研究快报

一种查找 DNA 顺序间同源片段的计算机程序及其在小麦贮藏蛋白基因研究上的应用

陶 芸 金惠生* 易 莹 顾其敏

（复旦大学生命科学学院，上海）

在对蛋白质和核酸一级结构数据处理中，常用到“两段顺序比较”的算法，目前一般都用 NWS 算法[1,2]，但这种算法的计算机内存耗用太大，理论上为 $\prod_{i=1}^{k} \mathrm{Li}$（$k$ 条链比较），多链比较如在多条 DNA 顺序中查找同源片段时，耗费大量机时，极不方便。我们设计了一种新的算法，使计算机的内存耗用大大降低，速度相应极大提高。基本思想如下：(1) 将 NWS 算法的两链比较的矩阵（两维空间）改成一链在另一链上的投影，投影的结果是得到一个布尔向量（一维空间）。对应的字符相同得 “T”，相异得 “F”。在此布尔向量中，顺序地取定长 LH 个元素，计 “T” 的个数，如果越过一阈值 C，则称两链间存在一同源片段（即长为 LH 的片段中至少有 C 个对应的核苷酸是相同的）。改变两链的相对位置（即每次将一链在另一链上向前移一个核苷酸的长度），又可投影得到一新的布尔向量，寻找新的同源片段。如此直到所有可能的投影都完成为止。可以看出，这里的布尔向量元素相当于 NWS 算法中的矩阵对角线元素。这样改进的结果可以大大节省内存。例如在长分别为 L_1 和 L_2 的两链中寻找 LH 的同源片段，NWS 的内存需求是 $L_1 \times L_2$，而我们算法的内存需求仅 $L_1 + L_2 + \mathrm{MIN}(L_1, L_2)$，$\mathrm{MIN}(L_1, L_2)$ 是用于存放两链投影得到的布尔变量所耗的空间。(2) 第 2 个改进是在多链间寻找同源片段，不需象 NWS 算法那样一次读入所有顺序（k 条），建立一个 k 维空间，而是将前 i 条链中找到的同源片段（共 Ki 组）Hi 与第 $i+1$ 条链比较（$i = 2, 3, \cdots k-1$），我们已证明[3] 这种方法与原链两两比较结果是一致的。改进的结果是内存耗用从 $(k+1)L_1L_2$[4] 减少到 $3L$ 左右，算法复杂度从 $O\left(\prod_{i=1}^{k} li\right)$ 减少到 $O\left(\mathrm{mn} + \mathrm{LH}\sum_{i=3}^{k} R_{i-1}l_i\right)$。可见无论从速度还是内存上，本算法都有极大改进。

根据上述算法我们在 IBM-PC 上完成一软件 Homology，使用此软件我们对小麦贮藏蛋白基因顺序进行了分析，取得了有意义的结果：从 11 条编码醇溶蛋白的 DNA 顺序间找到同源度为 72% 的一条 21 核苷酸长的同源片段。从 3 条编码完整高分子量谷蛋白的 DNA 顺序间找到一段 23 核苷酸长和一段 28 核苷酸长的同源片段，同源度分别为 100% 和 93%。而编码高分子量谷蛋白的 DNA 顺序与编码醇溶蛋白的 DNA 顺序间，当同源度大于 70% 时，找不到同源片段。

参 考 文 献

[1] Needleman, S. B. et al.: *J. Mol. Biol.*, 1970, **48**, 443.

[2] Sellers, P. H.: *J. Appl. Math. (Siam)*, 1974, **26**, 787.

[3] Jin, H. S. & Tao, Y.: *Nucl. Acids Res.*, 待发表。

[4] Krishnan, N. et al.: *Nucl. Acids Res.*, 1986, **14**, 543.

[本文于 1988 年 1 月 20 日收到]

* 复旦大学计算机科学系，上海。