34. Mining the Human Genome Using Protein Structure Homology

Randal R. Ketchem[12], Bruce A. Mosley[3], Scott L. Taylor[4], Steven R. Wiley[5], William C. Fanslow[6], Laurent Galibert[7], Peter R. Baum[8]

Keywords: gene mining, remote protein homology, four helical cytokine

1 Introduction

Remote homology detection (threading) using GeneFold [1] has been used to mine the human genome [2, 3] for novel proteins that are potential members of families of interest. Many protein sequences are of unknown function, and many of these are not easily classified by sequence comparison methods. We describe here a method for identifying proteins belonging to specific families, give an example of a group of proteins discovered using this method, and discuss their experimental validation.

2 Mining Method

Representative sequences from several known protein families were first run through GeneFold in order to obtain a list of hits characteristic for each family. This allowed for a clustering of the GeneFold hits from sequences of unknown function into possible known protein families. While threading produces a large number of false positives, analysis of the data was facilitated by linking each hit to its cluster in an internal database of known proteins (GeneBase). Thus, sequences that hit a structure template which is itself hit by the family of interest are displayed together, sorted by their GeneFold score and annotated as to their family classification, if known.

3 Results

Several possible four helical cytokines were identified using this method. One of these, which we will refer to here as IL-29 (IMX129840-1), when used as the basis for BLAST detected additional related potential exons tightly linked in the genome to IL-29. Electronic assembly predicted two additional full length molecules in both human and mouse (IL-28A (IMX129840-3) and IL-28B (IMX129840-2)) highly related to IL-29, as well as potential exons in the human genome downstream of IL-28A that could encode for a different C-terminal portion of an IL-28 molecule (IMX129840-4). GeneFold analysis suggested that the two additional full length molecules also were potential 4 helical bundle cytokines. The exon structure of the predicted molecules was confirmed by RT-PCR and the 5' and 3' determined by RACE PCR. Expression of the molecules by cloning and transfection was used to verify they are truly secreted proteins. Real Time PCR was

[1] Amgen Inc, 51 University Street, Seattle, WA 98101
[2] Department of Bioinformatics, ketchemr@amgen.com
[3] Department of Molecular Biology, mosleyba@amgen.com
[4] Department of Bioinformatics, taylorsc@amgen.com
[5] Department of Cancer Biology, wileys@amgen.com
[6] Department of Cancer Biology, fanslowb@amgen.com
[7] Department of Inflammation, glaurent@amgen.com
[8] Department of Autoimmunity, baump@amgen.com

used to determine expression of the molecule in tissues and cell types. We also will present preliminary results examining IL-28 activity.

References

[1] Jaroszewski, L., Rychlewski, L., Zhang, B. and Godzik, A. 1998. Fold Predictions by a Hierarchy of Sequence, Threading and Modeling Methods. *Protein Science* 7:1431-1440.

[2] Celera release r25, Celera Genomics, 45 West Gude Drive, Rockville, MD 20850.

[3] Venter J.C. et al. 2001. The sequence of the human genome. *Science* 291:5507 1304-51.