Statistical method for rapid homology search

Andrey A. Mironov and Nickolay N. Alexandrov

Institute for Genetics of Microorganisms, 1-st Dorozhnaya pr. 1, Moscow 113545, USSR

Received July 28, 1987; Revised and Accepted May 4, 1988

ABSTRACT

A new method for homology search of DNA sequences is suggested. This method may be used to find extensive and not strong homologies with point mutations and deletions. The running program time for comparing sequences is less than the dynamic program algorithms at least at two orders of magnitude. It makes possible to use the method for homology searching through over the nucleotide bank by personal computers.

INTRODUCTION

The high speed of DNA sequencing reveals wide possibilities for discovering new regularities in biology. One of the most important methods of DNA, RNA or protein texts analysis consists in finding homology regions. Two main conclusions can be drawn: 1) these sequences perform the same function and/or 2) similar DNA fragments are of the same origin. Thus, biopolymers texts comparison is necessary for solving the fundamental problems of molecular biology, namely the biopolymers structure/function relationship and evolutionary hierarchy in nature.

The most widely used sequence comparison algorithms determine the similarity by counting matches within the optimal (in some sense) alignment [1-3]. The obvious advantage of such an approach is its constructivity - similar regions are presented in a convenient form for following analysis. The disadvantages of this method are the subjective criteria of the optimal alignment and the large number of operations needed to provide this alignment. So, the comparison of two sequences with lengths $N_1$ and $N_2$ requires approximately $10^2 N_1 N_2$ operations.

For instance the homology search between a sequence $N_1=10^2$ and the entire nucleotide bank $N_2=10^7$ requires $10^{12}$ operations. The time needed for this comparison is equal $10^6$ seconds or one...
day running of quite fast computers with speed $10^7$ operations per second. Several attempts were made to reduce homology searching time [4]. But there is no yet universal method which permits to find any type of sequence similarity in nucleotide bank by personal computer. And one of the main reasons is ambiguousness of the term homology. It is necessary to use different methods to have a guarantee that you do not lose homologous regions. At this paper we preset a new method for searching long but not strong homologous regions. The method does not require sequence alignment also as one suggested by Blaisdell, B.E. [5].

METHOD

We can increase the rapidity of the aforementioned algorithms by applying the filtration with necessary conditions of homology. The most simple variant of such conditions is the similarity of nucleotide contents: if sequences are homologous then their nucleotide contents are also kindred (the opposite is wrong). This way of comparison may be used for preliminary selection of sequence pairs which are thought to be homologous and must be completed by the following construction of the optimal alignment. It is possible to create a number of comparison methods based on the reflection of DNA sequences into some vector space with defined distances between them. Here we describe one of such reflections and the corresponded algorithm.

Let $|S_i|$ be nucleotide sequence with length $N$, and $\varepsilon_i^j$ - characteristic function of the nucleotide of type $j$ at position $i$:

$$
\varepsilon_i^j = \begin{cases} 
0, & S_i = j \\
1, & S_i = \not{j}
\end{cases}
$$

Let then $p_j$ be the occurrence probability of $j$-type nucleotide. We will deal with normalized deviation of the dinucleotide content for nucleotides separated by $q$ positions:

$$
b_{q}^{k} = \frac{1}{\sqrt{(N-q)}} \sum_{i=1}^{N-q} (\varepsilon_i^k \varepsilon_{i+q}^k - p_k p_k) = \frac{1}{\sqrt{(N-q)}} \sum_{i=1}^{N-q} b_{q}^{k}
$$
The distance between sequences $S$ and $S'$ is defined by the following formula:

$$r(S,S') = \frac{1}{l} \sum_{q=1}^{l} \sum_{k \in A} \sum_{m=1}^{r} (b_q^k - b_q^m)^2$$

where $l$ is the parameter of the algorithm. If all $p_i$ are equal 0.25 one may easily calculate the mathematical expectation of the distance between random sequences:

$$\mu = M(r) = 15/8 = 1.875$$

Monte-Carlo modelling gives the same result.

Consider some important properties of this distance. Let two sequences $S$ and $S'$ be distinguished by a nucleotide at position $f$. Then the distance between sequences is:

$$r = \frac{1}{Nl} \sum_{q=1}^{l} \sum_{k \in A} (\sum_{i=1}^{f-q} \beta_{q,i}^k + \sum_{i=f+1}^{f-q} \beta_{q,i}^k - \sum_{i=f+1}^{f} \beta_{q,i}^k - \beta_{q,f-1}^k)^2$$

Taking into account that the first and fourth item in brackets are equal by simple calculations we obtain $r = 4/N$.

Let us suppose that $S'$ has been received from $S$ by deletion of nucleotides at positions $f,f+1,f+2,...,f+d$. The distance now will be

$$r = \frac{1}{Nl} \sum_{q=1}^{l} \sum_{k \in A} \left( \sum_{i=1}^{f-q} \beta_{q,i}^k - \sum_{i=f+1}^{f+d} \beta_{q,i}^k \right)^2$$

where $a = \sqrt{(N/(N-d))} = 1+2d/N$. By neglecting term $2d/N$ we obtain

$$r = \frac{1}{Nl} \sum_{q=1}^{l} \sum_{k \in A} \left( \sum_{i=1}^{f-q} \beta_{q,i}^k + \sum_{i=f+1}^{f+d} \beta_{q,i}^k - \sum_{i=f+1}^{f} \beta_{q,i}^k - \beta_{q,f-1}^k \right)^2$$

$$= \frac{1}{Nl} \sum_{q=1}^{l} \sum_{k \in A} (B_1 + B_2 + B_3 + B_4)^2$$
$B_t, B_s$ and $B_t, B_e$ are statistically independent giving the mathematical expectation of the distance

$$M(r) = (2(1-1)+d)/N$$

As if all the distance changes above have the same order of $1/N$ one may conclude that the distance between sequences characterizes the density of differences between them. Note that the statistical characteristics of $r$ do not depend on the length of sequences if $q << N$. If sequences $S_i$ and $S_j$ with length $N$ have a common fragment with length $L$ then

$$M(r) = \sqrt{(1-L/N) \ast \mu}$$

where $\mu = 15/8$.

It is clear that this method allows to find homology quite reliably in case when the size of the homologous region is more than $0.75N$. That is why the extensive sequences must be divided into short fragments with lengths equal to the expected homology size. To make the results independent from the way of division it is reasonable to use two divisions shifting at $N/2$. In this case the worst positions of homology regions will be shifted at $N/4$ and the homology will be found.

**ALGORITHM**

Now we can formulate the algorithm of the sequence comparison. Sequences to be compared are divided into fragments with length $N$, where $N$ is the minimal expected homology size. For each of the fragment pairs the distance $r$ is calculated. If $r$ occurs smaller then $r_0$ - the cutoff value, this sequences can contain homologous fragments.

By using the algorithm we found the convenient values of parameters: $l = 3$, $r_0 = 0.5$. Sequence comparison requires $-100N_tN_e/L^2$ operations. When $N_t = 1000$, $N_e = 10^7$, $L = 100$ the operation number equals $10^8$. The time of searching throughover the bank takes only 15 minutes when using a personal computer.

In order to test our algorithm all the coding regions from the All-Union bank [6] (above five millions nucleotides) have
been compared each with others. The coding regions were selected by two main reasons: first, they are quite extensive to illustrate the high speed of the algorithm; second, the results of the analysis should be easily interpreted. All sequences have been divided into fragments of 200 nucleotides with 100 nucleotides overlapping.

The parameters have the following values: $l = 3; r_0 = 0.4$. We have obtained 2000 sequence pairs which are similar in terms of the described metrics. Most of them have been then analyzed by dot-matrix method [7]. All these sequences belong to one of the two groups: 1) homologous sequences or 2) randomly close sequences without significant homology. The second group includes 100 pairs i.e. nearly five per cent of the total number of similar sequences.

In conclusion we want to underline the practical usage and simplicity of the method. For instance it may be used for updating the bank.

REFERENCES