以大肠杆菌为参照基因组的比较基因组学系统的建立

赵贵军 何智良 卢 阳 叶兰汀 王珣章 徐安龙*

(中山大学生物化学系, 广州生物信息中心, 广州 510275. * 联系人, ls36@zsu.edu.cn)

随着人类基因组计划及其他测序工作的顺利进行, 人们已经得到了大量的基因序列. 阐明这些序列的功能和意义成为功能基因组学的主要任务. 我们以大肠杆菌 *E. coli* 为参照基因组, 利用已公开基因组数据的24 种古细菌和真细菌蛋白质序列(可从 ftp://ncbi.nlm.nih.gov/genbank/genomes/bacteria 下载), 建立了一个基于互联网的可查询的比较基因组学研究平台(http://202.116.74.121:8080/database.htm). 该平台可用于研究同源基因在古细菌、真细菌中的分布、进化, 以及进行功能预测, 也可以查询对各种属特异的基因, 是对NCBI的COGs系统的一个改进. 在该数据库基础上, 进一步建立了原始基因的查询界面(http://202.116.74.108/DB_visualize.htm), 和各基因组总体直向同系物的动态变化图(http://202.116.74.108/chart.htm), 可用于查询古细菌和真细菌的共同保守基因, 并研究数据库中各细菌与大肠杆菌基因组的相似关系.

1 以 *E. coli* 为参照基因组的比较基因组学系统界面

把除 *E. coli* 外的所有细菌(古细菌和真细菌)种属中的所有蛋白质序列与 *E. coli* 的所有蛋白质序列作一对一的 BLAST 分析, 将 BLAST 分析结果全部载入数据库, 这样得到一个通过 SCORE 值关联的以 *E. coli* 基因组为联系和参照基因组的细菌种属基因组的 BLAST 比较数据库. 系统用 JAVA 开发, 运行在 SERVLET 平台上(http://202.116.74.121:8080/database.htm). 用户可以选择感兴趣的基因或全部基因进行排序. CLUSTAL_X 是一种多重比对工具, 并可以生成进化树文件.

目前数据库中还有以下几个 BLAST 种类:

(ⅰ) (*A. thaliana, D. melanogaster, C. elegans*) vs human. 研究从低等模式生物体到高等动物人类的基因进化及分布情况;

(ⅱ) (*M. gen, U. ure*) vs *M. pneu.* 研究这 3 种支原体中基因组的组成和基因分布;

(ⅲ) (*A. ful, M. the, P. abyssi, P. yro*) vs *M. jan.* 研究古细菌的基因组组成和基因进化;

(ⅳ) *P. yro* vs *P. abyssi.* 研究这两种热球菌属基因组的不同和相同之处.

这几种数据库目前均可任意选择, 在网页上会注明用户所选择的数据库.

2 原始基因的寻找

在已经构建的 *E. coli* 基因组与其他细菌基因组的BLAST数据库中, 每个 *E.coli* 基因都和其他基因组中有同源性的序列关联, 因此只要搜寻 *E. coli* 每个序列在其他基因组中出现的次数, 就可以统计出该种基因在细菌中的保守程度. 由于共有24个基因组和 *E. coli* 做比较, 因此出现的最大次数是 24, 即该基因在所有的 25 种细菌基因组中都存在. 如果它们的相似程度足够大的话(根据SCORE 值可以反映出来), 就可以认为这是一种非常保守的基因, 即原始基因. 根据该思路, 编制程序计算并将结果通过可查询的互联网界面反映出来(http://202.116.74.108/DB_visualize. htm). 经过检验, 发现 SCORE 值在 230 时, 能够找到在所有基因组中存在同源的 *E. coli* 基因.

3 细菌基因组总体直向同系物的动态变化

当 SCORE 值不同的时候, 各基因组所被选择出来的和 *E. coli* 的直向同系物的数量是不同的, SCORE 值越高, 表示同源程度越高, 返回的结果就少, 反之条件越宽松, 返回的结果越多. 但在同一个SCORE值下, 种属和 *E. coli* 亲源关系越接近, 序列同源性越高, 基因组中整体直向同系物的比例越高. 为证实这个假设, 编制程序, 计算在某个固定的 SCORE 值下, 各基因组中所有超出此条件的直向同系物在整个基因组序列中所占的比例. Internet 查询界面(http://202.116.74.108/chart.htm): 在 SCORE 框中可以输入任意的数值, 各数值之间用逗号隔开. 将在浏览器窗口返回反映满足这些SCORE 值的基因组中总体直向同系物的比例曲线.

本系统采用 JAVA 和 C++编成, 扩展功能强, 可移植性较好; 它直接面对互联网用户服务, 充分利用网络的共享特点, 系统的功能还将不断改进和扩充. 例如从基因组学的角度研究生物的特性和进化. 由于每个物种的遗传背景和生活环境不同, 其基因组的大小和表达的基因产物的种类、数量也将不同. 我们将进一步研究基因组水平基因产物按照功能分类的分布情况; 利用已知基因功能和序列上的同源性, 来发展预测未知基因功能的快速的基因组注释等方法.

致谢 本工作为国家自然科学基金重点资助项目(批准号: 69935020).

(2000-12-25 收稿)