

# 启发式序列比对算法种子长度及其灵敏度研究

丁茂华,徐永安,邵 明,李 谦

(扬州大学 信息工程学院,江苏 扬州 225009)

**摘 要:**序列比对是生物信息学中一个重要的研究方向,它可以确定两个或多个序列之间的相似性,进而判断其同源性并推测出序列间的进化关系。目前,启发式序列比对算法 BLAST 算法在实际问题用着重要应用。该算法中有一个参数叫做种子(Seeds),种子是控制比对速度和灵敏度的关键。但是种子的长度是基于经验而取的一个固定值,这个经验值并不适合于所有长度序列比对问题。因此,对于两条不同长度的序列之间实现启发式比对就需要取合理长度的种子,以便实现高效快速的比对。文中应用概率随机的思想对不同长度序列比对的种子的长度进行了分析,在此基础上对一定长度下种子的比对灵敏度做出了计算。通过理论推导和实验分析一定灵敏度下种子长度的计算结果是可行且有效的。这就给在高灵敏度(灵敏度几乎等于动态规划算法)下实现快速启发式序列比对的优化提供了保证。

**关键词:**启发式比对算法;种子长度;灵敏度

**中图分类号:**TP391

**文献标识码:**A

**文章编号:**1673-629X(2013)02-0097-04

doi:10.3969/j.issn.1673-629X.2013.02.024

## Study on Seed's Length and Sensitivity of Heuristic Sequence Alignment Algorithm

DING Mao-hua, XU Yong-an, SHAO Ming, LI Qian

(College of Information Engineering, Yangzhou University, Yangzhou 225009, China)

**Abstract:** sequence alignment is an important direction in bioinformatics, which is to determine the similarity for two or more sequences and thus determines the homology and evolutionary relationship between the deduced sequence. At present, the heuristic sequence alignment algorithm has important applications in the practical problems. This algorithm has a parameter called seeds, which is to control the speed and sensitivity of the algorithm. But the length of the seed is based on the experience and takes a fixed value. Obviously, this value is not suitable for the actual problem of all length sequence. It studies the length of the seed by the probability theory, and, on this basis, studies the sensitivity of the algorithm. The heuristic sequence alignment algorithm can be used nearly as the same sensitivity of dynamic programming algorithm in theoretical analysis and experimental analysis. In this paper, the probability theory is used for the analysis of the length of the seed. The analysis of seed's length under a certain sensitivity is feasible and effective in theoretical analysis and experimental. This will ensure the sequence alignment algorithm with the high speed and high sensitivity, which can be used nearly as the same sensitivity of dynamic programming.

**Key words:** heuristic sequence alignment algorithm; length of the seed; sensitivity

## 0 引 言

随着生物信息数据的爆炸性增长,极大地促进了生物信息学的发展。利用计算机信息处理技术和算法对生物序列数据进行分析,以帮助生物学家进行判断和决策<sup>[1,2]</sup>。

文中研究的内容即是生物信息学的一个重要分支:序列比对(Sequences Alignment)。序列比对可以用

于判断序列间的相似程度,进而判别序列的同源性,同时推测出序列之间的进化关系等,这些在生物学上有着重要的意义。传统的序列比对算法如 Needleman-Wunsch 算法<sup>[3]</sup>和 Smith-Waterman 算法<sup>[4,5]</sup>由于时间复杂度和空间复杂度的原因很少应用于实际问题。

目前,基因序列比对序算法有些启发式算法(如 BLAST 算法)是非常快速的,但是比对的准确率以及灵敏度都有待于提高。这是因为启发式序列比对算法一般都是靠牺牲一定的灵敏度来提高序列比对速度的。由此可见,目前序列比对研究中面临的主要问题就是如何研究和设计即能实现快速比对又能具备高灵敏度的算法<sup>[6-9]</sup>。

收稿日期:2012-06-11;修回日期:2012-09-14

基金项目:国家自然科学基金资助项目(70972040)

作者简介:丁茂华(1988-),男,硕士研究生,主要研究方向为生物信息处理、算法优化;徐永安,副教授,博士,硕士生导师,主要研究方向为计算机图形学、科学计算可视化、视觉测量、逆向工程。

## 1 双序列比对算法

双序列比对算法研究开始于 20 世纪 70 年代,发展至今已经有 40 多年。序列比对又称联配就是对两条生物序列进行比较,找出它们之间的最大程度的相似。尽管在此期间有很多学者提出了很多算法,但是算法的核心思想基本上基于动态规划的详尽式的比对算法和启发式比对算法。下面对这两种类型的算法做出简要介绍。

### 1.1 动态规划序列比对算法

应用动态规划的详尽式的算法最为代表性的是 Needleman-Wunsch 算法和 Smith-Waterman 算法。Needleman 和 Wunsch 在 1970 年提出的 Needleman-Wunsch 算法中。该算法的主要思想是通过构建一张二维表,然后运用动态规划从该表格左上角直至右下角进行打分比较,最后再从该表格的右下角至左上角进行回溯。这是一种全局比对,最优比对中包括了全部的最短匹配序列。Smith-Waterman 是一种局部比对算法,它是在 Needleman-Wunsch 算法的基础上发展而来的。这种比对的路径不需要到达搜索图的尽头,只需要在内部开始和终结。Smith-Waterman 算法先用迭代方法计算出两个序列的所有可能相似性比较的分值,然后通过动态规划的方法回溯寻找最优相似性比对。Smith-Waterman 局部比对算法与前面介绍的 Needleman-Wunsch 全局比对算法的时间复杂度和空间复杂度相同,即均为为  $O(mn)$ 。因此,这两个算法因其时间复杂度和空间复杂度的原因很少应用于实际问题。

### 1.2 启发式序列比对算法

启发式序列比对算法主要用于实际生物信息学中的数据库相似性检索的,它能在基因数据库中找出在一定程度相似的核酸或者蛋白质。目前,使用较多的数据库相似搜索算法 BLAST 算法<sup>[10-12]</sup>。BLAST 算法是 Altschul 等在 1990 年提出的,该算法的基础是序列局部比对算法-Smith-Waterman 算法。

BLAST 算法的主要步骤是:

- 1) 首先给定一个较短的匹配模式,该匹配模式包括  $w$  个连续匹配,在序列中搜索满足该模式的增强点;
- 2) 对增强点采用动态规划算法进行最优得分延伸,从而找到高分片段对(HSP)。

BLAST 基本思想是:通过产生数量更少的,但质量更好的增强点来提高比对的速度。它是建立在严格的统计学基础之上,它集中于发现具有较高相似性的局部比对。BLAST 算法是一个基于动态规划理论的近似算法,它通过寻找 HSP 片段较大的减少了所需的时间和内存空间。该算法的主要缺点是若同源性搜索匹

配过程中所得的片段增大,可以减少搜索时间,但是搜索的敏感性不能得到保证;同时若减少同源性搜索匹配片段,搜索时间又会较大的增加。可以简单的说 BLAST 算法的就是种子和延伸。可见种子在 BLAST 算法中的地位是多么的重要。关于种子的内容留在文中的下面部分作出详细介绍。

### 1.3 两种类型算法的矛盾

通过以上分析详尽的序列比对算法和启发式比对算法之间最大的矛盾就是比对速度和比对灵敏度之间的矛盾。Needleman-Wunsch 算法和 Smith-Waterman 算法虽然具有高灵敏度,但是速度很慢且占用内存空间;BLAST 算法虽然是快速,但是灵敏度相对较低。因此有必要研究一种算法在快速比对的情况下保证比对灵敏度。

## 2 相关研究工作

### 2.1 种子的概念及其作用

在启发式序列比对算法中,有一个非常重要的概念那就是“种子”。种子就是一些比较可信或者比较有代表性的序列集合,在这些序列的基础上,进行一些序列比对,可以获取更多的目标序列。种子是启发式算法中一个关键,只有找出一定数量的种子才能调用下一阶段动态规划算法。种子的作用相当于详尽的比对算法的部分关键路径,这实际是利用快速比对找出关键路径。

种子的长度  $w$  是调节比对速度和灵敏度的阀门,也就是说,  $w$  值越大,比对速度越快,但灵敏度越低;  $w$  值越小,比对速度越慢,但灵敏度越高。这是因为种子的长度相对于序列长度是非常短的,那么,无间隙的比对所需要的时间远远小于动态规划比对的时间。

基因序列的长度短的就几个碱基组成,长的可以达到成百上千的长度,而传统的种子的长度是基于经验而确定的。通常当比对的序列为蛋白质时种子长度为 3;比对序列为核酸种子长度为 11。但是实际问题的序列的长度是有长有短的,对于长度不一样的基因序列的比对都取上面的经验值显然不合适的。因此这种默认长度显然不适合所有序列比对的种子的长度。而且根据经验给出的种子长度,缺乏种子质量的度量。因此,对于种子定量性研究就有必要了。

下面就介绍了一种基于概率随即思想来分析种子的长度和灵敏度。

### 2.2 随机性种子和非随机性种子

定义 1:在启发式序列比对算法中由于两对比序列随机性所产生的种子称为随机性种子。

定义 2:在启发式序列比对算法中由于两对比序列非随机性,而由于生物信息本生特性所产生的种子

称为非随机性种子。

例如:序列  $S_1 = \text{AGAAAACGTA}$  和序列  $S_2 = \text{AGAG-GAACCTA}$ ,假定种子长度  $w = 3$ 。图1是  $S_1$  和  $S_2$  无间隙比对找到的种子为两个子片段 AGA 和 AAC。假如 AGA 片段能够反应生物学本身的特性那么该种子称为非随机性种子;假如 A A C 片段能够不能够反应生物学本身的特性那么该种子称为随机性种子。基因序列比对需要的是非随机性种子而不是随机性种子。

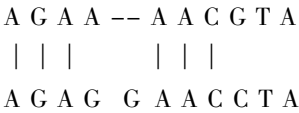


图1 种子片段示意图

性质1:在启发式序列比对算法中种子的随机性是不可以消除的。

该性质是由于启发式序列比对算法、无间隙比对的不确定性决定的。

2.3 种子计算

定义3:一条生物序列  $S$  中的字符由某个有限字符集合  $\Sigma$  确定。

定义4: $z$  表示  $\Sigma$  集合中元素种类的总和。

例如:核酸字符集合为  $\{A, C, G, T\}$  或  $\{A, C, G, U\}$ ,那么  $z=4$ ;蛋白质字符集合为  $\{A, M, V, L, I, P, F, W, G, S, T, C, Y, N, Q, K, H, R, D, E\}$ ,那么  $z=20$ 。

性质2:假设序列中每个都是相互独立的,那么元素种类为  $z$  的  $\Sigma$  集合长度均为  $w$  两个序列完全一样的概率  $P_w$  为:

$$P_w = (1/z)^w \tag{1}$$

例如:集合为  $\{A, C, G, T\}$  或  $\{A, C, G, U\}$  两条长度  $w = 5$  两个序列完全一样的概率为:

$$P_w = (1/z)^w = (0.25)^5 = 0.0009765625$$

集合为  $\{A, M, V, L, I, P, F, W, G, S, T, C, Y, N, Q, K, H, R, D, E\}$  两条长度  $w = 5$  两个序列完全一样的概率为:

$$P_w = (1/z)^w = (0.05)^5 = 0.0000003125$$

性质3:对于两比对序列长度为  $m$  和  $n$  它们可以长度为  $w$  取  $(m - w + 1) \times (n - w + 1)$  对公共子序列。

例如:序列  $S_1 = \text{ACGTG}$  和  $S_2 = \text{AGCT}$  在  $w = 3$  的时候公共子序列的对数为:  $(m - w + 1) \times (n - w + 1) = (5 - 3 + 1) \times (4 - 3 + 1) = 6$  对。

即为:ACG 与 AGC ,ACG 与 GCT ,CGT 与 AGC , CGT 与 GCT,GTG 与 AGC,GTG 与 GCT。

推论1:对于对于两比对序列长度为  $m$  和  $n$  不产生随机性性种子的概率是:

$$P_1 = 1 - (1 - P_w)^{(m - w + 1) \times (n - w + 1)}$$

$$= 1 - [1 - (1/z)^w]^{(m - w + 1) \times (n - w + 1)} \tag{2}$$

2.4 灵敏度分析

基于启发式比对算法的速度是大大提高了,但它 是通过牺牲灵敏度来提高速度的。而种子长度的大小  $w$  决定了比对的速度和灵敏度。上面已经介绍过随机性种子和非随机性种子这两个概念,只有非随机性种子是一种基于序列本身特性而产生的种子,而随机性种子是种子的随机特性产生的。因此,在序列比对问题中需要的是非随机种子而不是随机种子。根据定义1、定义2 和性质1 可知,种子的随机性具有不可消除性,但是随机性的大小是可以控制的。控制随机性的最好的办法就是引入概率分析,因此通过概率的大小可以控制种子的随机性。因此下面就对两个序列的产生长度为  $w$  的随机性种子做以讨论。

可以根据上述理论两条序列的灵敏度的计算公式为:

$$fsens = 1 - (1 - p_w)^{(m - w + 1) \times (n - w + 1)} \tag{3}$$

$fsens$  的取值从0 到1。 $fsen$  取值越大表示灵敏度越高,当其值等于1 的时候表示灵敏度取得最大值,即动态规划也是一样的灵敏度。但是在启发式序列比对算法中  $fsens=0$  和  $fsens=1$  是不存在的。这是因为这两种情况是属于确定性事件,这个与性质1 是矛盾。可以按照需要的灵敏度来控制种子长度。通过控制灵敏度的控制可以找出满足灵敏度最短的种子长度。

3 实验数据与分析

3.1 实验数据

根据上述灵敏度计算公式,表1 和表2 是计算灵敏度  $fsens \geq 0.99$  的两种相同长度序列长度种子  $w$  的取值,这里只列出序列长度部分序列长度相的计算结果。表3 和表4 分别是两条序列长度为500 不同灵敏度下种子核酸和蛋白质序列种子长度。

表1 灵敏度为0.99 不同长度蛋白质序列种子的长度

序列长度	种子长度	两者比值
3 ~ 10	3	1 ~ 3.3
11 ~ 183	5	2.2 ~ 36.6
184 ~ 1000	7	26.3 ~ 142.9

表2 灵敏度为0.99 不同长度核酸序列种子的长度

序列长度	种子长度	两者比值
8 ~ 18	7	1.1 ~ 2.5
19 ~ 59	9	2.1 ~ 6.5
60 ~ 215	11	5.5 ~ 19.5
216 ~ 833	13	16.6 ~ 64.1
834 ~ 1000	15	55.6 ~ 66.7

3.2 实验数据分析

通过以上实验数据可以得出如下结论:

1)表 1 和表 2 随着序列比对长度的增加种子的长度也在增加。无论蛋白质序列还是核算序列都是满足这个性质。

表 3 两条序列长度为 500 不同灵敏度下种子蛋白质序列种子长度

灵敏度	种子长度	灵敏度	种子长度
0.1	5	0.6	5
0.2	5	0.7	5
0.3	5	0.8	5
0.4	5	0.9	5
0.5	5	0.99	7

表 4 两条序列长度为 500 不同灵敏度下种子核酸序列种子长度

灵敏度	种子长度	灵敏度	种子长度
0.1	9	0.6	11
0.2	9	0.7	11
0.3	9	0.8	11
0.4	11	0.9	11
0.5	11	0.99	15

2)表 1 和表 2 无论比对序列的长短,在核酸种子长度  $w$  都是集中在 10 左右,蛋白质种子长度都是集中在 4 左右。这个实验值与经验值所取的值  $w = 11$  (核酸)和  $w = 3$  (蛋白质)基本上是一致的。

3)表 3 和表 4 表明了对于相同的序列比对情况,灵敏度要求越高的种子长度需要越长。

4)表 1 和表 2 在如此高的灵敏度长度下种子长度相对远远小于序列长度的,尤其是长度较大的序列更为明显。这就对启发式序列比对算法有指导意义。

3.3 指导意义

利用启发式的序列比对算法更符合生物信息学本身的特性,这是因为利用动态规划算法的打分是人为制定的,从而从某种程度上破坏生物学本身的特性。对于大序列比对的种子的长度可以取序列长度的很小的一部分而保持比对质量很高。因此根据以上的分析对于启发式序列比对算法中,可以用很短的种子来得到高灵敏度的序列比对算法。

高灵敏度快速序列比对算法流程图如图 2 所示。该算法可以既拥有动态规划算法的高效率,同时又可以实现快速比对。

4 结束语

文中针对启发式序列比对算法中的一个重要参数—种子进行了研究。应用概率随机的思想对种子的长度计算问题进行了研究,同时针对一定长度下种子的比对灵敏度做出了量化计算。通过理论推导和实验分析得出了种子长度计算公式和灵敏度的计算公式。通过分析证明可以保证为序列比对算法的优化提供保

证。

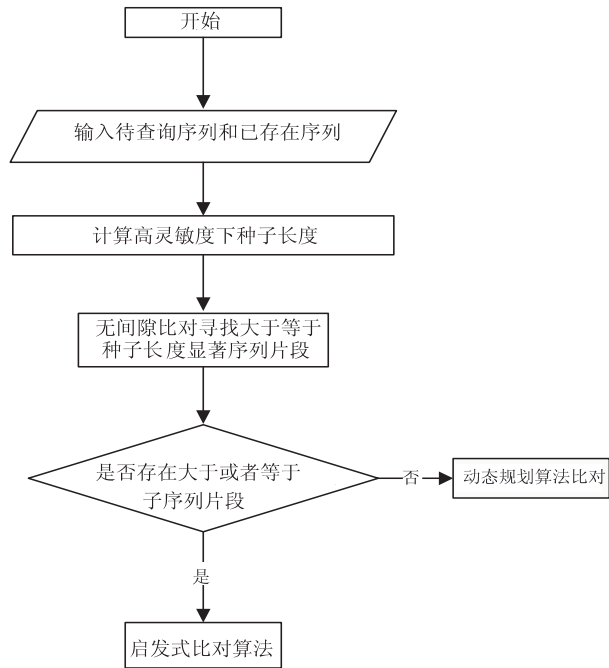


图 2 高灵敏度快速序列比对算法

参考文献:

[1] 许忠能.生物信息学[M].北京:清华大学出版社,2008.

[2] 詹超,胡江洪.SVM在基因表达数据分类中的研究和应用[J].计算机技术与发展,2006,16(3):107-109.

[3] Needleman S B,Wunsch C D. A general method applicable to the search for similarities in the amino acid sequence of two proteins[J]. J Mol Biol,1970,48(3):443-453.

[4] Smith T, Waterman M. Identification of common molecular subsequence[J]. Journal of Molecular Biology, 1981, 147(1):195-197.

[5] Smith T F, Wateman M S,Fitch W M. Comparative biosequence metrics[J]. J Mol Evol,1981,18(1):38-46.

[6] 唐玉荣.生物信息学中一个优化的全局双序列比对算法[J].计算机应用,2004,24(Sup):307-308.

[7] 李昭.存储约束条件下的序列联配算法[J].微电子学与计算机,2002,19(6):1-5.

[8] 魏大木,陶宏才.序列比对算法简单研究[J].微计算机信息,2011,27(4):201-203.

[9] 吴德敏,陈俊.双序列比对的算法研究[J].计算机工程与应用,2008,44(36):48-51.

[10] Lipman D J,Pearson W R. Rapid and sensitive protein similarity searchers[J]. Science,1985,227(4693):1435-1441.

[11] Altschul S F,Gish W,Miller W,et al. Basic local alignment search tool[J]. Journal of Molecular Biology, 1990, 215(3):403-410.

[12] Altschul S F,Madden T L,Schaffer A A,et al. Capped BLAST and PSI-BLAST: new generation of protein database search programs[J]. Nucleic Acids Res,1997,25(17):3389-3402.

## 启发式序列比对算法种子长度及其灵敏度研究

作者: [丁茂华](#), [徐永安](#), [邵明](#), [李谦](#)  
作者单位: [扬州大学 信息工程学院, 江苏 扬州225009](#)  
刊名: [计算机技术与发展](#)  
英文刊名: [Computer Technology and Development](#)  
年, 卷(期): 2013 (2)

本文链接: [http://d.g.wanfangdata.com.cn/Periodical\\_wjtz201302026.aspx](http://d.g.wanfangdata.com.cn/Periodical_wjtz201302026.aspx)