核酸序列数据库上的联机分析处理方法*)

王 敞 陈增强 袁著祉
(南开大学信息技术科学学院 天津300071)

Method of Online Analytical Processing on Nucleotide Sequences Database

WANG Chang CHEN Zeng-Qiang YUAN Zhu-Zhi
(Institute of Information Technology Science, Nankai University, Tianjin 300071)

Abstract This paper proposes a new method of Online Analytical Processing on EMBL Nucleotide Sequences Database. This scheme is used to automatically restore flat file data into relational database, which is then converted into OLAP's data marts. Both the quality and speed of analysis will be greatly improved based on the data marts. We believe that this method is a powerful and flexible tool and can be seen as successful application of data mining in molecule Biology.

Keywords OLAP, Nucleotide sequences database, Data warehouse, Bioinfomatics, Data marts

1. 前言

2000年6月26日,被誉为生命"阿波罗计划"的人类基因组计划,终于完成了工作草图,它预示着完成人类基因组计划已经指日可待,也预示着基于序列的生物学时代已经到来。截止目前为止,仅登录在欧洲生物信息中心 EMBL 核酸序列数据库中的序列总量就已接近200亿碱基对。与其同步的还有数据库中蛋白质数目的增长。例如,在 SWISS-PROT 蛋白质数据库中的序列现在已经接近了11万个,此外,还有一万多种蛋白质的空间结构以不同的分辨率被测定。在这类基础数据之上派生、整理出来的数据库更是高达几百个,这一切已经构成了一个生物学数据的海洋。与正在以指数方式增长的生物学数据相比,人类相关知识的增长却相对缓慢。数据并不等于知识,为了从大量生物学数据中挖掘出知识就需要综合运用数学、计算机科学和生物学的各种工具,来阐明和理解大量数据中所包含的生物学意义。

我们针对目前被广泛引用的 EMBL Nucleotide Sequence 数据库,提出了针对它的联机分析处理(OLAP)方法。我们的方案可以被认为是将高级数据分析处理技术应用于生物信息学中的一次成功探索。

2. 核酸序列数据库

EMBL 核酸序列数据库存储了欧洲范围内最全的核酸序列信息。这个数据库是由欧洲生物信息学研究所维护的。该数据库的信息来自基因序列中心,科学家个人,欧洲专利局,还有通过与日本 DDBJ 数据库,美国 GenBank 数据库信息交换得到的数据。上述三个数据库是目前世界上该领域最权威的数据库,为了保持三个库之间的同步,每天这三个数据库都要进行信息的更新。这三个数据库还遵守共同的规范来规定信息条目的内容和格式。这种规范保持了三个数据库的兼容,也保持了这三个数据库对生物信息学软件的兼容性。EMBL 核酸序列数据库目前存储了含有接近200亿个碱基对的1700万条核酸记录信息,这个数目仍在迅速增长。

EMBL 核酸序列数据库的数据是通过文件的形式发布的。文件遵循固定的格式,在每行开始,都有特殊字段标记该行的数据类别。这种文件由简单的英文写成,可以被人直接阅读,也可以通过序列分析软件读取。该文件包含的数据可以分为以下四类:

1. 标志性和描述性信息 包括条目名称,分子类型,分类,序列长度,查找号,序列标识,序列版本,入库日期,最后改动日期,描述信息,关键字信息,分类法信息和到相关数据库的链接等。

2. 引用信息 包括具体的引用细节(相关文献的作者,题目,出处,评论,还有序列提交者的相关信息)。

3. 特征信息 对序列数据进行注解。描述核酸序列本身的特征信息和更详细的注解信息。

4. 序列信息 包括序列长度,序列的组成成分,序列校验结果和序列数据本身。

(关于库结构的更进一步信息可以参阅文[5,6])

随着分子生物学的发展,现在的生物学研究正在逐渐转向基于序列的生物学研究模式。大量的跟序列有关的操作已经构成了新生物学的核心。这种序列包括基因序列,核酸序列,氨基酸序列等。对于如此大规模的序列集合,单纯的依靠简单的文件形式存储显然满足不了分析的需求。各种对数据的复杂操作至少要依赖于关系数据库系统的存储模式。更进一步,随着数据库技术的不断发展,单纯的关系数据库在数据集成,复杂数据查询和数据分析等方面的缺点也逐渐显现出来。这样将更适合于复杂查询和分析操作的联机分析处理(OLAP)应用于生物序列数据库中,就成了这些数据库发展的必然趋势。

3. 联机分析处理(OLAP)概述

OLAP 是一项经常与数据仓库一起讨论的技术,它提供了一种信息系统结构使得对数据的访问非常灵活。它还可以用多种方法对数据进行切片、分割,动态地考察汇总数据和细节数据的关系。

*)本文得到国家自然科学基金(60174021)资助。王 敞 硕士研究生,主要研究领域为数据挖掘,生物信息学,人工智能。陈增强 教授,博士生导师,主要研究领域为智能控制。袁著祉 教授,博士生导师,主要研究领域为智能控制。

OLAP 与传统的联机事务处理(OLTP)有很大区别,OLAP 主要是用于高层的数据统计分析或者复杂的、汇总型的数据访问,它提供汇总和聚集机制,在不同的粒度级别上存储和管理信息。OLAP 通常采用星型模型进行多维数据集数据的组织,对数据采取只读操作。OLAP 的出现使得一些在传统数据库上难于进行的复杂操作变得简单高效,它被认为是目前众多数据处理技术中的前沿。

OLAP 是基于多维数据集合模型(数据集市)的。该模型将数据看作数据立方体(Data Cube)形式。数据立方体允许以多维形式对数据建模和观察。立方体由维和事实定义。一般地,维是一个关于想要记录的透视。每个维有一个表与之相关联,该表称为维表,它进一步描述维。维表可以由用户或专家设定,或者根据数据分布自动产生和调整。通常,多维数据模型围绕中心主题组织,该主题用事实表表示。事实是以数值度量的。事实表包括了事实的名称和数值,以及每个相关维表的关键字。

4. 核酸序列数据库上的 OLAP

由于核酸序列的信息是以文件格式(Flat File)发布的,因此要想对核酸序列数据库进行联机分析处理,首先就要把以文件存储形式发布的数据库数据转存到关系数据库中。然后,才能以关系数据库为基础,建立多维数据集合(数据集市),并进行 OLAP 操作(过程如图1所示)。

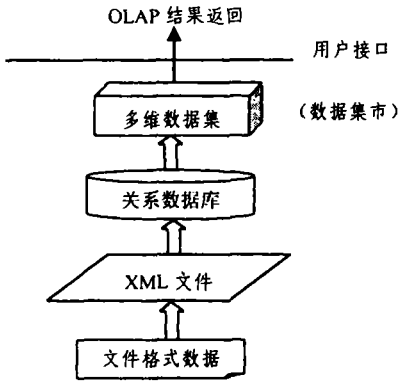

图1 核酸序列数据库上的 OLAP 实现方案

4.1 将文件格式转存成关系数据库格式

4.1.1 *库结构* 进行 OLAP 的第一步是把原来的文件存储形式的序列信息转存到关系表中。为此,我们需要首先设计关系数据库的结构。我们参考了大量相关文献和数据库中存储信息的生物学背景,结合多维数据集的主题和实现上的细节信息提出了我们的库结构。我们的关系数据库包含了 Main 表, Properties 表, Reference Paper 表, Reference Database 表,Feature 表,Comment 表,Date 表,Sequence Data 表,Organism Class 表,共9张表,这些表完全包含了原来 EMBL 核酸数据库的全部数据。

4.1.2 *将原有文件形式的数据信息导入到关系数据库中* 在这里,我们参考了文[3]中所提出的方法。即在文件和数据库中间,提供 XML 格式的文件,作为一个过渡。原来文件的数据先导入 XML 文件中,再利用数据库提供的接口把 XML 文件里面的信息导入到关系数据库中。

4.2 把关系数据库数据存储到多维数据集

4.2.1 *多维数据集合的组织* 我们把 Main 表作为多维数据集的事实表,维度表有四个,分别是 Properties 表, Reference Paper 表,Date 表和 Organism Class 表。各维度表与事实表靠外键建立关系。维度采用星型结构创建(如图2所示)。数据的存储采用混合 OLAP 方式,这种方式将数据保留在关系表中而将聚合存于数据立方体内。

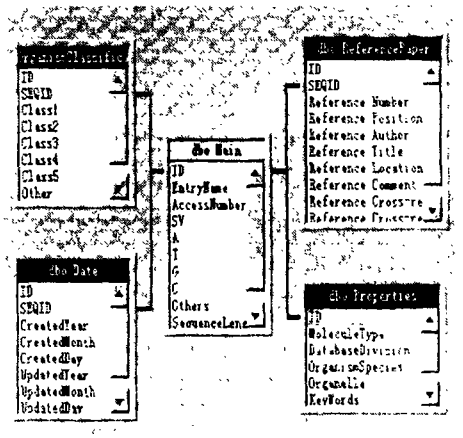

图2 多维数据集的星型架构

4.2.2 *联机分析处理实现的功能* 这里,我们主要着眼于对数据库中核酸序列的分布规律和每类核酸的大体组成成份的分析。因为 EMBL 核酸序列数据库含有近200亿个碱基对,所以用常规方法要想得到全面的关于核酸分布规律和每类核酸组成成份的统计分析结果是很困难的。我们通过适当建立维度和度量,利用 OLAP 可以很方便和快捷地完成这类任务。

4.3 核酸数据库上的 OLAP 实现

我们用微软 SQL Server 2000 OLAP Services 实现了我们的方案。在实践中,我们设定了 Molecule Type, Database Division, Organism Species, Organelle, Key Words, Reference Location, Created Date, Updated Date, Organism Classification 共9个维度。每个维度下面又分若干级别,每个级别又包含各自数目不同的成员,这构成了一个层次的结构。例如,Database Division 只包含一个级别,这个级别含有 EST, PHG, FUN, HTC 等17个表示子类的成员。而对于 Created Date 维度则包含三个级别,分别对应着年,月,日。我们选用核酸序列的个数累加值,最大序列长度,最小序列长度,Adenine、Guanine、Cytosine、Thiamine 四种碱基在序列中的数目累加值,最大数目,最小数目共十五个变量作为事实表的度量值。这种维度与度量的选择基本可以满足大部分复杂的统计分析要求。

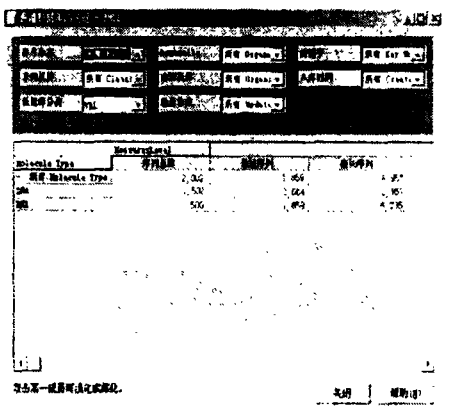

图3 联机分析处理界面

空间分析组件的研究与设计

杨占龙　汪林林
（重庆邮电学院　重庆400065）

Research and Design about Component of Spatial-Analyzer

YANG Zhan-Long　WANG Lin-Lin
(Chongqing University of Posts & Telecommunication, Chongqing 400065)

Abstract　Components GIS(ComGIS), a new technology based on Components Software, is a mainstream in GIS technology. In this paper, the phases in which software of GIS develops are discussed in detail, and the merits and trends of ComGIS are pointed out. Also, expatiates how to design the component of Spatial-Analyzer function in ComGIS.

Keywords　Geographic information system, ComGIS, Component, Spatial-analyzer

1. GIS 的组件化趋势

地理信息系统(GIS)技术正处于一个重要的发展时期，新概念和新产品层出不穷。在GIS蓬勃发展的今天，GIS的组件化趋势日益明显，已经成为GIS的重要发展方向之一。从发展历程看，GIS可以划分为图1所示的几个发展阶段。

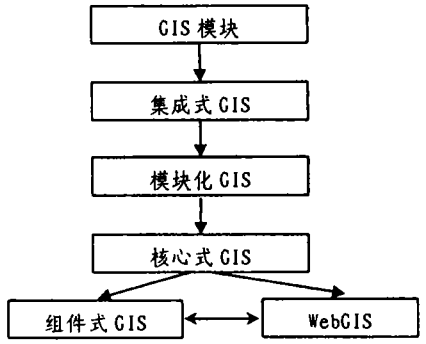

图1　GIS 软件发展历程

在GIS发展的早期阶段，由于受到技术的限制，GIS软件往往是只能满足于某些功能要求的一些模块，没有形成完整的系统，而各个模块之间不具备协同工作的能力。

随着理论和技术的发展，各种GIS模块走向集成，逐步形成大型GIS软件包(GIS Package)，我们暂称之为集成式GIS (IntegratedGIS)，如ESRI的Arc/Info、Genasys的GenaMap等均为集成式GIS的代表。

另一类GIS为随后出现的模块化GIS(Modular GIS)，代表软件有Intergraph的MGE等。模块化GIS的基本思想是把GIS按照功能划分为一系列模块，运行于统一的基础环境之上(如MicroStation)。但无论是集成式GIS或是模块化GIS，都很难与管理信息系统(MIS)以及专业应用模型集成高效、无缝的GIS应用。

为解决集成式GIS与模块化GIS的缺点，提出了核心式GIS(Core GIS)的概念。但核心式GIS被设计为操作系统的基本扩展，也不适应可视化程序设计的潮流。

随着计算机软件技术的发展，GIS组件化发展到了一个全新的阶段，出现了组件式GIS(Components GIS，缩写为ComGIS)。组件式GIS基于标准的组件式平台，各个组件之间不仅可以进行自由、灵活的重组，而且具有可视化的界面和

杨占龙　硕士研究生，从事GIS系统、数据库系统、计算机算法分析、计算机网络等方面的研究。汪林林　教授，从事GIS系统、数据库系统、计算机算法分析、计算机网络等方面的研究。

我们的核酸序列数据库上联机分析处理的界面如图3所示。

结论　本文针对当前广泛采用的EMBL核酸序列数据库，提出了有针对性的联机分析处理实现方案。该方案给出了从文件格式的数据到关系数据库再到OLAP多维数据集的一系列数据转存实现方法。并利用转存之后的数据进行了核酸序列分布规律和各类别核酸序列组成成份的统计分析。该方案可以有效地解决原始数据文件分析速度慢，分析能力差的弱点，可以被视作将联机分析处理(OLAP)技术应用于生物信息学的一次成功探索。另外，该方案也完全适用于蛋白质数据库(如Swiss-Prot)。

参 考 文 献

1　Inmon W H. Building the Data Warehouse Second Edition. John Wiley&Sons Inc, 1996
2　Melinda G, Song, Il-Yeol. Data warehouse design for pharmaceutical drug discovery research Axel. In: Intl. Conf. on Database and Expert Systems Applications - DEXA. Sep. 1997
3　Xie Guochun, DeMarco R, Blevins R, Wang Yuhong. Storing biological sequence databases in relational form Bioinformatics 16: 288～289
4　Microsoft Corporation Microsoft SQL Server 2000 Analysis Services Microsoft Press, 2000
5　Stoesser G. The EMBL nucleotide sequence database. Nucleic Acids Research, 2001, 29(1)
6　European Bioinformatics Institute. EMBL Nucleotide Sequence Database: User Manual Release 70, March 2002