第 34 卷 第 8 期 计 算 机 工 程 2008 年 4 月
Vol.34 No.8 Computer Engineering April 2008
·博士论文· 文章编号：1000—3428(2008)08—0038—03 文献标识码：A 中图分类号：N945

语义异构生物数据源中的数据集成与更新

杨 森[1]，夏 燕[1]，曹顺良[2]，邓绪斌[1]，朱扬勇[1,2]
(1. 复旦大学上海(国际)数据库研究中心，上海 200433；2. 上海生物信息技术研究中心，上海 200235)

摘 要：针对生物数据源的分布性、异构性和动态性等特性，探讨生物信息技术服务支撑系统整体解决方案，构建基于基因本体的信息集成模式以实现生物语义学上的数据集成。设计一种以半结构化形式规范生物元数据及基于 MD5 算法的增量更新技术，用以解决通用扩展性和效率问题，实现生物数据仓库中数据的共享并提高管理效率。

关键词：基因本体；半结构化；增量更新；MD5 算法

Data Integration and Update in Semantic Heterogeneous Biological Data Sources

YANG Sen[1], XIA Yan[1], CAO Shun-liang[2], DENG Xu-bin[1], ZHU Yang-yong[1,2]
(1. Shanghai (International) Database Research Center, Fudan University, Shanghai 200433;
2. Shanghai Center for Bioinformatics Technology, Shanghai 200235)

【Abstract】 For the characters of distribution, heterogeneity and dynamic of biological data, a resolution of the service system for bioinformatics technology is presented, and an approach of biological data integration based on Gene Ontology(GO) is proposed in order to realize biological semantic integration. Semi-structured incremental updating method to standardize biological metadata with MD5 algorithm to improve the updating efficiency is designed, which resolves the data sharing and the efficiency of data management in biological data warehouse.

【Key words】 Gene Ontology(GO); semi-structured; incremental update; MD5 algorithm

随着基因组测序工作的蓬勃发展以及基因芯片等技术的快速发展及普及应用，生物数据呈指数级增长。由于生物数据源具有分布性、异构性等特性，因此从多个生物数据源中获取生物信息变得愈加困难。

1 生物信息技术服务支撑系统体系架构

近年来，为了应对生物技术高速发展而引发的数据存储、分析等的应用需求，各种依生物学需求而开发的数据库不断涌现。这些在不同应用背景下开发的数据管理系统存在着异构性、自治性、重叠性、管理复杂性增大等问题，消除上述问题需要构建一个面向生物数据的可扩充平台来整合异构生物数据源。

生命科学的发展不断产生类型众多的海量数据，不断倍增的生物数据对数据管理系统提出更高要求，即如何快速响应生命科学研究手段的发展、及时管理并整合这些生物数据。针对海量生物数据应制定相关数据提交流程、开发数据浏览及访问工具，为用户提供一个统一的访问接口，实现多种类型数据的便捷访问，并依照一致化风格展现生物数据。共享是生物数据的另一个重要特征。数据管理系统不仅应提供交互型查询、浏览功能，还应提供灵活的数据下载及批量访问等功能。除了通过 HTTP 等方式下载完整数据集外，还可以使用 Web Service 接口实现批量访问，这些工具都能提高数据间的共享。

针对生物信息技术领域的应用需求，图 1 给出一个面向生物数据的生物信息技术服务支撑系统总体架构。系统采用基于B/S的*N*层体系架构，分为资源层、平台支撑层、平台服务层、系统门户层及用户界面层。并使用了基于基因本体(Gene Ontology, GO)[1]的生物数据整合模式，采用基于GO集成模式消除概念、术语间的混乱，拟合具有不同应用背景知识的数据管理系统建立者之间的理解差异，构建数据通信、共享、互操作及集成的基础。

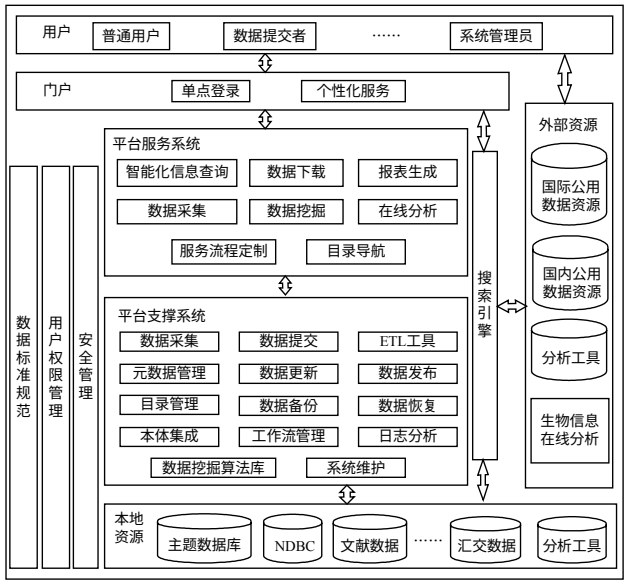

图 1 生物信息技术服务支撑系统总体架构

基金项目：国家自然科学基金资助项目(60573093)；上海市重大科技项目(02DJ14013)

作者简介：杨 森(1968－)，男，博士后，主研方向：数据挖掘，生物信息学；夏 燕，硕士研究生；曹顺良，博士；邓绪斌，博士研究生；朱扬勇，教授、博士生导师

收稿日期：2007-05-20 **E-mail：**syang0755@eyou.com

另外，系统提供在线和离线 2 种服务方式：对于需要大量计算资源的BLAST等应用，采用在线服务；对于可借助下载免费工具的应用，通过FTP方式实现离线服务。系统的实现采用数据仓库体系结构[2]，构建面向分布式异构生物数据源的数据集成系统，提供数据抽取、分析等工具，其体系结构如图 2 所示。

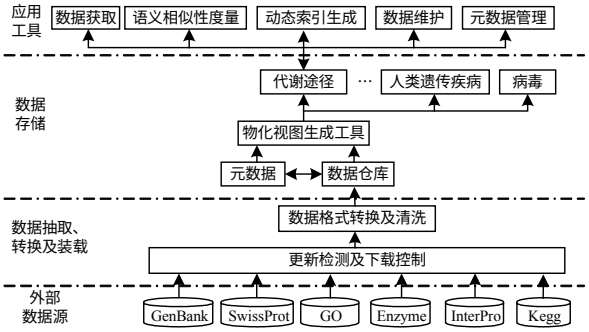

图 2　生物数据集成系统的结构

2　基于 GO 的生物数据集成

许多生物数据以半结构化数据形式存在，生物数据源又具有异构性、动态性，对其有效管理须应对以下问题：(1)生物数据的动态性造成概念关系的复杂性，使得无法预先完全定义概念间的相互关系；(2)同一个概念从不同数据源中可得到完全不同的属性信息；(3)很难从数据源中获悉复杂关系属性信息。为此，采用基于Ontology的方式实现异构生物数据源之间的整合。“ Ontology是对共享概念体系进行明确的形式化规范说明 ”[3]。GO已成为生物应用领域的标准[4]，为实现生物知识共享和重用提供了形式化的知识表示、明确的领域词汇及语义、完整的领域模型以及对该领域的共同理解。

对于图 2 中的外部生物数据源，以 GO 为黏合剂将 GenBank 的核酸及 DNA 等异构生物数据源中的数据集成到一个统一数据模式的生物数据仓库中，其中，对于外部数据源，定义一个范式以导入数据；对于主题数据库，设计一个实体视图，从多数据源中抽取数据。异构生物数据通过所构建的 DB2GO, DBREF 和 GO Similarity 表关联起来的，关联关系如图 3 所示。

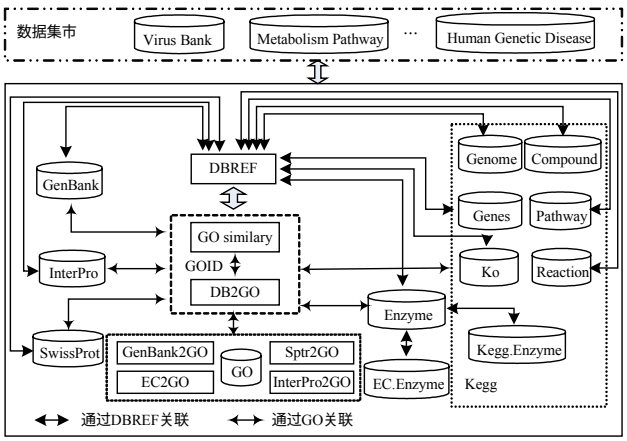

图 3　生物数据仓库中各数据库关联图

(1)DB2GO 记录了 GO 和不同数据库条目之间的关联关系，通过 Gene-ID 将不同数据源中数据项关联起来，较好地解决了数据源间语义异构问题。

(2)将多个外部数据源中的交叉引用关联信息集中存储在 DBREF 表中，实现多数据库之间的快速查询与调用。

(3)通过记录 GOID 的语义及其路径信息,实现 2 个 GOID 或 2 个 GO 术语之间的快速比较。GO Similarity 表用于存储语义相似性结果，实现语义相似性快速检索。

DB2GO 表从基因产物的注释信息角度建立不同数据源之间的数据联系；DBREF 表采用序列比对方式，从基因产物的序列信息角度建立不同数据源之间的数据联系；GO Similarity 表从基因产物间语义相似性的角度建立数据库之间的数据联系，由此实现异构生物数据源之间生物语义学上的数据集成。因此，可以设计浏览 GO 模式等多种基于 GO 的语义查询方式，扩充生物数据分析手段。借助全局模式 Ontology 所定义的概念之间的语义关系可推导出隐含的语义信息，实现带有一定逻辑推理功能的智能查询。

3　集成系统中的生物数据更新

生物数据多半是半结构化数据，是一种结构隐含、无规则、不严谨的自描述型数据[5]，除了具有隐含模式信息、不规则结构及弱类型约束等通常的半结构化数据特性外，还有其自身特点：(1)生物数据源数目繁多、格式类型众多，一种数据源就可能用一套元数据描述；(2)生物数据文件是由标签和相关值构成的具有一定规则的序列，多半采用非标准标签标记的纯英文文本文件；(3)一个文本文件通常由记录构成，记录与记录之间由标签分割，不同的数据源所采用的标签集及其语义不同；(4)每一条记录由多个具有特定意义的对象构成，记录中的对象具有层次性，对象与对象之间的次序相对固定；(5)数据对象往往具有多层嵌套结构，对象成分可能存在缺失、多次重复、无序等情况。虽然半结构化文本格式能够表示复杂、具有层次结构的信息，但存在缺乏索引机制、不支持多用户并发操作等问题，因此，需要一种适用于生物数据的数据模型。

定义 1 元数据是描述数据及其环境的数据。

在异构生物数据源集成系统进行数据更新过程中，生物元数据是指生物数据源的结构及位置、增量控制信息、数据更新目标位置、更新日志等。生物数据源数据更新可分为：(1)数据变化，而元数据不变；(2)数据和元数据均发生变更。

定义 2 对于那些无法用全局一致、上下文无关的文法描述的数据，若存在一个有序划分，可以从前 i 个部分的信息中推导出第 i+1 部分的信息，这种局部一致的文法称为半结构化文法。

定义 3 给定节点集 N，设 Σ 表示 N 中节点的命名集合，Q 表示 N 中节点的值集合，则 Σ 上的元树是一个三元组 $T(t,f,v)$，其中，t 为一棵树，t 中的节点均属于 N；f 为节点命名函数，其将 t 中的节点与 Σ 中的名字映射起来；v 是节点的取值函数，其将树中的节点与 Q 中的值对应起来。

定义 4 Σ 上的元树类型为一个三元组 Γ (Σ, tr, g)，其中，tr 为一棵树；g 为 Σ 中的元素分类函数，对于任意 $\alpha \in \Sigma$，函数 $g(\alpha)$ 表示 α 所属类别。将 Γ 中的类别集合记为 Σ'，则 tr 中的节点均在 Σ' 中。若元树 t 可用 Γ 抽象表示，则记为 $t \models \Gamma$，并将所有可用 Γ 抽象表示的元树记为 $der(\Gamma)$。

给定元树 t 及元树类型 Γ $(t \models \Gamma)$，对于元树类型中的一个节点，节点标记 w 表示该节点的重复次数，其表示符为 1,+, *。对于任意 $c \in \Sigma'$，c 的 w 次重复表示为 $c^w = \alpha_1, \alpha_2, \cdots, \alpha_n$，其中的 $\alpha_i (i = 1, 2, \cdots, n)$ 均为元树 t 中的节点，属于类别 c，w 的表示意义为：$w = 1$ 表示类别 c 在元树 t 中有且仅有一个对应

节点；$w=+$ 表示类别 c 在元树 t 中至少有一个对应节点；$w=*$ 表示类别 c 在元树 t 中的对应节点数没有限制。

定义 5 若给定元树类型 $\Gamma=(\Sigma,tr,g)$ 和 tr 中的 2 个节点 c_1，c_2，且 $c_1\neq c_2$，则称三元组 $(c_1,cond,c_2)$ 为 Γ 上的一条规则，其中的 $cond$ 表示一个条件，即当满足条件 $cond$ 时，c_1 引发 c_2。

采用数据仓库模式来整合生物数据存在着数据更新不及时等问题，因此，使用半结构化形式来规范生物元数据，以增量更新方式实现生物数据的即时更新。外部生物数据源的增量文件可分为全量增量文件、增量增加文件(包括纯增量、积累增量)2 类。

在图 2 中，从 GenBank 等外部数据源中抽取的数据存放在关系数据表中，如 GenBank 的抽取数据存放在 Main 表等关系表中，其中的字段类型可为 key, Normal 和 Sequence 等。Genbank 的 Main 表的结构如下：

AC	EntryName	Length	…	Flag	Sequence

图 4 给出了描述数据源抽取结果的结构元树。

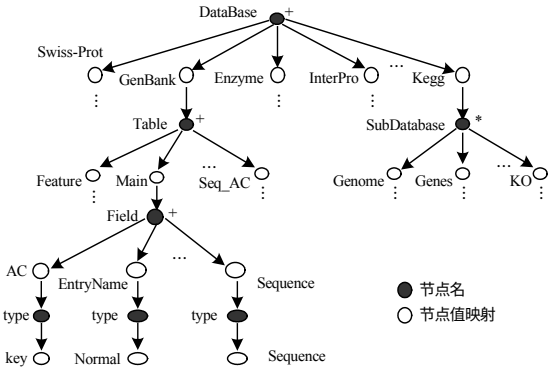

图 4　生物数据源抽取结果结构元树

在处理易变生物数据源时，使用半结构化形式规范生物元数据，实现节点的动态增添、分裂或合并等操作：

(1)增添新数据源时，只须依据元树类型节点 w 的取值情况，在相应元树的适当位置处添加新节点即可，无须变动元树类型。

(2)若现存的生物数据源结构或抽取结果结构发生变化，元树类型中的节点则随生物数据源结构或抽取结果结构的变化进行分裂、合并，而元树类型结构不会发生显著变化。在图 5 中，若数据源只有纯增量文件，则删除节点 DB_Source_Inc 下的节点，再合并节点 DB_Source_All 和节点 DB_Source_Inc 即可。

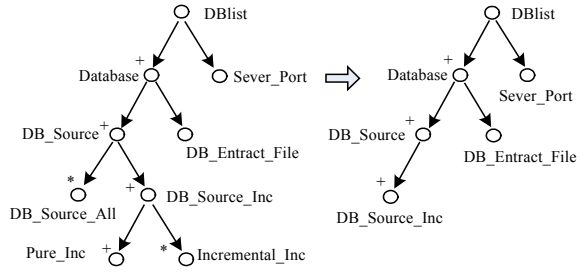

图 5　生物数据源位置及结构元树变迁图

(3)当数据源的更新频率发生变化时，依据元树类型节点上所设定的阈值(如 $\|\alpha_i-\alpha_j\|<\varepsilon$)，对自动更新的探测频率加以调整。

在复杂、异构生物数据源数据更新过程中，需要进行大量的增、删、改操作，时间、空间消耗极大。因此在更新过程中，使用MD5 算法[6]减少对所传递消息中数据变更与否的判定及校验，结合所传递消息中的特征位信息来加快生物数据的更新过程。生物数据更新过程的实现流程如下：

(1)自动下载模块 Autoftp 将最新增量下载到指定目录；

(2)下载结束后，主控进程 Mainctrl 启动数据更新进程；

(3)Wrapper 向 Updatae_Server 发送包含 MD5 签名的更新消息；

(4)Updata_Sever 端口捕获消息至 msg；

(5)　While(允许增量修改){

(6)　　从 msg 中解析出<数据库名，数据库记录列表>；

(7)　　If(该数据库允许修改){

(8)　　获取该 Entry 的 MD5 及关键字 key；

(9)　　校验对应数据库的 MD5 表；

(10)　　根据关键字，MD5，删除标记，对该条目进行修改、插入等相应操作；}}

(11)关闭 Updata_Server 端口的监听装置。

其中，Autoftp 下载的增量文件包括纯增量文件、累积增量文件和全量增量文件。纯增量文件直接送入包装器 wrapper 进行数据抽取，并向数据更新服务器逐条传送已抽取的条目。对于累积增量或全量文件，则用前后增量数据快照比对生成纯增量文件。

用半结构化形式来规范生物元数据，可以较好地应对生物数据源的多变性，在不改变元生物数据基本结构的情况下实现高效应对，并具有良好的扩展性和灵活性。

4　结束语

生物数据源的分布性、异构性、动态性等特性决定了生物数据管理的复杂性。本文面对海量生物数据集成问题，探讨了构建一个生物信息技术服务支撑系统的整体解决方案；在生物语义学一致性上，建立起基于 GO 的生物数据整合模式，实现异构生物数据源之间数据关联及集成；针对生物数据源的动态性、易变性，提出一种基于半结构化形式规范生物元数据及运用 MD5 算法的生物数据增量更新技术，较好地解决了数据更新问题。本文为生物研究者提供了一个高效的数据集成平台，以期达到数据共享、提高效用的目的。

参考文献

[1] Ashburner M, Ball C A. Gene Ontology: Tool for the Unification of Biology[J]. Nature Genet, 2000, 25(1): 25-33.

[2] Sperley E. The Enterprise Data Warehouse: Planning, Building and Implementation[M]. [S. l.]: Prentice Hall PTR, 1999.

[3] Borst W. Construction of Engineering Ontologies[D]. Enschede, Holland: University of Twente, 1997.

[4] Martucci D, Masseroli M. Gene Ontology Application to Genomic Functional Annotation, Staticstical Analysis and Knowledge Mining[M]. [S. l.]: IOS Press, 2004: 108-131.

[5] Suciu D. An Overview of Semistructured Data[J]. SIGACT News, 1998, 29(4): 28-38.

[6] Rivest R. The MD5 Message Digest Algorithm[S]. RFC 1321, 1992.