Nhunt: new program for DNA sequence similarity searching

Yury A. Pekov[1], Sergei S. Spirin[2]

[1]Faculty of Bioengineering and Bioinformatics of Moscow State University, Moscow, Russia
[2]Belozersky Institute of Moscow State University, Moscow, Russia

Introduction

DNA sequence similarity searching in databases is one of the most important problems of bioinformatics.

Background

In the case of coding sequences the program TBLASTN is successfully used. In the case of noncoding sequences, a number of programs is used, e. g. FASTA, BLASTN and discontiguous MEGABLAST. But each of these programs has some significant disadvantage.

Aim

The aim of this work was to create Nhunt computer program for DNA sequence similarity searching that would exceed both FASTA and BLASTN in sensitivity. An original algorithm for diagonal selection was applied, which allows to adjust the ratio "speed / sensitivity".

Methods

Algorithm

The algorithm for alignment construction is based on FASTA algorithm, but devoid of inherited disadvantages. Formula for E-value calculation is based on Karlin — Altschul extreme value distribution. Its parameters were fitted using a large random database.

Low complexity problem

Assuming a Bernoulli nucleotide model any scoring matrix can be written in the form

$$s_{ij} = \frac{1}{\lambda} \ln(\frac{q_{ij}}{p_i p_j}),$$

where $p_i = p_j = \frac{1}{4}$ and q_{ij} is a frequency of i,j-nucleotide matching. But for low complexity regions it is reasonable to use adjusted scoring matrix, that prevents alignments to be high-scored by accidental causes. We set

$$\tilde{q}_{ij} = \tilde{p}_i \tilde{p}_j + c r_{ij},$$

where $\tilde{p}_i$ are nucleotide frequencies, and c is a parameter that is dependent on $\tilde{p}_i$ and allows to get adjusted scoring matrix $\tilde{s}_{ij}$. In Bernoulli model $c = 1, \tilde{q}_{ij} = q_{ij}$. The calculation of c parameter is based on conservation of expected score:

$$\sum_{i,j} p_i p_j s_{ij} = \sum_{i,j} \tilde{p}_i \tilde{p}_j \tilde{s}_{ij}.$$

This equation is resolved by numerical methods.

Results

Comparison with FASTA

FASTA version: 36.3.4, FASTA parameters (in command line): `-f -10 -g -5, ktup = 3.`
Nhunt and FASTA programs were run for searching homologues of *E. coli* tRNA in 5 archaean genomes.

Table 1: Number of found alignments with E-value less than given

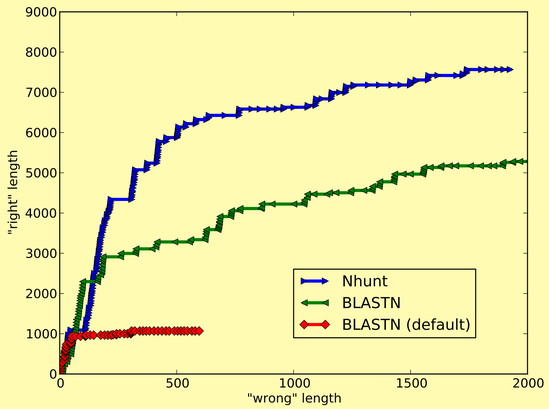
Program	Run time	10^{-19}	10^{-3}	0.1	3	20
FASTA	5.5	1	6	14	27	49
Nhunt	1.4	2	8	17	42	129

Comparison with BLASTN

BLASTN version: 2.2.24, BLASTN parameters (in command line): `-W 7 -F F -r 5 -q -4 -G 10 -E 6.`
Comparison with discontiguous MEGABLAST program showed that it concedes BLASTN in sensitivity. Nhunt and BLASTN programs were run for searching homologues of *E. coli* miscRNA in three bacterial genomes.

Table 2: Number of found alignments with E-value less than given

Program	10^{-10}	10^{-6}	10^{-4}	0.1	20
B. cereus					
BLASTN	2	5	14	146	859
Nhunt	2	6	15	238	4135
P. aeruginosa					
BLASTN	4	6	11	97	369
Nhunt	4	8	13	187	2060
Y. pestis					
BLASTN	23	29	34	112	684
Nhunt	25	30	41	167	2796



Nhunt and BLASTN programs were also run for searching homologues of *E. coli* tRNA in five archaean genomes.
Then coordinates of found homologues were compared with coordinates of annotated archaean tRNA and lengths of "wrong" and "right" fragments were calculated for each alignment.

In plot above one point corresponds to one alignment. X and Y axes values are respectively total lengths of "wrong" and "right" fragments of alignments with E-value less than some value.
BLASTN (default) is BLASTN with default parameters.

Low complexity example

•Example alignment:

```
Query:  ctgtttaccaggtcaggtccggaaggaagcagccaaggcagatgacgcgt
        ||||  ||||| | |   ||| ||  || |||||||  || |||  || |
Sbjct:  ctgtgaaccagcttatcgccgcaatcaaacagccaaatcatatgcagcat
Identity = 33/50 (66%)
Strand:  Plus/Minus
```

•Alignment features without scoring matrix adjustment:

```
Score:  97.0 (31.8 bits), E-value:  0.1586
```

•Alignment features after scoring matrix adjustment:

```
Initial score:  103.5 (33.7 bits),
adjusted score:  64.80 (22.3 bits), E-value:  112.932
Query frequences:  A: 0.26, T: 0.17, G: 0.33, C: 0.24,
Subject frequences:  A: 0.33, T: 0.21, G: 0.17, C: 0.29,
Adjusted substitution matrix (subject letters are in rows,
query letters are in columns):
    A      T      G      C
A  3.6   -3.5   -1.5   -2.2
T -3.8    6.3   -2.7   -4.3
G -5.2  -22.1    4.8   -5.9
C -2.4   -4.2   -1.7    4.1
```

Realization and availability

The program is realized on C programming language.
Executable files for Linux x86 and amd64 architectures, as well as the source code of the program are accessible in Internet: http://mouse.belozersky.msu.ru/~bennigsen/nhunt.html

Conclusions

1. We have created Nhunt computer program for DNA sequence similarity searching. This program has been successfully tested on a set of genomes.
2. We have proposed a new method of scoring matrix adjustment for low complexity regions.
3. It was shown that for all examples Nhunt exceeds FASTA program both in sensitivity and in speed. Nhunt also exceeds BLASTN in sensitivity.

Acknowledgements

The work is partly supported by the grant #10-07-00685-a of Russian Foundation of Basic Research