Metabolic Engineering 3, 285–288 (2001)
doi:10.1006/mben.2001.0195, available online at http://www.idealibrary.com on IDEAL®

WEB SITE REVIEW

Review of the TEIRESIAS-Based Tools of the IBM Bioinformatics and Pattern Discovery Group

Anthony P. Burgard, Gregory L. Moore, and Costas D. Maranas[1]

Department of Chemical Engineering, The Pennsylvania State University, University Park, Pennsylvania 16802

Received July 25, 2001; accepted July 27, 2001

INTRODUCTION

The use of pattern discovery techniques is becoming widespread in solving problems involving text mining, protein structure characterization and prediction, promoter signal detection, gene expression analysis, and others (Rigoutsos *et al.,* 2000). The Tools of the IBM Bioinformatics and Pattern Discovery Group provide an interface for the application of the pattern discovery algorithm TEIRESIAS (Rigoutsos and Floratos, 1998) to various problems in computational biology. Pattern discovery techniques identify "interesting" patterns of events (e.g., amino acids, nucleotides, gene expression levels, etc.) that appear a number of times above a threshold in a particular set of data (Rigoutsos *et al.,* 2000). Interesting patterns include one or more consecutive events or two or more events separated by an arbitrary number of events or wild-card characters. For example, the sequence AT..G matches ATCAGCAC and GCATTGGC at positions 1 and 3, respectively, where the "." symbol represents a wild-card character. The tools are accessed via http://cbcsrv.watson.ibm.com/Tspd.html and link to a variety of utilities including protein annotation with the Bio-Dictionary, gene expression analysis, sequence pattern discovery, and multiple sequence alignment through the MUSCA algorithm (Parida *et al.,* 1998). In addition, links are provided to Bio-Dictionaries of 17 genomes, Bio-Dictionary-based annotations of 12 genomes, tools that allow cross-genome searches for proteins of specific function (e.g., calcium-binding, dehydrogenase) or which contain specific features (e.g., phosphorylation site, hydrogen-bond donor site), text mining tools, association discovery tools, etc. In this Web site review, we focus on the protein annotation and gene expression tools. The definitions and default values of the user-defined parameters for the various applications are summarized in Table 1.

[1] To whom correspondence and reprint requests should be addressed. Fax: (814) 865-7846. E-mail: costas@psu.edu.

PROTEIN ANNOTATION WITH THE BIO-DICTIONARY

A central task in pattern discovery studies is the identification of patterns that are necessary to deduce membership of a sequence in a protein family. Protein annotation is valuable in many different studies spanning a range of metabolic reconstructions of newly sequenced genomes to the assessment of paternal diversity in directed evolution experiments. Rigoutsos *et al.* (1999) analyzed the entire GenPept database with the TEIRESIAS algorithm generating a collection of patterns named the Bio-Dictionary. The patterns identified cover almost all of the processed sequence space database and include both intra- and interfamily signals. Since then, the Bio-Dictionary is computed every few months by processing the Swiss-Prot/TrEMBL database (Bairoch and Apweiler, 2000). The input to the Protein Annotation Web site is a single amino acid sequence in FASTA format. After sequence input, the query is searched for Bio-Dictionary patterns. Each pattern has been annotated based on publicly available information including both family [e.g., ProSite (Hofmann *et al.,* 1999)] and structural [e.g., Protein Data Base (Berman *et al.,* 2000)] databases. For example, a pattern could be a predicate for family membership, a metal-binding domain, or a structural feature such as a helix or turn. Two Web browser frames are produced as output. The top "similarity" frame reports the protein families that match the query in order of descending score, and clicking on each classification

1096-7176/01 $35.00

Metabolic Engineering 3, 285–288 (2001)
doi:10.1006/mben.2001.0195

TABLE 1

Description of Parameters (http://cbcsrv.watson.ibm.com/Tutorial/helps.html)

Option	Description	Default
Maximum brackets	Maximum number of brackets in a pattern when carrying out equivalency-based pattern discovery	100
L	Minimum number of literals (i.e., non-wild-card characters) in a pattern	—
W	Maximum extent spanned by any *L* consecutive literals in a pattern	—
K	Desired minimum support for a pattern (i.e., find all patterns appearing at least *K* times)	2
Q	Desired maximum support for a pattern	2, 147, 483, 647
Bins[a]	Allows control of the pattern discovery resolution	30
Min positions–similarity	Minimum number of positions shared between the query and another sequence in the SwissProt/TrEMBL database	15
Threshold values[b]	Minimum required support for a location	2
Max reported results[b]	Maximum number of reported results	500
X-axis magnification[b]	Number of pixels that represent an amino acid position in the final plots	1

[a] Available only with the gene expression analysis tool.
[b] Available only with the protein annotation tool.

provides a plot (e.g., Fig. 1) showing the score of each amino acid position based on its similarity to the family definition.

A click on the link below the plot issues a SRS query to the Expasy (www.expasy.ch) server in Switzerland to output a list of the proteins that are members of a family; each of these results is linked to its respective Swiss-Prot/TrEMBL entry at http://www.expasy.com. The bottom "features" frame lists features that have been identified in the processed query, including active sites, binding sites, post-translationally modified sites, signals, various domains, etc.

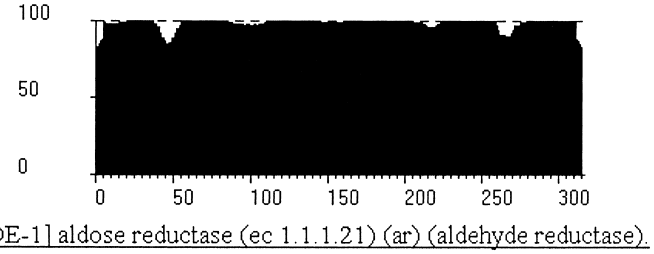
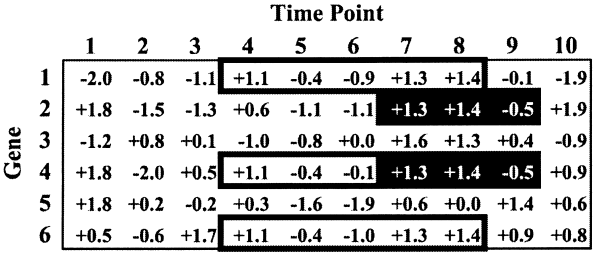


FIG. 1. Highest scoring "similarity" match for *H. sapiens* aldose reductase (default options).



Parameters: *L* = 3, *W* = 4, *K* = 2

Two patterns found:

(i) **[0.99,1.12][-0.44,-0.31].[1.25,1.38][1.38,1.51] in genes 1,4,6**

(ii) **[1.25,1.38][1.38,1.51][-0.57,-0.44] in genes 2,4**

Plot of genes 2 and 4:

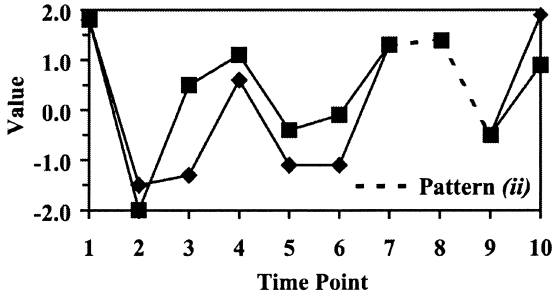


FIG. 2. Example of a gene expression analysis tool (default options).

Metabolic Engineering **3**, 285–288 (2001)
doi:10.1006/mben.2001.0195

GENE EXPRESSION PATTERN DISCOVERY

Another major application of TEIRESIAS accessed via the IBM Web site is the analysis of gene expression data. These data are often the log-transformed mRNA concentration ratios between the cell line of interest and a reference sample. Thus, inductions and repressions of identical magnitude yield values with opposite signs, while unaffected expression levels generate log ratios with a value of zero. Input is often a M by N matrix where the (i, j)th entry is the level of expression of the ith gene in the jth examined species, experimental condition, or time point. Associations between subsets of genes with subsets of conditions are typically sought either to suggest a relationship between genotype and phenotype or to distinguish cell types and biological states.

An example of the application of TEIRESIAS to a (6×10) matrix is shown in Fig. 2. The output is a list of patterns composed of multiple numeric intervals. Highlighting a specific pattern and then clicking on the Sequences button lists the input with the pattern underlined for easy identification, while the Plot button provides a graph of the expression ratios of each gene of the pattern over j time points or conditions. This Web tool also provides three possible options—*Use Derivatives, Smooth Input,* and *Inverse Regulation*—for transforming the raw input before analysis. *Use Derivatives* essentially assigns a "+," "−," or "=" symbol to each numerical input depending on whether the value increases, decreases, or stays the same with respect to the previous jth condition or time point. *Smooth Input* allows the smoothing of the numerical input by running an averaging filter before applying the pattern discovery. Finally, *Inverse Regulation* doubles the input set size (one component of original input and the second of original input with switched signs) enabling the identification of genes that are oppositely regulated over a particular interval.

STRENGTHS AND WEAKNESSES

The key strength of the Tools of the IBM Bioinformatics and Pattern Discovery Group is the application of the powerful yet versatile pattern discovery algorithm TEIRESIAS toward a variety of bioinformatics problems. TEIRESIAS is more efficient than many previously proposed algorithms (Brazma *et al.,* 1995) in that it avoids complete enumeration of the sequence space and provides results that are as specific as possible. In other words, if the pattern AT.G appears a threshold number of times in the input, reporting A..G is redundant and is avoided by TEIRESIAS. Other strengths stem from the advantages that pattern discovery has over earlier approaches. For example, protein annotation is typically conducted by constructing a multiple sequence alignment of assumed family members that is searched for conserved residues. This method has resulted in databases such as PRINTS (Attwood *et al.,* 1998), BLOCKS (Henikoff *et al.,* 1999), PROSITE (Hofmann *et al.,* 1999), and Pfam (Bateman *et al.,* 2000). However, by assigning proteins to a family before analysis, this approach only identifies intrafamily patterns, a problem overcome by applying pattern discovery over an entire database. In addition, clustering-based frameworks applied previously to expression data analysis assign a given gene to a single cluster and tend to form clusters of genes whose expression profiles are globally similar (e.g., agree across all or most of the sampled conditions or time steps). On the other hand, pattern discovery with TEIRESIAS extends clustering-based methodologies since it permits a given gene to participate in more than one cluster and enables the discovery of situations where a similar behavior spans only a limited time interval (Rigoutsos *et al.,* 2000). One potential drawback of the IBM Web site is that the scoring system and actual scores utilized to depict family similarity (Fig. 1) are not yet accessible. However, the scores are informative qualitatively and the scoring system will be addressed in an upcoming paper (Rigoutsos, personal communication). In summary, the IBM Tools provide an excellent platform for utilizing TEIRESIAS to address a surprisingly diverse set of bioinformatics challenges.

REFERENCES

Abola, E., Sussman, J., Prilusky, J., and Manning, N. (1997). Protein data bank archives of three-dimensional macromolecular structures. *Methods Enzymol.* **277**, 556–571.

Attwood, T., Beck, M., Flower, D., Scordis, P., and Selley, J. (1998). The PRINTS protein fingerprint database in its fifth year. *Nucleic Acids Res.* **26**(1), 304–308.

Bairoch, A., and Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28**, 45–48.

Bateman, A., Birney, E., Durbin, R., Eddy, S., Howe, K., and Sonnhammer, E. (2000). The Pfam protein families database. *Nucleic Acids Res.* **28**(1), 263–266.

Brazma, A., Jonassen, I., Eidhammer, I., and Gilbert, D. (1995). "Approaches to the Automatic Discovery of Patterns in Biosequences," technical report, Department of Informatics, Univ. Bergen, Norway.

Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I., and Bourne, P. (2000). The protein data bank. *Nucleic Acids Res.* **28**(1), 235–242.

Henikoff, S., Henikoff, J., and Pietrokovski, S. (1999). Blocks+: A non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics* **15**(6), 471–479.

Hofmann, K., Bucher, P., Falquet, L., and Bairoch, A., (1999). The PROSITE database, its status in 1999. *Nucleic Acids Res.* **27**(1), 215–219.

Parida, L., Floratos, A., and Rigoutsos, I. (1998). MUSCA: An algorithm for constrained alignment of multiple data sequences. *In* "Proceedings of the 9th Workshop on Genome Informatics," Tokyo, Japan.

Rigoutsos, I., and Floratos, A. (1998). Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm. *Bioinformatics* **14**(1), 55–67.

Rigoutsos, I., Floratos, A., Ouzounis, C., Gao, Y., and Parida, L. (1999). Dictionary building via unsupervised hierarchical motif discovery in the sequence space of natural proteins. *Proteins Struct. Funct. Genet.* **37**(2).

Rigoutsos, I., Floratos, A., Parida, L., Gao, Y., and Platt, D. (2000). The emergence of pattern discovery techniques in computational biology. *Metab. Eng.* **2**, 159–177.