Genome Analysis and Gene Discovery Tools

Christian Iseli[1,2] Christian.Iseli@licr.org
C. Victor Jongeneel[1,2] Victor.Jongeneel@licr.org
Marco Pagni[1,3] Marco.Pagni@isrec.unil.ch
Thomas Junier[1,3] Thomas.Junier@isrec.unil.ch
Philipp Bucher[1,3] Philipp.Bucher@isrec.unil.ch

[1] Swiss Institute of Bioinformatics, Chemin des Boveresses 155, 1066 Epalinges, Switzerland
[2] Office of Information Technology, Ludwig Institute for Cancer Research
[3] Swiss Institute for Experimental Cancer Research

Keywords: gene discovery, EST assembly, similarity, search, sequence motifs, molecular databases

1 Introduction

A working draft of the human genome will be publicly available in a few months, a finished version within about two years. However, it will take considerably longer to reconstruct the exact amino acid sequence of all human proteins from these data, by a combination of in silico gene prediction and continued mRNA sequencing. Another bottleneck will be the computer-based prediction of hypothetical functions for all human genes, which could greatly accelerate progress in human physiology and drug target discovery. Over the last few years, the two Lausanne-based groups of the Swiss Institute of Bioinformatics (SIB) have developed a number of sequence analysis tools (e.g., ProfileScan and ESTScan) that may help to make sense out of raw genome data.

2 Method and Results

There are few places that offer a comprehensive data collection and related set of tools allowing researchers to tackle the problems of gene prediction and discovery in an integrated manner [4]. To make this task easier, we make available to the scientific community at large a set of tools that address these needs via our ftp and www servers `www.isrec.isb-sib.ch`.

2.1 Available data collections

A comprehensive non-redundant protein collection, assembled from the SwissPROT, TrEMBL, GenPep, PIR, worm, and yeast collections. It is built using a 64-bit CRC method, where only one representative entry for all exactly matching sequences is kept. If two sequences differ, even by one amino acid, both are kept. Access to the annotation part of all entries is possible via a CRC-based retrieval mechanism.

The SIB-contigs, which are an attempt to produce assembled mRNA sequences, starting from UniGene clusters and raw EST data. The contigs are assembled by a mixture of Phrap, CAP, and ad-hoc scripts. We have contigs for human, mouse, rat, and drosophila EST sequences.

The translation of the SIB-contigs into putative proteins (TrEST). Once assembled, the contigs are analyzed by ESTScan (see below) for CDS prediction and frameshift correction, and the predicted CDS are translated into protein sequences. The TrEST database is a rich (but still error-prone) source of yet undocumented expressed proteins.

The coding regions present in the trGen database are predicted from human genomic sequences larger than 10,000 bp present in the HUM and HTG sections of the EMBL database. After vector masking, these are analyzed using Genscan [1], and the putative genes are translated into proteins.

The radiation hybrid maps are a useful complement to genetic and clone maps, as they can include non-polymorphic markers and are also powerful enough to order unresolved genetic clusters of polymorphic STSs. We maintain a Blast-searchable database of EMBL sequences for which mapping data are available in the Radiation hybrid database at the EBI (`http://www.ebi.ac.uk/RHdb/`).

Randomized or shuffled databases are tools to evaluate the statistical significance of match scores reported by sequence similarity, HMM, and profile search algorithms. We offer a collection of different types of randomized DNA and protein sequence databases of different sizes, to be used for score calibrations and assessment of statistical parameters.

Finally, we extract a set of complete 16S ribosomal RNA sequences from the prokaryotic division of the EMBL database, using a customized profile search method. This can be used for bacterial identification as well as for phylogenetic studies.

2.2 Available online tools

The following tools are available from our software page: `http://www.isrec.isb-sib.ch/software/`.

A patched version of BLAST which offers finer control over some parameters like, e.g., zero cost gap extension, and specification of statistical parameters. The WWW interface offers a number of features not available at other sites, such as the possibility to mask repetitive elements in a DNA sequence on the fly using a fast method based on Claverie's xblast program.

Hardware-accelerated Smith-Waterman, HMM and profile searches. Profiles are computer-readable descriptions of sequence motifs that are shared by a group of related sequences or sequence regions (domains). The Pftools package implements the generalized profile [3] method developed by us, which is closely related to the type of hidden Markov model commonly applied in molecular sequence analysis. The source code distribution available from our ftp site includes programs to construct profiles, to search sequence databases with profiles, to scan a profile library with a sequence, and to perform frame-shift tolerant searches of a DNA database with a protein profile. The WWW server allows the scanning of single protein or DNA sequences for the occurrence of motifs defined in public pattern, profile, and HMM collections.

The CpGpred tool displays CpG-islands, isochores, and some other statistical properties of large genomic DNA sequences according to user-defined parameters.

Hits is a database devoted to protein domains. It is also a collection of tools for the investigation of the relationships between protein sequences and motifs. These motifs are defined by a heterogeneous collection of predictors, which currently include regular expressions, generalized profiles and hidden Markov models.

The program ESTScan [2] was written to recognize the coding frame in ESTs, and a fortiori in assembled contigs. ESTScan detects most frame shifts, and predicts their position with an error of a few amino acids.

References

[1] Burge, C. B., and Karlin, S., Finding the genes in genomic DNA, *Current Opinion in Structural Biology*, 8(3):346–54, 1998.

[2] Iseli, C., Jongeneel, C.V., and Bucher, P., ESTScan: A program for detecting, evaluating, and reconstructing potential coding regions in EST sequences, *Proc. 7th ISMB*, 138–148, 1999.

[3] Bucher, P., Karplus, K., Moeri, N., and Hofmann, K., A flexible motif search technique based on generalized profiles, *Comput. Chem.*, 20:3–24, 1996.

[4] Jongeneel, C.V., Searching the EST databases: panning for genes, *Briefings in Bioinformatics*, in press, 2000.