WEB Based GENOME Analysis Tools in ALIS Project

Kazuo Takehana [1] take3@tokyo.jst.go.jp
Takehiko Ito [2] takehiko@mri.co.jp
Tetsushi Yada [3] yada@ims.u-tokyo.ac.jp
Mika Hirakawa [1] mika@tokyo.jst.go.jp
Hiroko Yamaguchi [1] yamako@tokyo.jst.go.jp
Junko Shimada [1] sjunko@tokyo.jst.go.jp
Katsuji Matsumura[1] matsumur@jst.go.jp
Masako Kuroda [1] kuroda@jst.go.jp

[1] Bioinformatics Division, Advanced Databases Department, Japan Science and Technology Corporation(JST), 5-3 Yonbancho, Chiyoda-ku, Tokyo 102-0081, Japan
[2] Department of Informatics and Mathematical Science, Mitsubishi Research Institute, INC.(MRI), 3-6 Otemachi 2-Chome, Chiyoda-ku, Tokyo 100-8141, Japan
[3] Genome Informatics Team, Human Genome Research Group, Genomic Sciences Center (GSC), The Institute of Physical & Chemical Research (RIKEN), 4-6-1 Shiroganedai, Minato-ku, Tokyo 108-8639, Japan

1 Introduction

In recent, sequencing and related technologies are progressing amazingly, so flood of sequencing data is produced day by day, and the role of Bioinformatics becomes more and more important. Though computational power increases and many genome analysis programs are maintained, their character based usage make biologists suffered to handle them.

In JST, targeting on ease of use, we have developed "Genome Analysis Tool" as part of ALIS (Advanced Life-science Information Systems) Project (`http://www-alis.tokyo.jst.go.jp/`).

These tools are WEB based systems so that anyone can use them without particular knowledge about computer programs. A user enters query sequence and necessary parameters through WEB browser, the data is analyzed on JST's super computer. The result is returned to user's E-mail address, and displayed graphically and interactively at our WEB site (`http://www-scc.jst.go.jp/sankichi/`).

2 Genome Analysis Tools

At present, we provide the following 5 systems as genome analysis tools.

2.1 Multiple DNA Sequence Alignment

The multiple alignment in this system uses Tree-based Round-Robin Iterative Algorithm (TRRIA) [1].

The method is a combination of progressive and iterative methods, which means a longer computation time causes a more precise multiple alignment than the result of compared programs currently considered practical, such as PileUp in the GCG package and ClustalW developed by EMBL.

2.2 GeneHacker

GeneHacker is a system for gene structure prediction in microbial genomes using hidden Markov model (HMM) [2]. An HMM adopted in GeneHacker describes the start codon and its downstream di-codon frequencies in protein coding regions, and can identify the regions in uncharacterized DNA sequences.

This system offers 8 models suited for analyzing the sequences of a wide range of prokaryotes. These models are; *Bacillus subtilis*, *Escherichia coli*, *Haemophilus influenzae*, *Helicobacter pylori*, *Methanococcus jannaschii*, *Mycoplasma genitalium*, *Mycoplasma pneumoniae*, *Synechocystis sp.*.

2.3 Motif Extraction / Motif Search

This is a system for automatic extraction of motifs or search motifs that occur frequently on a set of unaligned DNA sequences [3].

2.4 GeneWalker

GeneWalker is a human gene identification system. It analyzes a given DNA sequence, identifies likely signal sequences (TATA, CAAT, etc.) by similarity matching (using weight matrices), finds fragments that look like start codons, stop codons, acceptor sites, and donor sites, etc., and calculates Coding Potential values at each base position based on a statistical analysis of the local segment, which indicates the likelihood of the base being in a coding region.

After the analysis, GeneWalker tries to predict sub-sequences which are likely to be promoters, exons, or terminators, and to compile local information into global predictions of regions.

The GeneWalker viewer displays not only the final predictions, but various other intermediate information GeneWalker produced, such as likely signal sequences, likely start and stop codons, donor and acceptor sites, and coding potential values along the sequence.

GeneWalker was trained by GENSCAN 's training data set of 380 Human sequences, and tested by the set of 570 vertebrate gene sequences constructed by Burset & Guigo as a standard test set of gene finding.

2.5 Homology Search

This system provides the homology search program BLAST (BLAST 1.4 & 2.0) via WEB. Search available databases are SWISS-PROT (release 36.0) and GenBank (the latest release and daily-updated entries).

References

[1] Hirosawa, M., Totoki, Y., Hoshida, M., and Ishikawa, M., Comprehensive study on iterative algorithms of multiple sequence alignment, *Comput. Appl. Biosci.*, 11:13–18, 1995.

[2] Yada, T., Totoki, Y., Ishii, T., and Nakai, K., Functional prediction of *B. subtilis* genes from their regulatory sequences, *Proc. Fifth Int. Conf. Intell. Syst. Mol. Biol.*, 354–357, 1997.

[3] Yada, T., Totoki, Y., Ishikawa, M., Asai, K., and Nakai, K., Automatic extraction of motifs represented in the hidden Markov model from a number of DNA sequences, *Bioinformatics*, 14:317–325, 1998.