

8 用户自定义选项功能的增强

UCSC Genome Browser提供的用户自定义选项功能允许用户按照自己设定的条件只浏览感兴趣的数据结果。

新增了两种用户自定义数据时提交的数据格式。

用户自定义数据的内在表示法现在是基于专用机器的数据库表，而不再是以前所使用的基于文件系统的执行命令。

最近几年，Genome Browser中最受广大用户欢迎的工具就是“作业保存（session-saving）”功能，用户使用该功能能够保存并共享多个浏览器配置文件以供日后再次使用。与“作业保存”功能相关的用户选项功能也得到了增强。

因为有许多网页设置，还有几乎无数的设置与显示选项组合配置，所以Genome Browser使用了cookies来保存这些设置。

9 未来的发展方向

UCSC将持续收录新的脊椎动物和非脊椎动物模式生物的基因组数据。他们会与NCBI和Ensembl保持密切合作，为获得及处理新数据制定出标准化的流程，以保证所有机构的数据都是一致的。

他们希望能为2×覆盖率的基因组计划（2×coverage species project）提供包含44种物种基因组的多重比对，还希望能为最新的人类基因组数据提供扩展的多重比对。千人基因组计划（1000 Genomes project）获得的数据也将收录进变异信息库，还将收录重组热点的高分辨率图谱（high-resolution maps of recombination hotspot）。

此外，GBD还计划对网站进行进一步的升级。例如扩展网站工具的实用性和可配置性，升级isPCR功能，能让用户查询“RNA空隙（RNA space）”来比对被内含子隔开的序列，还将继续升级用户自定义系统使其支持混合格式数据。

在接下来的几年中，我们计划能让用户在Genome Browser上直接输入信息，就像wiki做的那样，这样各种“基因专家”们就能在网站上发表评论、数据和其它各种信息。最后要说的是，UCSC一直在与全球传染性疾病预防组织（Global Solutions for Infectious Diseases）合作开发限制访问的医药数据系统，例如HIV基因组数据和临床数据。我们还与几个研究机构合作建立了癌症基因组网站（cancer genomics browser）。如果网站的访问和保密问题解决了，我们将很快推出该网站。

原文检索：Nucleic Acids Research, 2009, Vol. 37, Database issue D755–D761

 YORK/编译

五 人类代谢和信号转导途径数据库

人类生物学反应及信号通路数据库Reactome

反应组学（Reactome, <http://www.reactome.org>）是一个汇集了由专家撰写，经同行评阅的有关人体内各项反应及生物学路径的文章的数据库，该数据库相当于一个有效的数据资源以及电子图书。该库目前发布了共计2975个人类蛋白、2907项生物学反应以及4455个引用文献。该数据库为人们提供了一个全新的从整体

水平上对生物学途径进行研究的工具，同时，它也是一个改良的搜索及数据挖掘工具，可以简化与生物学途径相关的数据搜索与研究。此外，对用户提供的高通量数据组进行分析，也变得更为简单。目前，由于直系同源预测方法的改进，反应组学数据库也开始收录其它模式生物的数据了，现在通过与其它数据库合作和人工注释方式，已经收录了包括拟南芥 (*Arabidopsis*)、水稻 (*Oryza sativa*)、果蝇 (*Drosophila*) 及原鸡 (*Gallus gallus*) 等22种模式物种的反应组学数据。反应组学的数据库内容和相关软件都是开源共享，免费使用的。

1 更广泛的覆盖人类生物学反应途径

最新版本反应组学数据 (版本26, 于2008年9月公布) 覆盖了UniProt数据库中20,000个经人工注释过的人类蛋白质中大约12.5%的蛋白质，这一数量在过去三年中增长了2.7倍。反应组学数据库中对46个主要人类生物学研究领域，如细胞凋亡、HIV、流感病毒生活周期、DNA复制、转录、生理止血机制、碳水化合物代谢途径等进行了注释，并对与OMIM疾病表型相关的1005个蛋白质的正常功能进行了记录 (<http://www.ncbi.nlm.nih.gov/omim/>)。

2 更完善的工具、软件及数据模型

2.1 修订后的直系同源预测方法

用第二代OrthoMCL聚类处理方法 (clustering procedure, http://reactome.org/electronic_inference.html) 对OrthoMCL DB (<http://www.orthomcl.org/cgi-bin/OrthoMclWeb.cgi>) 中高质量的基因组数据进行处理，可以发现22个在进化上与人类有分歧物种的直系同源基因。新方法只将每个基因最长的转录片段作为研究对象，并采用基因为基础的方法代替蛋白质为基础的方法，编制用于OrthoMCL的Ensembl 标识符及在反应组学中应用的UniProt的序列号。这些改进提高了在不影响准确率的情况下使用电子注释时的成功率。

2.2 经改善的对大规模数据组进行分析、数据挖掘以及建立模型的工具

2.2.1 Skypainter

目前的Skypainter版本可以帮助用户直观地、更有效地看到通过大规模试验鉴定出的基因之间的功能关系 (图1A及B)。

2.2.2 Biomat

新开发投入使用的Reactome BioMart 工具 (www.biomart.org) 进一步简化了研究人员进行数据挖掘、交叉数据库分析以及大规模基因功能分析等工作。用户可以在不同的数据组 (如生物学反应途径、生物学反应、参加反应的复合体等) 之间进行选择，设定关键字，对要查找的数据特性进行限制，以便进行数据过滤，缩小查找范围。例如，对Reactome database进行搜索，用户便可以找到所有有某种蛋白质的复合体 (图1C) 参与的信号通路。而且，用户还可以将Reactome数据组中的“复合体”信息与UniProt proteome相连，从而找到这些复合体所有组份的序列。

2.2.3 Reactome

数据模型中的改变。研究人员尽可能地将数据模型最小化，以利于数据管理和使用。新增功能之一是有有关一个“黑匣子 (black box)” 反应类别，可以帮助用户找到那些特征定义不完全的生物学元件；另外一个新的功能是“另一细胞元件 (entityOnOtherCell)”，则可以对发生在两个细胞内的生物学元件及存在的复合物进行描述。

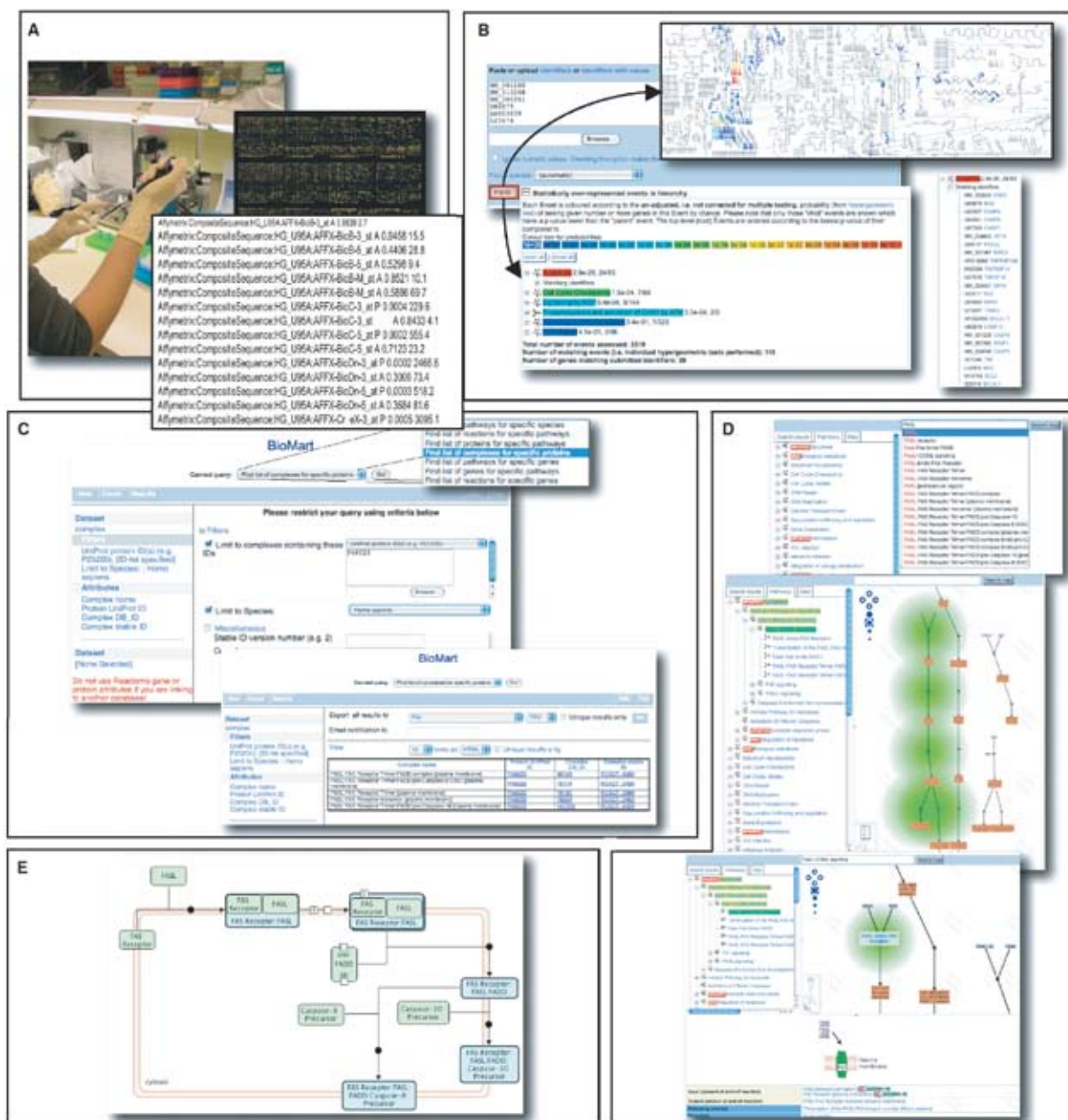


图1 使用Reactome对芯片等大规模试验数据进行数据分析应用实例。

A使用SkyPainter工具分析结果。

B图中的反应箭头根据用户提供的基因列表中参与该反应的基因数目不同而使用不同的颜色进行标记。统计学上“最热门”的反应突出表示。这些反应也会以列表的形式进行展示。该图中，凋亡途径是热门途径。

C BioMart能用来了解更多的基因信息。为了找出包含目的蛋白的复合体，要先选则复合体数据（名称、标识符等），输入蛋白在Uniprot中的标识符这些限制条件。然后点击左侧底部的“Dataset”按钮，用户就可以在Uniprot等多个数据库中进行搜索了。这样就可以获得例如蛋白质序列、注释信息等多种资料。图中，发现了大量的FASL:FAS受体复合体。

D 实体水平信号通路可视化工具用来发现有FASL:FAS受体复合体参与的信号通路。搜索FASL可以得到一个以分层方式呈现的列表。选择的信号通路和FasI/CD95L蛋白等信息以绿色高亮显示。反应途径以箭头形式表现。点击节点可以看到反应名称。分子以盒子形式展示，分子名称可以通过拖动显示。在图中的反应列表选择一个反应后可以在图中高亮显示出来。在图中选择一个反应也会在列表中高亮显示出来。左边还有缩放、滚动工具。还可以在网页左侧底部选择用“鹰眼模式”看信号通路图。拖动图中的盒子可以选择进行缩放操作的焦点位置。整个信号通路图都可以在网页中通过点击和拖动操作进行移动。点击网页底部的可折叠细节按钮可以了解更多有关该信号通路的细节。

E 部分使用网站上的新工具绘制的FASL/CD95L信号通路图。图中显示了该反应的亚细胞定位、参与反应的分子以及这些分子的化学计量学信息、翻译后修饰信息、结合位点等信息。

3 未来的发展方向

Reactome项目的长期目标是，为用户提供一个直观表现发生在生物体内的反应及反应途径的图表。为了实现这一目标，Reactome已经研发出了第二代整体水平反应途径可视化工具：通过更强大的“导航”功能，包括扩大缩小画面、滚动鼠标移动页面，对生物学过程进行高亮标注等，这个版本的工具可以为用户展示参与反应的分子之间更加细致的关系（图1D）。

这种对参加反应的分子以及反应过程进行“鸟瞰”式展示的方式，使得用户可以简便地获知每个参加反应的生物分子的细节信息。使用网页上改进后具备自动完成功能的搜索功能可以直接搜索反应名称、信号通路名称、个体名称、标识符和GO分子功能名称等。

Reactome项目正在开发兼具强大导航功能和可视化信号通路搜索功能的用户使用界面。为了达到这个目的，他们最近开发了SBGN工具（<http://www.sbgm.org>），使用SBGN就可以做出图1E那样的图。目前，他们在<http://brie8.cshl.edu:8080/ReactomeTools/AuthorTool/authorToolLaunch.html>网站上发布了该工具的测试版供大家测试。

原文检索：Nucleic Acids Research, 2009, Vol. 37, Database issue D619–D622

 YORK/编译

六 人类基因疾病数据库

癌症相关突变激酶信息数据库MoKCa



图片来源：MoKCa

当人类罹患癌症时，蛋白激酶家族成员是发生基因突变最多的一类基因。这些发生突变的激酶基因和被激活的激酶基因也是抗癌治疗最有效的作用靶点。MoKCa数据库（<http://strubiol.icr.ac.uk/extra/mokca>）是一个有关癌症相关突变激酶信息的数据库，库中收录了大量与突变激酶基因相关的结构注释信息和功能注释信息，以及当这些激酶发生突变时会产生怎样的结果的预测信息。肿瘤组织和肿瘤细胞系体细胞突变数据被标记在相应的蛋白质晶体结构域上。在蛋白质结构域象形图上以及蛋白质3D结构图上还对发生突变的氨基酸位点进行了突出标记。此外，在网页上还会显示与每一个突变有关的数据以及与该突变分子功能有关的专业注释信息。MoKCa数据库就是将蛋白质资源与功能注释资源相结合，并且加上例如结构域或磷酸化位点等结构和功能方面的注释信息，以此打造一个收录了每一个潜在的癌症相关突变激酶以及伴有权威肿瘤生物学家和结构生物学家注释信息的数据库。

原文检索：Nucleic Acids Research, 2009, Vol. 37, Database issue D824–D831

 筱玥/编译