 3.30E+08 | 8.52 | 4.10E+00 | 0.61 | - | 15019791 |
| 81 | Q47112 | P09881 | 4.10E+08 | 8.61 | 2.00E-01 | -0.70 | - | 15019791 |
| 82 | P04419 | Q03708 | 5.20E+08 | 8.72 | 7.30E+00 | 0.86 | - | 15019791 |
| 83 | P09882 | P04482 | 5.90E+08 | 8.77 | 1.05E+01 | 1.02 | - | 15019791 |
| 84 | P09882 | P13479 | 6.20E+08 | 8.79 | 3.99E+01 | 1.60 | - | 15019791 |
| 85 | P09882 | Q03708 | 7.80E+08 | 8.89 | 8.00E-01 | -0.10 | - | 15019791 |
| 86 | P04419 | P09881 | 7.90E+08 | 8.90 | 3.00E+01 | 1.48 | - | 15019791 |

