一个基于软件设计模式的生物信息存储模式

杨进才, 赵 森, 刘小姣, 胡金柱
(华中师范大学 计算机科学系, 武汉 430079)

摘 要: 为了消除各生物信息学数据库之间的模式异构问题,根据生物信息的存储现状,提出了一种存储模式。该模式从物种、类别、基本信息、功能和测序方法五个方面对数据中的信息进行抽象。运用了软件设计模式的思想,通过"派生""组装"等面向对象的方法生成与模式对应的 XML schema 文件。抽象出的存储模式不但能使数据之间的关系更加紧密,而且可以形成交叉索引的完整生物信息体系。

关键词: 生物信息抽象; 生物数据存储模式; 设计模式; 可扩展标记语言
中图分类号: Q811.4 **文献标志码**: A **文章编号**: 1001-3695(2010)07-2598-04
doi:10.3969/j.issn.1001-3695.2010.07.055

Storage pattern of bio-information based on software design patterns

YANG Jin-cai, ZHAO Sen, LIU Xiao-jiao, HU Jin-zhu
(*Dept. of Computer Science, Huazhong Normal University, Wuhan* 430079, *China*)

Abstract: For eliminating the pattern heterogeneous between bio-information databases, this paper proposed a storage pattern according to the current storage status of biological data, which abstracted data information from five aspects, including species, category, basic information, function, and sequencing method. Generated the corresponding XML schema files by using concepts of design patterns and some object-oriented means like "deriving" and "assembling". This storage pattern can not only make the relationships between biological data more close, also can form a complete biological information system with cross-index.

Key words: bio-information abstraction; biological data storage pattern; design patterns; XML

目前,越来越多的生物基因(组)已测序完成,生物信息学数据呈指数增长。当前主流的生物信息数据库包括核酸序列数据库(GenBank、EMBL、DDBJ 等)、蛋白质序列数据库(PIR、PROSITE 等)、三维分子结构数据库(PDB、SCOP 等)[1]。由于这些数据库中数据的存储模式千差万别,使得生物信息学数据中存在严重的模式异构现象。模式异构是指同样的生物信息在不同的数据库中采用不同的属性集与不同的结构。因为目前国际上还没有形成生物数据存储模式的统一标准,对于模式异构问题尚无很好的解决方法[2]。关于生物信息学数据的整合也存在大量的讨论和研究,文献[3,4]中阐述了生物信息学数据整合的困难所在。但绝大多数研究都是在现有数据存储模式的基础上加以整合,争取较大限度地对其进行数据挖掘,虽然在一定程度上解决了数据的语法和语义异构,但所得结果并不能完全满足大多数领域学者的要求。本文提出了一种基于软件设计模式思想的生物信息学数据的存储模式 SCIFS,其代表从五方面抽象生物信息学数据中所含信息(species,category,information,function and sequencing),能最大程度地解决数据存储的模式异构问题,随之从根本上避免了非同源数据间的语法及语义异构问题。本文使用 XML(extensible markup language,可扩展标记语言)描述 SCIFS 的具体实现。

1 生物信息学数据与 XML

1.1 XML 在生物信息学中的应用

XML 是一种半结构化的数据模型,以一种开放的自我描述方式定义数据结构在描述数据内容的同时能突出对结构的描述,从而体现出数据之间的关系[5]。这正是当前生物数据所需要的描述语言,使用 XML 可更好地规范生物数据的结构,方便数据之间的交流与集成。当前流行的生物信息学数据库中大多都已提供了 XML 格式的数据下载[6,7],但均为"单方向"的,即只有当用户需要 XML 格式的生物数据时,各大数据库才提供 XML 文件,但在各数据库之间仍然存在着"数据鸿沟"。而要从根本上解决这些问题,就必须充分发挥 XML 的特点,从数据的存储层面杜绝数据异构现象。

1.2 XML 在新型存储模式中的优势

本文提出的 SCIFS 存储模式(下文将进行详细描述)借用了软件设计模式的思想,将数据结构模块化,利用派生、继承等手段生成灵活多变且结构统一的数据存储模式。XML 对数据结构的描述和本身灵活多变的特点,使其十分适用于这种组装模式,为新型存储模式的应用提供了技术基础。

收稿日期: 2009-12-10; **修回日期**: 2010-01-18

作者简介: 杨进才(1967-),男,湖北咸宁人,教授,硕导,博士,主要研究方向为生物信息学、现代数据库理论与技术; 赵森(1984-),男,山东青岛人,硕士研究生,主要研究方向为生物信息学(nesoahz@yahoo.com.cn); 刘小姣(1982-),女,湖北麻城人,硕士研究生,主要研究方向为移动数据库; 胡金柱(1947-),男,湖北宜昌人,教授,博导,主要研究方向为分布式信息系统、软件工程.

2　生物信息学数据的模式抽象

2.1　软件设计模式思想

软件设计模式简单来说就是在一个环境中,将遇到的问题抽象出来,提出解决方案,当在其他环境中遇到类似的问题时,可直接利用已有的方案来解决当前问题。将设计模式的思想引入生物信息学数据存储,可为数据中信息的抽象提供理论依据,并使之更加规范。

2.2　对于生物信息学数据存储需求的分析

当前生物信息学数据的存储大多根据(以 DNA 序列为例)测序者、参考文献、所在生物体组织、序列本身等信息进行存储。虽然这些信息是不可或缺的,但对于要使用此序列的生物学家或医学家来说,其中很多信息是无用的。虽然当前有众多检索系统可以根据不同的关注角度进行查询,如 SRS 和 Entrez 等,但其采用的方法都是建立分散在不同的异构数据库中数据之间的简单链接,数据本身没有有机地整合在一起,迫使用户仍要手动地在大量的查询结果中寻找有用的信息[8]。这就需要将生物学家甚至于医学家、药物学家等各个领域学者对于生物信息的不同需求进行抽象,在提交和存储生物信息数据时对其进行详细的分类描述。如此一来,用户在搜索数据时就能更有针对性,并能提高搜索效率。

当然,这只是针对序列功能的单一抽象,本文提出的存储模式从五个方面抽象生物数据信息,每一方面看做一个“拟类”,“拟类”中的方法则是把抽象出来的信息用 XML 进行描述,并根据此“拟类”的特点和内容设计大量“派生拟类”。然后对部分“派生拟类”进行组装,形成最终的数据存储模式。之后的数据提交和存储即看做“拟类”的实例化。通过这种方式,不但使数据的描述更加详细、全面,并且可使海量数据形成层次分明、枝干交错的大型结构体系,便于各领域学者交叉检索。同时,因为“类”的封装性,每个“拟类”中描述的信息对其他“拟类”是独立的,用户对某一或某几方面信息的检索结果将更加准确,大大减少冗余结果。

2.3　SCIFS

SCIFS 分别是模式中五个“拟类”的英文名称首字母,包括物种—species、类别—category、基本信息—information、功能—function、测序方法—sequencing。这一节将详细介绍 SCIFS 存储模式中各“拟类”的设计目的(确定模式中的对象)、抽象过程(设计对象的类)和具体实现(对象的实现)。

为方便起见,下文中用名词“类”代替“拟类”。但要注意的是,“拟类”并非严格意义上的类,这里只是借用面向对象中“类”的部分概念。

2.3.1　物种类

1)目的　设计此类的目的是从来源上区别不同的生物信息学数据。生物信息学中的每一条数据均来自某一生物体,这也是其最基本的特征。将物种信息单独抽象出来可使生物信息学数据更加“生物化”,并方便对属于同一分支的生物进行比较。同时,从一定程度上将生物演化的理论加入到生物信息学数据中,方便了对 DNA、RNA 的演化方面的研究。

2)抽象过程　物种类的抽象根据现实中生物学对物种的分类理论,将其按照物种划分为五界(类):原核生物界、原生生物界、植物界、动物界、真菌界,其下又分为门、纲、目、科、属、种(均视为派生类)。以“人”为例,如图 1 所示。

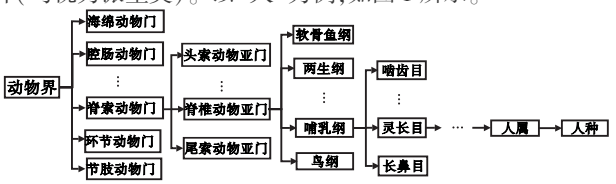


图1　人的物种分类树

由于篇幅所限,图中省略了真兽亚纲、真猴亚目、窄鼻猴次目、类人猿超科、人科。

因为每一种生物最终都属于某一“种”,将界、门、纲、目、科、属定义为抽象类[9]并包含抽象方法,即对每一层次生物特点的说明。要注意的是,物种类的最终实例化与具体实现(应用 SCIFS 存储数据)有所区别,存储时并不用具体方法覆盖每条数据所继承的所有抽象方法,而是为不同的界、门、纲等仅进行一次特征描述,将这些描述与数据本身分开存储,使用户在需要时可以方便查看。这就避免了来自相似生物的数据中存在大量的重复信息。

3)具体实现　物种类中包含的信息对应于生物信息学数据中的一个模块,此模块由一系列 XML 元素(element)构成,准确地描述了与当前数据对应的物种分类。在提交数据时,用户选择提交数据所属的“种”后,存储系统自动生成物种类的 XML 模块。

以“人”为例,当提交人体的 DNA、RNA 或蛋白质等序列时,用户选择“人种”,提交系统会自动生成类似图 1 的树状结构,并以文字形式在 XML 文件中进行描述。为了突出数据的结构,以下物种类的部分 schema 显示了与数据对应的 XML schema 文件(由于篇幅所限,对于 schema 文件的描述将省略 schema 的部分定义,只关注与抽象类的实现相关的内容)。

```
species. xsd
…
〈xs:element name = “Species”〉
  〈xs:complexType〉
    〈xs:sequence〉
      〈xs:element name = “kingdom” type = “xs:string”/〉
      〈xs:element name = “division” type = “xs:string”/〉
      〈xs:element name = “class” type = “xs:string”/〉
      〈xs:element name = “order” type = “xs:string”/〉
      〈xs:element name = “family” type = “xs:string”/〉
      〈xs:element name = “genus” type = “xs:string”/〉
      〈xs:element name = “species” type = “xs:string”/〉
    〈/xs:sequence〉
  〈/xs:complexType〉
〈/xs:element〉
…
```

对于其他物种的 DNA、RNA 序列数据同理,只需知道此生物所属的具体“种”,物种类即可生成(派生出)精确的物种分类树。

2.3.2　类别类

1)目的　当前生物信息学中包括 DNA、RNA、蛋白质序列等不同性质的数据,设计类别类是为了区别生物数据本身。之所以用统一的存储模式来存储 DNA、RNA 和蛋白质序列数据,是因为这三类数据之间存在十分紧密的联系,如 DNA 是某些 RNA 合成的模板,而 DNA 和 RNA 又载有氨基酸的合成密码等。使用统一的存储模式为以后科学家研究这几种数据间的关系提供了便利。

2)抽象过程　DNA、RNA 及蛋白质的种类繁多,如信使

RNA、转运 RNA、胶原蛋白、球蛋白等。按照化学成分的不同，将蛋白质进行分类；RNA 则根据其功能分为四类。由于目前尚没有 DNA 的统一分类，在存储模式中只需标注为 DNA 数据即可，如图 2 所示。

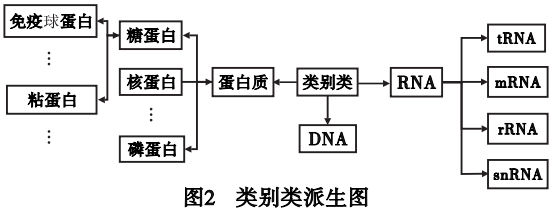


图2　类别类派生图

3）具体实现　此类对应的 schema 文件命名为 category. xsd，使用〈choice〉指示器完成在类别中选择一条派生路径的作用。这里的实现与物种类略有不同，由于核酸和蛋白质的分类深度不一，不像物种类中存在统一的门、纲、目、科、属、种等划分，而且不断地有新类别蛋白质或核酸被发现，使用〈choice〉指示器使存储模式更加灵活。

2.3.3　基础信息类

此类是沿袭现有生物信息学数据的主要数据域构成的，包括当前各大生物信息学数据库维护信息中的大部分字段。如序列发现者、测序者、组织来源、相关文献、统一索引序号、注释、序列具体位置等。设计此类的目的主要是为了让新存储模式可以兼容旧数据模式，在此前提下，对此类不进行细化派生，保持其字段的平等性。此类生成的 schema 文件命名为 infor. xsd。其格式可参照当前各大数据库的 schema 文件[10]。

2.3.4　功能类

1）目的　功能类在这五部分中较为重要，其目的是描述不同领域学者对生物信息学数据中所包含信息的不同需求，对于日后的检索和数据挖掘起着关键作用。

2）抽象过程　对生物信息学数据的研究并不仅仅局限于序列本身，对其外延信息的研究种类繁多，如 DNA 序列所包含的基因功能、与致病基因的关系、蛋白质的作用、如何改变基因表现型等。对于所有这些研究（应用），根据领域的不同进行划分，然后在各领域中再进行细化。完成这些工作首先要对生物信息学所设计的各个领域进行调查，了解各领域学者对生物信息学数据的需求，然后用尽量少的派生类涵盖尽可能多的应用，如图 3 所示。图 3 中简要列出了生物信息学数据在生物学、医学、农业等领域中的部分研究方向。

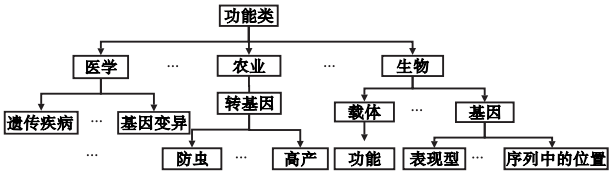


图3　功能类派生图

3）具体实现　使用与类别类相似的实现方法，在用户提交数据时，在功能类对应的 XML 文件模块中选择一项或多项功能域进行描述，在日后的研究中，还可由其他领域专家为此数据添加其他应用。如此一来，生物信息学数据所含信息从传统的以基本信息（基础信息类）为主过渡到了以功能描述为主。功能类生成的 schema 文件为 function. xsd。

2.3.5　测序类

1）目的　当前对于核酸测序技术虽已较为成熟，如已用 Shotgun 法测得了人类基因组[11]，但许多测序算法仍存在不可避免的误差，如基于 Hamilton 路径的拼接算法、基于 Euler 路径拼接算法[12]等，这些算法中所涉及到的 NP 完全问题，以及克隆 DNA 序列 read 的过程中所产生的变异或测序误差，都会导致最终数据有所偏差。设计此类的目的一是可以使研究序列拼接算法的学者更方便地查询及比对实验结果；二是若某种算法大大提高了精度，可轻易地检索到所有用此算法测序的核算或蛋白质序列，可进行比对甚至更新数据。

2）具体实现　生成的 XML 模块仍以选择的方式为主，列出当前对于 DNA、RNA 和蛋白质的主要测序手段，用户在提交数据时进行选择。因为测序算法的不断改进，将此模块设计为可扩展的也尤为重要。与测序类对应的 schema 文件为 sequencing. xsd。

3　模式的组装规则

在提交数据时，由物种类、类别类、基础信息类、功能类、测序类共同派生出一个类。此时“调用”类中的方法，即形成五个包含多个数据域的 XML 模块，等待用户选择或填充信息，同时将五部分的 schema 文件按目标数据的要求集成到一起。但此时所形成的存储模式并不是完整的 SCIFS 存储模式，SCIFS 还应满足以下规则及约束：

a）为保持最终数据格式的统一性，须严格规定模块和数据域的顺序。

b）为避免重复提交相同数据，需要为数据设置由多个数据域组成的“主键”（primary key）。选择类别和基础信息组成数据的“主键”，惟一地标示一条生物信息学数据。当用户提交数据时，如果数据库中已存在此数据，用户只需更新现有数据的功能类、测序类或者基础信息类中的部分数据域即可。

c）因为模式中存在可扩展的类，如类别类、功能类等，当用户针对某一数据的某些 XML 模块进行扩充时，所添加的数据域对于其他数据域应是透明的，即不会影响原有的存储模式。

在满足以上三条约束的情况下，生成的才是较严格的 SCIFS 存储模式。

此外，在之前的五个类中，最后注明的 schema 只是合并时所用到的 schema 文件，但对于每个类生成的 XML 模块，与其对应的并不只是单一的一个. xsd 文件，如类别类，其本身按照不同派生路径可建立多个. xsd 文件，最终组装时再根据目标文件确定属于此类的哪些. xsd 文件合并成所需要的 category. xsd。这样可以保证与生成的 XML 文件对应的 schema 规模最小，也使存储模式更具有灵活性。

4　模式规则的有效性证明

对于 DTD 文件来说，XML 模式存在两个主要约束：a）模式的结构有效性。要求 XML 模式是能够被实例化的。b）模式的结构良好性。要求模式定义中的所有元素定义是有意义的，即元素定义都能被实例化[13]。而由于 schema 与 DTD 的差别，仅需要给出一种针对 schema 方式的基本约束——结构有效性。需要明确的一点是，这里所提到模式中存在的问题完全由模式结构的设计所产生，而与语义无关。

定义　给定一个 XML 模式 S，当且仅当 S 中不存在嵌套定义时，且至少存在一个 XML 文档 T，T 满足 S，则称模式 S 是结构有效的。

当存在嵌套定义时，XML 模式将失去实际意义，即模式中

某些元素无法被实例化。不满足结构有效性的 schema,元素 *A* 与 *C* 嵌套引用,导致此 XML 模式中存在无限循环,如下所示:

```
〈xs:element name = "A"〉
  〈xs:complexType〉
    〈xs:sequence〉
      〈xs:elementname = "B" type = "xs:string"/〉
      〈xs:element rel = "C" minOccurs = "1"/〉
    〈xs:sequence〉
  〈/xs:complexType〉
〈/xs:element〉
〈xs:element name = "C"〉
  〈xs:complexType〉
    〈xs:element ref = "A" minOccurs = "1"/〉
  〈/xs:complesType〉
〈/xs:element〉
```

本文提出的 SCIFS 模式严格按照结构有效性的定义,对于模式中各部分进行抽象描述,使得其中所有元素均有明确的子孙或兄弟关系而不会出现嵌套引用的现象。这不仅使此模式具有理论意义,而且为下一步根据模式设计具有严格约束的生物信息数据库打下良好的基础。

5　结束语

将软件设计模式思想引入生物信息学数据存储,借用面向对象的概念,很好地将各方面的信息抽象为独立的类,使数据所含信息更加丰富且条理清晰,方便用户进行交叉检索,从而为海量的生物信息构建起更加发达也更加灵活的体系结构。由于此存储模式对于信息抽象要求较高,需要对大量领域进行详细的需求分析,这就要借助许多领域专家的帮助,使信息的抽象和类的设计更加精确。

参考文献:

[1] 张晓东,张传富,彭科峰,等. 生物信息学数据库研究进展[J]. 生物信息学,2006,4(3):143-145.

[2] CHEN Qing-feng, CHEN Yi-ping, ZHANG Cheng-qi. Detecting inconsistency in biological molecular databases using ontologies[J]. Data Mining and Knowledge Discovery,2007,15(2):275-296.

[3] PASQUIER C. Biological data integration using semantic Web technologies[J]. Biochimie,2008,90(4):584-594.

[4] LI Xiao, ZHANG Yi-zheng. Bioinformatics data distribution and integration via Web services and XML[J]. Genomics Proteomics & Bioinformatics,2003,1(4):299-303.

[5] 顾天竺,沈洁,陈晓红,等. 基于 XML 的异构数据集成模式的研究[J]. 计算机应用研究,2007,24(4):94-96.

[6] WANG Li-chuan, JEANJACK R, ROBINSON A. XEMBL: distributing EMBL data in XML format[J]. Bioinformatics,2002,18(8):1147-1148.

[7] MIYAZAKI S, SUGAWARA H, GOJOBORI T, *et al*. DNA data bank of Japan(DDBJ) in XML[J]. Nucleic Acids Research,2003,31(1):13-16.

[8] 杨文,韩涛,孙志茹. 生物信息学序列库与文献库的整合模式浅析[J]. 情报理论与实践,2008,31(1):112-115.

[9] ERICH G, RICHARD H, JOHNSON R, *et al*. Design patterns: elements of reusable object-oriented software[M]. [S. l.]: Addison Wesley Longman,2002:10-12.

[10] NCBI XML schema file index[EB/OL]. (2009-11-02) [2009-12-20]. http://www.ncbi.nlm.nih.gov/data_specs/schema/.

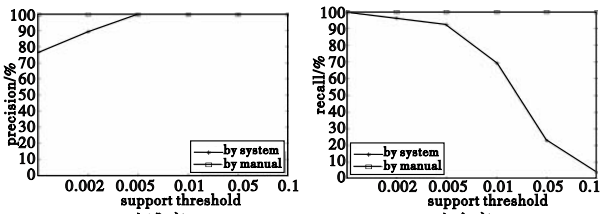
[11] ISTRAIL S, GRANGER S G, LILIANA F, *et al*. Whole-genome shotgun assembly and comparison of human genome assemblies[J]. Proc National Academy of Sciences,2004,101(7):1916-1921.

[12] 骆志刚, 方小永, 丁凡. DNA 序列拼接的研究进展及挑战[J]. 计算机工程与科学,2007,29(8):127-132.

[13] 郑华利, 钟昊, 郭汉英. XML 模式的结构规范化研究[J]. 系统工程与电子技术,2007,29(1):78-81.

(上接第 2586 页)由图 2(a)可以看出,在支持度阈值达到一定数值后(本例中是 0.005),系统可以获得 100% 与本体没有冲突的具有一致性规则,这里采用过高的阈值必将导致信息丢失率的大幅上升,且对于有效性而言并没有太大意义。这个实验也表明:通过适当的转换,可以一次性将得到的规则进行批量映射到本体,而无须再次检测,从而为使用 KDD 技术自动扩建本体提供基础。

图 2(b)主要显示了系统的查全率,在裕度 = 0.01 的情况下(支持度 = 0 时,裕度 = 0)得到的测量结果。



(a)查准率　　(b)查全率

图2　实验2结果

以上的实验确实证明了当选用适量的样本及合适的阈值参数,基于样本空间的规则一致性判则和 MARCMAO 算法具有较高的可行性及一定的有效性。

4　结束语

本文通过 KDD 自动扩建本体过程中出现的规则一致性问题进行详细阐述,建立数学模型,并提出相应的基于先验规则优先的 MARCMAO 算法,把本体规则和关联规则的一致性保持与维护映射到样本空间,利用一定的训练样本去校验新规则与本体中的先验规则之间的矛盾、冗余、蕴涵关系,解决新规则和先验规则的规则一致性问题。实验表明算法可操作性强、精度高、收敛速度快,具有较高的可行性和有效性。

参考文献:

[1] BASTIDE Y, PASQUIER N, TAOUIL R, *et al*. Mining minimal non-redundant association rules using frequent closed item sets[C]//Proc of the 1st International Conference on Computational Logic. London, VK:Springer-Verlag, 2000:972-986.

[2] 赵波,冯洁. 本体中继承关系的形式化表示及其应用[J]. 计算机工程与设计,2008,29(1):154-156.

[3] CHEN Xu-hui, LU Jun, LIU Zhong-yuan. Assistance ontology of quality control for enterprise model using data mining[C]//Proc of IEEE Industrial Engineering and Engineering Management. Singapore: IEEE Computer Society,2007:602-606.

[4] HAN Jia-wei, KAMBER M. Data mining: concepts and techniques[M]. 北京: 机械工业出版社,2001:162-167.

[5] 董俊,王锁萍,熊范纶,等. 基于多维关联规则的本体规则扩展方法研究[J]. 模式识别与人工智能,2009,22(5):756-762.

[6] GRAU B C, MOTIK B, HORROCKS I, *et al*. OWL 2 Web ontology language: model-theoretic semantics[EB/OL]. (2008-04-11). http://www.w3.org/TR/2008/WD-owl2-semantics-20080411/.

[7] 董俊,王锁萍,熊范纶. 基于本体的领域知识重用方法研究[J]. 计算机应用研究,2009,26(12):4546-4561.

[8] CARVALHEIRA L C C, GOMI E S. A method for semi-automatic creation of ontologies based on texts[M]. Berlin: Springer-Verlag,2007:150-159.