doi:10.3969/j.issn.1672-5565.2014.02.02

基于 WEB 的关键蛋白质预测平台

郑瑞清,李 敏*
(中南大学信息科学与工程学院,长沙 410083)

摘 要:关键蛋白质是指那些在蛋白质相互作用网络中承担重要作用、移除后会使蛋白质复合物功能丧失并导致生物无法存活的节点。随着蛋白质数据库的不断完善和高通量技术的发展,使得通过计算方法的关键蛋白预测得到广泛应用。针对目前软件多为桌面应用程序、用户难以迅速适应的情况,本文设计并实现了一个基于 WEB 的关键蛋白质预测平台 Essential Protein Finder(EP Finder)。该平台集成了 *DC*、*BC*、*CC*、*EC*、*LAC*、*SC* 和 *NC* 7 种关键蛋白质预测算法,还提供包含 *SN*、*SP*、*PPV*、*NPV*、*ACC*、*F* 和折刀曲线图在内的 7 种评估方法。平台对蛋白质网络图、算法运行及评估结果提供了可视化展示。该平台具有良好的扩展性。

关键词:蛋白质网络;关键蛋白质;WEB 平台
中图分类号:Q753 **文献标志码**:A **文章编号**:1672-5565(2014)-02-084-06

A WEB-based platform for predicting essential proteins

ZHENG Ruiqing, LI Min*
(*School of Information Science and Engineering*, *Central South University*, *Changsha* 410083, *China*)

Abstract: Essential proteins are those proteins playing an important role in the protein-protein interaction network, and removal of these proteins will result in a loss of protein complexes′ function and death of organism. The improvement of protein databases and the development of the high-throughput technology have made the computational methods be widely used in prediction of essential proteins. Because the current software is mostly desktop applications, it is difficult to be adapted quickly. Therefore, this paper designs and implements an WEB based protein prediction platform, Essential Protein Finder (EP Finder). It not only integrates seven essential protein prediction algorithms like *DC*, *BC*, *CC*, *EC*, *LAC*, *SC* and *NC*, but also provides seven evaluation measures, including *SN*, *ACC*, *PPV*, *SP*, *NPV*, *F* and a chart of *AUC*. The Platform provides the visual display of protein-protein interaction network, results of algorithms and evaluations. This platform is very scalable as new algorithms and layout methods can be added conveniently.

Keywords: Essential protein; Protein-protein interaction network; Web platform

蛋白质在生物体的生命活动中起着至关重要的作用。关键蛋白质是生物生命活动所必须的蛋白质,通过生物技术手段剔除造成相应的蛋白质复合结构体的功能丧失,从而导致生物死亡[1]。因此识别关键蛋白质在研究生物体的生命活动中具有重大意义。近年来,随着高通量技术的发展[2],蛋白质相互作用的数据大量产生,为通过计算手段预测关键蛋白质方法提供了条件。

研究已表明,蛋白质的关键性和它在相互作用网络中对应结点的拓扑属性紧密相关[3]。根据 H. Jeong 等[4]提出的"中心性-致死性"法则,即当一个蛋白质参与的相互作用越多,这个蛋白质在生物体的生命活动中的作用也越大。基于这种思想提出了一系列中心性算法,常见的比如度中心性(Degree centrality)[4],接近度中心性(Closeness centrality)[5],介数中心性(Betweenness centrality)[6],特征向量中心性

收稿日期:2013-09-07;**修回日期**:2013-11-20.
基金项目:国家自然科学基金(No.61370024);教育部新世纪优秀人才支持计划(NCET-12-0547)资助。
作者简介:郑瑞清,男,硕士研究生,研究方向:复杂网络数据挖掘、生物信息学。
*通信作者:李敏,副教授,研究方向:生物信息学;E-mail: limin@mail.csu.edu.cn.

(Eigenvector centrality)[7],信息中心性(Information centrality)[8],子图中心性(Subgraph centrality)[9]和局部连通性(Local average connectivity-based method)[10]。

目前,相关的工具多是桌面应用程序,对于用户而言,存在着安装、操作繁琐。鉴于这些情况,本文提出并设计了一个新的基于 Web 的关键蛋白预测平台。该平台集成更多的关键蛋白质预测算法以及提供相应的评估方法,实现了蛋白质网络、算法结果、评估结果的可视化显示。

1　EP Finder 总体设计

EP Finder 是一个基于 WEB 的平台,因此在设计上分成两大部分,包括前台的页面和后台的服务。EP Finder 平台采用了 Strut2 框架,能方便地控制页面的跳转和数据传递,前台页面将数据提交给对应的服务类,然后由服务类调用本地的方法来实现具体的功能模块。在前端页面的设计中,采用 Jquery 框架+HTML+CSS 的组合技术,使系统具有简洁、友好的界面。在前台和后台数据交互上选取了 javascript 处理起来更加方便的 json 格式。由于在结果展示部分的用户不需进行不同的操作,为了使用户能有更好的体验,采用了 Ajax 技术实现异步传输。

EP Finder 平台集成的功能主要包括以下几个方面:

(1)蛋白质相互作用网络文件的上传、保存和读取。

(2)提供 *DC*、*BC*、*CC*、*LAC*、*NC*、*EC* 和 *SC* 算法进行关键蛋白质的预测。

(3)提供算法处理并排序后的节点列表,并提供结果的压缩文件供用户下载。

(4)分析算法结果的可靠性,提供包括 *SN*(敏感度)、*PPV*(阳性预测值)、*ACC*(准确率)、*SP*(特异性)、*NPV*(阴性预测值)和 *F*-测度,同时生成相应的折刀曲线[11]。

(5)提供蛋白质相互作用网络的整体展示和局部展示,并给用户提供不同的网络布局方式。

(6)提供关键蛋白质预测算法和蛋白质网络布局方式的接口,方便二次开发和扩展。

2　系统详细设计与实现

整个平台从结构上来分,可分为前台的页面和后台的服务。前台页面用于提供算法任务的提交和结果的显示。页面后台的服务可以根据功能分为关键蛋白算法分析、构造网络图、算法评估和网络图展示(查询)。本章中将从页面和服务以及服务中各个功能模块来介绍平台的详细设计。

2.1　JSP 页面设计

整个系统的页面共分为三个大部分:index.jsp、wait.jsp 和 result.jsp。index.jsp 页面主要的任务在于提供平台信息介绍、任务提交等;wait.jsp 是一个过渡界面,用来缓冲任务提交和运行完成之间的真空期。result.jsp 提供算法运行结果的展示、下载,蛋白质网络图的展示以及评估结果的展示。

2.1.1　Index.jsp 设计

Index.jsp 的作用是提供任务信息的提交和平台相关内容的介绍。Logo 下方为页面的主体,包括两个部分,左边的是菜单栏,右边的是显示框,用来引用其他网页。通过对菜单栏中的每个项设置单击事件,来切换显示框中显示的内容。iframe 中用来显示的页面可以分为交互的任务信息提交页面(Content.jsp)和其他一些无交互的介绍页面。在任务提交页面中,为防止用户的误操作或者任务信息填写错误,该页面还用实现了简单的校验,其中为防止用户输入无效的 E-mail 地址,通过正则表达式匹配进行了简单的校验,表达式如下:"/^[0-9a-zA-Z_\.]+@\w+\.\w+\.? \w+ $/",这样可以减少一些后台的压力。

2.1.2　Result.jsp 页面设计

Result.jsp 承载了整个任务的结果展示的任务,整个页面在实现各项功能的时候和后台的交互非常频繁,比如结点的搜索、局部图的产生和评估结果的生成,页面上实现的功能包括:运算结果的压缩文件下载;蛋白质网络全局图、局部图的展示;关键蛋白质列表上传和评估结果的展示;提供蛋白质网络图生成的相关参数选择;提供用户选择的算法列表,实现不同算法结果显示的切换。

Result.jsp 主要的功能展示集中第三个框中。在这个框中采用了一些 jQuery 提供的插件,包括 jquery-ui、jqgrid 等。在图片的显示区域中,由于受到页面整体大小的限制,采用了 jquery-ui 的 tabs 插件,提供不同标签页之间的切换。算法和评估结果的可视化展示部分采用了 jqgrid 插件。算法展示表格通过和后台的交互来显示不同算法的结果,并提供对结果的筛选。

网络图生成模块会根据用户在算法结果的表格中选择的节点以及生成节点的邻居节点的级数和布局方式,通过 AJAX 的方式将参数传给后台的服务类 PartImageDeal,该类会返回一个状态,根据这个状态,来确定是否添加新的面板。

评估面板需要用户选择输入的参数 N,这个参数 N 表示用户设定运行结果中的前 N 个结点作为关键蛋白质。除此之外,评估还需要真实的关键蛋白质数据,系统为用户提供了两种方法:选择物种或者上传文件。如果上传文件,那么选择物种框会失效,文件上传使用了 ajaxfileupload 插件,实现了异步传输。评估在后台运行完成后,评估表格和折刀曲线图会出现在对应的位置。

2.2 服务器端设计

服务器端承担了整个系统的大部分工作,在设计中,为了减轻前台的负担,并且降低系统的响应时间,牺牲了一些服务器的空间。服务器端的类主要分为两部分:Action 类、功能类和数据类。

Action 类主要是实现前台参数的接受并调用功能类中的方法进行运行,并反馈结果,主要的类如表 1 所示。功能类根据功能可分为算法模块、图片布局模块、评估模块、读写文件模块等。为了使平台具有良好的可扩展性,关键蛋白质识别算法的模块中编写了 AlgoMethod 接口。新增的算法只需继承 AlgoMethod 接口、实现其中的 processNode 方法,并在根目录下 XMLConfig 文件夹中的 Config.xml 中注册,就可以自动在系统中调用。由于大部分关键蛋白质算法在运算的过程中都需要使用蛋白质网络的邻接矩阵,因此设计了一个继承 AlgoMethod 的抽象类 AbstractAlgo,该抽象类实现构建加权邻接矩阵和非加权邻接矩阵的方法。图片布局模块和算法模块相似,编写了 ImageLayout 接口,新增加的也需要继承该接口,实现 layout 方法,并在 Config.xml 中注册。评估模块需要生成折刀曲线(AUC 曲线),在这里调用了 jfreechart.jar 来生成折线图。

表 1 EP Finder 中的 Action 类

Table 1 The Action classes in EP Finder

类名	描述
RunAlgo	调用算法运行和局部图生成模块的 Action
TableInfo	处理用户对算法进行条件查询的 Action
DealPara	处理用户提交任务参数并保存到服务端的 Action
AssessAction	处理 result.jsp 中评估任务的 Action
DealConfig	初始化 result.jsp 内容的 Action
GetStatus	用于处理 wait.jsp 查询任务进度的请求
PartImageDeal	处理 resul.jsp 提出的生成局部图的请求

数据类主要是一些为提高系统响应时间和效率设计的一些数据结构。EP Finder 设计中主要的数据类有 ProteinNode 和 SessionNode。ProteinNode 是节点类,用于保存蛋白质节点的信息,主要的属性有 name、param、ranking。为了使画图和布局两部分松耦合,在 ProteinNode 额外增加了两个属性 X、Y,来表示节点在网络中的 X、Y 坐标。SessionNode 是 session 中存储任务信息的数据结构。session 处于服务器的内存中,并且它的生命周期为用户连接成功到浏览器关闭。它的主要属性包括 Nodes、ResultNodes、Algoname、SearchFiled、SearchString 和 SearchOper。其中 Algoname 是此时在前台表格中展示结果的算法名称;Nodes 是对应 Algoname 算法结果的所有节点信息;ResultNodes 表示用户查询结果的节点引用集合;SearchFiled、SearchString 和 SearchOper 这 3 个字段用来存储用户的查询条件,SearchFiled 是指查询的字段,SearchString 是该字段的值,SearchOper 是指符号,如大于、小于等。SessionNode 可以极大地减少 EP Finder 对用户的查询请求的响应时间。ResultNodes 可以减少用户翻页请求和局部图生成请求的响应时间。并且,ResultNodes 保存的是节点的引用,因此占用的空间极少。SearchFiled、SearchString 和 SearchOper 3 个属性的设置可以减少相同的查询操作,只有当查询的 3 个条件中的一个改变,系统才会重新查询。

3 EP Finder 集成的关键蛋白质预测算法

平台包含的关键蛋白质算法都是比较常用的中心性预测算法,这些算法都是根据蛋白质相互作用网络抽象成普通网络中所含的拓扑属性。

(1)度中心性(Degree centrality)[4] 表示对每个节点根据它的度进行排序,简单的说就是对每个节点边的个数进行排序,如公式(1),其中 d_u 是节点 u 在图 G 中的度:

$$DC(u)=d_u。 \tag{1}$$

(2)网络中心性(Network centrality)[11] 表示的节点 u 的所有临边的聚集系数之和(见公式 2),其中 N_u 表示的是 u 邻接节点的集合,$Z_{u,v}$ 是包括边 (u,v) 的三角形个数,d_u 和 d_v 是 u 和 v 在图 G 中的度:

$$NC(u)=\sum_{v\in N_u} ECC(u,v)=\sum_{v\in N_u}\frac{Z_{u,v}}{min(d_u-1,d_v-1)}。 \tag{2}$$

(3)局部连通性(Local average connectivity-based method)[10] 是表示节点 u 的邻接点的平均局部连通性(见公式 3),其中 N_u 表示的是 u 邻接节点

的集合,C_u 是 N_u 对应的子图。$\deg C_u(\omega)$ 是在子图 C_u 中 ω 对应的度。

$$LAC(u)=\frac{\sum_{\omega\in N_u}\deg\ C_u(\omega)}{|N_u|}。\qquad(3)$$

(4)介数中心性(Betweenness centrality)[6]表示经过 u 的两点间最短路径与两点间的最短路径个数的比值(见公式 4)。

$$BC(u)=\sum_s\sum_t\frac{\rho(s,u,t)}{\rho(s,t)}\quad s\neq u\neq t。\qquad(4)$$

其中$\rho(s,t)$是表示点 s 与 t 之间最短路径总数,$\rho(s,u,t)$表示其中经过中间节点 u 的个数。

(5)接近度中心性(Closeness centrality)[5]是节点 u 到其他节点路径长度平均值的反比(见公式 5),其中 $\mathrm{dist}[u,v]$表示 u 到 v 的最短距离。

$$CC(u)=\frac{N-1}{\sum_v\mathrm{dist}[u,v]}。\qquad(5)$$

(6)子图中心性(Subgraph centrality)[9]表示的是 G 的包含 u 的子图个数。子图越小,权值越大。计算 SC 如公式(6)所示,其中 $\mu_1(u)$表示的是在 u 点长度为 l 的闭环个数,$\alpha_i(1\leqslant i\leqslant N)$是 R^N 型矩阵 A 的标准正交基,并且与特征值 $\lambda_j(1\leqslant j\leqslant N)$相对应。同时,$\alpha_v(\mu)$是 α_v 的第 u 个组成成分(见公式 6)。

$$SC(\mu)=\sum_{l=0}^{\infty}\frac{\mu_l(u)}{l!}=\sum_{v=1}^{N}[\alpha_v(u)]^2e^{\lambda v}。\qquad(6)$$

(7)特征向量中心性 Eigenvector centrality)[7]表示的是节点 u 的最主要的特征向量(见公式 7),其中 $\alpha_{\max}$表示的是节点 u 在矩阵 A 中最大特征值所对应的特征向量。

$$EC(u)=\alpha_{\max(u)}。\qquad(7)$$

4　EP Finder 集成的评估方法

平台中包含的关键蛋白质识别算法的评估方法包含了在实际使用中主要的评估参数和图表,分别是敏感度(*SN*)、阳性预测值(*PPV*)、准确率(*ACC*)、特异性(*SP*)、阴性预测值(*NPV*)和 *F* 度测试(*F*),还有折刀曲线(*AUC*)。评估方法是通过实际的蛋白质关键性列表与算法所得的蛋白质关键性进行比较计算[12]。假设以下参数:*TP* 表示计算认为的关键蛋白质的确是关键蛋白质的数目,*FP* 表示计算认为的关键蛋白质并非是关键蛋白质的数目,*TN* 表示计算认为的非关键蛋白质的确是非关键蛋白质的数目,*FN* 表示计算认为的非关键蛋白质是关键蛋白质的数目,那么可将 *SN*、*ACC*、*PPV*、*SP*、*NPV* 和 *F* 测试的定义如表 2 所示。

表 2　算法评估方法

Table 2　Evaluation methods for different algorithms

评估指标名称	评估方法描述	评估方法表达式
敏感度(*SN*)	算法对关键蛋白正确识别的比例	$SN=\frac{TP}{TP+FN}$
阳性预测值(*PPV*)	选择的蛋白质被正确预测为关键蛋白质的比例	$PPV=\frac{TP}{TP+FP}$
准确率(*ACC*)	正确预测的蛋白质(包括关键蛋白和非关键蛋白)占所有蛋白质的比例	$ACC=\frac{TP+TN}{ALL}$
特异性(*SP*)	非关键蛋白被正确预测成非关键蛋白的比例	$SP=\frac{TN}{TN+FP}$
阴性预测值(*NPV*)	算法预测的非关键蛋白的正确率	$NPV=\frac{TN}{TN+FN}$
F 度测试(*F*)	*SN* 和 *PPV* 的调和平均值	$F=\frac{2*SN*PPV}{SN+PPV}$

折刀曲线(AUC)是一个折线图,其中横坐标表示蛋白质个数,纵坐标表示这些蛋白质中关键蛋白质的个数。图中每一条折线均代表一种算法。

5　EP Finder 展示

EP Finder 主要包括三个页面 index.jsp、wait.jsp 和 result.jsp。index.jsp 主要提供任务提交入口和平台信息的介绍的功能,其中在任务提交页面中,用红色 * 表示的栏目都必须填写。蛋白质相互作用网络数据提交可以通过上传文件或者将数据粘贴到 Input Data 栏目中。通过改变 index.jsp 包含的显示框中的链接,可以让 index.jsp 主体部分显示不同的页面内容,如图 1 所示。

Result.jsp 是 EP Finder 最主要的视图部分。result.jsp 需要为用户提供算法结果、网络局部图和评估方法的可视化展示。用户可以在提交任务时选择的算法之间进行切换,以显示不同的结果,并且,用户可以根据自己的需要显示前几个关键蛋白质。评估模块需要用户提交真实的关键蛋白质列表或者

选择特定的物种,并且由用户输入一个数值参数 N,用来标明算法结果中的前 N 个节点为关键节点。网络图展示的节点对象是根据用户选择的在当前算法表格中的内容而定,假如用户此时选择前 10 个节点,那么就生成这 10 个节点的局部图。在局部图可以生成具有一级邻居节点或者无邻居节点两种蛋白质网络。result.jsp 整体显示如图 2 所示。

图 1 index.jsp 页面

Fig.1 Page of index.jsp

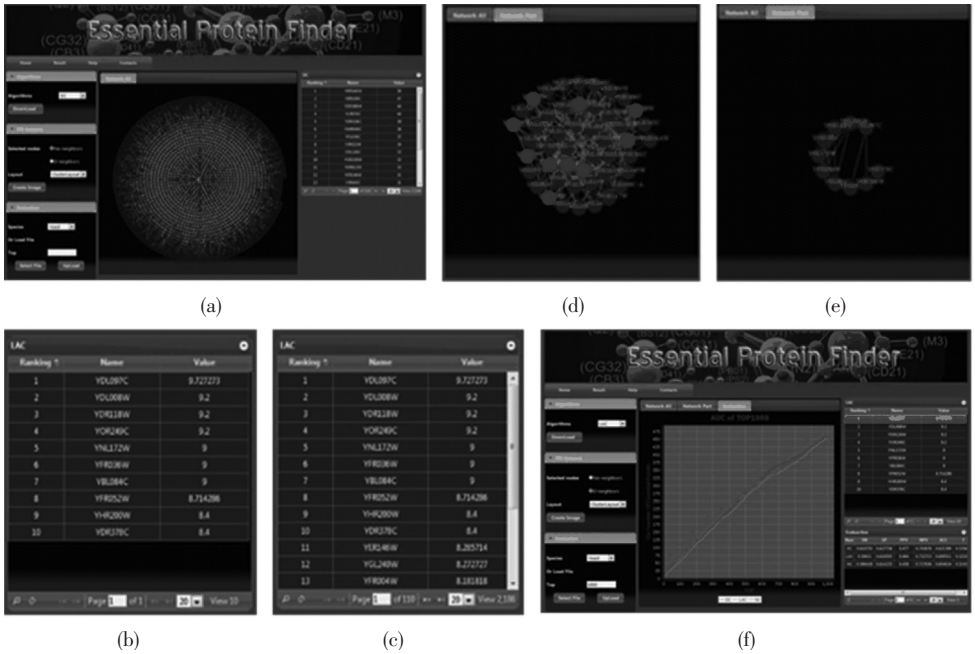

图 2 result.jsp 页面效果

Fig.2 Page of result.jsp

注:(a) result.jsp 初始化状态;(b) LAC 算法结果;(c) LAC 算法前 10 个节点;(d) Top10 的无邻居节点局部图;(e) Top10 的 1 级邻居节点局部图;(f) 评估结果显示。

Notes:(a) The initial view of result.jsp;(b) The result of LAC;(c) The Top10 nodes of LAC;(d) The topography of top10 nodes without neighbors;(e) The topography of top10 nodes with 1 level neighbors;(f) The result of evaluation.

6 结 论

EP Finder 既是一个用于对蛋白质相互作用网络进行可视化分析的平台,也是一个用于对多个预测算法进行比较分析的系统。它集成了 *DC*、*BC*、*CC*、*EC*、*LAC*、*SC* 和 *NC* 共 7 个关键蛋白质预测算法,并包含了 *SN*、*SP*、*PPV*、*NPV*、*ACC*、*F* 和折刀曲线图在内的 7 种评估方法,并且平台为用户提供了算法结果、评估结果和蛋白质网络的可视化展示。用户可根据自己的需要切换算法、选择局部图的内容和查询蛋白质节点。平台具有良好的可扩展性,设计了算法和布局方法的接口。然而,本平台还存在不足之处,例如在蛋白质相互作用网络数据的格式上,平台采用的是自定义的"结点 *A* 结点 *B*"的格式,而非主流数据库提供的 Tab 或者 XML 格式,因

此文件格式转换的功能需要在之后的再次开发中进行补充和升级。同时随着关键蛋白研究的进一步深入,更多的算法被提出,EP Finder 的算法种类也将在之后进行不断的增加。

参考文献(References)

[1] ELIZABETH A W, DANIEL D S, ANNA A, et al. Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis [J]. Science, 1999, 285(5429): 901-906.

[2] CHRISTIAN V M, ROLAND K, BEREND S, et al. Comparative assessment of large-scale data sets of protein-protein interactions [J]. Nature, 2002, 417(6887): 399-403.

[3] ALBERT-LÁSZLÓB , ZOLTÁN N O. Network biology: understanding the cell's functional organization [J]. Nature Reviews Genetics, 2004, 5(2): 101-113.

[4] JEONG H,MASON S P,BARABÁSI A L, et al. Lethality and centrality in protein networks [J]. Nature, 2001, 411(6833): 41-42.

[5] STEFAN W, PETER F S. Centers of complex networks [J]. Journal of Theoretical Biology, 2003, 223(1): 45-53.

[6] MALIACKAL P J, AMY B,DONALD E I, et al. High-betweenness proteins in the yeast protein interaction network [J]. Journal of Biomedicine and Biotechnology, 2005, 2005(2): 96-103.

[7] PHILLIP B. Power and centrality: A family of measures [J].The American Journal of Sociology, 1987, 92(5): 1170-1182.

[8] KAREN S , MARVIN Z. Rethinking centrality: Methods and examples [J]. Social Networks, 1989, 11(1): 1-37.

[9] ERNESTO E, JUAN A R-V. Subgraph centrality in complex networks [J]. Physical Review E, 2005, 71(5): 056103.

[10] LI Min, WANG Jianxin, CHEN Xiang, et al. A local average connectivity-based method for identifying essential protein from the network level [J]. Computational Biology and Chemistry 35,2011:143-150.

[11] ALEXANDER G H, PAUL J D, JEREMY M F, et al. Computational prediction of essential genes in an unculturable endosymbiotic bacterium, wolbachia of brugia malayi [J]. BMC Microbiology, 2009, 9: 2400.

[12] HERMINIA I. Network centrality, power, and innovation involvement: Determinants of technical and administrative roles [J]. Academy of Management Journal, 1993, 36(3): 471-501.