The web server of IBM’s Bioinformatics and Pattern Discovery group

Tien Huynh, Isidore Rigoutsos*, Laxmi Parida, Daniel Platt and Tetsuo Shibuya1

Bioinformatics and Pattern Discovery Group, IBM TJ Watson Research Center, PO BOX 218, Yorktown Heights, NY 10598, USA and 1Exploratory Technology, IBM Tokyo Research Laboratory, 1623-14, Shimotsuruma, Yamato-shi, Kanagawa 242-8502, Japan

Received February 13, 2003; Revised and Accepted April 8, 2003

ABSTRACT

We herein present and discuss the services and content which are available on the web server of IBM’s Bioinformatics and Pattern Discovery group. The server is operational around the clock and provides access to a variety of methods that have been published by the group’s members and collaborators. The available tools correspond to applications ranging from the discovery of patterns in streams of events and the computation of multiple sequence alignments, to the discovery of genes in nucleic acid sequences and the interactive annotation of amino acid sequences. Additionally, annotations for more than 70 archaeal, bacterial, eukaryotic and viral genomes are available on-line and can be searched interactively. The tools and code bundles can be accessed beginning at http://cbcsrv.watson.ibm.com/Tspd.html whereas the genomics annotations are available at http://cbcsrv.watson.ibm.com/Annotations/.

INTRODUCTION

‘Pattern discovery’ or ‘data mining’ methods have emerged as an alternative and effective approach for tackling a variety of biological problems. Although these methods have been used in the context of traditional computer science activities for more than a decade, their use in analyzing biological data began only recently.

For the last several years, the focus of the research activity in the Bioinformatics and Pattern Discovery group at IBM’s TJ Watson Research Center has been on theoretical and applied aspects of pattern discovery with an emphasis on tackling problems from molecular biology. During this time, we have designed and published a number of methods for generic pattern discovery and applications (1–6), multiple sequence alignment (7,8), gene discovery (9), protein annotation (5,10–13), comparative molecular moment analysis (14,15) etc. Implementations for many of these methods have been web-enabled and are accessible through the group’s web server. Executable code bundles for several operating system and processor combinations can also be downloaded from the same server. Finally, we make available content in the form of annotations for >70 complete genomes; provided graphical user interfaces and searching software facilitate the interactive study of these annotations.

In summary, the tools that are available on our server permit the discovery of patterns in the following types of datasets: amino acid sequences, nucleic acid sequences, integer streams, gene expression data (time series as well as association experiments), natural text, invariant representations of molecular shape and electrostatics, and others. The tools can also be used to generate multiple sequence alignments of amino acid, nucleic acid or generic textual input sets, to discover genes in nucleic acid sequences, and to interactively annotate amino acid sequences. In what follows, we briefly describe the various services and genomic content found on our web server; extensive information for each one of the tools, detailed help files and a full-length tutorial are available on-line.

WHAT IS AVAILABLE ON THE SERVER

The desired tool can be selected by clicking on the tab with the relevant header (Fig. 1). As seen in Figure 1, the typical tool page will include the following logical sections: ‘parameters’, ‘options’, ‘input sequences’, ‘references’, and ‘other relevant links’, the latter appearing always on the left side of each page. The parameters are multi-valued entities and are used to control the actual computation. The options, on the other hand, are binary in nature and are used to enable/disable various filtering steps, to modify attributes of the computation or to affect the format of the output. The references section enumerates those of the group’s publications which describe the corresponding tool. The ‘other relevant links’ section includes pointers to tool-specific pages (e.g. ‘About this Engine’ links to a detailed explanation of the tool, the underlying algorithm, the tool’s capabilities, the input and output formats, recommended values for the various parameters, etc.), to server-specific pages (e.g. ‘News’, ‘License Terms’, ‘Brief/Full Tutorial’), and to pages with general information. Notably, on each tool page we provide email links

*To whom correspondence should be addressed. Tel: +1 9149451384; Fax: +1 914 945 4104; Email: rigoutso@us.ibm.com

Nucleic Acids Research, Vol. 31, No. 13 © Oxford University Press 2003; all rights reserved
that permit the users to send us questions, comments and suggestions pertaining to the tools and the server’s operation.

Pull-down menus provide access to sample input datasets and common settings whereas the options and parameters are automatically initialized to default values. The user’s interaction with the server is facilitated by real-time checking of the various settings; this in turn enables the identification of logical and other conflicts and the prompting of the user with the appropriate suggestion. Context-sensitive help is also available on each tool page.

Computational requests are given 10 min of wall-clock processing time, an interval that is more than sufficient for handling a very large spectrum of problems and problem sizes. Computations that exceed this time limit are terminated and the user is notified to this effect. Note that the server will not prevent a user from issuing several simultaneous requests from multiple browser windows: each request will be given its own 10-min processing time limit. For security reasons, all of the submitted input and generated output files are deleted automatically 30 min after their creation.

Users can download executable codes for a variety of operating systems and processor combinations from http://cbcsrv.watson.ibm.com/download.phtml.html; note that our code updates closely follow the web enablement and publica-
tions of each new method. The downloadable codes and provided web services can be used free of charge by users that are not-for-profit work (e.g. academic or personal research); a license is required for use in commercial, for-profit environments and additional details on the terms and conditions can be found on our web site.

We next describe briefly each of the available tools using the same headers as the tabs which appear in Figure 1; the corresponding URL is also listed.

**Sequence pattern discovery** (http://cbcsrv.watson.ibm.com/Tspd.html). This is our implementation of Teiresias, a two-phase, combinatorial algorithm for general purpose pattern discovery (1–3). The user can run Teiresias in either ‘exact discovery’ or in ‘equivalence’ mode. In the latter mode, the user must provide groups of characters which can replace one another (=‘character equivalences’) and are potentially overlapping. Commonly used equivalences (e.g. ‘chemical’ and ‘structural’) are available through a pull-down menu; for non-amino acid inputs the user can define equivalences by simply typing the character groups in the provided window. The input set can be amino/nucleic acid sequences and general character streams composed of printable ASCII characters.

The discovered patterns are reported in groups of 100 at a time and a mechanism is provided for the user to navigate through this output. Individual patterns can be selected and their highlighted instances displayed in the processed input. Filtering tools based on pattern properties such as minimum/maximum support, minimum/maximum density, log-probability estimates, etc., are also available on the output page together with a link providing access to the complete set of discovered patterns. For amino acid inputs, individual patterns can be searched within SWISS-PROT/TrEMBL: the output is a list of strings showing the phylogenetic domain, database id, accession number, function and organism name for all the database entries containing an instance of the pattern under consideration — each output string is a dynamic link to the SWISS-PROT/TrEMBL record of the respective sequence. A working session can be seen in Figure 2.

**Multiple sequence alignment** (http://cbcsrv.watson.ibm.com/Tmsa.html). This tool represents our implementation of the MUSCA algorithm which allows the user to align multiple streams of letters so as to reveal salient features that are potentially present in the considered input (7,8). The user-controlled options and parameters are the same as in the previous tool.
with the addition of optional hydropathy coloring of the generated output.

Gene expression analysis (time series) (http://cbcsrv.watson.ibm.com/Tgea.html). With this tool one can analyze datasets that track the induction/repression of genes over time and in response to environmental changes. The input is given as a $N \times M$ matrix of space-separated real numbers each representing log-expression ratios. Teiresias is used to discover all relationships which simultaneously involve subsets of the rows and subsets of the columns in this matrix. This is a special case of the generic problem of association discovery and a discussion on the importance of discovering such patterns appears in the literature (5). This tool can carry out discovery on the raw expression ratios and on the signs of the derivatives of the expression ratios (see also 5). The degree of quantization that is applied to the input data can be changed interactively. Also, for each discovered pattern, the user can generate plots of the corresponding input rows or highlight a pattern’s instances in the original input (Fig. 3). Currently, only patterns that are synchronized in time (lock-step) are reported; the tool neither reports nor seeks patterns corresponding to cascaded signals.

Gene expression analysis (association data)/Association discovery (http://cbcsrv.watson.ibm.com/Tad.html). Unlike the previous tool which is appropriate for time series inputs, this tool is used to process more general inputs representing association data; a $N \times M$ matrix of space-separated integers is now expected at the input. For example, the rows of this matrix could correspond to plants or tissues, the columns to the genes of interest, and the $(i, j)$ cell to the expression level of gene $j$ in tissue/plant $i$. The tool outputs the associations which are present in the input, i.e. the correlations of subsets of rows that are deduced from the presence of similar data values among subsets of columns. The reported associations are provably maximal in composition and length, i.e. all of the columns that are involved in defining a cluster will be identified and reported (see 1–3 for a discussion on the concept of maximality and 5 for a discussion of this problem’s computational difficulty).

The general association discovery problem has also numerous instances outside the biological context. Indeed, the entries of the $N \times M$ input set can be a mix of categorical and numerical values corresponding to a multitude of units. Our desire to make this tool as flexible as possible for more users and the impossibility to anticipate all instances of such a diverse spectrum of inputs have led us to expect that the user will remap the original data points to a set of integers in a manner which best reflects his/her understanding of the data source(s).

Comparative molecular moment analysis (CoMMA) (http://cbcsrv.watson.ibm.com/Tco.html). CoMMA utilizes information from moments of molecular mass and moment expansions of electrostatic potentials up to and inclusive of second order, to generate molecular descriptions that remain invariant under Euclidean transformations (14,15). QSAR studies have shown that CoMMA descriptors do predict activity and that molecules with similar indices can be expected to behave similarly even if their topological structures are somewhat different. CoMMA descriptors provide a succinct representation of the three-dimensional (3D) distribution of mass, shape and charge without any need for 3D-alignment and molecular superposition; the descriptors can thus be used in molecular comparisons, an essential step in computer-aided drug discovery.

This tool can process inputs that contain one or more molecules in the MOL2 format (Mol2 File Format, TRIPOS Associates Inc., http://www.tripos.com/custResources/mol2Files/mol2.pdf) and generates a set of CoMMA descriptors for each molecule. Charges must have been computed for each molecule. When multiple input molecules are provided, the user can also run Teiresias on the CoMMA descriptors to determine associations among the various molecules as these are captured by the commonalities among subsets of the molecules’ CoMMA representations.

Integer pattern discovery (http://cbcsrv.watson.ibm.com/Tipd.html). This tool is our integer-based implementation of Teiresias and is meant for pattern discovery on event streams of positive integers. It is meant to tackle inputs where the underlying alphabet contains more distinct symbols than the printable ASCII set; in fact, up to $2^{31} - 1$ distinct integers can be used to form inputs. The number of problems that can be solved using this version of Teiresias is very large: virtually any data mining problem that can be cast as a pattern discovery task on streams of positive integers can be tackled by this tool. For the special case where the integers correspond to natural language words we have created the next tool.

Natural text mining—word units (http://cbcsrv.watson.ibm.com/Twwpd.html). This tool uses the integer version of Teiresias to process natural text (not necessarily English) and treats each input word as the unit of information. The provided input should be in ASCII, and in free format (e.g.
cut-and-pasted documents from a news web site). Input pre-filtering discards punctuation marks and uses carriage returns as event stream separators. No linguistic processing of any kind takes place during the reading of the input and the determination of the vocabulary: e.g. ‘go’ and ‘goes’ are treated as two distinct words. Once the vocabulary is determined, the textual input is mapped to streams of integers that is then processed by the integer-based version of Teiresias. Upon completion of the discovery phase, the integers in the patterns are replaced by the corresponding vocabulary words and word-based patterns are reported to the user.

Natural text mining—symbol units (http://cbserv.watson.ibm.com/Tspd.html). The only difference between the previous and this tool is that here it is the individual character that is treated as a unit of information. Processing a large body of natural text in this manner is expected to uncover pieces of words, whole words, and small phrases all of which are valid for the natural language that the text represents (5).

Protein annotation (http://cbserv.watson.ibm.com/Tpa.html). This tool allows the user to automatically annotate a given amino acid sequence using our Bio-Dictionary-based approach (11,13). The tool determines and reports local and global similarities between the query and the contents of the SWISS-PROT/TrEMBL database (16), phylogenetic domain membership as a function of position within the query, the location and nature of domains, active sites, post-translationally modified sites, PUBMED references which are relevant to the query, etc.

The output page currently contains three frames (Fig. 4). The top frame contains a link to the set of patterns that were found in the processed query sequence and were used to annotate it; also included are plots showing the likeness of the query to archaeal, bacterial, eukaryotic, viral and unclassified sequences as a function of amino acid position. The middle frame shows local and global similarities which are shared by the query and other entries in the database from which the Bio-Dictionary was built. Also reported in this frame are relevant PUBMED references; with the help of a text-clustering tool, a hierarchy of phrases contained in the title/abstract of these PUBMED references can be generated simplifying the analysis (17,18).

Finally, the bottom frame contains plots that show the type and location of features which have been identified in the annotated query; these features include binding domains, transmembrane domains, active sites, signals, etc. The plots’ y-axis indicates the degree of confidence in what the plot describes, and all plots are ordered in a manner that will list narrow (in terms of extent) and better-conserved regions before wide but not-so-well-conserved ones. The captions of all plots are ‘active’ and when selected a query is issued to the Expasy server’s SRS interface (http://www.expasy.org/srs5bin/cgi-bin/wgetz?) in order to determine which entries from SWISS-PROT/TrEMBL contain the string corresponding to the selected caption.

Gene identification (http://cbserv.watson.ibm.com/Tgi.html). This tool provides access to the implementation of our dictionary-based approach to finding genes in prokaryotic genomes (9). The input page provides the user with the ability to select ‘start’ codons while the output consists of a graphical interface that allows the user to zoom in/out and navigate over the processed nucleic acid sequence. Discovered genes are displayed using score-based colors and clicking on a gene will provide information about the gene’s location within the processed input sequence, its reading frame, amino acid translation, etc. The graphical interface also permits the user to annotate the amino acid translation of a selected gene using our annotation tool—a screen shot of a working session is included in Shibuya and Rigoutsos (9).

Genome annotations (http://cbserv.watson.ibm.com/Annotations/). We have used our protein annotation method to process and annotate putative protein sequences from >70 complete genomes. These annotations are accessible through our web server. Within each genome, a gene’s annotation can be retrieved either using the gene’s accession number or with the help of regular expressions (13). On-line documentation includes several examples and information on how to construct these regular expressions. Searches across genomes are also available through http://cbserv.watson.ibm.com/TsearchAllGenomes.html.

Human cytomegalovirus (HHV5) annotation (http://cbserv.watson.ibm.com/virus/). In recent work, we used two distinct methods to annotated the human cytomegalovirus (HCMV) genome. The first method employed the ProCeryon program (ProCeryon, a program for fold recognition and protein structure analysis, ProCeryon Biosciences, http://www.proceryon.com, 1999–2001) and generated a structural and functional
hypothesis for each putative gene using a threading-based scheme (19). The second, more recent effort re-annotated HCMV using our Bio-Dictionary-based approach that exploits only sequence-based information (20). Three viral strains were annotated in each study: (i) AD169 which was sequenced in 1990; (ii) Towne; and (iii) Toledo (21,22). Our results as well as natural language summaries for each putative protein have been included in a system that we designed and implemented on IBM’s DB2 database system and which can be seen in Figure 5.

**PLANNED EXTENSIONS AND UPCOMING TOOLS**

We are currently working on and will soon be making available a parallel version of Teiresias for both shared-memory and message-passing architectures; a utility for creating JPEG files from such output pages as those of the protein annotation, multiple sequence alignment and gene discovery tools; and tools for determining irredudant motifs in an input (4), finding tandem repeats in nucleic acid sequences (23), discovering flexible patterns (6) and determining the fine structure details of transmembrane helices.

**DISCUSSION**

We have described the capabilities of a Bioinformatics web server designed and maintained by our group. The server provides remote access to web-based implementations of methods that the group’s members and collaborators have published over the years and is operational around the clock. Code bundles are also available for download.

The various tools revolve around the concept of pattern discovery and provide solutions to a large variety of problems from biology and elsewhere. For each tool, detailed descriptions that include information on setting options, parameters, input formats etc. can be found on-line. Real-time controls check for errors and inconsistencies at all stages of the user’s interaction with the server while context-sensitive help is available on all pages.

Additionally, graphical user interfaces and software permit the interactive searching of our on-line annotations for >70 complete archael, bacterial and eukaryotic genomes. Finally, for the HCMV genome an IBM DB2-based system permits web access to our published analyses of three HCMV strains.

**ACKNOWLEDGEMENTS**

The authors thank Stephen Chin-Bow, Mashama McFarlane and Kyle Jensen for their contributions to the development of various aspects of the tools and graphical user interfaces.

**REFERENCES**


