基于多 Agent 的生物信息数据整合系统2BioAgent1

庄永龙[1], 马 飞[1], 周 敏[2], 沈 岩[2], 李衍达[1]
(1. 清华大学生物信息学研究所 教育部生物信息学重点实验室, 北京 100084
2. 中国协和医科大学基础医学院 100005, 国家人类基因组研究中心 100176)

摘 要: 随着人类基因组计划的实施, 生命科学研究已进入后基因组时代. 人们基于指数形式增长的核酸、蛋白质序列和结构等数据, 开发了数百种不同类型的数据库. 由于不同数据库存贮和检索方式的极大差异, 给研究者对它们的整合应用造成了一定的难度. 本文建立了一个基于多 Agent 的生物信息数据整合系统2BioAgent, 通过信息采集 Agent、信息整合 Agent、用户 Agent 的协调, 完成数据抽取、数据标准化、数据存储、数据融合、Web 显示等工作流程, 以实现数据整合的自动化. 同时利用 BioAgent 系统, 开发了人类精神分裂症相关基因的突变信息多层次定位数据库.

关键词: 生物信息学; 数据整合; XML; 多 Agent
中图分类号: Q811. 4 文献标识码: A 文章编号: 037222112 (2005) 0120078205

BioAgent: a Biological Data Integration System Based on Multi2Agent

ZHUANG Yong2long[1], MA Fei[1], ZHOU Min[2], SHEN Yan[2], LI Yan2da[1]
(1. Institute of Bioinformatics of Tsinghua University, The Key Laboratory of Bioinformatics of Ministry of Education, Beijing 100084, China;
2. Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences, Beijing 100005, China Chinese National Human Genome Center, Beijing 100176, China)

Abstract: With the exponential growth of biological data, there exist several challenges to effectively using such information for further research. Biological data are often complex, heterogeneous, hierarchical and dynamic, and may need to be integrated. We developed a software system, namely BioAgent, based on Multi2Agent system and XML technology for data integration, querying and customization. BioAgent consists of Information Extraction Agent, Intelligent Integration Agent, User Interface Agent and Information Customized Agent. We also developed a database of variants located on several levels based on BioAgent system.

Key words: bioinformatics; data integration; XML; multi2agent

1 引言

20 世纪 90 年代, 随着人类基因组[1]和其他模式生物基因组如小鼠、水稻、玉米、拟南芥测序计划的全面实施, 生物信息数据指数增长. 目前, 专门收集生物数据库信息的 DBCAT (http://www.infobiogen.fr/services/dbcat/)[2] 已收集的各种类型数据库已达 511 个, 因此, 如何管理、存贮、查询和开发利用这些多样的生物信息数据已成为生物信息学发展中的一个重要问题[3, 4].

首先, 呈指数增长的数据对数据的管理和处理提出了更高的要求. 数据更新速度加快, 同时需要将原始的数据及修改记录存档; 数据类型复杂多样, 包括基因组信息、染色体、基因突变、遗传疾病、分类学、比较基因组、基因调控和表达、放射杂交、基因图谱、基因芯片数据等, 以及新的分析方法产生的新数据类型, 如新的蛋白结构分类方法产生的蛋白分类结构数据; 数据描述标准缺乏统一, 导致存贮信息的数据结构存在很大差异, 包括文本文件、关系数据库、面向对象数据库等, 而且有很多数据信息是描述性的信息, 不是结构化的信息表示, 如 OMIM(http://www.ncbi.nlm.nih.gov/omim/) 中的临床疾病描述; 数据库中存在大量的数据冗余、数据错误以及数据的不一致现象[5]: 如数据载体形态变化时出错(一般是指由出版物转为其他载体形态), 实验室或测序机构提交的数据有误和由数据库工作人员创建数据项或对序列进行功能注释时出错等.

其次, 如何更方便, 更快捷地将各种不同的数据资源进行整合利用更是一个亟待解决的问题. 由于各个数据库存储的信息特点不尽相同, 存储查询传输的方法和标准更是大相径庭. 虽然在数据库内部, 信息的组织和一定的查询、传输方法配合, 一般可以获得较高的效率和信息使用质量, 但跨数据库查询的效率很低, 也给用户的研究带来许多不便, 同时, 这也严重影响了生物信息学的深入研究[4]. 对大多数用户来说, 为获得足够的信息, 都需要进行跨数据库查询, 例如, 在基因组

收稿日期: 2003209205; 修回日期: 2004204228
基金项目: 国家 973 项目(No. 2001AA221071); 科技部重大项目前期研究专项(No. 2001CCA01400); 国家自然科学基金重点项目(No. 60234020)

学、蛋白组学的研究中, 由于基因组、蛋白组中的信息互相关联、互为补充, 因此, 对不同数据库进行整合就显得非常重要[6].

2 数据整合

数据整合不是数据库的简单叠加, 更重要的是寻找不同类型数据之间的相互联系, 将这些数据库中彼此联系的信息整合起来, 形成一个构架于这类数据库之上的数据整合平台, 为生物信息学的研究构建信息平台[7,8]. 一类相似的生物问题所需要整合的生物信息存在很大交集, 针对研究对象, 从综合数据平台中挖掘与之最为相关的信息, 可以帮助我们阐明不同数据之间的联系. 数据整合主要包括四个部分: (1) 统一的数据描述模型: 将各种生物信息学数据以统一的数据格式描述; (2) 数据的获取: 通过计算机编程实现各种不同类型的数据库数据的自动获取; (3) 数据管理和数据融合: 对获取的生物信息数据实现智能融合和方便的数据管理; (4) 用户界面: 进行复杂的智能化的数据查询, 同时对整合后的生物信息数据给出更直观更多样化的显示. 数据标准化、数据整合克服了不同的生物数据库数据结构、信息提取过程的不统一等问题, 从而可以更深入的对生物信息数据进行深入的数据挖掘.

211 数据描述标准化

由于生物学数据库种类多样, 各种数据描述方式也存在很大差异, 同时新的生物技术又产生新的生物数据类型, 使数据的描述标准化显得非常重要, 因此人们开始研究生物数据的描述和交换统一标准[9]. 扩展标记语言(eXtended Marker Language, XML) (http://www.xml.com/) 作为一种数据描述语言, 具有模块化、可定制、通用性、简单方便等特点, 在数据标准化方面起着越来越重要的作用. 随着 Internet 的深入发展, 也必将成为生物信息数据交换的标准规范. 在生物信息学领域中, 如 BSML (Bioinformatics Sequence Markup Language) (http://www.BSML.org/), BIOML(Biopolymer Markup Language) (http://www.bioml.com/BIOML/) 等在生物信息学数据的标准化方面已做出有益的探索.

XML 用于生物信息学具有以下优点[10~12]:

(1) 易于定义复杂数据类型: 使用 XML 可以定义不同类型的复杂生物信息学数据类型, 无论是文件类型定义(Document Type Definitions, DTD) 或者大纲(Schema) 都可以定义不同的复杂的数据类型. XML schema 的特性更适用于生物信息学的应用.

(2) XML 的灵便性: 使得容易修改各种复杂的数据类型, XML 的数据与数据类型定义的分离性, 使得增加新的元素或者属性, 只要在 DTD 或者是 Schema 文件中进行修改, 而不需要修改数据本身.

(3) 数据交换: XML 可以成为不同的软件之间交换数据的标准, 生物信息学的分析方法和新软件层出不穷, 通常一个分析方法会使用另一个分析方法的结果作为输入.

(4) 数据整合: 生物信息学数据库采用一套生物信息学的 XML 定义规则, 进行跨数据库查询整合查询后的结果, 将变得更加方便.

(5) 数据互联: XML 的 XPointer(http://www.w3.org/TR/xpath) 和 Xlink(www.w3.org/TR/xlink/) 技术可以提供更好的互联技术, 使得不同对象之间的相互复杂引用和连接变得更加容易. 一个元素可能对应多个引用和连接.

(6) 更容易与数据库融合: 数据库可以直接使用 XML 类型的数据, ORACLE9i (http://www.oracle.com/) 和 SQL Server2000(www.microsoft.com/sql/default.asp) 都提供了内嵌的 XML 支持.

(7) XML 是一个开放的、通用性的语言, 不依赖于任何操作系统.

212 数据整合系统

美国生物技术信息中心(National Center for Biotechnology Information, NCBI) 开发的 Entrez(http://www.ncbi.nlm.nih.gov/entrez/) 检索系统以及欧洲生物信息学研究所(European Bioinformatics Institute, EBI) 开发的序列查询系统(Sequence Retrieve system, SRS) (http://srs.ebi.ac.uk/) 是目前生物信息学研究中使用最频繁的整合信息检索系统. 利用 Entrez 系统, 用户可以方便地检索 GenBank (http://www.ncbi.nlm.nih.gov/Genbank/) 核酸数据, 还可以检索蛋白质序列数据, 三维结构信息, 文献信息, 疾病相关信息, 突变信息, 基因组图谱数据等. Entrez 满足了一定范围内的信息查询要求, 但是 NCBI 也不是包罗万象的万能数据库, 比如信号转导、信号通路等信息, 就不存在 NCBI 所包含的数据库系统, 如果想查找这方面的信息, 还必须到 KEGG (http://www.genome.ad.jp/kegg/kegg.html) 等数据库去查找. SRS 通过一个统一的查询界面, 可以对数百种生物学数据库进行统一的查询, SRS 从各种生物学数据库中查找符合条件的相关信息, 然而 SRS 只提供每个数据库的独立信息, 将所有的相关信息排列显示, 没有对不同来源的数据进行数据整合, 各种类型的数据库依然是作为单一的个体, 没有实现数据的整合.

CIF(Corporate Information Factory)[13] 提出了一个生物信息系统构架结构, CIF 通过从不同的数据源整合生物信息学数据, 使用数据库仓库管理这些信息, 并将这些信息交专家系统进行处理的一个完整的体系, 实现了数据整合、数据标准化, 从而提供了一个企业级管理生物信息数据的方法. ISYS[14] 也提出了一个整合的生物信息数据平台, 由 NCGR 发布的一个基于网络的, 具有即插即用性质的平台, 用 JAVA 语言开发. 其他如 GeneMine (http://www.bioinformatics.ucla.edu/genemine/), BioDataServer[15], BSML, BioNavigator (www.bionavigator.com/), 等在生物信息整合方面做了积极的研究工作, 开发了相应的数据整合系统.

213 多 Agent

Agent(http://agents.umbc.edu/) 起源于人工智能领域, 是包含了诸如知识、信息、承诺和能力等智力状态的实体, 一个封装的有独立功能的模块, 它能够接受和处理其他 Agent 发送来的消息, 也能向其他 Agent 发送消息. 由多个代理按一定结构组织起来的系统称为多代理系统(Multi2Agent System, 简称 MAS). 由于一个典型的 agent 具有下列属性: 自治性、交互性、针对环境性、面向目标性, Agent 还可以具备的其他特性包

括自适应性、通信能力(包括协商、协作)等特性,因此在生物信息数据整合方面也得到了很好的应用. GABAagent[16]系统构建了一个搜索 GABA 受体相关的基因,蛋白序列等相关信息的单 Agent 系统. 基于 DIF2CAF[17]多 Agent 体系框架,Keith Dechker 开发了一个用于疱疹病毒的基因组信息注释和数据整合系统[18]. K. Bryson 开发了基于多 Agent 的 GeneWeaver 系统[19],用于生物信息的数据整合和数据管理.

3 基于 Multi2Agent 的数据整合系统22BioA2gent

本文开发了基于多 Agent 的数据整合系统模型 BioAgent,该系统通过信息搜索 Agent,信息整合 Agent,用户接口 Agent,可定制 Agent 的相互协作实现对生物信息数据库的自动化的数据搜索和数据整合,并在该系统的基础上,开发了人类精神分裂症相关基因的突变信息在多层次的定位数据库. 在数据整合过程中,采用 XML 作为统一的数据描述语言;BioAgent 系统可以动态扩展,可以根据特定 Agent 需求的改变,开发新的信息整合 Agent,实现新的信息整合需求. 系统结构框图如图 1 所示.

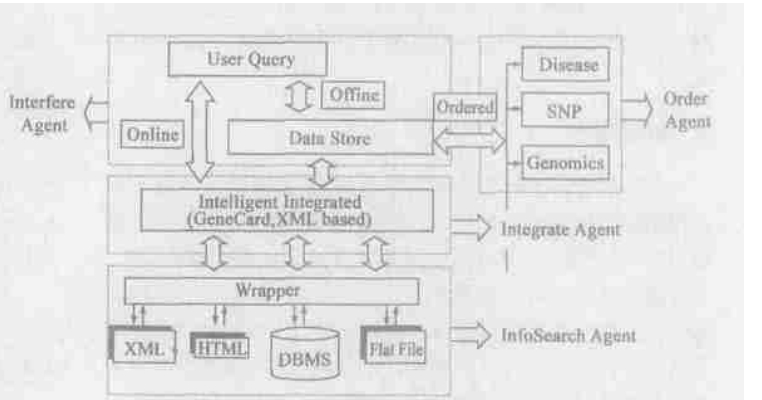

图 1 BioAgent系统结构图

311 信息搜索 Agent:(InfoSearch Agent)

信息搜索 Agent,整合不同类型的、以不同格式存贮的生物信息数据,包括各种类型的公开数据库和本地数据资源,转化为本地数据存贮,实现数据资源的本地化. Wrapper 是获取不同生物数据资源的接口,将不同的生物数据资源,包括超文本文件、关系数据库、文本文件、XML 文件格式等,转化为基于 XML 的数据描述模型[20].

BioAgent 系统采用 Perl(http://www.cpan.org/CPAN.html)编程实现. Perl 语言由于在字符串处理方面的优点,在生物信息学领域得到广泛应用[21]. 信息搜索 Agent 对 GenBank、PDB(http://www.rcsb.org/pdb/)、SwissProt(http://www.expasy.org)、dbSNP(http://www.ncbi.nlm.nih.gov/snp/)、UniGene(http://www.ncbi.nlm.nih.gov/unigene/)、HGVbase(http://hgvbase.cgb.ki.se/)、OMIM、KEGG 等数据库编写了数据获取接口.

312 信息整合 Agent(Integrate Agent):

信息整合 Agent 根据不同类型的数据之间的相互关系,按照特定的数据整合方式,形成一个集中的统一的数据模型,对整合后的数据基于 XML 的生物信息数据模型,存到数据库中,为下一步的数据挖掘提供操作方面的数据源. 经过智能的数据整合,将相关的生物信息数据转化为基于 XML 的统一描述的生物信息数据存贮,克服了不同的生物学数据库数据结构、信息提取过程的不统一等问题,能够更加深入地进行数据挖掘方面的研究[18]. 在生物学数据中还有很多信息是描述性的信息,而不是结构化的信息表示,如何从非规范化的数据描述中得到关键的有用的信息,这也是一个非常关键的研究内容,本系统结合针对文本信息的处理技术,对复杂多样的生物信息描述进行信息的抽取.

在处理复杂的生物信息学资源的时候,将数据信息资源从三个角度来组织数据,实现/以线串点0、/以网串点0、/以面串点0 的立体的信息整合:

(1)/以线串点0:以基因为中心,利用现有的基因卡片(GeneCard)概念[22],将相关的基因染色体定位信息、Gene 的序列和功能、mRNA 的序列和功能、蛋白的序列和功能、表达信息、疾病相关等信息,以该基因的 GeneCard 为主线,从而将该基因相关的生物信息资源以线状的方式进行连接.

(2)/以网串点0:以相互作用关系为中心,以新陈代谢,信号转导中的相互作用,蛋白和蛋白的相互作用为桥梁,将蛋白,酶的相互作用信息按照特定的作用关系组织起来,相互作用关系中的每一个元素都将对应于相应的 GeneCard 信息,从而构建一个网状的生物信息数据资源[23,24].

(3)/以面串点0:以进化为中心,将不同物种看作是生物进化的不同的横断面,不同的物种所对应的同源基因则构成了不同面上的点对应,两个将不同物种的同源基因构建不同的进化同源关系,通过进化主线,以不同的物种作为一个面,从而产生了一个立体状的生物信息资源组织图[25].

313 用户接口 Agent:

用户接口 Agent 设计用户使用界面,处理复杂灵活的输入,给出数据整合后的查询结果,提供图表式的显示方式. 对查询请求的处理,可以分为在线和离线两种方式:(1)在线处理,对输入进行语义分析处理,按照用户的需求在相关的生物信息学数据库中进行数据整合;(2)离线处理,对输入的处理同上,不同处在于该输入请求是否已经位于已经整合的数据资源中,如果是已经整合的信息,就直接的调用该信息.

314 信息定制 Agent:

可以针对不同目的的生物信息数据定制功能研究,制定不同的数据搜索、数据整合策略. 新的定制需求可能会需求新的数据源,那就需要增加相应数据库的信息搜索 Agent 的 Wrapper;如果不需求新的数据源,只是对已有的数据资源以新角度进行研究,则需要改变信息整合 Agent 的相应功能,以实现新的数据整合目的. 信息整合的可定制,这也是信息整合的一个发展方向,希望更具有人性化,更能满足个性的要求. 实现不同需求的定制搜索,对用户的研究需求进行深入分析,根据用户的研究需求,可以自定制一些特定需求的数据模块,这些数据模块可能包括不同的侧面:基因组侧面、蛋白质组侧面、模式生物、疾病研究等. 不同的数据模块可以用于不同的

研究目的, 为该研究提供尽可能多的知识, 给下一步的基于生物信息数据的数据挖掘提供更好的数据基础.

4 系统特点

BioAgent 是一个基于多 Agent 相互协作的生物信息搜索、智能数据整合的生物信息数据整合系统, 系统的设计基于开放的, 模块化的, 跨平台的设计思想, 本系统具有以下特点:

(1) 跨平台性: 基于开放性、扩展性、兼容性、易操作性等原则, 采用了 Apache+ Mysql+ XML + Php+ Perl 的体系结构, Apache(http: / / www. apache. org) 是基于 Gnu(http: / / www. gnu. org) 的 Webserver, Mysql(http: / / www. mysql. com) 是基于 Gnu 的数据库, Php(http: / / www. php. net) 是基于 Gnu 的程序语言, 通过这样的组合, 实现系统的跨平台, 可以在 Windows、Linux、Unix 等系统下安装运行.

(2) 自动化: 可以自动去搜索新的数据资源, 并可及时更新已有的数据, 可以根据一个基因的名称或 NCBI 的 LocusLink 号(http: / / www. ncbi. nlm. nih. gov/ locuslink/) , 就自动获取该基因相关的所有信息.

(3) 可定制性: 可自行修改数据获取的数据源, 也可以通过编程接口, 编程实现新的功能.

(4) 标准化: 采用统一的基于 XML 的数据描述模型, 将生物学数据用统一的 XML 语言描述.

(5) 模块化: 本系统基于 Multi2Agent 设计, 因此具有非常好的模块化, 便于新的功能的添加.

(6) 多角度的数据透视: 提供了多种角度对数据资源进行描述.

(7) 良好的扩展性: 可以方便的添加新的数据获取 Agent 中的数据库 Wrapper 接口, 也可以方便地添加数据整合 Agent 的新功能.

5 应用实例2基于突变信息的数据整合

在基于 BioAgent 系统的基础上, 开发了人类精神分裂症相关基因的突变信息整合数据库2VSD[26] (http: / / www. chgb. org. cn/ vsd. htm) , 实现了突变信息在基因组层次, 转录层次, 蛋白序列, 二级结构, 三级结构, 以及新陈代谢, 信号转导, 疾病相关等全方面的定位,/ 以线串点, 以网串点0, 并且将 SNP 在不同层次的定位都以图示方式给出, 提供了非常友好的用户界面.

信息获取 Agent 以 Perl 编程, 获取包括 GenBank、KEGG、Locus、UniGene、dbSNP、HGVbase、OMIM、SwissProt、InterPro 等数据库的相关信息. 信息整合 Agent 以突变信息在不同层次的定位为主线, 对各个方面的数据进行了数据整合, 以 dbSNP、HGVbase、OMIM 数据为基础, 将 SNP 在基因, 转录 mRNA 的位置进行定位, 以及非同义 SNP 对应的蛋白位置; 结合 SwissProt 中蛋白的 variant 和 feature 信息, 将非同义 SNP 在 Swissprot 中蛋白的 feature 所在位置进行定位; 结合综合二级结构数据库 InterPro, 将非同义 SNP 在 SwissProt 中蛋白的不同二级结构域进行定位, 见图 2; 结合 PDB 中的三级结构信息, 将非同义 SNP 在三级结构信息中进行定位; 同时结合 OMIM 中的突变信息; 结合新陈代谢信息, 将突变信息在网络层次定位. 用户接口 Agent 提供良好的用户使用界面, 可以查询基因、蛋白、酶、OMIM、染色体分布、SNP 等相关信息, 并且这些信息提供丰富的相互联接.

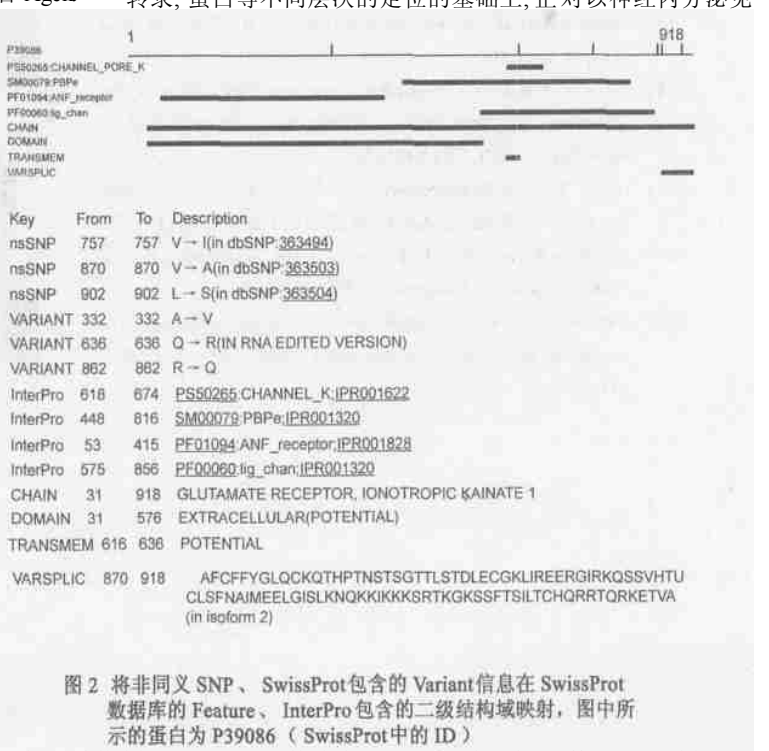

图 2 将非同义 SNP、SwissProt包含的 Variant信息在 SwissProt 数据库的 Feature、InterPro包含的二级结构域映射, 图中所示的蛋白为 P39086 (SwissProt中的 ID)

6 结论

数据整合是生物科学、计算机科学、信息科学之间学科交叉的前沿课题, 通过搜索技术、自然语言理解技术、计算机技术实现智能化地用于基因、蛋白质功能的专业化的生物信息数据整合. 目前的生物数据查询系统, 基本上都是提供简单词语的逻辑组合查询, 而且获取的信息比较单一, 因此往往满足不了人们的查询要求, 人们需要更加灵活的输入方式, 希望计算机能够提供更多的综合信息, 并且提供更好的界面显示. 对这些复杂多样的生物学数据资源进行智能的数据整合, 有助于我们更加深入地了解模式生物的基因组、蛋白质组信息和内部的相互作用机制; 可以更好地帮助生物医学工作者快速寻找已有的各种类型的生物学数据, 给他们提供整合后的全方位的相关信息, 使种类多样的生物数据进行优化整合、系统化, 形成一个综合的数据平台, 从而可以节省大量数据转换工作, 能更方便地应用数据挖掘技术和一些信息学中的方法加以分析. 因此生物信息学的数据整合研究成果, 能够促进数据的共享, 加快生物信息学的发展.

本文开发了基于 Multi2Agent 的生物数据整合系统2BioAgent, 并且在该系统基础上, 实现了精神分裂症相关基因的突变信息的数据整合. 目前该系统在实现了突变信息在基因组, 转录, 蛋白等不同层次的定位的基础上, 正对以神经内分泌免

疫网络为中心的调控网络的相关数据资源进行数据整合.

参考文献:

[1] Wolfsberg TG, et al. A user. s guide to the human genome[J]. Nature Genetics, 2002, 32(1): 1- 79.

[2] Discala C, et al. DBcat: a catalog of 500 biological databases[J]. Nucleic Acids Research, 2000, 28(1): 8- 9.

[3] Robbins RJ. Report of the invitational DOE workshop on genome informatics i: community databases[J]. Journal of Computational Biology, 1994, 1(3): 173- 190.

[4] Reichhardt T. It. s sink or swim as a tidal wave of data approaches[J]. Nature, 1999, 399(6736): 517- 520.

[5] Bishop MJ, et al. DNA and protein sequence analysis: a practical approach[M]. Oxford: IRL Press, Oxford University Press, 1997, 11- 13.

[6] 李衍达. 信息与生命[J]. 化学通报, 2001, 10: 601- 608.

[7] Martin AC. Can we integrate bioinformatics data on the Internet[J]. Trends Biotech, 2001, 19(9): 327- 328.

[8] Hubbard T. Biological information: making it accessible and integrated (and trying to make sense of it)[J]. Bioinformatics, 2002, 18(Suppl 2): 140- 148.

[9] Barillot E, et al. A proposal for a standard CORBA interface for genome maps[J]. Bioinformatics, 1999, 15(2): 157- 169.

[10] Fr d ric Achard, et al. XML, bioinformatics and data integration[J]. Bioinformatics, 2001, 17(2): 115- 125.

[11] Gilmour. Taxonomic markup language: applying XML to systematic data[J]. Bioinformatics, 2000, 16(4): 406- 407.

[12] Emmanuel Barillot, et al. XML: a lingua franca for science[J]. Trends Biotechnol, 2000, 18(8): 331- 333.

[13] Inmon WH, et al. Corporate information factory[M]. New York: John Wiley & Sons, 2001.

[14] Siepel A, et al. ISYS: a decentralized, component-based approach to the integration of heterogeneous bioinformatics resources[J]. Bioinformatics, 2001, 17(1): 83- 94.

[15] Freier A, et al. BioDataServer: a SQL-based service for the online integration of life science data[J]. In Silico Biol, 2002, 2(2): 37- 57.

[16] Abdelkrim Rachedi, et al. GABAagent: a system for integrating data on GABA receptors[J]. Bioinformatics, 2000, 16(4): 301- 312.

[17] John RG, et al. DECAF-A flexible multi agent system architecture[J]. autonomous agents and multi-agent systems, 2003, 7(1- 2): 102- 113.

[18] Keith Decker, et al. A multi-agent system for automated genomic annotation[A]. Proceedings of the Fifth International Conference on Autonomous Agents[C]. Montreal, Quebec, Canada, 2001, 433- 440.

[19] Bryson K, et al. Agent Interaction for Bioinformatics Data Management[J]. Applied Artificial Intelligence, 2001, 15(10): 917- 947.

[20] Lacroix Z. Biological data integration: wrapping data and tools[J]. IEEE Trans Inf Technol Biomed, 2002, 6(2): 123- 128.

[21] Jason E, et al. The bioperl toolkit: perl modules for the life sciences[J]. Genome Research, 2003, 12(10): 1611- 1618.

[22] Rebhan M, et al. GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. Bioinformatics[J]. 1998, 14(8): 656- 664.

[23] Minoru Kanehisa, et al. The KEGG databases at genomeNet[J]. Nucleic Acids Research, 2002, 30(1): 42- 46.

[24] Ananko EA, et al. GeneNet: a database on structure and functional organization of gene networks[J]. Nucleic Acids Research, 2002, 30(1): 398- 401.

[25] Tatusov RL, et al. The COG database: new developments in phylogenetic classification of proteins from complete genomes[J]. Nucleic Acids Research, 2001, 29(1): 22- 28.

[26] Min Zhou, Yonglong Zhuang, et al. VSD: A database of schizophrenia related genes focusing on variations. (Joint authors). Human mutation, 2004, 23(1): 1- 7.

作者简介:

庄永龙 男, 1977 年 6 月出生于江苏省, 1999 年获清华大学自动化系工学学士学位, 现于清华大学自动化系攻读模式识别与智能控制专业博士, 研究方向为生物信息学. Email: zhuangyl99@mails.tsinghua.edu.cn.

李衍达 男, 1936 年 10 月出生于广东省, 自动化系教授, 中国科学院院士, 信号处理与智能控制专家, 国务院学位委员会委员, 研究方向为: 信号处理理论、生物信息学. Email: daulyd@mail.tsinghua.edu.cn.