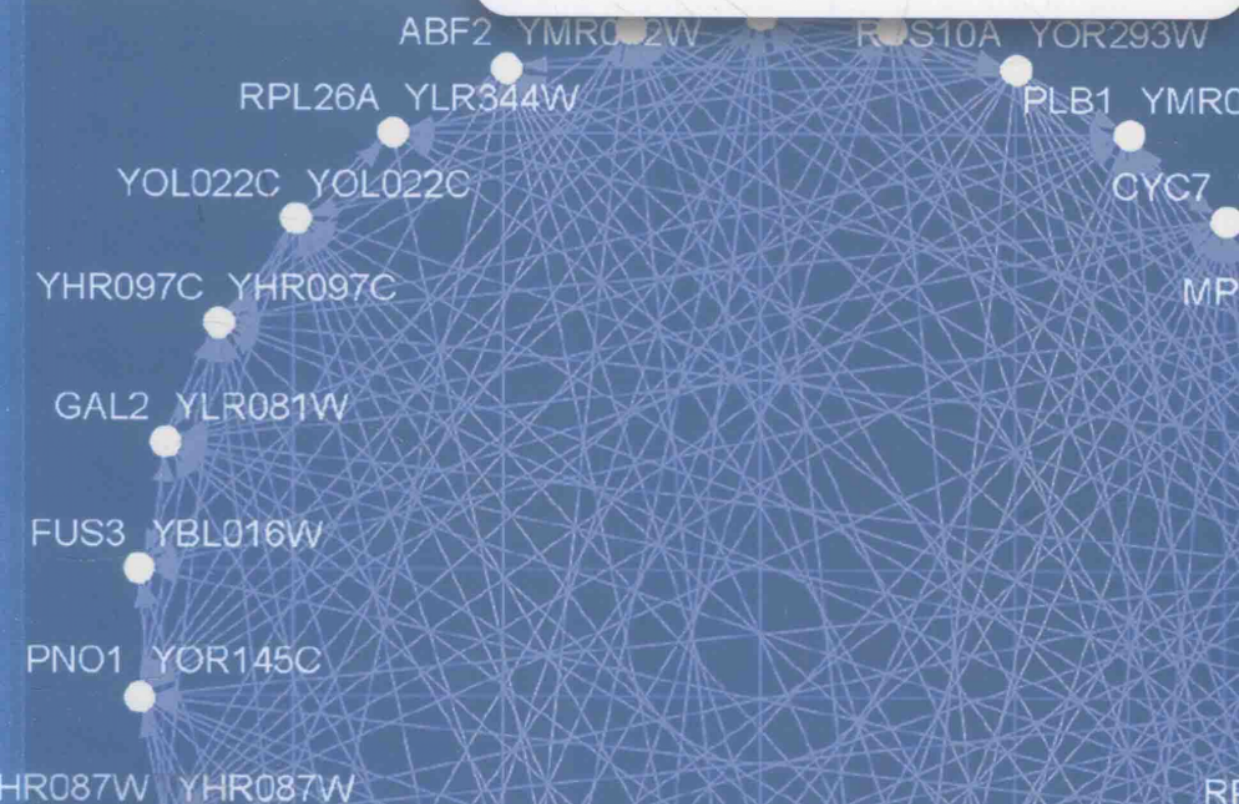


系统生物学基础: 如何使用Cell Illustrator 和通路数据库

システム生物学がわかる! —セルイラストレータを使ってみよう—

Foundations of Systems Biology: Using Cell Illustrator and Pathway Databases

(日) 土井淳 长崎正朗 齐藤步 著 李 晨 译
松野浩嗣 宫野悟



YDL226C



ZHEJIANG UNIVERSITY PRESS
浙江大学出版社

PDE1 YGL248W

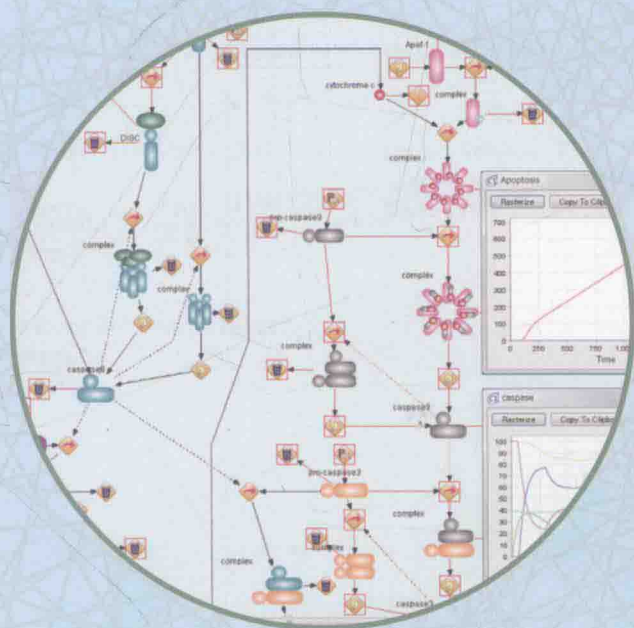
RPS9B

系统生物学基础: 如何使用Cell Illustrator 和通路数据库

システム生物学がわかる! -セルイラストレータを使ってみよう-

Foundations of Systems Biology: Using Cell Illustrator and Pathway Databases

(日) 土井淳 长崎正朗 齐藤步 松野浩嗣 宫野悟 著
李 晨 译



共立出版社



ZHEJIANG UNIVERSITY PRESS
浙江大学出版社

图书在版编目(CIP)数据

系统生物学基础：如何使用Cell Illustrator和通路数据库 / (日) 土井淳等著；李晨译. —杭州：浙江大学出版社，2016.11

书名原文：Foundations of Systems Biology: Using Cell Illustrator and Pathway Databases
ISBN 978-7-308-15520-5

I. ①系… II. ①土… ②李… III. ①生物学—系统科学—应用软件 IV. ①Q111-39

中国版本图书馆CIP数据核字(2016)第002125号

系统生物学基础——如何使用Cell Illustrator和通路数据库

(日) 土井淳 长崎正朗 齐藤步 松野浩嗣 宫野悟 著

李 晨 译

责任编辑 阮海潮

责任校对 潘晶晶 秦 瑕

封面设计 杭州林智广告有限公司

出版发行 浙江大学出版社

(杭州市天目山路148号 邮政编码 310007)

(网址: <http://www.zjupress.com>)

排 版 杭州林智广告有限公司

印 刷 杭州日报报业集团盛元印务有限公司

开 本 787mm×1092mm 1/16

印 张 9

字 数 208千

版 次 2016年11月第1版 2016年11月第1次印刷

书 号 ISBN 978-7-308-15520-5

定 价 85.00元

版权所有 翻印必究 印装差错 负责调换

浙江大学出版社发行中心联系方式: 0571-88925591; <http://zjdxbs.tmall.com>

浙江省版权局著作权合同登记图字：11-2015-185号

© 2007 Atsushi Doi, Masao Nagasaki, Ayumu Saito, Hiroshi Matsuno, Satoru Miyano
Shisutemu Seibutugaku ga waku! Seruirasutore-ta wo tsukatte miyou

ISBN: 978-4-320-05658-9 was originally published in Japanese language by KYORITSU SHUPPAN CO., LTD. in 2007.

This translation is published by arrangement with KYORITSU SHUPPAN CO., LTD., Tokyo, Japan.

All rights reserved. No part of this book may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording or by any information storage and retrieval system, without permission in writing from KYORITSU SHUPPAN CO., LTD.

Cell Illustrator is the property of Tokyo University and is distributed worldwide by BIOBASE GmbH.

TRANSPATH is a registered trademark of BIOBASE GmbH, Halchtersche Strasse 33, Wolfenbüttel 38304 Germany.

British Library Cataloguing in Publication Data.

A catalogue record for this book is available from the British Library
Library of Congress Control Number: 2009922124

译者序

近些年来,系统生物学(Systems Biology)这个学科已经成为国内除生物信息学以外的另一个重要的利用计算机进行分析和解读生物有机体奥秘的研究热点。如何从系统水平,从一个复杂生物学网络的水平全面地研究各分子及其之间存在的相互作用成为“后基因组”生物学研究的关键目标。

Cell Illustrator是一个可以用于构建及模拟复杂生物学网络动态变化信息的软件。最新的5.0版本,其源代码包括4000个class及200万行程序,具有模型构建、仿真、网络可视化、最优布局、高性能并行计算、数据库在线输入、多语言输出等功能。Cell Illustrator项目的开发始于1998年,最开始是由日本山口大学理工学部的松野浩嗣(Hiroshi Matsuno)教授与东京大学宫野悟(Satoru Miyano)教授团队共同开发,最初的名字叫作Genomic Object Net(GON),后来为了让生物学实验工作者也能更直接地接受,将名称更改为Cell Illustrator。

我从2004年开始读博士学位时便加入了开发团队。Cell Illustrator设计的初衷是希望生物学者和医学工作者通过简单拖动鼠标,就可以直观地构建通路或网络模型,然后根据计算机的“虚拟实验”,观察各分子浓度的动力学变化,预测实验结果。Cell Illustrator的计算体系是基于一个叫Petri网理论的数学语言,为了减少科研工作者们对于数学和统计学知识的陌生感,在开发过程中,我们结合生物学家和基础医学科学家们的意见,不断地修改操作界面和功能。用户友好的界面为科学家们获得创新性成果提供了良好的平台。

2007年,团队开发人员出版了日文版的利用Cell Illustrator进行系统生物学研究的教材,后翻译成英文出版。同团队的韩国籍研究人员也将其翻译成韩文出版。当时我就萌生了将其翻译成中文的念头,同时,我也发现国内还没有类似本书内容的参考书,但一直到2012年7月回国后才开始着手翻译,至今才得以出版,希望不会太晚。

借此机会感谢日本东京大学宫野悟教授、日本山口大学松野浩嗣教授、葛崎伟教授和日本东北大学长崎正朗教授在我做博士后和读博士学位期间的指导,他们不仅在专业上给予我多方面的指导,而且在学术研究等其他方面给予我许多的训练机会。

感谢我在浙江大学基础医学院的PI——Dante Neculai教授对我工作的长期支持。感谢我的研究生吴谦、李燕妮和应羽燚同学在翻译过程中的帮助。本书的出版得到了国家自然科学基金青年科学基金项目“in silico生物分子网络动力学参数高速与高精度自动化估计的研究”(31301100)、教育部高等学校博士学科点专项科研基金(新教师类)“基于高阶Petri网的定量生物学通路可视化模型调试验证技术的研究”(20130101120005)和浙江省自然科学基金(LY17C060003, LZ16C050001)的资助。本书在翻译过程中,还得到了我在东京大学的前同事韩国淑明女子大学郑恩娥博士的关怀和帮助,在此也一并表示感谢。

简易版Cell Illustrator操作软件可在Windows、Linux和Mac OS X上使用,该软件在作者提供的软件包里,可向本书译者索取,邮箱地址为chenli2012@zju.edu.cn。

由于能力有限,直至完稿,我仍觉得有许多不足之处,希望学术同仁不吝赐教,以便改正。

李 晨

2016年9月21日

目 录

第1章 绪 论

1.1 胞内活动	1
1.1.1 转录、翻译和调控	1
1.1.2 信号传导路径和蛋白质	2
1.1.3 代谢作用和基因	2
1.2 胞内反应和通路	3

第2章 通路数据库

2.1 主要的通路数据库	4
2.1.1 KEGG	5
2.1.2 BioCyc	6
2.1.3 Ingenuity Pathways Knowledge Base	6
2.1.4 TRANSPATH	7
2.1.5 ResNet	9
2.1.6 Signal Transduction Knowledge Environment (STKE): 细胞信号的数据库	9
2.1.7 Reactome	10
2.1.8 Metabolome.jp	10
2.1.9 归纳与总结	10
2.2 通路的显示软件	11
2.2.1 Ingenuity Pathway Analysis (IPA)	11
2.2.2 Pathway Builder	11
2.2.3 Pathway Studio	11
2.2.4 Connections Maps	12
2.2.5 Cytoscape	12
2.3 通路的文件格式	13
2.3.1 基因本体 (Gene Ontology, GO)	13

2.3.2	PSI MI	13
2.3.3	CellML	13
2.3.4	SBML	13
2.3.5	BioPAX	14
2.3.6	CSML/CSO	14

第3章 通路仿真软件

3.1	仿真软件后台	16
3.1.1	架构：确定性，概率性，还是混合的？	16
3.1.2	通路建模的方法	17
3.2	主要的仿真软件工具	17
3.2.1	Gepasi/COPASI	17
3.2.2	Virtual Cell	18
3.2.3	Systems Biology Workbench (SBW), Cell Designer, JDesigner	18
3.2.4	Dizzy	18
3.2.5	E-Cell	18
3.2.6	Cell Illustrator	18
3.2.7	总结	20

第4章 开始使用Cell Illustrator

4.1	安装Cell Illustrator	21
4.1.1	操作系统和硬件要求	21
4.1.2	Cell Illustrator的版本介绍	22
4.1.3	安装和运行Cell Illustrator	22
4.1.4	证书的安装	23
4.2	Cell Illustrator的基本概念	23
4.2.1	基本概念	23
4.2.2	实体	24
4.2.3	过程	27
4.2.4	连接	28
4.2.5	元素之间的连接规则	29

4.2.6 带有图标的元素	30
4.3 启动Cell Illustrator并创建和编辑模型	31
4.3.1 添加元素	32
4.3.2 模型编辑和画布上的操作	33
4.4 模型仿真	35
4.4.1 仿真时的设置	35
4.4.2 图表的设置	36
4.4.3 执行仿真	37
4.5 仿真的参数和规则	37
4.5.1 用离散实体和离散过程创建模型	38
4.5.2 用连续实体和连续过程构建模型	42
4.5.3 离散和连续的概念	44
4.6 使用带有图标的元素进行通路建模	44
4.7 用Cell Illustraor来构建通路的模型	47
4.7.1 自然降解 (Degradation)	47
4.7.2 转移 (Translocation)	49
4.7.3 转录 (Transcription)	51
4.7.4 结合 (Binding)	53
4.7.5 解离 (Dissociation)	55
4.7.6 抑制 (Inhibition)	57
4.7.7 酶促反应下的磷酸化 (Phosphorylation)	59
4.8 总结	62

第5章 通路建模和仿真

5.1 信号通路的建模	66
5.1.1 主要参与者: 配体和受体	66
5.1.2 构建在EGF刺激下经EGFR引起的信号传导通路的模型	66
5.2 代谢类模型的构建	77
5.2.1 化学反应方程式和对应通路的表现方法	77

5.2.2 米氏方程Cell Illustrator的通路表现方法 77

5.2.3 构建糖酵解通路的模型 78

5.2.4 糖酵解通路模型的仿真 91

5.2.5 模型改良 91

5.3 构建基因调控网络的模型..... 95

5.3.1 体内时钟和昼夜节律 95

5.3.2 小鼠昼夜节律的基因调控网络 96

5.3.3 小鼠昼夜节律基因网络的模型化 97

5.3.4 根据仿真结果产生假说 108

5.4 小结 112

第6章 有关各种通路介绍

6.1 酵母的基因调控网络 113

6.2 基因网络的解析方法..... 115

6.2.1 基因网络的可视化 115

6.2.2 基因网络的布局..... 115

6.2.3 通路搜索功能 117

6.2.4 提取子网络的功能 119

6.2.5 两个子网络的比较 119

6.3 系统生物学的其他功能..... 121

6.3.1 通路语言：CSML 3.0和CSO..... 122

6.3.2 SaaS技术 122

6.3.3 通路参数搜索 123

6.3.4 更快速的仿真 124

6.3.5 输出通路模型到编程语言 124

6.3.6 通路布局算法 125

6.3.7 通路数据库管理系统 126

6.3.8 更多的可视化：图标的自动生成 128

写在最后..... 130

第1章

绪 论

系统生物学 (Systems Biology) 的主要目标是实现“对生命体进行系统理解”。这个短语意味着什么呢？利用计算机，我们可以对“信号传导路径” (signaling pathway)、“基因调控网络” (gene regulatory network) 和“代谢调控网络” (metabolic pathway) 做些什么呢？本书将成为有这样疑问和兴趣的人的第一本入门参考书。本书的优势在于，理解这些内容既不需要读者有任何微积分方程的背景知识和经验，也不需要掌握计算机编程。通过阅读这本书并使用软件Cell Illustrator，读者就能够构建用于仿真 (simulation) 的复杂生物学通路 (biological pathway)。首先在本章中，我们将阐述构成这些生物学通路的基础知识。

1.1 胞内活动

每个细胞内都发生着各种各样的活动。不同的分子履行着它们各自不同的职能，为细胞的生存和分裂提供必需的能量和蛋白质。在细胞表面，不同的分子接受着来自外部的不同刺激。这就类似于一个多样化的人类社会，有些蛋白质传导信号，有些蛋白质接收信号，有些分子严格履行为细胞制造能量的责任，而有些则可以帮助其他分子进行代谢。

1.1.1 转录、翻译和调控

细胞的功能开始于由DNA调控合成的蛋白质，并通过大量不同蛋白质的相互作用实现该细胞功能。整个过程包括首先在细胞核中以DNA形式编码的遗传信息，然后经历一个叫作转录的过程产生了mRNA。核糖体接着将mRNA翻译成蛋白质。所产生的蛋白质有着不同的功能。有些蛋白质在合成后转移到核内，通过与DNA上的特定位点结合，来调控某个基因的表达，这类调控通常发挥活化或者抑制的作用。在前一种情况下，由于起到基因上调的作用，基因表达得更为活跃。在后一种情况下，基因是下调的，因此有可能不会被表达。由此可见，并不是所有基因都一直处于表达状态。即使在同一个人身上，根据细胞类型，也存在着不同的基因表达模式的细胞。除此以外，

新发现的一种叫miRNA的RNA，它能够影响基因的表达调控。

专栏 1

小 RNA

众所周知，蛋白质构成了细胞功能的主体。正如上面所提到的，根据分子生物学的中心法则，是按照由DNA转录成mRNA，再由mRNA翻译成蛋白质的顺序产生的。但是事实上，从DNA转录而来的RNA中并不全都是mRNA，其中还包括许多不知功能的部分。这种RNA长久以来都被认为是RNA垃圾并没有被足够重视。

然而，在1993年，科学家们发现了一种并不翻译成蛋白质的能抑制特定基因表达的RNA。进入21世纪之后，也陆续在其他各种有机体中发现了类似的现象，这些RNA便被称为小RNA（miRNA）。miRNA很短，大概只有20~25bp长。miRNA由细胞核向细胞质移动，在它成熟之前会经过几个过程与蛋白质结合形成复合物，并和与自己部分序列互补的mRNA相结合，使mRNA失去功能，也就是说，miRNA具有抑制、阻碍特定蛋白质进行翻译的功能。在植物中也有类似的物质，被称作小干扰RNA（short interfering RNA, siRNA），能阻碍病毒RNA的转录。科学家们一直在研究小RNA片段的职能。据说，事实上地球上第一个有功能的类似核酸的分子正是RNA。核酸携带着信息，可以这么说，核酸是生命之本。从生物学角度来说，一个进化得很好的生物不会去维持任何毫无用处的机制或系统，因为维持一个系统是需要成本的。

综上所述，我们可以说生物网络很复杂，我们必须记住除了蛋白质以外还存在其他许多功能分子。

1.1.2 信号传导路径和蛋白质

有些蛋白质在产生之后就被分泌到细胞壁外，并向其他细胞传导信号。这些被叫作配体（ligand）的蛋白质能传导信息，而其他被称作受体（receptor）的蛋白质则可以接受这些信号。配体和受体的三维结构是互补的，类似于一把生物锁和一把钥匙。因此，一个配体仅能和与它形状匹配的受体结合。在接受配体之后，受体被激活并把信号传导给另一个蛋白质。这个蛋白质则依次激活其他蛋白质。这个传导信号的网络被称作信号通路（signaling pathway）或者信号转导路径（signal transduction pathway）。这些信号到达细胞核内并引起以上提到的基因调控作用。

1.1.3 代谢作用和基因

细胞新陈代谢发生的大量化学反应需要一些必需的复合物，如ATP、氨基酸和糖

类。举个例子，乙醇被代谢为乙醛，乙醛又变成乙酸。除了完整的反应物以外，这些反应需要基因表达的酶的参与。

1.2 胞内反应和通路

代谢通路是一个由多个反应组成的网络，信号转导路径和基因调控网络也是如此。我们通常称这种网络（network）为一个通路（pathway）。一般情况下，在教科书和通路数据库中，这些通路以基因和它的产物所组成的网络示意图的形式被展示。

图1.1显示的是基因调控关系的一个实例。基因MDM2抑制了能激活基因Bax的基因p53。这些连接基因的箭头（ $\perp$ ， $\rightarrow$ ）显示了基因间的不同关系。箭头 $\perp$ 代表抑制，箭头 $\rightarrow$ 代表活化。

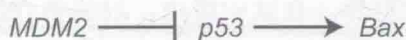


图1.1

图1.2是信号通路的一个例子。配体FasL携带着细胞凋亡信号。受体Fas与配体FasL结合，并且通过激活Caspase 8来传导信号。在信号通路示意图中，箭头代表着蛋白质之间的结合以及产生磷酸化作用的化学反应。

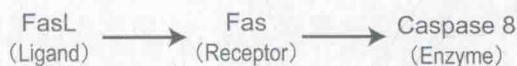


图1.2

从乙醇转变到乙酸的通路通常如图1.3所示。箭头依次连接代谢产物。每一个箭头代表着某一个代谢反应。尽管在这个示意图里各种必需的酶被省略，但它们仍然会作为一部分被包括在示意图中。

本书将以这些通路作为对象，利用计算机（in silico）学习理解生物系统作用机制的方法。

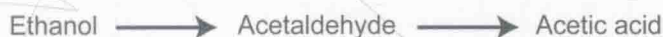


图1.3

第2章

通路数据库

通路信息可以从大量的数据库中获得，它的范围可以从由专业管理人（curator）构建的高质量的数据库开始到海量数据库，其中涵盖了大量假定存在的通路。这些通路是通过对摘要进行自然语言处理和文本挖掘所获得的。不管是商业还是公共数据库，因为数据库在大小、质量或者属性上存在差异，所以用户都需根据自己的目的选择正确的数据库。在这一章中，我们将介绍一些主要的通路数据库，这些数据库展示了以文献信息为基础的包括代谢、基因调控网络和信号通路的示意图。这一章也将涉及一些用来生成、编辑和分析这些通路的软件。

2.1 主要的通路数据库

全世界遍布着各种不同的通路数据库。每一个数据库都强烈地反映着设计者各自的意图和目的。有些数据库有详尽的代谢通路，有些数据库有详尽的信号通路。大多数数据库由专业管理人创建，他们通过阅读文献，提取通路信息并将其与数据库中的通路示意图进行整合。其他的一些数据库则使用自然语言处理和文本挖掘技术，对文献中的各类生物学关系（如基因调控关系）进行提取并将其并入数据库。这一章将介绍一些以代谢通路和信号传导路径为中心的数据数据库。

XML（eXtensible Markup Language）数据格式经常被用来描述通路信息，它在不同的数据库中有不同的形式，计算机和人都能容易地阅读这种格式。接下来的例子显示了“一个叫‘masao nagasaki’的人将会在2007年4月1日发表一场‘id’为5的演讲”。这个信息的XML格式如下：

```
<lecture id="5">
  <date>2007-04-01</date>
  <person>masao nagasaki</person>
</lecture>
```

在接下来的章节中，会出现以“…ML”结尾的词语，读者可以认为是以XML形式

描述的通路信息。这个结尾简单地表明了通路信息存储在XML的某些变化体中。在这本书中，我们将不深入讨论XML的细节。

专栏 2

什么是XML？

XML是一种可自我扩展的标记语言。它的固有名称是可扩展标记语言 (extensible markup language)。标记语言会使用一个句子结构对信息进行罗列和分类。XML在1996年由XML工作团队开发，这支团队是国际标准化组织W3C的一员。当文档创建者自由决定文档构造的规则并共享合格规则时，可以使用XML这个标准式样而不局限于特殊的构造，因而可以在各种应用中自由交换这些文档。

2.1.1 KEGG

KEGG (Kyoto Encyclopedia of Genes and Genomes) (<http://www.kegg.jp/>) 是由京都大学生物信息学中心和东京大学人类基因组中心联合开发的包含一系列小数据库的综合性数据库。这个数据库已经有超过十年的历史了。正如它的名字“百科全书”，KEGG数据库包含了对生物系统进行理解的必要信息，比如基因组序列和化学信息等 (图2.1)。以收集包括环境信息在内的所有和生物系统相关知识为目标的KEGG，也将会成为名副其实的“百科全书”。KEGG的“通路部分”主要由代谢通路组成。对于非商业用途的用户，使用许可证是免费的，而对于有商业用途的用户，需通过Pathway Solutions Inc. (<http://www.pathway.jp/>) 购买许可证。

KEGG的独特在于它对酵母、小鼠及人类代谢通路的聚焦和覆盖。最近，它又扩充了关于细胞周期和细胞凋亡的信号传导通路。专业管理人阅读并总结相关文献后创建新的通路。这些信息以通路示意图的形式展现在浏览者眼前。我们可以搜索从物质A到B的代谢通路是否存在，也可以搜索这样一个反应所需要的酶。除此之外，数据库还提供了相关知识的链接，如基因组序列、位点及反应条件等。这个数据库是以KEGGML的形式储存数据的。由于这些通路以GIF文件的形式显示，所以使用者不能方便地编辑通路信息。

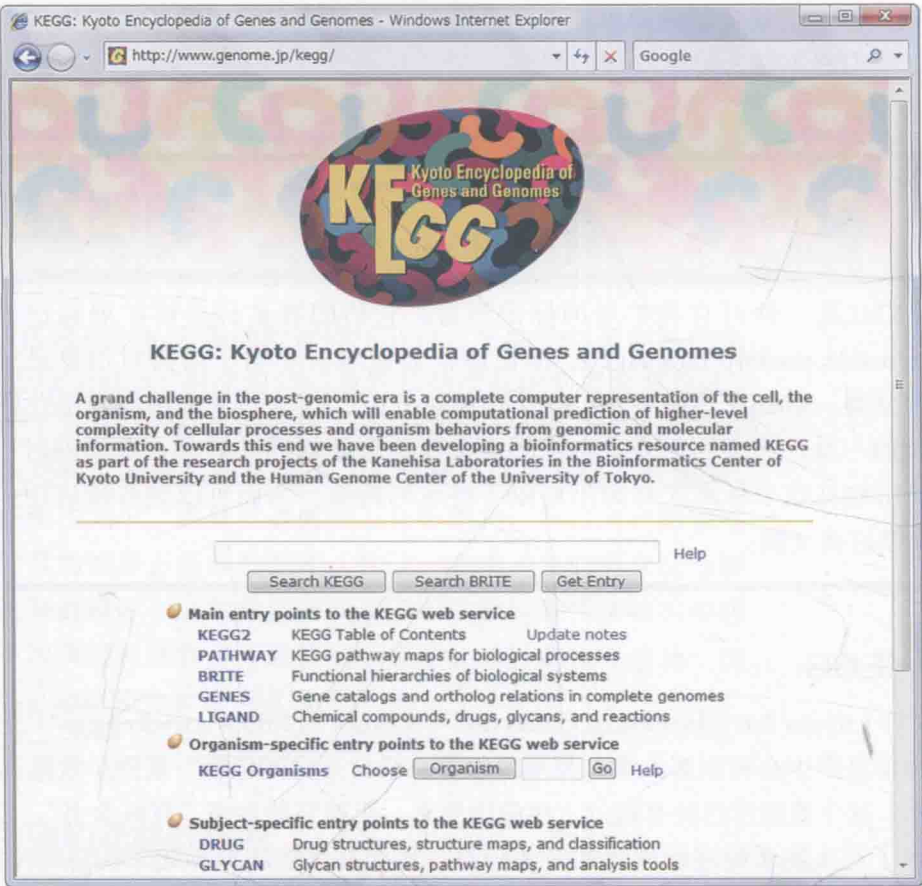


图 2.1

2.1.2 BioCyc

BioCyc (<http://www.biocyc.org/>) 是由SRI国际组织 (SRI International) 提供的通路数据库。作为一个优质数据库，它专注于代谢通路，这些通路最初是由SRI International的生物信息学研究小组创建的。与BioCyc相关的还有EcoCyc、MetaCyc、HumanCyc数据库。学术用途和非商业用途的许可证均免费。与列出的众多生物相比，人类和大肠杆菌是数据库中主要的研究对象。EcoCyc是一个以大肠杆菌代谢通路为主的数据库，其中的反应是以化学反应的形式表示的。EcoCyc也包含了少量信号通路，同样也是由专业管理人从文献中抽取通路信息，并以专用格式来描述的。

另外，BioCyc也列出了代谢通路上游的基因调控信息，也就是代谢通路和基因编码的酶以及它与调控者之间的关联信息。在BioCyc上，它根据不同层次的细节差异来对通路进行展示。在最详尽的层次上，代谢产物均以化学反应式的形式显示。

2.1.3 Ingenuity Pathways Knowledge Base

Ingenuity Pathways Knowledge Base (IPKB) 是由Ingenuity Systems Inc. (<http://www.ingenuity.com/>) 创建的通路数据库。包括学术和非商业目的在内的所有许可证都需

要收费。数据库由基因调控和信号通路组成。专业管理人从文献中提取信息至数据库,近年来它已涵盖了人类、小鼠和大鼠的基因信息(截至2008年5月,该网站宣称已有13600个人类基因、11000个小鼠基因和6600个大鼠基因)。数据库利用了Ingenuity Pathways Analysis (IPA) 软件(稍后会介绍)来观察并分析通路信息,因此通过网页浏览器是无法进入IPKB的。就像KEGG和BioCyc一样,IPKB利用自己的内部格式来存储数据。然而,与KEGG和BioCyc不同,IPKB允许通过IPA来编辑通路。这些被编辑的数据可以以图形的形式输出,比如SVG格式。

2.1.4 TRANSPATH

TRANSPATH是由BIOBASE (<http://www.biobase-international.com/>) 创建的有关基因调控和信号通路的数据库。不管是非盈利的用途还是商业用途,使用最新版本的数据库都需要缴纳费用。但是,TRANSPATH会以一个试用版的形式提供某些旧数据给学术用户(<http://www.gene-regulation.com/>)。除了TRANSPATH、BIOBASE,还有一个提供转录因子的数据库TRANSFAC和一个提供蛋白质数据的数据库 PROTEOME。BIOBASE同时提供了一个能分析这些数据库的软件ExPlain。

TRANSPATH和上述列出的这些数据库一样,都是由专业管理人创建的,因此保持着优质水准。通路也用自己专有的格式罗列。如果用户具有许可证,可以从网页浏览器上浏览到各种通路。除此之外,用户可以下载以文本格式储存的数据。例如,I- κ B的磷酸化过程如下所示:

```
IkappaB-alpha, IkappaB-beta : p50 : RelA +
ATP-IKK-alpha {p}: IKK-beta {p} : (IKK-gamma) 2
-> IkappaB-alpha, IkappaB-beta {ps}:p50:RelA +
ADP (phosphorylation)
```

每个反应都有一个文献的链接来保证它的存在。因此,理解每一个生化反应代表了什么是十分容易的。图2.2显示了IL-1通路是如何在网页浏览器上展现的,图2.3则展示了来自网页浏览器的TRANSPATH的反应信息。(截至2008年5月,网站表示已经涵盖主要包括人类、小鼠和大鼠在内总量为135563个的反应。)