大理学院毕业设计

基于蛋白质相互作用网络的数据挖掘

Data mining based on protein-protein interaction network

学　　院：工程学院

项目组成员：浦恩禄、张孟娇、吴高林

专　　业：生物医学工程

年级（班级）：2011 级（1）班

指 导 教 师：张俊鹏

起 止 日 期：2014-09-15 至 2015-05-20

制表日期：2014 年 9 月 15 日

[摘要]：

蛋白质相互作用是所有活体细胞相互作用组学系统的核心，蛋白质相互作用网络和它们包含的信息为疾病的发生机理提供了更好的解释。因此，蛋白质相互作用网络的挖掘有助于研究活体细胞的生物机理。基于此，本设计选用 HPRD、HIR 和 BioGRID 三个具有代表性的 PPI 数据库作为蛋白质相互作用网络数据源。为了挖掘蛋白质相互作用网络的网络模体，我们使用 CyClus3D 对其进行 3 点网络模体挖掘。为了理解网络模体所组成的子网络是否具有生物学意义，在线工具 DAVID 用来进行 GO(Gene Ontology)生物过程和 KEGG(Kyoto Encyclopedia of Genes and Genomes)信号通道分析。实验结果表明：网络模体挖掘能够简化蛋白质相互作用网络的分析，并且能够集中分析关键性的蛋白质。

[关键词]：蛋白质；蛋白质相互作用；网络模体；富集分析

[Abstract]:

Protein-Protein Interactions(PPIs) are at the core of the entire interactomics system of all living cells, and PPI networks and the information they contain provide an insight into the causes of diseases. Thus, mining PPI networks could facilitate understanding biological mechanism within living cells. In this work, we select three representative PPI databases (HPRD, HIR, and BioGRID) as data source of PPI networks. To identify 3-node network motifs of these PPI networks, CyClus3D software is used. To further understand the underlying biological meaning of sub-network consisting of 3-node network motifs, the online tool DAVID is used to conduct GO (Gene Ontology) biological process and KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways at significant level. The results show that network motif could reduce analysis of PPI networks, and could focus on analyzing key proteins.

[Key Words]: Protein; Protein-protein interaction; Network motif; Enrichment analysis

目录

项目组成员分工

学生姓名	学号	具体分工
浦恩禄	2011166134	基于蛋白质相互作用网络的模体挖掘
张孟娇	2011166123	基于蛋白质相互作用网络的富集分析
吴高林	2011166106	蛋白质相互作用网络数据的收集与分析

0 引言

近年来，生物学家认识到生物体生命活动的绝大多数过程是由许许多多的生物分子在特定条件下共同相互作用所引起的，借助分子相互作用所形成的生物网络，对影响某些生物过程的诸多特性进行具体的研究，能够了解自然界生物体的奥秘。生物体细胞的进化过程，其实是由多种分子共同参与来完成的，这些分子有蛋白质，DNA 以及各种代谢物[1]。现如今，人们利用一些先进的生化技术手段，经过实验、计算等方法获得了许多生物分子之间相互作用数据，组成了庞大的生物网络，比如代谢路径(Metabolic Pathway, 简称 MP)，基因调控网络(Gene Regulatory Network, 简称 GRN)，蛋白质相互作用网络(Protein-Protein Interaction Network, 简称 PPIN)和信号传导网络(Signaling Network, 简称 SN)等[2]。目前，存放这些生物网络数据的数据库已经有了明显的发展，数据也正在不断更新和完善，其中蛋白质相互作用网络是这些生物网络数据中很重要的一类数据。

蛋白质(Protein)是生命的物质基础，作为生物体生命活动的直接执行者和生命现象的直接体现者，参与了生命的几乎所有过程，如遗传、发育、能量代谢、应激等[3]。尽管有一些蛋白质在完成一定功能时可以以独立的形式发挥作用，但是大部分的蛋白质都是与其他蛋白质形成复合物来发挥作用，即多个蛋白共同作用完成某一功能。蛋白质相互作用[4]（Protein-Protein Interaction，简称 PPI）在生物体的生命活动中起核心作用，它是 DNA 复制、转录、翻译、物质代谢、信号传导以及细胞周期控制等正常生理过程的基础，同时在一些病理的过程中也起着重要的作用。因此，对于 PPI 的研究将有利于我们更好的了解生命体中的生物过程。

随着系统生物学实验的快速发展，人们已经得到了大量基因组水平上的 PPI 数据，并且数据量还在持续增长。这些数据成为 PPIN 构建和生物学通路分析的主要资源。根据这些数据库中提取的 PPI 数据，我们可以进一步从这些数据所形成网络的拓扑特性来研究蛋白质之间的联系。在 PPI 网络中，用一个节点(Node)来表示一个蛋白质，而连接两个节点的边(Edge)表示蛋白质之间的作用关系。

PPIN 同其他生物化学网络一样属于大型复杂网络，复杂网络的基本构造单元是网络模体。网络模体[5]（Network Motif）是一些在真实网络中重复出现的具有特定拓扑结构特征的小规模模式单元，它在真实网络中出现的频率要远高于随

机网络(Random Network)[6]，并且在网络的发展过程中扮演着重要的角色。因此，网络模体挖掘有利于揭示复杂网络的更多细节和内容。目前，挖掘网络模体的工具和算法主要有 CyClus3D[7]、NetMatch[8]、Mfinder[9]、MAVisto (Motif Analysis and Visualization tool)[10]和 NeMoFinder[11]等。对使用模体挖掘后的子网络进行深入的生物过程（Biological Process, BP）和信号通道（Signal Pathway, SP）显著性分析，将有助于深入理解生物体的内在作用机制。

为进一步研究网络模体所构成的子网络（Sub-network）是否有生物学意义，有必要对该子网络进行富集分析（Enrichment analysis）。富集分析主要分为 GO（Gene Ontology）[12]功能显著性分析和 KEGG（Kyoto Encyclopedia of Genes and Genomes）[13]信号通道显著性分析。GO 是由基因本体联合会(Gene Ontology Consortium)所建立的一个数据库，它主要包括生物过程 (Biological Process，简称 BP)、分子功能 (Molecular Function，简称 MF)和细胞组成(Cellular Component，简称 CC)三个分支。GO 动态词汇表包含了能描述所有真核生物的术语，已经被广泛运用于多种生物体的注释，分子功能以及蛋白质的生物作用等方面。KEGG 是一个关于基因组信息、基因功能分析、细胞过程、酶促反应以及生物化学物质的在线数据库。目前，DAVID（http://david.abcc.ncifcrf.gov/）软件是富集分析的一个标准化在线软件。

PPI 网络的数据挖掘作为蛋白质组学的研究内容之一，近年来受到越来越多研究人员的关注。用高效的信息分析和数据挖掘手段从繁杂的蛋白质组数据中找出内在联系，对揭示蛋白质的功能及相互作用关系具有极其重要的意义。如图 0.1，本设计将从三个层次对蛋白质相互作用网络进行数据挖掘。首先，收集和清理蛋白质相互作用数据集，包括 HPRD[14]、HIR[15]、BioGRID[16]三个数据库。其次，在生物信息学软件平台 Cytoscape（2.8.2）中对蛋白质相互作用网络进行可视化，并且利用 CyClus3D 对其进行网络模体挖掘。最后，利用 DAVID 在线软件对网络模体所形成的子网络进行富集分析。

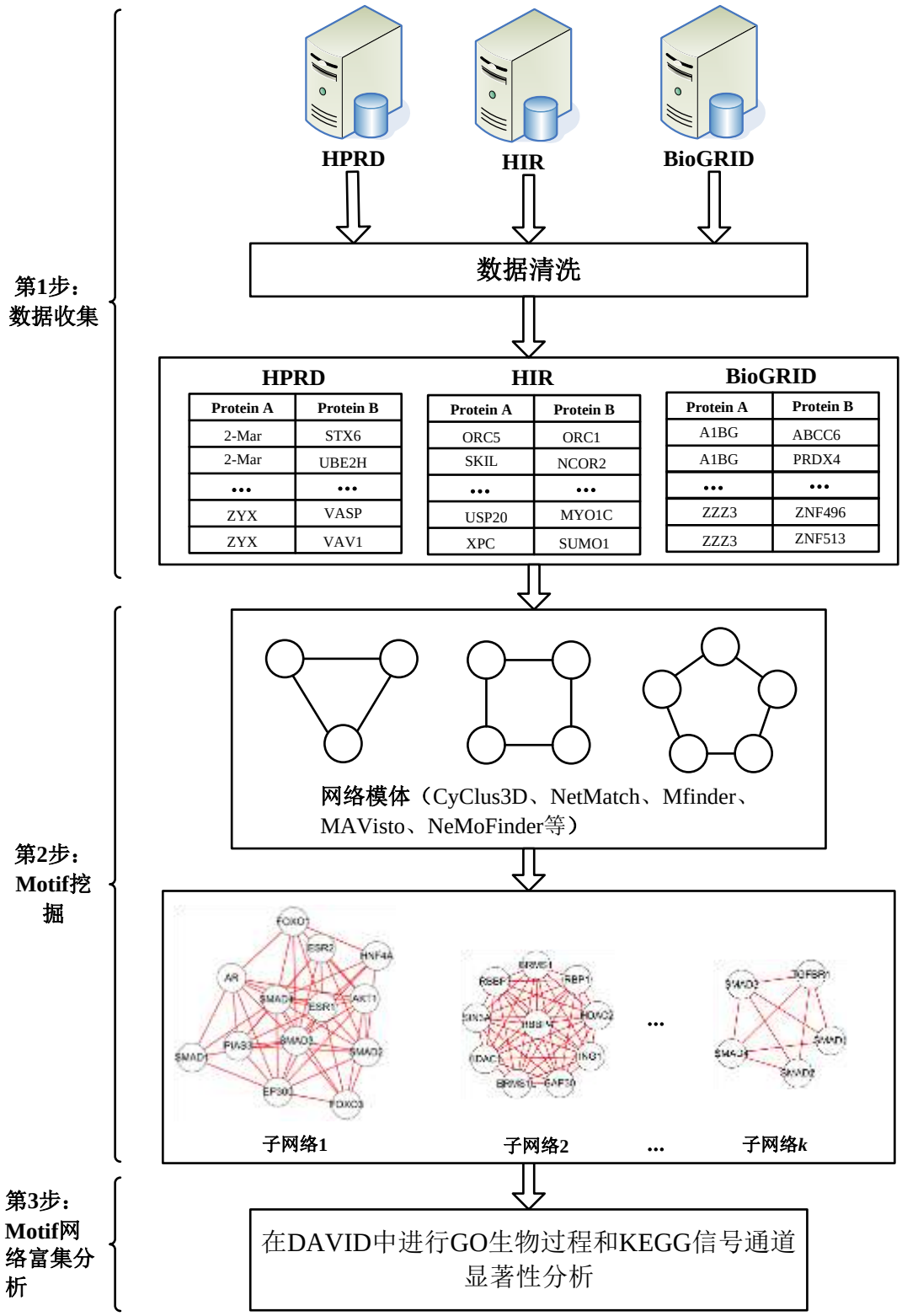


图 0.1 基于蛋白质相互作用网络的数据挖掘流程图

1 蛋白质相互作用数据

1.1 蛋白质相互作用概述

一个蛋白质相互作用(PPI)关系是指一个蛋白质与另一蛋白质之间的相互关系，它是细胞生命活动中最主要的组成部分。它不仅是正常生理过程如 DNA 复制、转录、翻译、物质代谢、信号传导以及细胞周期控制的基础，而且在病理过程中起着重要的作用[17,18]。可以说，生物体中几乎所有的生物过程都是通过蛋白质间的相互作用来精确地执行的。

蛋白质分子之间的相互作用方式以及形式是多种多样的，有蛋白质—蛋白质相互作用、蛋白质—DNA 相互作用、蛋白质—RNA 相互作用等。相互作用的蛋白质分子可以是直接接触(即物理相互作用)，也可以是不接触的，而仅仅表现为功能上关联(即遗传相互作用)。功能上的关联是指两个或多个蛋白质参与同一代谢途径或者它们都属于同一个生物学过程，因此这样关联的一些蛋白质具有相似或相同的生物功能。

1.2 蛋白质相互作用数据库

近几年，人们通过高通量实验方法筛选、计算方法预测以及文献挖掘等方法，得到了大量的 PPI 数据，并由此构建了很多内容丰富并且日益更新的 PPI 数据库，这些数据库为我们进一步深入研究 PPI 网络提供了数据支持，表 1.1 列出了一些常用的 PPI 数据库及其相关信息。这些海量的 PPI 数据为系统地探索生命体活动提供了丰富的数据来源的同时也面临着一些新的挑战，例如采用什么样的计算方法去处理、挖掘、分析、并有效利用这些数据。因此，如何从这些海量数据中挖掘出有效的生物信息成为生物医学领域的一大研究热点。

另一方面，每个 PPI 数据库都有自己独有的特性，并且数据库之间的 PPI 重合率很低，如何选择合适的 PPI 数据库往往通过经验。由于在选择 PPI 数据库方面没有统一的标准，本文根据 PPI 数据库在其他文献中使用的频繁程度，选择了 HPRD，HIR 和 BioGRID 三个数据库作为数据源。这三种数据库之间的 PPI 重合数如图 1.1 所示，从图中可知，三个数据库之间的 PPI 重合数只有 7631。

表 1.1 一些常用 PPI 数据库

数据库名称	PPIs 数量	蛋白质个数	PPI 数据类型	下载地址
HPRD	36,852	9,444	实验型	http://hprd.org/
HIR	69,586	11,366	实验和预测型	http://www.cls.zju.edu.cn/hir/faces/download.xhtml
BioGRID	168,054	16,522	实验和预测型	http://thebiogrid.org/download.php
DIP	1,067	804	实验型	http://dip.doe-mbi.ucla.edu/dip/Main.cgi
IntAct	10,244	4,614	实验型	http://www.ebi.ac.uk/intact/
MINT	11,367	4,975	实验型	http://mint.bio.uniroma2.it/mint/Welcome.do
REACTOME	5,960	970	实验和预测型	http://www.reactome.org/
PINA2	166,776	5,211	实验和预测型	http://cbg.garvan.unsw.edu.au/pina/home.do
ConsensusPathDB	221,328	154,537	实验和预测型	http://consensuspathdb.org/
HitPredict	239,584	49,071	实验和预测型	http://hintdb.hgc.jp/htp/

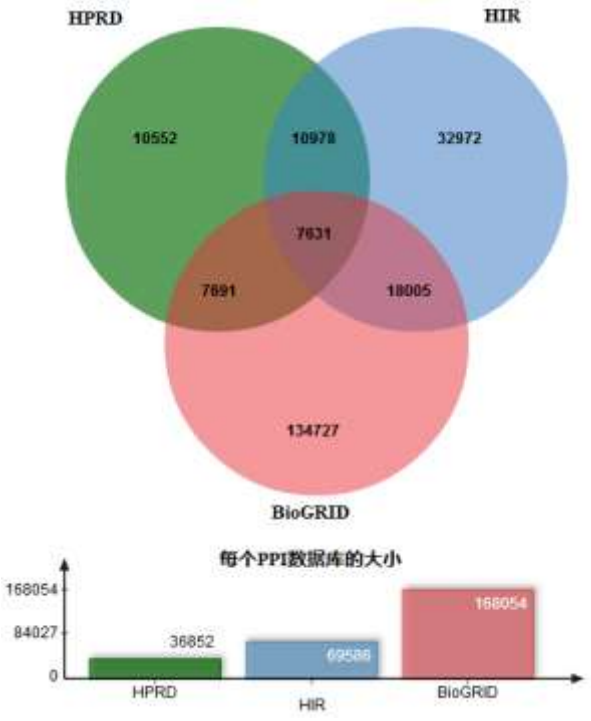


图 1.1 HPRD、HIR 和 BioGRID 三种 PPI 数据库之间的 PPI 重合个数

1.3 小结

对蛋白质相互作用关系(PPI)的认识和研究可以帮助我们更好的认识细胞生命活动的过程，揭示生物的生理和病理现象。利用 PPI 数据库提供的丰富数据源，可以从整体的角度探索生命体的活动规律，发觉其中蕴含的蛋白质的不同生物学功能。

尽管人们已经得到了大量丰富的 PPI 数据，建立了很多个物种的数据库，但是对蛋白质相互作用的研究一直是生物数据网络的研究热点，而在蛋白质功能的鉴定或预测、证实和应用等方面仍存在诸多挑战，主要包括如何利用现有的实验技术水平和预测方法来得到高质量的蛋白质相互作用数据，对已经产生的或将要生成的蛋白质相互作用数据的客观评价，进而利用这些已有的数据和其他生物学相关知识的关系来提高数据的可信度，以及如何从海量的 PPI 数据中发现更多具有生物学意义的信息。因此，选择高质量的 PPI 数据库作为研究对象是很重要的。目前，各种 PPI 数据库重合率都不高，往往通过经验去选择所要研究的 PPI 网络数据源。本设计选择了 HPRD，HIR 和 BioGRID 三个代表性的数据库作为研究对象。

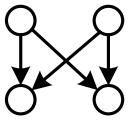
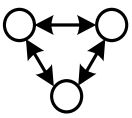
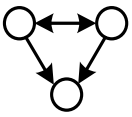
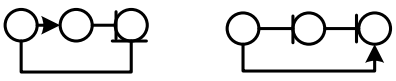
2 网络模体挖掘

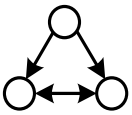
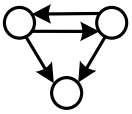
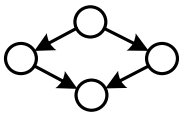
2.1 网络模体定义

网络模体是一些在真实网络中重复出现的具有特定拓扑结构特征的小规模模式单元，它在真实网络中出现的频率要远高于随机网络(Randomized Network)。它们是具有特定性质的小规模子图，也是构建整个生物网络的基本单元。这些小规模的子图多次出现于整个生物网络的不同部位，研究发现：它们在蛋白质相互作用网络、基因调控网络、万维网、人际关系网络等多种复杂网络中都可以找到这些多次出现的小规模子图结构。网络模体的发现对蛋白质相互作用关系组成的生物基础大网络的研究和探究具有非常重要的作用。如表 2.1 所示，罗列了部分网络模体示意图。

网络模体是蛋白质网络层次结构中重要的一环，是 PPI 网络的重要组成部分，如何高效而准确的从复杂的 PPI 网络中识别这些局部特征，对分析研究整个网络以及理解生物系统的组织结构有极其重要的意义，为此出现了一系列挖掘网络模体的聚类算法，用于大型网络中模体的识别。

表 2.1 7 种网络模体类型及其示意图

模体类型	模体示意图
Bi-fan	
Protein clique	
Interacting transcription factors that co-regulate a third gene	
Feed-forward loop (FFL)	 Coherent type 1 Coherent type 2  Coherent type 3 Coherent type 4

	 Incoherent type 1 Incoherent type 2  Incoherent type 3 Incoherent type 4
Co-regulated interacting Proteins	
Mixed-feedback loop between transcription factors that co-regulate a gene	
Biparallel	

2.2 网络模体挖掘算法

Milo 等人[6]提出网络模体的概念以来，人们研究出了许多针对于网络模体的算法，各种新型的模体挖掘算法也被陆续应用于大量复杂生物网络中，并且这些算法已经得到了很好的运用和发展，而这些新算法的出现无疑推动了蛋白质相互作用网络以及其他生物网络的研究进程。如表 2.2，罗列了部分具有代表性的模体挖掘算法，本设计使用 CyClus3D 进行网络模体挖掘。

表 2.2 10 个具有代表性的网络模体挖掘算法

算法	最大模体大小	软件使用平台	网址链接
CyClus3D	3	Cytoscape	http://bioinformatics.psb.ugent.be/software/details/CyClus3D
NetMatch	5	Cytoscape	http://baderlab.org/Software/NetMatch
mfinder	7	Windows2000, WindowsXP, Linux OS	https://www.encodeproject.org/software/mfinder/
MAVisto	8	All platforms	http://mavisto.ipk-gatersleben.de/

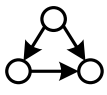
NeMoFinder	12	N/A	http://dl.acm.org/citation.cfm?id=1150418
FANMOD	8	Windows32-bit, Mac,Linux	http://dl.acm.org/citation.cfm?id=1182262
Grochow-Kellis	15	N/A	http://link.springer.com/chapter/10.1007/978-3-540-71681-5_7
Kavosh	>8	Windows32-bit, Linux	http://www.biomedcentral.com/1471-2105/10/318
MODA	>8	N/A	http://omictools.com/moda-s5727.html
NetMODE	6	Windows32-bit	http://www.netmode.ntua.gr/main/index.php

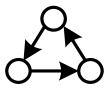
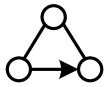
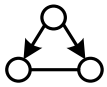
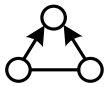
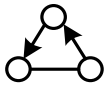
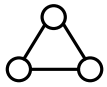
2.3 CyClus3D 软件

在系统生物学领域，细胞被建模为一个由多种相互作用形成的集群网络，例如蛋白质—蛋白质，蛋白质—DNA，蛋白质—代谢物，或遗传交互作用。在图形理论中，紧密连接节点趋向于集群，许多的算法被开发用于网络模体的识别，而不同的算法差异较大，因此开发了 CyClus3D 聚类算法用于模体的识别，CyClus3D 是 Cytoscape 软件平台的一个插件，利用 3D 谱聚类算法实现三节点网络模体挖掘。它适合于大型集群网络中网络模体的聚类分析。

该算法的每个网络模体定义了一个在异构数据类型之间的一种关系，与一个独特的信息处理角色和功能解释。因此，CyClus3D 的识别模块是由多重相互作用类型所组成的，如信号或补偿途径机制以及稳定蛋白质复合物的识别。CyClus3D 能挖掘的网络模体类型如表 2.3 所示，由于蛋白质相互作用关系为无向关系，因此采用模体 7 的 3 点模体类型实现网络模体挖掘。

表 2.3 CyClus3D 模体类型及其示意图

3 点模体类型	**模体示意图**
模体 1	

模体 2	
模体 3	
模体 4	
模体 5	
模体 6	
模体 7	

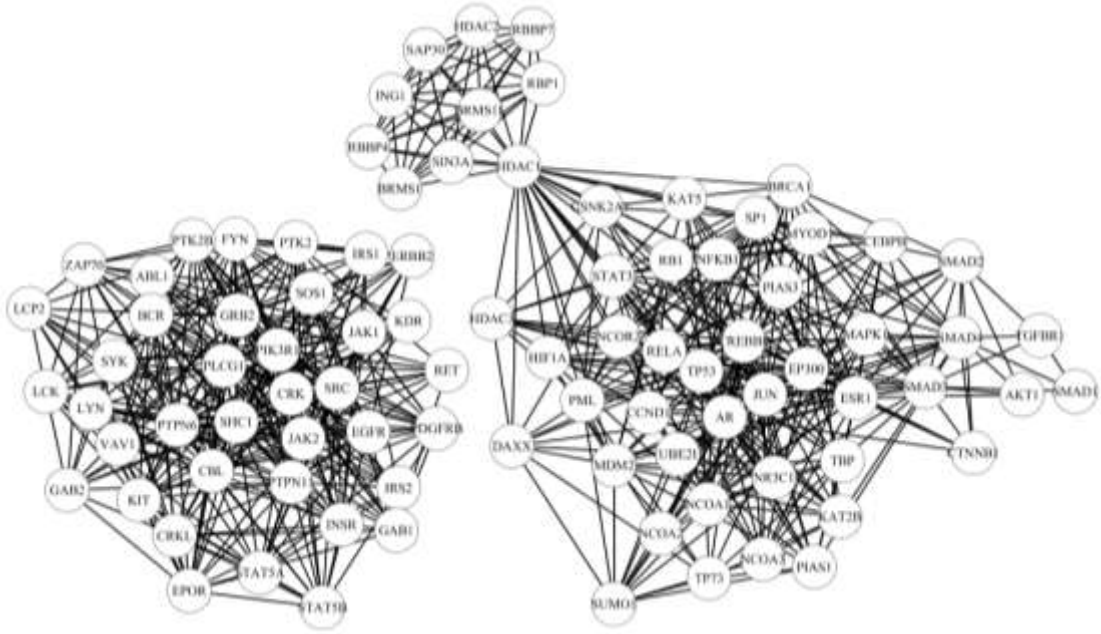
2.4 基于网络模体的数据挖掘

为了显示 CyClus3D 软件挖掘的网络模体，我们采用 Cytoscape 对网络模体进行重新组合，生成由不同网络模体组成的子网络。Cytoscape[19]软件是一个开源的用于网络显示和分析的生物信息学软件平台，是一个可视化的生物信息学分析软件，软件的核心部分提供了一些基本功能，包括网络的显示、布局、查询等。Cytoscape（http://www.cytoscape.org/）能适用于其他分子构件和相互作用的分析，但其最强大的功能还是用于大规模蛋白质—蛋白质相互作用、蛋白质—DNA 和遗传交互作用的分析。

本设计将分别采用 HPRD、HIR、BioGRID 三个 PPI 数据库作为模体挖掘的数据源，然后基于 CyClus3D 对这三种蛋白质相互作用数据进行网络模体挖掘。

2.4.1 基于 HPRD 数据源的网络模体挖掘

如图 2.1 所示，基于 HPRD 数据集所挖掘的网络模体子网络聚类效果比较好（聚类系数=0.619），蛋白质与蛋白质之间的路径长度比较短（特征路径长度=1.790），并且子网络很稀疏（网络密度=0.174）。

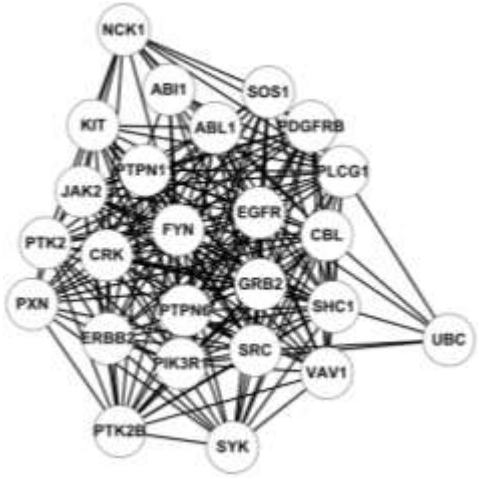


聚类系数（Clustering coefficient）	0.619	特征路径长度（Characteristic pathlength）	1.790
网络直径（Network diameter）	4	相邻数量平均值（Avg. number of neighbors）	15.467
网络半径（Network radius）	2	节点数（Number of nodes）	90
网络集中（Network centralization）	0.213	网络密度（Network density）	0.174
最短路径（Shortest paths）	4058(50%)	网络异质性（Network heterogeneity）	0.437

图 2.1 基于 HPRD 模体网络子图

2.4.2 基于 HIR 数据源的网络模体挖掘

如图2.2所示，基于HIR数据集所挖掘的网络模体子网络聚类效果比较好（聚类系数=0.744），蛋白质与蛋白质之间的路径长度比较短（特征路径长度=1.330），并且子网络比较密集（网络密度=0.667）。



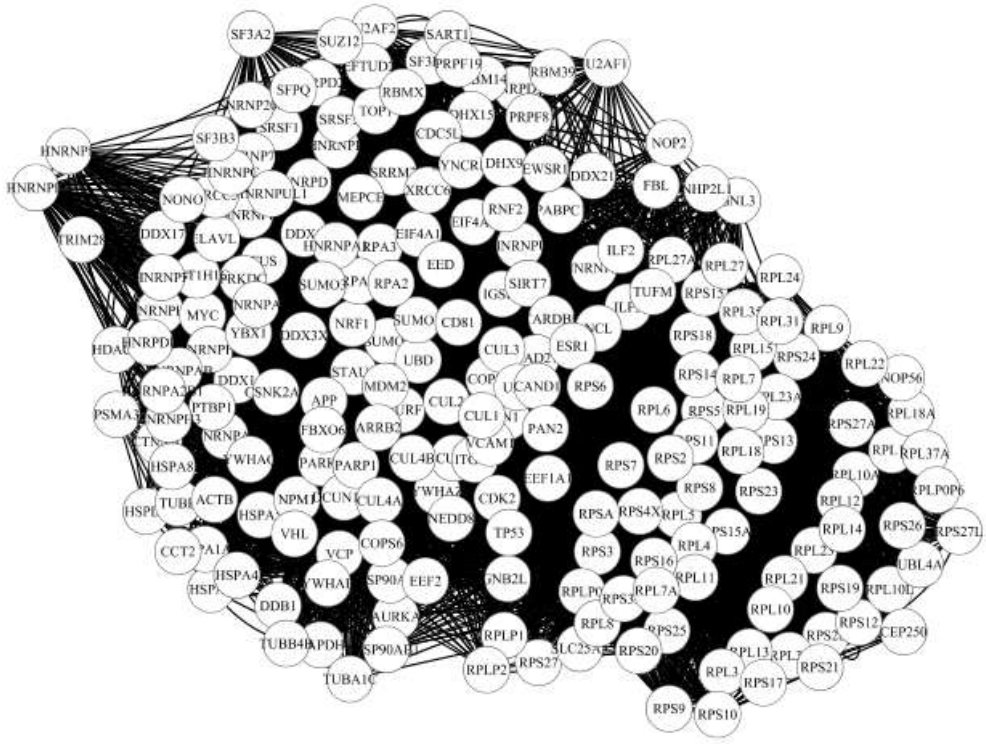
聚类系数（Clusteringcoefficient）	0.744	特征路径长度（Characteristic path length）	1.333
网络直径（Network diameter）	2	相邻数量平均值（Avg. number of neighbors）	16.0

网络半径（Network radius）	2	节点数（Number of nodes）	25
网络集中（Network centralization）	0.317	网络密度（Network density）	0.667
最短路径（Shortest paths）	600(100%)	网络异质性（Network heterogeneity）	0.257

图 2.2 基于 HIR 模体网络子图

2.4.3 基于 BioGRID 数据源的网络模体挖掘

如图 2.3 所示，基于 BioGRID 数据集所挖掘的网络模体子网络聚类效果比较好（聚类系数=0.526），蛋白质与蛋白质之间的路径长度比较短（特征路径长度=1.677），并且子网络比较稀疏（网络密度=0.323）。



聚类系数（Clusteringcoefficient）	0.526	特征路径长度（Characteristic path length）	1.677
网络直径（Network diameter）	2	相邻数量平均值（Avg. number of neighbors）	69.107
网络半径（Network radius）	2	节点数（Number of nodes）	215
网络集中（Network centralization）	0.679	网络密度（Network density）	0.323
最短路径（Shortest paths）	46010(100%)	网络异质性（Network heterogeneity）	0.448

图 2.3 基于 BioGRID 模体网络子图

2.5 小结

本节简要介绍了网络模体的概念、网络模体挖掘算法，以及基于 CyClus3D 分别对 HPRD、HIR 和 BioGIRD 三种 PPI 数据库进行网络模体挖掘。网络模体是生物系统中结构化的功能模块，在子网络中对应着不同的功能关系，使用聚类分析是识别这些模块的有效手段，而蛋白质网络的可视化使得聚类分析更加方

便。CyClus3D 算法适合于大型集群网络中网络模体的聚类分析，用于生物功能的识别。目前，网络模体识别已经取得了长足的发展并得到了广泛的应用，如在人类信号网络中，网络模体挖掘已经被用于一些癌症患者的识别，如乳腺癌。网络模体也为理解一些基因在基因调控中的功能角色提供了更好地解释，另外还可用于 PPI 网络中蛋白质—蛋白质相互作用的预测。

3 蛋白质相互作用的模体网络富集分析

3.1 GO 生物过程显著性分析

GO(Gene Ontology)即基因本体，是由基因本体联合会(Gene Ontology Consortium)所建立的一个数据库，其目的是建立一个适用于多物种的，对基因和蛋白质功能进行限定和描述的，并能随着研究不断深入而更新的语义词汇标准。GO 是多种生物本体语言中的一种，一共有 3 个结构化的网络，用于描述基因的产物以及基因注释。包含所有真核生物信息，如基因和蛋白质在细胞改变中所扮演的角色方面的知识。从生物过程、分子功能、细胞组成这三个方面对对基因进行分类、定义和注释。

通过 DAVID[20]在线对蛋白质相互作用的模体网络进行 GO 生物过程显著性分析。如表 3.1-3.3，只罗列了基于 3 个 PPI 数据库的前 10 个显著性 GO 生物过程，详细的 GO 生物过程显著性分析结果见附件 1。

表 3.1 基于 HPRD 的前 10 个显著性 GO 生物过程

术语	蛋白质数	显著性 *p* 值	Benjamini 校正 *p* 值
GO:0010604~positive regulation of macromolecule metabolic process	48	2.43E-33	4.29E-30
GO:0007167~enzyme linked receptor protein signaling pathway	34	1.94E-30	1.72E-27
GO:0031328~positive regulation of cellular biosynthetic process	42	5.12E-30	3.02E-27
GO:0009891~positive regulation of biosynthetic process	42	9.11E-30	4.02E-27
GO:0007242~intracellular signaling cascade	51	3.28E-29	1.16E-26
GO:0010557~positive regulation of macromolecule biosynthetic process	40	2.46E-28	7.25E-26
GO:0051173~positive regulation of nitrogen compound metabolic process	38	3.58E-26	9.04E-24
GO:0010628~positive regulation of gene expression	35	3.74E-24	8.26E-22
GO:0010033~response to organic substance	37	2.56E-23	5.02E-21

GO:0045935~positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	35	3.79E-23	6.71E-21

表 3.2 基于 HIR 的前 10 个显著性 GO 生物过程

术语	**蛋白质数**	**显著性 *p* 值**	**Benjamini 校正 *p* 值**
GO:0007169~transmembrane receptor protein tyrosine kinase signaling pathway	17	1.21E-23	1.04E-20
GO:0007167~enzyme linked receptor protein signaling pathway	18	1.39E-22	5.98E-20
GO:0018108~peptidyl-tyrosine phosphorylation	10	3.31E-17	9.54E-15
GO:0018212~peptidyl-tyrosine modification	10	5.03E-17	1.09E-14
GO:0007166~cell surface receptor linked signal transduction	21	3.12E-14	5.38E-12
GO:0006793~phosphorus metabolic process	16	4.54E-12	6.53E-10
GO:0006796~phosphate metabolic process	16	4.54E-12	6.53E-10
GO:0016310~phosphorylation	15	6.45E-12	7.95E-10
GO:0007242~intracellular signaling cascade	17	9.98E-12	1.08E-09
GO:0007243~protein kinase cascade	12	1.00E-11	9.62E-10

表 3.3 基于 BioGRID 的前 10 个显著性 GO 生物过程

术语	**蛋白质数**	**显著性 *p* 值**	**Benjamini 校正 *p* 值**
GO:0006414~translational elongation	68	2.09E-101	2.47E-98
GO:0006412~translation	72	7.25E-64	4.29E-61
GO:0006396~RNA processing	69	1.99E-44	7.84E-42
GO:0008380~RNA splicing	49	3.23E-37	9.56E-35
GO:0016071~mRNA metabolic process	53	3.36E-36	7.96E-34
GO:0006397~mRNA processing	50	8.24E-36	1.63E-33
GO:0000375~RNA splicing, via	35	1.48E-30	2.50E-28

transesterification reactions			
GO:0000377~RNA splicing, via transesterification reactions with bulged adenosine as nucleophile	35	1.48E-30	2.50E-28
GO:0000398~nuclear mRNA splicing, via spliceosome	35	1.48E-30	2.50E-28
GO:0022613~ribonucleoprotein complex biogenesis	28	1.18E-19	1.74E-17

3.2 KEGG 信号通道显著性分析

KEGG(Kyoto Encyclopedia of Genes and Genomes)，翻译为京都基因与基因组百科全书，是一个关于基因组信息、基因功能分析、细胞过程、酶促反应以及生物化学物质的在线数据库(http://www.genome.ad.jp/kegg/)，主要包括 GENES database、PATHWAY database 和 LIGAND database。

各个数据库中包含了大量有用信息。KEGG 的PATHWAY 数据库整合当前在分子互动网络（比如通道，联合体）的知识和高级功能信息，包括了生物细胞生化过程，如代谢、信号传递、物质的跨膜转运、细胞周期等过程的图解分析，还包括同系保守的子通路等信息；KEGG的GENES数据库提供关于基因组信息的知识和蛋白质的相关知识；KEGG的另一个数据库LIGAND，包含了酶分子、酶反应以及化学物质等方面的知识。

通过 DAVID 在线对蛋白质相互作用的模体网络进行 KEGG 信号通道显著性分析。如表 3.4-3.6，只罗列了 Benjamini 校正 *p* 值小于 0.05 的前 10 或 4 个显著性 KEGG 信号通道，详细的 KEGG 信号通道显著性分析结果见附件 2。

表 3.4 Benjamini 校正 *p* 值小于 0.05 的前 10 个显著性 KEGG 信号通道，PPI 数据集为 HPRD

术语	**蛋白质数**	**显著性 *p* 值**	**Benjamini 校正 *p* 值**
hsa05200:Pathways in cancer	42	1.04E-30	7.83E-29
hsa05220:Chronic myeloid leukemia	26	2.74E-29	1.03E-27
hsa04012:ErbB signaling pathway	19	6.33E-17	2.78E-15
hsa05215:Prostate cancer	18	2.26E-15	4.16E-14
hsa05212:Pancreatic cancer	16	3.09E-14	4.63E-13
hsa04722:Neurotrophin signaling	19	4.78E-14	5.97E-13

pathway			
hsa04520:Adherens junction	15	1.79E-12	1.92E-11
hsa05210:Colorectal cancer	15	6.25E-12	5.86E-11
hsa04660:T cell receptor signaling pathway	16	1.58E-11	1.31E-10
hsa05221:Acute myeloid leukemia	13	1.59E-11	1.19E-10

表 3.5 Benjamini 校正 *p* 值小于 0.05 的前 10 个显著性 KEGG 信号通道，PPI 数据集为 HIR

术语	**蛋白质数**	**显著性 *p* 值**	**Benjamini 校正 *p* 值**
hsa04012:ErbB signaling pathway	13	3.30E-16	1.87E-14
hsa04510:Focal adhesion	13	9.70E-12	2.72E-10
hsa04650:Natural killer cell mediated cytotoxicity	11	9.15E-11	1.71E-09
hsa04660:T cell receptor signaling pathway	9	1.20E-08	1.68E-07
hsa05220:Chronic myeloid leukemia	8	2.32E-08	2.60E-07
hsa05200:Pathways in cancer	12	4.58E-08	4.27E-07
hsa04062:Chemokine signaling pathway	10	5.32E-08	4.25E-07
hsa05214:Glioma	7	2.43E-07	1.70E-06
hsa04722:Neurotrophin signaling pathway	8	7.67E-07	4.77E-06
hsa04664:Fc epsilon RI signaling pathway	7	8.80E-07	4.93E-06

表 3.6 Benjamini 校正 *p* 值小于 0.05 的前 4 个显著性 KEGG 信号通道，PPI 数据集为 BioGRID

术语	**蛋白质数**	**显著性 *p* 值**	**Benjamini 校正 *p* 值**
hsa03010:Ribosome	65	1.79E-86	1.33E-84
hsa03040:Spliceosome	28	2.38E-17	8.80E-16
hsa04120:Ubiquitin mediated proteolysis	12	0.001506	0.036482749
hsa04110:Cell cycle	11	0.00253	0.045791482

3.3 小结

本节简单介绍了 GO 和 KEGG 数据库，分别对 HPRD、HIR 和 BioGIRD 三种 PPI 数据库进行 GO 生物过程和 KEGG 信号通道显著性分析。得出它们的显著性 *p* 值，*p* 值（校正方法为 Benjamini 法）表示一个蛋白质集合被随机赋予一个 GO 生物过程或 KEGG 信号通道的概率。*p* 值越小，说明富集性越好，就越能证明这个 GO 生物过程或 KEGG 信号通道不是随机出现的，而是为了完成某一特定的生物功能或 KEGG 信号通道而出现的，有显著的统计学意义。

基于HPRD、HIR和BioGIRD三个PPI数据库的显著性GO生物过程和KEGG信号通道基本不一样，但是都能够挖掘一些具有代表性的 GO 生物过程和 KEGG 信号通道，甚至可以挖掘出相同的生物过程。而同一个生物过程或 KEGG 信号通道在不同的 PPI 数据库中经过 DAVID 挖掘出来的蛋白质数却不同，例如：基于 HPRD 和 HIR 两个 PPI 数据库都能够挖掘出酶联受体蛋白信号通路（GO:0007167~enzyme linked receptor protein signaling pathway），在数据库 HPRD 中参与过程调节的蛋白质有 GRB2、STAT5A、ERBB2、STAT5B、PML、KIT、SRC 等 34 个(见附件 1)，而在数据库 HIR 中，参与过程调节的蛋白质有 EGFR、GRB2、ERBB2、CBL、ABI1、KIT、PXN 等 18 个(见附件 1)。又例如，基于 HPRD 和 HIR 两个 PPI 数据库都能够挖掘出 KEGG 癌症信号通道，在数据库 HPRD 中参与 KEGG 癌症信号通道蛋白质有 GRB2, STAT5A, ERBB2, STAT5B, PML, NFKB1, KIT 等 74 个(见附件 2)，而在数据库 HIR 中，参与过程调节的蛋白质有 EGFR, PTK2, PLCG1, GRB2, ERBB2, SOS1, CBL 等 24 个(见附件 2)。所以，为了挖掘更多具有显著性的 GO 生物过程和 KEGG 信号通道，选择更加全面的 PPI 数据库是一个急需解决的问题。

4 结束语

生物完成各种生命活动、实现生物功能需要蛋白质的参与，这些生物功能的实现都涉及到蛋白质相互作用，并且蛋白质相互作用是其他生物学网络的基础，所以对 PPI 的鉴定以及分析将会越来越重要。

如何得到高质量的数据并充分利用这些数据是分析 PPI 网络的重要方向，而海量的 PPI 数据也正在不断的完善。本文选取了 HPRD、HIR、BioGRID 三个 PPI 数据库进行数据收集，在 Cytoscape 软件平台中，基于三点网络模体 CyClus3D 模体挖掘算法对 HPRD、HIR、BioGRID 蛋白质相互作用数据库进行数据挖掘，进一步用 DAVID 在线工具对网络模体组成的子网络进行 GO 生物过程和 KEGG 信号通道显著性分析，发现了一些具有代表性的 GO 生物过程和 KEGG 信号通道。通过显著性 p 值不仅可以看出某些生物过程或通路中蛋白质的富集程度，并且在 DAVID 中还可以查看参与控制调节这些生物过程的基因。PPI 网络属于大型的复杂网络，从大量的 PPI 数据中挖掘出更多具有生物学意义的信息，有利于更好的了解生物体的生命活动中蛋白质之间的相互作用关系，从整体水平上理解某些疾病的细胞机制，发现不同的生物途径，进而预测一些蛋白质可能的生物功能。

参考文献

[1] Yamada T, Bork P. Evolution of biomolecular networks—lessons from metabolic and protein interactions[J]. Nature Reviews Molecular Cell Biology, 2009, 10(11): 791-803.

[2] Yeger-Lotem E, Sattath S, Kashtan N, et al. Network motifs in integrated cellular networks of transcription–regulation and protein–protein interaction[J]. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101(16): 5934-5939.

[3] Lodish H, Berk A, Matsudaira P, et al. Molecular Cell Biology (5th ed.)[M]. New York, New York: WH Freeman and Company, 2004.

[4] Waugh DF. Protein-protein interactions[J]. Advances in protein chemistry, 1954, 9:325-437.

[5] Shen-Orr SS, Milo R, Mangan S, et al. Network motifs in the transcriptional regulation network of Escherichia coli[J]. Nature genetics, 2002, 31(1): 64-68.

[6] Milo R, Shen-Orr SS, Itzkovitz S, et al. Network motifs: simple building blocks of complex networks[J].Science, 2002, 298(5594):824-827.

[7] Audenaert P, Van Parys T, Brondel F, et al. CyClus3D: a Cytoscape plugin for clustering network motifs in integrated networks[J]. Bioinformatics, 2011, 27(11): 1587-1588.

[8] Ferro A, Giugno R, Pigola G, et al. NetMatch: a Cytoscape plugin for searching biological networks[J]. Bioinformatics, 2007, 23(7): 910-912.

[9] Kashtan N, Itzkovitz S, Milo R, et al. Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs[J]. Bioinformatics, 2004, 20(11): 1746-1758.

[10] Schreiber F, Schwöbbermeyer H. MAVisto: a tool for the exploration of network motifs[J]. Bioinformatics, 2005, 21(17): 3572-3574.

[11] Chen J, Hsu W, Lee ML, et al. NeMoFinder: Dissecting genome-wide protein-protein interactions with meso-scale network motifs[C]//Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2006: 106-115.

[12] Ashburner M, Ball CA, Blake JA, et al. Gene Ontology: tool for the unification of biology[J]. Nature genetics, 2000, 25(1): 25-29.

[13] Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes[J].

Nucleic acids research, 2000, 28(1): 27-30.

[14] Prasad T S K, Goel R, Kandasamy K, et al. Human protein reference database—2009 update[J]. Nucleic acids research, 2009, 37(suppl 1): D767-D772.

[15] Zhou X, Chen P, Wei Q, et al. Human interactome resource and gene set linkage analysis for the functional interpretation of biologically meaningful gene sets[J]. Bioinformatics, 2013, 29(16): 2024-2031.

[16] Stark C, Breitkreutz B J, Reguly T, et al. BioGRID: a general repository for interaction datasets[J]. Nucleic acids research, 2006, 34(suppl 1): D535-D539.

[17] Del Sol A, O'Meara P. Small‐world network approach to identify key residues in protein‐protein interaction[J]. Proteins: Structure, Function, and Bioinformatics, 2005, 58(3): 672-682.

[18] Del Sol A, Fujihashi H, O'Meara P. Topology of small-world networks of protein–protein complex structures[J]. Bioinformatics, 2005, 21(8): 1311-1315.

[19] Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models ofbiomolecular interaction networks[J]. Genome research, 2003, 13(11): 2498-2504.

[20] Huang D W, Sherman B T, Lempicki R A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources[J]. Nature Protocols, 2009, 4(1):44-57.

致谢

时光匆匆，转眼便是大学毕业时节。值此论文完成之际，谨向我们的导师张俊鹏老师表示最诚挚的敬意和最衷心的感谢!本文就是沿着张老师指引的科研方向、在张老师的悉心指导下完成的。在论文写作的过程中，遇到了很多的问题，但是在张老师的耐心指导下，问题得以很好的解决。正是由于他在百忙之中多次审阅全文，对细节进行修改，并为本文的撰写提供了许多中肯而且宝贵的意见，本文才得以顺利完成。张老师学识渊博，视野广阔。在他身上我们学到的不仅是前沿的专业知识和科学研究的方法，更重要的是刻苦钻研的精神和严谨的科研作风，这些都将使我们受益终生。

再次感谢所有关心、支持和帮助过我们的老师和同学们，从开始进入课题到论文的顺利完成，一直都离不开老师、同学、朋友给我们热情的帮助，在这里请接受我们诚挚的谢意。是你们帮助我们成长进步，包容我们的缺点和不足，给我们勇气和力量，非常感谢！感谢我们的家人和好友，他们的关爱和支持永远是我们前进的最大动力，在任何时候，他们都给予我们最大的鼓励和支持，我们感谢他们!

最后，向审阅我们论文和参加答辩的老师们表示衷心的感谢，感谢你们抽出宝贵的时候参加我们的论文答辩会，感谢你们对论文不当之处提出的宝贵意见。