第35卷 第24期 计 算 机 工 程 2009年12月
Vol.35 No.24 Computer Engineering December 2009
·软件技术与数据库· 文章编号：1000—3428(2009)24—0057—03 文献标识码：A 中图分类号：TP311.12

基于 XML 的 DNA 公共数据模型

杨进才，金 蕾，胡金柱
(华中师范大学计算机科学系，武汉 430079)

摘 要：针对目前 DNA 数据组织与处理中存在的数据异构问题，提出一个基于 XML 的 DNA 公共数据模型(DCDM)。该模型具有很强的可扩展性，能克服一般公共数据模型的作用范围小的缺点，可用于构建 DNA 研究领域统一的 DNA 数据描述模式。实验结果表明，该模型能解决 DNA 数据异构中的语义异构。

关键词：数据异构；语义异构；DNA 公共数据模型；XML 技术

DNA Common Data Model Based on XML

YANG Jin-cai, JIN Lei, HU Jin-zhu
(Department of Computer Science, Huazhong Normal University, Wuhan 430079)

【Abstract】Aiming at the problem of data isomerism existing in DNA data organizing and processing, a DNA Common Data Model(DCDM) based on XML technology is proposed in this paper. Compared to the ecumenical common data models which have narrow range of application, DCDM has much stronger expansibility, which could be the uniform mode of DNA data description. Experimental result shows that the model can eliminate semantic isomerism in DNA data isomerism.

【Key words】data isomerism; semantic isomerism; DNA Common Data Model(DCDM); XML technology

1 概述

目前生物学研究已进入后基因时代，产生的巨量的生物基因数据有待分析与处理。然而主要的 DNA 数据库并没有一个完全统一的数据库格式，这就导致 DNA 数据异构问题的存在。

数据的异构分为语法的异构和语义的异构[1]。DNA 数据异构问题是生物数据整合研究中的重点。随着 XML 日渐成为 Web 上数据交换的事实标准，很多 DNA 数据库开始采用 XML 作为生物数据的描述语言。XML 的使用基本上能解决生物数据在语法上的异构问题，但语义的异构问题依然存在。

如何消除 DNA 数据间的异构是 DNA 数据整合中的关键问题。现阶段的解决方案主要有：生物数据置标语言[2]，核酸元数据模型[3]。然而这些方法都存在不足：由于很多研究机构推出各自的数据描述标准，从而形成很多置标语言，且研究领域的不同导致了其使用的局限性，并且各生物置标语言之间也并没有统一，因此，并不能从根本上解决生物数据在语义上的异构问题。运用元数据技术可以为各专业数据库提供统一、规范的结构化描述，为生物数据的集成与共享提供了实现手段。然而，元数据标准只是基于本专业领域知识建立的统一的关于数据的描述格式，它并没有从数据管理的角度考虑如何更有利于知识的管理与发现。

上述方法都只能实现特定类型数据间转换，如何能等价地实现更多不同类型数据转化是 DNA 数据整合的关键，因此，本文提出一个基于 XML 的 DNA 公共数据模型(DNA Common Data Model, DCDM)以解决 DNA 数据异构中的关键问题——语义异构。本文设计的 DCDM 具有很强的可扩展性，可用于构建 DNA 研究领域统一的 DNA 数据描述模式。

2 DNA 公共数据模型

DNA 数据语义异构主要包括：命名冲突和结构冲突。前者对于同一概念或数据的命名不同，后者对同一信息的描述采用不同的结构。因此，对于这些语义异构的数据，采用统一数据模型描述的方式，将原始数据模式通过公共数据模型进行翻译，生成的输出模式再通过模式集成生成统一模式，这样，就可以完成异构数据的整合，消除语义异构[4]。本文提出的 DNA 公共数据模型就是通过希望建立统一的 DNA 数据描述模式来实现 DNA 数据的整合，消除语义异构。

2.1 DCDM 框架

XML 能够描述半结构化的数据，而 DNA 数据基本上都是以文档形式存在的，使用 XML 对 DNA 数据加以描述，能够统一数据描述格式，消除语法异构。因此 DCDM 基于 XML 而构建。但是因为 DNA 数据库不但语法上异构，而且语义上也是异构的，即使采用同一种模型描述数据模式也无法实现数据的集成，因此不可能用一个数据模式涵盖 DNA 数据需要表达的信息。基于上述理由，DCDM 框架如图 1 所示。

DCDM 由 2 个部分组成：DNA 数据模式库(以下简称模式库)和 DNA 数据模式映射库(以下简称映射库)。模式库分别由 3 个库构成：DNA 数据类型库，DNA 数据组件库，DNA 数据子模式库(以下简称类型库、组件库、子模式库)；模式映射库是子模式包含的数据类型和数据组件的模式映射。

基金项目：湖北省科技攻关计划基金资助项目(2007AA101C49)
作者简介：杨进才(1967－)，男，博士，主研方向：生物信息学；金 蕾，硕士研究生；胡金柱，教授、博士生导师
收稿日期：2009-05-20 **E-mail：**yuki925@qq.com

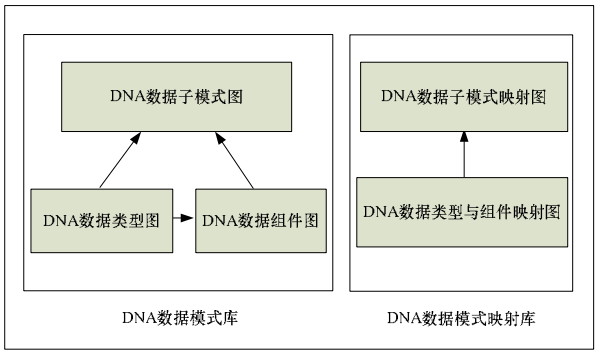

图 1　DCDM 框架

2.2　DCDM 的功能与构建

DCDM 的生物数据模式库包含类型库、组件库、子模式库，这 3 个库都是依据 DNA 数据之间的关系而构建的。以 DNA 序列数据为例说明 DCDM 的功能与实现。图 2 列出了与 DNA 序列相关的几项数据及其数据关系。

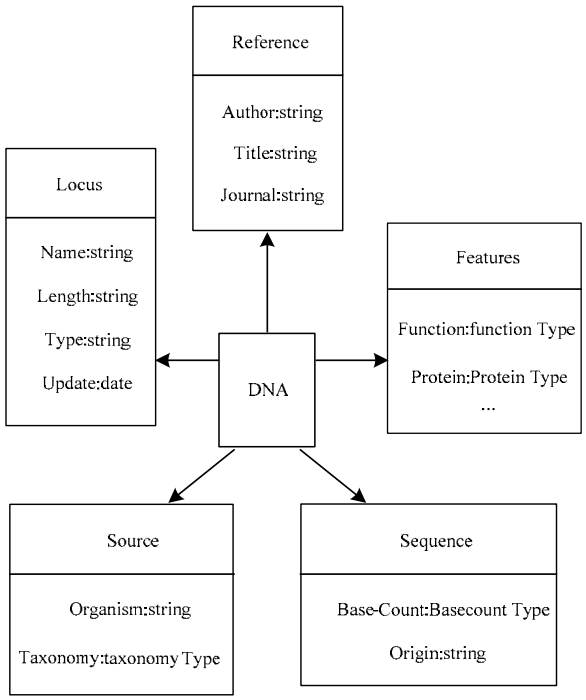

图 2　DNA 数据关系

DCDM 把 DNA 数据类型分为简单数据类型和复杂数据类型，简单数据类型不包含子元素，它的数据类型都属于 XML SCHEMA 基本的数据类型。

简单数据类型主要是为了统一数据标识以及对数据类型的定义。如图 2 中的 Author 表示文献的作者名，在 EMBL 数据库中则用 RA 表示，而有的数据库 Author 由 first-name 和 last-name 组成，如果 DCDM 定义它为 Author Type 类型，数据类型为字符型，则可以消除语义上的冲突；复杂数据类型的数据包含若干子元素，子元素之间的关系是聚合关系中的整体一部分关系。

复杂类型除了统一数据标识以外还主要是统一数据的描述内容及结构，如图 3 中的 Taxonomy 表示序列来源分类学的位置，它由 Cell-type, kingdom, phylum, subphylum, class, infraclass 等子元素组成，这些子元素结合在一起表示分类学的位置，因此，可以定义为 Taxonomy Type 类型。类型库中简单数据类型和复杂数据类型是根据对现有的 DNA 数据库的数据模式的分析以及对 DNA 数据关系的分析而定义的。就 DNA 数据组织与处理的整体而言，可以制定 DNA 数据类型的标准，所有 DNA 数据的描述都遵循这个标准，从根本上消除因标识名不同和数据类型与结构不同而引起的语义上的异构。

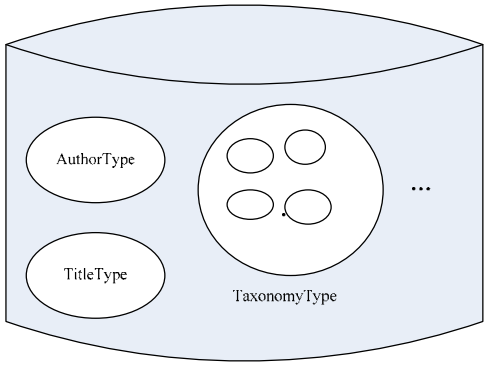

图 3　DNA 数据类型库

组件库中的数据组件是根据生物学知识将描述某一对象的数据按一定结构集合在一起形成的，构成数据组件的子元素都是类型库中定义的数据类型。如图 4 中的 Reference 就可以定义为由 Author, Titie, Joumal 构成的数据组件，将对象统一定义成为一个标准，从而消除因数据描述内容不同而引起的语义异构。

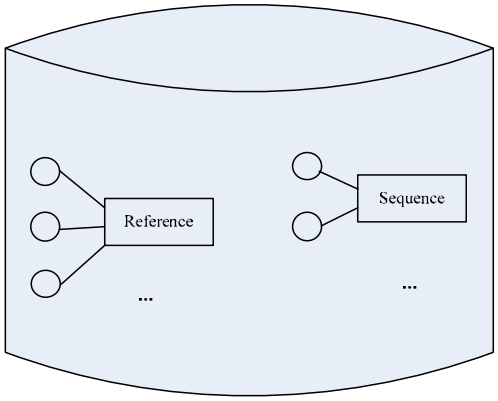

图 4　DNA 数据组件库

子模式库是根据 DCDM 类型库和组件库定义的应用模式的集合。通过构建子模式库可以以一致的描述方式对异构数据源模式进行描述，如图 5 所示。

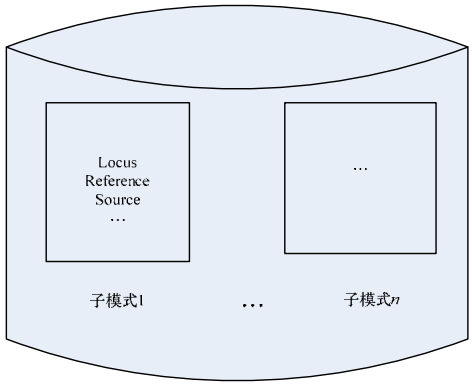

图 5　DNA 数据子模式库

DCDM 的映射库主要是帮助 DNA 数据库解决模式翻译的问题，其中存放的是模式库到数据源的映射关系，由模式映射 XML 文档组成，如图 6 所示。

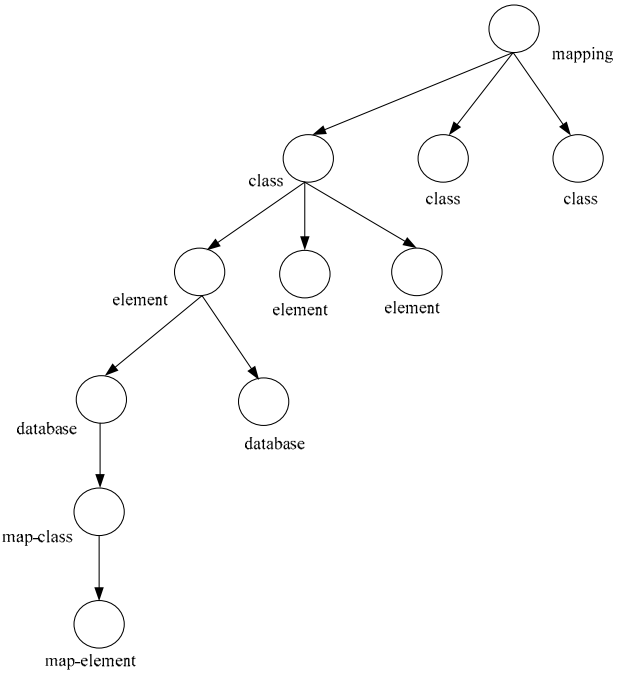

图 6　模式库到数据源的映射文档结构

其中，mapping 是文档根节点；class 是模式库中的数据类型或数据组件，例如 locus, reference 等；element 是 class 所包含的元素，例如 locus 中包含 name, length, type，database 是对应的数据库的名称，例如 EMBL；map-class 和 element 是数据源对应的数据类型或组件和所包含的元素，例如 Locus 中的 name 对应 EMBL 数据库 ID 中的 AC。

2.3　验证实验

以 DCDM 到 EMBL 和 GenBank 数据库的映射为例，模式库到数据源的映射文档如下：

```
<mapping>
<class name="Locus">
<element name="name">
<database address="http://www.EMBL.com">
<map-class name="ID">
<element> AC </element>
</map-class>
</database>
<database address="http://www.Genbank.com">
<map-class name="Locus">
<element> Accession </element>
</map-class>
</database>
</element>
<element name="length">
<database address="http://www.Genbank.com">
<map-class name="Locus">
<element> length </element>
</map-class>
</database>
</element>…
<!--节约空间，后面定义类似-->
```

整合前对 EMBL 的查询结果文档如下:

```
<database address=”http://www.Embl.com”>
<ID><AC>u00096</AC></ID>
<OS>Escherichia coli</OS>
</database>
```

对 Genbank 的查询结果文档如下：

```
<database address=”http://www.Genbank.com”>
<Locus><name>u00096</name>
<length>4639221BP</length></Locus>
</database>
```

通过 DCDM 模式整合以后的结果文档如下：

```
<result>
<Locus>
<name>u00096</name>
<length>4639221BP</length>
</Locus>
<source>
<taxonomy> Escherichia coli</taxonomy>
</source>
</result>
```

实验结果证明，整合前的查询文档是各来源数据库的格式，整合后的查询结果文档是 DCDM 模式库的格式。通过对比与分析，证实了 DCDM 可以实现 DNA 异构数据的整合。

3　结束语

本文通过分析 DNA 数据之间关系，确定了 DCDM 的框架，基于 XML 提出 DCDM 的具体实现方法，并通过实验验证了该模型的有效性。一般的公共数据模型的作用范围相对很小，而本文设计的 DCDM 具有很强的可扩展性，可以用来构建 DNA 研究领域统一的 DNA 数据描述方式。DCDM 相当于一个生物数据的字典，具有强大的数据描述能力。因此，任何存在于生物数据库的 DNA 数据都可以参照 DCDM 描述。通过 DCDM 能在一定程度上解决数据语义异构的问题。

参考文献

[1] Philippi S. Light-weight Integration Molecular Biologica Databases[J]. Bioinformatics, 2004, 20(l): 51-57.

[2] 习夏燕, 张忠平, 朱扬勇. 生物语义相似性在数据仓库中的应用与实现[J]. 计算机应用与软件, 2005, 22(11): 5-7.

[3] 张基温, 杨叶勇. 基于 XML 的核酸序列元数据模型[J]. 计算机工程, 2004, 30(21): 76-77.

[4] 王　芳. 基于标准规范的生物标本信息的整理、整合和共享技术研究[D]. 北京: 北京林业大学, 2007.

编辑　金胡考