

基因组规模代谢网络模型自动化修正

吴晓红¹, 薛卫², 张梁^{*1}, 石贵阳¹

(1. 粮食发酵工艺与技术国家工程实验室, 江南大学, 江苏 无锡 214122; 2. 南京农业大学 信息科学技术学院 江苏 南京 210095)

摘要: 基于 KEGG 在线数据库以及 6 个蛋白质区间预测数据库, 对基因组规模代谢网络模型进行了自动化修正。作者提出了蛋白质区间预测结果的权重打分机制, 同时利用图像处理算法确定可信度高的特异性反应。上述修正的研究均在 *Spathaspora passalidarum* NRRL Y-27907 基因组规模代谢网络精炼过程中得到运用实施, 对于提高模型构建效率意义重大。

关键词: 基因组规模; 代谢网络; 断点补齐; 图像处理; 区间预测

中图分类号: TP 391; Q 939 **文献标志码:** A **文章编号:** 1673-1689(2017)09-0982-08

Auto-Refinement of Genome-Scale Metabolic Network Model

WU Xiaohong¹, XUE Wei², ZHANG Liang^{*1}, SHI Guiyang¹

(1. National Engineering Laboratory for Cereal Fermentation Technology, Jiangnan University, Wuxi 214122, China; 2. College of Information Science and Technology, Nanjing Agricultural University, Nanjing 210095, China)

Abstract: KEGG online database and six subcellular prediction databases have been studied for the process of auto-refinement. The weighted scoring mechanism was proposed to analyze the results of subcellular prediction databases, using image processing algorithm to determine high credibility specific reaction. As an illustration example, all of the automatic methods were implemented in the process of genome-scale metabolic network refinement of *Spathaspora passalidarum* NRRL Y-27907, which confirmed that these methods can improve the efficiency of model reconstruction.

Keywords: genome-scale, metabolic networks, gaps supplement, image processing, subcellular prediction

随着基因组高通量测序数据的涌现以及大量的生物学数据的产生, 代谢网络模型构建成为研究生物信息学的热点之一。代谢网络构建是一个耗时费力的过程, 因此许多自动化构建的工具随之应运而生。通常这些自动化工具侧重关注代谢网络粗模型的构建如 metaSHARK^[1]和 AUTOGRAPH^[2], 其次关注代谢网络模型的模拟过程, 如 CellNetAnalyzer^[3]、

OptFlux^[4]和 COBRA Toolbox^[5], 只有少量的自动化工具是针对代谢网络模型的精炼过程。目前能够提供代谢网络模型自动化精炼过程的工具有 Model SEED、Pathway Tools、RAVEN 和 SuBliMinaL。

代谢网络的模型构建包括粗模型的构建、模型的精炼、数学模型的转换、模型的预测验证四个过程。一个高质量的代谢网络模型, 应达到模型模拟

收稿日期: 2015-03-02

基金项目: 江苏省自然科学基金项目(BK2012363, BK2011153)。

* 通信作者: 张梁(1978—), 男, 江苏无锡人, 工学博士, 教授, 博士研究生导师, 主要从事代谢工程方面的研究。

E-mail: zhangl@jiangnan.edu.cn

引用本文: 吴晓红, 薛卫, 张梁, 等. 基因组规模代谢网络模型自动化修正[J]. 食品与生物技术学报, 2017, 36(09): 982-989.

结果和生物实际生长表型一致,否则要不断的重复精炼修正过程,直到模拟与表型一致。模型的精炼修正无疑是代谢网络模型构建过程中最耗时耗力的过程,现有模型精炼工具并不能真正实现真菌代谢网络模型精炼过程的自动化。模型的精炼过程必须包括漏洞代谢的填补、反应区间定位等。Model SEED^[6]和 Pathway Tools^[7]只能提供原核生物的代谢网络模型的精炼自动化过程,不能提供反应区间的定位。RAVEN^[8]和 SuBliMinaL^[9]是基于 Wolf PSORT 蛋白质区预测数据库实现自动化定位区间的程序。但是 Wolf PSORT^[10]只是基于氨基酸组成特征的在线预测数据库。研究表明,基于氨基酸组成、二肽和物理化学三种综合特征的蛋白质区间定位预测结果更为准确^[11]。

利用作者所在实验室自动化构建全基因组代谢网络模型的程序,自动构建了 *Spathasporapassalidarum* NRRL Y-27907 全基因组规模代谢的粗模型。以 *S. passalidarum* NRRL Y-27907 的基因组规模代谢网络模型的精炼过程为例,以简单、面向对象的 Java 语言为基础,对精炼过程中人工冗杂的断点补齐的方法进行了研究,提出

了一种基于 KEGG^[12]在线数据库自动化填补漏洞反应的方法,并利用权重打分机制分析,6 个真菌蛋白质定位数据库预测 *S. passalidarum* NRRL Y-27907 的结果,在保证模型中反应的物种特异性的同时,实现了真菌代谢网络模型精炼的自动化。自动化修正的流程见图 1。图中进程 g、进程 n、进程 o 为一个小的流程循环。进程 g 中判断反应包含断点,则进入进程 h,查找该反应在注释图谱中对应的坐标,并在进程 i 中读取此坐标,在进程 j 中判断此坐标是否为特异性反应,如果是,则在进程 p 中记录该反应。如果不是,则在进程 l 中判断此坐标是否为最后一个坐标,如果是最后一个坐标,则进入进程 n,即进入进程 g、进程 n、进程 o 该流程循环。如果不是最后一个坐标,则进入进程 m,读取下一个坐标,判断此坐标是否为特异性反应,重复此循环直至将所有的特异性反应都被找出,进入进程 q,进行模型修正。在进程 r 中判断模型中是否已经包含此反应,若已经包含,则回到进程 n,即进入进程 g、进程 n、进程 o 该流程循环,检查下一条反应。若不包含此反应,则进入进程 s,将此反应加入到模型中。

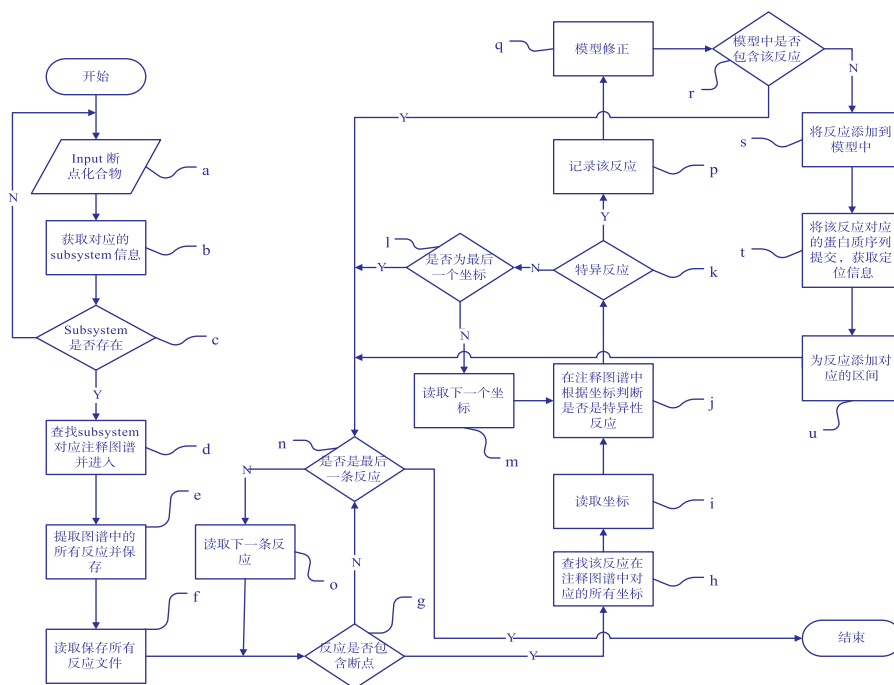


图 1 自动补齐断点流程

Fig. 1 Process of the auto-refinement of gap

1 自动填补网络漏洞

采用柴文平^[13]等人的方法构建了 *S. passalidarum* NRRL Y-27907 代谢网络粗模型。构建的代谢网络粗模型需要进一步精细化与修正,最终完成一个高质量的基因组规模代谢网络模型。

1.1 代谢网络漏洞查找

模型导入到装有 COBRA 工具包和 GLPK 线性规划器的 Matlab 中,将模型转化为计算机可读的格式 (SBML) 才能进行代谢网络漏洞查找。通过 xls2model 程序将模型 Excel 表读取为计量学 S 矩阵。 S 矩阵(828×984)表示该模型由 828 个代谢物和 984 个反应组成。同时通过 GapFind 程序完成代谢漏洞的查找,其中上游漏洞代谢物有 44 个,下游漏洞代谢物有 128 个。

1.2 基于 KEGG 网络爬虫反应

KEGG 是代谢网络构建常用数据库,含有多个在线子数据库,其中 REACTION 数据库包含迄今为止发现的所有生化反应。各个子数据库的网页数据格式比较统一明确,方便人们进行远程服务器访问。但是,KEGG 数据库更新频繁,各个子数据库不能够免费下载,需要付费使用。而在基因组代谢网络断点补齐过程中,因为数据信息量浩大,频繁访问远程服务器比较耗时耗力。因此,实现一种批量在线获取并存取数据的方法意义重大。

1.2.1 方法概述 利用超文本转移协议和 Java 控件 HttpClient 相结合,实现对网页中特定信息的抓取 KEGG 提供物种特异性基因组信息以及所有反应式信息查询网页,通过一定的 URL (Uniform Resource Locator,统一资源定位符)格式地址发送 HTTP 请求并获取网页中的基因信息。在漏洞填补的过程中需要访问大量不同的网络资源,获取相关的基因信息,由于数据量较大且人工操作比较繁琐,这里利用 Java 控件 HttpClient 实现爬虫技术,抓去符合特定条件的网络资源。HttpClient 是 Apache Jakarta Common 下的子项目,可以用来提供高效的、最新的、功能丰富的支持 HTTP 协议的客户端编程工具包,并且它支持 HTTP 协议最新的版本和建议。利用 HttpClient 访问具体的 URL 地址,获取服务器端返回的获取 html 内容,html 内容由标题、js 代码、正文、相关链接、声明等区域组成,而有用信息只出现在正文中的各种 html 标签标记内,分析 html 标签

并获取特定的网页信息。

1.2.2 漏洞填补算法实现

1)获取注释图谱:提交物种基因组蛋白质序列至 KAAS 自动注释服务器,获取注释信息,下载 html 和 text 格式。

2)查找包含断点的注释图谱:根据 Matlab 软件中 GapFind 程序返回的漏洞代谢物列表,在代谢网络模型 Excel 格式中确定代谢物的反应途径,依据 KASS 注释返回的途径图谱找到包含漏洞代谢物的所有反应。

注释返回的 KEGG 代谢途径为包含糖代谢等在内的 110 个途径。查找包含断点的代谢图谱的流程见图 2。具体思路和伪代码步骤如下:

A: 获取断点化合物所对应的 Subsystem 信息,记为 sub。

B:向注释查询网页 URL 地址发送 HTTP 请求。

C: 如果服务器端响应代码为 HTTPStatus.SC_OK 则正常响应,否则继续请求,获取 html 正文内容。

D:分析 html 内容,设 i 为行号,由第一行开始遍历<table></table>标签对中的每一行,

For i from 1 to n

{

if(该行中第二个<td></td>标签中的内容与 sub 相等)

{

提取对应的第一个<td></td>标签中的内容,记为 KO;

}

else

忽略该行,遍历下一行;

}

E: 根据 D 中的 KO 号得到满足条件图谱的 URL 地址,向 URL 地址发送 HTTP 请求得到服务器端响应的网页图片记为 T1,T1 即为整个网络结构图,其中绿色酶号表示包含断点的特异性反应。

F:点击 T1 左上角途径方框,进去包含所有反应页面 page1,网页中每一个 EC 号对应图谱中的一个具体反应,它的 URL 地址指向具体的反应方程式。

G: 获取 page1 中所有 EC 号对应的反应,设 ec_num 为每一个 EC 号,从第一个开始

```

For ec_num from 1 to n
{
  获取 ec_num 在 page1 中对应的 K 号;
  定位到 ec_num 所对应的 URL 地址, 分析<
table></table>标签中的信息;
  获取具体的反应方程 pro_rec;
  将 K 号和 pro_rec 写入文件 ec_k.txt;

}
H: 获取包含断点代谢物的 KEGG 反应。
1.1 按行读取 ec_k.txt 中内容, i 为行号, content
为每行内容
1.2 循环遍历
  For i from 1 to n
  {
    if(content 的反应中包含断点化合物)
    {
      将具体的反应方程 pro_rec 和对应的 K 号写入
      EC_K_Break.txt;
      转到下一行;
    }
    else
      忽略该行, 遍历下一行;
  }

```

EC_K_Break.txt 保存包含断点化合物的

EC, K 号的信息。

3) 查找 EC_K_Break.txt 中每个 K 对应的坐标
根据 K 号获取其在 T1 中对应的坐标, 判断特
异性反应。

A: 重复 E 中的 HTTP 请求, 获取 T1 服务器端
响应的 html 代码, 其中标签为图片 T1, 每个<
area>中的 title 后为 K 号, coords 后为 K 对应的坐标
(x1, y1, x2, y2)。

B: 读取 EC_K_Break.txt 中每行 i 中的 K 号记
为 k_num, 分别到 html 的标签中遍历每一行<
area>, 行号为 j, 找到 K 号对应的 coords 坐标。

```

For i from 1 to n
{
  获取 K 号 k_num;
  For j from 1 to n
  {
    if(title 中的 K 号和 k_num 相同)
    {
      获取 coords 后对应的坐标 position;
    }
    else
      查找下一行;
  }
}

```

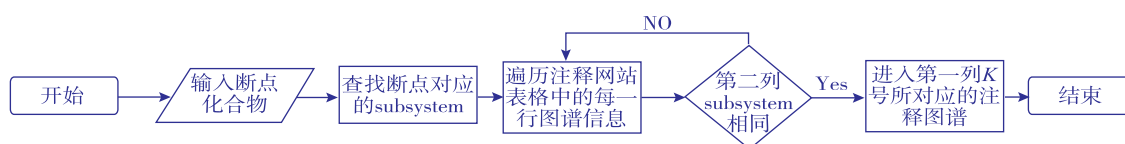


图 2 断点代谢途径定位

Fig. 2 Orientation of gap metabolic pathway

1.3 判断特异性反应

KEGG 所有的反应都包含在通路数据库 (PATHWAY database) 中, PATHWAY 图谱上有颜色标记的酶号是指这个物种特定的基因或酶, 只有有颜色标记的酶号表示的反应才是具有该物种特异性的反应, 也才能添加到代谢网络模型中。在代谢网络模型中添加非特异性的反应会改变整个代谢途径和代谢物流量, 进而使模型模拟的结果偏离实验数据, 影响模型的准确性和可信度。

构建代谢网络模型需要提取代谢途径中的特

异性反应, 图中特异性反应对应的酶号所在的方形框有颜色标记。因此通过网络爬虫技术获得方形框的位置列表, 定位到某酶号所在的方形框后需要选取框内的像素点, 读取其颜色值, 如果颜色分量 RGB 均为 0 或 255, 则没有颜色标记, 反之则有。代谢网络特异性反应获取流程见图 3。

基本思路为:

根据得到的 position 坐标读取 T1 对应点的 RGB 色彩值。

Picture(Key: 酶号; Value: 代谢网络图中所有方

形框的坐标向量集 $\{V_1, V_2, \dots, V_n\}$)

For i from 1 to n

{

If(某酶号所在的方形框)

{

沿方形框的长边内侧逐一选取像素点, 读取其颜色值;

If 颜色分量 RGB 均为 0 或 255 then 没有颜色标记

else 有颜色标记;

If 有颜色标记 then 该酶号对应的是特异性反应

do 将反应加入菌的代谢网络模型中;

else 舍弃该酶号对应的反应。

}

}

反应式漏洞填补

遍历 new_rec.TXT 中每一个反应, 查看模型中是否存在, 存在则不处理, 否则添加。

A: 读取 new_rec.TXT 中每行反应记为 new_rec, i 为行号

For i from 1 to n

{

if(模型中不包含 new_rec)

{

将 new_rec 添加到模型中;

}

else

忽略该反应, 查找下一条反应;

}



图3 特异性反应获取流程

Fig. 3 Process of getting the specific reaction

2 获取反应区间定位

细胞是生命活动的基本单位, 它由执行不同机体功能的称为亚细胞的各部分组成, 如细胞膜、细胞核、线粒体、高尔基体、内质网等。亚细胞功能是由位于其中的蛋白质执行的, 蛋白质所在的亚细胞称为蛋白质的亚细胞位置^[14]。蛋白质必须转运到其应在的亚细胞位置上才能正确行使其功能, 否则就会出现机体功能紊乱, 正确合理的蛋白区间定位是高质量模型构建的基础, 见表1。

表1 真菌蛋白质亚细胞预测数据库

Table 1 Database for subcellular localization of fungal proteins

数据库名称	数据库网站
PSORT II ^[17]	http://psort.hgc.jp/form2.html
BaCeILO ^[18]	http://gpcr2.biocomp.unibo.it/bacello/pred.htm
Cello ^[19-20]	http://cello.life.nctu.edu.tw/
EpiLoc ^[21]	http://epiloc.cs.queensu.ca/
SLPFA ^[22]	http://sunflower.kuicr.kyoto-u.ac.jp/~tamura/slpfa.html
Euloc ^[23]	http://sunflower.kuicr.kyoto-u.ac.jp/~tamura/slpfa.html

确定一条蛋白质的亚细胞位置称为蛋白质亚细胞定位^[15]。蛋白质亚细胞定位的传统方法是通过生物化学实验, 如射线晶体衍射电子显微镜核磁共振等方法进行测定^[16]。实验方法精确度高, 但费时耗力代价昂贵, 而且对难于结晶的蛋白质来说, 实验方法不再有效。借助于先进高效的计算机自动化数据处理技术, 出现了一些蛋白质定位预测网站。结合 *Spathasporapassalidarum* NRRL Y-27907 的生理生化性质和蛋白质特征提取方法、算法和准确性等, 选取了6个真菌生物蛋白质区间预测网站, 自动化提取分析网站的预测结果, 在权重打分机制的基础上得到最佳的蛋白质定位区间。这6个网站是基于蛋白质的氨基酸组成、伪氨基酸组成、二肽、生物化学特征或是四种特征的综合。

2.1 区间定位算法实现

A: 对每条反应获取对应的 KO 号。

B: 将 A 中的 KO 号在 KASS 注释结果中查找基因号, 并在本地下载 *Spathasporapassalidarum* NRRL Y-27907 蛋白质序列库提取其对应的蛋白质序列。

C:将蛋白质序列提交到对应网站的表单中,获取返回的定位信息。

D:获取定位区间的信息并填入反应式中。

在获取具体反应的区间信息过程中,需要将反应所对应的蛋白质序列提交到网页的表单中,提交后返回具体的区间定位信息,此处会遇到两个问题:1) 表单提交过程中不支持大量蛋白质序列自动

提交。由于模型中蛋白质序列数量较大,在有的网站中获取定位信息时不支持大量序列的一次性提交而只能分别提交单个序列获得定位信息,在提交过程中任务量大且人工耗费时间长。2)大量蛋白质序列提交耗费时间长,在网站中提交多个序列后等待服务器端反馈的定位信息耗费时间太长,甚至会发生无响应等问题,见图4。

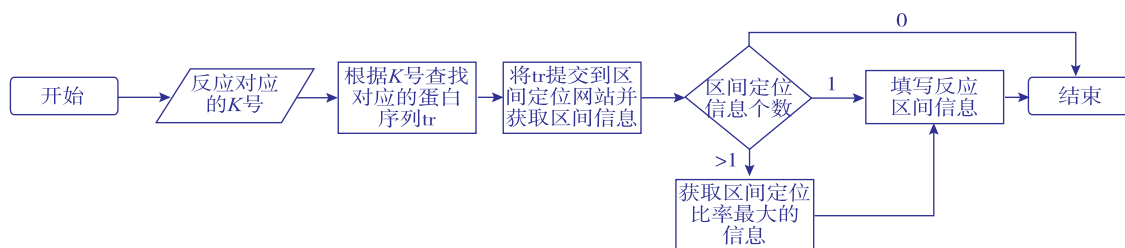


图4 反应亚细胞定位流程

Fig. 4 Process of subcellular localization

HttpClient 支持访问特定的 URL 地址,获取服务器端返回的 html 信息,并且能够分析 html 中 form 表单中的信息,实现内容的自动提交。由于涉及到的定位页面所有的表单提交方式都是 POST 提交,利用 HttpClient 中的 PostMethod 方法实现 post 提交。表单中的元素赋值过程:获取表单中需要赋值的元素标签,以蛋白质序列元素赋值标签为例,标签为<textarea name="input">,标签名称为 input,利用 PostMethod 中的 addParameter 方法为具体标签元素赋值,形式如下 PostMethod.addParameter("input",protein);其中 protein 为蛋白质序列。提交过程:调用 HttpClient 的 executeMethod 方法将赋值后的表单提交到服务器并获取服务器返回的 html 信息,其中具体的定位信息包含在 html 文本中,分析提取即可。

2.2 权重打分机制分析蛋白质区间定位

6 个蛋白质区间预测网站的训练数据集不同,不能单纯依据文献里公布的预测结果准确率来预测 *Spathasporapassalidarum* NRRL Y-27907 各个蛋白质区间的准确率。同时这 6 个网站各个区间预测的准确性也有区别。因此我们整合了 RH2427^[24]和 PK7579^[25]数据集,组成一个包含 12 个蛋白质区间,每个区间包含 100 条蛋白质序列的真菌蛋白质数据集,计算出 6 个网站基于新的真菌数据集上的权重。

权重是一个相对的概念,针对某一指标而言。某一指标的权重是指该指标在整体评价中的相对

重要程度。研究 6 个蛋白质预测网站指标体系权重计算,反映各个蛋白质预测网站在预测结果中的重要性程度的数量。具体计算步骤如下:

第一步,统计每个预测网站各个区间预测正确的蛋白质序列个数。

第二步,计算出每个预测网站的平均识别正确数量,设 $X\{X_1, X_2 \cdots X_{12}\}$ 为每个预测网站 12 个区间的正确预测区间个数,则每个网站平均识别正确数量为: $D=(X_1+X_2+\cdots+X_{12})/12$, 计算结果见表 2 最后第二列。

第三步,计算 6 个预测网站的权重,计算结果见表 2 最后一列。cello 权重 = $89.3/(89.3+62.4+72.7+62.0+56.4+85.1)=0.208$; Psort II 权重 = $62.4/427.9=0.146$; Epiloc 权重 = $72.7/427.9=0.170$; Bacello 权重 = $62.0/427.9=0.145$; SLPFA 权重 = $56.4/427.9=0.132$; Euloc 权重 = $85.1/427.9=0.199$ 。

根据各个蛋白质区间预测网站的权重,采用加权投票的方式计算每个蛋白质序列最佳的蛋白质区间定位。本研究中采用的权值是每一个蛋白质区间预测网站的准确率,而不是对预测结果中区间票数的简单加和,这样就充分考虑到每一个蛋白质区间预测网站的准确率,有区别的对待了每一个预测网站,更符合实际预测结果。加权投票方式可以采用以下公式^[26]表示:

$$V_i = \sum_{n=1}^N w_n * f_n^i (i=1, 2, \cdots, c),$$

表 2 数据库的权重

Table 2 Weight of database

预测网站	chlo	cytop	cytos	er	extr	golgi	lyso	mito	nuc	pero	plas	vacu	acc	wei
1	96	86	88	83	84	85	93	91	96	86	85	98	89.3	0.208
2	-	67	-	42	65	-	-	58	78	61	66	-	62.4	0.146
3	-	75	-	66	75	65	-	97	77	83	60	56	72.7	0.170
4	-	71	-	44	44	63	-	70	80	-	-	-	62.0	0.145
5	40	56	-	48	59	46	-	74	72	-	-	-	56.4	0.132
6	81	89	61	85	95	70	86	97	86	87	90	94	85.1	0.199

注:1表示 cello;2表示 Psort;3表示 Epiloc;4表示 Bacello;5表示 SLPFA;6表示 Euloc

其中, V_i 表示第 i 条蛋白序列的判决区间结果; w_n 为第 n 个蛋白区间预测网站的权重, 其中保持 $\sum_{n=1}^N w_n=1$; f_n^i 表示第 i 条蛋白质序列在第 n 个区间预测网站上的预测结果, 其为 m, c, n, e 等 12 个区间中的一个值(实际预测结果中待预测蛋白质可能只对应这 12 个区间值中的某几个), N 表示选取的真菌蛋白质区间预测网站个数, c 表示所要预测的蛋白质序列数。当对输入的待测蛋白质序列做判决时, 把预测蛋白质区间在每一类区间的得票量排序, 把待测蛋白质序列划分到得票量最大的区间所在的类。当加权计算后, 预测结果中对应多个得票量最大的区间时, 则认定该蛋白质对应多个蛋白质预测区间。

下面通过具体的实例说明一下采用加权方式计算每个蛋白质序列最佳蛋白质区间定位的步骤。如 SPAPADRAFT_69954 编码的蛋白质序列上传到 6 个预测网站, 返回的预测结果是 cello、Epiloc、SLPFA 预测的区间为 cytos, Psort II 预测的区间为 er, Bacello 预测区间为 mito, Euloc 预测结果为 golgi。

第一步那么根据各个网站的权重和公式计算得 $y=0.208\text{cytos}+0.146\text{er}+0.170\text{cytos}+0.145\text{mito}+0.132\text{cytos}+0.199\text{golgi}=0.51\text{cytos}+0.146\text{er}+0.145\text{mito}+0.199\text{golgi}$

第二步, 预测结果在每一类区间得票量排序, $0.51>0.199>0.146>0.145$, 蛋白质序列在各个区间的概率为 $\text{cytos}>\text{golgi}>\text{er}>\text{mito}$, 则最佳蛋白质预测区间为 cytos。

3 讨论

根据 TCDB 和 TransportDB 数据库返回的转运

反应信息, 添加不同细胞器间的代谢物转运反应, 使整个代谢物网络模型连接起来。将精炼后的模型转化为计算机可读的格式(SBML)进行模拟分析。通过 xls2model 程序将模型 Excel 表读取为计量学 S 矩阵。模型由包含 828 个代谢物和 984 个反应的粗模型, 扩充到包含 873 个代谢物和 1 243 个反应的精细模型。

在补充断点的过程中, 对于模型新添加的代谢物信息, 自动化从网站上抓取, 节省了大量的时间和劳力。代谢物信息则包括代谢物缩写(mets)、代谢物全称(metNames)、代谢物分子式(metFormulas)、带电荷数(metCharge)、代谢物分区(metCompartment)、在 KEGG 数据库中的编号(metKEGGID)、在 PubChem 数据库中的编号(metPubChemID)、在 EBI 数据库中的编号(metChEBIID)等。

基于 Matlab 平台的 COBRA 工具, 打开 GLPK 线性规划器, 输入 `[allGaps, rootGaps, downstreamGaps] = GapFind(model, false, false)` 命令, 补齐后的模型, 不存在任何断点, 说明此普适性自动化程序的可行性。

4 结语

国内外虽然对代谢网络自动化修正也做了多方面的研究, 但仍然不能够实现完全自动化修正真菌代谢网络, 代谢网络模型构建过程中仍然需要手工添加模型特异性反应和反应区间。作者提出基于 KEGG 数据库的网络爬虫和自动化提取蛋白质区间预测结果, 并加权计算分析结果, 能够在修正一个相对精炼模型的过程中实现计算机技术与代谢网络模型修正的最大化结合, 减少了修正过程中大量的劳力与时间, 提高了代谢网络构建的效率与精确性。

参考文献:

- [1] PINNEY J W, SHIRLEY M W, MCCONKEY G A, et al. MetaSHARK: software for automated metabolic network prediction from DNA sequence and its application to the genomes of *Plasmodium falciparum* and *Eimeria tenella*[J]. **Nucleic Acids Research**, 2005, 33(4): 1399-1409.
- [2] NOTEBAART R A, van Enckevort F H, FRANCKE C, et al. Accelerating the reconstruction of genome-scale metabolic networks [J]. **BMC Bioinformatics**, 2006, 13(7): 296.
- [3] STEFFEN K, JULIO S R, ERNST D G. Structural and functional analysis of cellular networks with CellNetAnalyzer[J]. **BMC Systems Biology**, 2007, 1(1): 1-13.
- [4] ROCHA I, MAIA P, EVANGELISTA P, et al. OptFlux: an open-source software platform for in silico metabolic engineering[J]. **BMC Systems Biology**, 2010, 14(1): 45-57.
- [5] SCHELLENBERGER J, QUE R, FLEMING R M T, et al. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0[J]. **Nature Protocols**, 2011, 6(9): 1290-1307.
- [6] HENRY C S, DEJONGH M, BEST A A, et al. High-throughput generation, optimization and analysis of genome-scale metabolic models[J]. **Nature Biotechnology**, 2010, 28(9): 977-982.
- [7] KARP P D, PALEY S M, ROMERO P. The pathway tools software[J]. **BMC Bioinformatics**, 2002, 18: S225-S232.
- [8] AGREN R, LIU L M, SHOAIE S, et al. The RAVEN Toolbox and its use for generating a genome-scale metabolic model for *Penicillium chrysogenum*[J]. **PLoS Computational Biology**, 2013, 9(3): 1-16.
- [9] SWAINSTON N, SMALLBONE K, MENDES P, et al. The subliMinaL toolbox: automating steps in the reconstruction of metabolic networks[J]. **J Integr Bioinform**, 2011, 8(2): 186.
- [10] PAUL H, KEUN-JOON P, TAKESHI O, et al. WoLF PSORT: protein localization predictor[J]. **Nucleic Acids Research**, 2007, 35: W585-W587.
- [11] LU Z, SZAIFRON D, GREINER R, et al. Predicting subcellular localization of proteins using machine-learned classifiers[J]. **BMC Bioinformatics**, 2004, 20(4): 547-556.
- [12] KANEHISA M, GOTO S. KEGG: Kyoto encyclopedia of genes and genomes[J]. **Nucleic Acids Research**, 2002, 28(1): 27-30.
- [13] CHAI Wenping, XUE Wei, ZHANG Liang, et al. Research on the auto-reconstruction of genome-scale metabolic network model [J]. **Journal of Food Science and Biotechnology**, 2014, 33(9): 957-967. (in Chinese)
- [14] 韩榕. 细胞生物学[M]. 北京: 科学出版社, 2011: 55-106.
- [15] 叶子弘. 生物信息学[M]. 杭州: 浙江大学出版社, 2011: 179-223.
- [16] LIU Liyuan, CHEN Yuehui, MA Bingxian, et al. Prediction of protein subnuclear location using evolutionary fuzzy K-nearest neighbors and its ensemble[J]. **Journal of University of JINAN**, 2010, 24(4): 376-379. (in Chinese)
- [17] NAKAI K, HORTON P. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization [J]. **Trends in Biochemical Sciences**, 1999, 24(1): 34-36.
- [18] ANDREA P, PIER L M, FARISELLI P, et al. BaCellLo: a balanced subcellular localization predictor[J]. **BMC Bioinformatics**, 2006, 22(14): 408-416.
- [19] YU C S, LIN C J, HWANG J K. Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions[J]. **Protein Science**, 2004, 13(5): 1402-1406.
- [20] YU C S, CHEN Y C, LU C H. Prediction of protein subcellular localization[J]. **Proteins: Structure, Function and Genetics**, 2006, 64(3): 643-651.
- [21] BRADY S, SHATKAY H. EpiLoc: a (working) text-based system for predicting protein subcellular location[J]. **Pacific Symposium on Biocomputing**, 2008(13): 604-615.
- [22] TAMURA T, AKUTSU T. Subcellular location prediction of proteins using support vector machines with alignment of block sequences utilizing amino acid composition[J]. **BMC Bioinformatics**, 2007, 8(1): 466-478.
- [23] CHANG TH1, WU L C, LEE T Y. EuLoc: a web-server for accurately predict protein subcellular localization in eukaryotes by incorporating various features of sequence segments into the general form of Chou's PseAAC[J]. **Journal of Computer-Aided Molecular Design**, 2013, 27(1): 91-103.
- [24] GARG A, RAGHAVA P S. ESLpred2: Improved method for predicting subcellular localization of eukaryotic proteins[J]. **BMC Bioinformatics**, 2008, 9(1): 1-10.
- [25] CHOU K C, SHEN H B. Euk-mPLoc: A fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites[J]. **Journal of Proteome Research**, 2007, 6(5): 1728-1734.