Novel Gene Discovery with Sequence Profile Comparison

Weizhong Li[1], Lukasz Jaroszewski[2], Adam Godzik[3]

Keywords: genome annotation, sequence profiles, sequence homology

1 Introduction.

Despite a human genome draft being available for over two years, the number of protein coding genes is still a matter of debate and novel genes are continuously being found. Homology based methods, which predict new genes by comparing an entire genome with known genes, have played an important role [1]. These methods can find real genes that may be systematically missed by other methods such as those that employ *ab initio* gene prediction. However, homology based methods also give rise to huge amount of false positive and pseudo-genes.

In the past years, we have been developing sensitive algorithms to efficiently detect more remote homologies and that implement comprehensive approaches to validate predictions [2-5]. Our key techniques are sequence and profile alignment tools and sequence clustering methods. Here, we present a novel gene discovery system that incorporates with these protein sequence and genome analysis tools. This system has been applied in identifying several novel proteins families, some of which represent potential new drug targets.

2 Methods.

In general, homology based approaches use known genes as queries to search against a genome to identify fragments that may encode genes. Typically, a very large number of fragments with statistically significant similarities to known genes could be found, which in itself is an interesting and not completely understood phenomenon. These fragments need to be ranked in order to produce a reasonable number of fragments that can advance to more detailed analysis such as gene assembly.

Our gene discovery system incorporates several processes: a) protein sampling, b) protein clustering and analysis, c) profile building, d) profile-sequence and profile-profile search against genome, e) fragment characterization, f) backward profile-sequence and profile-profile search, g) fragment filtering, prioritizing and validating, h) full gene assembly, and i) experimental validation. A list of algorithms, programs, and resources employed in this system are listed in table 1.

Given one or more known proteins the system a) first searches the **NR+** database with an intermediate sequence or profile search to retrieve all the close and remote homologues of this protein family; b) selects representative sequences of the family (depending upon diversity within the family) by applying an appropriate clustering tool such as **PSI-Clstr.** These representative

[1] Quorex Pharmaceuticals, Carlsbad, California, E-mail: wli@quorex.com
[2] San Diego Supercomputer Center, La Jolla, California, E-mail: lukasz@sdsc.edu
[3] The Burnham Institute, La Jolla, California, E-mail: adam@burnham.org

sequences are then used to retrieve a conserved sequence pattern from the multiple sequence alignment; c) then builds sequence profile for each representative protein from **rep-NR+**; d) performs profile-sequence and profile-profile search against **PCF** or **PCF-Profile** to collect genomic fragments; e) calculates a broad array of properties of fragments, such as sequence complexity, exon probability, relative positions to known genes and **EST**s, match score of specific sequence pattern. This calculation incorporates publicly available genome annotation databases; f) annotates the fragments by comparing them against known protein profile databases such as **PDB-Profile** and **Pfam-Profile**; g) filters out random hits and hits corresponding to known genes, and it also ranks fragments according to all available calculated results. The more detailed analyses are performed to the top ranking fragments; h) full length genes are assembled; i) fragments and assembled gene are tested experimentally.

CD-HIT	Protein sequence clustering algorithm with very high speed to handle medium to high homology on very huge database
BLAST-Clstr	BLAST-based clustering algorithm which can handle medium to low homology
PSI-Clstr	PSI-BLAST-based clustering algorithm which can handle very remote homology
Saturated-BLAST	Intermediate sequence and profile search algorithm, which can effectively explore diverse protein family. It also offers multiple sequence alignment and clustering.
FFAS	Sensitive profile-profile alignment algorithm
NR	Public protein database from NCBI
NR+	NR plus other proteins predicted or annotated from some specific genomes
REP-NR+	Representative protein families from NR+ prepared with CD-HIT
Genome	Complete human genome sequence
Annotation	Publicly available genome annotations such as NCBI genome, Ensembl and Goldenpath
PCF	Putative coding fragments, translated peptides from human genome
PCF-Profile	Sequence profiles for PCF
EST	Public EST database
PDB-Profile	Sequence profiles for PDB sequences
Pfam-Profile	Sequence profiles for Pfam database
Other	Sequence profiles calculated from other public databases

Table 1: A list of algorithms and resources.

3 Applications.

We have used this system to identify several novel protein families such as putative apoptotic proteins and kinases.

References

[1] Li, W. and Godzik A. 2002. Discovering new genes with advanced homology detection. Trends in Biotechnology. 20:315-316.

[2] Li, W., Jaroszewski, L. and Godzik A. 2002. Database clustering strategies improve PSI-BLAST remote homology recognition while cutting down on search time. Protein Engineering, 15:643.

[3] Li, W., Jaroszewski, L. and Godzik A. 2002. Tolerating some redundancy significantly speeds up clustering of large protein databases. Bioinformatics, 18:77-82

[4] Li, W., Pio, F., Pawlowski, K and Godzik A. 2000. Saturated BLAST: An automated multiple intermediate sequence search used to detect distant homology. Bioinformatics, 16:1105-1110

[5] Rychlewski, L., Jaroszewski, L., Li, W. and Godzik A. 2000. Comparison of sequence profiles. Strategies for structural predictions using sequence information. Protein Science, 9:232-241