

基于 Web 的水稻芯片数据注释和分析平台

陈迪俊^{①②†}, 张帆^{①†}, 吴超^①, 李霞^{①*}, 陈铭^{②*}

① 哈尔滨医科大学生物信息科学与技术学院, 哈尔滨 150086;

② 浙江大学生命科学院, 杭州 310058

† 同等贡献

* 联系人, E-mail: lixia@hrbmu.edu.cn; mchen@zju.edu.cn

收稿日期: 2008-08-26; 接受日期: 2008-12-17

国家自然科学基金(批准号: 30500106, 30570424, 30571034, 30600367 和 30771326)、国家高技术研究发展计划(批准号: 2007AA02Z329 和 2008AA10Z125)、国家重点基础研究发展计划前期研究专项(批准号: 2008CB517302)、浙江省科技支撑和引导计划(批准号: 2007C22025)及黑龙江科技攻关计划(批准号: GB03C602-4)资助项目

摘要 国际水稻基因组测序计划(IEGSP)顺利完成, 水稻基因的研究也进入了后基因组研究阶段. 水稻基因芯片数据注释分析是一项重要的功能基因组学研究内容, 它为理解水稻基因的生物学意义提供了帮助. 本研究开发了一个基于 Web 的水稻基因芯片数据注释和分析平台(RiceChip), 它比同类的注释数据库更加全面快捷. 本平台共由 5 个功能模块组成: BioChip 模块为水稻基因表达数据提供快速检索和高级检索, 可依次按照 Probe Set ID, Locus ID, Analysis Name 等字段进行检索; BioAnno 模块整合多个生物学数据库, 为水稻基因提供基因功能、蛋白质结构、生物代谢途径以及转录调控等方面的注释信息; BioSeq 模块则收集水稻基因组的序列信息, 支持对水稻基因与芯片探针的序列查询; BioView 模块是系统图形可视化的核心模块, 提供友好的访问界面与结果输出, 方便研究人员使用; BioAnaly 模块结合 R/Bioconductor 统计分析工具提供高通量芯片数据的在线分析. 本系统从不同的方面依次提供了数据检索、基因注释、序列分析、数据可视化和数据分析等功能, 其数据收集的全面性与功能分析的强大性在同类水稻基因芯片数据注释和分析平台中都较突出.

关键词

水稻
基因芯片
注释和分析平台
功能基因组

水稻在人类饮食结构中占有主导地位, 在全球半数以上人的饮食结构中占 80%. 随着 2005 年国际水稻基因组测序计划(IEGSP)的完成^[1]和中国科学院基因组研究中心完成了水稻两个亚种粳稻(*Oryza sativa* L.ssp.japonica)和籼稻(*Oryza sativa* L.ssp.indica)全基因组工作框架图的测序, 已破译的水稻基因组图谱表明, 水稻基因组共包含 37544 个基因, 比人类基因还要多, 分布于 12 个染色体上. 对水稻基因组研究而言, 水稻基因组的测序仅仅是第一步, 完全识别

出水稻的基因及其功能, 以及研究这些基因如何编码产生蛋白质, 是摆在科学家面前的更大挑战. 利用功能基因组学的理论和手段, 挖掘出基因背后隐藏的生物学意义, 了解基因的功能, 可为揭示农作物重要性状的遗传控制的分子机理, 并从基因组水平上对这些作物实现遗传改良铺平道路. 水稻基因芯片数据注释和分析是一项非常重要的功能基因组学研究内容, 它为理解基因芯片数据的生物学意义提供了很大帮助.

目前, 国内外很多组织和研究机构都提供了水稻的相关基因功能注释数据库, 例如, 美国基因研究所水稻基因组注释数据库(TIGR Rice Genome Annotation)^[3,4]、日本水稻分子生物学百科全书计划(Knowledge-based *Oryza* Molecular biological Encyclopedia, KOME)^[5,6]、谷类比较基因组数据库(Gramene: A Resource for Comparative Grass Genomics)^[7]、日本水稻注释计划数据库(Rice Annotation Project Database, RAP-DB)^[8,9]、植物基因组数据库(Plant Genome Database, PlantGDB)^[10~12]、谷类small RNAs 数据库(CSRDB: Cereal small RNAs Database)^[13]及miRBase^[14~16]的水稻microRNA注释数据库等。

然而, 大部分水稻基因注释系统都有各自的不足之处, 如某些系统提供的各种注释信息分散在各地, 未很好地整合到一起; OsCAS(*Oryza sativa* Chips Annotation System)^[17]很好地整合了水稻基因芯片的注释信息, 但还存在注释信息不丰富、更新不及时、图形用户界面(GUI)不够友好等问题。总的来说, 这些注释系统存在如下的一些不足: (i) 只是针对某个领域的注释; (ii) 不支持批量注释; (iii) 注释信息更新不及时; (iv) 整合注释系统少见, 资源分散在异地, 没能有效地整合到一起; (v) 图形用户界面(GUI)不友好, 结果单调, 缺少图形显示, 交互性不强; (vi) 没有结合水稻芯片数据进行分析的注释系统。

本研究结合浙江大学生命科学学院仪器与技术服务平台(Equipment and Technology Service Platform, ETSP)Affymetrix芯片系统的水稻基因芯片数据, 开发了一个基于Web的水稻基因芯片数据注释和分析平台——RiceChip。在注释方面, 根据Affymetrix提供的NetAffx^[18]基因注释信息, 有效地整合了包括GeneBank, TIGR Rice, Gramene, InterPro^[19~21], KOME, KEGG^[22~24], GO^[25], CSRDB, miRBase等在内的数据库, 注释内容涉及基因功能、序列信息、染色体定位、蛋白质结构、生物学通路以及转录调控等方面, 是一个相对完整的水稻基因芯片数据注释系统; 在数据分析方面, 提供水稻基因芯片数据批量检索、数据可视化、数据筛选和下载、常见统计学分析和芯片数据挖掘等功能。并且本系统将基因芯片数据的注释和分析有效地结合在一起, 组成了一个内容丰富、功能强大的整合分析平台。

1 系统介绍

RiceChip 整合了多种异构数据库, 最终将它们统一存储到 MySQL 数据库管理系统中, 为基因注释和数据分析做准备。在数据获取和预处理过程中, 本系统提供了一些特定功能的 Perl 程序包, 方便数据的自动更新和预处理, 同时确保数据的完整性和一致性, 也为以后系统的更新提供方便。

RiceChip系统框架的中间层提供了5个核心子模块: BioChip专为芯片数据的检索而设计, 同时提供链接到其他模块的接口; BioAnno用于对芯片数据进行注释, 提供关于水稻基因和芯片探针的多种注释信息; BioSeq提供了关于基因和芯片探针的序列信息, 此外还提供了—个启动子查询子程序; BioView是系统数据可视化的核心模块, 提供 Probe View, Chromosome Distribution View, miRNA View, MA散点图和统计图等多种可视化功能; BioAnaly则结合R/Bioconductor^[2]统计分析工具, 提供数据分析的功能, 其中包括多种统计学分析, 网络分析、结构预测、数据挖掘技术等。这些模块并不是完全独立的, 而是相互结合在一起, 最终目的是通过对芯片数据的多角度和多层次的注释分析, 挖掘出隐藏在数据背后的生物学意义。

本系统主要采用Java语言作为开发技术, 基于成熟的Struts和Hibernate技术, 开发了一个基于Web的水稻基因芯片数据注释和分析平台。在数据获取和处理方面, 结合了Perl语言强大的文本处理能力; 在数据分析方面, 则采用了R统计分析工具和Bioconductor专业的芯片数据分析包; 同时, 还参考了Java与Perl在生物信息学应用方面的扩展分析程序包, BioJava^[26]和BioPerl^[27,28]。

RiceChip 系统的整体框架示意图如图1所示, 整个系统分为3层: 底层是数据库层, 主要提供了芯片数据和注释数据; 中间层是系统功能模块层, 提供芯片数据注释和分析功能模块; 顶层是用户视图层, 不同的用户可以根据需要选择自己关注的分析模块。

1.1 数据获取与处理

数据的获取和前期处理是本系统最关键的一部分, 它直接影响到后续芯片数据的分析和处理。

(1) 水稻芯片数据。本系统的芯片数据来源于

浙江大学生命科学学院 Affymetrix 芯片系统的 126 张水稻基因芯片数据.

(2) 水稻注释数据. 本系统的注释数据来源于不同的公众数据库, 这些数据库包括 NetAffx, TIGR Rice, Gramene, KOME, GO, InterPro 和 miRBase, 各

个数据库可提供的注释信息如表 1 所示. 本系统的 BioAnno 模块设计的目的就是整合这些公众数据库, 为水稻基因提供更加全面的注释信息.

(3) 数据处理. 芯片系统和各个公众数据库提供的

数据大部分是文本数据, 这些数据存在冗余和不

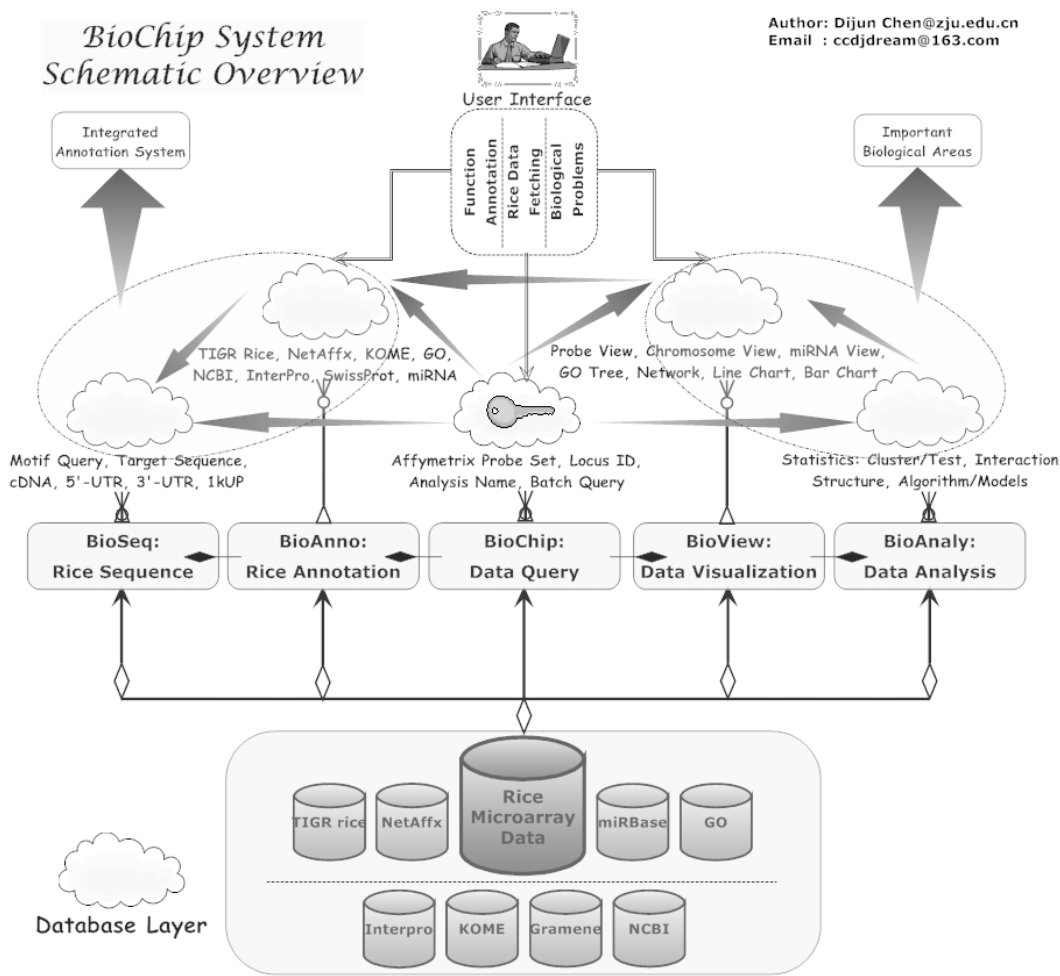


图 1 RiceChip 系统整体框架示意图

表 1 各种公众数据库提供的注释信息^{a)}

数据库	注释内容
NetAffx	NCBI 注释 ^{a)} 、GO 注释、TIGR 注释、染色体位置、结构域注释
TIGR Rice	NCBI 注释 ^{a)} 、TIGR 注释、KOME 注释、染色体位置
Gramene	NCBI 注释 ^{a)} 、GO 注释、染色体位置、KEGG 通路注释
KOME	NCBI 注释 ^{a)} 、KOME 注释、GO 注释、结构域注释
GO	GO 注释
InterPro	结构域注释
miRBase	小分子 RNA 调控

a) NCBI 注释是指注释信息来源于 NCBI 的 GenBank, UniGene, dBEST, RefSeq, LocusLink, Entrez Gene 等数据库

一致的信息, 而且有些公众数据库不提供数据下载。因此, 本系统提供了一些 Perl 脚程序, 用来提取和预处理这些数据, 然后把它们统一存储到 Mysql 数据库管理系统中。

1.2 模块设计

本系统共分为 5 大模块: BioChip, BioAnno, BioSeq, BioView 和 BioAnaly。

(1) BioChip 模块。BioChip 主要提供水稻基因表达数据的检索和获取等功能, 包括简单查询、批量查询和高级查询, 提供的查询字段主要有 Affymetrix Probe Set ID, Locus ID, Analysis Nam 等。数据结果主要以表格形式提供, 而且不同表达值的探针用深浅不同的颜色标注。为保证芯片数据的安全性, 还专门为本模块设计了一个安全系统, 获得认可之后的用户可以直接下载这些数据进行分析。

BioChip 模块还提供了链接到本系统其他模块(包括 BioAnno, BioSeq, BioView 和 BioAnaly)的接口, 以便对这些数据进行进一步的注释和分析。

(2) BioAnno 模块。BioAnno 是芯片数据注释的核心模块, 这些注释数据来源于多个公众数据库(表 1), 内容涉及到 NCBI 注释、GO 注释、TIGR 注释、KOME 注释、染色体位置信息、结构域(InterPro)注释、KEGG 通路注释和 miRNA 调控注释等方面的信息, 还提供了链接到 BioSeq 模块和 OsCAS 水稻芯片注释系统的接口。

(3) BioSeq 模块。BioSeq 设计的目的是为水稻基因提供完备的序列信息, 目前主要涉及到探针靶序列、基因上游 1 kb 序列、cDNA 序列、5'UTR 和 3'UTR 等序列信息; BioSeq 还附有一个基因启动子^[29,30]信息查询子模块。

(4) BioView 模块。BioView 是系统数据可视化的核心模块, 提供了 Probe View, Chromosome View, miRNA View, MA 散点图和多种统计图。

Probe View 支持批量查询, 结果用图形显示不同探针的信号值在所关注的芯片实验中的对比情况, 这样可以直观地显示出某些基因是否发生了差异表达; Chromosome View 直观地显示基因在染色体上的位置以及不同实验之间某个基因表达值的比较情况, 这样可以比较直观地看出在染色体的某些区域, 这些基因表达的整体效应——是否都是高表达或者低

表达; miRNA View 意在直观地显示水稻 miRNA 的信息; 统计图主要包括线图和直方图, 可以方便地看出某些基因在不同实验中的表达趋势。同时, BioView 也提供了链接到其他模块的接口。

(5) BioAnaly 模块。BioAnaly 提供数据分析的功能, 为本系统的重点和难点所在, 其中包括多种统计学分析, 网络分析、结构预测、建模和算法实现等。相关功能有待进一步完善。BioAnaly 也提供了链接到其他模块的接口。

1.3 系统实现

本系统主要用 Java 语言开发, 采用了基于 Struts 的 MVC 设计模式, 即模型层(Model)关注模型的建立和算法的实现, 视图层(View)只关注系统用户图形界面(GUI)的设计, 而控制层(Control)则关注系统流程的控制。这种设计模式使得系统层次清晰、可扩展性好。整个系统的流程图如图 2 所示, BioChip 是系统的核心模块, 其查询得到的数据可在本系统上作进一步分析, 包括功能注释(BioAnno 模块和 BioSeq 模块)、数据分析(BioAnaly 模块)、数据可视化(BioView 模块)等。图中的箭头表示了在本系统可以进行的分析流程及各个模块之间的关系。

1.4 特点介绍

本研究吸收了当前基因注释系统的优点, 再结合现有的芯片数据, 开发了数据分析功能模块, 使得本系统更具特色。总的来说, 相对于其他基因注释系统, 本系统主要有以下特点:

(1) 功能多样性。本系统在整合多种基因注释系统的基础上, 再结合现有的芯片数据, 添加了数据分析模块。具有数据检索、数据呈现、基因注释和数据分析等多样功能。

(2) 使用方便性。经过多年的发展, Bioconductor 已经成为芯片数据处理、分析、注释及可视化的标准工具。然而, 对于几乎没有编程经验的生物学家来说, R/Bioconductor 命令行式的使用方式过于复杂。本系统结合计算机技术, 将底层的操作方法以可视化的图形用户界面(GUI)展现给用户, 极大地方便了用户的使用。

(3) 结果可视化。本系统对查询和分析结果提供多样可视化功能, 这些可视化效果包括数据表格、

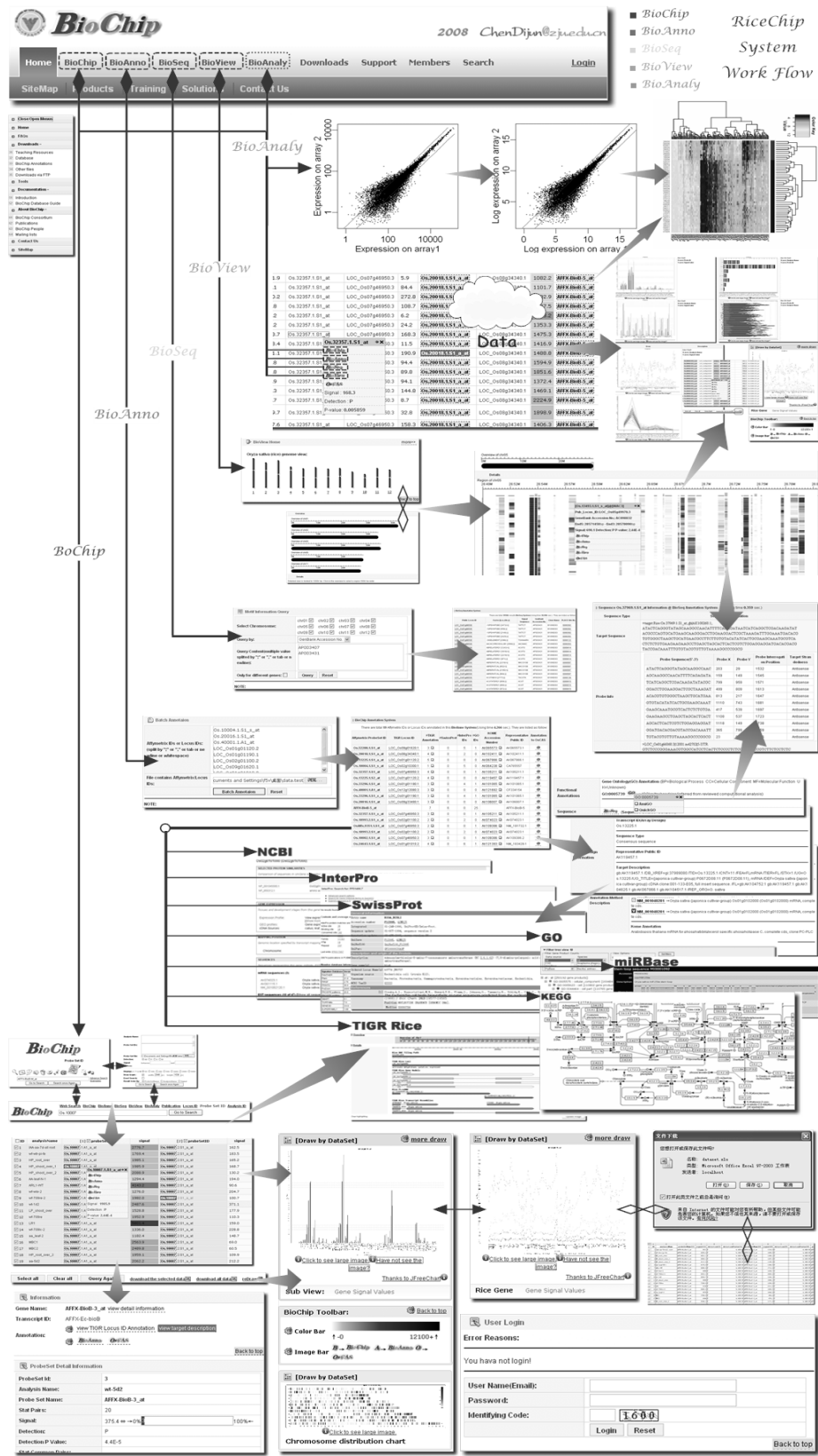


图 2 RiceChip 系统流程图

各类统计图、染色体定位图、探针标识图等, 直观地呈现了数据结果.

(4) 系统安全性. 为保证芯片数据的安全性, 本系统专门为 RiceChip 模块开发了一个安全系统, 通过注册就可以方便使用本系统检索、分析和下载数据.

(5) 界面国际化. 为方便用户使用, 本系统支持中文和英文界面. 如图 3 所示, 图 3(A)是系统顶部导航部分, 图 3(B)是系统左侧功能菜单部分.

2 结果

本研究构建了一个基于 Web 的水稻芯片基因注释和数据分析系统. 其中, 芯片注释模块为水稻基因和芯片探针提供了序列信息、基因功能、蛋白质结构、生物学通路以及转录调控等方面的注释信息, 并且为这些注释提供了多种可视化界面; 数据分析模块结合 R/Bioconductor 统计分析工具, 实现了高通量芯片数据的在线分析, 并且为分析结果提供多种统计图. 本系统是一个开放可扩展的芯片数据在线分析



图 3 RiceChip 系统中文和英文界面

平台, 随着新的分析工具的出现, 更多的数据分析模块可以继续整合到本系统之中; 而且, 伴随着多种新的生物学问题的出现, 本系统还可以针对某些具体问题提供相应的分析模块.

在此举例说明各个模块的部分分析结果. Bio-Chip 模块的数据检索图(图 4)、BioAnno 模块的批量

探针注释信息(图 5)、BioView 模块的探针分布图(图 6)和 BioAnaly 模块的数据统计分析结果图(图 7).

图 4 中根据探针表达值大小的不同, 用深浅不同的颜色来标示, 点击单个探针可以查看相应的注释信息.

图 5 的注释内容包括 NCBI 注释、GO 注释、

Os.2010.1.S1_at	7252.8	Os.2010.8.1.S1_a_at	401.5	Os.2010.4.1.S1_at	2905.4	Os.2010.2.1.S1_s_at	2217.6	Os.2010.2.1.S1_at
Os.2010.1.S1_at	7221.4	Os.2010.8.1.S1_a_at	336.8	Os.2010.4.1.S1_at	3055.1	Os.2010.2.1.S1_s_at	180.7	Os.2010.2.1.S1_at
Os.2010.1.S1_at	643.0	Os.2010.8.1.S1_a_at	328.9	Os.2010.4.1.S1_at	4807.5	Os.2010.2.1.S1_s_at	6.8	Os.2010.2.1.S1_at
Os.2010.1.S1_at	640.9	Os.2010.8.1.S1_a_at	258.7	Os.2010.4.1.S1_at	4412.1	Os.2010.2.1.S1_s_at	18.9	Os.2010.2.1.S1_at
Os.2010.1.S1_at	379.6	Os.2010.8.1.S1_a_at	104.5	Os.2010.4.1.S1_at	7860.2	Os.2010.2.1.S1_s_at	283.8	Os.2010.2.1.S1_at
Os.2010.1.S1_at	119.0	Bio-Chip	233.0	Os.2010.4.1.S1_at	978.7	Os.2010.2.1.S1_s_at	34.1	Os.2010.2.1.S1_at
Os.2010.1.S1_at	78.5	Bio-Anno	480.0	Os.2010.4.1.S1_at	4403.4	Os.2010.2.1.S1_s_at	357.0	Os.2010.2.1.S1_at
Os.2010.1.S1_at		Bio-Seq	435.3	Os.2010.4.1.S1_at	4233.4	Os.2010.2.1.S1_s_at	2814.5	Os.2010.2.1.S1_at
Os.2010.1.S1_at	969.4	Bio-View	633.1	Os.2010.4.1.S1_at	276.4	Os.2010.2.1.S1_s_at	396.3	Os.2010.2.1.S1_at
Os.2010.1.S1_at	325.7	OsCAs	290.9	Os.2010.4.1.S1_at	1933.9	Os.2010.2.1.S1_s_at	45.1	Os.2010.2.1.S1_at
Os.2010.1.S1_at		Signal: 258.7	556.9	Os.2010.4.1.S1_at	4226.5	Os.2010.2.1.S1_s_at	2512.8	Os.2010.2.1.S1_at
Os.2010.1.S1_at		Detection: P	708.6	Os.2010.4.1.S1_at	424.7	Os.2010.2.1.S1_s_at	49.3	Os.2010.2.1.S1_at
Os.2010.1.S1_at	342.6	P-value: 2.44E-4	310.0	Os.2010.4.1.S1_at	5003.2	Os.2010.2.1.S1_s_at	485.7	Os.2010.2.1.S1_at
Os.2010.1.S1_at	188.2	Os.2010.8.1.S1_a_at	299.0	Os.2010.4.1.S1_at	5323.3	Os.2010.2.1.S1_s_at	456.5	Os.2010.2.1.S1_at
Os.2010.1.S1_at	150.8	Os.2010.8.1.S1_a_at	676.0	Os.2010.4.1.S1_at	2310.4	Os.2010.2.1.S1_s_at	574.2	Os.2010.2.1.S1_at
Os.2010.1.S1_at	541.1	Os.2010.8.1.S1_a_at	639.8	Os.2010.4.1.S1_at	2246.8	Os.2010.2.1.S1_s_at	524.6	Os.2010.2.1.S1_at
Os.2010.1.S1_at	504.2	Os.2010.8.1.S1_a_at	384.8	Os.2010.4.1.S1_at	2698.0	Os.2010.2.1.S1_s_at	162.2	Os.2010.2.1.S1_at
Os.2010.1.S1_at	6829.4	Os.2010.8.1.S1_a_at	609.0	Os.2010.4.1.S1_at	684.9	Os.2010.2.1.S1_s_at	955.4	Os.2010.2.1.S1_at
Os.2010.1.S1_at	5140.9	Os.2010.8.1.S1_a_at	133.8	Os.2010.4.1.S1_at		Os.2010.2.1.S1_s_at	1196.3	Os.2010.2.1.S1_at
Os.2010.1.S1_at	5604.3	Os.2010.8.1.S1_a_at	354.9	Os.2010.4.1.S1_at	2119.9	Os.2010.2.1.S1_s_at	1348.8	Os.2010.2.1.S1_at
Os.2010.1.S1_at	6357.2	Os.2010.8.1.S1_a_at	249.7	Os.2010.4.1.S1_at	332.1	Os.2010.2.1.S1_s_at	1594.6	Os.2010.2.1.S1_at
Os.2010.1.S1_at	2033.1	Os.2010.8.1.S1_a_at						

图 4 BioChip 模块数据检索图

BioChip Annotation System								
There are total 19 Affymetrix IDs or Locus IDs annotated in this BioAnno System(Using time 4.266 sec.). They are listed as follow:								
Affymetrix ProbeSet ID	TIGR Locus ID	#TIGR Annotation	#SwissProt	#InterPro IDs	#GO IDs	KOME Accession Number	Representative Public ID	Annotation to OsCAs
Os.52200.1.S1_at	LOC_Os09g01620.1	3 ☐	0	5	1	AK065573 ☐	AK065573.1	☐
Os.20018.1.S1_a_at	LOC_Os08g34340.1	4 ☐	0	2	0	AK102411 ☐	AK102411.1	☐
Os.13225.1.S1_a_at	LOC_Os01g01120.2	5 ☐	0	0	0	AK067866 ☐	AK067866.1	☐
Os.10004.1.S1_s_at	LOC_Os02g10300.1	1 ☐	0	5	0	AK064238 ☐	CA765557	☐
Os.32357.1.S1_at	LOC_Os07g46950.3	4 ☐	0	0	1	AK105211 ☐	AK105211.1	☐
Os.13225.1.S2_at	LOC_Os01g01120.2	4 ☐	0	0	0	AK119457 ☐	AK119457.1	☐
Os.33296.1.S1_at	LOC_Os01g01190.1	3 ☐	0	4	1	AK101065 ☐	AK101065.1	☐
Os.40001.1.A1_at	LOC_Os12g12080.3	3 ☐	0	0	0	AK121692 ☐	CF334154	☐
Os.33296.1.S1_x_at	LOC_Os01g01190.1	3 ☐	0	4	1	AK101065 ☐	AK101065.1	☐
Os.20016.1.S1_at	LOC_Os09g33480.1	3 ☐	0	0	0	AK106007 ☐	AK106007.1	☐
AFFX-BioB-5_at		7	6	0	25		AFFX-BioB-5	☐
Os.32357.1.S1_x_at	LOC_Os07g46950.3	3 ☐	0	0	1	AK105211 ☐	AK105211.1	☐
Os.18993.2.S1_x_at	LOC_Os02g01100.2	2 ☐	0	2	0	AK074023 ☐	AK074023.1	☐
OsAffx.9351.1.S1_at	LOC_Os07g46950.3	2 ☐	0	0	1	AK109386 ☐	NM_191732.1	☐
Os.18993.2.S1_at	LOC_Os02g01100.2	2 ☐	0	2	0	AK074023 ☐	AK074023.1	☐
Os.10002.1.S1_at	LOC_Os07g46950.3	3 ☐	0	0	1	AK109386 ☐	AK109386.2	☐
Os.24643.1.S1_at	LOC_Os01g01010.2	4 ☐	0	6	1	AK121393 ☐	NM_183429.1	☐

图 5 BioAnno 批量探针注释信息

TIGR 注释、KOME 注释、染色体位置信息、结构域 (InterPro) 注释、KEGG 通路注释和 miRNA 调控注释等。

图 6 给出了染色体不同位置上的探针表达值; 图 7 给出了不同实验条件下的探针表达值。

3 讨论

本系统还在进一步完善之中, 未来有以下几个方面需要加以改进和实现:

(1) 完全整合 GO(Gene Ontology)数据库, 并实现系统的 GO 分析;

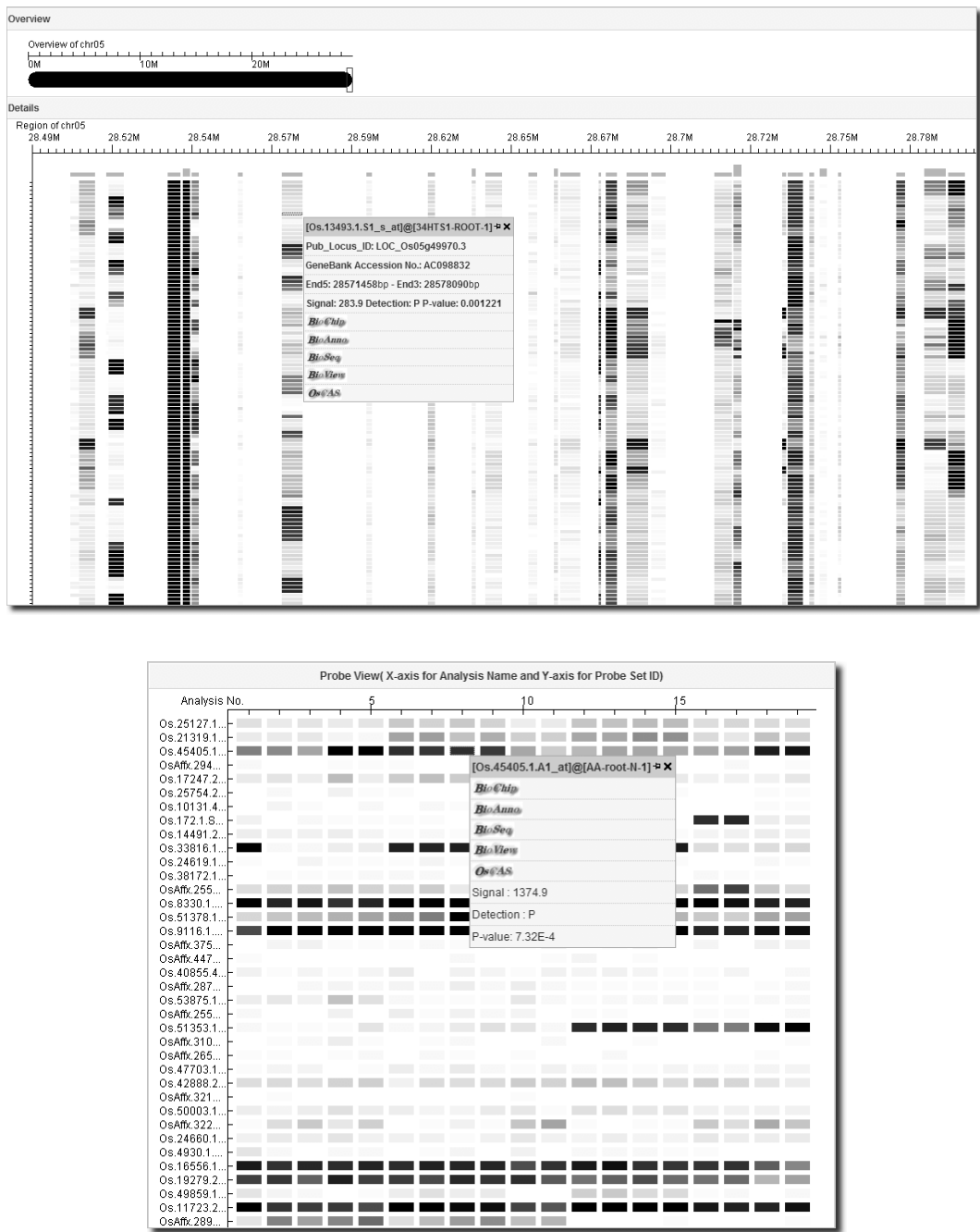


图 6 BioView 模块探针分布图

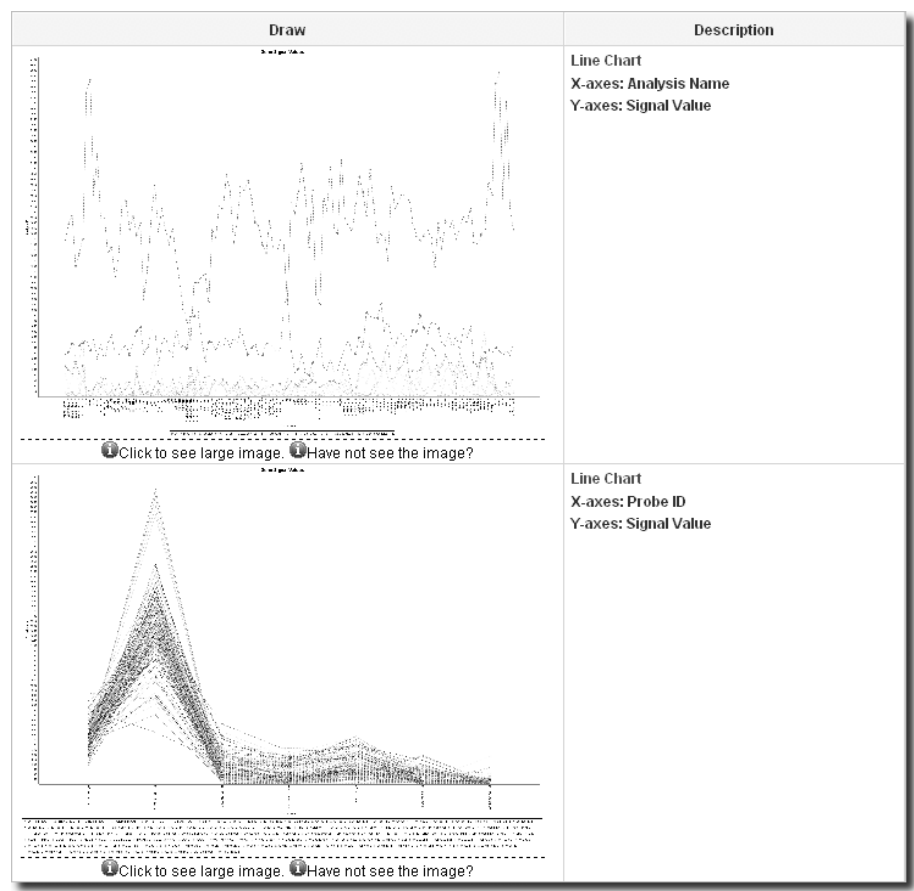


图 7 BioAnaly 模块数据统计分析图

- (2) 实现水稻的 SNP 注释;

(3) 在序列分析方面, 实现基于 BLAST 序列比对的分析功能;

(4) 系统可视化技术还有待加强, 实现 GO 树状图、互作网络图、结构预测图等;
- (5) 整合 R/Bioconductor 在芯片数据分析方面的工具, 实现芯片数据在线的多种生物信息学分析功能;

(6) 整合其他物种的生物信息学数据库(如拟南芥、玉米、大麦等), 提供这些物种的基因注释信息, 实现跨物种注释功能。

参考文献

1 International Rice Genome Sequencing Project. The map-based sequence of the rice genome. Nature, 2005, 436(7052): 793—800[DOI]

2 Gentleman R C, Carey V J, Bates D M, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol, 2004, 5(10): R80[DOI]

3 Yuan Q, Ouyang S, Wang A, et al. The institute for genomic research Osa1 rice genome annotation database. Plant Physiol, 2005, 138(1): 18—26[DOI]

4 Ouyang S, Zhu W, Hamilton J, et al. The TIGR Rice Genome Annotation Resource: improvements and new features. Nucleic Acids Res, 2007, 35(Database issue): D883—887[DOI]

5 Satoh K, Doi K, Nagata T, et al. Gene organization in rice revealed by full-length cDNA mapping and gene expression analysis through microarray. PLoS ONE, 2007, 2(11): e1235[DOI]

6 Rice Full-Length cDNA Consortium, National Institute of Agrobiological Sciences Rice Full-Length cDNA Project Team, Kikuchi S,

- Satoh K, Nagata T, et al. Collection, mapping, and annotation of over 28,000 cDNA clones from japonica rice. *Science*, 2003, 301(5631): 376—379[DOI]
- 7 Jaiswal P, Ware D, Ni J, et al. Gramene: development and integration of trait and gene ontologies for rice. *Comp Funct Genomics*, 2002, 3(2): 132—136[DOI]
- 8 Rice Annotation Project, Tanaka T, Antonio B A, Kikuchi S, et al. The Rice Annotation Project Database (RAP-DB): 2008 update. *Nucleic Acids Res*, 2008, 36(Database issue): D1028—1033[DOI]
- 9 Ohyanagi H, Tanaka T, Sakai H, et al. The Rice Annotation Project Database (RAP-DB): hub for *Oryza sativa* ssp. japonica genome information. *Nucleic Acids Res*, 2006, 34(Database issue): D741—744[DOI]
- 10 Schluter S D, Dong Q, Brendel V. GeneSeqer@PlantGDB: Gene structure prediction in plant genomes. *Nucleic Acids Res*, 2003, 31(13): 3597—3600[DOI]
- 11 Duvick J, Fu A, Muppirala U, et al. PlantGDB: a resource for comparative plant genomics. *Nucleic Acids Res*, 2008, 36(Database issue): D959—965[DOI]
- 12 Dong Q, Schluter S D, Brendel V. PlantGDB, plant genome database and analysis tools. *Nucleic Acids Res*, 2004, 32(Database issue): D354—359[DOI]
- 13 Johnson C, Bowman L, Adai A T, et al. CSRDB: a small RNA integrated database and browser resource for cereals. *Nucleic Acids Res*, 2007, 35(Database issue): D829—833[DOI]
- 14 Griffiths-Jones S, Grocock R J, van Dongen S, et al. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res*, 2006, 34(Database issue): D140—144[DOI]
- 15 Griffiths-Jones S. miRBase: the microRNA sequence database. *Methods Mol Biol*, 2006, 342: 129—138
- 16 Griffiths-Jones S, Saini H K, van Dongen S, et al. miRBase: tools for microRNA genomics. *Nucleic Acids Res*, 2008, 36(Database issue): D154—158[DOI]
- 17 He F, Shi Q Y, Chen M, et al. RiceDB: a Web-Based Integrated Database for Annotating Rice Microarray. *Rice Sci*, 2007, 14(4): 256—264
- 18 Liu G, Loraine A E, Shigeta R, et al. NetAffx: affymetrix probesets and annotations. *Nucleic Acids Res*, 2003, 31(1): 82—86[DOI]
- 19 Apweiler R, Attwood T K, Bairoch A, et al. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res*, 2001, 29(1): 37—40[DOI]
- 20 Mulder N J, Apweiler R, Attwood T K, et al. InterPro: an integrated documentation resource for protein families, domains and functional sites. *Brief Bioinform*, 2002, 3(3): 225—235[DOI]
- 21 Apweiler R, Attwood T K, Bairoch A, et al. InterPro--an integrated documentation resource for protein families, domains and functional sites. *Bioinformatics*, 2000, 16(12): 1145—1150[DOI]
- 22 Ogata H, Goto S, Fujibuchi W, et al. Computation with the KEGG pathway database. *Biosystems*, 1998, 47(1-2): 119—128[DOI]
- 23 Nakao M, Bono H, Kawashima S, et al. Genome-scale gene expression analysis and pathway reconstruction in KEGG. *Genome Inform Ser Workshop Genome Inform*, 1999, 10: 94—103
- 24 Ogata H, Goto S, Sato K, et al. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 1999, 27(1): 29—34[DOI]
- 25 Holland R C, Down T A, Pocock M, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 2000, 25(1): 25—29[DOI]
- 26 Holland R C, Down T A, Pocock M, et al. BioJava: an open-source framework for bioinformatics. *Bioinformatics*, 2008, 24(18): 2096—2097[DOI]
- 27 Stajich J E, Block D, Boulez K, et al. The Bioperl toolkit: perl modules for the life sciences. *Genome Res*, 2002, 12(10): 1611—1618[DOI]
- 28 Stajich J E. An Introduction to BioPerl. *Methods Mol Biol*, 2007, 406: 535—548[DOI]
- 29 Higo K, Ugawa Y, Iwamoto M, et al. PLACE: a database of plant cis-acting regulatory DNA elements. *Nucleic Acids Res*, 1998, 26(1): 358—359[DOI]
- 30 Higo K, Ugawa Y, Iwamoto M, et al. Plant cis-acting regulatory DNA elements(PLACE) database: 1999. *Nucleic Acids Res*, 1999, 27(1): 297—300[DOI]