基于置换距离度量的蛋白质多序列比对算法性能评估*

高 峰[1]，李防震[1]，王 珺[2]，董骝焕[2]
（1. 山东经济学院计算机科学与技术学院//山东省数字媒体技术重点实验室，山东 济南 250014；
2. 中国科学院-马普学会计算生物学伙伴研究所，上海 200031）

摘 要：蛋白质多序列比对是一种重要的生物信息学工具，在生物的进化分析以及蛋白质的结构预测方面有着重要的应用。各种比对算法在这个领域都取得了很大的成功，但是每种算法都有其固有的缺陷。提出置换距离法，对当前流行的几种蛋白质多序列比对算法进行对比评价。由于置换距离法仅关注于不同蛋白质间进化距离的相对次序，而不考虑这些进化距离之间的细微差异，因而得到的评价结论更具有鲁棒性。另外，采用最长公共子序法度量置换距离可以比较准确的反映不同置换之间的差异性。基于该算法，对 Dialign，Tcoffee，ClustalW 和 Muscle 多序列比对算法进行了性能评估。

关键词：多序列比对；置换距离；最长公共子序；进化距离
中图分类号：Q7 **文献标志码**：A **文章编号**：0529－6579（2011）02－0087－06

Performance Assessment of Protein Multiple Sequence Alignment Algorithms Based on Permutation Distance Measurement

GAO Feng[1], *LI Fangzhen*[1], *WANG Jun*[2], *DONG Liuhuan*[2]
(1. School of Computer Science and Technology//Shandong Provincial Key Laboratory of Digital Media Technology, Shandong Economic University, Jinan 250014, China;
2. CAS-MPG Partner Institute for Computational Biology, CAS at Shanghai, Shanghai 200031, China)

Abstract: Protein multiple sequence alignment is an important bioinformatics tools. It has important applications in biological evolution analysis and protein structure prediction. A variety of alignment algorithms in this field have achieved great success. However, each algorithm has its own inherent deficiencies. In this paper, permutation distance is proposed to evaluate several protein multiple sequence alignment algorithms that are widely used currently. As the permutation distance method only concerns the relative order of different protein evolutionary distances, without taking into account the slight difference between the evolutionary distances, it can get more robust evaluations. In addition, the longest common subsequence method can well define the distances between different permutations. Using these methods, we compared and assessed Dialign, Tcoffee, ClustalW and Muscle.

Key words: multiple sequence alignment; permutation distance; longest common subsequence; evolutionary distance

蛋白质多序列比对是蛋白质组学研究中的一项 重要而具有基础性的研究课题，在整个生物信息学

* 收稿日期：2010－04－14
基金项目：国家自然科学基金资助项目（30600121，10701070）；山东省优秀中青年科学家科研奖励基金资助项目（2007BS09002）
作者简介：高峰（1988年生），男，硕士生；通讯作者：李防震；董骝焕；E-mail：fzli1976@gmail.com；dlh@picb.ac.cn

和系统生物学研究中亦是至关重要的一步。比对算法的好坏直接影响到其它生物信息学研究结果的准确性，低质量的比对算法甚至会给相关研究带来导致错误结论的危险。因此，研究对蛋白质多序列比对算法的评估方法就显得尤为重要。

目前，蛋白质多序列比对算法的打分方法有很多[1-2]，比如：WSP（weighted sum-of-pairs），最大似然，最小熵等等。现在最流行的打分方法是WSP，但是在使用CluatalW，Tcoffee等方法的比对测试中，63%的比对都比正确的比对有较高的WSP值，也就是说WSP分数与比对质量之间的对应是比较差的。后来Thompson等[3]引入两种计分来比较多序列比对：列分数和SPS（sum-of-pairs score）。

列分数是计算参考比对与待检验比对序列中相同列的个数。这种衡量反映了程序正确比对所有序列的能力，所以任何一条序列的比对错误既可导致比对得分为零。SPS计算待检验比对与参考比对间相同的残基对的比例。后来，Karplus等[4]进一步修改了这种打分方法，给它赋予了权值。Lassmann等则对SPS方法做了另外的修改，提出了交叠分数（overlap score）法。该方法把参考比对中的所有残基对及待检验比对中的所有残基对分别存入两个集合中，然后计算这两个集合的交与并的比值。所有这些评估算法都基于序列比对结果的局部性质，因而都存在着易受数据噪声影响的缺陷。

通常情况下，研究者得到的DNA和氨基酸序列数据总是含有噪声的。噪声主要来源于测序过程，目前的DNA和氨基酸测序技术还难以达到100%的准确度；另外，测序后的序列数据在存储和传输过程中也有可能引入误差，例如，由于操作人员的粗心导致的字母输入错误。因此，对序列数据的处理程序，需要有一定的容错能力和抗噪声能力。

本文提出了一种蛋白质多序列比对算法的性能评估新方法，即置换距离法。由于置换距离法仅关注于不同蛋白质-蛋白质进化距离之间的相对次序，而不考虑这些进化距离之间的细微差异，因而具有更强的鲁棒性。它能够克服氨基酸序列数据中噪声的影响，对多种蛋白质序列比对算法做出客观公正的评价。

另外用最长公共子序法度量置换距离，能够比较准确的反映出置换距离之间的差异性。基于该算法，我们对Dialign，Tcoffee，ClustalW和Muscle多序列比对算法进行了比较评估。

1 材料与方法

1.1 数据源

1.1.1 BALIBASE数据库数据 BALIBASE（Benchmark Alignment Database）是针对多序列比对问题而设计的数据资源，它提供的所有多序列比对都是来源于三维结构的叠加，是目前用作多序列比对结果比较的标准平台，我们应用BALIBASE作为我们的数据源。BALIBASE包含了八个等级的参考系列，由于其中的RV6-RV8还不成熟，所以我们只应用了RV1-RV5共142组多序列进行测试。

1.1.2 ROSE软件生成的数据 ROSE软件可以生成蛋白质序列进化的概率模型。通过进化树，可以经过插入、删除和替换，从同一祖先生成具有一定相似性的序列。在人工的进化过程中，我们可以指定序列长度，进化距离，进化祖先等多项参数，来产生具有一定生物意义的序列。ROSE生成的序列适合于多序列比对，也广泛应用于生物进化关系的预测。我们将其作为另一个数据源。

1.2 蛋白质多序列比对算法介绍

本文选取Dialign，Tcoffee，ClustalW和Muscle这四种目前比较流行的蛋白质多序列比对算法进行性能评估，从而验证我们算法的有效性。

Dialign是最近几年才发展起来的一个多序列比对算法[5]。它注重于寻找多条序列中相似的区域，是一种基于段的渐进比对算法。它对于非全局相关，只是局部相似的序列非常有效。在序列比对软件当中，它的准确性是非常高的，但是运行时间很长。而且输出格式不够灵活，使得比对比较耗时，并需要进一步处理比对后的结果。

Tcoffee是一个多序列比对包[6]。给定一组蛋白质或者DNA序列，它便可以生成多序列比对。Tcoffee可以连接多序列或者一对序列，也可以连接全局比对序列或者局部比对序列到一条序列中。不论数据源来自哪里，它都能通过剩余的部分来判断得出每个位置与新序列联系的紧密性。这种紧密型通常是序列比对准确性的指示器。另外，它也提供多种格式的输出结果。

Clustal W可以进行蛋白质与核酸的多序列比较[7]，分析不同序列之间的相似性关系，还可以绘制进化树。由于其灵活的输入输出格式、方便的参数设定和选择以及良好的可移植性，使得ClustalW在蛋白质与核酸的序列分析中得到了广泛应用。它可以用来发现特征序列，进行蛋白分类，证明序列间的同源性，帮助预测新序列二级结构与三

级结构，确定 PCR 引物，以及用于分子进化分析。

Muscle 在比对过程中采取了两次迭代过程来提高比对的精度[8]。对于一对序列 Muscle 提供了两种距离测量方法：一种针对未比对序列的 *kmer* 距离和一种针对已比对序列的 Kimura 距离。Muscle 算法会首先进行一次渐进式比对，然后利用迭代精细法对多序列比对结果进行进一步的优化。Muscle 采用的是基于进化树的分组迭代法。

1.3　置换距离法介绍

设 Ω 为 n 个自然数构成的集合，即 $\Omega = \{1,2,3,\cdots,n\}$。集合 Ω 到其自身上的一一映射叫做 Ω 上的一个置换[9]。集合 Ω 上的所有置换构成置换群，共有 $n!$ 个元素。

例 1：集合 $\Omega = \{1,2,3\}$ 共有 $3! = 6$ 个置换，分别为 $\{1,2,3\}$，$\{1,3,2\}$，$\{2,1,3\}$，$\{2,3,1\}$，$\{3,1,2\}$，$\{3,2,1\}$。其中 $\{1,2,3\}$ 称为恒等置换。

本文中置换的生成过程如下：从数据库中提取一组序列数据，设含有 m 个氨基酸序列。由于数据库自身已给出各组序列的参考比对，我们可基于这 m 个序列的参考比对生成任意一对序列之间的进化距离。显然，m 条序列共有 $n = m(m-1)/2$ 个两两组合（称为二元组），构成一个 n 维的距离向量。我们把这 n 个距离由小到大进行排序，并分别给予编号 $1 \sim n$，得到 n 个自然数（构成集合 Ω）的恒等置换 $\pi_0 = \{1,2,\ldots,n\}$，同时得到这 m 条序列的 n 维二元组集合到自然数集合 Ω 上的一一映射。

选取一个待考察的比对算法，如 Dialign，对这组序列重新进行比对。然后按与前面相同的过程，基于此比对结果生成 n 维的进化距离向量，并把这 n 个距离由小到大进行排序。需要注意的是，这里得到的 n 个二元组的排列顺序与前面是不一样的。我们把这 n 个二元组按与前面相同的映射关系映射到集合 Ω 中，即得到集合 Ω 的对应于该比对算法的置换 $\pi_1 = \{\pi(1),\pi(2),\ldots,\pi(n)\}$。

因此，对于每个比对算法，我们都可得到 n 个自然数集合 Ω 关于这个算法的置换。

例 2：假设从数据库中提取一组氨基酸序列，含有 3 条序列 a，b 和 c。3 条序列共有 $3 \times (3-1)/2 = 3$ 个两两组合，其进化距离分别记为 $D(a,b)$，$D(b,c)$，$D(c,a)$。

按照数据库给出的参考比对得到的进化距离排序设为：$D(a,b) < D(c,a) < D(b,c)$。则通过从集合 $\{(a,b),(b,c),(c,a)\}$ 到集合 $\Omega = \{1,2,3\}$ 的一一映射：$(a,b) \to 1$，$(a,c) \to 2$，$(b,c) \to 3$，得到集合 Ω 的恒等置换 $\{1,2,3\}$。

接着使用某个比对算法得到的距离排序设为：$D(c,a) < D(a,b) < D(b,c)$。按上面的映射关系，即得到集合 Ω 的置换 $\{2,1,3\}$。

这样，通过进一步测量不同置换之间距离的大小，就提供了一种评价各种多序列比对算法性能的有效方法。采用该方法，数据的细节差异性被排序后的数字所隐藏了，我们关心的已不再是原始的数据，而是置换后的数字串。因而，置换距离法仅关注于不同蛋白质-蛋白质进化距离之间的相对次序，而不考虑这些进化距离之间的细微差异，因而具有更强的鲁棒性。

1.4　置换距离的最长公共子序度量

关于置换的距离问题，人们已发展了多种度量方法，我们采用最长公共子序（Longest Common Subsequence，LCS）法度量置换之间的距离。子序即子序列，是与原字符串具有相同出现顺序的字符的子集。本文所处理的字符串是由自然数组成的数字串。

例 3：[3,5,7] 是 [1,2,3,4,5,6,7] 的子序。

两条串所共有的最长子序的长度可用来衡量两条串的相似性[10]。设原串长度都为 n，则当两条串完全相同时，其最长公共子序长度达到最大值 n；当两条串字符顺序完全相反时，其最长公共子序长度达到最小值 1。我们定义最长公共子序距离为 n 减去最长公共子序的长度，用 d_{lcs} 表示，那么 d_{lcs} 的范围就是：$[0,n-1]$。

规范化的最长公共子序距离为最长公共子序与 n-1 的比值，即

$$\hat{d}_{lcs} = \frac{d_{lcs}}{n-1} \tag{1}$$

最长公共子序可以通过一种动态规划方法实现[11]，算法复杂度为 $O(n^2)$。首先介绍 LCS 相似度[12]：字符串 a 和 b 的 LCS 相似度是 a 和 b 间的最大相同子序的长度，记为 $\mathrm{LCS}(a,b)$。显然 $\mathrm{LCS}(a,b)$ 越大，a，b 越相似。$\mathrm{LCS}(a,b)$ 的动态规划计算公式为

$$\mathrm{LCS}(i,j) = \begin{cases} 0 & i = 0 \text{ 或 } j = 0 \\ 1 + \mathrm{LCS}(i-1,j-1) & a_i = b_i \\ \mathrm{MAX}(\mathrm{LCS}(i-1,j),\mathrm{LCS}(i,j-1)) & a_i \neq b_i \end{cases} \tag{2}$$

其中 i 和 j 分别为字符串 a 和 b 中字符位置的标记，$LCS(i,j)$ 表示两条子串 $[a_1,\ldots,a_i]$ 和 $[b_1,\ldots b_j]$ 的最长公共子序，则 $LCS(a,b)$ 就是字符串 a 和 b 的最长公共子序。然后，基于 $LCS(a,b)$ 可计算得两条串 a，b 间的最长公共子序距离 d_{lcs}，为

$$d_{lcs}(a,b) = n - LCS(a,b) \quad (3)$$

代入公式（1），即可求得规范化的最长公共子序距离 $\hat{d}_{lcs}(a,b)$。

使用最长公共子序的方法来比较两个置换，比较灵敏准确。由于它面对的处理数据是数字串，而传统的比较方法面对的处理数据是比对过的氨基酸序列。相对来说，最长公共子序所对应的数据量比较小，因而可以达到比较快的速度。

1.5 处理过程

1.5.1 *序列比对* BALIBASE 数据库数据和 ROSE 生成的数据均是 fasta 格式的序列，这是蛋白质序列信息的通用存储格式。从 BALIBASE 数据库的物种类别中，我们得到了 142 组数据。使用 ROSE 软件，我们又生成了 3 721 组序列，指定序列长度范围：[50,650]，步长为 10；进化距离范围：[0,300]，步长为 5。所有序列数据都已给出了参考比对。对于两部分数据，我们分别使用 Dialign，Tcoffee，ClustalW 和 Muscle 这 4 种多序列比对算法比对各组序列。从而每组序列数据都可以生成 5 种比对结果（包括参考比对）。比对后的序列，可以显示出一些相似的区域以及差异较大的区域，反映了序列在进化上的保守区和非保守区。

1.5.2 *计算距离矩阵* 使用 phylip 3.69 程序包中的 protdist.exe 程序可以基于比对后的每组序列分别生成进化距离矩阵[13]。protdist.exe 程序提供了多达 5 种基于比对序列计算进化距离的模型，本文采用其缺省设置 Jones-Taylor-Thornton（JTT）模型[14]。对于含 m 条序列的分组（各组的 m 值可能是不同的），得到的进化距离矩阵是 $m \times m$ 维的，每个元素对应一个序列对。显然，这个矩阵是对称的，且对角线元素为 0，因而只有 $m(m\text{-}1)/2$ 个独立的非零元素，对应 m 条序列的 $m(m\text{-}1)/2$ 个两两组合。这样，每组序列都可以利用 protdist.exe 程序产生 5 个距离矩阵，分别为参考比对距离矩阵、Dialign 距离矩阵、Tcoffee 距离矩阵、ClustalW 距离矩阵和 Muscle 距离矩阵。

1.5.3 *排序并生成置换* 每个 $m \times m$ 维的进化距离矩阵可提取出一个 $n = m(m\text{-}1)/2$ 维的二元组向量，其元素为序列的两两组合。按前面所介绍的过程对每个二元组向量进行排序和映射，即得到每组序列数据在自然数集合 $\Omega = \{1,2,\ldots,n\}$ 的 5 种置换，分别记为：

参考比对产生的置换 $\pi_0 = \{1,2,\ldots,n\}$，为恒等置换；

Dialign 产生的置换 $\pi_d = \{\pi_d(1),\pi_d(2),\ldots,\pi_d(n)\}$；

Tcoffee 产生的置换 $\pi_t = \{\pi_t(1),\pi_t(2),\ldots,\pi_t(n)\}$；

ClustalW 产生的置换 $\pi_c = \{\pi_c(1),\pi_c(2),\ldots,\pi_c(n)\}$；

Muscle 产生的置换 $\pi_m = \{\pi_m(1),\pi_m(2),\ldots,\pi_m(n)\}$。

1.5.4 *计算置换距离* 利用公式（1）分别计算四种置换（π_d,π_t,π_c,π_m）与恒等置换 π_0 之间的距离，称为置换距离，即 $\hat{d}_{lcs}(\pi_d,\pi_0)$，$\hat{d}_{lcs}(\pi_t,\pi_0)$，$\hat{d}_{lcs}(\pi_c,\pi_0)$，$\hat{d}_{lcs}(\pi_m,\pi_0)$。所得的置换距离可作为评价各比对算法性能好坏的依据。比对算法的置换距离越小，则表明该算法性能较好，反之则表明算法性能较差。通过大数据量的统计计算，可综合评价各种算法的优劣。

2 试验结果

2.1 基于 BALIBASE 数据库的评价结果

图 1 表示在 BALIBASE 数据库的六类数据中，各比对算法表现出最优的比例。可以看出在所有 6 种类别的数据中，Tcoffee 方法的表现都是最优的，Dialign 次之，且这两个算法的获胜比例远远超过 ClustalW 和 Muscle。在该图中，Dialign 和 Tcoffee 占了绝对优势，Muscle 和 ClustalW 的表现只局限于很小的数据范围。

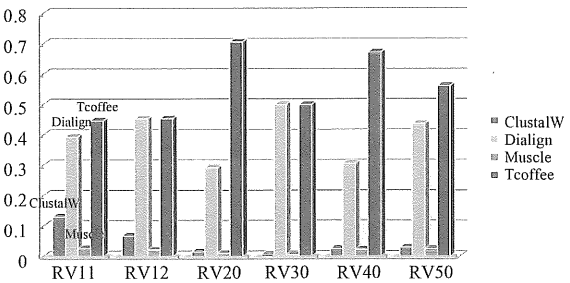

图 1 BALIBASE 数据库数据处理的最优比例图

Fig. 1 Percentage of 'wins' for BALIBASE database

图 2 为 4 种算法在各类别数据上的置换距离数值统计图，注意图中较小的数值对应着算法较好的性能表现。可以看出，Tcoffee 和 Dialign 的数值相

近，前者略优于后者，Muscle 和 ClustalW 的表现差不多，Tcoffee 和 Dialign 的数值显著小于 Muscle 和 ClustalW，这些结论与图 1 也都是一致的。这些结果与其它评价方法所得结果是一致的[8]。

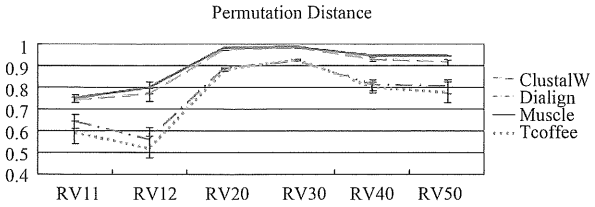

图 2　BALIBASE 数据库数据处理的各算法置换距离图

Fig. 2　Permutation distances of four algorithms based on BALIBASE database

2.2　基于 ROSE 数据的评价结果

我们使用 ROSE 软件生成的数据进行实验，结果见图 3 和图 4。图 3 为获胜算法的分布示意图，其中横坐标为 ROSE 软件的序列长度参数，纵坐标为进化距离参数。由图 3 可以看出，在所有区间 Dialign 方法的最优比例都是最高的，Muscle，ClustalW 和 Tcoffee 的比例都比较小。

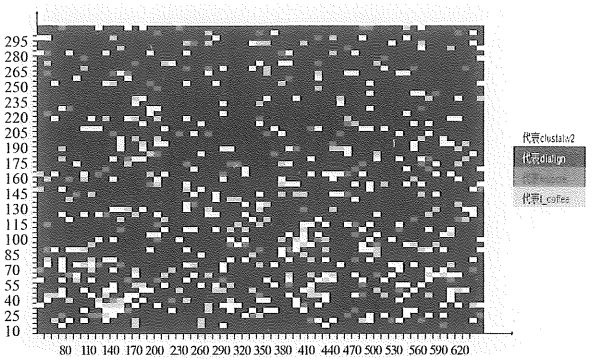

图 3　基于 ROSE 数据的各算法获胜分布示意图

Fig. 3　Distribution of 'wins' for ROSE database

图 4 为基于 ROSE 数据得到的置换距离统计图。可以看出 Dialign 算法表现最好，显著超过其它几种算法。ClustalW 算法、Muscle 算法与 Tcoffee 算法表现差不多，它们的置换距离数值都处于较高的水平，这些结论与图 3 也是一致的。

3 讨　论

在蛋白质多序列比对算法的性能评估方法的研究中，我们提出了置换距离法。置换距离法能够克服氨基酸序列数据中噪声的影响，对多种蛋白质多序列比对算法进行了客观公正的评价；用最长公共子序法度量置换距离比较准确的反映出了不同置换之间的差异性。计算结果与其它序列比对评价算法所得结果是一致的[8]，另外，计算结果表明各算法的性能表现与处理的实际数据有关，不同算法具有不同的数据偏好。

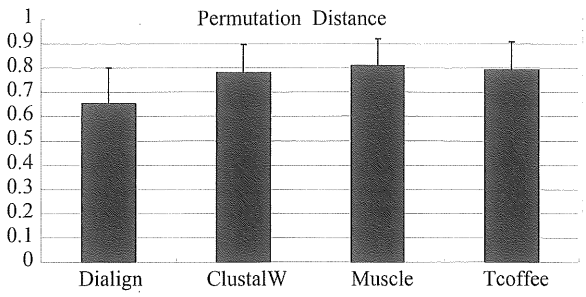

图 4　ROSE 生成数据的各算法置换距离统计图

Fig. 4　Permutation distances of four algorithms based on ROSE database

我们分别用 2 个数据库评价 4 种算法，得到的结论不尽一致。综合来看，在 BALIBASE 数据库上 Dialign 和 Tcoffee 表现较好，但 Dialign 又稍逊于 Tcoffee；在 ROSE 数据上 Dialign 占绝对优势，其它 3 种算法表现差不多都处于较低的数据范围内。但需要指出的是，由图 2 和图 4 可以看出，实验结果具有显著的标准差。考虑这个因素之后，我们认为，总的来看，在这 4 个多序列比对算法中，Dialign 表现出最佳的性能。

分析 Dialign 在 BALIBASE 上的表现稍差的原因，可能与 Dialign 属于概率模型有关。BALIBASE 数据来自真实的生物学实验，而 ROSE 数据是根据概率模型人工产生的。因而，同样基于概率模型的 Dialign 在 ROSE 数据中会表现更好一些。

本文采用了最长公共子序法来度量置换之间的距离，事实上，目前存在着多种置换距离度量方法。采用其他的置换距离度量方法，如编辑距离，或许也可以达到较好的效果。目前人们发展了大量蛋白质序列比对算法，本文仅选取了其中 4 种进行评估，计算结果初步证明了我们所采用的置换距离法的可行性。我们希望将来用置换距离法对目前存在着的各种序列比对算法做一个全面系统的性能评价研究，以考察不同比对算法所适用的数据范围，从而对其他研究者在选择序列比对算法时提供参考。另外，本文仅讨论了蛋白质多序列比对算法的评估问题，但置换距离法显然也适用于评估各种 DNA 多序列比对算法，表明该方法具有广泛的适用性。

参考文献：

[1] LASSMANN T, SONNHAMMER E L L. Quality assessment of multiple alignment programs[J]. FEBS, 2002, 529:126 - 130.

[2] OSAMU G. Multiple sequence alignment: algorithm and applications[J]. Adv Biophys, 1999,36: 159 - 206.

[3] THOMPSON J D, PLEWNIAK F, POCH O. A comprehensive comparison of multiple sequence alignment programs[J]. Nucleic Acids Res, 1999,27: 2682 - 2690.

[4] KARPLUS K, HU B. Evaluation of protein multiple alignments by SAM-T99 using the BALIBASE multiple alignment test set[J]. Bioinformatics, 2001, 17: 713 - 720.

[5] MORGENSTERN B. DIALIGN: multiple DNA and protein sequence alignment at BiBiServ[J]. Nucl Acids Res, 2004,32: 33 - 36.

[6] CEDRIC N, DESMOND G H, JAAP H. T-Coffee: a novel method for fast and accurate multiple sequence alignment[J]. J Mol Biol, 2000, 302:205 - 217.

[7] YANG Jing, LI Chengyun, WANG Yunyue, et al. Computational analysis of signal peptide-dependent secreted proteins in saccharomyces cerevisiae[J]. Agricultural Sciences in China, 2006, 5 :221 - 227.

[8] 谷俊峰,王希诚,赵金城. 多序列渐进式比对算法研究与比较[J]. 生物信息学, 2005, 2:73 - 76.

[9] 尚骅. 代数运算与自然数[EB/OL]. (2002 - 05 - 23)[2010 - 04 - 12]. http://media. open. edu. cn/media_file/rm/ip2/2002_5_23/gdds/gdds2/htm/gdds05. htm.

[10] MARC S, KENNETH S. Permutation distance measures for memetic algorithms with population management[C]. The Sixth Metaheuristics International Conference. Vienna, Austria, 2005.

[11] HIRSCHBERG D S. A linear space algorithm for computing maximal common subsequences[J]. Communications of the ACM, 1975,18: 341 - 343.

[12] 张阳, 李建良,胡正国. News Grouper: 一个自动抽取重要新闻的软件工具[J]. 计算机工程, 2002,28 :83 - 84.

[13] Phylip 软件包[EB/OL]. [2010 - 04 - 12]. http://evolution. gs. washington. edu/phylip. html.

[14] JONES D T, TAYLOR W R, THORNTON J M. The rapid generation of mutation data matrices from protein sequences[J]. Comput Appl Biosci, 1992,8 :275 - 282.