蛋白质关系网络可视化系统的研究与实现

王 健 谢 冬 杨志豪 林鸿飞
(大连理工大学计算机科学与技术学院 大连 116024)

摘 要 蛋白质关系网络的研究在生物医学领域中已成为一个热点。研究者通过对蛋白质关系网络进行分析和聚类,能够发现其中的复合体,进一步理解细胞组织原理。在对关系网络进行分析的过程中,将网络拓扑显示为图形,以直观地表示出关系网络的结构,便于对比聚类方法,辅助关系网络的研究。利用网络建模与可视化工具包JUNG设计并实现了一个蛋白质关系网络可视化系统,它能够解析多种格式的蛋白质关系网络数据,集成了几种有效的图聚类算法,并实现了一种基于蛋白质功能标注的发现复合体的聚类算法。用户能够通过二维网络视图方便地观察原始网络和聚类后的结果。

关键词 蛋白质关系网络,网络可视化,图聚类,JUNG工具包
中图法分类号 TP391 **文献标识码** A

Research and Implementation on Visualization System for PPI Network

WANG Jian XIE Dong YANG Zhi-hao LIN Hong-fei
(School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China)

Abstract Research on protein-protein interaction (PPI) network has become a hot spot in the biomedical domain. With analysis and clustering application on PPI network, researchers detect complexes for a better understanding of cellular organizations. Visualizing PPI network in analyzing process can intuitively present the organization of the network, which makes easy comparison of clustering methods and promotes researches on PPI network. We designed and implemented a PPI network visualization system with JUNG, a network modeling and visualization library. We also integrated several effective graph clustering algorithms into the system, and implemented a protein complex detection method based on protein function annotation. The original network and clusters can be navigated conveniently in two-dimensional network views.

Keywords PPI network, Network visualization, Graph clustering, JUNG toolkit

1 引言

随着对蛋白质组学研究的深入,蛋白质关系网络越来越多地受到研究者的关注。由于蛋白质通常并不是单独起作用,而是通过相互作用关系组成一定的功能结构来发挥作用,因此在这些相互作用关系构成的网络中发现这些功能团与复合体,能够帮助生物医学专家预测蛋白质功能,了解生物功能的分子机制,从而进一步理解生物学行为和设计药物。蛋白质关系网络属于无尺度网络,由少量度数较高的结点和大量度数较低的结点构成,因此在进行蛋白质关系网络的分析和聚类时,研究者能够以图形的形式直观地观察到网络结构,以便针对其拓扑结构特点进行分析,这有助于蛋白质关系网络的研究。

目前已有数十种关系网络可视化工具,如Pajek[1]适用于将通用领域的关系网络的数据进行可视化,它具有多种二维网络图形的布局算法用以动态地显示网络图形,内置了多种聚类算法,使用其特有的文件格式存取网络数据。而生物医学领域也有专用的关系网络可视化工具,如Cytoscape[2]适用于对大规模生物实体关系网络进行可视化和分析,它支持有向和无向加权图的数据,内置了强大的布局显示方式,适用于显示分子作用关系和基因表达数据,并具有插件架构,便于用户对其功能进行扩展。Ondex[3]能够对多种类别的生物学实体构成的网络进行整合,在数据整合和过滤方面的功能较为完善。Pivot[4]则用二维无向图来显示蛋白质关系网络,内置了4个物种的蛋白质的功能注释,能够标识同源物并连接在线数据库。这些软件都提供了较好的人机交互和用户界面,并内置了多种布局算法和一些常用的网络分析功能,对生物学关系网络的研究者提供了较好的帮助。在国内也有针对关系网络可视化的研究[5],对社会关系网络、电力网络等复杂网络的可视化技术进行了探讨[6],但目前的可视化系统还较少。

到稿日期:2011-01-23 返修日期:2011-04-21 本文受国家自然科学基金(60673039,60973068,61070098),国家社科基金(08BTQ025),国家863高科技计划(2006AA01Z151),教育部留学回国人员科研启动基金和高等学校博士学科点专项科研基金(20090041110002)资助。

王 健(1967-),女,副教授,CCF高级会员,主要研究方向为生物医学文本挖掘,E-mail:wangjian@dlut.edu.cn;谢 冬(1986-),男,硕士生,主要研究方向为蛋白质关系网络聚类与可视化;杨志豪(1973-),男,博士,副教授,主要研究方向为生物医学文本挖掘;林鸿飞(1962-),男,博士,教授,博士生导师,主要研究方向为搜索引擎、文本挖掘、情感计算和自然语言理解。

随着近年来蛋白质关系数据量的增长，利用计算方法进行蛋白质关系网络分析和聚类成为计算机科学的研究者将基于图的聚类算法应用于实际领域的合适的研究方向。新的蛋白质关系网络的聚类方法不断提出，而现有软件对数据格式和利用计算方法进行实验的功能支持不够完善，缺乏针对蛋白质关系网络聚类以及复合体发现的可视化软件，编写现有工具的插件需要做大量编程工作。因此需要一个更易扩展的关系网络可视化工具，让研究者能够方便地向其中添加特定的功能和分析特定的数据，对比不同的聚类方法。

针对以上需求，本文设计实现了一个蛋白质关系网络的可视化系统，并集成了几种常用的蛋白质关系网络聚类方法，针对网络拓扑结构选择了适当的网络图形布局算法。另外利用 JUNG 网络建模与可视化工具包[7] 提供了灵活的网络数据模型和编程接口，方便用户添加自己的算法。

2 系统框架与实现

本系统的主要功能是对蛋白质关系网络进行可视化，并集成蛋白质关系网络的聚类算法，系统主要分为关系网络模型和网络浏览器两部分，其框架设计如图 1 所示。

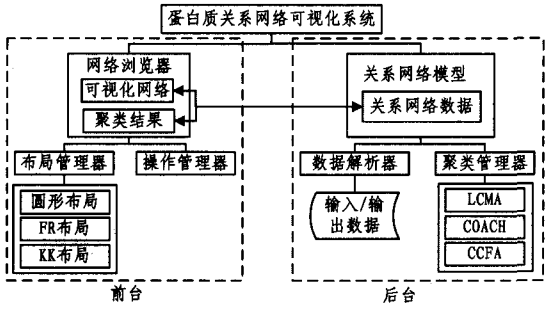

图 1 系统框架

当系统载入关系网络数据后，首先由数据解析器将输入数据解析并用系统中的数据结构为关系网络建立模型，然后生成可视化的图形并显示在网络浏览器中。用户可以通过操作管理器对网络进行全面的观察，并用聚类管理器调用一种或几种聚类算法对关系网络或其子图进行聚类分析，对比不同的复合体数据集来评价不同的方法。

2.1 蛋白质关系网络模型

蛋白质关系网络模型是系统的后台部分，主要提供了系统的输入输出、数据存储和数据处理功能。

2.1.1 输入与输出

数据解析器负责处理输入输出数据。由于系统的输入数据有蛋白质关系网络数据、标准复合体数据、蛋白质功能标注以及结点与边的属性数据等多种格式，因此由数据解析器将这些数据解析并构建为系统内部的数据模型，供网络浏览器进行网络图形的显示、聚类管理器进行分析，并将网络浏览器中显示的图形与聚类分析的结果导出为多种格式的输出文件。数据解析器能够解析 DIP[8]、BioGRID[9]、MIPS[10] 等常用蛋白质关系数据库中使用的数据格式，如 PSI-MI、XIN 等，另外还能解析带有可视化信息的数据格式，如 GraphML 格式的数据。系统的输出有可视化展示的二维网络图形，数据解析器能够将其保存为图像或大多数可视化软件都支持的 GraphML 格式文本，另外还有用户修改过的关系网络数据，以及通过聚类分析得到的聚类结果，能够与其他的软件使用的数据格式兼容。

另外，数据解析器还集成了一些实用功能，如根据不同的属性对网络中的结点和边进行过滤，构成子网，以及解析 MIPS 蛋白质功能标注[10]、GO[11] 基因本体标注等常用的蛋白质功能信息数据库的数据，有助于研究者方便地获取蛋白质的详细资料分析聚类结果。系统中内置了 GO 基因本体数据库中的蛋白质功能标注数据，使用户不需要连接网络就能够快速地了解蛋白质功能信息。

2.1.2 聚类分析

聚类管理器是系统中主要的数据分析和处理部分，负责加载不同的聚类分析方法，对导入系统中的蛋白质关系网络进行聚类，并将得到的聚类结果用列表的形式展示给用户，用户可以通过从列表中选择并创建子图的方式来查看某个特定的聚类结果。另外它还能依据标准复合体数据和蛋白质功能标注数据对不同的方法得到的结果进行对比评价。

聚类管理器实现了系统中主要的数据处理功能，其中集成了 MCL[12]、RNSC[13]、LCMA[14]、COACH[15]、CMC[16] 等几种重要的发现蛋白质复合体的图聚类算法，另外实现了一种基于蛋白质功能标注的蛋白质复合体发现方法 CCFA[17]，用以探测关系网络中的复合体，使蛋白质关系网络的研究者能够利用多种方法发掘网络中的复合体，并对不同方法的性能进行比较。系统还提供了聚类结果与标准答案的对比功能，用以计算准确率、召回率、F 值以及敏感度、阳性预测值、精度等聚类评价指标并显示不同方法的对比结果。

2.2 网络浏览器

系统的前台部分是一个网络浏览器，为用户提供对网络视图的浏览和操作。浏览器将关系网络显示为二维无向图形，若关系网络数据中的蛋白质或蛋白质关系具有特定的属性，如权重、相似度等，用户可以选择将这些属性通过图形中结点或边的标签和外观表示出来。

布局管理器提供了多种网络布局算法，如圆形布局、力导引布局等，为用户提供了不同的网络图形外观，使用户能够根据不同的网络拓扑结构选择不同的结点与边的布局方式，更好地观察网络的结构。

操作管理器管理对网络视图操作的响应，为用户提供了鼠标和快捷菜单等方式来对关系网络图形进行浏览和操作，除了基本的拖拽、旋转、缩放等功能之外，用户可以方便地增加、删除网络中的结点和边，并选择部分结点和边查看其属性、改变外观或创建子图。

2.3 系统实现

本系统利用 Java 语言实现，在关系网络数据建模与可视化方面主要利用了 JUNG 工具包。JUNG 是一个 Java 语言实现的开源工具包，它提供了一组通用性强的编程接口，用于对图或网络的数据结构进行建模、分析和可视化。在数据模型方面，JUNG 给出了多种类型的图的数据结构，如有向图、无向图、K 部图等，创建和使用图的模型时只需定义图中结点和边的数据类型即可利用这些图的模型。在可视化方面，JUNG 提供了将网络模型中的数据显示为结点和边构成的二维图形以及修改结点和边的颜色、形状等外观的方法，另外还提供了多种网络图形的布局方法。

在系统数据解析器将输入的关系网络数据解析后转化为 JUNG 中定义的无向图模型，然后调用网络浏览器创建可视

化对象来显示网络。系统前台和后台的数据中网络模型是一致的，对网络视图中的数据进行修改会同时改变后台的网络模型。用户可以自己定义结点和边的属性，并利用 JUNG 提供的网络模型更方便地实现自己的聚类方法。

3 关键技术

3.1 数据模型设计

蛋白质关系网络数据不仅包含网络本身的拓扑信息，有的还包含网络中的结点和边的属性，这些属性描述了蛋白质和蛋白质关系的某种性质，如功能相似性、亲和度等，很多聚类分析算法会利用这些信息。系统需要导入关系网络、网络中蛋白质与蛋白质关系的属性，另外还要根据用户的需求将这些属性通过网络图形中的结点和边的颜色、形状等显示属性表现出来，因此要设计合理的数据模型来整合和管理这些信息。由于后台部分主要处理结点和边的属性，而前台部分主要涉及显示属性，因此为了使数据结构简洁，节省存储空间，保证数据处理和显示的效率，本系统在后台和前台分别设计了两种不同的数据模型来储存结点和边的信息。

后台中的结点定义为一个包含蛋白质名和蛋白质属性向量的类。由于不同的蛋白质关系网络数据源中包含的蛋白质属性信息的种类和数量不同，为了节省每个蛋白质结点的存储空间，将属性信息存放在可变长度的向量中。类似地，后台中的边也由其序号和属性向量组成。结点中的蛋白质名和边的序号是唯一区分不同结点和边的依据。系统中的关系网络就是以这两个类的实例构成的无向图模型。

前台中的结点和边的显示属性是通过一组映射器生成的，这组映射器决定了结点和边的哪些属性显示为什么样的显示属性。而存储显示属性的目的是将网络图形某一时刻的显示状态保存下来以便用户查看，因此前台只记录下映射器和每个结点的显示位置即可。另外前台部分还单独维护一个列表，将那些不是通过映射器生成，而是用户对个别结点和边手动修改过的显示属性保存下来。

3.2 功能扩展

由于来自不同数据源的蛋白质关系网络数据格式多种多样，新的聚类方法也在不断被提出，因此为了能够兼容地集成现有的和今后提出的聚类算法，本系统为用户提供了易扩展的接口，使关系网络的研究者可以方便地添加自己编写的数据解析和聚类分析程序。

本系统为用户提供了两种方式来扩展数据解析和处理功能。首先，系统给出了通用的数据解析和聚类分析的 Java 抽象类，使用 Java 语言的编程者可以在简单了解系统中的关系网络数据模型的基础上，编写数据解析或聚类分析的程序来继承对应的抽象类，实现其中的抽象方法，这种方式能够与系统结合得更紧密，使算法在系统中的执行效率更高。另一种方式是直接向系统中添加数据解析或聚类算法的可执行文件，给出可执行文件的路径、执行参数和输出结果的路径，让系统导入并执行。这种方式更加灵活，用户可以添加任何可执行程序，输出文件只要满足系统能够解析的文件格式即可。系统中集成的聚类算法都是通过这两种方式实现的。

3.3 CCFA 聚类算法

本系统实现并集成了一种基于蛋白质功能标注的复合体发现方法 CCFA。

蛋白质复合体通常存在于关系网络的稠密区域中，即复合体内的蛋白质结点相互间的边较多，而与复合体外的蛋白质间的边较少。目前的复合体发现方法主要都是通过挖掘关系网络中密度较大的区域实现的。由于同一个复合体中的大部分蛋白质都具有相同或相近的生物学功能，因此可以将蛋白质功能标注信息引入复合体发现中。Gavin 等人[16]提出了蛋白质复合体的 core-attachment 结构特征，即复合体的结构主要由核心和附属物蛋白质构成，复合体的核心蛋白质相互作用关系较多，构成的子图密度较大，附属物蛋白质则表现为与核心蛋白质联系较密切而与其它蛋白质联系不密切。CCFA 算法主要根据这种结构特点对关系网络中的蛋白质复合体进行挖掘，主要分为 3 步：

(1) 挖掘出关系网络中所有的极大完全子图，根据 MIPS 蛋白质功能标注信息过滤掉其中具有相同功能标注的蛋白质较少(占子图中蛋白质总数的比例小于 0.7)的子图；

(2) 对完全子图进行扩展，向这些完全子图添加具有相同功能的邻结点，若该结点加入后图的密度大于一个阈值(0.6)，则加入这个结点，构成复合体核心部分；

(3) 向核心部分添加附属物，将核心部分的邻结点中与核心部分半数以上的蛋白质之间有关系的作为附属物蛋白质加入复合体。

经过以上步骤得到的复合体预测结果比没有引入蛋白质功能信息的方法得到的结果有良好的改善。

4 系统运行

4.1 系统环境

系统基于 JDK1.5 编写，利用了 JUNG2.0 工具包，以及 JUNG 工具包所必需的 Commons Collections4.01 和 Colt1.2.0，它们分别用于对 Java 容器类的扩展和提供高效的数学运算，以便支持 JUNG 工具包中内置的网络分析算法。

本系统由于使用 Java 语言开发，因此具有跨平台特性，可以在多种操作系统上运行，运行环境需要 JRE1.5 以上的版本。

4.2 运行实例

本系统的用户界面主要由控制面板和网络浏览器两部分组成，如图 2 所示，左侧为网络树状图、聚类结果以及过滤面板，中间为用不同布局算法显示的 Gavin 蛋白质关系网络数据集[18]及其两个子图，下方的详细信息栏用于显示用户所选中的蛋白质的功能标注信息。用户可以利用本系统对不同数据格式的关系网络进行可视化，通过不同的布局算法和过滤器对关系网络的拓扑结构和蛋白质的属性进行观察，利用系统中集成的聚类算法对网络进行聚类分析，导入标准复合体数据对聚类算法的结果进行评价和对比。

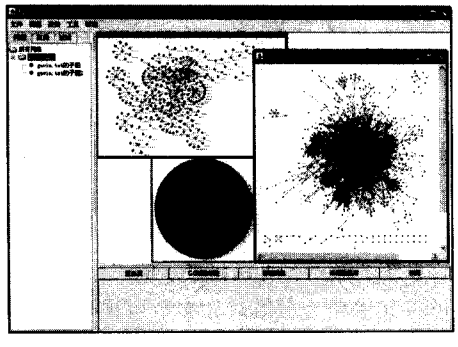

图 2 系统概貌

若要对 Gavin 数据集进行聚类分析，发现其中的复合体，首先应将关系网络数据集载入到系统中。左侧面板中的树状图对应了右侧浏览器中打开的每一个网络的视图。浏览器提供了一个多窗口的桌面展示每个已载入的关系网络。对选定的网络进行聚类分析，左侧的聚类标签中就会显示出得到的所有聚类结果，列表中每一项对应一个聚类结果，并列出了其中包含的蛋白质结点，选中其中的一项，在右边的视图中就会选中对应的结点，如图 3 所示。

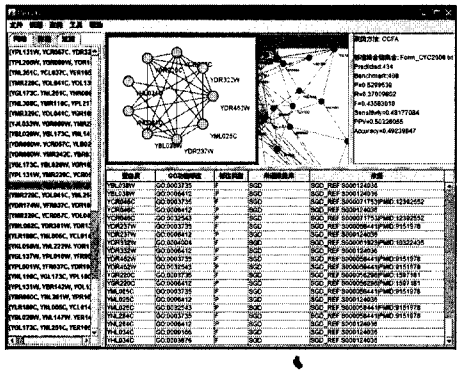

图 3　CCFA 聚类算法的结果

用户可以创建子图并进一步查看这些结点的信息。执行完聚类算法之后就可以导入标准复合体数据对聚类算法的结果进行评价。右侧的信息栏中给出了 CCFA 算法在 Gavin 数据集上对比 CYC2008 复合体数据集得到的评价结果，由系统计算出预测复合体个数、实际复合体个数、准确率、召回率、F 值等评价指标。

图 3 中间的子图是利用 CCFA 算法预测出的复合体集合中的一个，这个聚类结果没有匹配到实际的复合体，但我们通过选中这些蛋白质结点来查看系统内置的 GO 数据库中的蛋白质功能标注信息，发现这 9 个蛋白质中有 8 个的分子功能标注为 GO:0003735，生物学过程标注为 GO:0006412，因此它们组成的可能是一个尚未被发现的复合体，同时也预示着 YHL034C 可能具有与其他几个蛋白质相似的生物学功能。这样通过详细信息栏可以使研究者方便地发现这个潜在的复合体，进一步锁定研究对象进行下一步实验，而不需要搜集大量文献进行对比。

本系统在 CPU 主频 2GHz、内存空间 1GB、操作系统为 Windows XP SP3 的计算机上进行了性能测试。当结点数少于 1000、边数少于 3000 时，关系网络的载入和鼠标操作的延迟小于 1s，而在结点数和边数多于以上值时对网络视图进行拖拽、旋转以及切换网络布局时会有一定的延迟，这主要是受硬件配置的影响。另外网络模型中结点和边的属性数据的差异以及输入数据格式的不同也会影响载入的速度。

结束语　本文利用 JUNG 工具包设计实现了一个蛋白质关系网络可视化系统，它能够将不同格式的关系网络数据导入并用二维网络图形进行展示，提供了不同的布局方式显示网络的拓扑结构、对比聚类实验结果等实用功能，还提供了易扩展的编程接口，内置了 GO 蛋白质功能标注数据，集成了几种常用的蛋白质关系网络聚类算法；另外实现了一种引入蛋白质功能标注的复合体发现算法 CCFA，用以有效地发现关系网络中的蛋白质复合体。与现有系统相比，本系统的主要特点是简便易用和易扩展，拥有很多蛋白质关系网络聚类方面的实用功能，适用于蛋白质关系网络和复合体发现的研究。经过实验证明，本系统能够较好地为蛋白质关系抽取和关系网络聚类分析的研究者提供对其进行可视化和辅助实验分析的功能。

在今后的工作中系统的功能还可以进一步扩展，针对不同的聚类算法设计更合理的布局方式，使用户能够更直观地查看并对比不同聚类算法得到的结果，还可以加入聚类方法编辑功能，让用户能够在系统中修改聚类方法的步骤，方便对比实验结果。

参 考 文 献

[1] Batagelj V, Mrvar A. Pajek-program for large network analysis [J]. Connections, 1998, 21(2): 47-57

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

[3] Kohler J, Baumbach J, Taubert J, et al. Graph-based analysis and visualization of experimental results with ONDEX[J]. Bioinformatics, 2006, 22(11): 1383-1390

[4] Orlev N, Shamir R, Shiloh Y. PIVOT: protein interactions visualization tool[J]. Bioinformatics, 2004, 20(3): 424-425

[5] 王柏，吴巍，徐超群，等. 复杂网络可视化研究综述[J]. 计算机科学，2007，34(4)：17-23

[6] 孙扬，蒋远翔，赵翔，等. 网络可视化研究综述[J]. 计算机科学，2010，37(2)：12-18

[7] O'Madadhain J, Fisher D, Smyth P. Analysis and visualization of network data using JUNG[J]. Statistical Software, 2005, 10: 1-35

[8] Xenarios I, Rice D W, Salwinski L, et al. DIP: The database of interacting proteins[J]. Nucleic Acids Research, 2000, 28: 289-291

[9] Stark C, Breikreutz B J, Reguly T, et al. BioGRID: a general repository for interaction datasets[J]. Nucleic Acids Research, 2006, 34: 535-539

[10] Mewes H W, Frishman D, Gruber C, et al. MIPS: a database for genomes and protein sequences[J]. Nucleic Acids Research, 2000, 28: 37-40

[11] Harris M A, Clark J, Ireland A, et al. The Gene Ontology (GO) database and informatics resource[J]. Nucleic Acids Research, 2004, 32: 258-261

[12] Enright A J, Van Dongen S, Ouzounis C A. An efficient algorithm for large-scale detection of protein families[J]. Nucleic Acid Research, 2001, 30: 1575-1584

[13] King A D, Przulj N, Jurisica I. Protein complex prediction via cost-based clustering[J]. Bioinformatics, 2004, 20(17): 3013-3020

[14] Li X, Tan S H, Foo C S, et al. Interaction graph mining for protein complexes using local clique merging[J]. Genome Informatics, 2005, 16(2): 260-269

[15] Wu M, Li X, Kwoh C K, et al. A core-attachment based method to detect protein complexes in PPI networks[J]. BMC Bioinformatics, 2009, 10: 169

[16] Liu G, Wong L, Chua H N. Complex discovery from weighted PPI networks[J]. Bioinformatics, 2009, 25(15): 1891-1897

[17] 安波. 基于蛋白质关系网络的蛋白质络合物抽取研究[D]. 大连：大连理工大学计算机科学与技术学院，2010

[18] Gavin A C, Aloy P, Grandi P, et al. Proteome survey reveals modularity of the yeast cell machinery[J]. Nature, 2006, 440(7084): 631-636