

基因组规模代谢网络模型的自动化重构

柴文平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: 对基于 KEGG 在线数据库、Uniprot-MetaCyc 数据库, 以及同源比对 3 种构建基因组规模代谢网络模型的方法进行了自动化研究。同时提出了基于反应式字符频度直方图的马氏距离比对算法, 并应用于模型整合和模型核心反应的识别。上述自动化方法的研究均在树干毕赤酵母基因组规模代谢网络重构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; 自动化重构; 直方图

中图分类号: TP 391; Q 939 **文献标志码:** A **文章编号:** 1673—1689(2014)09—0957—09

Research on the Auto-Reconstruction of Genome-Scale Metabolic Network Model

CHAI Wenping¹, 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: Three kinds of methods for reconstruction of genome-scale metabolic network model, which were based on KEGG online database, Uniprot-MetaCyc databases and homologous alignment, have been studied for the process automation. Meanwhile, it has proposed an algorithm called Mahalanobis distance which was used to calculate the distance between reactions character frequency histogram. This algorithm can be used to the auto-integration of the draft model reactions and identify the core breakpoint. As an illustration example, the all automatic methods were implemented in the process of genome-scale metabolic reconstruction of *Scheffersomyces stipitis*, and confirming that these can improve the efficiency of model reconstruction.

Keywords: genome-scale, metabolic networks, auto-reconstruction, histogram

随着物种基因组测序的完成以及大量生物学数据的产生, 系统生物学研究技术也日益成熟。系统生物学能够模拟和推测复杂生物体行为, 其中网络模型模拟是最主要的模拟方法^[1-2]。系统生物学的

网络模型种类包含代谢网络^[3]、转录调控网络^[4]、信号转导网络^[5]和转录翻译网络^[6]等。而其中的基因组规模代谢网络模型(Genome Scale Metabolic Model, GSMM)已经成为系统生物学不可或缺的研究工具。

收稿日期: 2013-12-05

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, 主要从事代谢工程研究。E-mail: zhangl@jiangnan.edu.cn

它通过整合基因组学、文献组学、蛋白质组学等组学数据,建立由基因-蛋白质-生化反应(G-P-R)关联组成的特定生物代谢网络,是从全局深刻理解其生理特性与定向调控工业微生物生理功能的重要平台^[7]。

截止到2013年11月,已经公布了98个生物全基因组代谢网络模型,其中细菌65个,古细菌6个,真核生物28个^[8]。而GOLD数据库公布的已完成测序物种包含2698个,其中细菌2384个,古细菌163个,真核生物151个^[9]。前者数量远远小于后者,造成这种情况的原因除了对很多物种生理生化机制了解较少之外,更重要的是重构代谢网络过程需要大量的人工,非常耗时耗力^[10]。虽然已经出现了一些通过自动获取信息初步重构网络的软件平台,如SEED服务器与GEM System软件等^[11-13],但其中仍然需要大量的人工操作与修正工作,如SEED注

释的基因序列需要人工进行,同NCBI注释的基因序列相一致,以便后续其他构建方法的补充与修正。此外,大量生物信息学数据库的出现也对进一步实现网络自动化重构提出了挑战。因此,代谢网络重构的自动化研究,已成为推动代谢网络发展的重大课题。

文中以构建树干毕赤酵母(*Scheffersomyces stipitis* 或者 *Pichia stipitis*)^[14-15] CBS 6054 的基因组规模代谢网络模型为例,以简单、面向对象的Java语言为基础,对代谢网络自动化重构的方法进行了研讨,提出了一种基于KEGG在线数据库来自动化构建初模型的方法,并对基于Uniprot-MetaCyc本地数据库,以及亲缘物种同源比对构建初模型的方法,和整合过程进行了自动化研究,达到代谢网络模型构建的最大程度自动化。自动重构流程如图1所示。

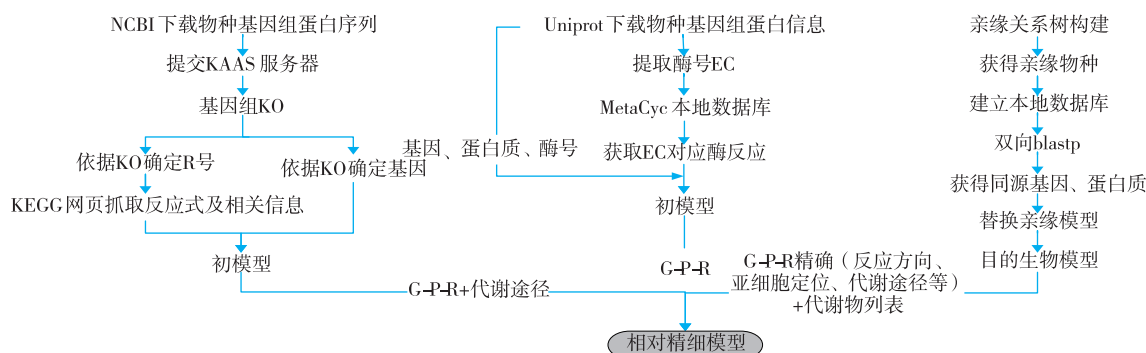


图1 模型自动重构流程

Fig. 1 Process of the auto-reconstruction of model

1 代谢网络自动重构

基因组规模代谢网络模型构建过程主要涉及^[16]:初模型反应列表的获得;模型精细化;转换数学模型;模拟与应用。一般认为构建出G-P-R反应列表就等于完成了代谢网络的初模型^[10]。虽然构建G-P-R反应列表过程比较容易理解,但由于基因组规模的大量生物信息学数据涌现,反应列表的构建反而成为构建过程中最繁琐耗时耗力的一部分^[17]。所以借助计算机技术来达到高效构建反应列表、提高代谢网络模型构建效率的作用。

KEGG (Kyoto Encyclopedia of Genes and Genomes) 与 MetaCyc (<http://metacyc.org/>) 是模型构建中最常用的网络数据库,包含了物种的基因与基

因组、酶促反应及其代谢途径和化合物等相关信息^[18-19],是网络构建过程中网络数据信息的主要来源。而同源比对这一比较基因组学策略可以快速找到亲缘菌株之间的遗传关系及对应的生化信息,这一策略是构建全新微生物代谢网络模型的可靠信息来源。

1.1 基于KEGG在线数据库的模型构建

KEGG作为代谢网络构建常用数据库,其包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^[18]。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。而在重构基因组代谢网络过程中,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。

1.1.1 方法概述 超文本转移协议(Hypertext transfer protocol,HTTP)是一种详细规定了浏览器和万维网服务器之间互相通信的规则,通过因特网传送万维网文档的数据传送协议。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,返回网页 html 脚本含有基因组信息或者反应式相关信息。html 脚本由标题、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栏的<table>标记内。如果对于每个查询网页均全面分析,将大大降低效率。因此作者提出基于正文先验位置的网页分析方法,获取 html 的正文信息中的<form>标志后第一个<table>在 html 脚本字符串中的起始位置 begin_pos, </form>标记前最近一个</table>结束位置 end_pos, begin_pos 至 end_pos 即正文先验位置,只处理 begin_pos 至 end_pos 脚本串信息即可获得反应等信息。具体可采用 JAVA 控件 NekoHTML 分析 html 脚本中每个节点数据,并用正则表达式提取相关信息。

1.1.2 算法实现

1)确定 KO 以及 R 号:提交物种基因组蛋白质序列至 KAAS 自动注释服务器,下载 KEGG BRITE Database 中 KO-R 列表至本地为 K-R.xls,依次读取返回的 KO 号,自 K-R.xls 确定相应的 R 号,写入 KR.xls 中。

2)依据 KO 获得基因蛋白质信息:

向反应式信息查询网页 URL 地址发送 HTTP 请求。

服务器响应代码串为 Gene_string, 分析提取 html 正文信息中的<form>标志后的<DIV>节点,节点数为 DIV_number。

设 $n=1$ 。对第 n 个节点,依据 JAVA 正则表达式 " $K+\backslash d\{5\}$ " 提取对应的 KO 号。

读取 KR.xls 中的每个 KO,与上步中 KO 比较,若相等,设置 JAVA 正则表达式 " $sign+\backslash_+\backslash d+$ " 提取 GENE,设置 JAVA 正则表达式 " $\backslash s+(\backslash w+)+\backslash ;$ " 提取 PROTEIN。一并写入表中 KO 对应行。

n 增加 1, 如果 $n \leq \text{DIV_number}$, 重新执行;否

则,结束。

3)依据 R 号获得反应式及相关信息:

读取上步生成的 KR.xls。

设 $i=1$ 。

读取 KR.xls 的第 i 行 R 值,设置 KEGG 服务器访问地址为 "http://www.genome.jp/dbget -bin/www_bget?+R 值",发送 http 请求。

获取 html 格式脚本,服务器响应代码串为 reaction_string。如果 $i=1$,首先计算正文先验位置,在 reaction_string 中分析获取 html 的正文信息中的<form>标志后第一个<table>在 html 脚本字符串中的起始位置 begin_pos 并保存。在 reaction_string 中查询获得 end_pos 值,得到<table>...</table>内容字符串 content_string。

用 NekoHTML 读取 content_string 中 Name、Definition、Equation、Enzyme、Pathway、Orthology 字段值并写入该表对应行中。

i 增 1,如果 $i \leq \text{"P-K.xls 表行数"}$,重新执行;否则,结束。

4)实例结果:以树干毕赤酵母为例,实现流程如图 2 所示。系统运行大约 30 min,得出的初模型包含 786 条反应及关联基因(549 个)、酶及代谢途径等信息,不仅节省了大量的人力和时间,而且保证了数据的最新性。



图 2 KEGG 网页挖掘数据实现流程

Fig. 2 Procedure of the excavation of data based on the web page of KEGG

1.2 基于 Uniprot-MetaCyc 本地数据库的模型构建

MetaCyc 数据库具有可信度高、信息全面、易访问、免费用于学术研究等优点^[20]。相较于 KEGG 来说,也存在格式繁复、网页打开速度慢等缺点,所以不容易网页查找与数据抓取。因此可构建 MetaCyc 本地数据库进行数据提取。MetaCyc 数据同样存在无物种特异性问题,只能依据物种特异性信息从数据库中查找提取特定的反应。UniProt^[21]是信息最丰富、资源最广的蛋白质数据库。它包含高质量的、手工注释的、非冗余的数据集。可以在 UniProtKB 中找到物种特异性的基因组注释的蛋白质相关详细信息。

1.2.1 方法概述 通过 UniProt 中的酶号 (EC number) 搭建一条连接 UniProt 与 MetaCyc 数据库的桥梁, 从而获取物种特异性的基因组反应信息。UniProt 数据库 (<http://www.uniprot.org/>) 中下载物种特异性基因组蛋白质注释信息至本地, 建立 Excel 数据库 DBU。下载 MetaCyc 数据库至本地, 建立本地 Excel 数据库, 其中包含有 EC 和酶学反应号 (ERN) 的 DBMR 子数据库和有 ERN 和对应酶学反应式 (ER) 的 DBME 子数据库。

设 DBU 中有 N 条有 EC 的蛋白质信息, $N>1$ 。DBMReaction 中有 M 条有 EC 和对应 ERN 的信息, $M>1$ 。DBMEnzymes 有 K 条 ERN 对应的 ER 的信息, $K>1$ 。 $E_i=(EC_i, G_i, P_i)$ 表示第 i 条 EC 对应的基因 (G) 和蛋白质 (P) 信息, $1\leq i\leq N$; $F_i=(EC_i, ERN_i)$ 表示第 i 条 EC 对应的 ERN 信息; $H_i=(ERN_i, ER_i)$ 表示第 i 条 ERN 对应的 ER 信息; $EFH_i=(EC_i, G_i, P_i, ERN_i, ER_i)$ 表示第 i 条初模型反应信息 (酶号、基因、蛋白质、酶学反应编号、酶学反应式)。求取某个 EC 对应的反应信息表示为 $EFH_i=E_i\cup F_i\cup H_i, 1\leq i\leq N$ 。设 $EFH_0=(EC_0, G_0, P_0, ERN_0, ER_0)$ 表示整个初模型, 则将问题转化为求解各反应信息的合集, 即求解 $EFH_0=\sum EFH_i$ 。

1.2.2 方法实现 Uniprot 与 MetaCyc 数据库数据均有文献支持、手工注释录入的特点, 可信度较高。其中, MetaCyc 数据库数据为无物种特异性的酶及酶学反应信息。为了获得物种特异性的 G-P-R 关系列表, 则需由 Uniprot 获得物种特异性基因组蛋白质注释信息, 依据信息中的酶号信息自 MetaCyc 数据库获得相应的酶学反应信息。初模型 G-P-R 列表获得流程如图 3 所示。

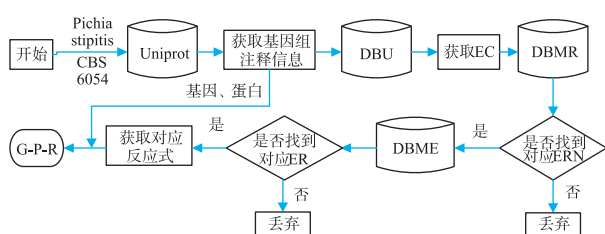


图 3 基于 UniProt-MetaCyc 的 G-P-R 获取流程

Fig. 3 Process of getting the G-P-R list based on UniProt-MetaCyc

1.3 基于同源物种的模型构建

比较基因组学的策略可以快速找到两物种间

的遗传关系与对应的生化反应信息。基于局部比对算法的搜索工具 (Basic Local Alignment Search Tool, BLAST)^[22] 对目的生物与其亲缘物种的基因组序列信息进行双向比对, 可推测大部分同源序列的基因功能^[23]。在比对过程中蛋白质序列比基因序列有更高的保守性, 所以基于蛋白质序列的同源比对有效性也更高。在蛋白质序列双向比对 (Blastp) 过程中, 通过设置一定的期望值 (E)、相似度 (Identity) 与匹配序列长度来确定两者是否为同源序列。当确定两个蛋白质序列为同源序列后, 可推测两者具有相似的酶学功能, 进而推测其可以作用于同一反应。

1.3.1 方法概述 通过亲缘关系树的建立, 查找目的生物的亲缘物种, 并下载亲缘物种的高质量代谢网络模型, 建立本地数据库。依据目的生物与亲缘物种的基因组蛋白质序列双向比对结果, 和每条序列的登录号 (Accession Number), 自动获得蛋白质对应的基因, 进而根据基因-酶-反应的关联与亲缘物种的全基因组规模代谢网络模型, 获得目的生物的规模代谢网络模型。利用 Java 分别调用 Poi_3.7.jar 和 tm_extractors_0.4.jar 工具包函数, 对 excel 和 word 进行操作。

1.3.2 方法实现 以树干毕赤酵母 CBS 6054 为例, 选择亲缘菌株为巴斯德毕赤酵母 (*Pichia pastoris*) 和酿酒酵母 (*Saccharomyces cerevisiae*), 并分别以酿酒酵母的 iMM904^[24] 与 iLL672^[25]、巴斯德毕赤酵母的 iPP668^[26] 与 iLC915^[27] 4 个高质量模型数据 (excel) 为同源比对模型。利用 BLAST 软件对本地数据库蛋白质序列文件进行双向 Blastp, 设置筛选 E 值为 $1e-40$ 。利用 JAVA 语言编写程序对信息量巨大的 word 结果文件进行筛选处理。同时, 基于登录号自对应的基因蛋白质列表中获得该序列对应的基因与蛋白质信息。依据比对上的基因, 进一步替换 iMM904、iLL672、iPP668 与 iLC915 4 个模型中的基因, 获得 4 个树干毕赤酵母基因组规模代谢网络初模型, 分别包含有 809、613、1 105、1 413 条反应。

1.4 初模型的比较和整合方法简述

上述 3 种自动化重构方法构建的树干毕赤酵母各个初模型, 除了包含 G-P-R 列表之外, 还包含一些构建精细模型的附加信息。不同初模型内容比较如表 1 所示。

表 1 自动化重构的不同模型比较

Table 1 Comparison of the different auto-reconstruction models

自动化方法	基因	蛋白质	反应	酶号	代谢途径	反应方向	区室划分	代谢物信息
KEGG Online Database	549	502	786	426	有	无	无	无
Uniprot-MetaCyc Database	800	755	582	479	无	有	无	无
Homologous alignment (four)	485/479/ 540/811	412/424/ 463/703	809/613/ 1105/1413	无/316/ 无/487	有	有	有	有

基于上述 3 种自动化方法构建的初模型中包含有大量重复反应信息,需对初模型进行整合获得格式统一、信息量充足、无重复反应的反应列表^[28]。整合过程中遇到的最大问题就是反应式中化合物格式不统一,一般表示为 4 种情况:亲缘物种替换的模型中同一种化合物以不同的简写形式表达;配位化合物由于配位单元顺序不同,造成了名称的不同;某些化合物为同一种物质,可能有不同的名称;化合物中出现一些符号,使得不能够识别为同一种化合物。由此,作者提出了基于化合物数据库和反应式字符频度直方图特征比对两种方法来达到模型反应自动化整合。

1.4.1 基于化合物数据库的模型整合 构建了整合 KEGG 和 MetaCyc 的化合物数据库。KEGG Compound 列表中包括每个化合物的 KEGG ID,即 C 号(字母 C+5 个数字)以及对应的化合物不同的名称。MetaCyc 数据库中也包含了每个化合物不同的表达形式,以及与 KEGG 数据库相对应的 C 号。两者中部分化合物有数据库特异性,所以整合两个数据库建立本地化合物数据库。

编写程序对整合模型中的每个化合物,建立对化合物数据库的映射关系。即查找替换对应的 C 号,然后比对 C 号形式的反应式中反应物与生成物的异同来确定是否为同一反应(反应式中化合物均为简写表达形式的模型数据,需先替换为代谢物列表中对应的全称形式)。

基于化合物数据库的模型整合并不能映射全部相同反应,原因可分为以下 3 种:相同反应因 H 或 H₂O 的缺失或冗余而不能判定为同一反应;某些配位化合物由于配位单元顺序不同,或者化合物中一些符号而造成了不能成功替换 C 号;反应方向不同造成反应物与产物识别错误。

1.4.2 基于字符频度直方图特征的模型整合 针对上述情况,作者提出了一种判断两个反应是否为相同反应的新方法。该方法通过提取反应式字符频

度直方图特征,进一步计算直方图间的马氏距离^[29]来实现。化合物的化学式核心由英文字母与阿拉伯数字组成,建立每个反应式的直方图,26 个英文字母与阿拉伯数字(0—9)为横坐标,36 个元素出现频次为纵坐标,计算直方图之间的距离(马氏距离),设置一定的阈值,当距离小于该阈值时表示为相同反应。

在进行判断之前,程序先对化合物区室标志如 [c]或[m]进行比较,因为同一反应可能存在不同的细胞区室,若区室相同依据反应式直方图特征进行马氏距离计算。在统计过程中,自动将大写字母转化为小写字母进行统计,且自动忽略标点符号或特殊字符(如,、-等)等。在计算之前还应移去无意义的干扰词。干扰词分为 3 类:不存在生物意义的词汇(如 alpha,beta 等);出现频度非常大的词汇(如 a、the 等);单词长度小于 2 的词汇。

马氏距离表示数据的协方差距离,它是一种有效的计算两个未知样本集的相似度的方法。设 $X\{X_1, X_2 \cdots X_n\}$ 和 $Y\{Y_1, Y_2 \cdots Y_n\}$ 为总体中抽取的样本,则 X, Y 两组样本之间的马氏距离为^[30]

$$D(X, Y) = \sqrt{(X_1 - Y_1)^2 + (X_2 - Y_2)^2 + \cdots + (X_n - Y_n)^2}$$

针对上述基于化合物数据库不能匹配的相同反应出现的原因,文中在计算马氏距离的时候采用模糊匹配值。当距离值相差 2.828 之内被认为是相同的,否则是不同的。当判断两个反应为相同反应后,程序接着判断该反应对应的基因是否相同,若不同,则将不同的基因保存于同一反应后,方便下一步进行人工判断取舍。

1.4.3 实施实例 初模型整合过程中,对相同反应的取舍依据各个初模型信息全面的高低:同源比对模型>KEGG 模型>MetaCyc 模型,但整合后的化合物格式统一以 KEGG 为标准,以同一 C 号对应不同的化合物表达形式的第一个为准。以树干毕赤酵母为例,6 个初模型基于化合物数据库整合后的模型包含 1 878 条反应以及对应的 956 个基因,经字符

频度直方图特征的模型整合方法补充,最终获得 1 531 条反应。经人工逐一检查,发现无重复反应,结果理想。结果模型中大部分反应来自基于同源比对

构建的模型,所以其中囊括了精细模型包含的基因、反应、酶、酶号、代谢途径、反应方向、亚细胞定位,以及代谢物的附加信息等内容。整合流程如图 4 所示。

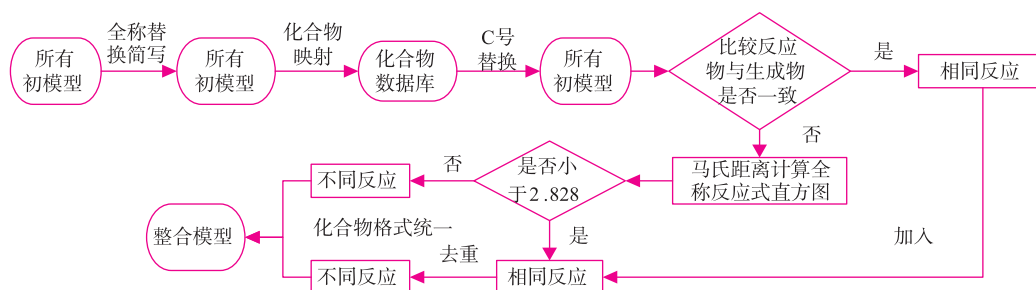


图 4 初模型自动整合流程

Fig. 4 Auto-integration process of the draft model

2 核心反应的识别

构建基因组规模代谢网络模型的核心代谢途径就是糖代谢,不管是脂肪酸还是蛋白质,或者是多糖,最终都要转化为糖类进行能量代谢与产物合成。EMP 糖酵解 (Glycolysis) 是糖代谢过程的第一步,存在于所有生物体内;TCA 三羧酸循环 (Tricarboxylic acid cycle) 是需氧生物体内普遍存在的代谢途径;PPP 磷酸戊糖途径 (Pentose phosphate pathway) 不是机体产能的方式,但生成具有重要生理功能的 NADPH 和 5-磷酸核糖。文中基于上述模型整合的两种方法,提出了一种基于基本核心代谢途径^[31] (EMP、TCA、PPP) 完善初模型的方法,该方法通过已有核心反应列表对模型数据的识别,可以快速确定模型是否包含完整的基本核心代谢途径。该方法适用于绝大多数生物体。核心反应识别流程如图 5 所示,基本核心代谢反应列表见表 2。

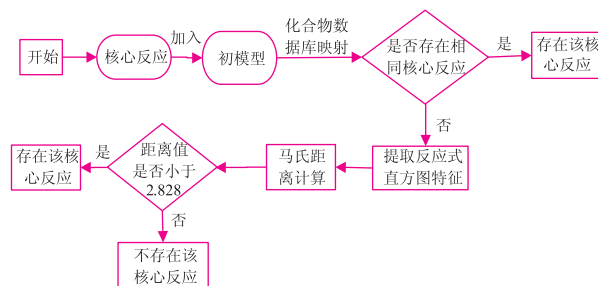


图 5 核心反应识别流程

Fig. 5 Process of Identification of core reaction

运用本方法对上述构建出的树干毕赤酵母代

谢网络模型进行核心反应识别,结果无核心断点,每条核心反应都可以在模型反应列表中找到相同反应。进一步人工查看发现无误,说明了该方法的有效性,同时也说明上述 3 种构建初模型并整合为相对精细模型方法的数据完善性。结果中,基于化合物数据库方法识别的核心反应数为 20 条,马氏距离计算反应式字符频度直方图补充识别了剩余 5 条反应,而这 5 条反应未被第一步识别的主要原因为反应式中出现了 H。

3 讨论

基于 KEGG 网页抓取数据在提高了模型构建效率的同时,保证了数据的最新性和全面性(包含反应式、酶、基因、代谢途径等),但提取的反应无方向性,全部表示为可逆。基于 Java 实现的 Uniprot-MetaCyc 模型构建的方法保证了程序的跨平台通用性;将两个数据库本地化也降低了远程 web 访问的时间。但构建出的模型出现一个基因对应多个反应的情况,需要进一步进行核对筛选,同时确定同工酶或聚合酶。基于同源菌株模型构建的代谢网络模型的真实性和信息更全面,包括代谢反应区室的划分,代谢途径的确定,化合物完整信息列表,方向的可逆性等。所以整合上述 3 种方法构建出的初模型,以同源比对结果模型为目标进行整合,当前两种反应列表中有相同反应时,以同源菌株模型构建的代谢网络模型为准,这样同时也确定了大部分反应与化合物的附加信息,减少了模型修正的大量工作。

基于化合物数据库与字符频度直方图特征两

表 2 基本核心代谢反应列表

Tabel 2 List of the basic core metabolic reactions

反应式	代谢途径
ATP + D-Glucose $\rightarrow$ ADP + D-Glucose-6-phosphate + H	EMP
D-Glucose-6-phosphate $\rightleftharpoons$ D-Fructose-6-phosphate	EMP
ATP + D-Fructose-6-phosphate $\rightarrow$ ADP + D-Fructose-1,6-bisphosphate + H+	EMP
D-Fructose-1,6-bisphosphate $\rightleftharpoons$ Glycerone phosphate + D-Glyceraldehyde-3-phosphate	EMP
D-Glyceraldehyde-3-phosphate $\rightleftharpoons$ Glycerone phosphate	EMP
D-Glyceraldehyde-3-phosphate + Orthophosphate + NAD+ $\rightleftharpoons$ 3-Phospho-D-glyceroyl phosphate + NADH + H+	EMP
ADP + 3-Phospho-D-glyceroyl phosphate $\rightleftharpoons$ ATP + 3-Phospho-D-glycerate	EMP
3-Phospho-D-glycerate $\rightleftharpoons$ 2-Phospho-D-glycerate	EMP
2-Phospho-D-glycerate $\rightleftharpoons$ Phosphoenolpyruvate + H ₂ O	EMP
ADP + Phosphoenolpyruvate + H+ $\rightarrow$ ATP + Pyruvate	EMP
CoA + NAD+ + Pyruvate $\rightarrow$ Acetyl-CoA + CO ₂ + NADH	TCA
Acetyl-CoA + H ₂ O + Oxaloacetate $\rightarrow$ Citrate + CoA + H+	TCA
Citrate $\rightleftharpoons$ Isocitrate	TCA
Isocitrate + NADP+ $\rightarrow$ 2-Oxoglutarate + CO ₂ + NADPH	TCA
2-Oxoglutarate + CoA + NAD+ $\rightarrow$ Succinyl-CoA + NADH + CO ₂	TCA
ATP + CoA + Succinate $\rightleftharpoons$ ADP + Orthophosphate + Succinyl-CoA	TCA
Succinate + FAD $\rightleftharpoons$ FADH ₂ + Fumarate	TCA
Fumarate + H ₂ O $\rightarrow$ (S)-Malate	TCA
(S)-Malate + NAD+ $\rightleftharpoons$ Oxaloacetate + NADH + H+	TCA
D-Xylulose-5-phosphate $\rightleftharpoons$ D-Ribulose-5-phosphate	PPP
ATP + D-Ribose-5-phosphate $\rightleftharpoons$ AMP + 5-Phospho-alpha-D-ribose-1-diphosphate + H+	PPP
D-Ribulose-5-phosphate $\rightleftharpoons$ D-Ribose-5-phosphate	PPP
Sedoheptulose-7-phosphate + D-Glyceraldehyde-3-phosphate $\rightleftharpoons$ D-Erythrose-4-phosphate + D-Fructose-6-phosphate	PPP
D-Ribose-5-phosphate + D-Xylulose-5-phosphate $\rightleftharpoons$ Sedoheptulose-7-phosphate + D-Glyceraldehyde-3-phosphate	PPP
D-Erythrose-4-phosphate + D-Xylulose-5-phosphate $\rightleftharpoons$ D-Fructose-6-phosphate + D-Glyceraldehyde-3-phosphate	PPP

种方法的整合,弥补了彼此整合过程中的缺憾,提高了反应式的匹配率。比如,前者不能判别电荷不平衡引起的相同反应,以及因分子式中特殊符号或分子机构顺序不同造成的同种化合物不能识别;后者通过计算每个反应式的字符频度直方图特征,设定一定的阈值,快速找到前者不能匹配的同种反应。反过来,前者能够识别某些只相差几个字母或数字的不同反应。比如 D-果糖-6-磷酸转化为 D-果糖-1,6-二磷酸,即第三条核心反应(ATP + D-Fructose-6-phosphate $\rightarrow$ ADP + D-Fructose-1,6-bisphosphate + H+),这条反应的缺失直接导致糖酵解途径中断,进而使得模型不能够生长。但模型中也包含 D-果糖-6-磷酸转化为 D-果糖-2,6-二磷酸的反应(ATP + D-Fructose-6-phosphate $\rightarrow$ ADP + D-Fructose-2,6-bisphosphate + H+),由于两个

反应只有一个数字的不同,基于反应式字符频度直方图特征的整合,两者被识别为相同反应。

整合后的树干毕赤酵母模型需要进一步精细化与修正,通过文献数据以及其他网络数据信息进行完善。基于 Matlab 平台的 COBRA 工具包函数,可以对其进行反应式电荷平衡检测、代谢漏洞查找与填补,以及后续的完整模型分析等,最终完成一个高质量的基因组规模代谢网络模型。

4 结 语

国内外虽然对代谢网络自动化重构也做了多方面的研究,但仍然不能够实现完全自动化构建代谢网络,代谢网络构建过程中仍然存在一些无法避免的人工操作与修正工作。作者提出的 3 种自动化重构代谢网络并整合为一个相对精细模型的方法,

以及依据化合物数据库映射和字符频度直方图特征整合初模型并识别核心反应的应用,能够在构建一个相对精细模型的过程中实现计算机技术与代

谢网络构建的最大化结合,减免了构建过程中大量的人力与时间,提高了代谢网络构建的效率及其精确性。

参考文献:

- [1] Veliz-Cuba A, Jarrah A S, Laubenbacher R. Polynomial algebra of discrete models in systems biology [J]. **Bioinformatics**, 2010, 26(13): 1637–1643.
- [2] Van Norman J M, Benfey P N. Arabidopsis thaliana as a model organism in systems biology [J]. **Wiley Interdisciplinary Reviews: Systems Biology and Medicine**, 2009, 1(3): 372–379.
- [3] Reed J L, Vo T D, Schilling C H, et al. An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR)[J]. **Genome Biol**, 2003, 4(9): R54.
- [4] Shlomi T, Eisenberg Y, Sharan R, et al. A genome-scale computational study of the interplay between transcriptional regulation and metabolism[J]. **Molecular Systems Biology**, 2007, 3(1): 1.
- [5] Hyduke D R, Palsson B Ø. Towards genome-scale signalling-network reconstructions [J]. **Nature Reviews Genetics**, 2010, 11(4): 297–307.
- [6] Thiele I, Jamshidi N, Fleming R M T, et al. Genome-scale reconstruction of *Escherichia coli*'s transcriptional and translational machinery: a knowledge base, its mathematical formulation, and its functional characterization [J]. **PLoS Computational Biology**, 2009, 5(3): e1000312.
- [7] Oberhardt M A, Palsson B O, Papin J A. Applications of genome-scale metabolic reconstructions [J]. **Mol Syst Biol**, 2009(5): 320.
- [8] Systemsbiology[EB/OL]. [2013-11-21]. <http://systemsbiology.ucsd.edu/InSilicoOrganisms/OtherOrganisms>
- [9] Genomes OnLine Database[EB/OL]. [2013-11-21]. <http://www.genomesonline.org/cgi-bin/GOLD/bin/gold.cgi>.
- [10] Thiele I, Palsson B O. A protocol for generating a high-quality genome-scale metabolic reconstruction [J]. **Nat Protoc**, 2010, 5(1): 93–121.
- [11] 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.
- [12] DeJongh M, Formsma K, Boillot P, et al. Toward the automated generation of genome-scale metabolic networks in the SEED[J]. **BMC Bioinformatics**, 2007, 8(1): 139.
- [13] Arakawa K, Yamada Y, Shinoda K, et al. GEM System: Automatic prototyping of cell-wide metabolic pathway models from genomes[J]. **BMC Bioinformatics**, 2006, 7(1): 168.
- [14] Agbogbo F K, Coward-Kelly G. Cellulosic ethanol production using the naturally occurring xylose-fermenting yeast, *Pichia stipitis* [J]. **Biotechnology Letters**, 2008, 30(9): 1515–1524.
- [15] Jeffries T W, Van Vleet J R H. *Pichia stipitis* genomics, transcriptomics, and gene clusters [J]. **FEMS Yeast Research**, 2009, 9(6): 793–807.
- [16] Zou W, Zhou M, Liu L, et al. Reconstruction and analysis of the industrial strain *Bacillus megaterium* WSH002 genome-scale in silico metabolic model[J]. **Journal of Biotechnology**, 2013: 1.
- [17] Dreyfuss J M, Zucker J D, Hood H M, et al. Reconstruction and validation of a genome-scale metabolic model for the filamentous fungus *Neurospora crassa* using FARM[J]. **PLoS Computational Biology**, 2013, 9(7): e10003126.
- [18] Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes[J]. **Nucleic Acids Research**, 2000, 28(1): 27–30.
- [19] Caspi R, Foerster H, Fulcher C A, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases[J]. **Nucleic Acids Research**, 2008, 36(suppl 1): D623–D631.
- [20] Caspi R, Altman T, Dale J M, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases[J]. **Nucleic Acids Research**, 2010, 38(suppl 1): D473–D479.
- [21] Bairoch A, Apweiler R, Wu C H, et al. The universal protein resource(UniProt)[J]. **Nucleic Acids Research**, 2005, 33(suppl 1): D154–D159.

- [22] Pertselmidis A, Fondon J W, John W. Having a BLAST with bioinformatics and avoiding BLASTphemy[J]. **Genome Biol**, 2001, 2(10):1.
- [23] Becker S A, Palsson B Ø. Genome-scale reconstruction of the metabolic network in *Staphylococcus aureus* N315: An initial draft to the two-dimensional annotation[J]. **BMC Microbiology**, 2005, 5(1):8.
- [24] Mo M L, Palsson B Ø, Herrgard M J. Connecting extra-cellular metabolomic measurements to intracellular flux states in yeast[J]. **BMC Systems Biology**, 2009, 3(1):37.
- [25] Kuepfer L, Sauer U, Blank L M. Metabolic functions of duplicate genes in *Saccharomyces cerevisiae* [J]. **Genome Research**, 2005, 15(10):1421-1430.
- [26] Chung B K S, Selvarasu S, Camattari A, et al. Research Genome-scale metabolic reconstruction and in silico analysis of methylotrophic yeast *Pichia pastoris* for strain improvement[J]. **Microbial Cell Factories**, 2010(9):50
- [27] Caspeta L, Shoaie S, Agren R, et al. Genome-scale metabolic reconstructions of *Pichia stipitis* and *Pichia pastoris* and in silico evaluation of their potentials[J]. **BMC Systems Biology**, 2012, 6(1):24.
- [28] Dreyfuss J M, Zucker J D, Hood H M, et al. Reconstruction and validation of a genome-scale metabolic model for the filamentous fungus *Neurospora crassa* using FARM[J]. **PLoS Computational Biology**, 2013, 9(7):e1003126.
- [29] Bedrick E J. Graphical modelling and the Mahalanobis distance[J]. **Journal of Applied Statistics**, 2005, 32(9):959-967.
- [30] Leuven E, Sianesi B. PSMATCH2; Stata module to perform full mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing[J]. **Statistical Software Components**, 2012;1.
- [31] Riemer S A, Rex R, Schomburg D. A metabolite-centric view on flux distributions in genome-scale metabolic models [J]. **BMC Systems Biology**, 2013, 7(1):33.

会议信息

会议名称(中文): 中国生物工程学会 2014 年学术年会

开始日期: 2014-11-07

结束日期: 2014-11-10

所在城市: 浙江省 温州市

主办单位: 中国生物工程学会

承办单位: 温州医科大学

议题: 生物技术与健康生活

摘要截稿日期: 2014-09-30

全文截稿日期: 2014-09-30

联系人: 常丽娟 15801343134

联系电话: 010-64807678

E-MAIL: xh@im.ac.cn

会议网站: <http://www.biotechchina.org/Notice/show/id/139>

会议名称(中文): 第五届媒介生物可持续控制国际论坛

开始日期: 2014-11-03

结束日期: 2014-11-05

所在城市: 山东省青岛市

主办单位: 中华预防医学会、中国疾病预防控制中心

承办单位: 中华预防医学会媒介生物学及控制分会、山东省疾病预防控制中心、青岛市疾控中心

联系人: 吴海霞 任东升

联系电话: 010-58900741

传真: 010-58900739

E-MAIL: meijieluntan@gmail.com, vectorforum.sina.com

会议网站: <http://www.chinavbc.cn/Item/list.asp?id=1540>