·医学数据分析与信息处理方法·

计算分子生物学与基因组信息学

郭政　李霞　李晶　编著

黑龙江科学技术出版社

前言

近年来生物医学科学发展的重要特点是开始同时注重生命特性的质和量两方面及其内在逻辑联系的的规律性研究，作为科学逻辑分析与表达语言的数学方法及作为信息处理、分析技术的强大计算机工具在生物医学科学中得到了日益广泛、深入的应用。联合国科教文组织预言，生命科学与数学的结合将可能在21世纪产生最深刻的科学发现。

分子生物学是现代生物医学科学中最深刻和最重要的基础之一，发展极为迅猛。计算分子生物学(Computational Molecular Biology)是近年来发展起来的一门由分子生物学和数学、计算机信息处理技术相结合的交叉学科，它涉及分子生物学的定量建模、理论计算等各方面，也包括生物信息学(bioinformatics)的内容，即发展相关的软件工具和数据库用于有效地进行信息的获取、存储和分析等。特别地，随着当前国际生命科学领域内最重要的基因组计划研究项目的发展，基因组信息学(genome informatics)已成为计算分子生物学和生物信息学中的最具活力的领域。本书主要介绍计算分子生物学中的基因组信息学方法和技术。

1953年Watson和Crick发现了DNA分子双螺旋结构，从而奠定了人类了解自身遗传结构的基础。人类基因组计划的目标是分析人类DNA的结构，确定5～10万个人类基因的位置。整个工作需要对人类基因组作图，确定人类基因组中约30亿碱基对的顺序，包括分析基因的核苷酸顺序结构、识别基因的编码区、测定基因的位置及功能等。与此同时，为了提供揭示人类基因组功能的比较信息，还要研究一组模式生物的DNA。由于通过人类基因组计划将得到存在于DNA中的调控人体发育、生长和各种表型表达的整套指令，所以其完成将使人类对自身有一个根本的了解，并为人类疾病的防治开辟崭新的途径。

基因组信息学涉及基因组信息的获取、存储、分析和解释等各方面，包括发展和有效地利用基因组相关的软件工具和数据库以便处理物理图、遗传图、表达图和序列等信息，并将这些数据资料进行比较分析以解释基因组的信息，例如预测功能基因、调控区及对序列进行同源性分析以揭示生物大分子的分子结构、功能和进化关系等。基因组计划产生的大量作图与测序数据等信息为基于理论推理的新的生物学研究模式提供了基础。人们可以通过对大量信息的规律性分析，从理论推测出发，由实验追踪或验证理论假设。

近年来，随着Internet网络的发展普及，越来越多的生物信息学数据库和分析软件与Internet联结，充分利用网络资源对开展生物信息学研究以及进行人类基因组计划的DNA序列分析非常重要。本书第一章以WWW资源为主简要介绍一些与Internet网络信息和资源利用有关的基本概念和方法，并在附录中介绍一些分子遗传学概念。

为了发现新测定的DNA序列上的功能区域，常用的方法是将该序列与同种或异种生物中已知功能的序列进行相似性比较，获得同源和功能类型相似的线索。可在序列同源性分析的基础上建立蛋白质之间的进化关系等。蛋白质序列的同源性分析也是蛋白质功能、结构预测和分子设计的一个基础。各种功能信号识别，包括发现新基因及基因编码区等是基因组信息学的重要内容。本书第二章介绍序列比较与功能信号检测分析的一些方法。

人类基因组计划的最基本的结果是得到一个由4个元素A、C、G、T串接组成的长度为3×10^9的一维链。在这个链上不仅包含有制造人类全部蛋白质的信息，还要有按照特定的时空模式把这些蛋白质装配成为生物体的调控信息。如何找到这些信息的编码方式、调节规律，将是人类基因组研究的重要任务。为了深入阐明DNA上信息的运作规律，近年来已逐渐采用了各种统计与信息学分析、复杂性与分维等非线性工具用于DNA序列信息分

析。语言学和密码学方法也介入了 DNA 序列分析，用于研究 DNA 序列的语言特征。随着许多全基因组及其合适注释信息的积累，已开始发展了一些统计和仿真方法用于基因组比较，分析功能类描述的基因组类似性、基因与其产物间的互作网络等。第三章介绍一些基因组序列的统计与信息学分析方法。

人类基因组分析的前期基本任务是“读出”人基因组全部核苷酸的顺序。通过构造遗传图和物理图等可以确定 DNA 标记、基因在 DNA 序列上的位置。测序就是得到每一个 DNA 分子的碱基串接次序。第四章介绍一些与人类基因组制图与测序有关的方法。

随着基因组研究的深入，怎样把基因与多种多样的表型联系起来是一个非常重要的问题。用连锁分析的方法粗略定位基因后，可以再采用定位克隆、候选基因等策略筛选出致病基因。通过进一步研究基因与表型的关系，以及该基因编码蛋白质的结构和功能，就可以逐步揭示人类基因组的奥秘。随着 STR 等遗传标记的发现，已有大量的遗传标记可用于遗传性状的连锁分析。对于简单 Mendel 遗传的单基因性状，采用第四章介绍的基因组遗传图的构建方法，通过经典的 lods 分析就可以完成连锁定位分析。更大的挑战在于寻找诸如糖尿病、哮喘、高血压、精神分裂症等复杂性状的相关基因。近年来发展了许多复杂性状的定位克隆分析方法，另外应用均匀分布于基因组的大量标记进行扫查定位已成为复杂性状相关基因定位的一种重要手段。扫查定位所面临的算法及与实验成本密切相关的最优化策略也已成为目前积极研究的课题。第五章介绍复杂性状连锁定位分析与定位克隆策略的有关方法。

可以预见，当基因组大规模测序工作基本完成以后，积累的大量数据将使基因组信息学尤其是比较基因组学研究在人类基因组研究中的重要性更为突出。第六章介绍一些主要与人类基因组序列分析有关的分子生物学数据库和分析软件。 Internet 联接上的许多多功能、复合、开放的新型数据库系统还整合了数据库与数据分析软件，为开展基因组分析(包括所谓的基因组后分析)提供了新的途径。

人类基因组计划是当前国际生命科学领域内一项最引人注目的研究项目，作为其重要组成部分的基因组信息学的发展极为迅速，知识和方法的更新很快，尤其是各种数据库中的信息内容和含量更新更快。在这样的前提下，本书所介绍的内容只能说是引论性的。

计算分子生物学(包括基因组信息学)涉及分子生物学和数学、计算机信息处理技术等很多领域，限于作者目前的水平，加之时间仓促，书中必有许多不足和谬误，希望能得到读者的指正。本书对分形和混沌、孤立子、自组织理论与耗散结构理论等非线性方法的应用尚未介绍，对书中介绍的各种较成熟的分析方法及分析工具也较少给出详细的使用实例，而只是介绍其基本原理与概况。我们希望能与感兴趣的读者交流，有机会完善本书或再提供一本实用的工作手册，后者应该对生物医学工作者更有实用价值。

多年来，我们有幸得到了我校生物遗传教研室许多遗传学家和我校其他专家、领导的帮助与鼓励。我们的工作也得到了黑龙江省自然科学基金委的大力支持。谨在此表示深切的谢意！

作 者

1998 年 6 月

目录

第一章 Internet 网络信息资源和利用 …… 1
一、 Internet 网络信息和资源 …… 1
二、 WWW 浏览器和搜索引擎 …… 3
三、 MEDLINE 数据库检索服务 …… 6
四、网络出版物 …… 7
五、 OCLC 联机计算机图书馆中心 …… 8
六、电子论坛 …… 9
七、分子生物信息学数据库和分析软件 …… 13
7.1 序列库检索 Web 服务器 …… 14
7.2 序列类似性分析等 Web 服务器 …… 15
7.3 蛋白质三维结构分析 Web 服务器 …… 18
7.4 序列检索 E-mail 服务器 …… 20
7.5 序列类似性比较 E-mail 服务器 …… 21
7.6 文件传输协议 FTP 服务器 …… 22
7.7 Gopher 服务器 …… 23
附录：一些分子遗传学的基本概念 …… 24

第二章 序列比较与功能信号检测 …… 38
一、 DNA 与蛋白质序列的相似性分析 …… 38
1.1 相似性得分与空位(Gap)罚分 …… 38
1.2 基本动态规划算法 …… 40
1.2.1 全局相似性 …… 41
1.2.2 局部相似性 …… 42
1.2.3 更一般的 Gap 罚分 …… 43
1.3 *k*-best 配准 …… 43
1.4 多序列配准 …… 44
1.5 近似的模式匹配 …… 45
1.6 基因重组和序列内部重复问题 …… 46
1.7 点矩阵作图与窗口过滤技术 …… 46
1.7.1 简单点矩阵作图 …… 46
1.7.2 窗口过滤技术 …… 47
1.8 基因组长序列比较 …… 48
1.9 相似性与差异性测度的对偶性 …… 48
1.10 相似性和同源 …… 49
二、氨基酸指标与相似矩阵（突变矩阵） …… 49
三、一个序列与数据库的比较 …… 52

3.1 数据库搜索的算法工具 …… 52
3.2 序列类似性的统计显著性 …… 54
3..2.1 Monte Carlo 仿真法 …… 54
3.2.2 BLAST 得分显著性的 Karlin-Altschul 公式 …… 55
3.2.3 局部配准的统计显著性 …… 56
3.2.4 短序列配准的显著性评价 …… 57
3.2.5 核酸序列比较的显著性评价 …… 57
3.3 算法的敏感性与准确度（选择性） …… 58
3.4 有空隔配准的BLAST程序与位置特异的迭代BLAST程序 …… 59
3.4.1 有空隔配准的BLAST程序 …… 60
3.4.2 BLAST 对位置特异的分值矩阵的迭代应用 …… 61
四、蛋白质序列对 DNA 数据库翻译的所有可能序列进行比较 …… 63
五、蛋白质家族与蛋白质分类 …… 67
5.1 蛋白质家族与超家族 …… 67
5.2 蛋白质分类的方法 …… 68
5.2.1 Blocks 分类方法 …… 68
5.2.2 加权特征标纹分类方法 …… 70
5.2.3 标纹(motif)数据库的基准模式 …… 70
5.2.4 Profile 方法 …… 71
六、分子进化钟与进化树 …… 71
6.1 分子进化钟 …… 72
6.2 进化树 …… 73
6.2.1 序列进化树 …… 73
6.2.2 转换/颠换比(TI/TV)估计 …… 77
6.3 构建进化树的图论原理 …… 78
6.4 结构进化树 …… 80
6.5 *mt*DNA 序列分析与歧异度 …… 82
6.6 从 DNA 多态性模式重建祖先序列 …… 85
七、蛋白质序列模式和序列结构域模式 …… 86
7.1 基准序列(序列模式)：标纹、标志、指纹和位点 …… 86
7.2 序列结构域与模式匹配方法 …… 86
7.2.1 频率表方法 …… 87
7.2.2 权值矩阵法：Profile 分析 …… 88
八、核酸的信号检索与结构预测 …… 88
8.1 限制性酶切位点和固定序列模式检索 …… 88
8.2 短寡聚核苷酸序列的随机出现机率 …… 90
8.3 编码区 DNA 寡聚体出现频率 …… 92
8.4 核酸序列的特殊信号检索 …… 95
8.4.1 基准序列频率表和权值矩阵法 …… 96
8.4.2 分析候选结合位点的 ConsInspector 程序 …… 97

8.4.3 搜索基因组中特殊模式的 PatScan 程序 …… 98
九、基因识别与翻译 …… 100
9.1 开放阅读框架分析 …… 100
9.2 编码区识别 …… 100
9.2.1 碱基组成偏歧法 …… 101
9.2.2 密码子使用法 …… 101
9.2.3 密码子偏歧法 …… 102
9.3 基因识别 …… 102
9.3.1 GenLang 基因识别 …… 103
9.3.2 GRAIL 基因识别 …… 105
9.4 基因识别的一些相关程序 …… 106
9.4.1 发现和屏蔽重复 …… 107
9.4.2 序列相似性与标纹数据库搜索 …… 107
9.4.3 整合的基因识别 …… 109
9.4.4 序列片段的编码区分析 …… 113
9.4.5 其它功能信号识别 …… 114
9.5 编码序列翻译 …… 115
十、 RNA 标纹识别和局部结构配准 …… 116
10.1 信号搜索：概率方法 …… 116
10.2 信号搜索：模式匹配方法 …… 117
10.3 *t*RNA 的二级结构预测 …… 118
10.4 RNA 序列的局部结构配准 …… 119
十一、 DNA 序列分析软件包 X-HUSAR …… 120
十二、蛋白质结构预测与分子设计 …… 121
12.1 蛋白质结构预测 …… 122
12.2 蛋白质二级结构预测 …… 124
12.3 合理药物分子设计 …… 124

第三章 基因组序列的统计与信息学分析 …… 129
一、核苷酸序列的相关与信息学分析 …… 129
1.1 核苷酸序列相关性的 Markov 熵分析 …… 129
1.2 核苷酸序列的长程相关与非线性方法 …… 130
1.3 长程互作对 DNA 的结构和可变性的作用 …… 131
1.4 重复对熵的影响 …… 132
1.5 编码片段的相互信息 …… 133
1.6 DNA 序列的模式结构 …… 134
1.7 语言学复杂性测度 …… 135
1.8 非编码区（“ Junk ” DNA ）基因组序列 …… 136
二、密码子指纹与密码子使用偏歧 …… 138
2.1 单、双核苷酸的相对丰度和基因组指纹 …… 139

2.2 密码子频率和密码子指纹 …… 139
2.3 基因间和基因类间的异质性 …… 140
三、编码 DNA 片段的长度与 GC 含量 …… 145
四、重叠基因的信息论问题 …… 146
五、*mt*DNA 的核苷酸组成偏歧(组成模式) …… 147
六、定向突变压力对 *mt*DNA 进化的影响 …… 148
七、功能相关基因在两个基因组间或内部的聚类关系 …… 150
7.1 基因组比较与基于功能组成的物种间的比较 …… 150
7.2 两个细菌基因组间或内部的聚类关系 …… 150
八、真核生物的基因表达调控(表达促进网络) …… 152
8.1 相对同义密码子使用值与密码子适应指数 …… 152
8.2 信息聚类方法与自身一致信息聚类 …… 153
8.3 碱基组成及相关性与基因表达的关系 …… 154
九、基因与其产物间的完全互作网络 …… 156
9.1 转录因子结合位点的出现概率 …… 157
9.2 识别转录因子结合位点聚类的显著性 …… 158

第四章 人类基因组制图与测序 …… 161
一、人类基因组制图 …… 161
1.1 遗传图 …… 161
1.2 物理图 …… 162
1.3 序列图 …… 162
1.4 转录图（表达图）与 *c*DNA 文库构建 …… 163
二、基因组遗传图的构建方法 …… 164
2.1 检测连锁与估计重组率 …… 164
2.2 估计相对图距和推测多位点顺序 …… 166
2.2.1 图距与交叉干涉 …… 166
2.2.2 推测多位点顺序 …… 167
三、基因组物理图谱与测序 …… 169
3.1 克隆与克隆库 …… 170
3.2 随机克隆重叠构图 …… 171
3.3 指纹方案的评价 …… 172
3.4 锚定法作图 …… 173
3.5 检测重叠的 Bayes 方法 …… 173
3.5.1 重叠构型 …… 174
3.5.2 重叠检测 …… 175
3.6 由随机克隆的指纹法组装物理图 …… 176
3.7 用 YAC 克隆构造人类基因组图谱的策略设计 …… 176
3.8 采用高冗余度的亚克隆库 …… 177
3.9 Contig 图或克隆定序 …… 178

3.10 直接作图法…… 179
3.11 有序鸟枪测序作图的仿真分析…… 179
3.12 大规模 DNA 测序 …… 180
3.12.1 定向测序法 …… 180
3.12.2 随机测序法 …… 181
3.12.3 复合测序法 …… 181
3.12.4 自动化 DNA 测序…… 182
3.13 寡聚核苷酸引物和探针设计程序…… 182
3.14 定位克隆的流水线鸟枪策略…… 184
3.15 放射杂交作图和 FISH 作图…… 186

第五章 复杂性状连锁分析与定位克隆策略 …… 191
一、复杂性状（疾病）…… 191
1.1 复杂性状的特点 …… 192
1.2 改善复杂性状遗传作图的一般途径 …… 192
1.3 微卫星DNA标记与基因型测定…… 194
二、lods连锁分析…… 195
2.1 遗传异质性 …… 196
2.1.1 遗传异质性检验…… 197
2.1.2 多位点遗传异质性对连锁分析的影响…… 197
2.1.3 复杂疾病多位点系统分析…… 198
2.1.4 家系内异质性…… 199
2.2 外显率未知时的连锁分析 …… 199
2.3 检验效能的仿真分析方法与实验设计 …… 200
2.3.1 检验效能的仿真分析方法…… 200
2.3.2 实验设计…… 201
2.3.3 标记位点等位基因频率的影响…… 202
2.3.4 诊断阈值的影响…… 202
2.4 多模型分析与多重比较的检验效能与显著性水平 …… 202
2.5 有关精神类疾病的几个问题 …… 203
2.6 单子女家系与连锁不平衡 …… 204
2.7 生物协变量 …… 204
2.8 连锁分析程序 …… 205
2.9 应用实例 …… 207
三、等位基因共享连锁分析…… 209
3.1 血缘一致IBD法…… 210
3.1.1 受累同胞对IBD连锁分析与检验效能 …… 210
3.1.2 单受累同胞对IBD连锁分析与检验效能 …… 211
3.1.3 扩展的亲属对IBD连锁分析 …… 212
3.1.4 在近亲婚配群体抽样…… 213
3.2 状态一致连锁分析法 …… 214

3.2.1 受累同胞对状态一致IBS法 …… 214
3.2.2 受累亲属对状态一致IBS法 …… 214
3.2.3 使用未发病家系成员的标记信息的IBS法 …… 216
3.2.4 X连锁位点上的受累亲属对状态一致IBS法 …… 216
3.3 一组亲属共享IBD等位基因的连锁分析 …… 217
3.4 分析两个连锁的易感位点的IBD方法 …… 218
3.5 用ASP法分析与同一位点关联的两种疾病的关系 …… 219
3.6 受累同胞（亲属）对法与lods连锁分析 …… 219
3.7 多标记位点连锁分析的一般似然函数框架 …… 221
3.8 结合参数与非参数分析的GENEHUNTER …… 222
3.9 区间作图法 …… 225
3.10 使用体细胞组成杂合性丢失（LOH）数据 …… 225
3.11 ASP法的检验效能与基因组扫查 …… 226
3.12 应于 GMS 技术进行基因组描查 …… 228
3.13 多位点定位策略的成本与效益 …… 230
四、疾病关联与连锁不平衡分析 …… 230
4.1 疾病与遗传标记关联的传统方法 …… 231
4.1.1 列联表检验与 Woolf 相对风险 …… 231
4.1.2 非随机婚配群体中的关联分析 …… 232
4.1.3 疾病与高度多态性位点上的等位基因的关联性分析 …… 233
4.1.4 单体型关联分析 …… 234
4.2 基于家系资料的关联分析 …… 234
4.2.1 单体型相对风险关联分析 …… 234
4.2.2 传递/不平衡 TDT 检验 …… 238
4.2.3 TDT 关联与连锁分析及其扩展 …… 239
4.3 关联与连锁分析的关系 …… 242
4.4 核心家系构形描述和分析 …… 246
4.5 基因组扫查由远缘相关亲属共享的 IBD 区域 …… 248
4.6 在混合群体定位疾病位点 …… 250
4.7 单体型构建分析 …… 252
4.8 连锁不平衡作图 …… 253
4.8.1 单体型频率与不平衡系数计算 …… 255
4.8.2 连锁不平衡测度与简单不平衡作图 …… 256
4.8.3 连锁不平衡测度与物理距离的关系 …… 259
4.8.4 基于群体模型的连锁不平衡精细作图 …… 259
4.8.5 用多个多态标记根据似然函数方法进行基因定位 …… 265
4.8.6 连锁不平衡基因组等阶扫查分析 …… 266
五、基因组扫查的统计显著性 …… 267
六、数量性状的基因定位 …… 268
6.1 同胞对等位基因共享方法 …… 269
6.1.1 一致与不一致同胞对的抽样策略与检测效能 …… 269

6.1.2 筛查抽样策略与优化实验设计…… 274
6.1.3 基因组扫查定位及位点与性状类型关系的筛查…… 276
6.2 lods 连锁分析的效能…… 277
6.3 实验杂交与多基因性状包括 QTLs 作图…… 278
七、定位克隆与高精度基因定位…… 279
7.1 定位克隆与定位候选策略 …… 279
7.2 高精度连锁定位实验分析 …… 282
八、与基因作图和连锁分析有关的Internet网络资源…… 284

第六章 分子生物学数据库和分析软件 …… 295
一、核苷酸序列与基因组数据库…… 296
1.1 GenBank 数据库…… 296
1.2 EMBL 核苷酸序列库与 EBI 网络服务…… 299
1.3 DDBJ 数据库…… 300
1.4 密码子使用与核苷酸信号数据库… …… 301
1.5 基因组序列数据库 GSDB …… 304
1.6 人类基因组数据库 GDB …… 305
1.7 几个模型生物基因组数据库 MGD 、 ECDC 、 NRSub …… 307
1.8 基因组的图形交互显示和检索、浏览工具资源…… 308
1.9 基因表达索引数据库 Genexpress Index …… 310
二、蛋白质序列与模式、同源性数据库…… 313
2.1 蛋白质序列数据库 PIR-International …… 313
2.2 蛋白质序列数据库 SWISS-PROT …… 314
2.3 Prosite 、 Blocks 、 PRINTS 和 SBASE 数据库…… 316
2.4 MIPS 蛋白质序列、同源数据和 yeast 基因组信息数据库…… 322
2.5 数据库中存在的问题 …… 323
三、基因组数据库搜索与相关基因比较识别 …… 325
3.1 基因组特异的BLAST和FASTA序列搜索…… 325
3.2 PEDANT基因组浏览器…… 326
3.3 识别人类 cDNA 与 Drosophila 基因产物的同源性的 DRES 搜索引擎 …… 327
3.4 基因组比较交互参考数据库 XREFdb …… 328
四、基因和分子的互作和调节通路信息数据库…… 329
4.1 基因和基因组百科全书数据库 KEGG …… 329
4.2 E.coli K-12 基因组和代谢通路数据库…… 330
4.3 E.coli 基因及其产物的数据库 GenProtEC …… 333
4.4 果蝇的遗传和分子数据的数据库 FlyBase …… 334
五、RNA核苷酸序列数据库…… 335
六、线粒体 DNA 数据库 MITOMAP 与 MmtDB …… 336
七、免疫球蛋白、 T 细胞受体、 MHC 的整合数据库 IMGT …… 337
八、突变数据库…… 337

九、放射杂交作图数据库 Rhdb …… 340
十、限制酶数据库 REBASE 与分子探针数据库 MPDB …… 341
十一、蛋白质结构数据库 …… 342
十二、其它遗传学与分子生物学资源 …… 344

第一章 Internet 网络信息资源和利用

Internet 国际互联网络是现代信息社会中巨大的信息传输及处理系统，在 Internet 上可以获取关于各个领域几乎无所不包的丰富信息。可以认为信息资源是 Internet 网络应用的前提和实质内容， Internet 网络为我们提供了获得各种丰富的信息资源的机会。

近年来，随着 Internet 网络的发展普及，越来越多的生物信息学数据库和软件与 Internet 联结，而且绝大多数网上资源可免费检索或下载下来使用。又由于新的网络信息查询工具 WWW 及其浏览软件 Mosaic 和 Netscape 等的出现，网络信息的检索越来越方便、快速。这为我国开展生物信息学研究以及进行人类基因组计划的 DNA 序列分析提供了捷径。特别是在当前我国生物信息学数据库引入不够、软件很不丰富的情况下，充分利用网络资源尤为重要。诺贝尔奖金获得者 Gilbert 指出：“基于全部基因都将知晓并以电子技术可操作的方式驻留在数据库中，新的生物学研究模式的出发点应是理论的。科学家将从理论推测出发，然后再返回到实验中去，追踪或验证这些理论假设。……生物学家不仅必须成为计算机学者，而且也要改变他们研究生命现象的途径。……我们必须把个人计算机联入世界联网，这样才可能了解数据库中每天发生的变化，并实时地彼此进行通讯联系。”

本章以 WWW 资源为主简要介绍一些与 Internet 网络信息和资源利用有关的基本概念和方法，并在附录中介绍以后各章常涉及到的一些分子遗传学概念。

一、 Internet 网络信息和资源

通常根据提供网络服务的方式，将 Internet 信息和资源分为电子邮件资源、电子论坛和网络消息组资源、 FTP 资源、 Gopher 资源和 WWW 资源等。特别是 WWW(亦称 Web)作为环球信息资源网，其信息和资源的含量极大，而且增加非常迅速。

已开发了许多搜索 Internet 信息的软件，人们特别称之为搜索引擎(search engines)。例如，在 Netscape “ Net Search ”页面下列出了许多搜索引擎。这些搜索引擎软件都将网络信息按照其性质分成基本相似的 10 余类，以方便和加快对网络资源的检索。例如， Infoseek 公司的 Infoseek 软件将网络信息分成 12 大类，而各大类又进一步分成若干个较小的相关专题(topics)。例如，在 Science 大类下分成农业、数学和自然科学等 9 个专题，后者又细分为生物科学、化学、地球科学和物理学 4 个分专题，而生物科学又再分为生物化学、生物信息学、生物科学杂志、分子生物学和动物学等 27 个支专题。在 Health 大类下分成了公共卫生、性卫生和医学等 20 个相关专题，其中医学又进一步分成解剖学、牙医学、内科学、药学和医学院校等 13 个子专题，每个子专题又有多个网点供浏览。

Infoseek 等搜索引擎除了可以帮助用户按它们对 Web 资源的分类一层一层地浏览网上信息外，还提供输入任意检索词检索网点及文献的功能。常用的利用网络信息并参与发布消息的方式有以下几种：

1.WWW 浏览器和搜索引擎： WWW(World Wide Web)的含义是“环球网”，也称“万维网”、 3W 、 Web 。 WWW 是一个基于超文本(Hypertext)方式的信息检索服务工具。通过将位于全世界 Internet 网上不同地点的相关信息有机地编织在一起， WWW 提供一种友好的信息查询接口，用户仅需提出查询要求，而到什么地方查询及如何查询则由 WWW 自动

完成。 WWW 全球网络信息检索系统通过超级文本链接，能够存取全世界几千个使用不同网络协议(如 WWW 、 FTP 、 WAIS 、 Gopher)的服务器所储存的文献和信息，包括文字、图像、电影和声音。

WWW 服务器主要有两类客户程序，其一是 Lynx，它是全屏幕浏览，使用箭头键、关键字符增亮显示；其二是 NCSA Mosaic 、 Netscape 和 Microsoft Explorer，它们是带有图形用户接口的灵活和强有力的浏览软件。 WWW 与传统的 Internet 信息查询工具 Gopher 、 WAIS 最大的区别是，它展示给用户的是一篇篇文章，而不是那种时常令人费解的菜单说明。因此，用它查询信息具有很强的直观性。另外， WWW 还可提供其它传统的 Internet 服务，包括 Telnet 、 FTP 、 Gopher 和 Usenet News(Internet 的电子公告板服务)。通过使用 WWW，一个不熟悉网络使用的人也可以很快成为 Internet 的行家。因此，自 Web 服务器于 1993 年问世以来，很快便上升为 Internet 信息流的主要载体，而且所占网络信息的比重还将越来越大。利用它们可以很方便地检索环球网上的信息。

WWW 的成功在于它制定了一套标准的、易为人们掌握的超文本开发语言 HTML 、信息资源的统一定位格式 URL 和超文体传送通信协议 HTTP 。

(1)HTML(Hyper Text Mark-up Language)即超文体标记语言，是 WWW 的描述语言。设计 HTML 语言的目的是为了能把存放在一台电脑中的文本或图形与另一台电脑中的文本或图形方便地联系在一起，形成有机的整体，人们不用考虑具体信息是在当前电脑上还是在网络的其它电脑上。这样，用户只要使用鼠标在某一文档中点取图标，Internet 就会立即转到与此图标相关的内容上去，而这些信息可能存放在网络的另一台电脑中。 HTML 文本是由 HTML 命令组成的描述性文本， HTML 命令可说明文字、图形、动画、声音、表格、链接等。 HTML 的结构包括头部(Head)、主体(Body)两大部分。头部描述浏览器所需的信息，主体包含所要说明的具体内容。

(2)要在 Web 上检索特定的服务器资源，需要知道相应的 Web 服务器的网络地址，一般称为 URL(Uniform Resoure Locator)或“统一资源定位符”。统一资源定位符 URL 是 WWW 页的地址，它从左到右由下述部分组成：(i)Internet 资源类型(Scheme)：指出 WWW 客户程序用来操作的工具。如“ http://”表示 WWW 服务器，“ ftp://”表示 FTP 服务器，“ gopher://”表示 Gopher 服务器，而“ new:”表示 Newsgroup 新闻组。(ii)服务器地址(Host)：指出 WWW 页所在的服务器域名。(iii)端口(Port)：有时对某资源的访问来说，需给出相应的服务器提供端口号。(iv)路径(Path)：指明服务器上某资源的位置(其格式与 DOS 系统中的格式一样，通常由目录/子目录/文件名这样的结构组成)。与端口一样，路径并非总是需要的。

WWW 上的服务器都是区分大小写字母的，所以必须注意正确的 URL 大小写表达形式。 URL 地址格式排列为： scheme://host:port/path 。例如 http://www.cnd.org/pub/HXWZ 就是一个典型的 URL 地址。客户程序首先看到 http(超文本传送协议)，便知道处理的是 HTML 链接。接下来的 www.cnd.org 是站点地址，最后是目录 pub/HXWZ 。又如对于 ftp://ftp.cnd.org/pub/HXWZ/cm9612a.GB， WWW 客户程序需要用 FTP 去进行文件传送，站点是 ftp.end.org，然后去目录 pub/HXWZ 下，下载文件 cm9612a.GB 。如果上面的 URL 是 ftp://ftp.cnd.org:8001/pub/HXWZ/cm9612a.GB，则 FTP 客户程序将从站点 ftp.cnd.org 的 8001 端口连入。

2.网络通讯组或电子公告牌(又称电子论坛)：例如，生物学家电子论坛有 100 多个生物医学专题讨论组(或消息组)，用户通过发送电子邮件到 BIOSCI/bionet 服务器(biosci-

server@net.bio.net)报名参加这些消息组。这样用户的电子邮件信箱中将不时地收到所参加的消息组的信息，同时用户可以参加讨论或提出自己的问题以求得别人的帮助或解答。

3.电子邮件：利用 Internet 网络上的 E-Mail 功能，可以接收和发送电子邮件。这些邮件可以是一封信，也可以是数据和软件程序等。例如，早期的核酸序列数据库 GenBank 和 EMBL 主要通过向电子邮件文件服务器发送一定格式的电子邮件，用户可检索到序列记录或将用户靶序列针对序列数据库进行序列类似性比较，从而检出同源序列。

4.Telnet：通过 Telnet 登录到远距离的主计算机上，用户的计算机可以暂时成为主机的一个终端，从而利用主机上的资源，但是用户必须有主机认可的用户名和口令。许多大型数据库(如 Medlars 文献库)和软件系统(如用于生物大分子序列分析的软件包 GCG)可以经由此种方式访问。远程登录检索是指在网络通讯协议的支持下，用户的计算机成为远程计算机终端而进行检索的过程。

5.文件传输协议(FTP)：它是一种实时的联机服务，使用时，用户先要登录到远程的 FTP 服务器主机上。许多大型公用数据库提供“匿名文件传输(anonymous FTP)”服务，即用户在登录时可用 anonymous 作为用户名，用自己的 e-mail 地址作口令。在登录后，用户可进行与文件搜索和文件有关的操作，如显示文件目录、改变当前工作目录、设置传输参数和传送文件等。“匿名” FTP 方式不要求用户事先从服务器主机获得特定的用户名和保密字，因此它对获得公用数据库和软件很有用。例如，通过美国全国生物技术信息中心(NCBI)的 FTP 服务器可以免费得到最新的核酸序列数据库 GenBank 和几十个其它分子生物学数据库，以及如序列对序列数据库类似性检索程序 BLAST 和多重序列比较程序 MACAW 等多个序列分析软件，通过美国华盛顿大学的 FTP 服务器可以免费获得分子进化树程序 PHYLIP 。

除上述 5 种方式外， Gopher 和 WAIS 等网络工具都可以用于相应网络信息的检索。一种资源可以用多种方式来访问，例如几乎目前所有的网络工具都可以访问 DNA 序列数据库 GenBank 。其中 WWW 具有强大的检索功能，既可以访问 Web 服务器资源，也可以发送电子邮件、参加网络通讯组和访问 FTP 、 Gopher 、 WAIS 等服务器资源，已成为 Internet 资源的主要载体和服务方式。

可以使用图形界面的 WWW(或称 Web)浏览器 Netscape Navigator 、 Internet Explorer 、 Mosaic 、 HotJava 和 Lynx 等资源浏览工具以方便、生动地获得和浏览 Web 上信息资源。利用 Internet 网络上的 Telnet 和 FTP 功能，还能直接使用远程电脑主机的软件系统，以及丰富的信息资源。对于科学研究的课题、论文、图书馆的藏书和各种科学杂志等图像文字资料，都可以使用 Yahoo 、 WAIS 、 Archie 、 Veronica 、 Jughead 等工具，依据关键字查询和检索到它们。

二、 WWW 浏览器和搜索引擎

WWW 浏览器(Browsers)是一种 WWW 客户程序。 WWW 浏览器主要有两类版本：第一类是以 Lynx 为代表的面向字符操作的 WWW 客户程序，主要供不具备图像和声音功能的电脑终端或采用仿真终端方式工作的电脑用户使用。第二类是 Netscape Navigator 、 Internet Explorer 、 Mosaic 、 HotJava 等面向多媒体电脑工作站的 WWW 客户程序，它们可以在 PC 机 Microsoft Windows 、 Apple Macintosh 机以及 Unix 操作系统 X-Window 软件平台上运行。在多窗口的界面上，用它们不但可以浏览文本信息，还可以显示与文本内容相配合的图像、

影视和声音。浏览器自身的功能越来越强大，从某种意义上来说，浏览器扮演着现今操作系统的角色。

WWW 浏览器的最基本目的在于让用户在自己的电脑上检索、查询、采掘、获取 Internet WWW 上的各种资源。随着 Internet 的飞速发展，浏览器的功能在不断的扩充和更新。归纳起来浏览器具备以下几种基本功能：①检索查询功能：浏览器读入 HTML 文档，解释 HTML 所描述的图表、声音、动画、表格，以及进一步的链接信息，利用超文本传输协议(HTTP)，可在任意 WWW 服务器上畅游。②文件服务功能：能在下载文档时实时查阅该文档，并可利用 HTTP 去跟踪感兴趣的链接。可以将正在查阅的文档随时保存、打印、浏览等等。③热表管理：浏览器应能够自动记住用户刚刚访问过的 WWW 地址，称为“热表”。当用户想要回到刚才曾访问过的某一 WWW 中，用户可以从热表中快速地切换。④建立自己的首页(Home Page)：当用浏览器启动 Internet 上某一 URL 地址上的某一文档文件时，由 Internet 器首先显示的那个文档，叫做首页。在首页中，可以加入表征用户特点的图形或图像，列出最常用的一些链接。浏览器提供了很好的接口，可以利用 HTML 和 HTTP 在 WWW 服务器上方便地制作出自己的首页。⑤提供其它 Internet 服务：浏览器除了完成自己基本的查询浏览信息功能外，还提供其它 Internet 服务，如 FTP 、 Gopher 、 WAIS 、 Telnet 、 Usenet 上的 NNTP(网络新闻传输协议)及 E-mail 等。

目前已有几十种浏览器，它们大多为免费或共享软件，可以在 Internet 上方便地获取。代表性的浏览器有 Netscape Navigator 、 Microsoft Internet Explorer 、 Mosaic 、 HotJava 等。Mosaic 是最早的 WWW 浏览器，能解释 WWW 中的 HTML 文档，并能把 HTML 文档包容的信息以统一的方式显示出来，它最早运行在 Sun 工作站 X-Window 图形环境上，后来又推出了 Mosaic 的 Macintosh 及 Windows 版本。 Netscape Navigator 提供了一个与 Mosaic 相似但比它更实用的图形界面。它是第一个优先快速显示文本和图形的浏览器，也是第一个在收到整页文档前就允许查看页首的浏览器。 Navigator 支持新闻组(News Group)，同时还可在同一窗口上支持 HTTP 、 FTP 和 Gopher 。更重要的是，它支持对 HTML 的增强功能，改进了整个 WWW 上的网点设计。下面简要介绍 Netscape Navigator(简称 Netscape)的主要功能及其操作方法。

在 Windows95 下连接 Internet 后，在 Netscape 的图标下按动鼠标运行 Netscape ，屏幕上就会出现 Netscape 主画面。 Netscape 页是用超文本语言(HTML)编写的，集文字和选单(又称为链)为一体。 Netscape 页中选单项(链)是高亮度显示的带下划线的文字，带有有色框的图像或图标。 URL 内嵌在链中，所以链将一页和另一页相连。 Netscape 页自上向下由以下几个部分组：标题条、下拉工具条按钮、地址域、目录按钮、页正文、状态域以及传输进程条。标题条显示当前页的标题；下拉选单提供诸如文件保存、打印、属性设置等操作；工具条按钮一般可用来改变当前 Netscape 页的属性；地址域显示前页的 URL 地址；目录按钮则提供用于浏览 Internet 的工具。用户可在 Option 选单中分别将工具条、目录按钮、地址域设置为不显示，从而扩大正文显示面积。状态域显示当前操作的反馈信息。传输进程条显示页传输的进度。

(1)使用 Netscape 浏览 WWW 资源：当用户启动 Netscape 时，屏幕显示某一主页。主页可由用户设置，默认为 Netscape 主页。用户可浏览当前页，或撳击某个链将所连页显示在屏幕上浏览。通过撳击，用户在浏览过程中可轻松地从一台机器访问到另一台机器，而无需知道机器、文件所处位置，从而实现在 Internet 中的漫游。配备多媒体电脑的用户不仅可看到

文字、图像或动画，而且还可以听到声音。

用户浏览的过程以页标题表的形式保存在历史表(History)中。撳击其中一项，则可将对应页显示在屏幕上。Netscape 还提供了向后(Back)和向前(Forward)功能：撳击 Back 按钮，则可将当前页的前页显示在屏幕上；撳击 Forward 按钮，则可将当前页的后页显示在屏幕上。Netscape 支持用户为经常访问页或当前来不及访问的页设置书签(Bookmarks)，当要多次调阅热点资源时，只需在书签文件中按 Bookmark/Go to Bookmarks 进行选择，省去了每次输入地址的麻烦。页书签的使用提供了永久、快速检索页的手段。Netscape 提供了多种设置书签文件的方式：用户在浏览过程中可执行 Bookmarks/Add Bookmarks 为当前页设定书签，也可将别人的书签输入到自己的书签文件中或将自己的书签输出给别人。通过执行 Windows/Bookmarks，用户可维护书签文件。

Netscape 还提供了用户直接键入 URL 地址以浏览对应页的方法。用户可在 Location 域或 Open 对话框(执行 File/Open Location)内的 Open Location 域内键入 URL，再按回车后，该资源的 Homepage 主页将会出现在屏幕上。利用主页中的超级链接(Hyperlink)，移动鼠标处的图标，可以很方便地调出该资源的内容。用户在传输下页的过程中，如因等待时间过长或其它原因想终止传输，可点击红色 stop 按钮。

(2)使用 Netscape 的搜索机制(搜索引擎)：用户可通过输入文件的 URL 地址找到文件，也可通过漫游 Internet 得到自己感兴趣的文件。前者要求用户必须知道文件的 URL 地址，后者未必能使用户顺利、快速地得到所需要的文件。针对上述情况，Netscape 提供了许多搜索机制(搜索引擎)，允许用户根据主题词、页标题等搜索相应的页、新闻或文档。

按 Net Search 按钮，用户便可在屏幕上看到 Netscape 提供的 Net Search 目录表。其中 Accufind 用来搜索信息库、新闻、书以及 Internet。通过使用 Airsii，用户可输入搜索词对文件内容进行搜索。SHAREWARE.COM 可让用户很容易地找到 Internet 上的软件，以便浏览及下载。Infoseek、Yahoo 可被用来按关键字对文件内容进行搜索，以找到最相关的匹配、相关的话题以及来自于流行杂志的新闻和观点，它常被用来寻找 E-mail 地址、公司文件等。WhoWhere 可让用户快速搜索网上的人或机构，从而得到相应的 E-mail 地址或其主页地址。

(3)使用 Netscape 发送电子邮件/新闻组：用户可用 Netscape 发送邮件、新闻稿。点击当前页右下方的信封图标或 File/New Mail Message，用户可在随即出现的 Netscape Mail 或 Netscape message Composition 对话框内，填入接收方的邮件地址、邮件的主题词以及正文，并将其发送出去。用户还可让邮件附带某一文件或当前页的内容发送。通过执行 Windows/Netscape News，用户在随即出现的 Netscape News 对话框内不仅可发送、收看邮件，而且可参与新闻组讨论。

(4)设置 Netscape 的属性：为了让用户以自己喜爱的方式访问 Internet，Netscape 允许用户设置或更改当前操作的属性。通过执行 Option/Preference，用户可选择字体、字的大小、页的背景色、链被激活前后的颜色，用户还可设置 Netscape 启动主页。用户可执行 Option/Show Toolbar、Option/Show Locaton、Option/Show Directory Button 让 Toolbar、Location 或 Directory Button 显示或不显示在屏幕上。如不显示，则可以扩大当前页的显示区域。Netscape 页由文字、图像或图标组成。传输图像要化较长时间。用户可执行 Option/Auto Load Images 关闭自动图像装入，这样在传输过程中图像均被图标取代，文件传输的速度大大提高，用户可手动恢复图标所在处的图像。

Netscape 还提供许多其它功能：①储存资源功能： Netscape 中的 File/Save 允许将看到的超级文本以及图像储存到本地机上，并可随时调阅。②打印功能： Netscape 中的 File/Print 功能可以将超级文本图文并茂地打印出来。③调用 Internet 其它服务：通过简单的设置，Netscape 允许调用 Internet 的其它服务，如远程登录到别的主机，访问 Gopher 、 FTP 服务器。方法是在 URL 地址处输入 Gopher 、 FTP 服务器地址。

Home Page 首页是一种用超文本标记语言(描述性语言)将信息组织好，再经过相应的解释器或浏览器翻译出的包括文字、图像、声音、动画等多种信息的组织方式。 Home Page 的传输方式是将原代码和与 Home Page 有关的图形文件、声音文件放在一台服务器(称 WWW 服务器)查询。

三、 MEDLINE 数据库检索服务

MEDLINE 数据库是美国国立医学图书馆 MEDLARS 系统中规模最大、权威性最高的著名医学文献数据库。它收录 1966 年至今的世界 70 多个国家出版的生物医学期刊约 3600 种，另外于 1976 ～ 1981 年还收录了图书专著。 MEDLINE 数据库实际上是由《 INDEX MEDICUS 》(《美国医学索引》)、《 INTERNATIONAL NURSING INDEX 》(《国际护理学索引》)及《 INDEX TO DENTAL LITERATURE 》(《牙科文献索引》)三种重要的医学索引组成。内容涉及基础医学、临床医学、生物学、心理学及神经精神疾病、动植物、微生物学及情报科学等多学科领域。

MEDLINE 医学文献数据库提供了多种方式的检索服务，如光盘、远程登录及在 WWW 上的检索等。在 WWW 上检索 MEDLINE 与联机检索和光盘检索相比，用户不必记忆大量的标识符和检索词的组配方法，就能完成医学文献的检索。

可以使用图形界面的 WWW(或称 Web)浏览器 Netscape 或文字界面的 Web 浏览器 lynx 在 Web 上免费检索 MEDLINE 数据库，相应的 MEDLINE Web 服务器的网络地址(URL 或"统一资源定位符")有两个，分别是：

http://www.healthgate.com/HealthGate/MEDLINE/search.shtml

http://www.healthgate.com/HealthGate/MEDLINE/search-advanced.shtml

分别对应着 MEDLINE 的普通检索界面和高级检索界面。用户要检索 MEDLINE 时，就需要将相应的 URL 键入 Web 浏览器的地址栏内。

1.MEDLINE 普通检索界面：使用很容易，只要在检索窗口键入一个关键词或自由词、主题词、作者名、基因符号、时间、词组或一个自由词与布尔逻辑运算符所组成的检索式，也就是在检索窗口键入一个符合自己检索要求的概念，然后按回车键或文献检索按钮，就可以得到以这个概念为中心的 MEDLINE 数据库中的医学文献。其中基因标识符号(GS)是 1991 年新增加的字段。 GS 字段收录了文献中出现的基因名称的"标识符号"或缩略形式，如：Ghox-lab， pulC， pyrB 。建立该字段的目的并不是为了将其建成一个权威的和标准化的基因名称表。该字段最多可含 25 个值，即每条记录最多可收录 25 个基因标识符。为了方便基因文献的检索， 1992 年 NLM 推出了基因标识符号片段检索这一新的功能，它使得该字段的检索具有类似文本词检索的能力。所谓文本词或自由词是作者在文献中使用的未经规范化处理的语言。当没有适当主题时，文本词检索是一条可供选择的主题检索途径。文本词是作者使用的语言，检索时必须考虑到作者可能使用的各种同义词、不同的拼法、形容词和名词形