水稻矮缩病毒基因组数据库的构建*

王建民 罗静初 李 毅 曲 红 吴光耀 顾孝诚**

(北京大学生命科学学院 北京 100871)

摘 要:二级数据库的构建是生物信息学新的重要领域。目前部分生物的基因组序列测定完成后,正在进行广泛而深入的结构和功能研究,使二级数据库的重要性显得日益突出。水稻矮缩病毒是一种在日本、中国和东南亚感染水稻的病原微生物,给农业生产造成很大损失。根据国际和国内对水稻矮缩病毒基因组的研究,利用已有的基因序列和结构、功能等方面的数据,以计算机网络为载体,参考国际通用数据库的格式,尝试建立一个简洁的、友好的通用性好而且专用性强的二级数据库 水稻矮缩病毒基因组数据库。希望能够为研究普通水稻矮缩病毒的粒子结构、基因表达调控、致病机理和防治方法提供一个良好的工具,为从事水稻矮缩病理论和应用研究的工作者提供方便和帮助,并为探索二级数据库的构建积累经验。

关键词:生物信息学,二级数据库,水稻矮缩病毒,基因组

中图分类号:Q332 文献标识码:A 文章编号:0001-6209(2001)01-0043-06

目前,以人类基因组计划为代表的各种物种的基因组学(Genomics)研究工作正在全面展开,根据美国国家生物信息中心(NCBI)的资料,目前有 729 种生物正在或已经完成了基因组的测序工作,并且收录到数据库中。其中古细菌、细菌和真核生物有 30 余种,其余大部分为病毒。国内中国科学院微生物研究所、遗传研究所和生物物理研究所合作已经基本完成了一种嗜热细菌的全基因组测序工作,并开始建立序列数据库。在与农业相关的基因组研究中,除了以主要粮食作物(如水稻、玉米等)为代表的大规模的研究外,还有与这些物种相关的病原生物的基因组研究,后者在研究病原生物和宿主之间的相互作用机理和病虫害防治中均有重要作用。生物信息学(Bioinformatics)作为一门新兴的学科,在基因组学的研究中有重要作用,其中数据库的构建就是一个基本方面。基因组数据库目前以一级数据库(序列数据库)为主,但是随着基因组研究的深入,越来越多基因的结构和功能得到阐明,因此建立二级数据库已经成为一个研究热点。二级数据库(Secondary Database)是根据研究任务的需要,通过搜索、查询已知数据库的信息,进行加工整理、系统化,构建成专用的数据库。目前关于二级数据库的确切定义还不很确定,常用名称有:以任务为导向的数据库(Job-oriented Database)、升值的数据库(Value-added Database)、专家数据库(Specialist Database)或以知识为基础的数据库(Knowledge-based Database)等。

水稻矮缩病毒(Rice Dwarf Virus,RDV)属于呼肠孤病毒科(Reovirudae)的植物呼肠

* 国家自然科学基金委资助项目(39993420,39870027)

** 通讯作者

作者简介:王建民(1974-),男,甘肃省嘉峪关市人,北京大学生命科学学院研究生,硕士,主要从事生物信息学研究。本库网址:http://www.cbi.pku.edu.cn/rdv/

收稿日期:2000-01-27,修回日期:2000-03-17

孤病毒属(Phytoreovirus),在自然情况下由叶蝉以持久方式传播,并可在昆虫体内繁殖。流行于中国南方、日本和东南亚,对农业生产造成很大损失[1~3]。RDV 病毒颗粒为 20 面体,有双层外壳[4,5];其基因组由 12 条双链 RNA(dsRNA)片段组成,除片段 12 外都只有一个开放读码框(ORF)[6]。日本已经对多种分离物(isolate)进行了研究并测定了基因组序列,国内福建分离物的基因组全序列都已经全部测定,并在此基础上正在进行结构功能方面的研究,使构建二级数据库成为必要。

我们利用国内外水稻矮缩病毒基因组研究的数据,尝试构建一个二级数据库。首先希望能够为研究普通水稻矮缩病毒的粒子结构、基因表达调控、致病机理和防治方法提供一个良好的工具,其次希望以此来探索构建二级数据库的经验。

1 水稻矮缩病毒基因组数据库的结构

1.1 数据库的基本结构

根据水稻矮缩病毒基因组以及其他相关信息,首先建立数据库的基本框架(图 1)。为便于与国际上的生物学数据库相衔接,数据库内容用英文编写。

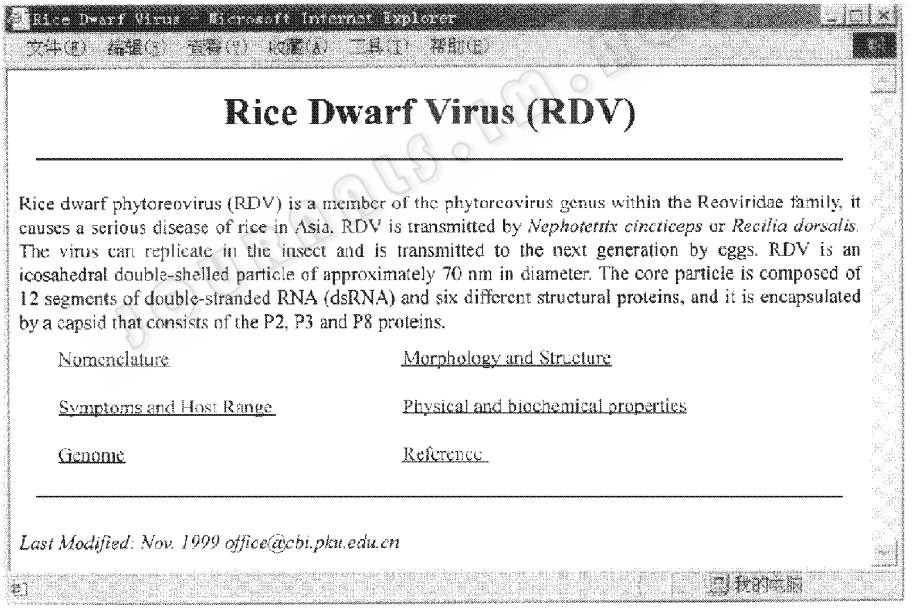

图 1 水稻矮缩病毒基因组数据库的主页

Fig. 1 Homepage of RDV Genome Database

1.2 数据库的具体结构框架如下

1 **Nomenclature**

1.1 Acronym

1.2 Taxonomy

1.3 Strain

1.4 ICTV decimal code

1.5 Isolates

2 **Morphology and Structure**[7]

2.1 Morphology

2.1.1 Diameter

2.1.2 Shell Thickness

2.1.3 Shape

2.1.4 Outer Capsid

2.1.5 Triangular Number (T)

2.2 Crystal Parameters

2.2.1 Space Group

2.2.2 Cell Parameters

2.3 3 -D Structure

2.3.1 CryoEM 50Å Resolution Structure[4]

2.3.2 CryoEM 25Å Resolution Structure[5]

3 **Symptoms and Host Range**

3.1 Natural Host Range and Symptoms

3.2 Experimental Host Range and Symptoms

3.3 Diagnostic,Propagational and Assay Host Range

3.4 Susceptible Hosts

3.5 Insusceptible Hosts

3.6 Transmission

3.7 Geographic Distribution

3.8 Ecology and Control

3.9 Diagnostic and Methods

4 **Physical and Biochemical Properties**

4.1 Properties of particles in sap

4.2 Purification method

4.3 Particle morphology

4.4 Physical properties

4.5 Biochemical properties

4.6 Antigenic Properties

4.6.1 Serological Relationships

4.6.2 Diagnosis

5 **Genome**

5.1 Structure[8~18]

5.2 Comparison of 5′terminus and 3′terminus of 3 phytoreoviruses

5.3 Function[1,2,19~21]

6 **Reference**

2 水稻矮缩病毒基因组数据库的内容

本数据库包括 80 年代以来的基因组研究内容,以及从 1895 年报道该病毒以来的基本信息,如 RDV 的形态、结构、理化性质等,以保证数据库的完整性。本数据库包括了与国际上相关数据库的链接。

2.1 RDV 的主要术语

包括首字母缩写、分类学地位、以及各种分离物。目前日本已经完成多种分离物的研究,中国福建分离物的全部序列已经测定,云南和浙江分离物正在研究中。

2.2 RDV 的形态和结构

包括形态学和结构两方面的内容。形态学信息有病毒颗粒的各种参数,如直径、形状、衣壳厚度等,并有病毒颗粒在电子显微镜下的照片和病毒的模式图;结构信息则有病毒颗粒的晶体学参数和通过冷冻电镜(CryoEM)获得的 50Å 和 25Å 分辨率的三维结构。

2.3 RDV 感染水稻的症状和宿主范围

说明在自然条件和实验条件下,RDV 的宿主范围和被感染并患病的水稻的症状(包括图片),另外还有病毒的传播方式、地理分布和生态控制方面的信息。

2.4 RDV 的理化性质

物理性质概括了病毒在植物提取液中和提纯后的性质,如沉降系数等;生化性质指出了病毒中核酸和蛋白质的比例以及病毒 RNA 的基本信息;还有一些抗原性质的信息。

2.5 RDV 的基因组

这是水稻矮缩病毒基因组数据库最主要的部分,不但包括了基因组的序列还有序列分析和结构功能方面的内容,共分三个部分:

2.5.1 结构 根据来自不同分离物的样品,将已经测序的所有片段以表格的形式表示,建立与每一个片段的链接。每个片段的序列数据格式与 GenBank 相似;根据文献的内容和 RDV 基因组研究的需要,加入了如下内容:主要是实验方法和每个片段性质的信息,并且根据分析结果,加入了二级结构的一些内容。数据的基本格式如表 1 所示。

表 1 RDV 基因组片段的数据格式

Table 1 RDV genome segment data format

数据项(Fields)	说明(Explanation)
定义(DEFINITION)	对基因组片段的简短定义
GenBank 注册号(GenBank ACCESSION)	该片段在 GenBank 中的编号,同时可以作为 RDV 基因组数据库到 GenBank 的链接
关键词(KEYWORD)	说明该片段性质和特征的关键词
参考文献(REFERENCE)	有关该片段研究的文献。包括作者(AUTHORS),题目(TITLE),杂志(JOURNAL)以及与 MEDLINE 或 RDV 基因组数据库文献部分的链接
实验方法(METHOD)	实验用的主要方法,如 PCR 的引物序列,反应体系和反应条件;克隆测序的载体、方法;序列分析的软件等

续表 1

数据项(Fields)	说明(Explanation)
性质(FEATURES)	包括了该基因组片段的一些特性,如长度,同源性分析结果,开放读码框(ORF)和二级结构
碱基计数(BASE COUNT)	说明片段中 A、C、U、G 的数目
序列(SEQUENCE)	该基因组片段的全部序列

2.5.2　三种植物呼肠孤病毒正链 RNA5 '和 3 '末端保守序列分析 以表格的方式将三种植物呼肠孤病毒 水稻矮缩病毒(RDV)、伤瘤病毒(WTV)和水稻瘿矮病毒(RGDV)正链 RNA5 '和 3 '端进行比较。

2.5.3　功能 根据在聚丙烯酰胺凝胶电泳(PAGE)中的迁移率,水稻矮缩病毒基因组的 12 个片段从大到小依次分别为 S1－12,其中 S1、S2、S3、S5、S7 和 S8 编码结构蛋白 P1、P2、P3、P5、P7 和 P8,其他的片段都编码非结构蛋白,S11 编码 2 个蛋白 Pns11a 和 Pns11b,S12 则有四个 ORF。有些蛋白的结构和功能已经研究清楚,有些仍在研究。每个蛋白都依次列出,并且采用了与结构基因组部分类似的数据格式,也加入了方法研究、二级结构分析和三维结构的信息。

2.6　参考文献

将有关 RDV 研究的国内外文献列出,包括仅有摘要的和可全文下载的各种文献。

3　讨论

水稻矮缩病毒基因组数据库是一个简洁、无冗余而且专用性强的二级数据库。根据数据库构建中的一些体会,我们考虑了当数据库规模变大时可以采取下述策略。

首先是数据库系统问题。在水稻矮缩病毒基因组数据库中,我们并没有采用专门的数据库系统,而采用了比较简单的超文本链接标记语言(HTML)和通用网关接口(CGI)技术。当数据库规模扩大时,就必须考虑数据库系统。在目前世界上比较大的生物学数据库中,有一部分使用商业化的数据库系统软件,如 EMBL 数据库采用 Oracle,GDB 数据库采用 Sybase,而有一部分则使用自己开发的数据存储格式和数据操作程序,如 SRS 数据库查询系统。根据需要,这两种方法都可行,使用商业软件成本比较高,但通用性好,自己开发则需要有更强的开发实力,并且要考虑国际化和通用性问题。目前我们正研究在数据库系统下构建生物数据库的方法。

另外,就是关于数据库中是否要包括实验方法的问题。在水稻矮缩病毒基因组数据库中,我们根据需要,在数据库中加入了实验方法和一些性质方面的信息。当数据库规模较大,数据项增多时,就应该采用新的方法 将实验方法的内容去除,或者可以直接从文献部分查找 也可以对实验方法专门建立一个数据库,并和序列等数据建立链接。

在数据库的文献部分,我们收集摘要和部分全文。当数据库的规模扩大,摘要内容很多时,可以建立以与其他文献数据库(如 Medline)链接为主的文献系统,有些国内发表的论文在国际文献数据库中没有,可以单独建立一个数据库来处理,以保证其完整性。

目前这个数据库已经基本建成,随着水稻矮缩病毒的研究进展,数据库的内容也会不

断更新。希望水稻矮缩病毒基因组数据库能够为从事该病毒研究的科学工作者提供方便和帮助,并能够为探索二级数据库的构建积累经验。

参 考 文 献

[1] Xu H, Li Y, Chen Z L, *et al*. *Virology*, 1998, **240**: 267～272.
[2] H. H. Zheng, Y. Li Chen Z L, *et al*. *Thero Appl Genet*, 1997, 94: 522～527.
[3] 李 玮,李 毅,陈章良,等. 应用基础与工程科学学报, 1994, 2(2): 109～115.
[4] Zhu Y, Hemmings A M, Iwasaki K, *et al*. *Journal of Virology*, 1997, 71(11): 8899～8901.
[5] Lu G, Zhou Z H, Matthew L, *et al*. *Journal of Virology*, 1998, 72(11): 8541～8549.
[6] Zhang F, Li Y, Chen Z L, *et al*. *Acta Virologica*, 1997, 41: 161～168.
[7] Murao K, Uyeda I, Ando Y, *et al*. *Virology*, 1996, 216: 238～240.
[8] Suzuki N, Watanabe Y, Kusano T, *et al*. *Virology*, 1990, 179: 446～454.
[9] Suzuki N, Tanimura M, Watanabe Y, *et al*. *Virology*, 1990, 179: 455～459.
[10] 高 谦,欧阳新,刘 玮,等. 植物学报, 1990, **31**(1): 13～18.
[11] 李 玮,李 毅,陈章良,等. 病毒学报, 1995, **11**(1): 56～62.
[12] 李 毅,薛志宏,陈章良,等. 病毒学报, 1994, **4**: 339～345.
[13] 刘一飞,李 毅,陈章良,等. 病毒学报, 1994, 10(3): 246～250.
[14] 鲁瑞芳,李 毅,陈章良,等. 微生物学报, 1998, **39**(4): 305～314.
[15] 曲 林,李 毅,朱玉贤,等. 病毒学报, 1995, **11**(5): 271～275.
[16] 曲 林,李 毅,陈章良,等. 微生物学报, 1996, **36**(5): 335～343.
[17] 肖 锦,李 毅,陈章良,等. 生物工程学报, 1996, **12**(3): 361～366.
[18] 肖 锦,李 毅,陈章良,等. 微生物学报, 1998, **38**(5): 348～358.
[19] Suzuki N, Kusano T, Matsuura Y, *et al*. *Virology*, 1996, **219**: 471～474.
[20] Tomaru M, Maruyama W, Kikuchi A, *et al*. *Journal of Virology*, 1997, **71**(10): 8019～8023.
[21] Mao Z J, Li Y, Chen Z L, *et al*. *Arhives of Virology*, 1998, **143**: 1831～1838.

CONSTRUCTION OF RICE DWARF VIRUS GENOME DATABASE*

Wang Jianmin Luo Jingchu Li Yi Qu Hong Wu Guangyao Gu Xiaocheng**

(*College of Life Sciences*, *Peking University*, *Beijing* 100871, *China*)

Abstract: Secondary database construction is an important subject in the field of bioinformatics. As the full genomic sequences of some organisms are being completed and followed by structural and functional studies, construction of secondary database becomes essential on the agenda. The rice dwarf virus (RDV) is a pathogen infecting rice in China, Japan and the Southeastern Asia region and leading to considerable economic loss. Based on the data generated from recent genomic research and earlier biochemical studies scattered in various primary databases and scientific journals, we have constructed a compact, user-friendly and non-redundant job-oriented secondary database. This work will provide compiled useful information for plant molecular biologists as well as in achieving preliminary experiences in secondary database construction.

Key words: Bioinformatics, Secondary database, Rice Dwarf Virus, Genome

* The work was supported by National Nature Sciences Foundation (39993420, 39870027)

** To whom correspondence should be addressed