遗传 HEREDITAS (Beijing) 2007 年 8 月, 29(8): 1023－1026
ISSN 0253-9772 www.chinagene.cn

DOI: 10.1360/yc-007-1023

GSDS: 基因结构显示系统

郭安源, 朱其慧, 陈新, 罗静初

北京大学生物信息中心, 北京大学蛋白质工程和植物基因工程重点实验室, 北京大学生命科学学院, 北京 100871

摘要: 构建了一个用于绘制基因结构示意图的网站系统(http://gsds.cbi.pku.edu.cn/)。用户可提交核酸序列、NCBI 核酸序列号或基因外显子位置信息, 得到基因结构示意图; 并可指定在基因结构图上标注某些特定区域。系统允许用户同时输入多个基因, 并指定输出次序和标注区域。结果可用位图和矢量图两种图形格式显示。点击位图格式结果, 可以查看相应序列。系统提供中英文两种用户界面。

关键词: 基因结构示意图; 生物信息网络工具; 计算机图形

GSDS: a gene structure display server

GUO An-Yuan, ZHU Qi-Hui, CHEN Xin, LUO Jing-Chu

Center for Bioinformatics, Peking University, Beijing 100871, *China*

Abstract: We developed a web server GSDS (Gene Structure Display Server) for drawing gene structure schematic diagrams. Users can submit three types of data：CDS and genomic sequences, NCBI GenBank accession numbers or GIs, exon positions on a gene. GSDS uses this information to obtain the gene structure and draw diagram for it. Users can also designate some special regions to mark on the gene structure diagram. The output result will be PNG or SVG format picture. The corresponding sequence will be shown in a new window by clicking the picture in PNG format. A Chinese version for the main page is also built. The GSDS is available on http://gsds.cbi.pku.edu.cn/.

Keywords: gene structure display; bioinformatics web tool; computer graphics

基因结构示意图可直观地显示基因结构, 常用于基因可变剪切、基因家族分析。NCBI 基因数据库(http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene)、TIGR 模式生物基因组数据库(http://www.tigr.org/)都提供在线基因结构图。我们开发的拟南芥、水稻和杨树转录因子数据库(http://planttfdb.cbi.pku.edu.cn/)也用图形方式提供了各转录因子基因结构[1, 2]。 此外, NCBI 剪接序列比对在线工具(Splign, http://www.ncbi.nlm.nih.gov/sutils/splign/)[3], 用于 cDNA 序列与基因序列比对, 同时显示基因结构图。用户可提交一条基因序列和 5 条以下 cDNA 序列, 并以可移植的网络图象文件格式(Portable Network Graphics, PNG)显示基因结构。进行基因家族分析时, 经常需要得到该基因家族所有基因的结构示意图。此外, 有时需要对所得基因结构示意图进行编辑处理, 如缩放、标注等。但 PNG 是一种位图格式, 难以进行上述不失真的编辑, 因此有必要提供可编辑的矢量图格式(Scalable Vector Graphics, SVG)

收稿日期: 2006−12−09; 修回日期: 2007−02−04

基金项目: 国家重点基础研究发展规划(973 计划)项目(编号: 2003CB715900), 教育部博士后基金(编号: 20060390012), 国家高技术研究发展计划项目(国家 863 计划)和国家科技大平台项目资助 [Supported by the Key Project of Chinese National Programs for Fundamental Research and Development (973 Program)(No. 2003CB715900), MOE (No.20060390012), Chinese National Programs for High Technology Research and Development(863 Program) and China High-Tech Platform Program]

作者简介: 郭安源(1980–), 男, 江西人, 博士研究生, 专业方向: 生物信息学。E-mail: guoay@mail.cbi.pku.edu.cn

通讯作者: 罗静初(1947–), 男, 上海人, 教授, 博士生导师, 研究方向: 生物信息学。Tel: 010-62757281；E-mail: luojc@mail.cbi.pku.edu.cn

的结果。目前，尚无绘制矢量图格式的基因结构示意图的工具或网站。为此，我们开发了基因结构显示系统(Gene Structure Display Server, GSDS)。该系统可根据用户提交的序列或 NCBI 序列号绘制基因结构示意图，除常用的位图，还可提供矢量图。考虑到部分用户的需求，该系统允许用户一次提交多条基因序列，并显示多个基因结构。为便于国内用户使用，本系统提供中英文两种用户界面(http://gsds.cbi.pku.edu. cn)。自 2006 年 11 月发布以来，已经有不少国内外用户使用。

1 数据提交和处理

GSDS 为用户提供 3 种数据提交方法：基因和 CDS 序列、NCBI 核酸序列号或基因外显子位置信息(图 1)。对于这 3 种不同输入数据，系统分别采取如下方法处理。当用户提交基因和 CDS 序列时，GSDS 执行 EMBOSS 软件包中的 est2genome 程序[4]，通过 CDS 序列与基因序列比对，得到 CDS 序列中各外显子在基因上的定位信息，从而绘制出该基因结构图。当用户提交 NCBI GenBank 序列号(包括 GI 号和 Accession 号)时，则用 NCBI 的 EFetch 工具(http://eutils.ncbi.nlm.nih.gov/entrez/query/stati-c/efe-tch_help.html)链接到核酸序列数据库 GenBank 中该序列条目。若该序列条目的序列特征表(Feature table)中含 CDS 项，则提取 CDS 中各外显子在基因上的定位信息。若无 CDS 项，则无法得到基因结构信息，GSDS 会在结果中提示该基因没有相应结构信息。当用户提交外显子位置和基因长度信息时，则直接利用这些信息绘制该基因结构示意图。

GSDS 还可标注一些特定区域，如功能区、重复区等。若用户指定这些特定区域在基因或 CDS 上的位置，GSDS 可在输出结果中用不同于外显子的颜色显示该区域。此外，GSDS 提供了另外一个选项，用户可提交基因 ID 次序，以得到按该顺序排列的输出结果。

GSDS 界面友好，使用方便(图 1)。各输入框附有样例按钮，点击该按钮可自动粘贴样例数据，便于初学者熟悉输入数据格式和使用方法。该系统还提供了详细的在线使用文档，包括各种输入数据范例、输出结果说明及 3 种不同输入数据的处理方法。GSDS 为国内用户提供了中文界面。

2 实现方法

基因结构显示系统 GSDS 基于开放源代码的 Linux 操作系统和 Apache 网络服务器。页面实现使用 HTML 和 PHP 脚本语言。位图格式结果图片使用 PHP 的 GD 图形库绘制，SVG 矢量图使用标准 XML 代码生成。后台数据处理使用了 Shell 和 Perl 脚本语言。

3 结果输出

本系统提供位图(PNG)和矢量图(SVG)两种格式基因结构示意图，以黑白或彩色两种方式输出，并提供了相应图例。对 PNG 位图结果提供序列同步显示功能，即用户点击某一外显子、内含子或 UTRs 区域，系统弹出新窗口，显示该区域序列。SVG 矢量图可在安装了 SVG 图形显示插件的网页浏览器中直接显示。用户也可将矢量图下载到本地，并用 Illustrator 等软件自由编辑或修饰，如按照某功能域对齐、添加文字标注、修改颜色等。用户还可选择图片显示宽度。如用户提交 GenBank 序列号，输出结果图片下方则用表格列出所提交序列条目的信息，包括序列号、基因名称、编码区位置，以及所属物种等。若用户提供的 GenBank 序列号有误，或该序列条目无 CDS 区注释信息，系统将在输出结果时给出提示信息。GSDS 同时提供内含子相位显示选项。

图 2 为利用 GSDS 系统绘制植物特异转录因子 WRKY 家族中 5 个成员基因结构图实例。图 2A 为 GSDS 系统在线输出结果，图 2B 则是将图片下载后用图形软件编辑后的结果， 图中 5 个基因 3 端外显子编码的 WRKY 功能域(图中标注为红色)是 DNA 结合结构域；基因 BK005038 和 BK005073 编码的 5 端 WRKY 功能域用黄色标注。基因 BK005038 第 1 个外显子编码甘氨酸和脯氨酸富集区用黑色标注。

4 讨 论

与 NCBI 的 Splign 基因结构显示网站相比，GSDS 系统具有以下几个特点。(1)Splign 用于单个基因结构图显示，用户每次只能提交一个基因序列；而 GSDS 既可用于单个基因，也可用于多个基因结构显示，用户可一次提交多个基因，系统同时生成所有基因结构图，并可按指定次序输出结果、按不同颜色标注特定功能域，并可标注内含子相位。上述功能可用于基因可变剪接和内含子相位特征、多

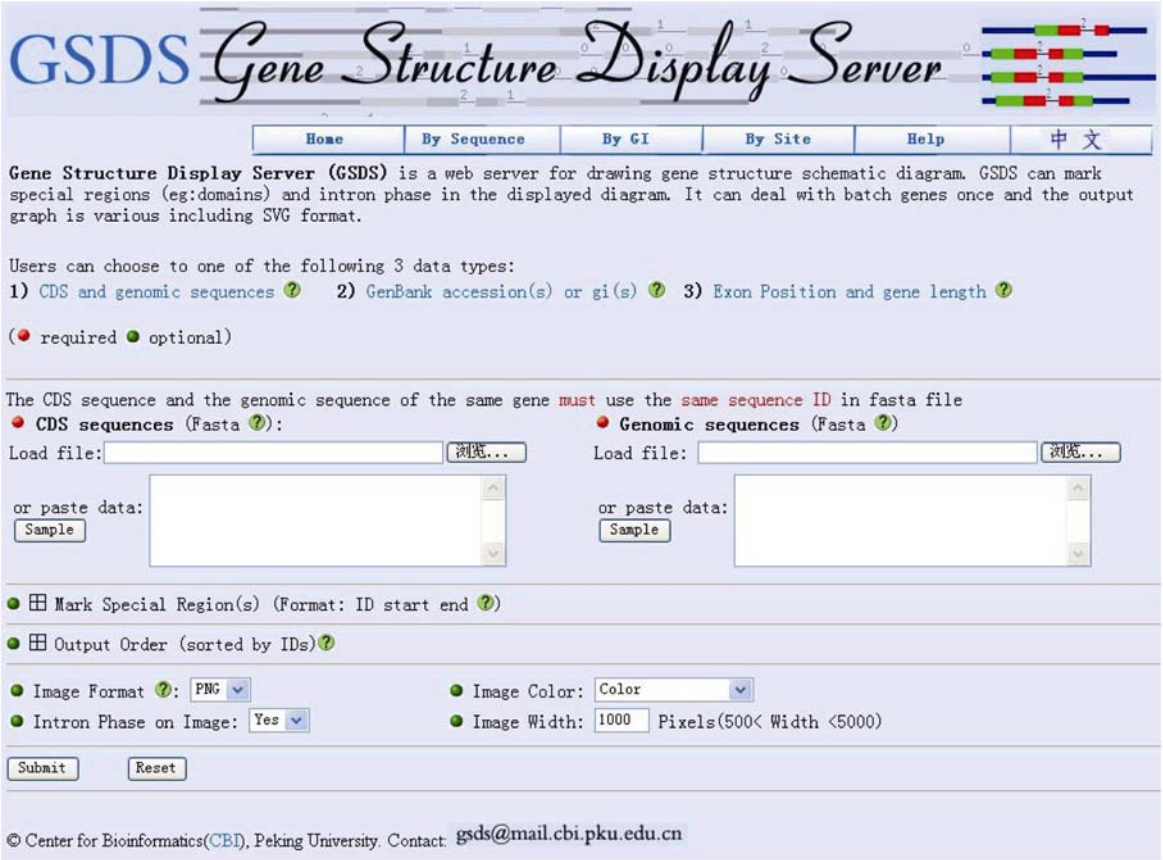

图 1　基因结构显示系统 **GDSD** 用户界面

Fig. 1　User Interface of the gene structure display system

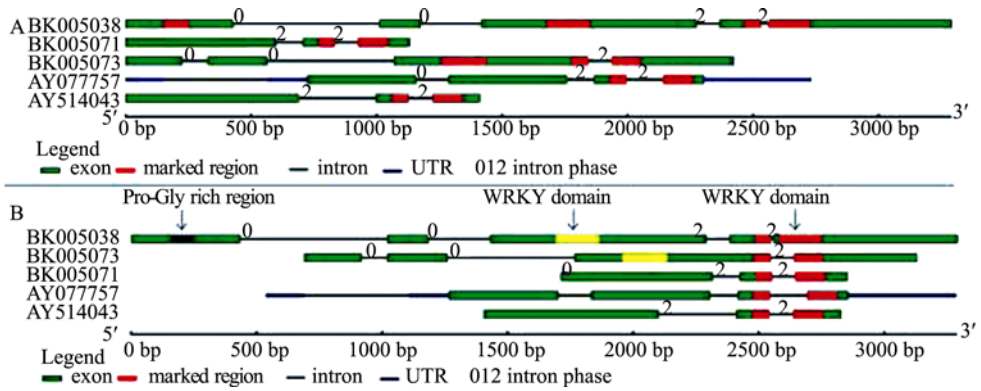

图 2　利用 **GSDS** 系统绘制植物特异转录因子 **WRKY** 家族 5 个成员的基因结构

A: GSDS 基因显示系统输出结果; B: 对图 A 进行编辑修饰后的结果。图中左侧为基因代码(如 BK005038); 绿色矩形框表示外显子, 蓝色线粗线表示 UTR 区, 黑色细线表示内含子, 数字 0、1、2 表示内含子相位; 红色矩形框为特定结构域, 可在递交作业时由用户设定, 并可通过编辑改变颜色(如图 B 中基因 BK005038 和 BK005073 的黑色和黄色矩形框)。

Fig. 2　The gene structure of five members of the WRKY transcription factor family generated by GSDS

A: The GSDS output of the gene structure; B: Edited with desktop computer graphics tools. Gene IDs (e.g. BK005038) are shown at the left. The green boxes denote exon, the blue lines show UTR and the black lines represent intron with intron phase shown as digits (0, 1, 2). The red boxes defined by user show special domains. The colors can be changed using desktop tools, e.g. the two yellow boxes of 5′-terminus WRKY domain of BK005028, BK005073, the black box located at the first exon of BK005038.

基因家族基因结构和保守结构域等研究。(2)Splign 系统只提供位图格式输出结果; GSDS 可生成 SVG 矢量图格式基因结构示意图, 用户可根据需要进行编辑修饰, 图形可任意缩放而不改变清晰度。(3)当提交 GenBank 序列号时, Splign 需同时提供基因序列和相应 cDNA 序列的 GenBank 序列号, 而 GSDS 只需提供基因序列, 系统会自动根据注释提取外显子及 UTR 位置信息。

基因结构显示系统 GSDS 在线网站为基因结构预测、基因进化、基因家族功能分析以及可变剪接等研究提供可视化工具。特别是 GSDS 系统提供的 SVG 矢量图格式, 为用户利用常用图形编辑软件进一步编辑提供了方便。如图 2 所示, 利用 GSDS 系统, 用户可快速生成拟南芥 WRKY 转录因子家族 5 个成员的基因结构, 并通过简单编辑, 清楚地显示该家族共有的 DNA 结合区 WRKY 功能域以及脯氨酸-甘氨酸富集区等其他特殊序列特征, 为研究该家族的分类、功能和进化提供简单实用的生物信息学工具。

参考文献(References):

[1] Gao G, Zhong Y, Guo A, Zhu Q, Tang W, Zheng W, Gu X, Wei L, Luo J. DRTF: a database of rice transcription factors. *Bioinformatics*, 2006, 22 (10): 1286–1287.

[2] Guo A, He K, Liu D, Bai S, Gu X, Wei L, Luo J. DATF: a database of Arabidopsis transcription factors. *Bioinformatics*, 2005, 21 (10): 2568–2569.

[3] Kapustin Y, Souvorov A, Tatusova T. Splign–a hybrid approach to spliced alignments. In: Proceedings of RECOMB 2004-Research in Computational Molecular Biology, 2004: 741.

[4] Olson SA. EMBOSS opens up sequence analysis. European Molecular Biology Open Software Suite. *Brief Bioinform*, 2002, 3 (1): 87–91.