第 32 卷 第 4 期
2013 年 7 月

华 中 农 业 大 学 学 报
Journal of Huazhong Agricultural University

Vol. 32 No. 4
July 2013,143～148

GSAS:针对基因序列的集成分析系统

黄 钰 李 平 李儒苗 魏小梅 胡 滨 王建勇

华中农业大学理学院,武汉 430070

摘要 以 J2EE 环境为基础构建 1 个基于工作流的基因序列分析系统 GSAS(gene sequence analysis system),该系统提供统一的数据、算法接口和可视化的图形界面,集成多种关键的序列分析算法和工具,同时实现异构数据的整合。GSAS 所采用的工作流机制为基因序列分析领域提供了协作研究的解决方案。

关键词 基因序列;序列分析;接口;集成;软件系统

中图分类号 TP 311.52 **文献标识码** A **文章编号** 1000-2421(2013)04-0143-06

目前,人类的基因序列图谱已经全部构建完成,基因组的研究也从先前的结构基因组学转向功能基因组学[1]。如何确定基因的功能,以及弄清全部的遗传信息成为当今生命科学最重要的研究领域之一。有关生命科学的基因序列分析研究中,随着计算机技术和相关实验技术的不断发展,不同实验室产生了大量异构的生物学基因序列数据,针对这些序列数据的分析以及注释算法工具也不断涌现出来,严重阻碍了不同实验室之间的实验数据以及相关分析方法的共享和交流[2]。因此,提供 1 个集成的分析系统成为必须[3]。

GeneMANIA (http://www. genemania. org)[4]是一款能独立做出快速、高效的基因功能分析和注释的工具,并对用户提供便利的 Web 界面。它能产生有关基因功能的假设、分析基因列表,扩展了使用有效的基因组学和蛋白质组学标识的且在功能上相似的基因的列表。此外,它还集成了一系列预测、注释算法工具,其功能扩展到可以让相关生物学家在其个人计算机的内存范围内对任意数量的基因组进行查询。在查询中产生的功能相关的与预测网络以及预测基因可以为更深入的分析产生一个注释性的 Cytoscape 网络。Pena-Castillo 等[5]提出一个标准化的收集小鼠功能基因组数据的组装工具,9 个生物信息学团队利用这些数据设置为独立训练分类,并产生预测的功能,基于 GO(gene ontology,基因本体)条款,为 21 603 个表现最好的小鼠基因提交了 1 组预测,并且确定了当前的功能基因组数据集的优缺点以及与相关功能预测算法的性能比较,该工具集成了一系列统计学预测和注释算法。然而,上述软件工具都没有提供一致的算法和数据接口,而且实现的算法和功能比较单一,使得研究人员需要使用多个软件包去分析这些海量的生物学序列数据,过程繁琐且难以掌握。同时,这些软件不具备可移植性。因此,笔者尝试构建 1 种基于工作流的基因注释分析系统 GSAS(gene sequence annotation system)。该系统基于 Java 环境,采用 C/S 模式(client/server,客户/服务器),以工作流机制提供一致的数据和算法接口,实现异构数据和算法的整合,为基因序列分析领域提供协作研究的解决方案。

1 研究方法

1.1 系统架构

工作流是将一项非常复杂的工作过程划分成单个独立的子过程,每个子过程的输出为下个子过程提供输入,将这些子过程通过一定的次序装配起来,就可以形成一条独立的数据分析流程。在工作流中将这些子过程称为节点。数据从进入第 1 个节点开始,按照顺序在各个节点中流动得到最终的分析结果。每个节点代表 1 个算法或者工具,都能对数据进行处理和分析。GSAS 的主要目的在于分析基因

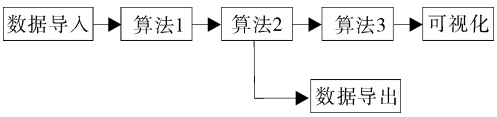
收稿日期:2013-03-19

基金项目:中央高校基本科研业务费专项(2012ZYTS021)、国家自然科学基金项目(61202304/F020606)和湖北省自然科学基金一般项目(2012FFB02801)

黄 钰,博士,讲师. 研究方向:系统生物学和生物信息学. E-mail:yhuang@mail. hzau. edu. cn

通讯作者:王建勇,硕士,副教授. 研究方向:生物信息学. E-mail:wjy01@mail. hzau. edu. cn

序列数据以及对结果的可视化，其核心思路是将序列数据的分析过程拆分为由一些算法节点连接而成的工作流程，即将相关的每个分析算法分解为一系列独立的节点，这些节点按照特定的方式连接，从而形成1个相对完整的序列数据分析过程(图1)。因此，单个的算法节点就成为分析流程的主要部件之一。



方框代表特定的算法节点 Each pan represents a specific algorithm node.

图1 GSAS的工作分析流程

Fig. 1 The analysis process of GSAS

为构建这种基于工作流机制的序列数据分析流程，GSAS拟应用Java 2平台企业版(Java 2 platform enterprise edition，J2EE)的架构模式[6-7]来搭建本系统。一般而言，J2EE使用3层或者4层的分布式应用模型，其中各个应用程序可以根据不同的功能划分为不同的组件，由这些组件所组成的J2EE应用系统可以根据需求安装于J2EE环境的各个层次。根据J2EE系统架构的划分，GSAS划分为相对应的3层，即客户端图形界面层、执行层和数据库层(图2)。

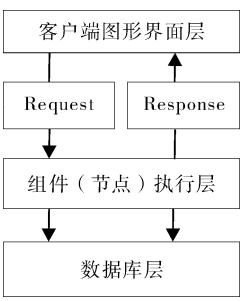


图2 GSAS的体系结构

Fig. 2 The architecture of GSAS

客户端图形界面层的主要功能是给客户提供一个友好的可视化图形界面，用户可以在自己的用户空间中添加自己的节点，还可以在工作流窗口构建基因注释分析工作流。

在执行层中，界面层的工作流被转换为具体的执行序列。该层主要表述工作流的逻辑模型，并且作为节点执行的虚拟机。执行层的主要功能是按照客户端界面用户构建的工作流，按照节点的连接次序，以及输入的数据来依次调用节点所对应的算法，然后将结果传递给下个节点作为下个节点的输入。该层中包含有很多的节点，例如算法节点、数据导入节点、可视化节点以及数据导出节点。此外，GSAS系统提供的软件开发包SDK(software development kit)定义了数据层和执行层之间的接口。

GSAS数据节点库由一系列算法功能节点组成，图形库由一系列可视化节点组成，二者构成本系统的数据库层。GSAS的这种多层应用模式不仅满足了构建数据分析流程的需要，并且也将数据细节和算法细节分离开来，研究者在客户端用户界面使用节点拖拉的方式来构建序列数据的分析流程，在每个节点的属性窗口修改相关的执行参数，最终的分析结果以图形化来显示或者输出保存。

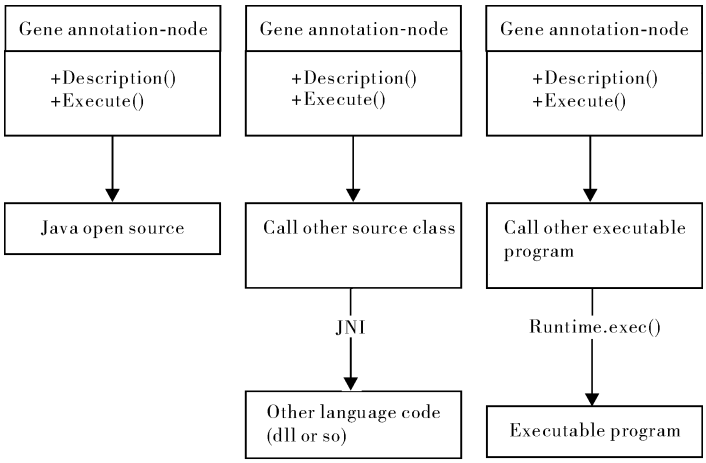
1.2 算法集成

为满足集成各种异构算法工具的需要，GSAS提供一致的应用程序接口API(application programming interface)。目前，算法工具主要有3种，即纯Java语言编写、其他语言编写、二进制可执行代码，3种算法都可以利用本系统提供的API快速集成，它们只需实现节点接口的2个方法：Property()和Run()。Property()方法必须实现节点的一些基本属性(节点名称、节点组名称等)。执行函数Run()则必须实现算法的具体功能。以上3类算法的集成如图3所示。

在GSAS系统中，基因序列数据分析模块中的节点主要可分为4类：数据输入节点、数据分析(算法)节点、可视化节点、数据输出节点。数据输入节点的功能是序列数据文件的输入，即将数据从客户端传至服务器。数据分析节点主要是对输入的序列数据进行数据格式处理、相关分析以及基因注释，目前已经集成了Phred、Phd2Fasta、Cap3、Consed、Primer3、Clustalw、MUSCLE、Blast等一系列相关算法和工具。可视化节点是把基因注释分析过程中的结果进行可视化，如将序列数据中的碱基的相关信息用图形显示出来，便于研究者分析。数据输出节点主要是将结果从服务器返回给客户端，并保存到相应的文件或者是数据库。

2 结果与分析

构建1组基因组注释的工作流(图4)。首先通过“LoadData”节点将各种主流测序仪(如ABI、SCF、ESD等)的峰图文件导入到服务端，然后将所需要的算法节点(可视化节点、数据导出节点等)由节点模块管理区拖拉至用户工程区，按照算法分析



纯 Java 代码的算法在方法 Execute()中被调用 Open source of algorithms programming with Java can be directly called in the method "Execute ()"；其他代码编写的算法采用 JNI(java native interface)集成 Open source of algorithms programming with other language can be integrated by adopting java naming interface (JNI) technology；二进制可执行代码采用 Java 外覆类来进行调用 For the binary files that can be executed，the "wrapper" container mechanism can be adopted.

图 3　算法工具集成示意图

Fig. 3　Integrative frame

的先后顺序将它们连接起来组成工作流程链。再按照算法需求通过节点属性编辑区设置参数，执行后就可通过可视化节点看到数据分析结果。其中"phred"和"phd2fasta"2 个节点将异构的数据进行一些基本的处理统一转化为 fasta 格式的文件，以供后续分析。

由图 4 可知，可以对 fasta 格式的序列文件进行各种分析和基因注释，通过"blast"节点可以对序列文件进行序列比对，通过"cap3"和"phrap"节点可以对序列进行基因序列的拼接和组装，然后通过"consed"节点，可以把序列拼接组装的结果以图形的方式显示出来，同时"consed"节点还可以对序列拼接的结果进行分析和修改。

2.1　序列拼接数据流程

以图 4 工作流中的序列拼接流程："LoadData"—"phred"—"phd2fasta"—"cap3"—"phrap"—"consed"为例来说明。首先，在"LoadData"节点的属性框中设置测序仪峰图文件的路径，再依次对其他各个算法节点设置运行参数，最后在"consed"节点执行工作流(图 5，图 6)。图 5 中第 1 行表示的是

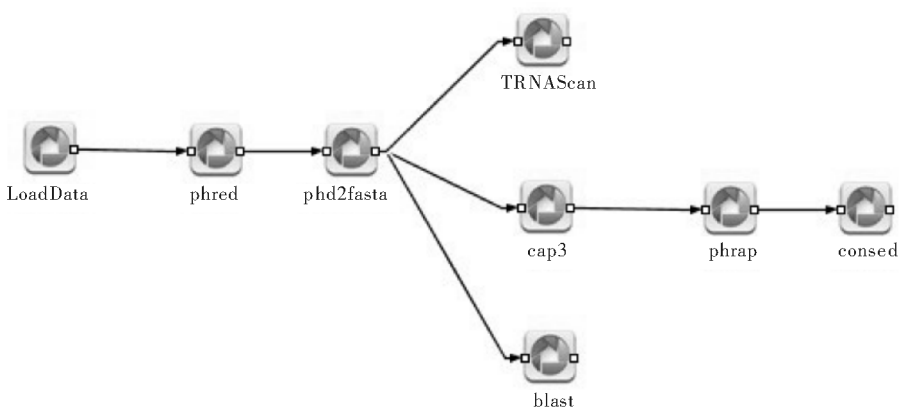


图 4　基因组注释工作流

Fig. 4　Genome annotation workflow

序列中碱基的位置，第1条序列是组装完成的contig序列，序列名为CONSENSUS。第2条和第3条是用于组装的序列，序列名字后的箭头表示的是测序的方向，可以查看每条用于组装的序列。图6为Q10F_A05.abi这条序列的峰图，图中曲线分别代表了不同的碱基以及各个碱基的分布位置。同时还可以用GSAS系统提供的其他统计学算法来检测CONSENSUS序列的拼接质量。

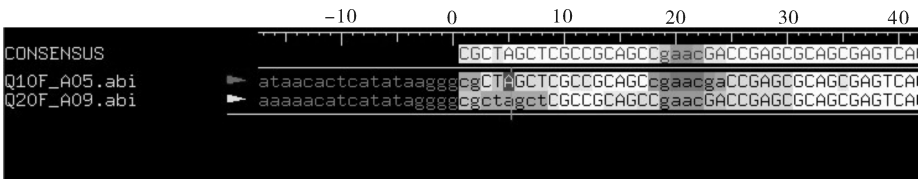


图5 序列拼接结果

Fig. 5 Sequence assembly results

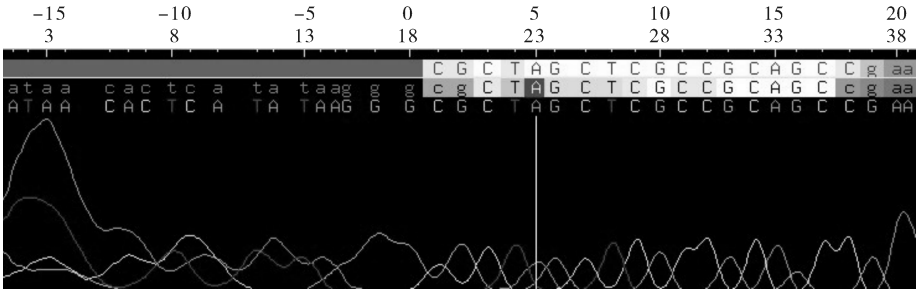


图6 序列拼接的序列峰图

Fig. 6 The peaks of the sequence assembly

2.2 基因组分析流程

利用图4工作流中的“tRNAScan”节点运行1个对猪链球菌基因组上的tRNA识别，部分运行结果见图7和图8。图7为预测的tRNA的信息，包括发现tRNA的序列的名字(sequence name)、发现的tRNA的个数、tRNA的起始和终止位置信息(tRNA Begin和Bounds End)、转运氨基酸的类型以及密码子(tRNA Type和Anti Condon)、内含子Intron的起始和终止位置信息(Intron Begin和Bounds End)以及预测的分值(Cove Score)。图8是运行“tRNAScan”节点后生成的tRNA的二级预测结果，即图7中的第1条tRNA的记录的tRNA的二级结构，gi|347750429|ref|NC_004350.2|表示序列的名字，.trnal(18 486～18 558)表示tRNA序列在基因组上的位置，tRNA序列长度为73 bp，转运氨基酸的类型为Ala，编码氨基酸的密码子TGC在tRNA序列的34～36位置上，在基因组中的位置为18 519～18 521，分值为73.48，第2行列出了tRNA的序列，最后1行指出了tRNA的二级结构信息，其中包括氨基酸接受区(GGGGCCT和CCTCGGA)、反密码区、二氢尿嘧啶区、TψC区和可变区。根据二级结构信息可以调用可视化节点，画出它的二级结构图(图9)，从图中可以看到除了氨基酸接受区外，其余每个区均含有1个突环和1个臂，最终可以对该二级结构图进行分析预测以及修改。

Sequence Name	tRNA #	tRNA Begin	Bounds End	tRNA Type	Anti Codon	Intron Bounds Begin	Intron Bounds End	Cove Score
gi|347750429|ref|NC_004350.2|	1	18486	18558	Ala	TGC	0	0	73.48
gi|347750429|ref|NC_004350.2|	2	22018	22090	Val	TAC	0	0	76.80
gi|347750429|ref|NC_004350.2|	3	22120	22192	Asp	GTC	0	0	74.51
gi|347750429|ref|NC_004350.2|	4	22220	22292	Lys	TTT	0	0	80.04
gi|347750429|ref|NC_004350.2|	5	22302	22383	Leu	TAG	0	0	62.25

图7 猪链球菌基因组tRNA预测结果

Fig. 7 Predicted results of *Streptococcus suis* genome tRNA

```
Predicted tRNA Secondary Structures:
gi|347750429|ref|NC_004350.2|.trna1 (18486-18558)       Length: 73 bp
Type: Ala       Anticodon: TGC at 34-36 (18519-18521)   Score: 73.48
        *    |    *    |    *    |    *    |    *    |    *    |    *    |
Seq: GGGGCCTTAGCTCAGCTGGGAGAGCGCCTGCTTTGCACGCAGGAGGtCAGCGGTTCGATCCCGCTAGGCTCCA
Str: >>>>>>>..>>>>........<<<<.>>>>>.......<<<<<.....>>>>>.......<<<<<<<<<<<.
```

图 8　猪链球菌基因组 tRNA 的二级结构预测结果

Fig. 8　The secondary structure prediction results of *Streptococcus suis* genome tRNA

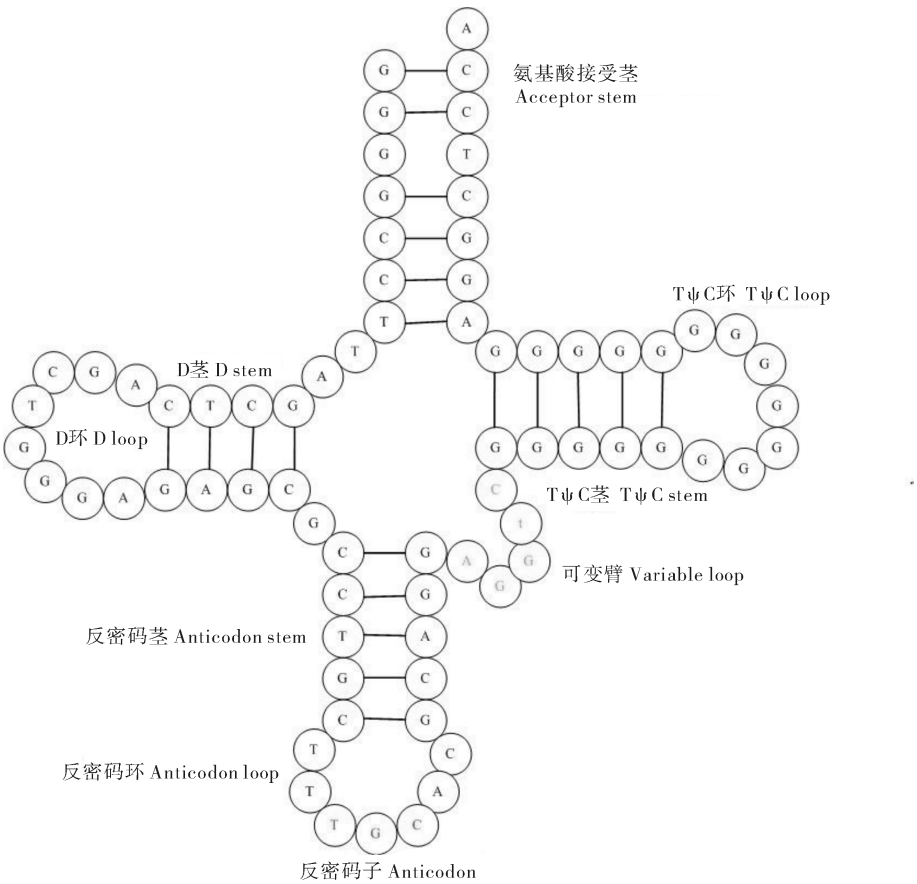


图 9　猪链球基因组中预测的第 1 条 tRNA 的二级结构图

Fig. 9　The secondary structure of the first tRNA of the *Streptococcus suis* genome forecast

3　讨　论

GSAS 系统为基因组数据的分析提供了一个集成的研究环境，与该领域内其他软件工具相比，该系统具有以下几个特点：可扩展性的 SDK 使得算法集成更为有效；客户端用户界面使得用户可以方便地构建序列分析流程，提高了研究效率；基于 Java 语言开发，因此可以方便地移植于不同的操作系统。GSAS 系统实现了对测序仪生成的峰图文件进行基本的数据处理，以及对经过基本数据处理的序列进行一些基本的分析，例如，基因序列的拼接和组装、序列比对、基因（组）的注释等。GSAS 目前集成的数据和算法工具非常有限，但是 GSAS 系统提供了一致的接口，在需要的时候可以方便地集成其他的软件算法工具，从而不断地向节点库中添加新的算法节点。GSAS 应用 J2EE 多层架构实现，可以满足复杂数据的分布式分析需求。对于海量的数据源以及相对复杂的算法节点，可以部署于服务器上来

进行计算，分析结果直接返回给客户端用户，因此，有限的计算机资源可以得到合理配置。

目前，相关研究领域的其他软件系统（GeneMANIA 等）并没有提供开放一致的算法集成接口。因此，当集成上述 3 类算法时，必须将源码重写，整个软件项目也需要在 OS 下重新编译，尤其是当集成二进制可执行代码时特别困难，基本需要全部改写。而在 GSAS 系统中，基于本系统提供的 SDK 下开发的各个算法节点，只需要 1 次编译，然后部署到指定的服务器端就可以使用，整个系统项目的代码不需做任何改变，体现了本系统良好的可扩展功能。同时，GSAS 系统目前还存在的一些问题。例如，如果节点之间传送的数据量非常大时，分析效率会降低。可行的解决方案是以被传送数据的地址来代替数据本身。

GSAS 系统软件可以联系通讯作者获取源代码以及相关使用手册。目前在 GSAS 的基因组序列数据分析模块中只实现了部分经典的算法。基于 GSAS 系统的开放式特征，新一代的测序组装以及分析软件也将被集成进本系统，如 Velvet[8] 等。在下一步的研究中，还将开发 GSAS 的 B/S 模式（browser/server：浏览器/服务器），用户可以直接在浏览器中发布，执行数据分析工作流程，并获取可视化结果。

参 考 文 献

[1] DULBECCO R. A turning point in cancer research：equencing the human genome[J]. Science，1986，231(4742)：1055-1056.

[2] GARDNER D. Neurodatabase. org：networking the microelectrode[J]. Nat Neurosci，2004，7(5)：486-487.

[3] HUANG Y，LI X N，LI Y L，et al. An integrative analysis platform for multiple neural spike train data[J]. J Neurosci Meth，2008，172(2)：303-311.

[4] MONTOJO J，ZUBERI K，RODRIGUEZ H，et al. GeneMANIA Cytoscape plugin：fast gene function predictions on the desktop[J]. Bioinformatics，2010，26(22)：2927-2928.

[5] PENA-CASTILLO L，TASAN M，MYERS C，et al. A critical assessment of *Mus musculus* gene function prediction using integrated genomic evidence[J]. Genome Biol，2008，9(S1)：S2.

[6] GRAWFORD W，KAPLAN J. J2EE 设计模式[M]. 刘绍华，毛天露，译. 北京：中国电力出版社，2005.

[7] BODOFF S，GREEN D，HAASE K，et al. The J2EETM turial [M]. 颜承，罗时飞，赵涌，等，译. 北京：中国铁道出版社，2003.

[8] ZERBINO D R，BIRNEY E. Velvet：algorithms for de novo short read assembly using de Bruijn graphs[J]. Genome Research，2008，18(5)：821-829.

GSAS：an integrative analysis system for sequence

HUANG Yu　LI Ping　LI Ru-miao　WEI Xiao-mei　HU Bin　WANG Jian-yong

College of Science，Huazhong Agricultural University，Wuhan 430070，*China*

Abstract　In analyzing gene sequences，various heterogeneous data analysis and a variety of software and algorithms should be integrated. The interactive graphical interface is a prerequisite. A software system named GSAS (gene sequence analysis system) based on Java-based J2EE environment was proposed. The system provides a unified data，algorithm interface and visualization graphical interface，not only integrates a variety of key sequence analysis algorithms and tools，but also achieves the integration of heterogeneous data. The workflow mechanism of GSAS provides a collaborative study solution for analyzing gene sequence.

Key words　gene sequences；sequence analysis；interface；integrative；software system

（责任编辑：陆文昌）