研究报告

中国微生物资源数据库（MRDC）

马俊才　赵玉峰　王大耜　张克莉　刘澎涛

（中国科学院微生物研究所，北京　100080）

摘要　中国微生物资源数据库 MRDC 是以中国的微生物资源为数据源的大型综合数据库，包括目前已建成的微生物性状库、微生物产品市场信息库、微生物名称库、微生物名词库、国际交流用 RKC 代码库及一个微生物学数值分类软件包 MINTS。

关键词　数据库技术；微生物资源的利用；计算机应用

当第四纪冰川覆盖全球时，中国某些地块曾免遭覆盖。中国有世界上独特的高盐、高温、高酸、高碱、高盐碱、高酸热、高干旱、高辐射、高浓度重金属离子等极端环境，形成了生物尤其是微生物特有的生境和生态系。国外生物学家称中国是世界的生物种质库。

多年来国家对微生物分类、生态、资源调查等研究和开发项目投入了大量人力物力。我国拥有实力雄厚近万人的微生物学专业队伍，从中国特有生境采集的菌种为试验材料的研究开发取得了硕果，发现了许多新的微生物属、种，积累的大量资源菌株和数据资料，成为我国和全人类的宝贵财富。

根据当代信息网络技术和数据库技术发展现状，结合国情和中国科技、经济发展战略，在“七五”计划期间，中国科学数据库决定建立我国的微生物学专业子库——中国微生物资源数据库(MRDC)。使之成为我国微生物信息交流的枢纽和供国际交流的窗口。

中国微生物资源数据库(MRDC)是我国微生物学界第一个不局限于某个菌种保藏中心、立足于整个中国微生物资源、所有配套软件自行开发的大型综合性资源数据库。到目前为止，MRDC 已建成了微生物性状库、微生物产品市场信息库、微生物名称库、微生物名词库、国际交流用 RKC 代码库及一个微生物学数值分类软件包 MINTS。

中国微生物资源数据库的建设分为两部分。一是，围绕数据准备的数据工程，二是，针对微生物资源数据库提出的各种功能实现软件设计。

（一）数据工程

1. 数据源评价及可信性估计[1]：微生物资源数据库在数据源的选择上是严格的。凡是入库的数据必须是产生数据源条件较好的单位。为照顾数据来源的多样性，除了采用国家一级刊物的资料外，其他符合要求的实验数据均有在本研究领域富有经验的专家负责。

2. 数据源的规范化[1]：在微生物资源数据库的建设过程中，数据源的规范化是关键，为此按细菌、放线菌、酵母菌等制定一系列规范或暂行规定以保证数据质量。

3. 建立合理的录入流程，保证数据加工及录入质量：为了确保数据质量，未经专家评审同意的数据不予采纳。采纳的数据都按照登录、选择、整理、编码、输入、复核这一基本流程，并采用平行录入逐级复检等技术手段。另外还编制了一系列数据录入和质量控制软件，用计算机进行质量控制。

（二）数据库系统

1. 系统的硬软件环境：硬件系统为长城 286 型微机，内存 1 兆字节，84 兆快速硬盘，彩色高分辨率显示器，1.2M 和 360K 软盘驱动器各一个，LQ 1500 打印机。

操作系统为 DOS3.3，配置 CW 中文录入软件，XE 中文字处理软件。

数据库管理软件，我们采用美国 Fox Software 公司 1987 年推出的微机数据库管理软

件-FoxBASE + 2.0 版。数值分类系统 MINTS 采用编译 BASIC 语言。

2. 系统的构成:

(1) 微生物性状库[2]: 微生物性状库是微生物资源数据库组成中的关键，微生物的开发利用无不建立在微生物的性状之上。

① 性状库特点: a. 数据库中的数据全部都是与中国微生物学家目前正在研究、开发的微生物资源有关的数据，每条记录的信息都与一个活菌种相对应; b. 数据项完整，菌株覆盖面大: 在详细记载微生物性状信息的同时，还特别着重存贮了有关菌株来源、分离基物及分离源等信息。分离基物涉及人体、温泉、土壤、海底淤泥、动物体内等; c. 软件功能集数据录入、检索、统计绘图和数值分类于一身，本库除包含一般的数据录入、性状检索功能以外，还重点增加了交叉复合检索和各项数据的统计绘图功能。此外，还沟通了数据库与数值分类软件包之间的联系，使数据库的检索结果能够自动进行数值分类。

② 性状库的数据结构: 该库由十个表(子库)组成，分别是: 概述表、别号表、用途表、培养基成分表、参考文献表、菌株历史表、菌株别名表、标准菌株说明表、菌株性状表、性状名称表。

概述表是概要记载每株菌的基本信息的表。在本表的设计中，我们将栖息地水平分为3个等级，既可以使用户对栖息地进行不同层次的检索，又可满足生态学对栖息地做各种统计分析。概述表中的上述信息基本上每株菌都具备，因此统一放在一个表中，每株菌占一条记录。

概述表的主要数据项包括: 菌号、拉丁菌号、鉴定单位、定名者、标准菌株否、分离省县、分离具体地、分离号、分离单位、分离者、分离日期、生态环境、具体栖息地、分离地经、纬度和高度、培养基号、培养温度、存活否、血清型、保藏单位、保藏者、保藏方式、提供单位、提供者、入库时间、更新时间等。

其它各表主要记录该菌株在其它保藏机构中的编号; 菌株的用途、开发状况以及开发者; 培养基成分及其制作方法; 参考文献刊物名称、文章涉及的领域以及文献语种; 菌株的历史; 标准菌株的说明; 菌株以前被鉴定过的名称; 菌株的各项性状名称、取值以及与 RKC 代码之间的关系。这些表的信息视菌株信息源而定，不一定每株菌必有一条记录。

③ 微生物性状库的主要软件功能:

从固定项检索: 主要包括经常检索的数据项，如从菌号、拉丁菌名、菌株用途、培养基编号、菌株别号、别名等进行检索。检索命中后，调出该菌的全部信息。

从性状检索: 可对所有性状任意进行与、或、非运算。检索命中后，调出该菌的全部信息。

复合检索: 是指对概述表(基本信息)中的所有数据项目进行任意条件的逻辑组配，同一次检索中关键字最多不得超过 10 个，即一次检索最多可使用 10 个数据项。例如，我们希望在经度 96—111，纬度 20—35 的地理环境(即我国的西南地区)中检索从"土壤"中分离的菌株，从回车到调出第一条符合条件的记录的反应时间约为 5 秒，整个检索过程约需 55 秒，命中 395 条记录。由于该项功能可作用于概述表(基本信息)中的所有项目，并且可以多个数据项目相复合，满足不同的检索要求。例如刚才检索的实例中，"西南地区"几个字并没有实际存贮在数据库中，但通过多个数据项复合检索，可达到检索西南地区菌株资源的目的。

统计功能: 设立本功能的目的一是，随时掌握库中的数据情况; 二是，及时纠正库中的错误数据。因为库中所有数据排序后，很容易发现错误数据(特别是名称或关键词类的拼写错误)。最后以数字、文字或柱状图形方式显示各种统计结果。该功能主要针对以下几方面:

a. 统计菌株的各种分离基物: 由于已将分离基物划分为三个等级，因此使作用于其上的统计功能成为可能。可使我们脱离具体分离基物，从较为宏观的角度对分离基物进行统计分析。例如，可以通过统计了解分离自"植物"的

或"土壤"的菌株数目，不论该菌株来自哪一种植物或哪一类土壤均被计数。

b. 统计菌株的分离地：由于将菌株的分离地分解为两个等级，即省及直辖市和具体分离地两个水平。使我们可以从"省"的水平上对分离地进行统计分析。例如我们可以随时统计全国三十一个省、市、自治区（台湾省资料暂未录入）分离的菌株数目。

c. 统计"属"或"种"的数目：由于将菌株的拉丁名作为一个单独的数据项来存贮，我们可以随时统计数据库中的菌株一共包含多少个"属"和"种"，以及隶属于每个"属"或"种"的菌株数。

d. 其它统计功能：除上述统计功能外，还可对菌株的分离、鉴定、提供、保藏者以及用途、保藏方式、参考文献、菌株别号等方面进行统计。

数值分类功能[3-5]：我们于 80 年代初期建成了微生物学数值分类软件包 MINTS，它除了包括常用的相似性计算公式和聚类方法外，还可以自动生成并绘制树状谱。几年来，用该软件包为几十个微生物分类学课题进行了几百次数值分类计算。用户用 MINTS 软件包计算的分类结果撰写的论文已达几十篇。在中国微生物资源数据库总体设计方案中，我们有意沟通了数据库与 MINTS 软件包之间的联系，系统可以自动将用户的任意一次检索结果所得到的菌株性状生成一个可供 MINTS 软件包识别并进行分类计算的数据文件。这项功能具有以下几种益处：

a. 用户可按自己意愿对数据库中数据进行统计分析、分类鉴定，或从数据库提取自己工作所需的菌株数据集合，使数据库不仅仅限于检索查询，从而提高了数据库的利用率。

b. 使数值分类计算摆脱了过去只能对数据提供者本人提供的数据做数值分类计算，无法对进行同类研究工作他人发表的结果或不同类别的菌株在同一个水平上同时做纵向或横向的分析对比。

c. 利用本功能不同的聚类策略，利用数据库存贮不同类别菌株繁多的属种性状，按分类学家的意愿进行多层次多方位的统计分析，为分类学家寻找新的分类系或鉴定路径提供重要依据。

(2) 微生物名称库：名称库是专门记录菌株的拉丁菌名和中文菌名及其命名者的数据库，既可以作为性状库中菌名标准译名的辅助库，又可以作为一个单独的菌名数据库使用。目前已收录部分细菌、放线菌和枝原体等菌名 6785 个。

主要数据项目有：拉丁菌名、命名者、中文菌名、菌名类型等。

主要软件功能有：从菌株的拉丁名/中文名检索、从菌株的中文种名/拉丁文种名检索、从拉丁文属名和种名的字首部分检索、从菌株命名者检索等。

(3) 微生物名词库：名词库目前是一个独立的数据库，它收录了经国家名词审定委员会审定的 1674 条微生物名词，作为微生物学中的规范化名词。当国家名词委员会继续公布新的微生物名词，本库的库容进一步扩大之后，我们将利用此库作为性状库中微生物性状描述的标准中英文词库。

主要数据项有：名词编号、中文名词、英文名词以及注释等。

主要软件功能：从中文名词/英文名词检索。

(4) RKC 代码库：RKC 代码是将微生物性状的自然语言描述转换为数值描述的一种编码。该代码目前已获得国际科学数据委员会（CODATA）的推荐。我们将 RKC 代码全部翻译成中文并存入数据库，目前已收录全部 RKC 代码 1 万余个。对于 MRDC 性状库中的每一条性状我们都确定了其最适合的 RKC 代码，为将来的国际数据交流奠定基础。

主要数据项有：RKC 代码、中文描述、英文描述。

主要软件功能：从指定的 RKC 代码检索，整节检索 RKC 代码，检索含指定的中文/英文关键词的 RKC 代码等功能。

(5) 微生物产品市场库：由于微生物产品具有十分可观的经济效益，因此为人们所重视。微生物产品市场库的建立旨在随时掌握我国微生物市场动态及微生物行业现状，为微生物产品的开发利用提供依据。

主要数据项有：微生物名称、微生物产物名称、用途类型、开发单位、开发现状、有无专利、开发工艺、直接信息源、间接信息源、入库时间等。

主要软件功能有：从微生物产品检索、从微生物检索、复合检索等功能。

目前中国微生物界最大的菌种保藏组织——中国微生物菌种保藏管理委员会 CCCCM，已开始在该系统内建立统一数据结构的菌种保藏用数据库。该库以中国微生物资源数据库 MRDC 的数据结构和软件功能为蓝本，由中国科学院微生物研究所牵头，联合卫生部、农业部、农业科学院、林业科学院、全国医药管理总局等单位共同进行。并将据此出版第一本英文版《中国菌种目录》。本数据库的建立，必将大大提高中国微生物界菌株数据的保存质量，并能进一步促进中国微生物资源的开发利用以及菌种保藏质量的提高。

由于中国微生物资源数据库 MRDC 的建设时间较短，不论在数据量的积累以及软件功能上都有待进一步加强和完善。我们深信具有中国特色的微生物资源数据库必将日趋完善，它对国内外微生物信息的交流和应用必将产生深远的影响。

参 考 文 献

1. 赵玉峰、马俊才等：《全国第一届科学数据库讨论会论文集》，第 169—178 页，科学出版社，1992。
2. 马俊才、赵玉峰等：微生物学通报 17(4)：210—214，1990。
3. 马俊才、赵玉峰：微生物学通报，13(5)：225—228，1986。
4. 马俊才、赵玉峰等：微生物学通报，16(1)：37—40，1989。
5. 马俊才、赵玉峰：《微生物分类和鉴定技术进展》，第 340—355 页，光明日报出版社，1989。

THE MICROBIAL RESOURCES DATABASE OF CHINA (MRDC)

Ma Juncai　Zhao Yufeng　Wang Dasi　Zhang Keli　Liu Pengtao

(*Institute of Microbiology Academia Sinica, Beijing* 100080)

MRDC is the first self-developing large integrated microbial resources database in the Chinese microbiological field. It includes Microbial Character Bank, Microbial Product Market Information Bank, Microbial Names Bank, Microbial Nomenclature Bank, RKC Code Bank for International exchange and one Microbiological Information Numerical Taxonomic System named MINTS.

Key words Utilization of microbial resources; Database technique; Application of computer