

基于 PC/Linux 蛋白序列分析系统的构建及应用^{*}

寸树健, 邓百煌, 邓日强, 杨彦臻, 吴文言

(中山大学生命科学学院 // 生物防治国家重点实验室 //
海洋生物功能基因组开放实验室, 广东 广州 510275)

摘 要: 介绍利用常见的 x86 PC, Linux 操作系统配合 BLAST、PHYLIB 和 EMBOSS 等常见的生物分析软件构建可以完成多种蛋白数据分析的大规模自动分析系统。该系统可自动完成从 DNA 序列中获取读码框、核酸序列向蛋白序列的转化、在蛋白序列数据库中查找同源序列、序列录入数据库以及蛋白序列的等电点、亲/疏水位点、二级结构等性质的分析、多个序列之间相似度和进化地位的分析, 并对输出的结果数据利用 Web 服务器进行发布。与传统的序列分析模式相比, 此分析系统大大加速了对大规模数据信息的分析和利用。

关键词: linux; blast; emboss; phylib; 生物信息学工作站

中图分类号: Q754 **文献标识码:** A **文章编号:** 0529-6579 (2004) 05-0080-03

随着现代生物大分子序列测定技术的飞速发展, 可获得的序列信息空前的丰富, 然而如何有效地处理这些信息, 从中发掘有价值的生物学规律已成了生物学工作者必须面对的课题。生物信息学作为存储、修复、分析、整合生物数据的学科, 是分子生物学与信息技术的结合体, 在近年来的生物学研究中发挥了越来越大的作用。在后基因组时代, 科研人员越来越倾向于在掌握核苷酸序列的基础上从蛋白质分子中探索生命现象的规律, 同时现代生物学的研究方式也逐渐由传统的“实验摸索→确立理论”向“理论假设→实验验证”转变, 其间生物信息学起到了重要的作用^[1]。运用生物信息学方法分析生物数据离不开一个工作平台, 我们经过多次实践成功构建了基于廉价的 PC/Linux 系统的工作站, 安装了 PHYLIB, BLAST, EMBOSS 等生物分析软件包, 并配合相关数据库和使分析工作自动化的 shell 脚本。经过测试表明此系统在稳定性、速度和易用性方面都能胜任常见的蛋白分析工作, 可高通量、自动化的完成大规模的蛋白数据分析。

1 工作站的配置

1.1 硬件环境

1. 8G Pentium4 CPU, 512M DDR266 内存, 7 200 r/min, ATA100/80G 硬盘。此外, 考虑到科学运算对系统稳定性和可靠性的要求, 硬件系统被置于室

温 (~25 °C) 环境下工作; 并构建冗余硬盘阵列 (RAID 1 方式) 加强磁盘系统的可靠性^[2]。

1.2 软件环境

RedHat Linux 8.0 操作系统并进行基本的系统设置^[3]; 生物分析软件包, 如 blast. linux. tar. Z, phylib 3.6.1.tar.gz 等。此外, 视研究内容的需要下载了相关的数据库压缩包, 比如 swissprot.tar.Z 及 nr.tar.Z 等, 相关介绍可以参看 NCBI 网站的在线说明文档 (<http://www.ncbi.nlm.nih.gov/blast/blast.cgi>)。

2 Linux/Unix 系统下常用的生物分析软件

2.1 PHYLIB

PHYLIB 软件包可进行分析的内容包括分子序列 (DNA 或蛋白质)、基因频率、限制酶切位点、进化距离、进化树等, 所支持的生物信息学分析方法涵盖 parsimony, distance matrix 和 likelihood methods, 并包括 bootstrapping 和 consensus trees, 可以提供全功能的进化分析服务^[4]。PHYLIB 以源代码和可执行文件的形式发布, 用户可从其网站上自由下载、安装和使用。

2.2 BLAST

BLAST^{*} (basic local alignment search tool) 是在生物信息学分析中经常使用的计算机程序之一, 它

^{*} 收稿日期: 2003-12-02

基金项目: 国家自然科学基金资助项目 (39970583, 30271580); 973 资助项目 (G1999012002); 广东省自然科学基金资助项目 (001216); 广东省科技计划重点资助项目 (A3050303, 2003C104003); 广州市科技计划资助项目 (2003J1-C0061)

作者简介: 寸树健 (1977 年生), 男, 硕士研究生; 通讯联系人: 吴文言; E-mail: ls83@zsu.edu.cn

将 DNA、蛋白质等生物大分子序列之间的同源性问题转化为字符串之间相似性的问题，进而利用计算机强大的计算能力解决分子生物学工作者过去极难解决的同源性分析问题。BLAST 使用启发式算法在相应的生物序列数据库中搜索与给定序列有相似性的数据项，并根据相似度的差异把具有生物学意义的搜索结果反馈给使用者^[5]。用户可以登录 NCBI 网站将需要的软件包下载到本地硬盘解压缩后即可使用。必需的软件包是 blast. linux. tar. Z 和 fmerge. linux. Z，前者是打包的 BLAST 程序组，包含多个字符界面下的应用程序；后者是数据库升级程序。

2.3 EMBOSS

EMBOSS (the european molecular biology open software suite) 是作为自由软件发布的多个生物学软件的集合，可自动识别并处理多种格式的数据，同时还提供了用于功能扩展的库文件，用户可在此基础上根据自己的需要开发软件。EMBOSS 软件包通过源代码方式发布，可通过操作系统中附带的编译器（如 GCC）将其编译后得到近百个字符界面的生物分析程序^[9]。

3 蛋白序列分析及结果

3.1 分析过程概述

下面以对虾白斑综合症病毒（white spot syndrome virus, WSSV）的蛋白序列为例，具体说明如何使用所构建的工作站进行分析。WSSV 是一种杆状病毒类似的、无包涵体的双链 DNA 病毒，基因组测序工作已经完成^[7]，但目前相关蛋白的研究成果还不多。通过大规模的蛋白数据分析，从中寻求 WSSV 蛋白的理化特性和生物学规律，可为今后更深入的研究提供理论指导。我们在所构建的分析工作站上完成两个主要的分析任务：一是从 WSSV 基因组中获取 ORF，转换为蛋白序列并与 nr 数据库进行同源性比对，寻找同源的蛋白序列，在此基础上利用 PHYLIB 进行进化分析；二是对所获得的蛋白序列进行物化性质的分析和结构、功能的预测。

分析隐藏于 ORF 序列中的蛋白信息，需要下面的几个主要步骤：

从 ORF 中寻找蛋白序列，可由 EMBOSS 中的 getorf 程序完成；对蛋白序列进行等电点分析，由 iep 程序完成；二级结构预测，由 gamier 程序完成；抗原位点预测，由 antigenic 完成，等等，分析项目视研究人员的需要而定。通过运行自动化处理脚本，所有 ORF 序列可按照预先指定的分析内容自

动完成，并将分析结果通过 Web 服务器发布到互联网上。用户可以通过在浏览器的地址栏中输入含有结果文件存放路径的 URL，获得其文件夹中的数据文件。

3.2 序列分析结果

WSSV 的蛋白序列与 nr 库进行同源比对分析，找到了若干与数据库中具有较高同源性的序列。其中部分已见文献报道，如发现了一个编码 715 个氨基酸残基的多肽的 ORF 与真核丝氨酸/苏氨酸蛋白激酶有着明显的同源性。使用 PHYLIB 进行进化分析，结果表明 WSSV 与几个主要的蛋白激酶家族之间没有明显的发生关系^[8]。还有一些属于尚未报道的最新分析结果，如：① 编号为 2311 的 WSSV 基因组 ORF 在翻译为蛋白序列后与 nr 库进行比对，发现与部分高等生物的胶原蛋白前体有着极高的同源性，而胶原蛋白可能在蛋白吸附过程中起着重要的作用，因此 ORF2311 编码的蛋白可能涉及病毒粒子与宿主细胞的侵染吸附。② 从 WSSV 基因组序列中获取 ORF 制成 FASTA 格式的集合文件，使用 BLAST 包中的 formatdb 程序创建 WSSV 的 ORF 序列数据库，将 WSSV 的所有 ORF 与此数据库做序列比对，结果发现有 3 个 ORF 之间存在约 40% 的蛋白序列相似性。进一步的理化性质分析揭示了这 3 个蛋白除了等电点存在一定差异外，在二级结构、抗原决定簇分布、理论值大小、信号肽位置等方面均有很高的相似性，并在相同的位置存在（预测）跨膜区，提示这 3 个蛋白在起源上可能有着共同的祖先，并且执行相似的生理功能。③ 通过对 WSSV 基因组的 GC 峰图和读码框在基因组中的分布两项数据进行分析，可以了解到 WSSV 倾向于在基因组 0.5 ~ 1×10⁵ 的范围内使用 GC 含量偏高的 ORF，暗示在此区域内高 GC 含量的 ORF 编码序列易转座、易突变或易发生同源重组^[9] 等等（其余分析结果未列出，另文发表）。另外，将本系统应用于水母 cDNA 文库 EST 序列的数据处理和分析，也获得了很好的结果^[10]。

4 讨 论

本文综合利用多个可免费获得的生物学分析软件包，在低价位的 PC/Linux 系统上构建可完成大规模、多任务蛋白分析的工作站级别分析平台。我们已经采用该套系统成功地分析了大量蛋白序列数据，获得了许多有意义的结果。

就示例的蛋白序列分析而言，由于目前基因组测序工作已经日益普遍地进行，因此此类序列分析方法也必将越来越多地得到应用。通过从 WSSV 的

基因组数据中获取 ORF 序列, 并从中读取其编码的蛋白序列, 可最大程度地了解 WSSV 在宿主体内的蛋白表达情况。在整体把握 WSSV 基因组编码蛋白的基础上, 采用现代蛋白质组学研究的思路, 对所有蛋白序列进行分析比较, 从中筛选出在计算机数据分析中具有研究价值的基因或蛋白序列进行深入的实验分析、验证, 从而最大程度地提高研究效率并降低科研成本, 这具有很大的现实意义。

传统的蛋白序列分析是基于浏览器的, 通过 WEB 访问提供序列分析服务的网站 (如 NCBI 和 EMBL 等) 来完成蛋白序列分析任务。但是在线分析的工作效率受限于网络条件和单任务处理模式, 这对动辄需要分析上千条序列信息的分子生物学工作者来说太过繁琐, 因而将分析程序下载至本地机运行显示出了其必要性。本系统处理此类分析任务, 只需要将合理编号的 DNA 序列文件存放至系统工作目录, 启动脚本文件后即可按照预先设定的流程从读取 DNA 序列开始直至计算结果发布全部自动完成, 并可视需要配合 FTP 文件服务器提供结果文件的下载。如果程序运行的时间较长且不易估计, 可在 shell 脚本末尾处加入发送邮件的系统指令, 在计算完成后邮件发送服务器将发送特定内容的邮件到用户地址, 用户通过接收邮件的方式即可得知序列分析已经完成。建议将数据分析人员与普通用户隔离开来, 这样既降低了用户使用 Linux 工作环境的难度, 又避免了多个用户直接操作系统可能带来的不安全因素, 可最大限度的提高工作站系统的稳定性与运行效率。

相关网址:

BLAST 的官方网址: <http://www.ncbi.nlm.nih/>

blast

PHYLIB 的官方网址: <http://evolution.genetics.washington.edu/phylib.html>

EMBOSS 的官方网址: <http://www.hgmp.mrc.ac.uk/Software/EMBOSS/>

参考文献:

- [1] MINORU K. 后基因组信息学[M]. 北京: 清华大学出版社, 2000.
- [2] 康腊梅, 陈基禄, 武慧琴. 廉价冗余磁盘阵列技术的新进展[J]. 电力情报, 2000(2): 20—29.
- [3] 顾云苏. Linux 服务器配置与优化[M]. 北京: 清华大学出版社, 2001.
- [4] CYNTBIA G, PER J. Developing bioinformatics computer skills[M]. 北京: 科学出版社, 2002.
- [5] SCOTT F, JON C H. Evolutionary analysis[M]. 2nd Ed. US: Prentice Hall, 2000.
- [6] 张帆. 在 Linux 上编译程序[J]. 软件工程师, 1999(4): 20—21.
- [7] 张成岗, 欧阳曙光, 张绍文, 等. 基于 PC/Linux 的核酸序列分析系统的构建及其应用[J]. 生物化学与生物物理进展, 2001, 28(2): 263—266.
- [8] LIU W J, YU H T, PENG S E, et al. Cloning, characterization and phylogenetic analysis of a shrimp white spot syndrome virus gene that encodes a protein kinase[J]. Virology, 2001, 289: 362—377.
- [9] 张洪钧. 基因组研究——多学科交叉和相互启发[J]. 科学, 2002, 54(6): 15—18.
- [10] YANG Y Z, CUN S J, XIE X J, et al. EST analysis of gene expression in the tentacle of *Cyanea capillata*[J]. FEBS Letters, 2003, 538: 183—191.

Construction and Application of Protein Sequence Analysis System Based on PC/Linux

CUN Shu jian, DENG Bai huang, DENG Ri qiang, YANG Yan zhen, WU Wen yan

(School of Life Sciences //State Key Lab for Biocontrol, Sun Yat sen University, Guangzhou 510275, China)

Abstract: This paper introduces how to set up a bioinformatics workstation, based on linux operating system on x86 PC, that can analyze protein sequences with some usual biosoftwares (BLAST, PHYLIB, EMBOSS, etc.) automatically in large-scale. Isoelectric point, hydropathy, secondary structure and many other properties of protein sequence can be predicted by such a system, and the output data can be published as web pages via web server. In comparison with traditional mode of sequence analysis, the system makes analytical process efficiently and publishes results with much lower price. Its application will accelerate analysis and utilization to large number of protein data.

Key words: linux; blast; emboss; phylip; bioinformatics workstation