PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2014, 41(2): 105~114
www.pibb.ac.cn

全细胞网络重建与细胞工厂设计 *

徐自祥　郑　平　孙际宾 **
(中国科学院天津工业生物技术研究所，中国科学院系统微生物工程重点实验室，天津 300308)

摘要　在各种组学及其相应的网络研究相对成熟的基础上，集成各组学网络的细胞整合型网络或全细胞网络将大大提高对生物表型的预测能力，并成为代谢工程决策的有力武器．本文在阐述了细胞工厂设计中应该考虑细胞整合网络之后，综述了细胞整合网络的重建、分析、设计方法方面的有关问题，并进一步就研究细胞整合网络涉及的数据库、软件平台、并行计算几方面的作用作了介绍．

关键词　细胞整合网络，系统生物学，合成生物学，生物信息学，细胞工厂，代谢工程
学科分类号　Q5，Q6，Q7　　**DOI**: 10.3724/SP.J.1206.2012.00530

随着基因组高速测序技术的快速发展，越来越多的物种基因组测序已完成，转录组、蛋白质组、代谢组等各种组学(Omics)测定技术逐渐进入了成熟、普及阶段．在后基因组时代，理解基因组所含信息的任务越来越迫切，从系统层面上大规模测量生命活动中各种组分的技术也逐渐发展起来．超越了 20 世纪占主导地位的基因 - 蛋白质 - 功能与表型的还原论研究模式，以生命体整体及其内组元间动态相互关系为研究对象的系统生物学正成为当今生物学研究领域的重点之一．系统生物学现阶段主要集中于微生物领域，其主要研究内容包括微生物系统的建模、分析、控制(改造)．由于目前对微生物系统的建模主要是生物网络形式，所以系统生物学当前主要表现为对生物学网络的研究．多种模式微生物在全基因组规模上的不同形式的网络模型已经建立，其中主要以代谢网络模型为主，其他关于代谢物 - 蛋白质相互作用网络、蛋白质 - 蛋白质相互作用网络、信号网络、调控网络等方面的研究也相对成熟．要全面地解读生命的本质，除了对细胞各种组分进行全面、定量地测量之外，还要特别重视研究各种组分之间的相互作用及其时空动态关系．在各组学研究的生物网络相对成熟的基础上，集成各组学网络的细胞整合型网络或全细胞网络开始出现．整合细胞中的各种组分，共同构成一个超大规模的相互作用网络，我们称之为细胞整合网络(cell integral network)，核心是代谢和调控的结合．细胞整合网络(代谢和调控)可以包含基因转录调控、酶活性的抑制和激活、动力学数据、代谢反应等各种信息，是反映细胞内各种组分和相互作用关系的全细胞模型．这种整合分析将大大提高对生物表型的预测能力，并成为代谢工程决策的有力武器．在后基因组时代，完美地获取了各种层次的组学数据之后，能否在细胞整合网络的层面上进一步整合、分析各方面的数据，成为理解生命、乃至重新设计生命的关键．

1　细胞工厂设计中研究细胞整合网络的必要性

生物炼制是以生物可再生资源为原料生产能源

* 国家重点基础研究发展计划(973)(2011CBA00804)，国家自然科学基金(31070037, 31370829)，天津科技支撑计划项目(11ZCZDSY08400)和国家高技术研究发展计划(863)(2012AA023402)资助项目．
** 通讯联系人．
Tel: 022-84861949, E-mail: sun_jb@tib.cas.cn
收稿日期：2013-03-27，接受日期：2013-07-04

与化工产品的新型工业模式[1]. 随着能源、资源、环境问题的日趋严峻，生物炼制已经成为世界各国的战略研究方向[2]. 生物炼制主要有两种途径，一是热化学加工，二是生物化学转化即微生物发酵. 所谓细胞工厂，有两方面含义，其一是指生物炼制所用的微生物菌株具有工厂的要素如同生产线，其二是指微生物菌株具有可设计性. 细胞工厂的设计是生物炼制的核心内容[3]. 由于生物炼制的中心是利用可再生资源为原料来生产我们需要的生命物质，这发生在微生物的代谢层次，所以细胞工厂设计的主要目标是发现从目的底物到目标产物代谢层次上的最优途径.

由于细胞代谢系统的复杂性，细胞代谢系统表现为一个规模庞大的复杂网络. 图 1 是黑曲霉的代谢网络，含有 2 349 个代谢物和 2 443 个生化反应[4]. 为达到细胞工厂就某些目的底物高产目标产物的要求，单纯研究某几个反应或某个途径是不够的. 实现细胞工厂的设计目标是对反映代谢系统的代谢网络进行重新布局.

Fig. 1　Metabolic network of *A. niger*

图 1　黑曲霉代谢网络

2 349 个代谢物和 2 443 个生化反应.

另一方面，在实现细胞工厂的设计中，我们的操作手段却不是直接对代谢网络来进行，而是采用基因工程的遗传修饰方法[3]. 对微生物基因组进行基因的删除、添加等操作，再利用从基因到蛋白质、酶的确定关系，进而实现对催化反应酶的控制，最终期望达到改变代谢网络布局的目的.

然而，从基因工程的实现手段到代谢网络重新布局的目标，这之间需要利用的信息过程是复杂的，其中至少应考虑三方面. a. 基因 - 蛋白质 - 代谢复杂的映射关系. 从基因到蛋白质再到蛋白催化反应催化酶的表达，在蛋白质层次有可能会形成复合体，基因之间还有或(OR)、与(AND)、非(NOT)的逻辑关系[5]. 图 2 中，琥珀酸脱氢酶(succinate dehydrogenase)可以催化 SUCDli 和 SUCD4，其蛋白 Sdh 同时需要(AND 关系即 &) 4 种多肽 sdhC、sdhD、sdhA、sdhB，它们分别由 b0721、b0722、b0723、b0724 基因编码，中间情形 XYLabc 需要 XylF、XylG、XylH 共同构成复合体，GAPD 情形中，GapA 和 GapC 具有或的关系(OR)，即只需一者出现即可. b. 复杂基因之间的调控关系. 从基因到蛋白质，这一过程受到的调控是多方面的，转录调控是主要形式之一，一个基因的表达往往受一个或多个调控基因(转录因子)的合作调控，调控基因有时也受其他调控因子的调控，调控过程也有或(OR)、与(AND)、非(NOT)的逻辑关系. 图 3 所示酵母的调控网络[6]中 4 410 个靶基因受到 157 个调控基因的控制，其调控关系达 12 837 个之多. 例如基因 ADE1(ORF 为 YAR015W)受到的调控布尔规则为“if ((BAS1 and PHO2) or GCN4) then ADE1”[6]. c. 从代谢到基因的反馈调控. 细胞中，某些基因的表达会受一些代谢物存在与否及其浓度的影响，从而构成从代谢层到基因层的负向反馈调控. 例如，酵母基因 AAC3(ORF 为 YBR085W)的表达受乙醇(ethanol，etoh)和 L- 乳酸(L-lactate，lac-L)影响，其受到的调控布尔规则为“if (not ROX1 or etoh[e] or lac-L[e]) then AAC3”[6]. 在控制理论上，负反馈控制往往有利于系统的稳定性.

从数学的角度来看，在细胞处于稳态的情形下，代谢网络的表型空间，即代谢网络化学计量学矩阵 S 与代谢网络流通量向量 v 所构成线性方程组 $S \cdot v=0$ 的解空间，会受到代谢网络本身的限制，如反应的方向性、反应流通量的边界、热力学条件等方面. 调控方面的因素，通过影响催化反应的酶，进一步限制了代谢网络表型空间大小. 调控因素的加入，这样从计算角度获得的代谢网络表型的解，将更为符合细胞的实际，从而达到更为准确的目的.

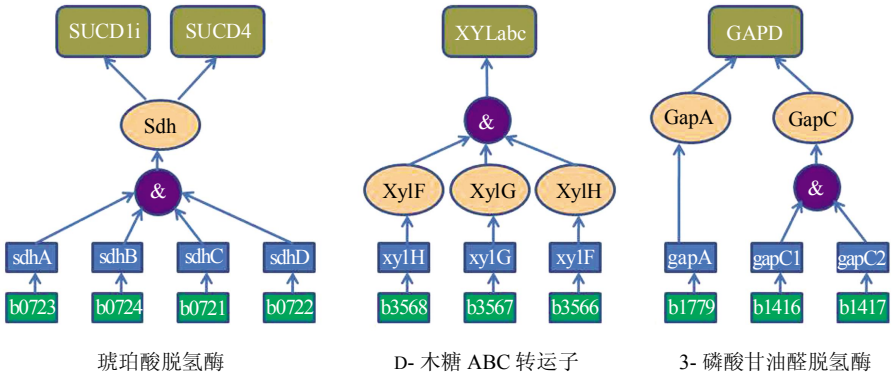

Fig. 2　Complex information flow in network in Yeast
图 2　基因-蛋白质-代谢的复杂信息流

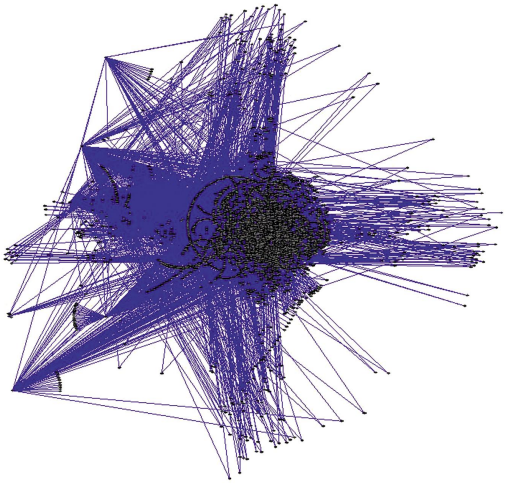

Fig. 3　Complex regulatory gene-protein-metabolism relationship
图 3　酵母的复杂调控网络

综合上述内容，作为生物炼制核心内容的细胞工厂设计，必须将细胞的代谢与调控两方面结合起来，才能更为客观和准确地预测细胞的行为，也才能提出有效的菌种改造方案. 图 4 给出了细胞代谢 / 调控整合网络的一个概念框架，其中代谢网络反映了细胞内代谢物之间相互作用和转化关系，其输入输出分别是培养媒介和生物质. 调控网络反映了基因与调控因子以及调控因子之间的相互作用. 从调控层次到代谢层次的正向作用通过蛋白质和催化反应的酶来实现，从代谢层次到调控层次的负向反馈作用是调控的本身内容之一.

2　细胞整合网络的建模与分析

随着基因组测序技术和功能基因组学的快速发展，目前多种模式生物在全基因组规模上的不同形式的网络模型已经建立，其中主要以代谢网络模型为主. 这些组学网络在建模方法上经历了三个主要阶段，早期是基于图论(graph theory)，目前最主要的是基于化学计量学(stoichiometrics)，未来的主要趋势是基于动力学(dynamics). 图论层次上，对各组学网络的研究主要是从复杂网络的角度研究其拓扑结构问题，包括节点的连结度、节点间平均距离、网络模块与功能关系等[7-8]，小世界网络和无标度网络是两个主要特点[9-10]. 计量学层次上，美国加州大学圣迭戈分校(UCSD)的 Palsson 研究组将线性优化引入代谢网络分析，提出了基于约束(constraint-based)的建模理念和 FBA(flux balance analysis)方法[11]，结合自动化和手工修饰[12]，该研究组在精致的基因组规模细胞代谢网络构建上做了很多贡献，发表了一系列模式生物的代谢网络[13]. 目前计量学上典型模式生物的基因组规模代谢网络已有 *A.niger*[4]、*E.coli*、*Chlamydomonas*、*M.barkeri*、*K. pneumoniae*、*G. metallireducens*、*B. subtilis* 等物种的代谢网络模型[14-19]. 计量学意义上基因组规模的转录调控网络还比较少，*E. coli*、*S. cerevisiae* 的基因调控网络模型[20-22]是主要代表. 其他关于代谢物 - 蛋白质相互作用网络、蛋白质 - 蛋白质相互作用网络、信号网络等方面的研究也有一些. 动力学

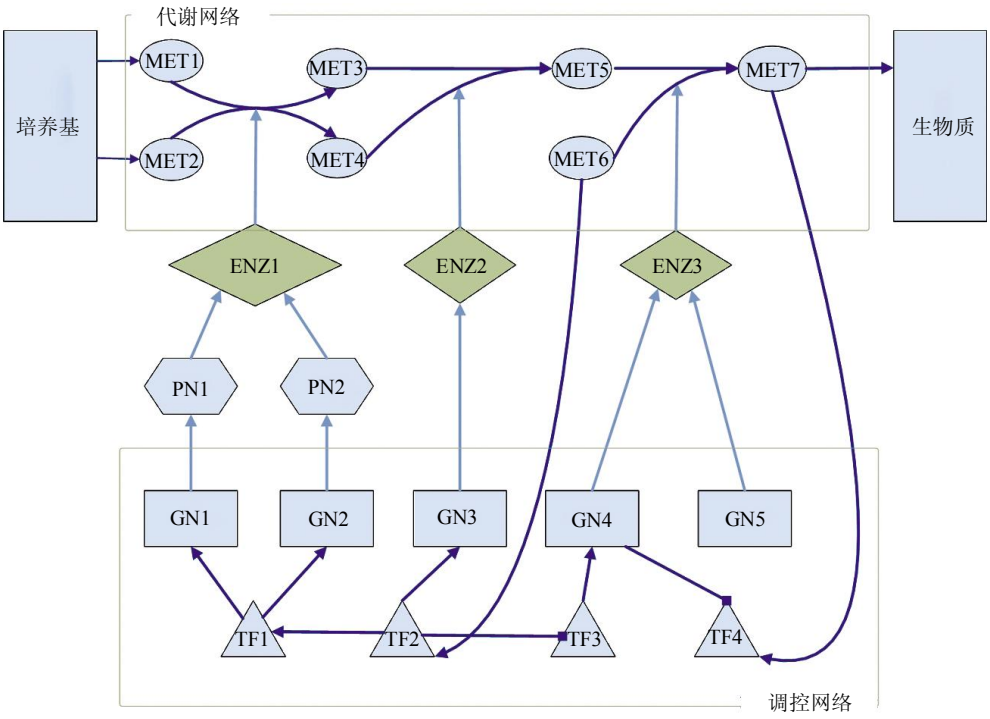

Fig. 4　A framework for metabolic/regulatory integral network

图 4　代谢/调控细胞整合网络的概念框架

MET: Metabolite; ENZ: Enzyme; PN: Protein; GN: Gene; TF: Transcription factor. 调控包括激活和阻遏.

层次上，由于细胞本身是动态的，所以本质上理想细胞模型应具有动力学描述这样的特性，计量学模型只能模拟细胞在稳态下的状态，难以考察细胞内的化学量随时间的变化情形．由于动力学数据的缺乏，任何形式的拥有数千个反应的细胞代谢动力学目前都无可行性．对 ODE (ordinary differential equation)形式细胞代谢动力学来说，方程形式没有统一的格式而且主要来自对文献的搜索获取，难度则更大．代谢系统的动力学方面，中央代谢和某些代谢途径的动力学得到很多关注，如大肠杆菌中央代谢的动力学模型[23]．细胞转录调控网络的动力学 ODE 模型，相比于代谢网络，非严格意义下可以有统一的方程形式．在大规模基因芯片(microarray)数据的支持下，有少数模式生物基因组规模的转录调控网络 ODE 模型发表，如 Richard Bonneau 等[24]建立了 *Halobacterium salinarum* NRC-1 的这类模型．

在各组学研究的生物网络相对成熟的基础上，集成各组学网络的细胞整合型网络开始出现．实际上在系统生物学诞生以前，以计算和仿真的方式从整体和动力学的角度研究细胞就已经开始，Virtual-CELL、E-CELL 是其中有名的代表[25-26]，但相对来说规模都很小且都还不够细致．基于整合网络的系统生物学研究，拥有底层生物学数据的支持，从而也就更为贴近细胞的真实情况．目前细胞整合网络的研究尚处于模型建立的初级阶段，整合的层面和网络的规模都十分有限，仅有少数几个模式生物的细胞整合网络模型有报道．UCSD 的 Palsson 研究组对 *S. cerevisiae* 和 *E. coli* 以布尔网络表示基因调控并整合了计量学意义上的代谢网络[27-28]；德国汉堡工业大学的 Zeng An-Ping 研究组在图论描述意义上就 *E. coli* 开展了整合基因调控、代谢、代谢物 - 蛋白质相互作用、蛋白质 - 蛋白质相互作用诸方面网络的研究[29]；本文作者的研究团队从计量学角度也正在尝试大肠杆菌细胞整合网络构建与分析、黑曲霉整合型细胞网络构建以及细胞整合网络等其他方面的研究工作．

虽然理想细胞模型应具有动力学特性，但大规模的联合组学实验数据是细胞动力学建模的根本动力．大规模、整合型的细胞动力学建模对组学实验

数据的要求，不仅在于数据规模的庞大，还需要时间序列的，尤其需要各组学联合实现时间上的同步．不是时间序列的，就谈不上动力学；时间同步的含义是，同一时间点上细胞有不同组学状态．大规模的联合组学实验，是一件困难的事情，需要先进的设备支持、成本也很昂贵．目前国际上这样的数据十分稀少．本文作者的研究团队对大肠杆菌构造过一系列的突变体如高产琥珀酸，并对最后高产菌株在代谢组学、蛋白质组学层次上进行了分析，分别获得了 6 个时间点上联合代谢组学和蛋白质组学的时间序列数据．

细胞网络分析方面，流量平衡分析(flux balance analysis，FBA)常被用于全基因组代谢网络的仿真，但 FBA 不包含基因调控的信息，且该算法假设系统处于拟稳态[30]．Palsson 研究组在对 *S. cerevisiae* 以布尔网络表达基因调控并整合代谢网络的基础上，以 rFBA(regulated flux balance analysis)方法对此整合网络进行了动态仿真的研究[27-28]．在 rFBA 的基础上，结合局部微分方程模型，Covert[31]和 Lee[32]分别整合了大肠杆菌和酵母的局部信号、代谢和调控网络，分别提出了 iFBA(integrated flux balance analysis)和 idFBA (integrated dynamic flux balance analysis)两种整合型细胞网络的模拟和仿真方法．rFBA、iFBA、idFBA 在描述调控过程上都是采用基于布尔逻辑规则的布尔网络模型，但布尔网络的 0-1 表达方式并不能很好地描述基因的调控状态，因为实际上基因的表达不能完全以是或非来描述，而应有不同表达水平的区别．细胞整合网络的分析方法总的来说尚处于起步阶段．

3 细胞整合网络的设计

随着人类基因组计划的完成，2000 年以后, 合成生物学一词在学术刊物及互联网上逐渐大量出现．合成生物学更为强调对生命系统的设计，研究内容主要有[33]：新的生物零件、组件和系统的设计与建造；对现有的、天然的生物系统的重新设计．近几年来，作为合成生物学技术基础的基因组测序技术及 DNA 合成技术正以指数增长速率发展，这正如大规模集成电路与计算机技术的发展一样，于是人们认为合成生物学将会像信息技术一样得到迅速发展，并将在能源、化学品、材料、疫苗等医药领域得到广泛应用，具有巨大的社会效益及经济效益．同时，在对人类认识生命、揭示生命的奥秘、重新设计及改造生命等方面具有重大的科学意义[33].

生物系统的改造和重新设计是系统生物学和合成生物学从认识自然走向改造自然的重要内容．目前微生物系统的改造主要是在代谢层面上基于基因删除、少量添加某些基因或过表达某些基因的代谢工程研究，尚未上升到结合调控的整体层面．这方面韩国的 Sang Yup Lee 研究组做了不少工作，如就生产苏氨酸、缬氨酸、琥珀酸对大肠杆菌进行的代谢工程改造[34-36]，以及对酵母调控网络的研究[22]. Jens Nielsen 研究组通过对酵母菌代谢网络的模拟提出了以代谢工程提高乙醇产量的策略[37]．由于细胞整合网络更真实地反映了细胞的实际情况，因而基于细胞整合网络的设计可以在生物工程实践中发挥更好的效果．

目前细胞整合网络的设计尚无统一和规范的方法可循．实际上，作为生物网络理论基础的异构网络理论(如 bipartite network)，在网络分析、可视化、设计等方面的研究还比较缺乏．已故著名学者钱学森先生曾指出系统科学和系统工程的技术基础可以是数学运筹学和控制理论[38]，所以作为一般方法论的数学优化技术和控制方法已经在生物系统的设计中崭露头角．细胞工厂的构建是实际生产中真正关心的问题，底物对产物的转化率和转化速度是两个核心．前者需要知道基因工程上的敲除靶点，基因敲除靶标预测是网络设计的这方面任务；后者需要知道基因工程上的强化或弱化靶点，获取产物合成途径能大大缩小基因操作的范围．

生物合成途径搜索，数学优化方法正是其基础．起先有基于图论和网络模型的代谢最短路径发现算法，如 k-shortest[39]．后来有为满足计量学物质平衡的途径发现算法，如 EFMs(基元模式)和 EPs (极端途径)，但这两种算法对大规模网络会存在组合爆炸问题[40]．CASOP 代表了一类基于代谢网络 elementary modes (EMs)的预测算法[41]，其优点在于结合了每个反应对产率的相对贡献和代谢网络容量，但正是由于全基因组代谢网络搜索其 EMs 会遇到计算上的组合爆炸问题，这限制了该方法对大规模网络的适用性．本文作者的研究团队开发了可避免这种组合爆炸同时又能满足计量学平衡的途径发现算法 FIND_SBPiLMN，并做成 web 服务[42]，对大肠杆菌全基因组代谢网络在有氧和无氧条件下从葡萄糖到琥珀酸的最短路径进行了计算，获得了较好的结果．本文作者的研究团队还拓展了从热力学考察代谢途径可行性的计算方法[43]，并利用到对 FBA 的多解选优问题[44].

作为重组 DNA 主要技术之一的基因敲除方法是用来提高菌株对某种产物转化率的主要手段．早期基因删除策略方案的制定主要依据对局部代谢途径的分析和实验的经验，随系统生物学和合成生物学的发展，利用细胞网络模型并结合不同的数学方法，代谢工程的遗传操作趋于理性化，代谢工程进入了系统代谢工程时代[45]．基因敲除靶标预测这类算法也有所发展，并有一些文献发表[46-51]，基因删除靶标的预测算法能大大减少实验的盲目性．为达到能从基因组规模进行预测，这些算法都以全基因组代谢网络作为细胞模型，FBA(本质是线性优化)是网络模拟的主要手段；另一方面，FBA 计算中为提高转化率的基因删除不能绕开细胞生长这一现实，所以这些算法都采用双层优化策略，如图 5 所示．上层是工业上的高产目标，以对某些操作是否实施(反应阻断、基因删除等)的 0-1 整型变量为控制变量；下层是以细胞生长最大化为目标的基因组规模上的 FBA．

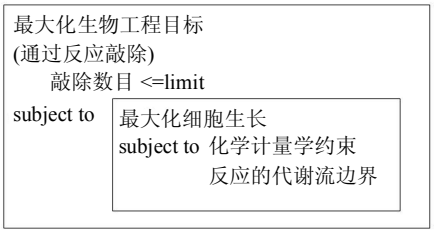

Fig. 5 Bi-level methdology of existed algorithms for the prediction of gene deletion

图 5 已有基因删除预测算法采用的双层优化思想

基因删除预测方法的主要代表有：a．OptKnock[46]是最先采用双层优化来解决工业上某产物产率最大化和 FBA 计算细胞生长最大化这两个目标的矛盾，求解方法是采用线性优化的对偶理论将双层优化问题转化为一个单层混合整数问题，从而利用已有优化软件进行求解．b．RobustKnock[47]比 OptKnock 更进一步之处是采用三层优化，最上层是一个 min-max 目标而下面两层与 OptKnock 等同，所以其主要思想是希望得到鲁棒性更好的解．c．美国 MIT 和哈佛医学院 Desmond 等的 GDLS 算法[48]与上述两种算法有所区别，它预测的不是要删除的反应(酶)，而是直接的酶基因，所依据的细胞模型是含 Gene-Protein-Reaction(GPR)映射关系的代谢网络，主要做法是删除这个酶涉及的任一基因就相当于删除该反应．d．其他：OptStrain[49]和 OptReg[50]在对融合非宿主反应、基因表达强度等方面扩展了 OptKnock 的功能；本研究团队开发了利用非线性优化来预测敲除靶标的算法，并能给出多个敲除方案[51]．e．本研究团队在参与中国科学院信息化专项项目“工业生物技术知识环境建设及其 e-Science 应用”[52]中，曾以遍历的方式，对酵母进行过多基因组合删除的计算，但遍历所有组合即使对低的组合数也需要相当长的计算时间．f．在细胞整合网络的基础上，我们正在开发一种能将代谢与调控结合起来，为提高目标产物产率，整体上有效预测靶基因删除的算法 RegKnock(target prediction for gene deletion based on integral metabolic- regulatory network，待发表)．

在对改造野生菌株为高产特定产物的工程菌中，从所构建的计量学静态网络模型，可以较易找到基因操作靶标以增加底物对产物的转化率．但就提高转化速度需要在动力学层次上解决不同反应的反应速度之间的协调问题，如催化不同反应的酶的浓度匹配，也即所谓发现基因工程上的强化或弱化靶点，这一直受到缺乏动态模型与方法的困惑，同时这也是目前很多工业菌株研发中遇到的普遍现象，所以基于动力学模型的细胞网络分析与设计方法是目前亟待解决的问题．MCA(metabolic control analysis)较早应用于考察代谢途径动力学中，某一步反应由于一个微小扰动，引起代谢系统中所指定的代谢通量或代谢物浓度的相对变化[7]，可以用来考察代谢系统内各反应节点对目标产物产出的灵敏度．Alvarez-Vasquez 等[53]以黑曲霉中央代谢动力学为基础，采用线性优化的策略，以目标产物产出速度为目标函数，约束内容包含酶浓度(constraints on enzyme concentrations)即各酶的许可浓度变化、总酶量(total enzyme concentration)即保证细胞活性的胞内酶总量、代谢物池(metabolite pools)即胞内代谢物总浓度，构建优化模型，最终得到匹配的系统中各酶浓度．

4 数据库和软件平台

4.1 数据库

研究细胞整合网络，无论是网络的构建还是对已建成网络的验证，或是对网络分析结果的考察，都离不开数据库的支持．目前与系统生物学相关的数据库已达到惊人的 260 多个(表 1)，涉及系统生

物学的各个环节，如代谢途径(metabolic pathways)、途径图表(pathway diagrams)、基因互作网络(genetic interaction networks)、信号网络(signaling pathways)、转录因子 / 基因(transcription factors / gene)、调控网络(regulatory networks)、蛋白质互作网络(protein-protein interactions)、蛋白序列相关(protein sequence focused)等.

Table 1 Quantitative distribution for the online databases related with systems biology

表 1 与系统生物学相关的在线数据库的数量分布

相关数据库	数量
代谢途径	59
途径图表	30
基因互作网络	6
信号网络	51
转录因子 / 基因	41
蛋白质互作网络	104
蛋白序列相关	16
其他	14
合计(无重复)	266

为更有效和方便地利用系统生物学数据库，本研究团队在全面了解了上述 260 个数据库(包括其名称、全名、类型、可获取性及费用要求、数据量的统计及统计时间)的基础上，本着这些数据库所提供数据的重要性、全面性、准确性、最新程度等方面原则，在中国科学院信息化专项的支持下，我们确定并整合了 40 多个重要的数据库[52].

4.2 软件平台

系统水平上各种生物测量技术的日益成熟，以及各国对系统生物学的高度关注和投入，使生物学数据的获取速度有了质的飞跃，自顶向下的数据驱动型研究方法得到了越来越多的支持. 而为了有效地分析和处理系统生物学研究中产生的大量数据，要将大量的数据转化为有用的知识，必须要有强大的生物学数据库的支持和生物信息学分析软件的帮助. 十余年的开拓，众多的生物信息学工具被开发出来，用于处理不同层面的数据，如基因组注释、比较基因组研究、时间序列的基因表达数据分析、调控网络的逆向工程(reverse engineering)、蛋白质组数据获取和评价、代谢网络重建和预测等. 可以说针对大多数环节，都有很多相应的软件可供选择，以辅助进行数据处理与分析. 目前涌现出一批比较贴近系统生物学的软件，如 Cytoscape、IdentCS、CellNetAnalyzer、SBtoolbox、VisANT、FBA、Fluxor、SimPheny、INSILICO Discovery、Metabologica、MicrobesFlux 等[54-60]，但它们仍主要用于分析代谢网络、进行代谢工程策略预测，特别是针对静态代谢网络. 尽管在每个环节上(如基因组分析、转录组分析、蛋白质组分析等)都有很多优秀的软件可供选用，目前还缺乏一些能够整合各种组学数据、真正从相互作用的细胞整合型网络水平上分析理解生命活动的软件工具.

在中国科学院信息化专项中，本研究团队全面了解了几百个当前系统生物学和生物信息学主要软件，包括它们的名称、功能(基因组、蛋白质组、转录组、代谢组等的分析；或是系统生物学各层面功能)、可获取性、费用要求等. 出于为细胞工厂的设计、开发、优化服务的原则，我们重点整合了工业生物技术相关的元基因组分析、基因组注释、转录组、蛋白质组和代谢组分析等方面的软件[52]. 同时初步开发了新型的面向细胞整合网络的数据整合、模型构建、可视化、分析、模拟、预测的软件平台[52]. 应用这一软件平台，系统生物学各种层次的数据将在细胞网络的框架下统一起来，得到有效系统的分析和可视化；系统性地把调控网络信息用于细胞工厂设计中，用于分析生物工艺动态过程，可以为大规模改造和控制细胞生理代谢、生物技术工艺过程优化服务.

4.3 大规模及并行计算

我们这里不去陈述高性能计算或并行计算技术本身的进展，主要关注其在细胞网络应用方面的影响. 从细胞网络构建角度看，所涉及的如数据库搜索[61-62]、基因组序列的拼接和注释[63]、大规模基因表达数据聚类[64-65]都可能用到高性能计算；从细胞网络的构建和分析角度看，高性能计算可能表现在如代谢结合调控的动态仿真研究，而采用如 Monte Carlo 方法等的随机模拟由于运算量大而用到并行计算最多[66]；从细胞网络的优化设计角度看，高性能计算可能表现在如基因组合删除研究、代谢途径发现. 在中国科学院信息化专项中，本文作者就酵母利用纤维素来高产乙醇这一具体问题，在酵母细胞网络模型的基础上，对酵母 600 多个非致死基因进行了大规模的基因组合删除研究[52]，k 个基因的组合数目达到 C_{600}^{k}，当然 k 是一个适当大小的数字，同时进行了酵母在多种培养基组合环境下细胞内代谢网络和基因表达状态的动态仿真研究[52]. 对

此类大规模的超算问题，高性能的计算条件和并行化的计算方法是解决问题的有效手段．在我们所开发的满足计量学平衡的途径发现算法 FIND_SBPiLMN 中，也使用了并行计算技术，大大提高了计算速度，节约了计算时间[42]．以上所述都会涉及大规模的数据分析、数据处理、方程组求解等，高性能计算是必不可少的技术支持．一些大型涉及生物计算的科研机构也正关注高性能计算对它们的促进作用，国际权威期刊《自然》(*Nature*)杂志上的新闻“Genome giant offers data service”报道，全球最大的基因组测序机构之一的华大基因正展望于云计算[67].

5 小结与展望

细胞整合网络是当前日益强调的系统生物学整合思想的具体体现，细胞整合网络的研究对生物学理论和生物工程的实践都将产生很大的推动作用．系统生物学结合了科学研究之数学、实验、计算三架马车，数学方法在细胞整合网络的研究中将发挥重要作用，正如马克思曾经指出：“一种科学只有在成功地运用数学方法时，才算达到真正完善的地步．”(注：《马克思恩格斯全集》，人民出版社 1962 年版，第 13 卷第 42 页)．数学模型作为实验数据和计算模型的纽带，一方面用于归纳湿法的实验数据，另一方面能用于解释干法的计算结果.

我们将积极投身于细胞整合网络的研究工作，同时也热情期待细胞整合网络在建模、分析、设计等各方面的优秀成果.

参 考 文 献

[1] Ragauskas A J. The path forward for biofuels and biomaterials. Science, 2006, **311**(5760): 484−489

[2] Kamm B, Kamm M. Principles of biorefineries. App Microbio Biotechno, 2004, **64**(2): 137−145

[3] 张延平, 李 寅, 马延和. 细胞工厂与生物炼制. 化学进展, 2007, **19**(7): 1076−1083
Zhang Y P, Li Y, Ma Y H. Prog Chem, 2007, **19**(7): 1076−1083

[4] Sun J, Lu X, Rinas U, *et al*. Metabolic pecularities of Aspergillus niger disclosed by comparative metabolic genomics. Genome Biology, 2007, **8**(9): R182

[5] Bernhard O. Palsson. Systems Biology: Properties of Reconstructed Networks. UK: Cambridge University Press, 2006: 40−42

[6] Balaji S, Babu M M, Lyer L M, *et al*. Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of yeast. J Mol Biol, 2006, **360**(1): 213−227

[7] 徐自祥, 孙 啸. 细胞代谢复杂网络研究进展. 生物信息学, 2009, **7**(2): 120−124
Xu Z X, Sun X. China J Bioinformatics, 2009, **7**(2): 120−124

[8] Ma H W, Zeng A P. Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. Bioinformatics, 2003, **19**(2): 270−277

[9] Barabasi A L, Albert R. Emergence of scaling in random networks. Science, 1999, **286**(5439): 509−512

[10] Jeong H, Tombor B, Albert R, *et al*. The large-scale organization of metabolic networks. Nature, 2000, **407**(6804): 651−654

[11] Francke C, Siezen R J, Teusink B. Reconstructing the metabolic network of a bacterium from its genome. Trends in Microbiology, 2005, **13** (11): 550−558

[12] Schellenberger J, Que R, Fleming R M T, *et al*. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. Nat Protoc, 2011, **6**(9): 1290−1307

[13] Schellenberger J, Park J O, Conrad T M, *et al*. BiGG: a biochemical genetic and genomic knowledgebase of large scale metabolic reconstructions. BMC Bioinformatics, 2010, **29**(11): 213

[14] Jeffrey D Orth, Tom M Conrad, Jessica Na, *et al*. A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism-2011. Mol Syst Biol, 2011, **7**: 535

[15] Chang R L, Ghamsari L, Manichaikul A, *et al*. Metabolic network reconstruction of Chlamydomonas offers insight into light-driven algal metabolism. Mol Syst Biol, 2011, **7**: 518

[16] Adam M Feist, Johannes C M Scholten, Palsson B Ø, *et al*. Modeling methanogenesis with a genome-scale metabolic reconstruction of *Methanosarcina barkeri*. Mol Syst Biol, 2006, **2**: 2006.0004

[17] Liao Y C, Huang T W, Chen F C, *et al*. An experimentally validated genome-scale metabolic reconstruction of Klebsiella pneumoniae MGH 78578, iYL1228. J Bacteriol, 2011, **193**(7):1710−1717

[18] Sun J, Sayyar B, Butler J E, *et al*. Genome-scale constraint-based modeling of Geobacter metallireducens. BMC Systems Biology, 2009, **3**: 15

[19] Oh Y K, Palsson B O, Park S M, *et al*. Genome-scale reconstruction of metabolic network in Bacillus subtilis based on high-throughput phenotyping and gene essentiality data. J Biol Chem, 2007, **282** (39): 28791−28799

[20] Gama-Castro S, Jiménez-Jacinto V, Peralta-Gil M, *et al*. RegulonDB (version 6.0) gene regulation model of *Escherichia coli* K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. Nucl Acid Res, 2008, **36** (Database issue): D120−124

[21] Salgado H, Santos-Zavaleta A, Gama-Castro S, *et al*. The comprehensive updated regulatory network of *Escherichia coli* K-12. BMC Bioinformatics, 2006, **7**: 5

[22] Kopp A, McIntyre L M. Transcriptional network structure has little effect on the rate of regulatory evolution in yeast. Mol Biol Evol, 2012, **29** (8): 1899−1905

[23] Kadir T A, Mannan A A, Kierzek A M, *et al*. Modeling and simulation of the main metabolism in *Escherichia coli* and its several single-gene knockout mutants with experimental verification. Microbial Cell Factories, 2010, **9**: 88

[24] Richard Bonneau. A predictive model for transcriptional control of physiology in a free living cell. Cell, 2007, **131**(7): 1354−1365

[25] Resasco D C, Gao F, Morgan F, *et al*. Virtual Cell: computational tools for modeling in cell biology. WIREs Syst Biol Med, 2012, **4**(2): 129−140

[26] Dhar P K, Takahashi K, Nakayama Y, *et al*. E-Cell: Computer Simulation of The Cell//Meyers R A. Systems Biology. German: Wiley-VCH, 2012

[27] Covertw M W, Palsson B Ø. Constraints-based models: regulation of gene expression reduces the steady-state solution space. J Theor Biol, 2003, **221**(3): 309−325

[28] Covert M W, Palsson B Ø. Transcriptional regulation in constraints-based metabolic models of *Escherichia coli*. J Biol Chem, 2002, **277**(31): 28058−28064

[29] Kumar B, Ma H, Zeng A P. An integrated cellular network of *Escherichia coli* and its structural analysis//University of California. Proceedings of Foundation of Systems Biology in Engineering. USA: University of California, 2005: 107−110

[30] Jeffrey D Orth, Ines Thiele, Palsson B Ø. What is flux balance analysis?. Nature Biotechnology, 2010, **28**(3): 245−248

[31] Covert M W, Xiao N, Chen T J, *et al*. Integrating metabolic, transcriptional regulatory and signal transduction models in *Escherichia coli*. Bioinformatics, 2008, **24**(18): 2044−2050

[32] Lee J M, Gianchandani E P, Eddy J A, *et al*. Dynamic analysis of integrated signaling, metabolic, and regulatory networks. PLoS Computational Biology, 2008, **4**(5): e1000086

[33] Keasling J D. Synthetic biology and the development of tools for metabolic engineering. Metabolic Engineering, 2012, **14**(3): 189−195

[34] Lee K H, Park J H, Kim T Y, *et al*. Systems metabolic engineering of Escherichia coli for L-threonine production. Molecular Systems Biology, 2007, **3**(149): 1−8

[35] Park J H, Lee K H, Kim T Y, *et al*. Metabolic engineering of *Escherichia coli* for the production of L-valine based on transcriptome analysis and in silico gene knockout simulation. Proc Natl Acad Sci USA, 2007, **104** (19): 7797−7802

[36] Lee S J, Lee D Y, Kim T Y, *et al*. Metabolic engineering of *Escherichia coli* for the enhanced production of succinic acid based on genome comparison and in silico gene knock-out simulation. Appl Environ Microbiol, 2005, **71**(12): 7880−7887

[37] Bro C, Regenberg B, Forster J, *et al*. In silico aided metabolic engineering of *Saccharomyces cerevisiae* for improved bioethanol production. Metab Eng, 2006, **8**(2): 102−111

[38] 钱学森. 论系统工程(新世纪版). 上海: 上海交通大学出版社, 2007: 141−144
Qian Xue-Sen. On systems engineering. Shanghai: Shanghai University Press. 2007: 141−144

[39] Xia D G, Zheng H, Liu Z, *et al*. MRSD: a web server for metabolic route search and design. Bioinformatics, 2011, **27**(11): 1581−1582

[40] Planes F J, Beasley J E. Path finding approaches and metabolic pathways. Discrete Applied Mathematics, 2009, **157** (10): 2244−2256

[41] Hadicke O, Klamt S. CASOP: a computational approach for strain optimization aiming at high productivity. J Biotechnology, 2010, **147**(2): 88−101

[42] FIND_SBPiLMN(Find Smallest Balanced Pathway in Large-scale Metabolic Networks)的 Web 服务. http://124.16.173.8/

[43] Xu Z X, Sun X, Sun J B. Construction and analysis of the weighted-network model of energy metabolism in *Escherichia coli*. PLoS ONE, 2013, **8**(1): e55137

[44] Zhu Y, Song J N, Xu Z X, *et al*. Development of thermodynamic optimum searching (TOS) to improve the prediction accuracy of flux balance analysis. Biotechnol Bioeng, 2013, **110**(3): 914−923

[45] Blazeck J, Alper H. Systems metabolic engineering: genome-scale models and beyond. Biotechnology J, 2010, **5**(7): 647−659

[46] Burgard A P, Pharkya P, Maranas C D. OptKnock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. Biotech Bioeng, 2003, **84** (6): 647−657

[47] Tepper N, Shlomi T. Predicting metabolic engineering knockout strategies for chemical production: accounting for competing pathways. Bioinformatics, 2010, **26**(4): 536−543

[48] Lun D S, Rockwell G, Guido N J. Large-scale identification of genetic design strategies using local search. Molecular Systems Biology, 2009, **5**: 296

[49] Pharkya P, Maranas C D. An optimization framework for identifying reaction activation/inhibition or elimination candidates for overproduction in microbial systems. Metab Eng, 2006, **8** (1): 1−13

[50] Pharkya P, Burgard A P, Maranas C D. OptStrain: a computational framework for redesign of microbial production systems. Genome Res, 2004, **14**(11): 2367−2376

[51] Maria G, Xu Z X, Sun J B. In-silico search for optimal fluxes and theoretical gene knockout strategies for *E. coli*. Chem Biochem Eng Q, 2011, **25** (4): 403−424

[52] 中国科学院信息化专项“工业生物技术知识环境建设及其 e-Science 应用”. http://dct.bioindustry.cn/dct/

[53] Alvarez-Vasquez F, Gonzalez-Alcon C, Torres N V. Metabolism of citric acid production by *Aspergillus niger*: Model definition, steady-state analysis and constrained optimization of citric acid production rate. Biotech Bioeng, 2000, **70**(1): 82−108

[54] Smoot M E, Ruscheinski J, Wang P L, *et al*. Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics, 2011, **27**(3): 431−432

[55] Sun J, Zeng A P. IdentiCS - identification of coding sequence and in silico reconstruction of the metabolic network directly from unannotated low-coverage bacterial genome sequence. BMC Bioinformatics, 2004, **5**: 112

[56] Steffen Klamta, Axel von Kamp. An application programming interface for CellNetAnalyzer. Biosystems, 2011, **105**(2): 162−168

[57] Schmidt H, Jirstrand M. Systems biology toolbox for MATLAB: a computational platform for research in systems biology. Bioinformatics, 2006, **22**(4): 514−515

[58] Hu Z, Wang Y, Chang Y C, *et al*. VisANT 3.5: multi-scale network

visualization, analysis and inference based on the gene ontology. Nucl Acid Res, 2009, **37**(Web Server Issue): W115-W121

[59] SimPheny™: http://www.genomatica.com/

[60] Feng X, Xu Y, Chen Y, *et al*. MicrobesFlux: a web platform for drafting metabolic models from the KEGG database. BMC Syst Biol, 2012, **2**: 94

[61] Bjornson R, Sherman A, Weston S, *et al*. Turboblast: a parallel implementation of blast built on the turbohub//IEEE International. Parallel and Distributed Processing Symposium. 2002

[62] Camp N, Cofer H, Gomperts R. High-throughput blast. SGI white paper, available at http://www.sgi.com (1998)

[63] Aluru S, Bader D A, Kalyanaraman A. High-performance computing methods for computational genomics//IEEE International. Parallel and Distributed Processing Symposium, 2007

[64] de Hoon M, Imoto S, Nolan J, *et al*. Open source clustering software. Bioinformatics, 2004, **20**(9): 1453-1454

[65] DU Z, Lin F. A hierarchical clustering algorithm for mimd architecture. Comput Biol Chem, 2004, **28**(5-6):417-419

[66] Li H, Petzold L. Efficient parallelization of the stochastic simulation algorithm for chemically reacting systems on the graphics processing unit. Int J High Performance Comp Appl, 2010, **24**(2): 107-116

[67] Callaway E. Genome giant offers data service. Nature, 2011, **475**: 435-437

Reconstruction of Whole Cell Network and Design of Cell Factory*

XU Zi-Xiang, ZHENG Ping, SUN Ji-Bin**

(*Key Laboratory of Systems Microbial Biotechnology, Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, Tianjin* 300308, China)

Abstract Based on the maturity of the researches of various kinds of Omics networks, integrated cell network which integrates Omics networks will greatly improve the prediction ability of biological phenotype, and will become a powerful weapon of metabolic engineering decisions. After expounding that integrated cell network should be considered in the design of cell factory, we gave a review about reconstruction, analysis, and design methods of integrated cell network. We also introduced several aspects of databases, software platforms, parallel computing which involved in the research of integrated cell network.

Key words cell integral network, systems biology, synthetic biology, bioinformatics, cell factory, metabolic engineering

DOI: 10.3724/SP.J.1206.2012.00530

* This work was supported by grants from National Basic Research Program of China (2011CBA00804), The National Natural Science Foundation of China (31070037, 31370829), Tianjin Research Program of Application Foundation and Advanced Technology (11ZCZDSY08400) and The National Hi-Tech Research and Development Program (2012AA023402).

**Corresponding author.

Tel: 86-22-84861949, E-mail: sun_jb@tib.cas.cn

Received: March 27, 2013 Accepted: July 4, 2013