生物医学信息基础课系列教材

医用生物信息学

理论与实践

李 林 主编

科学出版社

生物医学信息基础课系列教材

医用生物信息学理论与实践

主　编　李　林
副主编　李冬果　华　琳
编　者　高　磊　夏　翊
　　郑卫英　郑文新

科学出版社
北京

内 容 简 介

本书是面向医学研究生和本科生的一本生物信息学入门级读物. 全书编写力求通俗易懂、图文并茂，突出实用特色. 内容包含序列比对、基因芯片数据分析、基因注释与功能分析、SNP 数据分析与相关数据库、蛋白质组学数据分析、非编码 RNA 与复杂疾病、生物分子网络等.

本书要求读者具备医科本科的数学、计算机及生物化学基础知识. 本书可以作为基础医学、临床医学、预防医学、医学相关学科研究生或高年级本科生生物信息学课程教材，也可供医学或其他相关学科科技人员参考.

图书在版编目（CIP）数据

医用生物信息学理论与实践 / 李林主编. —北京：科学出版社，2019.4
生物医学信息基础课系列教材
ISBN 978-7-03-058555-4

Ⅰ. ①医… Ⅱ. ①李… Ⅲ. ①医学-生物信息论-医学院校-教材
Ⅳ. ①R318.04

中国版本图书馆 CIP 数据核字（2018）第 191850 号

责任编辑：张中兴 梁 清 张 晨 / 责任校对：杨聪敏
责任印制：张 伟 / 封面设计：迷底书装

科学出版社出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com
北京中石油彩色印刷有限责任公司印刷
科学出版社发行 各地新华书店经销
*
2019 年 4 月第 一 版 开本：720×1000 B5
2019 年 4 月第一次印刷 印张：11 1/4
字数：227 000
定价：59.00 元
（如有印装质量问题，我社负责调换）

前　言

生物信息学的研究对象是大规模的生物医学大分子数据. 该学科起源于基因组计划的开展而产生的人类 DNA 序列图谱，其得以发展依赖于新的高通量分子生物技术的出现和大量组学数据的产生. 新一代测序技术、新型质谱技术在基因序列、蛋白质组信息、功能组学信息的不断延伸与推进，产生了真正意义上的生物医学大数据. 对这些生物医学大数据的进一步研究与开发必将对生物医学问题的研究产生深远影响.

随着生命科学、医学科学的飞速发展，人类进入了探索生命奥秘的新时代. 科学家致力于从分子层面探索疾病成因与发展的分子机制. 当前精准医学和转化医学已成为生物医学研究所关注的焦点. 人类重大复杂疾病的分子生物医药数据，包括基因组、转录组、变异组、蛋白质组、代谢组、转录调控、蛋白质互作、病原生物全基因序列、临床病例资源、药物生物学活性、药物毒性、药物代谢动力学数据等大规模组学数据已经成为疾病和生命科学研究的核心资源. 医学相关学科的研究也逐渐涉及生物信息学的相关内容与技术. 例如，在生物力学研究中过去更多的是探索力与组织、细胞的形态间的相互作用. 当前的生物力学研究者不满足于此，为了揭示力相关疾病(如青光眼)，开始着手在蛋白质水平、基因水平上开展研究，因而必将借助于基因组、蛋白质组学的相关技术探索这些疾病的致病机制，以及临床中采用与力相关的处置后，探究在分子层面、细胞层面发生的变化及其与力的关联程度.

八年前，首都医科大学面向研究生开设了“生物信息技术概论”课程，面向本科生开设了“医学生物信息学”课程，然而在教学实践中一直没有发现一本较为合适的教材. 因此，在对多年授课经验与授课讲义和多媒体课件进行整理的基础上编写这本教材成为初衷. 同时，首都医科大学生物信息学研究团队十多年来一直活跃在生物信息教学和科研一线，使我们集全体人员智慧，融合经典生物信息学知识与技术、生物信息学新的进展编写一本针对性强的教材成为可能.

本书的目的是使医学本科生、研究生了解生物信息学的基础理论、生物信息学数据库、生物信息学常用算法及软件实现，熟悉生物信息学主要分支和使用生物信息学技术分析生物医学问题的一般方法. 以期使读者能够根据生物分子在基因表达调控中的作用，通过研究生物结构与功能相关的信息，加深对人类疾病的认识、改进对疾病的诊断与治疗方式.

全书分 7 章，第 1 章介绍生物信息学中最重要的基础工作之一，即序列比对，简要介绍序列比对的算法、BLAST 数据库的使用方法以及几种主流的多序列比对软件的使用. 第 2 章是基因芯片数据分析，主要介绍基因芯片数据预处理、聚类与分类分析方法及常用分析软件. 第 3 章是基因注释与功能分析，主要介绍两个常用数据库，即基因本体数据库和京都基因与基因组百科全书数据库，以及富集分析的主要内容. 第 4 章是 SNP 数据分析与相关数据库，主要介绍单核苷酸多态数据分析方法，如关联分析、互作分析等，同时介绍相关的数据库，如 dbSNP 数据库、dbGaP 数据库等. 第 5 章是蛋白质组学数据分析，介绍蛋白质组学的基本内容及数据分析方法，如蛋白质注释及功能预测，蛋白质相互作用网络构建及网络分析，也介绍了相关数据库的使用方法. 第 6 章是非编码 RNA 与复杂疾病，介绍复杂疾病中的非编码 RNA 调控分析的生物信息学方法. 第 7 章是生物分子网络.

本书是面向医学研究生和本科生专门编写的一本生物信息学入门级读物. 因此适合基础医学、临床医学及生物医学工程等医学相关学科的硕士、博士研究生及高年级本科生. 本书要求读者具备医科本科的数学、计算机及生物化学基础知识. 为了便于读者学习，我们力求使本书内容通俗易懂，做到图文并茂，同时也大幅度删减了烦琐的数学公式，为了让读者更清楚看到高清数据图表，本书在相应位置配备二维码，读者可以扫码观看细节图. 本书也可供医学或其他相关学科科技人员参考.

本书编写的初始动力来自于首都医科大学生物医学工程学科带头人刘志成教授和生物医学工程学院相关领导，在编写过程中得到了生物医学工程学院领导的不断鼓励和大力支持，在此，我们对他们表示深深的敬意和由衷的感谢. 同时也感谢对首都医科大学生物信息类课程建设和学科建设给予支持和帮助的人士，并向参与讨论的研究生们表示感谢.

生物信息学学科的特点之一是新进展层出不穷. 随着生物技术发展日新月异，生物信息学的方法与技术也高速发展. 由于编者水平和所涉猎范围的局限，书中肯定存在不足之处，希冀得到专家、同行和读者的批评指正，以使本书不断完善.

编　者

2018 年 1 月

试读结束：需要全本请在线购买：　www.ertongbook.com

目　录

第1章 序 列 比 对

序列比对(sequence alignment)的主要思想就是运用特定的算法找出两个或多个序列间产生最大相似性得分的空格插入和序列排列方案，对发现生物序列中有关功能、结构和进化信息有重要意义. 根据比对序列的个数，序列比对可分为双序列比对(pairwise sequence alignment)和多序列比对(multiple sequence alignment)；根据比对是着眼于全局还是局部，序列比对可分为全局比对(global alignment)和局部比对(local alignment). 本章简要介绍序列比对的算法，详细介绍 National Center for Biotechnology Information(NCBI)网站提供的 BLAST 数据库搜索的使用方法，以及几种主流的多序列比对软件的使用.

1.1 序列比对简介

序列比对是生物信息学中一项重要的基础工作. 一段 DNA 或蛋白质序列包含什么信息，以及和其他序列间存在什么关系，是研究人员遇到的首要问题.

序列比对运用特定算法找出两个或多个序列间产生最大相似性得分的空格插入和序列排列方案，对于发现生物大分子序列(如 DNA 或蛋白质等)中有关功能结构和进化的信息具有非常重要的意义. 在序列比对中，多序列比对可以发掘多个序列中的相似性信息，当两个序列不能很好地比对时，通过引入更多的序列，可有效地使两个难以直接比对的序列合理地关联起来.

1.1.1 同源、相似和距离

同源(homology)、相似(similarity)和距离(distance)的概念是序列比对和分析的基础. 同源是指两个序列享有一个共同的进化上的祖先. 同源是个定性的概念，没有度的差异. 与同源相关的两个概念是相似和距离，相似和距离是基于对序列中字符的精确比较，定量描述多个序列相似程度的度量. 相似性可以定量地定义为两个序列的函数，依据两个序列对应位置上相同字符的个数确定序列函数的值，值越大两个序列越相似. 距离也可以定量地定义为两个序列的函数，依据两个序列对应位置上差异字符的个数确定距离函数的值，值越小序列越相似. 例如，两条序列 a: ATTCGAGC，b: ATGCGATC. 编辑距离时一般用汉明距离(Hamming distance)表示，对于两条长度相等的序列，它们的汉明距离等于对应位置不同字符的个数，则 a、b 两序列的汉明距离为 2. 度量相似性时按照匹配计 1 分，不匹

配计 0 分的计分规则，相似性计分为 6.

在基因组测序中，同源性根据数据库搜索和序列比较确定. 同源分为垂直同源(ortholog DUS)和水平同源(paralog DUS)，垂直同源序列是指在种系形成过程中起源于一个共同祖先的不同种系中的 DNA 或蛋白质序列，水平同源序列是由序列复制事件产生的同源序列. 一般假定，同源序列具有相同的功能，但垂直同源和水平同源的基因功能未必总相同.

在基因组分析中有时同源和相似的关系很难确定. 一方面，同源序列的相似性可以很低，对于一个基因或蛋白质，进化可以产生物种间高度差异的碱基或氨基酸序列，但同时保持 DNA 序列、RNA 序列和蛋白质序列二级和三级结构的保守性；另一方面，非同源序列的相似性也可以很高，趋同进化可以产生物种间高度类似的碱基或氨基酸序列，它们对应于相同或相似的功能. 再者，由于氨基酸编码的冗余性，差异相当大的 DNA 序列也可产生差异相当小的蛋白质序列，这也是一种与同源无关的相似.

1.1.2 序列比对的作用

通过序列比对，可以确定一个蛋白质或核酸序列有哪些垂直同源序列或水平同源序列；确定哪些蛋白质或基因在特定的物种中出现；发现新基因；确定一个基因或蛋白质的变种；寻找对于一个蛋白质的功能和结构起关键作用的片段. 多序列比对还有更广泛的应用，例如，获得共性序列、序列测序、突变分析、种系分析、保守区段分析、基因和蛋白质功能分析等.

1.1.3 序列比对算法简介

用计算机进行序列比对，就是要找出两个序列的最长公共子序列，从而定量描述两个序列的最高相似度. 如对两条核酸序列 a: ATTCAGTCAGTA，b: ATTAGTCACGTA 进行如下比对.

```
a: ATTCAGTCAGTA          a: ATTCAGTCA--GTA
   |||     |||              |||  |||||  |||
b: ATTAGTCACGTA          b: ATT--AGTCACGTA
```

从上面两个比对可以看出，如果直接进行比对，公共序列只有 6bps，如果在合适的位置插入空格，那么得到的公共序列有 11bps，然而这只是两条 12bps 长的核酸序列，如果比对的序列长度增加，复杂性增加，那么找到最优的空格插入以及排列方式就会是一件非常复杂的问题，所以要借助合适的算法快速找到最优比对. 对于字符的插入或缺失产生的失配，可以通过引入空格使得原本可以对齐的字符对齐，但是对于替换引起的失配则需要考虑不同替换的意义，在序列比对中对于这类失配如何合理而精确地计分，就要考虑替换的各种情况，对于蛋白质序

试读结束：需要全本请在线购买： www.ertongbook.com

列还要考虑到氨基酸的不同理化性质，这就是所谓的计分矩阵. 下面简要介绍计分矩阵及其比对算法.

1. DNA 序列比对的替换计分矩阵

对于 DNA 和 RNA 序列，适用于 4 种碱基和 6 种彼此间替换关系的计分规则可用简单的替换计分矩阵(substitution matrix)来描述. 常用的 DNA 序列比对的替换计分矩阵有以下几种.

(1) 等价矩阵(unitary matrix)，是最简单的替换计分矩阵，其中，相同核苷酸间的匹配计 1 分，不同核苷酸间的替换计 0 分，如图 1-1A 所示，由于不含有碱基的任何理化信息和不区别对待不同的替换，所以在实际的序列比对中较少应用.

(2) 转换-颠换矩阵(transition-transversion matrix)，核酸的碱基按照环的结构特征分为双环的嘌呤(A、G)和单环的嘧啶(C、T)，环数不变的替换称为转换，如 $A \rightarrow G, C \rightarrow T$，环数发生变化的称为颠换，如 $A \rightarrow C, A \rightarrow T$ 等. 在实际进化过程中，转换发生的频率远比颠换高，如图 1-1B 所示的转换-颠换矩阵中，转换计–1 分，颠换计–5 分，反映了这种实际情况.

(3) BLAST 矩阵(BLAST matrix)，经过大量的实际比对发现，如果令被比对的两个核苷酸相同时计分为+5，不同时计分为–4，比对效果较好，如图 1-1C 所示，这个矩阵广泛应用于 DNA 序列比对，称为 BLAST 矩阵.

	A	T	C	G
A	1	0	0	0
T	0	1	0	0
C	0	0	1	0
G	0	0	0	1

A. DNA等价矩阵

	A	T	C	G
A	1	−5	−5	−1
T	−5	1	−1	−5
C	−5	−1	1	−5
G	−1	−5	−5	1

B. 转换-颠换矩阵

	A	T	C	G
A	5	−4	−4	−4
T	−4	5	−4	−4
C	−4	−4	5	−4
G	−4	−4	−4	5

C. BLAST矩阵

图 1-1　核苷酸转换矩阵

2. 蛋白质序列比对的替换计分矩阵

蛋白质序列由 20 种氨基酸构成，不同的氨基酸有不同的理化性质，会影响它们在进化过程中的相互替换性，如体积的差异、与水的亲和性等都会影响替换的概率. 因此，简单的计分系统是不够的，必须使用能够反映氨基酸的相互替换性的计分系统，常用的如：①等价矩阵，最简单的计分矩阵，相同的氨基酸计 1 分，不同的氨基酸计 0 分. ②遗传密码矩阵(genetic code matrix, GCM)，通过计算一个

氨基酸转变为另一个氨基酸所需的密码子变化的数目而得到，矩阵元素的值对应于代价. 如变化一个碱基就可以使一个氨基酸的密码子变化成另一个氨基酸的密码子，则这两个氨基酸的替换代价为 1，如需要两个碱基的改变，则替换代价为 2. 遗传密码矩阵常用于进化距离的计算，其计算结果可以直接用于绘制进化树，但在蛋白质序列比对，尤其是相似程度很低的蛋白质间的比对很少使用. ③疏水性矩阵(hydrophobic matrix)，在相关蛋白质之间，某些氨基酸可以很容易相互取代而不改变它们的生理生化性质，根据 20 种氨基酸侧链基团疏水性的不同及氨基酸替换前后理化性质变化的大小，制定了以氨基酸疏水性为标准的疏水性矩阵. 若一次氨基酸替换后疏水特性不发生大的变化，则这种替换得分高，反之得分低，适用于偏重蛋白质功能方面的序列比对. ④PAM 矩阵(point accepted mutation scoring matrix)，对于氨基酸之间的替换，对实际替换率的直接统计也可以导出合理的计分方法. Dayhoff 等研究了 34 个蛋白质家族，包括高度保守的和高度易突变的，根据对其氨基酸之间相互替换频率的统计得到了 PAM 矩阵，即可接受突变点和可接受突变百分比矩阵，该矩阵基于氨基酸进化的点突变模型，即如果两种氨基酸替换频繁，说明自然界易接受这种替换，那么这对氨基酸替换得分就应该高. PAM 矩阵是目前蛋白质序列比对中最广泛使用的计分方法之一. ⑤BLOSUM 矩阵(block substitution matrix)，由 Henikoff 首先提出的另一种氨基酸替换计分方法，也是通过统计相似蛋白质序列的替换率而得到的. PAM 矩阵是从蛋白质序列的全局比对结果推导出来的，而 BLOSUM 矩阵是从蛋白质序列块(短序列)的比对推导出来的. 基本数据来源于 BLOCKS 数据库，其中包括了局部多重比对，虽然没有使用进化模型，但优点在于可以通过直接的观察而不是通过外推获得数据. PAM 矩阵和 BLOSUM 矩阵都有许多不同的编号，这里的编号是指序列可能相同的最高水平，并且同模型保持独立性. 对于 PAM-n 矩阵，n 越小表示氨基酸变异的可能性越小，高相似序列间的比对应该选用 n 值小的矩阵，低相似序列间的比对应该选用 n 值大的矩阵；对于 BLOSUM-n 矩阵，n 越小则表示氨基酸相似的可能性越小，高相似序列间的比对应该选用 n 值大的矩阵，低相似序列间的比对应该选用 n 值小的矩阵.

3. 比对算法

用算法实现的两个序列的比对，就是找出两个序列最长的公共子序列，反映两个序列的最高相似度. 然而找出最长的共同子序列并不是一件容易的事. 动态规划(dynamic programming)算法是一种多阶段决策过程，通过将复杂问题分解为简单子问题进行求解的方法，通过动态规划算法可以实现序列间的比对. 动态规划的算法应用于生物信息源于 1970 年，首先由 S.Needleman 和 C.Wunsch 两人将

其应用于两条序列的全局比对，称为 Needleman-Wunsch 算法，后来 T.Smith 和 M.Waterman 两人于 1981 年对双序列的局部比对进行了研究，产生了 Smith-Waterman 算法.

基于动态规划算法可以实现双序列的全局比对、双序列的局部比对、多序列的全局比对以及多序列的局部比对. 对于多序列比对，由于动态规划方法的时间和空间的复杂性太高，人们发展了该算法的多种变体：①渐进多序列比对，首先使用动态规划算法构造全部 k 个序列的 $\binom{k}{2}$ 个配对比对，然后以计分最高的配对比对作为多序列比对的种子，按计分高低依次选择序列，逐渐向已构造的多序列比对中加入序列，形成一个树状结构的多序列比对结果. ②迭代法，在渐进多序列比对中，一个序列一经加入构造的比对结果，其配对比对便不再重新处理，对在比对中发现的错误或不适当的计分没有机会进行更正，这提高了比对的运行效率，但牺牲了准确性. 迭代法克服了渐进法中的不足，其基本过程是先用渐进多序列比对产生一个初始结果，再对序列的不同子集进行反复比对，并利用这些结果重新进行多序列比对，目标是改进多序列比对的总积分值. 迭代法常使用随机搜索，或者通过对比对结果重排来寻找更优的解，迭代持续直到比对计分值不再提高. ③基于一致性的方法，渐进多序列比对的基本方法是先产生全部的配对比对，然后根据配对比对的计分高低逐渐构造多序列比对. 基于一致性的方法则采用了另一种利用序列信息的方式，对每对序列中的每对字符计算如上的似然率. 根据基准测试数据的研究，基于一致性方法的多序列比对产生的结果常比渐进多序列比对产生的结果更加准确.

全局比对的共同特征是假定序列中所有对应的字符可以匹配，所有字符具有同等的重要性，空格的插入是为了使整个序列得到比对，包括使两端对齐，因此，更适合于比对高度相似且长度相当的序列. 局部比对不假定整个序列可以匹配，重在考虑序列中能够高度匹配的一个区段，可赋予该区段更大的计分权值，空格的插入是为了使高度匹配的区段得到更好的比对.

1.2 数据库搜索 BLAST

在分子生物学研究中，通过数据库搜索可以找出与新测定的碱基序列或氨基酸序列相似的序列，以推测未知序列是否与已知序列同源，具有哪些功能，属于哪个基因家族等信息. BLAST 是最常用的数据库搜索程序，本节介绍 BLAST 的功能及应用.

1.2.1 BLAST 简介

BLAST 是目前最常用的数据库搜索程序，国际知名的生物信息中心都提供基于 Web 版的 BLAST 服务. 本节介绍 NCBI 提供的数据库搜索程序 BLAST(https://blast.ncbi.nlm.nih.gov). NCBI 网站不仅提供在线服务，也可以下载安装本地 BLAST，但必须有本地的 BLAST 格式的数据库，可以直接下载，也可以通过提供的格式转换工具转换而得到.

1.2.2 BLAST 搜索页面的功能

NCBI 网站 BLAST 有五个基本搜索页面(图 1-2)，每个页面执行特定类型的序列比对. 表 1-1 简要介绍 blastn、blastp、blastx、tblastn、tblastx 这五个搜索页面，并给出这些搜索需要用到的公共数据库.

扫码看图

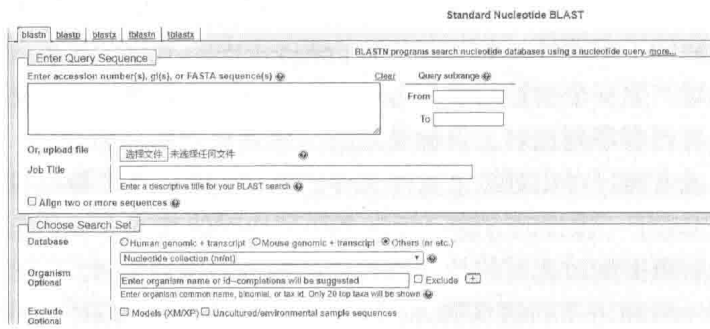

图 1-2　BLAST 基本搜索页面

表 1-1　BLAST 基本搜索页面

搜索页	查询序列和数据库类型	比对类型	程序和函数(默认函数用粗体)
Nucleotide blast (blastn)	核酸 vs 核酸	核酸 vs 核酸	**megablast:** 用于序列识别，物种内比较 **discontiguous megablast:** 用于跨物种比较，或用编码序列搜索 **blastn:** 用于搜索短序列，或跨物种比较
Protein blast(blastp)	蛋白质 vs 蛋白质	蛋白质 vs 蛋白质	**blastp:** 一般的序列识别和相似性搜索 DELTA-BLAST: 灵敏度高于 blastp 的蛋白质相似性搜索 PSI-BLAST: 通过迭代法搜索构建位点特异性计分矩阵 PSSM(Position specific scoring matrix)或搜索与查询序列亲缘关系较远的蛋白质序列 PHI-BLAST: 适用于带模式的短序列的查询，能够搜索到既和查询序列相配又和特定模式相配的数据库记录，能用来帮助判断这个蛋白质属于哪个家族

续表

搜索页	查询序列和数据库类型	比对类型	程序和函数(默认函数用粗体)
blastx	核酸(翻译后)vs 蛋白质	蛋白质 vs 蛋白质	**blastx:** 识别查询核酸序列编码的蛋白质产物
tblastn	蛋白质 vs 核酸(翻译后)	蛋白质 vs 蛋白质	**tblastn:** 搜索数据库中序列所编码的蛋白质，识别出和查询蛋白质序列相似的序列
tblastx	核酸(翻译后)vs 核酸(翻译后)	蛋白质 vs 蛋白质	**tblastx:** 基于编码能力识别与查询核酸序列相似的核酸序列

1.2.3　BLAST 常用搜索数据库

不同的 BLAST 程序使用的搜索数据库不同，主要分为核酸搜索数据库和蛋白质搜索数据库，下面分别加以介绍.

1. 常用核酸搜索数据库

搜索使用的核酸数据库见表 1-2.

表 1-2　BLAST 常用核酸搜索数据库

数据库	数据库内容
nr(nt)default	所有 GenBank + EMBL + DDBJ + PDB 序列，不包括 PAT、EST、STS、GSS、WGS、TSA 中的序列和相位 0, 1, 2 HTGS 序列，大部分非冗余
refseq_rna	NCBI Reference Sequence Project 中人工审核的序列(以 NM_, NR 开头)和预测序列(以 XM_, XR_开头)
refseq_genomic	NCBI Reference Sequence Project 中的基因组序列
refseq_representative_genomes	NCBI RefSeq Reference 和 Representative genomes 包括广泛的类群，如真核生物、细菌、古细菌、病毒及类病毒. 这些基因组有最小的冗余度，真核生物每个物种一个基因组，其他都是一个物种有不同的菌株，属于人工审核的基因组
chromosome	NCBI Reference Sequence Project 中的完整基因组和完整染色体序列
Human G+T	人类基因组序列最新版本中的基因组序列和人工审核的以及预测的 RNA 序列
Mouse G+T	鼠基因组序列最新版本中的基因组序列和人工审核的以及预测的 RNA 序列
est	GenBank + EMBL + DDBJ 中 EST 数据库中序列所构成的数据库
HTGS	未完成的高通量基因组序列；相位 0, 1, 2 HTGS 序列
wgs	全基因组鸟枪序列的装配片段
pat	来自 Patent division of GenBank 中的核酸序列

续表

数据库	数据库内容
pdb	来自 Protein Data Bank 的三维结构数据中的核酸序列
TSA	组装自 RNA-seq SRA 数据的转录组鸟枪序列装配
16S microbial	来自 Targeted Loci Project 的微生物 16S rRNA 序列

2. 常用蛋白质搜索数据库

搜索常使用的蛋白质数据库见表 1-3.

表 1-3　BLAST 常用蛋白质搜索数据库

数据库	数据库内容
nr default	非冗余 GenBank CDs 翻译序列 + RefSeq + PDB + SwissProt + PIR + PRF, 不包含 PAT, TSA 和 env_nr 的序列
refseq_protein	NCBI Reference Sequence Project 中的蛋白质序列
swissprot	最新的主要版本 UniProtKB/SWISS-PROT 蛋白质序列数据库(无新增的更新)
Landmark	Landmark 数据库包括广泛类群的代表基因组中的蛋白质组
pat	来自 Patent Division of GenBank 中的蛋白质序列
pdb	来自 Protein Data Bank 的三维结构数据中的蛋白质序列
env_nr	由宏基因组核酸序列注释的 CDS 翻译得到的蛋白质序列
tsa_nr	由转录组鸟枪装配注释的 CDSs 翻译得到的蛋白质序列

1.2.4　BLAST 标准核酸搜索页面简介

NCBI 所提供的核酸 Web 版的 BLAST 服务中，五个基本搜索页面和大部分的特定搜索页面都采用标准的搜索页面，下面对标准搜索页面的功能做简要介绍.

“Nucleotide-BLAST” 链接加载 “标准核酸 BLAST” 搜索页面(图 1-3). 页面顶端包含显示页面位置的选项卡(A)，页面的标题，一组在五个核心 BLAST 搜索页面间快速切换的选项卡(B)，恢复页面默认设置的链接和给自定义设置的搜索页面设置书签的链接(C). 默认的显示页面包含三个部分，功能描述如下.

1. 输入查询序列

主输入框(D)可输入不同格式的核酸查询序列. 对于一个查询序列 “Query subrange” 框(E)定义搜索使用的一个查询片段. 用纯文本文件保存的查询序列可

试读结束：需要全本请在线购买：　www.ertongbook.com