本 科 生 毕 业 论 文（设计）

文献综述外文翻译和开题报告

题目　重大疾病相关的蛋白质组学的模式识别

姓名与学号　秦 臻 3130000210

指导教师　庞天晓、张南松

年级与专业　四年级 统计学

所在学院　数学科学学院

一、题目：重大疾病相关的蛋白质组学的模式识别

二、指导教师对文献综述和开题报告的具体内容要求：

文献综述要求：

查阅与模式识别相关的文献。要求至少 8—10 篇以上（其中外文文献不少于3—5篇）。根据阅读的国内外文献撰写文献综述报告，要求根据主题展开，文献综述内容要切题，包括

（1）简述与模式识别分析相关的近几年的论文的研究内容和相关进展，掌握当前该领域的研究前沿，分析你毕业论文研究的内容与这些论文的差异和相关。

（2）分析论文拟采用的研究方法，包括理论方法和统计模拟技术等。

（3）简述各参考文献的创新性、存在的问题或未能解决的问题。

（4）要求翻译其中的一篇外文文献，结构完整，语句通顺。

开题报告要求：

（1）分析该课题的研究意义（包括理论意义和实际意义，并分析课题与本专业的关系）。

（2）根据文献综述分析课题的研究背景（即要解决什么问题？这些问题在其他文献中有没有讨论过？本文所讨论问题的角度

与已有参考文献中所涉及的问题的差异，课题的主要创新点是什么？）。

（3）选题的可行性分析（一般从数据的可获得性、研究技术的可行性等方面去说明）。

（4）主要研究内容（这部分要展开写，主要包括理论研究内容、实证分析内容和调研内容等）。

（5）根据研究的内容写出具体的实施计划。

（6）明确论文最后预期结果。

指导教师（签名）_____

2017 年 3 月 3 日

基于神经网络的 DNA 芯片数据分析研究综述

生物信息学是研究生物信息的采集、处理、存储、传播，分析和解释等各方面的学科，也是随着生命科学和计算机科学的迅猛发展，生命科学和计算机科学相结合形成的一门新学科。它通过综合利用生物学，计算机科学和信息技术而揭示大量而复杂的生物数据所赋有的生物学奥秘。在生物信息学中，DNA 芯片是其中一个热门的研究领域。DNA 芯片技术就是指在固相支持物上原位合成寡核苷酸或者直接将大量的 DNA 探针以显微打印的方式有序地固化于支持物表面，然后与标记的样品杂交，通过对杂交信号的检测分析，即可获得样品的遗传信息。是伴随“人类基因组计划”的研究进展而快速发展起来的一门高新技术。本文介绍了 DNA 芯片数据的特点以及分析 DNA 芯片数据的几种常见方法，其中基于机器学习理论的几种分析方法和算法的发展情况是本文的一个重点。

1 前言

基因表达的数据分析离不开多种分析手段和工具，只有用好的工具才能尽可能多地挖掘数据中的信息。为了满足分析系统的需要，两个主要类别的功能是必备的，一个是数据共享环境，另一个是功能强大的分析工具。前者允许一个研究团队贡献基因芯片实验数据库、相关说明数据、文献、实验流程等通用资料；后者可以从基因组数据库中获得临床或者生物学知识。统计学家一直致力于获得更有效的分析方法来解决后者的问题，于是机器学习方法被引入到 DNA 芯片数据分析之中。机器学习的观点是设计出一种像人类一样可以学习的机器，在复杂的环境中获得经验并从中获得智慧即从而在现有资料中挖掘出所需的信息。生物信息学所研究的课题涉及到从高度复杂的生物系统获得的大量数据中找到我们所需要的数据，因此机器学习对于研究生物信息学相关问题是适用的。

2 国内外现状

国内对于机器学习的研究很早就已经开始，而且具有以下几个特点。第一，我国对于机器学习理论的方法学研究在某些领域上具有国际领先的地位。机器学习所关注的一个非常重要的问题是对于泛化问题的研究。也就是说，尽可能推广机器学习的应用范围。我国科学家研究的集成学习的成果，在国际上享有盛誉。第二，机器学习技术在数据挖掘中体现出了十分重要的商业作用。在数据挖掘领

域，机器学习方法能够处理各种不同种类的数据，所以机器学习收到这个领域的重视。

然而，与机器学习在商业领域的应用相比，我国对于基因芯片的研究仍处于发展阶段。目前，核心科技主要掌握在国外的跨国公司，例如美国昂飞公司(Affymetrix， Inc.)，全球第一家生物芯片公司，提供“完整的基因芯片解决方案”。我国在的基因芯片发展中起到代工厂的作用，对于机器学习在 DNA 芯片中应用，我国落后于发达国家。

西方国家对于 DNA 芯片的研究是十分超前的。美国在开展人类基因组计划以后，于 1998 年正式启动基因芯片计划，美国国立卫生部、能源部、商业部、司法部、国防部、中央情报局等均参与了此项目。同时斯坦福大学、麻省理工学院及部分国立实验室也参与了该项目的研究和开发。英国剑桥大学、欧亚公司正在从事该领域的研究。世界大型制药公司尤其对基因芯片技术用于基因多态性、疾病相关性、基因药物开发和合成或天然药物筛选等领域感兴趣，都已建立了或正在建立自己的芯片设备和技术。

3 研究方向

国内主要的研究方法是支持向量机方法和隐马尔科夫模型方法，这两种方法也是机器学习中的经典方法，在国内期刊中有许多关于这个问题的研究报告和论文。其中大多数使用的是支持向量机方法和隐马尔可夫模型方法，

支持向量机(Support Vector Machine，SVM)是 Corinna Cortes 和 Vapnik 等于 1995 年首先提出的，它在解决小样本、非线性及高维模式识别中表现出许多特有的优势，并能够推广应用到函数拟合等其他机器学习问题中。支持向量机（SVM，还支持矢量网络）是与相关的学习算法有关的监督学习模型，可以分析数据，识别模式，用于分类和回归分析。一个支持向量机的构造一个超平面，或在高或无限维空间，其可以用于分类，回归，或其它任务中设定的超平面的。直观地，一个良好的分离通过具有到任何类(所谓官能余量)的最接近的训练数据。SVM 的关键在于核函数。低维空间向量集通常难于划分，解决的方法是将它们映射到高维空间。但这个办法带来的困难就是计算复杂度的增加，而核函数正好巧妙地解决了这个问题。也就是说，只要选用适当的核函数，就可以得到高维空间的分类函数。在 SVM 理论中，采用不同的核函数将导致不同的 SVM 算法。点的最大距离的超平面的一般实现中，由于较大的裕度下分类器的泛化误差。

隐马尔可夫模型（Hidden Markov Model，HMM）是统计模型，它用来描述一个含有隐含未知参数的马尔可夫过程。其难点是从可观察的参数中确定该过程的隐含参数。然后利用这些参数来作进一步的分析。隐马尔可夫模型（HMM）可以

用五个元素来描述，包括 2 个状态集合和 3 个概率矩阵：

1. 隐含状态 S；2. 可观测状态 O；3. 初始状态概率矩阵 π；4. 隐含状态转移概率矩阵 A；5. 观测状态转移概率矩阵 B。

隐马尔可夫模型实际上是标准马尔可夫模型的扩展，添加了可观测状态集合和这些状态与隐含状态之间的概率关系。

除了上述两种方法，由于神经网络具有运用已知认识新信息，解决新问题，学习新方法，预见新趋势，创造新思维的能力，所以我们将神经网络处理问题的方法介入进来，处理模式分类的问题。 神经网络的主要特点有:高度的并行性，高度的非线性全局作用，良好的容错性与联想记忆功能。十分强的自适应，自学习功能。

传统的分类识别方法有很多缺陷，尤其是对于非线性的识别非常困难，但是神经网络可以解决这列问题。而且神经网络模型有不同的处理方式，可以采用不同的算法灵活处理不同的 DNA 芯片数据。所以，人工神经网络就成为解决分析基因芯片数据的又一个有力的方法，而且这种方法具有很大的潜力，一旦取得重大进展，就可以产生极大的经济和社会价值。

4 进展情况

国内对于 DNA 芯片分析的发展还处于起步阶段，大部分研究人员专注于 DNA 计算，机器学习算法，DNA 芯片的制作以及工业化方面。目前，基于基因芯片的数据处理技术已经引起了国内外的广泛关注，基因表达谱分析已经成为生物信息学领域的重要课题。国外的许多研究机构针对基因表达谱已经提出了一些具有代表性的方法，并对若干种不同类型的的基因表达谱进行了研究，各种算法及基于这些算法的芯片产品已经陆续出现。当前国内对基于基因分析的研究还不完善，从事该项研究的人员较少，与国外存在着相当大的差距。

第一代微阵列分析方法在过去 5 年内已经证明了表达数据可以用在大量的类别发现和分类预测等生物问题中，包括癌症分类问题。用机器学习和统计方法来解决这些问题，包括区分形态，预测治疗后的效果，以及发现疾病的分子标记。现在基于微阵列的不同形态学的分类可以在很多情况下成功应用。预测诊断结果或药物反映的性能是有限的，但有些结果是很精确的。大多数微阵列分析结果仍然需要进一步实验验证和进一步的研究。现在很多人正在致力于这一方向。在一些情况下微阵列的分析结果被应用在临床诊断中。

当前主要使用的模式识别方法可以分为有监督和无监督两大类。无监督学习或称为聚类，是指基于数据集中的不同特征把数据样本进行聚类分析的识别方法。同样利用基因表达谱数据之间的相似性将样本尽可能的精确聚类。有监督识别方

法是指通过对已知类别学习样本进行训练从而构建分类器，再对未知样本进行分类的识别方法。这方面的尝试包括使用加权表决法，使用神经网络方法，还有支撑向量机法，以及贝叶斯方法，决策树等等。

5 存在问题

由于 DNA 芯片数据具有高维数，高噪声，高相关等特点，基于 DNA 芯片数据进行机器学习分析仍然面临很多困难，一些古典问题的解决方法不再实用，因此需要用很多具体措施提高机器学习分类器的性能，排除无关因素干扰。21 世纪虽然现在已经有许多较为成熟的算法，但性能大多都还达不到实用水平。首先最大的问题就是准确度不够，在训练集合较大而且是两类的情况下～般才达到80%左右，这样的精确程度对于医学诊断来说是远远不够的。

另外，微阵列数据应用到癌症诊断和分型时，从两类到多类问题的扩展还需要寻找更专用更有效的方法。再次，算法性能的好坏在很大程度上依赖于数据本身和初始条件的选择，需要寻找鲁棒性更好的算法。DNA 微阵列技术为生物学和医学研究带来前所未有的机遇的同时，其所产生的海量和复杂的微阵列数据却对现有的数据处理和分析方法提出了巨大的挑战。

第一，微阵列数据具有很高的维度（基因），通常有五千至一万五千维，而且这些基因维度之间又有非常复杂的关系。第二，实验的复杂和费用的昂贵导致微阵列数据具有较少的样本，并与巨大的基因数目构成不平衡的矛盾。这种矛盾造成大多数经典模式识别和机器学习方法不能被直接应用，比如，Fisher 线性分析所要求的总类内样本协方差矩阵将成为奇异阵。第三，微阵列数据天生具有高噪声和高变异等数据分析难点。第四，微阵列数据中大量有用变量被隐藏。这可能需要使用概率统计的方法以挖掘和推导这些潜在的生物信息。另外，当考虑时间问题（比如死亡时间，癌症复发时间）时，又会产生不期望的审查中止（right－censoring）以致数据分析变得更加困难。

《重大疾病相关的蛋白质组学的模式识别》
开题报告

1 课题意义

近几十年来，随着科学技术的迅猛发展，人类在科学研究和生产活动的许多领域都产生和积累了大量的数据，对这些数据进行分析以发掘数据中含有的有用信息，并用利用这些信息来指导未来的科研和生产活动，进而改善科研和生产的效率，几乎是所有领域的发展趋势之一。正是在这样的大背景下，利用统计软件进行生物 DNA 芯片数据分析在统计学和计算机科学中成为非常热门的话题。

随着大规模 DNA 芯片技术的发展，利用机器学习方法研究生物信息成为了可能。人们利用基因芯片在一次实验中可同时获得组织样本中成千上万个基因的表达数据，基于基因表达谱数据，采用模式识别与数据挖掘技术建立起有效的预测和分类模型，不仅有助于 DNA 模式的分类鉴别，还有助于其他有效数据的发现，有利于对基因特征不同的疾病确定正确的治疗方案，为研究生物信息学提供了强大的科学依据。

2 背景及可行性分析

DNA 微阵列技术所蕴涵的巨大科学价值不仅在于能够帮助人们探索生物体内基因调控及其相互作用的机理，更重要的是它联系了人类基因组序列与临床医学，为人类疾病的诊断和防治开辟了全新的途径。通过机器学习方法，人们可以分析 DNA 芯片中的数据，从而得出更多关于人类生命的有价值的线索。

机器学习从测试数据中自动分析获得规律，并利用规律对未知数据进行识别、分类和预测。机器学习与生物信息学之间的联系是一个长期而复杂的历史过程。Khan 等人在 2001 年使用人工神经网络技术进行 SRBCT 癌症的诊断与分类。他们的实验首先收集了 63 个病人用作训练样例，训练神经网络从而建立了 SRBCT 肿瘤的诊断模型。该模型在另外 25 个独立测试病例上进行诊断检验，结果成功诊断了全部测试样例。这说明通过神经网络方法及其推广，确实可以有效解决 DNA 芯片分析问题，事实上，这种方法也是广泛使用的方法之一。

统计软件是研究生物序列的重要工具，想要利用 DNA 芯片数据进行分析，统计软件必不可少。本论文以 R 软件为主要统计软件，R 是一套完整的数据处理、计算和制图软件系统。更重要的是基于 R 的一个用于分析高通量基因组数据的工具，Bioconductor。这个工具的应用功能是以包的集成形式呈现在用户面前，它提供的软件包中包括各种基因组数据分析和注释工具，其中大多数工具是针对 DNA 微阵列或基因芯片数据的处理、分析、注释及可视化的。

3 调研报告

微阵列基因芯片技术的成熟带来了基因表达谱数据分析方法的日新月异。从本质上讲，通过基因芯片技术实验所直接获得的是一个基因表达谱，基因芯片技术的实际应用就是通过对基因表达谱的生物信息学处理来实现的。本文采用的数据是使用测量相对荧光强度而获得的基因表达谱数据，为了进行随后的数据处理，首先要把基因表达谱数据从杂交图像中提取出来，基因表达谱数据是以矩阵的形式表示的，也即基因表达矩阵，矩阵的各行表示不同的基因，矩阵的各列表示不同的样本或实验条件。例如对于肿瘤芯片的数据，红光和绿光表示不同的 RNA，那么有以下矩阵：

$$x_{np} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}.$$

其中 n 表示基因个数，p 表示样本个数，矩阵元素是红光强度与绿光强度之比。通过基因芯片，就可以把复杂的人类 DNA 特征转化为矩阵形式，为接下来的分析工作提供便利。我们下面介绍基于人工神经网络的分析方法。

人工神经网络是一种计算模型，它以大脑的神经网络结构为原型。在大脑的简化模型中，它包含了大量的基本计算单元，也就是神经元，以复杂的通信网络形式彼此互联，通过他们，大脑才能够执行高度复杂的计算。人工神经网络是规则的计算结构，它模仿大脑的计算框架进行建模。

人工神经网络的基本构成单元就是神经元，它是一种有多个输入和一个输出的非线性单元，可以有反馈输入和阈值参数。连接模式是指神经元之间的连接关系，有单层、多层和循环连接模式。前两种连接模式构成的都是前向网络。第三种是包含反馈的连接模式。前馈型网络的输出只由当前输入、网络参数和结构决定，而循环网络的输出由当前输入和先前的输出两者、以及网络参数和结构决定，因此有短期记忆的性质。神经网络模型的基本特征是由其结构决定的，可归纳为：

1．非线性：非线性关系是自然界的普遍特性。人工神经元处于激活或抑制两种不同状态，这种行为在数学上表现为一种非线性关系。具有阈值的神经元构成的网络具有更好的性能，可以提高容错性和存储容量。

2．非局域性：一个神经网络通常由多个神经元连接而成。一个系统的整体行为不仅取决于单个神经元的特征，还要由单元之间的相互作用、相互连接所决定，通过单元之间的广泛连接模拟大脑的非局域性。联想记忆是非局域性的典型例子。

3．非定常性：人工神经网络具有自适应、自组织、自学习能力。神经网络处理的信息可以有各种变化，而且在处理信息的同时，非线性动力系统本身也在

不断地变化。

4．非凸性：一个系统的演变方向，在一定条件下将取决于某个特定的状态函数，例如能量函数，它的极值相应于系统比较稳定的状态。非凸性是指这种函数有多个极值，故系统具有多个较稳定的平衡态，这将导致系统演化的多样性。

好的算法也要有工具来实现。Bioconductor 中有很多精心设计、维护良好且广泛支持的与机器学相关的 R 程序包。与机器学习，生物信息学相关的 R 软件程序包功能十分强大。如常用的基因过滤软件包（genefilter），这是一种常用的滤除工具，包括缺省值数量，ANOVA 中的 P 值，Cox 模型等。另外丰富的注释工具，如注释软件包（annotation）可以帮助分析数据和直观显示产生的结果。

4 研究方案

这次研究主要需要使用四种研究方法。首先是文献研究法。文献研究法是根据一定的研究目的或课题，通过调查文献来获得资料，从而全面地、正确地了解掌握所要研究问题的一种方法。文献研究法被子广泛用于各种学科研究中。其作用有：能了解有关问题的历史和现状，帮助确定研究课题；能形成关于研究对象的一般印象，有助于观察和访问；能得到现实资料的比较资料；有助于了解事物的全貌。然后是模拟法。模拟法是先依照原型的主要特征，创设一个相似的模型，然后通过模型来间接研究原型的一种研究方法。GenBank 中有很多 DNA 芯片数据，可以用这些数据来进行分析。接着是数量研究法。数量研究法也称“统计分析法”，指通过对研究对象的规模、速度、范围、程度等数量关系的分析研究，认识和揭示事物间的相互关系、变化规律和发展趋势，借以达到对事物的正确解释和预测的一种研究方法。最后是探索性研究法。探索性研究法是高层次的科学研究活动。它是用已知的信息，探索、创造新知识，产生出新颖而独特的成果或产品。

神经网络是广泛使用的模式分类工具，在肿瘤基因表达谱的研究领域，也已经尝试了多种人工神经网络，如 BP 网络、概率神经网络、自组织特征映射等等。BP 网络是应用最广泛的一种有监督学习的神经网络口，也是一种单向传播的多层前向网络，网络除输入输出节点外，还有一层或多层隐含节点。它采用误差反向传播算法对网络权值不断调整，从而进行学习训练。在理论研究和实际应用中，人们最常用的是具有线性输出的单隐层网络。当网络的训练样本问题解决以后，网络的输入层节点数和输出层节点数便已确定。因此，多层前馈网络的结构设计主要是解决设几个隐层和每个隐层设几个节点的问题。理论分析证明，具有单隐层的前馈网络可以映射出所有非线性连续函数。网络除输入输出节点之外，还有一层或多层的隐层节点，同层节点中没有任何耦合。对予输入信息，要先向前传播到隐层的节点上，经过各单元的激活函数(又称为作用函数、映射函数等)运算后，把隐含节点的输出信息传播到输出节点，最后给出输出结果。根据生物学和

神经生理学，人脑中分布着大量的协同作用的神经元群体，同时大脑网络又是一个复杂的反馈系统，既包括局部反馈，也包括整体反馈及化学交互作用，聚类现象对于大脑的信息处理有着重要作用。在大脑皮层中，神经元呈二维空间排列，神经元的输入信号主要有两部分，一种是来自感觉组织或其他区域的外部输入信号；另一种是同一区域的反馈信号，形成信息交互。神经元之间的信息交互方式有很多种，但是邻近神经元之间的局部交互有一个共同的方式，就是侧向交互：以发出信号的神经元为圆心，对最邻近的神经元的交互作用表现为兴奋性侧反馈，对较远的神经元的交互作用表现为抑制性侧反馈，对更远的神经元的交互作用则又是弱兴奋。

在实际操作中，可以利用 SGD 算法启发式搜索最优解。在最小化风险函数的 SGD 算法之中，嵌套经典的 backpropagation 函数，让它与神经网络的非线性目标函数相关。图体积为 S(n)的神经网络能够描述所有运行时间为$O\left(\sqrt{s(n)}\right)$的假设类预测。反向传播算法（BP 算法）能够高效计算各个边上权重损失函数的梯度，从而简便计算。

5 论文框架结构

机器学习在 DNA 芯片数据分析中占据重要地位，而人工神经网络是一种重要的机器学习思想。本文围绕机器学习的效率和学习性能进行定性和定量分析，并将对数据分析的结果和原因进行探讨，最后归纳本文的基本结论，并引申出简要的应用前景。

全文共分五章，第一章为绪言，第二章为文献综述，第三章为 DNA 芯片数据，第四章为人工神经网络算法及实现，第五章为基本结论及应用方向，将结合第三、四章研究成果，归纳基本结论，提炼本文观点。

第一章 绪言

第一节 选题背景及意义

第二节 研究框架和研究方法

第二章 文献综述

第三章 DNA 芯片数据

第一节 DNA 芯片数据的特点

第二节 DNA 芯片数据的获取——重要的数据网站

第四章 人工神经网络算法及实现

第一节 算法的深刻原理

第二节 算法的几种具体实现

第五章 基本结论及应用方向

6 实施计划

首先学习基本方法，生物信息学经过一段时间的发展，形成了很多基本而且有效的方法，对这些方法的学习是之后研究的关键。利用方法解决问题模拟是研究中必不可少的过程。本论文不仅关注理论方法，同时关注在统计软件中的实际操作，和如何解决实际问题。思考和创新，通过研究和思考，发现传统方法的不足之处，提出新思路，新观点或者改进方法。

具体的实施方案如下表所示：

项目	时间	内容	提交材料
选择导师	2016 年 10 月	选择生物统计为论文方向	无
学习基础知识	2016 年 10 月至 12 月	学习《理解生物信息学》，了解 DNA 芯片	无
选题	2017年1月	选择 DNA 芯片分析的统计分析方法	无
准备材料	2017年2月	查找阅读文献资料	翻译,综述
开题答辩	2017年3月	在理解课题的基础上进行答辩	开题报告
研究	2017年4月	对算法进行理论研究	无
实验	2017年5月	利用数据检验结果	无
得出结论	2017年5月	从实验结果得出相应的结论，分析结论	无
结题答辩	2017年5月	结束课题，提交论文	毕业论文

7 预期结果

本论文期望以生物信息学为背景，以统计软件分析为主要方法，把机器学习中的一些算法和技术，如人工神经网络应用到生物 DNA 芯片数据分析中来。虽然本论文涉及了较多理论方面的内容，但是在研究过程中还是应该把软件分析当作核心工作，把理论与实践相结合。

首先在具体算法方面，针对无监督学习算法的多类别应用问题应该进一步深入研究，解决目前效果不佳、类别混杂严重的现状，获得更好的聚类结果，可以考虑将无监督方法与支撑向量机相结合来解决这一问题。采用理论较为成熟、应用较为广泛的无监督方法首先对基因进行聚类，确保结果具有一定的可信性，再用 SVM 对结果进行训练，对聚类结果的准确性加以反馈；利用反馈信息改进聚类方法。探索微阵列数据分析中利用无监督学习的结果指导有监督学习的过渡，提高微阵列数据统计分析的效能。

事实上，人工神经网络在 DNA 芯片数据分析中并没有占据绝对领导地位。一些相关文献对各种分类技术在癌症诊断与分类中的应用进行了比较。比如，Dudoit 等人在公开发表的癌症微阵列数据上系统比较了常用的分类器，而 Statnikov 等人在 11 个数据集上对各种分类器进行评价，结果发现支持向量机技术在对 11 个数据集的分类应用中具有最高的分类精度。而 Pochet 等在

分析了几种常用的微阵列数据分析方法后也得到相同的结论。

另外，Klivans 和 Sherstov 在 2006 年证明，即使我们允许独立学习，对于这些部分的相交部分也不能使用神经网络进行学习。所以学者认为，这意味着即使允许更大的网络或者其他的高效执行的激活函数，也不能通过神经网络的训练找到高效的算法。

虽然如此，计算机能力的提升和新的算法的提出，使神经网络的有效性获得了很大的突破。特别是深度学习网络已经在很多领域表现出十分出色的性能。本次论文研究的目的之一就是挖掘人工神经网络在数据分析方面的潜力，发展与支持向量机不同的 DNA 芯片分析方法，试图达到更好的性能。

在主要目标之外，还可以进行多分类器集成系统的实验。许多事实已经证实，一个集成系统在许多领域内都能取得比单个优秀的分类器更高的识别效果，因此多分类器集成系统的应用领域在不断扩大。多分类器集成系统的概念虽然提出较早，但真正在 1990 年以后才得到研究者的广泛重视。由于微阵列数据识别是典型的高维小样本问题，虽然研究者在不断研究设计某种分类模型，但基于单个精确分类器能够实现的预测精度和泛化能力总是有限的。相比之下，集成系统更容易解决这样困难的问题。如果人工神经网络结合其他方法，一定能够有效提升算法的效率，这是一个值得关注的进一步研究方向。

8 参考文献

[1] Leif E. Peterson, Mustafa Ozen, Halime Erdem, Andrew Amini, Lori Gomez, Colleen C. Nelson, and Michael Ittmann. Artificial Neural Network Analysis of DNA Microarray-based Prostate Cancer Recurrence[OL]. US:IEEE. 2005

[2] Paolo Arena, Maide Bucolo, Luigi Furtuna, Luigi Occhipint. Cellular neural networks for real-time DNA microarray analysis[OL]. US: MEMB. 2002

[3] Beatriz A. Garro a, Katya Rodríguez a, Roberto A. Vázquez. Classification of DNA microarrays using artificial neural networks and ABC algorithm[OL]. US: Applied Soft Computing. 2016

[4] Hieu Trung Huynh 1 , Jung-Ja Kim 2 and Yonggwan Won. Classification Study on DNA Microarray with Feedforward Neural Network Trained by Singular Value Decomposition[OL]. US: International Journal of Bio-Science and Bio-Technology. 2006

[5] Kyung-Joong Kim, Sung-Bae Cho. Evolving Artificial Neural Networks for DNA Microarray Analysis[OL]. US: IEEE. 2003

[6] Jagdish Chandra Patra, Lei Wang, Ee Luang Ang, and Narendra S. Chaudhari. Neural Network-Based Analysis of DNA Microarray Data[OL]. US: Proceedings of International Joint Conference on Neural Networks. 2005

[7] 孙啸等，R 语言及 BIOCONDUCTOR 在基因组分析中的应用[M]，科学出版社，2006

[8] 高山等，R 语言与 Bioconductor 生物信息学应用[M]，天津出版传媒集团，2014

[9] 皮埃尔·巴尔迪，生物信息学——机器学习方法[M]，中信出版社，2003

[10] 李瑶，基因芯片数据分析与处理[M]，化学工业出版社，2006。

生物信息学的统计分析：介绍

Warren J. Ewens, Gregory Grant

6.4.2 有间隙的全局比较和动态规划算法

假设我们给出一个由替换矩阵和线性间隙惩罚组成的评分方案，我们的目标是找到，可能的两个序列的全局比较（允许有间隙），得分最高的一个（或一些）。一种原则上的方法是列出所有可能的全局比较和他们的得分，然后标出最高得分。然而，当序列很长的时候，这种方法在计算上是不可行的，我们需要更加快速的算法。我们下面将介绍一种算法，但是首先我们要说明上面所说的穷举法是非常低效的，只需要考虑对于长度为 m 的序列$x = X_1X_2X_3 \ldots X_m$，和长度为 n 的序列$y = Y_1Y_2Y_3 \ldots Y_n$，他们之间可以产生多少全局比较。我们记这个数为 c(m, n)。因为匹配两个缺失项是没有意义的，所以插入缺失标记之间的比较是不允许出现的。

设 g(m, n)是有对齐的残留物对无视插入缺失标记相同的组合得到的组数。那么 g(m, n) < c(m, n)，这是 c(m, n)的一个下界。如下所示，我们可以计算 g(m, n)。

对于长度为 m 和 n 的两个序列，对准的残基的数目 k 是介于 0 和 min {m, n}之间的。而且，对于每个这样的 k，有$\binom{m}{k}$种选择与 y 的残基对应的 x 的残基的方法，有$\binom{n}{k}$种选择与 x 的残基对应的 y 的残基的方法。所以 k 个残基一共有$\binom{m}{k}\binom{n}{k}$种比较，所以

$$g(m,n) = \sum_{k=0}^{\min\{m,n\}} \binom{m}{k}\binom{n}{k}$$

从下述的问题 6.1 的结论，下面有

$$g(m,n) = \binom{m+n}{n}$$

特别的，当 m=n 时，

$$g(n,n) = \binom{2n}{n}$$

这个数随着 n 快速增长，斯特林近似公式（B.4），和更加直接的公式（B.5），显示出

$$\binom{2n}{n} \sim \frac{2^{2n}}{\sqrt{\pi n}}$$

所以两个长度为 1000 的序列之间的全局比较的数量 c(1000, 1000)满足

$$c(1,000,1,000) \geq g(1,000,1,000) \cong \frac{2^{2,000}}{\sqrt{1,000\pi}} \cong 10^{600}$$

这说明为什么检验所有的比较是不合适的。这个结果激发了对于寻找能够有效计算最佳得分和比较方法的算法的研究，这种算法不需要检验所有可能。一种这样的算法是 Needleman–Wunsch 算法（1970），我们还会介绍由 Gotoh（1982）引入的这种方法的另一个版本。这些都是动态规划算法，我们用这些算法来介绍动态规划的基本概念。

输入由两个长度为 m 和 n 的序列组成，

$$\mathrm{x} = X_1X_2X_3 \ldots X_m \text{和} \mathrm{y} = Y_1Y_2Y_3 \ldots Y_n$$

他们的元素是 N 个符号组成的字母表（对于 DNA 或者 RNA 序列，N=4，对于蛋白质序列，N=20）。我们假设有一个替代矩阵 S 和一个线性惩罚项 d，输出包括 x 和 y 之间的比较的最高得分和相对应的全局比较。

普遍的方法是把大问题分成同样类型的小问题，然后利用小问题的答案来建立最终的答案：这是任何动态规划算法背后的基本思路。在这个问题中，我们首先对于小的 x 和 y 的子序列寻找最高得分的比较，然后用前面的结果寻找得分最高的全局比较。我们记$x_{1,i}$为 x 最开始的一段$X_1X_2X_3 \ldots X_i$，相似的，我们记$y_{1,j}$为 y 最开始的一段$Y_1Y_2Y_3 \ldots Y_j$。对于 i=1,2,…,m 和 j=1,2,…,n，我们记 B(i, j)为$x_{1,i}$和$y_{1,j}$序列之间的比较的最高得分。对于 i=1,2,…,m，我们记 B(i, 0)为$x_{1,i}$和长度为 i 的空序列比较的得分，所以 B(i, 0)=-id。相似的，对于 j=1,2,…,n，我们记 B(0, j)为$y_{1,j}$和长度为 j 的空序列比较的得分，所以 B(0, j)=-jd。最后，我们设初始值 B(0,0)=0。这些计算得出一个$(\mathrm{m}+1)\times(\mathrm{n}+1)$的矩阵 B。B 的最后一行和最后一列的元素，记作 B(m, n)，是我们两个序列 x 和 y 的最高得分，并且这是我们想要我们的算法得到的结果之一。

这个过程的本质是递归地填充矩阵 B 的元素。我们已经有 B 的一部分元素，在(0, 0), (i, 0), (0, j)，对于 i=1,2,…,m，j=1,2,…,n。现在，我们从左上到右下，注意到最高的$x_{1,i}$和$y_{1,j}$之间的得分比较可以终止在三种可能的方式之一，分别是

$$\begin{matrix} X_i \\ Y_j \end{matrix}, \quad \begin{matrix} X_i \\ - \end{matrix}, \quad \text{or} \quad \begin{matrix} - \\ Y_j \end{matrix}$$

第一种情况下，B(i, j)等于$x_{1,i-1}$和$y_{1,j-1}$之间的最高得分与额外的 s(i, j)，表示X_i和Y_j的匹配，的和；也就是说，B(i, j)= B(i-1, j-1)+ s(i, j)。在第二种情况下，B(i, j)等于$x_{1,i-1}$和$y_{1,j}$之间的最高得分与额外的-d，表示插入缺失标记与X_i的比较。相似的，在第三种情况下，B(i, j)= B(i, j-1)- d。这些都是可能的选项，从而 B(i, j)是三个之中最高的，也就是说，

$$B(i,j) = \max\{B(i-1,j-1)+s(i,j), B(i-1,j)-d, B(i,j-1)-d\}$$

这样我们就递归的把每一个元素填入到矩阵 B 中，而且决定了 B(m, n)的值，也就是要求的最大得分。这个算法的运行时间显然是 O(mn)。为了找到具

这个得分的比较，我们必须保持跟踪，在每一步递归之中，在每一次三选一之中。如果我们对于找到唯一比较感兴趣，我们选择其中一个设置线索。一旦获得了 B(m, n)，通过对于线索的追踪，我们可以对于最高分重新建立一个比较。现在我们可以通过一个例子来描述这个过程。

例子。让 x=gaatct, y=catt，那么 m=6 并且 n=4。用 6.4.1 的例子中同样的方法，B 在途中给出了，我们用箭头表示每个元素来自何处。一个对准的最好成绩是在底部右边的元素了，是-2。追溯沿大胆的箭头，我们得到最高的得分比较。

$$\begin{array}{cccccc} g & a & a & t & c & t \\ c & - & a & t & - & t \end{array}$$

通过选择不同的箭追踪程序我们可以得到以下的其他路线，这也是得分最高，

$$\begin{array}{cccccc} g & a & a & t & c & t \\ c & a & - & t & - & t \end{array} \quad \text{and} \quad \begin{array}{cccccc} g & a & a & t & c & t \\ - & c & a & t & - & t \end{array}$$

接下来我们考虑的 Needleman–Wunsch 算法的修改，可用于解决其他种类的双序列比对问题。

	$-$		c		a		t		t
$-$	0	$\nwarrow$	-2	$\nwarrow$	-4	$\nwarrow$	-6	$\nwarrow$	-8
g	-2	$\nwarrow$	-1	$\leftarrow$	-3	$\leftarrow$	-5	$\leftarrow$	-7
a	-4	$\nwarrow$	-3	$\nwarrow$	0	$\leftarrow$	-2	$\leftarrow$	-4
a	-6	$\nwarrow$	-5	$\nwarrow$	-2	$\leftarrow$	-1	$\leftarrow$	-3
t	-8	$\nwarrow$	-7	$\nwarrow$	-4	$\nwarrow$	-1	$\nwarrow$	0
c	-10	$\nwarrow$	-7	$\nwarrow$	-6	$\nwarrow$	-3	$\nwarrow$	-2
t	-12		-9		-8		-5		-2

Figure 6.1.

6.4.3 用线性间隙模型拟合一个序列

在本节中，我们解决以下问题：给定两个序列，一个较长的和较短的一个，找到子序列的较长的一个，可以与较短的序列最好地比较，其中间隙允许。此过程是相关的，当我们有兴趣在序列中指定一个指定的图案时。

让$\mathrm{x} = X_1X_2X_3 \ldots X_m$和$\mathrm{y} = Y_1Y_2Y_3 \ldots Y_n$满足$\mathrm{n} \geq \mathrm{m}$。对于$1 \leq \mathrm{k} \leq \mathrm{j} \leq \mathrm{n}$，记

$y_{k,j}$是 y 的子序列$Y_k Y_{k+1} \ldots Y_j$，对于两个序列 u 和 v，记 B(u, v)是 u 和 v 之间的最高得分全局比较的得分。我们的目标是找到

$$\max\{B(\boldsymbol{x}, \boldsymbol{y}_{k,j}) : 1 \leq k \leq j \leq n\}$$

对于每一个 k 和 J 的选择，Needleman–Wunsch 算法的运行时间，在给定 B(x, $y_{k,j}$)的值时，是 O(m(j-k))。所以如果我们对于所有的 k 和 j 使用这个算法，然后对于所有这些选择取最大值，那么总共运行时间是 O(mn^3)，因为对于 J 和 K，有$\binom{n}{2}$中可能。我们现在提出另一种具有更好的运行时间的方法，大约 O(mn)的运行时间。

对于$1 \leq i \leq m$和$1 \leq j \leq n$，让 F(i, j)是对于所有的 1 和 j 之间的 K 的 B($x_{1,i}$, $y_{k,j}$)最大得分。也就是说，对于所有可能得分最高得分之间的初始段比较的 x 知道x_i和在某个 k 开始结束在y_j的 y，我们以 F(i, j)为最大的分数。对于所有的 1 和 n 之间的 j，这个值是 F(m, j)的最大值。为了找到这个，我们初始化定义 F(i, 0)=-id，对于$1 \leq i \leq m$，而且 F(0, j)=0，对于$0 \leq j \leq n$，因为删除 y 开头应明确无惩罚。然后我们递归地填充矩阵 F

$$F(i,j) = \max\{F(i-1,j-1) + s(i,j), F(i,j-1) - d, F(i-1,j) - d\}$$

这个公式背后的推理类似于后面。请注意，可能会有一个以上的值 J 给最大得分。为了恢复得分最高的比对 x 与子序列 y 我们可以保持追踪，就像在 Needleman–Wunsch 算法中。

6.4.4 局部线性模型的带隙比较

另一个有趣的比较问题是找到，给定两个序列，其中各自的子序列具有最高的得分比较（允许的间隙）。这被称为局部比较问题，它是合适的，当一个是在两个序列中寻求共同的模式/域。

在下面的假设中，我们使用的评分方案是这样的预期（或平均）的随机排列的分数是负的。如果这个假设不成立，那么长的比赛得分高的子序列之间可能只是因为他们的长度，使两长无关的子序列，可以给一个得分最高的对齐。显然，我们不希望这种情况发生。

对于$1 \leq i \leq m$和$1 \leq j \leq n$，我们定义 L(i, j)是 0 和所有可能的子序列 x 直到x_i和一个y直到y_j结束之间的比较的最高得分的最大值。也就是说，

$$L(i,j) = \max\{0, B(\boldsymbol{x}_{h,i}, \boldsymbol{y}_{k,j}) : 1 \leq h \leq i, 1 \leq k \leq j\}$$

当最大的 B($x_{1,i}$, $y_{k,j}$)是负数的时候我们想要 L(i, j)=0，是因为这种方法是明智的，总是删除第一部分的比较方式，如果这部分有一个负的分数，因为它只会降低整体得分的比较方式。然后， 0 和上式的最大值是 L(i, j)的最大值，

对于$1 \le i \le m$和$1 \le j \le n$。为了确定最大值我们再一次利用动态规划，通过初始化 L(i, 0)=0=L(0, j)，和计算，

$$L(i,j) = \max\{0, L(i-1,j-1) + s(i,j), L(i-1,j) - d, L(i,j-1) - d\}$$

然后，我们计算所有的 L(i, j)的最大值，对于$1 \le i \le m$和$1 \le j \le n$。在以前的最大化程序，有可能是一个以上的最高得分本地比较。找到一个得分最高的比较，我们按照前面描述的追踪程序。然而，对于这个算法，我们遇到一个 0 时我们停止这个过程。

图 6.2 显示了 L(i, j)矩阵在局部比较两个序列的长度为 7 和 10 的一个例子。在这个例子中，最佳局部比较的分数是 28，并且有唯一的子序列比较给出这个分数，箭头显示了这个事实，如下，

$$\begin{matrix} X_2 & X_3 & - & X_4 & X_5 \\ Y_5 & Y_6 & Y_7 & Y_8 & Y_9 \end{matrix}$$

		Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	Y_9	Y_{10}
	0	0	0	0	0	0	0	0	0	0	0
X_1	0	0	0	0	0	0	0	0	0	0	0
X_2	0	0	0	5	0	5	0	0	0	0	0
X_3	0	0	0	0	2	0	20	12	4	0	0
X_4	0	10	2	0	0	0	12	18	22	14	6
X_5	0	2	16	8	0	0	4	10	18	28	20
X_6	0	0	8	21	13	5	0	4	10	20	27
X_7	0	0	6	13	18	12	4	0	4	16	26

Figure 6.2.

6.4.5 其他间隙模型

上面讨论的算法有许多变种和扩展。例如，虽然上面使用的线性间隙模型在其简单性上有吸引力，但它通常不适合于生物序列。因为开放的间隙比它延长。因此，它往往不合适额外的惩罚间隙和第一个一样多的步骤，并使用一个更复杂的间隙成本，这意味着递推关系需要调整。例如，我们现在必须区分$x_{1,i}$与$y_{1,j}$，最后我对准一个插入缺失标记。这样的一个线形评分也将取决于 x 前X_i对准插入缺失标记的许多符号是如何的。假设在这样的一个线形的最后符号前

面的 x 我不对齐插入缺失（因此对准Y_j）是X_k。然后 i-k 符号从X_{k+1}到X_i对准插入缺失标记和一个得分最高的比较得分 B（k，j）+δ（i−k）。类似的推理必须适用于那些$x_{1,i}$和$y_{1,j}$之间的比较，Y_j对齐插入缺失标记结尾。所以（6.11）一定会下面的等式替代

$$\begin{aligned} B(i,j) = \max\{&B(i-1,j-1)+s(i,j),\\ &B(k,j)+\delta(i-k) : k=0,1,\ldots,i-1,\\ &B(i,k)+\delta(j-k) : k=0,1,\ldots,j-1\} \end{aligned}$$

初始值是 B(0，0)=0，B(i，0)= δ(i)，$1 \le i \le m$，B(0，j)= δ(j)，$1 \le j \le n$

因此，在一般情况下，找到两个序列的长度 m 和 n 之间的最高得分比较需要 O（$m^2n + mn^2$）操作，而不是 O（mn）的操作需要，如果间隙惩罚模型是线性的。这是因为对于 B 的每个单元格，我们现在需要考虑 i+ j + 1 以前的元素，而不是只有三个。

如果，然而，δ（l）满足一定条件，有算法，采取 O（mn）操作。这一个简单的例子是一个仿射差距模型，其中 δ（l）=−d−（l−1）e，一些（非负）d 和 e。d 被称为开间隙惩罚，并被称为差距扩大的惩罚。通常，e 被设置为小于 d。因此所有的差距步骤以外的第一个具有相同的成本，但他们每一个被处罚小于第一。现在，我们描述了一个动态规划实现的全局比对算法，这种情况下，其运行时间也是 O（mn）。

不是只用一个矩阵 B，该算法使用三个矩阵。让 x 和 y 成为我们想要比较的序列。我们使用相同的符号上面的$x_{1,i}$和$y_{1,j}$。$1 \le i \le m$和$1 \le j \le n$， S（i，j）表示$x_{1,i}$和$y_{1,j}$之间的最高得分的比较，给定条件在X_i和Y_j比较时结束。我们记$I_x(i,j)$为$x_{1,i}$和$y_{1,j}$之间的比较的最高得分，给定条件X_i和插入缺失标记比较时结束。最后，我们用$I_y(i,j)$为 $x_{1,i}$和$y_{1,j}$之间的比较的最高得分，给定条件在插入缺失标记和Y_j比较时结束。然后，如果我们假设删除不会直接跟随插入，对于$1 \le i \le m$和$1 \le j \le n$，我们有

$$\begin{aligned} &S(i,j)\\ &= \max\{S(i-1,j-1)+s(i,j), I_x(i-1,j-1)+s(i,j), I_y(i-1,j-1)+s(i,j)\} \end{aligned}$$

此时

$$I_x(i,j) = \max\{S(i-1,j)-d, I_x(i-1,j)-e\}$$

而且

$$I_y(i,j) = \max\{S(i,j-1)-d, I_y(i,j-1)-e\}$$

这些递推关系允许我们填充矩阵 S，I_x和I_y，一旦我们初始化$S(0,0) =$

$\mathrm{I}_x(0,0) = \mathrm{I}_y(0,0) = 0$，$\mathrm{S}(0,j) = \mathrm{I}_x(0,j) = -\mathrm{d} - (\mathrm{j}-1)\mathrm{e}$，$\mathrm{S}(i,0) = \mathrm{I}_y(i,0) = -\mathrm{d} - (\mathrm{i}-1)\mathrm{e}$，对于$1 \le \mathrm{i} \le \mathrm{m}$和$1 \le \mathrm{j} \le \mathrm{n}$。之后最高得分的比较的得分就由下面给出

$$\max\{S(m,n), I_x(m,n), I_y(m,n)\}$$

6.4.6 动态规划比对算法的局限性

上面所讨论的所有算法根据给定的评分方案得到准确的最高分数。然而，当一个人必须处理非常长的序列，如会发生，如果希望将给定的序列与每个序列中的一个大的数据库中的每个序列，O（mn）的时间复杂度可能不够好，以执行所需的搜索在一个可接受的时间量。因此，各种各样的算法已经被开发来克服这个困难，BLAST 是其中之一。这些算法使用启发式技术，以限制搜索的一个部分之间的可能序列的两个序列的方式，试图不错过的高得分比较。权衡的是，一个可能不一定找到最好的得分。BLAST 的各个方面将在第 10 章广泛讨论。

另一个考虑的是空间，因为内存的使用也可以是一个限制因素，在动态编程。在 Needleman–Wunsch 算法，例如，一个需要存储（m + 1）×（n + 1）矩阵 B.如果仅关心最好的得分值，而不想找一个得分最高的对齐，则不需要存储所有的细胞，由于 B（i，j）的值只取决于参赛作品最多的一回。因此，人们可以扔掉矩阵的行，是进一步回来。然而，通常人们想找到一个最高的得分对齐，不仅它的得分。有方法，允许一个这样做，在 O（m + n）空间，而不是 O（mn），这不超过一倍的时间，所以时间复杂度保持在 O（mn）。更多详情请参见第 2.6 节在杜斌等人（1998）或 9.7 节（1995）沃特曼。

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	−1	−2	−2	0	−1	−1	0	−2	−1	−1	−1	−1	−2	−1	1	0	−3	−2	0
R	−1	5	0	−2	−3	1	0	−2	0	−3	−2	2	−1	−3	−2	−1	−1	−3	−2	−3
N	−2	0	6	1	−3	0	0	0	1	−3	−3	0	−2	−3	−2	1	0	−4	−2	−3
D	−2	−2	1	6	−3	0	2	−1	−1	−3	−4	−1	−3	−3	−1	0	−1	−4	−3	−3
C	0	−3	−3	−3	9	−3	−4	−3	−3	−1	−1	−3	−1	−2	−3	−1	−1	−2	−2	−1
Q	−1	1	0	0	−3	5	2	−2	0	−3	−2	1	0	−3	−1	0	−1	−2	−1	−2
E	−1	0	0	2	−4	2	5	−2	0	−3	−3	1	−2	−3	−1	0	−1	−3	−2	−2
G	0	−2	0	−1	−3	−2	−2	6	−2	−4	−4	−2	−3	−3	−2	0	−2	−2	−3	−3
H	−2	0	1	−1	−3	0	0	−2	8	−3	−3	1	−2	−1	−2	−1	−2	−2	2	−3
I	−1	−3	−3	−3	−1	−3	−3	−4	−3	4	2	−3	1	0	−3	−2	−1	−3	−1	3
L	−1	−2	−3	−4	−1	−2	−3	−4	−3	2	4	−2	2	0	−3	−2	−1	−2	−1	1
K	−1	2	0	−1	−3	1	1	−2	−1	−3	−2	5	−1	−3	−1	0	−1	−3	−2	−2
M	−1	−1	−2	−3	−1	0	−2	−3	−2	1	2	−1	5	0	−2	−1	−1	−1	−1	1
F	−2	−3	−3	−3	−2	−3	−3	−3	−1	0	0	−3	0	6	−4	−2	−2	1	3	−1
P	−1	−2	−2	−1	−3	−1	−1	−2	−2	−3	−3	−1	−2	−4	7	−1	−1	−4	−3	−2
S	1	−1	1	0	−1	0	0	0	−1	−2	−2	0	−1	−2	−1	4	1	−3	−2	−2
T	0	−1	0	−1	−1	−1	−1	−2	−2	−1	−1	−1	−1	−2	−1	1	5	−2	−2	0
W	−3	−3	−4	−4	−2	−2	−3	−2	−2	−3	−2	−3	−1	1	−4	−3	−2	11	2	−3
Y	−2	−2	−2	−3	−2	−1	−2	−3	2	−1	−1	−2	−1	3	−3	−3	−2	2	7	−1
V	0	−3	−3	−3	−1	−2	−2	−3	−3	3	1	−2	1	−1	−2	−2	0	−3	−1	4

6.5 蛋白质序列与替代矩阵

6.5.1 引言

在 DNA 序列的研究中，简单的评分方法通常是有效的。然而，对于蛋白质序列来说，一些替换比其他的更有可能。任何比对算法的性能提高时，占这种差异。在所有的情况下，我们认为，更高的分数将代表更可能的替代。

有两种常用的方法来寻找替代矩阵。一个导致 PAM（接受点突变）矩阵族，和 BLOSUM 其他（块替代矩阵）的家庭。表 6.7 给出了一个典型的 BLOSUM 替代矩阵的一个实例（称为 BLOSUM62 矩阵）。在本节中，我们将讨论如何生成这些替换矩阵。

WWYIR	CASILRKIYIYGPV	GVSRLRTAYGGRK	NRG
WFYVR	CASILRHLYHRSPA	GVGSITKIYGGRK	RNG
WYYVR	AAAVARHIYLRKTV	GVGRLRKVHGSTK	NRG
WYFIR	AASICRHLYIRSPA	GIGSFEKIYGGRR	RRG
WYYTR	AASIARKIYLRQGI	GVGGFQKIYGGRQ	RNG
WFYKR	AASVARHIYMRKQV	GVGKLNKLYGGAK	SRG
WFYKR	AASVARHIYMRKQV	GVGKLNKLYGGAK	SRG
WYYVR	TASIARRLYVRSPT	GVDALRLVYGGSK	RRG
WYYVR	TASVARRLYIRSPT	GVGALRRVYGGNK	RRG
WFYTR	AASTARHLYLRGGA	GVGSMTKIYGGRQ	RNG
WFYTR	AASTARHLYLRGGA	GVGSMTKIYGGRQ	RNG
WWYVR	AAALLRRVYIDGPV	GVNSLRTHYGGKK	DRG

Table 6.8. A set of four blocks from the Blocks database

任何创建氨基酸替换得分矩阵的尝试都必须从可信任的一组数据开始。然后使用“可信”数据来确定哪些替换或多或少。矩阵，然后来自这些数据，使用（如我们将看到）方面的统计假设检验理论。

历史上，PAM 矩阵首先提出（1978），但由于 BLOSUM 矩阵的推导是比 PAM 矩阵比较简单，我们先考虑 BLOSUM 矩阵。