ICS 07.080
C 04

SZDB/Z

深 圳 市 标 准 化 指 导 性 技 术 文 件

SZDB /Z 92—2014

生物基因信息数据库建设与管理规范

2014－01－23 发布　　2014－02－01 实施

深圳市市场监督管理局　　发 布

目　　次

前　言

本标准按照GB/T1.1—2009给出的规则起草。

本标准由深圳市经济贸易和信息化委员会归口。

本标准负责起草单位：深圳华大基因研究院、深圳市标准技术研究院。

本标准主要起草人：张勇、严志祥、操利超、陈凤珍、肖萍、袁翠红、陈欢。

本标准为首次发布。

引　言

生物基因信息数据是21世纪的重要战略资源，生物基因信息数据库是促进生物基因数据共享和利用的重要基础平台，也是保护重要基因数据资源的有效手段。

随着人类基因组测序计划的完成，基因测序技术得到快速发展，特别是第二代高通量测序技术的出现，产生了大量的生物基因信息数据。西方发达国家一般都拥有具有代表性的生物基因信息数据库，如美国国家生物技术信息中心管理的核酸序列数据库、欧洲生物信息学研究所管理的核酸序列数据库及日本国家遗传学研究所管理的核酸序列数据库。在我国，虽然生物基因信息数据库的建设在快速发展，但是大多数生物基因信息数据库的数据量属于中小规模水平。目前国内还没有针对性的国家标准、行业标准和地方标准，导致在搜集和整理基因数据资源时缺乏依据，在建设生物基因信息数据库时难以保证数据的准确性、完整性和安全性。因此，利用标准化手段，制定生物基因信息数据库建设规范，可更好的指导基因信息数据库规范化建设，促进基因数据资源的共享和利用，助推我国生物产业快速健康发展。

生物基因信息数据库建设与管理规范

1 范围

本标准规定了与生物基因信息数据库建设相关的设备、环境的基本要求以及生物信息数据的处理方法和原则。

本标准适用于生物基因信息数据库的建设，以及生物基因信息数据的采集、处理、存储、备份和使用。

2 规范性引用文件

下列文件对于本文件的应用是必不可少的。凡是注日期的引用文件，仅所注日期的版本适用于本文件。凡是不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB 50052-2009 供配电系统设计规范

GB 50054-2011 低压配电设计规范

GB 50057-2010 建筑物防雷设计规范

GB 50116-1998 火灾自动报警系统设计规范

GB 50174-2008 电子信息系统机房设计规范

GB 50189-2005 公共建筑节能设计标准

GB 50222-1995 建筑内部装修设计防火规范

GB 50243-2002 通风与空调工程施工质量验收规范

GB 50254-1996 电气装置安装工程低压电器施工及验收规范

GB 50311-2007 综合布线系统工程设计规范

GB/T 50314-2006 智能建筑设计标准

GB 50370-2005 气体灭火系统设计规范

SJ/T 10796-2001 防静电活动地板通用规范

YD/T 1095-2008 通信用不间断电源（UPS）

3 术语与定义

下列术语和定义适用于本文件。

3.1

比对 alignment

指将两个或多个序列排列在一起,标明其相似之处。序列中可以插入间隔（通常用短线“-”表示）。对应相同或相似的符号（在核酸中是A、T（或U）、C和G，在蛋白质中是氨基酸残基的单字母表示）排列在同一列上。

3.2

期望值 E-value

比对软件中使用的统计值，表示因为随机性而获得等于或优于当前比对结果的可能次数。E 值越小，随机发生这一事件的可能性越小，比对结果越显著。

3.3

P 值 P-value

比对软件中使用的统计值，表示因为随机性而获得等于或优于当前比对结果的可能性。P 值越小，比对结果的可信度越大，P 值越大，比对结果来自随机匹配的可能性越大。

3.4

读长 read length

高通量测序仪产生的序列标签的长度。

3.5

重叠群 contig

基因组测序过程中将多个短的序列片段拼接成较长的连续片段。

3.6

支架 scaffold

基因组测序过程中，基于构建文库获得的一定大小片段两端的序列确定一些重叠群之间的顺序关系，将先后顺序已知的重叠群拼接得到的更长的连续片段。

3.7

间隙 gap

基因组测序时,有些序列不能被测定，在组装和拼接时两序列不能连接形成的空隙。

3.8

数据抽象化 data abstraction

将数据以它的语义来呈现出它的外观，但是隐藏起它的实现细节，用来减少数据的复杂度,使得工作人员只需要关注数据少数重要的部分。

3.9

基因组 genome

单倍体细胞中的全套染色体或全部基因。

3.10

转录组 transcriptome

在某一生理条件下，细胞内所有转录产物的集合，包括信使 RNA、核糖体 RNA、转运 RNA 及非编码 RNA。

3.11

生物信息分析数据 biological information analysis data

应用数学、信息学、统计学和计算机科学的方法对生物测序数据进行归纳和注释等分析后得到的数据。

3.12

键值对 key-value pair

一对属性名和属性值，属性名和属性值中间用等号连接的数据结构。

3.13

生物数据存储 biological data storage

将生物测序数据和生物信息分析数据以某种格式记录在计算机内部或外部存储介质上的过程。

3.14

直接附加存储 direct-attached storage

将存储设备直接连接在服务器内部总线上，数据存储设备作为整个服务器结构的一部分。

3.15

网络附加存储 network-attached storage

存储系统直接通过网络接口与网络直接相连，由用户通过网络访问的存储方式，而不是附属于某个特定的服务器。

3.16

存储区域网络 storage area network

一种通过光纤集线器、光纤路由器、光纤交换机等连接设备将磁盘阵列和磁带等存储设备与相关服务器连接起来的高速专用子网。

3.17

完全备份 full backup

指对某一个时间点上的所有数据或应用进行的一个完全拷贝。

3.18

增量备份 incremental backup

备份上一次备份（包含完全备份、差异备份和增量备份）后数据发生的所有变化。

3.19

差分备份 differential backup

备份上一次的完全备份后发生变化的所有数据。

3.20

局域网络备份 local area network backup

在局部区域网络中配置一台中央备份服务器，数据全部通过网络传输到本地备份服务器上进行备份。

3.21

无局域网备份 local area network free backup

在存储区域网络环境下,利用存储区域网络对数据进行传输和备份,而不用占用局部区域网络资源。

3.22

存储区域网无服务器备份 storage area network server-free backup

备份过程在存储区域网络内部完成，大量数据流无需流过服务器，可极大降低备份操作对生产系统的影响。

3.23

恢复点目标 recovery point objective

衡量灾备系统的重要指标之一。当灾难或紧急事件发生时，数据可以恢复到的时间点。

3.24

恢复时间点 recovery time objective

衡量灾备系统的重要指标之一。将信息系统“从灾难造成的故障或瘫痪状态恢复到可正常运行状态，并将其支持的业务功能从灾难造成的不正常状态恢复到可接受状态”所需时间。

3.25

持续数据保护 continuous data protection

一种在不影响主要数据运行的前提下，可以实现持续捕捉或跟踪目标数据所发生的任何改变，并且能够恢复到此前任意时间点的方法。

4 缩略语

下列缩略语适合于本文件。

CDP：持续数据保护（Continuous Data Protection）
CDS：编码序列（Coding Sequence）
DAS：直接附加存储（Direct-Attached Storage）
DNA：脱氧核糖核酸（Deoxyribonucleic Acid）
FTP：文件传输协议（File Transfer Protocol）
IT：信息技术(Information Technology)
LAN：局部区域网络（Local Area Network）
NAS：网络附加存储（Network-Attached Storage）
PUE：电源使用效率（Power Usage Effectiveness）
RNA：核糖核酸（Ribonucleic Acid）
RPO：恢复点目标（Recovery Point Objective）
RTO：恢复时间目标 （Recovery Time Objective）
SAN：存储区域网络（Storage Area Network）

5 生物基因信息数据库建设规划

5.1 建库目标

在建设生物基因信息数据库时应明确建库目标，包括数据类型、可能使用该数据库的人群及流量等。信息数据库应具有较好的灵活性、稳定性和可移植性，且对运行环境有较强的适应能力。

5.2 建库对象

建库对象是指生物基因信息数据，包括 DNA 和 RNA 数据。

6 数据库机构

6.1 伦理审查委员会

6.1.1 生物基因信息数据库建设过程中应设立伦理审查委员会，建立生物基因信息数据库的伦理准则，对生物基因信息数据来源、生物基因信息数据使用和共享等涉及到伦理问题的地方进行伦理审查。

6.1.2 伦理审查委员会的成员组成应涉及多学科和多部门，并且有能力对所有伦理问题进行审查和评价。

6.2 科学审查委员会

6.2.1 科学审查委员会应由生物信息分析领域和生物科学领域的专家组成。

6.2.2 科学审查委员会负责对生物基因数据的完整性、正确性及可用性进行审查。

6.3 设备管理机构

6.3.1 设备管理机构应由计算机硬件维护工程师和计算机软件维护工程师组成。

6.3.2 设备管理机构负责日常计算机硬件系统和计算机软件系统的搭建、运行和维护工作。

6.4 数据库管理机构

6.4.1 数据库管理机构负责数据库建设的规划、组织协调和运营。

6.4.2 数据库管理机构对数据库的工作有最终的审核和批准权。

6.5 数据库执行机构

6.5.1 数据库执行机构负责数据库日常工作的执行和落实。

6.5.2 数据库的日常工作包括数据采集、数据分析、数据存储、数据使用、数据备份、恢复管理、日常行政和财务工作。

7 数据库管理

7.1 数据采集

7.1.1 采集的数据包括从已有的生物基因信息数据库 FTP 下载的数据、通过试验终端产生的原始数据和对原始数据进行生物信息分析得到的数据。从已有生物基因信息数据库下载的数据要求标明数据来源，宜保持跟数据源一样的文件结构。采集的数据格式包括 FASTQ、FASTA、GFF 和其它特殊格式。

7.1.1.1 FASTQ 格式

FASTQ格式的序列以及质量信息均使用一个ASCII字符表示。FASTQ文件中每个序列应有四行：

——第一行是序列标识以及相关的描述信息，以“@”开头；

——第二行是序列；

——第三行由“+”开头，后面可以添加序列的描述信息；

——第四行是序列的测序质量值，每个字符与第2行每个碱基对应。

示例：

@SEQ_ID

GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT

+

!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65

7.1.1.2 FASTA 格式

FASTA序列由标准的IUB/IUPAC核酸代码表示,如表1所示。FASTA文件每行不应超过80个字符。一个文件可以有多条序列。FASTA 序列格式包括三个部分：

——第一部分为注释行，以“>”开头，后面是序列的名称和来源等信息；

——第二部分是标准的单字符标记的序列；

——第三部分为可选的“*”表示序列的结束，此符号也可不出现。

示例：

>gi|187608668|ref|NM_001043364.2| Bombyx mori moricin （Mor）, mRNA

AAACCGCGCAGTTATTTAAAATATGAATATTTTAAAACTTTTCTTTGTTTTTA

TTGTGGCAATGTCTCTGGTGTCATGTAGTACAGCCGCTCCAGCAAAAATACCT

表1 IUB/IUPAC 核苷酸代码表

代码	碱基	英文含义	中文含义
A		Adenine	腺嘌呤
G		Guanine	鸟嘌呤
C		Cytosine	胞嘧啶
T		Thymine	胸腺嘧啶
U		Uracil	尿嘧啶
R	(A/G)	puRine	嘌呤
Y	(C/T/U)	pYrimidine	嘧啶
M	(A/C)	aMino	腺嘌呤或胞嘧啶
K	(G/T)	Ketone	鸟嘌呤或胸腺嘧啶
S	(C/G)	Strong interaction	强相互作用碱基
W	(A/T)	Weak interaction	弱相互作用碱基
H	(A/C/T)	Not-G (H after G)	非鸟嘌呤
B	(C/G/T)	Not-A (B after A)	非腺嘌呤
V	(A/C/G)	Not-T/U (V after U)	非胸腺嘧啶
D	(A/G/T)	Not-C (D after C)	非胞嘧啶
N	(A/C/G/T)	aNy	不确定碱基
-		gap of indeterminate length	不确定长度的间隙

7.1.1.3 GFF 格式

GFF 格式一般用于基因序列注释，由 Tab 键隔开的 9 列组成：

——第一列为序列的编号，编号的有效字符为大小写字母（a-z和A-Z）、数字（0-9）和一些特殊

字符（.:^*$@!+_?-|）；

——第二列为注释信息的来源，比如“Genescan”和“Genbank” 等，可以为空，或用“.”代替；

——第三列为注释信息的类型，比如Gene、cDNA和mRNA等；

——第四列和第五列为开始与结束的位置，计数由1开始，结束位置不能大于序列的长度；

——第六列为得分，是注释信息可能性的说明，可以是序列相似性比对时的E-value值或基因预测的P-value值，“.”表示为空；

——第七列为序列的方向，“+”表示正义链，“-”表示反义链，“？”表示未知；

——第八列为起始编码的位置，仅对注释类型为“CDS”有效，有效值为0、1和2；

——第九列为序列的属性，以多个键值对组成的注释信息描述，键与值之间用“=”，不同的键值用“；”隔开，一个键可以有多个值，不同值用“,”分隔。

示例：

NC_015866.1	RefSeq	region	1	12781	.	+	.	ID=id0;Dbxref=taxon:1032845sylvarum;strain=054;
NC_015866.1	RefSeq	gene	396	1217	.	+	.	ID=gene0;Name=Rh054_00005;Dbxref=GeneID:11014636;
NC_015866.1	RefSeq	CDS	396	1217	.	+	0	ID=cds0;Name=YP_004763714.1;Parent=gene0;
NC_015866.1	RefSeq	gene	132	1641	.	+	.	ID=gene1;Name=Rh054_00010;Dbxref=GeneID:11013311;

7.1.1.4 其它特殊序列格式

在进行数据处理过程中，不同的生物信息分析软件得到的数据格式各不相同。这些数据分析结果也是重要的数据采集对象，如 bed、wig 和 sam 等格式文件。

7.1.2 不同数据格式之间可以用特定的软件进行相互转换。在数据采集过程中，可采集其中一种或多种格式的数据。

7.1.3 在数据采集过程中，应对数据进行严格的科学性审查。对于涉及人的基因信息数据，应进行伦理审查。

7.2 数据处理

7.2.1 数据处理对象为采集的生物信息数据，主要为 DNA 和 RNA 数据。

7.2.2 数据处理过程主要包括核酸序列数据格式转换和生物基因信息原始数据分析。

7.2.3 不同的数据类型采用不同的分析流程和分析软件，基因信息分析流程见图 1。

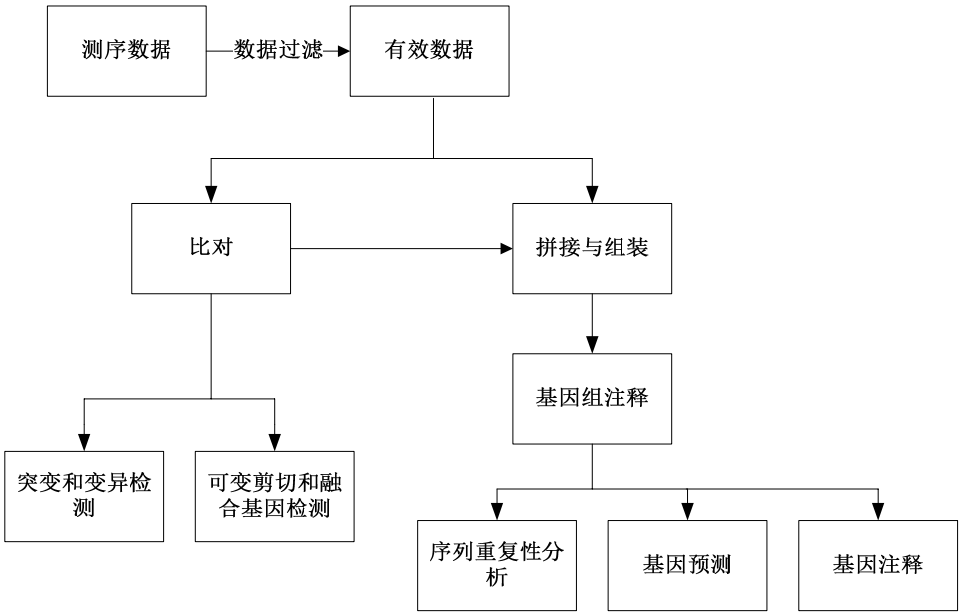

图1　基因信息分析流程图

注1：高通量测序仪的下机原始数据，应去接头，去除不符合要求的数据。

注2：序列比对，比较测序序列和参考序列的相似性或不相似性，寻找核苷酸的连续产生模式，找出基因序列中的信息成分。

注3：突变和变异检测，利用生物信息分析软件分析插入、缺失、倒位或转位和拷贝数变异等。

注4：含有参考基因组的 RNA 测序，应进行可变剪切及融合基因检测。

注5：序列拼接和组装，利用拼接组装软件将测序得到的读长拼接为重叠群，把重叠群拼接成支架。

注6：基因组注释，利用生物信息学方法和工具，对基因组所有基因的生物学功能进行注释，包括序列的重复性分析、基因预测和基因注释。

7.3　数据存储

7.3.1　存储方式

7.3.1.1　数据存储之前，应先对将要存储的数据进行抽象化，再进行分类。宜按下列几种方式分类：

a)　按照数据对象，可分为基因组学数据、转录组学数据、表观组学数据和小 RNA 数据等；

b)　按数据类别，可分为生物测序数据和生物信息分析数据；

c)　按数据的物种来源，可分为原生生物界、植物界、动物界和真菌界，也可进一步细分为门、纲、目、科、属和种；

d)　按数据重要性，分为保密数据和可公开数据。保密数据是指数据所有者要求进行保密的和重要的不能公开给公众的数据；可公开数据是除保密数据外可以对外公开的数据。

7.3.1.2　在存储中，所有存储内容应层次分明，条理清晰，可以综合运用 7.3.1.1 提到的几种方法。

7.3.1.3　存储数据时，应建立数据间的逻辑与物理对应关系。可用 DAS、NAS 和 SAN 三种存储方式中的一种或多种。针对保密数据要求采用 SAN 或者 NAS 存储方式，确保可靠性。

7.3.2　存储管理原则

存储管理原则包括：

a） 应根据存储内容的数据对象和重要性进行分类存储并做好记录，针对不同的对象和数据制定不同的数据保护策略实现不同的 RPO 和 RTO：对于重要数据和对象采用可实现 CDP 效果的存储方式，对于一般数据和对象采用备份或容灾存储技术，实现双份数据冗余；

b） 确定存放在当前数据库系统中的在线数据的有效性，及时清理过时和冗余数据，暂时不用的数据转移到备份专用存储介质中；

c） 应妥善保管存储有数据的备份专用存储介质，保证存放的物理环境，避免对备份数据的非授权访问；

d） 同一数据应有两个以上的副本，且存放在不同位置；

e） 硬件和软件更新换代时，应明确每一项数据的去留，做好转移或清理工作，针对存储介质的升级要实现无缝操作，跨代兼容；

f） 对于存储在备份专用存储介质中的数据，应保证能正常使用；

g） 对于需要归档的历史数据，应进行归档；

h） 所有存储事件均应做好存储日志并填写存储日志表。数据存储日志应包含数据类别、数据存放介质、数据时效性、数据完整性、数据更新情况和数据管理人员等。

7.4 数据使用管理

7.4.1 应按存储文件的用户分组和重要性设置相应的使用权限。文件的用户分组可分为文件属主、同组用户和其他用户，使用权限分为可读、可写、可执行和无权限，具体说明见表 2，可将不同的权限分给不同的分组用户。

表2 使用权限及相应说明

权限	符号	权限说明（在 linux 或 unix 系统下适用）
可读	r	文件可以被显示或者拷贝，只有读权限不能够删除或移动文件
可写	w	文件可以被修改、移动和删除
可执行	x	文件可以被执行
无权限	-	权限被拒绝

7.4.2 应由数据库管理机构任命系统管理员、操作管理员和一般管理员，其职责说明见表 3。

表3 数据库管理机构岗位及相应职责说明

岗位	职责说明
系统管理员	隶属于数据库管理机构，具有存储系统最高管理权限，负责存储系统的日常管理和运营工作
操作管理员	隶属于数据库执行机构，负责存储系统的日常维护和运行工作
一般管理员	隶属于数据库执行机构，负责数据使用账号审核和发放

7.4.3 除管理员账户外，一般账户通过向一般管理员提供申请表获得。账号申请表应包含申请人、申请时间、申请权限、申请说明、批准人和批准日期等。

7.5 数据备份管理

7.5.1 数据备份前，应确定备份策略。备份策略包括完全备份、增量备份和差分备份。

7.5.2 首次备份宜采用完全备份的方式，以后定期进行完全备份、增量备份或差分备份。

7.5.3 数据备份应填写数据备份申请表。数据申请表应包括备份的内容描述、申请人、申请时间、使用介质的标识、数据需要备份的原因和管理员审批意见等。

7.5.4 数据备份过程中，应为文件和文件夹创建备份标记，备份时只需选中那些有标记的文件和文件夹，备份完成后应立即清除备份标记。

7.5.5 数据备份宜采用如下流程：

a） 操作管理员收集并汇总备份需求；
b） 操作管理员根据备份需求制定备份方案；
c） 操作管理员提交备份方案给系统管理员审核，经系统管理员审批后，操作管理员方可进行数据备份；
d） 操作管理员在进行数据备份过程中，应填写备份日志并存档。数据备份日志应包括备份内容、备份时间、备份情况、介质标识、备份操作人员签名和管理员签收介质等。

7.5.6 备份管理应遵循以下原则：

a） 数据备份介质应采用性能可靠和不宜损坏的介质；
b） 数据备份应及时，并对数据进行异地备份；
c） 存放备份数据的介质应有明确的标识；
d） 备份介质应由操作管理员保管，任何人不得擅自取用，应填写介质管理日志。管理介质日志应包括介质标识、是否新介质、借出时间、归还时间、借用人员签字和归还确认等；
e） 应定期对备份数据进行验证；
f） 针对海量数据备份需要采用重复数据删除和压缩技术，减少备份窗口和备份介质容量使用。

7.6 数据恢复管理

7.6.1 数据恢复宜采用如下流程：

a） 发现数据损坏或丢失，需要恢复数据时，应填写数据恢复申请表并向操作管理员提交申请。数据恢复申请表应包括需恢复的内容、申请人、申请时间、使用介质的标识、数据需要恢复的原因和数据管理员审批意见等；
b） 操作管理员收集数据恢复需求并制定数据恢复方案；
c） 操作管理员提交数据恢复方案给系统管理员审核，经系统管理员审批并签字同意后，操作管理员开始进行数据恢复；
d） 操作管理员完成数据恢复后，应填写数据恢复日志并存档。数据恢复日志应包括恢复内容、恢复时间、恢复操作人员、恢复情况、介质标识和操作人员签名。

7.6.2 应妥善保存备份恢复的审批文档及备份恢复工作的日志。

7.7 数据库搭建

7.7.1 数据库设计原则

7.7.1.1 数据库设计应具有良好的性能,具有可移植性和可扩展性。

7.7.1.2 数据库中各种对象的命名和程序代码的编写应规范，数据库应用系统应能够适应不同的数据库平台。

7.7.1.3 合理设计表间关联，尽可能的降低数据的冗余，保证数据的一致性和完整性。

7.7.1.4 根据计算机硬件和网络设计确定情况对数据库进行逻辑设计和物理设计,根据应用系统的事物大小和服务器的性能调整数据库服务器的系统参数,选择合适的数据类型,尽可能使数据库性能达到最佳。

7.7.2 数据库搭建流程

7.7.2.1　建库前应进行全面的需求分析，全面考虑可能的扩充和修改,并设计远程数据接口。

7.7.2.2　概念结构设计要求设计出能真实反映客观事物的模型,同时让设计的模型能易于理解。

7.7.2.3　逻辑结构和物理结构设计应对数据进行规范化并进行数据流分析，使数据库的存储结构和配置达到最佳。

7.7.2.4　数据库实施阶段要求定义数据库的结构，对数据库进行试运行。

7.7.2.5　数据库的使用和维护阶段要求注意数据的转存和恢复,对数据库进行安全性和完整性控制;对数据库进行性能监督,分析和改造数据库。

8　硬件设备要求

8.1　电子信息系统

电子信息系统设备应满足基因信息数据采集、存储、运行和管理等要求，包括基础设施设备、IT设备、系统与数据设备和管理工具等。电子信息系统机房系统应符合 GB 501748 的规定。

8.2　不间断电源系统

不间断电源系统应符合 YD/T 1095 的规定。

8.3　供配电系统和照明系统

机房供配电系统和照明系统应符合 GB 50052、GB50054 和 GB 50254 的规定。

8.4　防震和防雷保护系统

机房应配防震和防雷保护系统。防雷保护系统应符合 GB 50057 的规定。

8.5　空调系统、新风和排风系统

机房应安装单独空调系统。空调系统可靠性等级应符合 GB50174 的 A 级标准。空调系统、新风和排风系统还应符合 GB50243 的规定。

8.6　给排水系统

机房应设给排水系统。给排水系统管道铺设应符合 GB 50174 的规定。

8.7　智能化系统

机房应设基础设施环境监控、安保监控、能源监测和综合显示控制系统等智能化系统，智能化系统应符合 GB/T 50314 的规定。

8.8　消防系统

机房应设由消防中心统一控制的自动火灾报警器，应采用气体灭火，不可使用液态或固态灭火。消防系统应符合 GB 50116、GB 50222 和 GB 50370 的规定。

8.9　综合布线系统

机房综合布线系统应符合 GB 50311 的规定。