BIOINFORMATICS

Vol. 20 no. 18 2004, pages 3326–3335
doi:10.1093/bioinformatics/bth392

MyWEST: My Web Extraction Software Tool for effective mining of annotations from web-based databanks

Marco Masseroli[1,*], Andrea Stella[1], Natalia Meani[2,3], Myriam Alcalay[2,3] and Francesco Pinciroli[1]

[1]Bioengineering Department, Politecnico di Milano, I-20133 Milano, Italy,
[2]IEO—European Institute of Oncology, I-20141 Milano, Italy and
[3]IFOM—FIRC Institute of Molecular Oncology, I-20139 Milano, Italy

Received on February 20, 2004; revised on June 16, 2004; accepted on June 30, 2004
Advance Access publication July 9, 2004

ABSTRACT

Motivation: High-throughput technologies create the necessity to mine large amounts of gene annotations from diverse databanks, and to integrate the resulting data. Most databanks can be interrogated only via Web, for a single gene at a time, and query results are generally available only in the HTML format. Although some databanks provide batch retrieval of data via FTP, this requires expertise and resources for locally reimplementing the databank.

Results: We developed MyWEST, a tool aimed at researchers without extensive informatics skills or resources, which exploits user-defined templates to easily mine selected annotations from different Web-interfaced databanks, and aggregates and structures results in an automatically updated database. Using microarray results from a model system of retinoic acid-induced differentiation, MyWEST effectively gathered relevant annotations from various biomolecular databanks, highlighted significant biological characteristics and supported a global approach to the understanding of complex cellular mechanisms.

Availability: MyWEST is freely available for non-profit use at http://www.medinfopoli.polimi.it/MyWEST/

Contact: masseroli@biomed.polimi.it

INTRODUCTION

Concerted efforts for deciphering the structure of many genomes have led to a growing amount of publicly available sequence data. Information describing individual genes and their encoded protein products continues to accumulate in many different databanks (Galperin, 2004), where data are usually stored in sets of text files or in relational databases. Most of these biomolecular databanks are easily accessible through heterogeneous Web interfaces but require expertise to be comprehensively queried. Few of them also provide FTP access to retrieve the entire dataset in a structured format, mainly in ASCII flat files. No public biomolecular databank using a relational database provides a direct remote access to the backend database used.

New high-throughput technologies—such as DNA microarrays, oligonucleotide arrays and serial analysis of gene expression—are generating massive datasets describing the behaviour of thousands of genes at once. At present, an important challenge is to find ways to exploit this large amount of information for understanding cellular mechanisms underlying complex phenotypes. It is, therefore, necessary to provide easily accessible bioinformatics tools capable of mining the increasing amount of biological information publicly available in biomolecular databanks, and automatically connecting gene expression data with the mined information.

Several approaches to create integrated access to the numerous available databanks have been proposed and utilized in different systems. The most popular are data warehousing [e.g. SRS (Etzold *et al*., 1996), NCBI/Entrez (Tatusova *et al*., 1999)], multi databases [e.g. BACIIS (Ben Miled *et al*., 2002), TAMBIS (Stevens *et al*., 2000), BioKleisli (Davidson *et al*., 1997)], federated databases [e.g. ISYS (Siepel *et al*., 2001), DiscoveryLink (Haas *et al*., 2001)], mediator based systems [e.g. BioDataServer (Freier *et al*., 2002)] and information linkage [e.g. GeneLynx (Lenhard *et al*., 2001), GeneCards (Rebhan *et al*., 1998), SOURCE (Diehn *et al*., 2003)]. Nevertheless, most of these require resources and expertise in order to be implemented and maintained. Others, such as GeneLynx, GeneCards and SOURCE, have been compiled as integrational databanks with a Web interface for easy querying, and for providing simple access to multiple biomolecular information resources. However, they have been created to collect and present data organized for individual nucleotide or amino acid sequences, and some of them (e.g. GeneLynx) only present a series of links to other external resources. Moreover, they are designed for human browsing and not for machine reading.

*To whom correspondence should be addressed.

Thus, these and other valuable databanks poorly adapt to biological interpretative analyses and knowledge discovery from vast datasets, which involve the comparative evaluation of multiple characteristics of many nucleotide and/or amino acid sequences at once. Hence, large dataset interpretation requires the support of new automated tools for mining and comparing the information of interest from the databanks where they are available.

For the purpose of mining information, accessing the data in the structured form inside the databank would be the best option. The FTP access that some databanks provide, however, requires local reimplementation and maintenance of the entire databank, which implies expertise and resources that only big research centres can afford. Moreover, the annotations usually used for the interpretation of high-throughput experiments reside in several different databanks, some of which do not provide access to their structured data. Thus, even locally reimplementing and maintaining many different databanks could not be enough for obtaining all available data. On the other hand, researchers generally do not need all the data present in a databank but only a specific subset. The best technological option would be accessing the data in XML format through Web services, however, at present still very few data providers offer access to their data via a Web services model.

Here, we present an automatic method for mining selected data of multiple nucleotide and/or amino acid sequences from different biomolecular databanks accessible through Web interfaces, organizing the extracted data in order to allow their integration to expression profiling results, and performing comparisons and further analyses on them. The method has been implemented in a prototype software package, called MyWEST (i.e. My Web Extraction Software Tool), made freely available to users at http://www.medinfopoli.polimi.it/MyWEST/.

SYSTEMS AND METHODS

In MyWEST, we used Java programming language to implement a mining method for automatically extracting data of interest from HTML pages of databanks, then structuring and storing the data in aggregated form. The method works as described in Figure 1:

- HTML pages containing data of interest are retrieved from biomolecular databanks accessible via Web.
- Using the Document Object Model (DOM) recommendation of the World Wide Web Consortium (http://www.w3.org/DOM/), each retrieved HTML page is parsed to separately identify data, HTML tags and other page elements such as Javascript functions or comments. Tags and data are used to represent the HTML page as a hierarchical tree structure, i.e. a data structure comprised of nodes containing either a HTML tag (tag node), or data inside the page (data node).

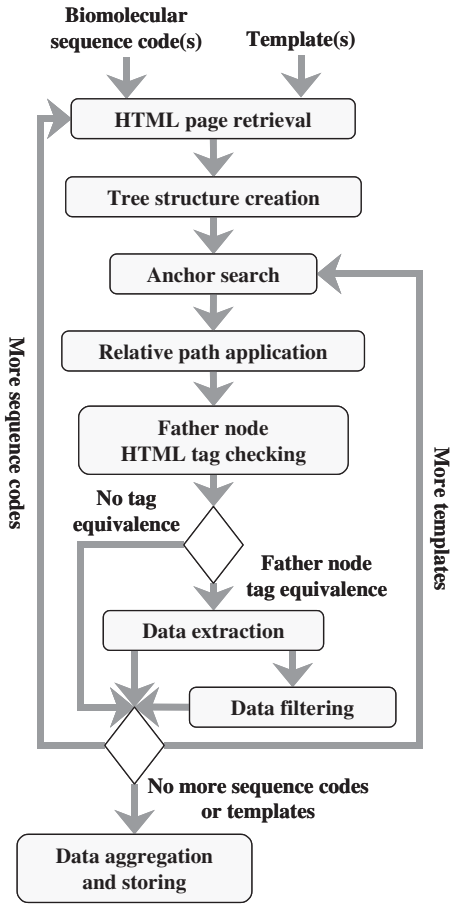

Fig. 1. Steps of the mining method for automatically extracting and aggregating data from different HTML pages of diverse biomolecular databanks.

- Using templates previously created on HTML pages with similar structure, the developed mining method automatically locates and extracts the different data of interest inside the created hierarchical tree structure of each retrieved HTML page.
- Data mined from multiple HTML pages, also of different biomolecular databanks, are structured and stored in aggregated form.

Each template provides the information for: (1) accessing the HTML pages of a biomolecular databank containing data of interest; (2) locating in a HTML page the exact position of the data of interest, identified by specific HTML tags; (3) extracting these data and those identified by the same HTML tags present in other pages with similar structure; and (4) performing filtering operations on the extracted data, if needed, and using some data to characterize the others.

Template creation and data mining

To create a template the user must provide some sequences of characters, selected on a reference HTML page, which can

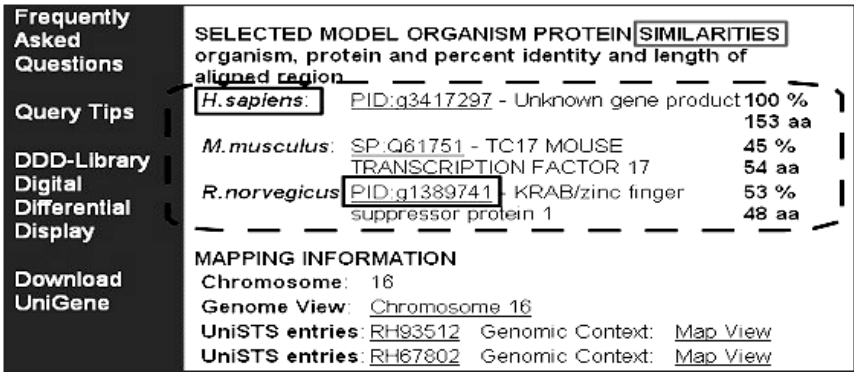

Fig. 2. Example of MyWEST extraction from the UniGene databank. A reference UniGene HTML page is where MyWEST allows visual selection of the data of interest. In the example, data of interest to extract are inside the dashed line, and the possible sequences to select for identifying the HTML tags containing the data to extract are highlighted inside solid line boxes. They are the anchor, i.e. a unique sequence of characters on the page (e.g. 'SIMILARITIES'), and other two sequences of characters, to be chosen among the data to extract (e.g. '*H.sapiens*' and 'PID:g1389741').

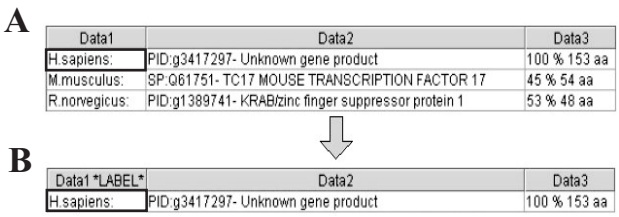

Fig. 3. Extracted data filtering: values in the 'Data1' column of the extraction result table (**A**) for the example in Figure 2 are chosen as labels and only the data with the '*H.sapiens*' label are considered (**B**).

identify the data to extract. Two template creation modalities have been defined: semiautomatic and manual.

Semiautomatic creation This modality enables the creation of templates for extracting sets of data structured as they appear formatted on a HTML page. For each dataset, the user must select three sequences of characters on a reference HTML page. The first sequence must be an anchor, i.e. a unique sequence of characters on the page. The other two sequences of characters must be selected among the data to extract. For example, in order to extract the data inside the dashed line from the UniGene databank HTML page in Figure 2, the sequence of characters 'SIMILARITIES', unique in the page, can be provided as an anchor. The other two sequences of characters can be '*H.sapiens*' and 'PID:g1389741'.

Within the tree structure representation of the reference page, the three selected sequences of characters allow to automatically identify the father tag node of the subtree structure containing the data to extract, and to automatically locate it inside the page. Thus, the relative path in the page tree structure from the anchor data node to the father node is defined and included in the created template together with the anchor and the HTML tag in the father node, used for correctness checking of the extracted data.

Extracted data filtering. When not all data in the subtree structure of the father node are of interest, a further template creation step allows the definition of two types of filtering operations on the data extracted and structured inside an extraction result table: (1) operations on the data and (2) operations on the table structure. The first allow considering the values in a column as labels, and filtering the extracted data according to selected label values only (Fig. 3). The second is useful when some table rows logically subdivide the extracted data in subtables (Fig. 4A), or when the cells in a table row contain the name of the column they belong to (Fig 4B). In the last case, the names of the table columns can be automatically assigned, better characterizing the data and enabling subsequent specific queries.

Manual creation This modality allows the generation of templates for extracting and structuring sparse data from a HTML page, independently of the way data are formatted on the page. Templates created through this modality are comprised of extraction units, one for each type of data to be extracted, representing the columns of an extraction result table. For each extraction unit, the user must select two sequences of characters on a reference HTML page. The first sequence constitutes an anchor, i.e. a label characterizing the data to extract and representing a unique landmark in the page

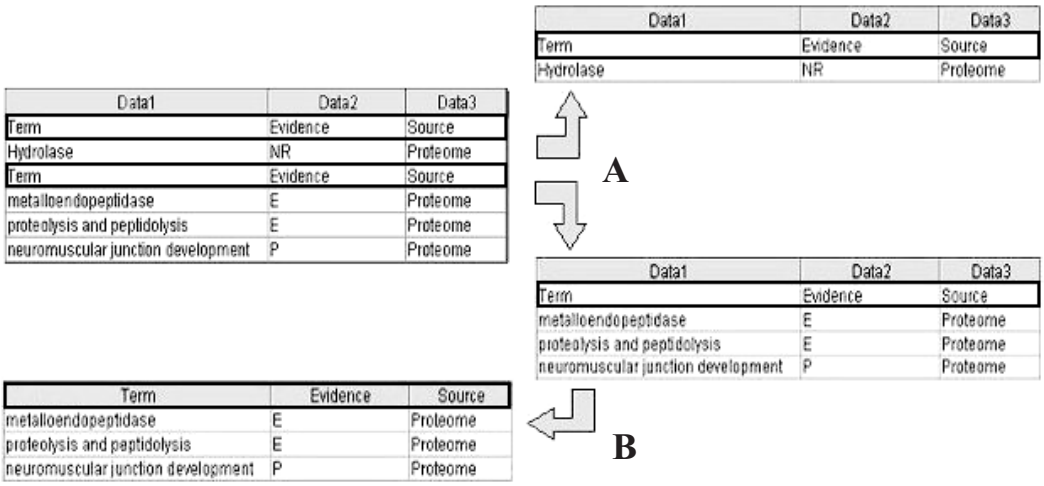

Fig. 4. Filtering on extraction result table structure. (**A**) The table is subdivided in two subtables using the rows containing the table column names as reference. (**B**) The values in the table first row are used as table column names.

for locating the data father node. The second sequence of characters represents the data of interest and enables identification of the father node within the tree structure of the reference page. Using these two sequences of characters, the relative path from the anchor data node to the father node identifying the data of interest is automatically defined, and a template is created with the same structure as those generated with the semiautomatic modality described above.

Data mining from multiple HTML pages and biomolecular databanks

To correctly apply the created templates for automatically mining data of interest from different HTML pages of the same biomolecular databank, some criteria on the structure of the databank Web pages need to be satisfied. The HTML pages must have the following features:

- the character sequence selected as anchor must be present, identical and unique in all HTML pages containing the data of interest;
- in all databank HTML pages, the HTML tag in the father node of the subtree structure containing the data to extract must not change;
- the data of interest must be univocally identified by the defined anchor, the relative path and the father node HTML tag stored inside the template to use;
- between the anchor and the father node identifying the data to extract, all databank HTML pages must have an unvaried structure, i.e. the relative path must not change.

Provided an adequate anchor is selected, the HTML pages of biomolecular databanks satisfy all these criteria. In fact, these HTML pages are dynamically and automatically created from the data contained in the database underneath and thus, all of them have a similar structure. Furthermore, the most restrictive hypothesis of invariable structure between the anchor and the father node of the data to extract is not relevant when the anchor is correctly selected inside or just beside the father node subtree structure containing the data of interest. Using different templates and anchors we performed many tests on the HTML pages of the UniGene (Schuler, 1997), LocusLink (Pruitt *et al*., 2001), Swiss-Prot (Boeckmann *et al*., 2003), SOURCE and GeneCards databanks. Only in few pages, when the anchor is selected not inside or beside the father node subtree structure containing the data to extract, changes in structure between the selected anchor and the father node can occur. In these cases, the controls implemented in the extraction algorithm (e.g. correctness checking of the HTML tag in the father node) prevent from mining irrelevant data. Moreover, when some data of interest are not found in a HTML page, a reference of that page and a possible reason for the unsuccessful extraction are noted in a log file. This gives the user the chance of revising only those HTML pages that presented problems during the automatic extraction.

Usually, each HTML page of a biomolecular databank contains all the information present in that databank about a single nucleotide or amino acid sequence, and can be retrieved using an identification code in that databank for that sequence (e.g. GenBank accession number, UniGene cluster ID, LocusLink ID and Swiss-Prot accession number). The mining method we developed can automatically extract data of interest from multiple HTML pages of a databank when the identification codes of nucleotide or amino acid sequences of interest are provided. To achieve extractions of different data from multiple HTML pages of a single or distinct

biomolecular databanks, different templates can be created and automatically applied in sequence. All mined data are aggregated and stored either in tab-delimited text files, or in a relational database, allowing both simple analyses and articulated queries on all aggregated data.

Mining validation

We evaluated MyWEST using a set of 729 clones resulting from the analysis of microarray experiments aimed at identifying genes that are differentially expressed in U937 cells after 4 h of treatment with 10^{-6} M Retinoic Acid (RA). As described below, the identified putative RA target genes were classified by mining for their annotations using MyWEST. In order to verify if the extracted data were sufficient and accurate, the same genes were also independently analysed. Literature search was performed for all known genes, and the obtained results were compared with those found through MyWEST extractions.

IMPLEMENTATION AND RESULTS

Mining method and software prototype

We created MyWEST, a new prototype software package that implements the developed method for mining information from the HTML pages of different Web-interfaced databanks and allows local aggregation and comprehensive analyses of all extracted data. The main characteristics of MyWEST are: (1) a Graphic User Interface with intuitive windows for an easy use adequate to biologists and physicians; (2) a module for template creation from any reference HTML page; (3) a module for automatic extraction of data from different HTML pages; (4) parametric functioning, which adapts the extraction performances by modifying parameter values inside extraction configuration text files; (5) log files that contain information about the data extractions performed and allow quick evaluation of results; (6) aggregation and storage of all extracted data, either in tab-delimited text files or in a relational database; and (7) a software agent module for updating the extracted data stored in the database.

Mined data database We designed a database schema that can aggregate and store in a relational database all the heterogeneous data extracted from different HTML pages and databanks. The schema is comprised of four tables (Fig. 5). The table MAIN stores the general information related to each extraction [i.e. extraction date, used template, name of mined biomolecular databank, used ID code of the considered nucleotide or amino acid sequence (e.g. GenBank accession number, Clone ID, UniGene Cluster ID, LocusLink ID, Swiss-Prot accession number)]. Two different tables, DATA and LINKS, contain the mined information concerning textual data or links, respectively. Table COLUMN_NAMES stores the names of the extraction result table columns, defined as described in the Methods section, which characterize the mined data.

The designed database schema enables comprehensive querying of all gathered data. In fact, each datum in the database is identified by the ID code of the nucleotide or amino acid sequence it refers to and that was used for extraction (e.g. the GenBank or Swiss-Prot accession number ID code). Besides, each extracted annotation is characterized by the name of the extraction result table column (stored in the database COLUMN_NAMES table), and/or by the other data mined with the annotation itself. When different types of ID codes for the same sequence must be used to mine data from different databanks, they can be extracted from one of the public databanks providing the ID codes of a nucleotide or amino acid sequence in different resources (e.g. GeneCards or SOURCE databanks). These different sequence ID codes are used inside the mined data database to link annotations mined from different resources.

As paradigmatic examples, in the sample databases provided with MyWEST we created two sets of queries in Structured Query Language, easily accessible through a graphic interface (see the User Guide section of MyWEST Web site). The first set includes general purpose queries, which allow simple mining of the annotations in the database just by specifying the desired keywords and/or the type of annotations among which to search (e.g. '*transcription*' and 'Term', respectively, for mining transcription related genes according to Gene Ontology (GO) controlled terms of functional categories mined from the LocusLink databank). The second set comprises articulated queries designed to provide comprehensive integrated views on the annotations mined from UniGene, LocusLink, Swiss-Prot, SOURCE and GeneCards databanks (Tables 1 and 2 and the Mining Results section of MyWEST Web site).

Mined data updating Using Java programming language, we created a software agent module for updating the information stored in the database of the mined data. The software agent utilizes the defined templates used to populate the database, and an identification code list of the nucleotide and/or amino acid sequences whose annotations need to be kept updated. At predefined intervals of time, the software agent automatically applies the extraction rules stored inside the templates, mines the available data of interest from the web pages of the sequences identified in the code list and stores the extracted data in the database, in case replacing their old versions. Thus, retrieved information presenting high temporal variability can also be kept updated and synchronized to those in the original databanks.

Mining validation and applications

We tested the efficacy and utility of MyWEST by mining from different databanks specific annotations of a set of 729 clones identified through microarray experiments as described in the

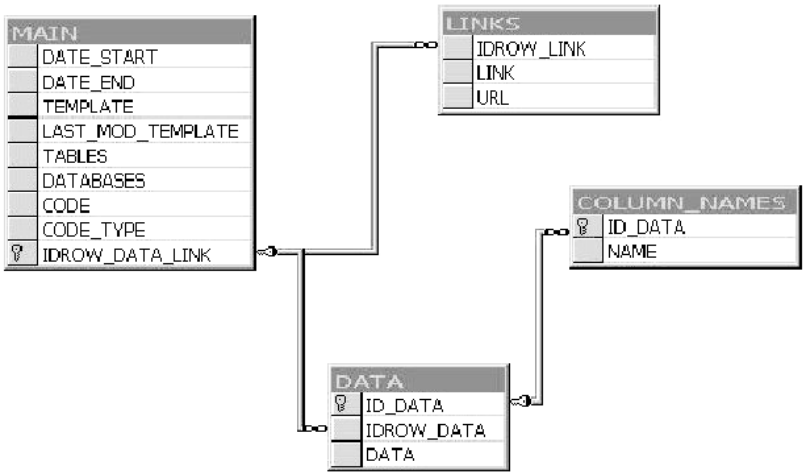

Fig. 5. Schema of the database aggregating and storing the mined data.

Table 1. Example of MyWEST mining results from the Swiss-Prot databank. Subcellular location, pathway and function of protein products of four of the identified genes with decreased (D) and increased (I) expression after 4 h Retinoic Acid treatment (RA 4 h)

RA 4 h	GenBank (or RefSeq) AN[a]	Swiss-Prot AN[a]	Gene symbol	Subcellular location	Pathway	Function	Gene title
D	NM_003921	O95999	BCL10	Cytoplasmic. Appears to have a perinuclear, compact and filamentous pattern of expression. Also found in the nucleus of several types of tumour cells		Promotes apoptosis, pro-caspase-9 maturation and activation of NF-κB via NIK and IKK. May be an adapter protein between upstream TNFR1-TRADD–RIP complex and the downstream NIK-IKK–IKAP complex	B-cell CLL/lymphoma 10
D	NM_006214	O14832	PHYH	Peroxisomal	Alpha-oxidation of 3-methyl branched fatty acids (phytanic acid); second step	Converts Phytanoyl-CoA to 2-Hydroxyphytanoyl-CoA	Phytanoyl-CoA hydroxylase (Refsum disease)
I	NM_005178	P20749	BCL3	Nuclear		Could be a transcriptional activating factor. Functions as a form of I-κB specific for NF-κB P50 subunit inhibiting its translocation to the nucleus	B-cell CLL/lymphoma 3
I	NM_002198	P10914	IRF1	Nuclear		Specifically binds to the upstream regulatory region of type I IFN and IFN-inducible MHC class I genes [the interferon consensus sequence (ICS)] and activates those genes	Interferon regulatory factor 1

[a] Accession number.

Table 2. Example of MyWEST mining results from the Swiss-Prot databank. Diseases related to two of the identified genes with decreased (D) and increased (I) expression after 4 h Retinoic Acid treatment (RA 4 h)

RA 4 h	GenBank (or RefSeq) AN[a]	Swiss-Prot AN[a]	Gene symbol	Disease	Gene title
D	NM_003921	O95999	BCL10	Defects in BCL10 are involved in various types of cancer	B-cell CLL/lymphoma 10
D	NM_003921	O95999	BCL10	Involved in a t(1;14)(p22;q32) chromosomal translocation recurrent in low-grade MALT lymphoma (mucosa-associated lymphoid tissue). Although the Bcl10/IgH translocation leaves the coding region of BCL10 intact, frequent BCL10 mutations could be attributed to the Ig somatic hypermutation mechanism resulting in nucleotide transitions	B-cell CLL/lymphoma 10
I	NM_002198	P10914	IRF1	Deletion or rearrangement of IRF1 are a cause of preleukaemic myelodysplastic syndrome (MDS) and of acute myelogenous leukaemia (AML)	Interferon regulatory factor 1

[a] Accession number.

Methods section, and by analysing the mined annotations to functionally classify the clones.

First, we used MyWEST to mine the descriptions and identification codes in several resources, and the genomic, proteomic, cytogenetic, phylogenetic, expression, structural, functional and disease annotations for the 729 putative RA target genes. As clone identifications we used the GenBank accession numbers provided with the microarray results to mine the UniGene, SOURCE and GeneCards databanks, whereas, we utilized the LocusLink IDs and Swiss-Prot accession numbers extracted from the SOURCE databank to mine the LocusLink and Swiss-Prot databanks, respectively. Tables 1 and 2 show examples of mined annotations for some of the identified differentially expressed genes (complete tables of mined annotations obtained applying the articulated queries created in the mined data database are available in the Mining Results section of MyWEST Web site).

To evaluate correctness and efficacy of the implemented mining method, the automatically mined annotations were visually compared with those in the HTML pages of the considered databanks. Accordingly to the defined extraction templates, we looked for false positive and false negative results—i.e. extracted data not of interest, and data of interest present in the HTML pages but not extracted, respectively. We found that no irrelevant annotations were mined (no false positives), and few annotations of interest present in the databank HTML pages were not mined (few false negatives). These latter were not mined mainly because of syntax errors or diverse incorrect structures in the HTML code of the Web page sections where they were located. Therefore, MyWEST proved efficient in specifically and rapidly mining the requested information for virtually all of the genes for which such information was actually available.

Next, we analysed the mined annotations and found that 513 of the 729 considered clones were classified genes (221 induced and 292 repressed by the RA treatment) and 216 were expressed sequence tags (ESTs) (118 induced and 98 repressed). Then, we attempted to classify the putative target genes according to function. Initially, among the mined annotations we selected only the GO classifications extracted from the LocusLink databank. Of the 513 identified genes, only 370 (199 induced, 171 repressed) presented GO annotations. To increase the number of functionally classified genes, we searched for other annotations, such as the RefSeq Summary and Gene Reference Into Function (GeneRIF) annotations mined from the LocusLink databank, and the Keywords and Function annotations mined from the Swiss-Prot databank. We retrieved information for 61 additional genes. Thus, we were able to obtain functional annotations from at least one source for 431 of the 513 identified genes.

To evaluate the feasibility of using the annotations mined with MyWEST for functionally classifying genes, we decided to search all extracted functional annotations for the functions presumably induced or repressed in the considered experimental condition. The RA treatment of U937 cells results in partial differentiation along the myelomonocytic lineage (Grignani *et al*., 1993). The analysis of differential gene expression at an early time point (4 h) of the RA treatment aims at identifying genes that are

Table 3. Example of comparison of MyWEST mining results from the LocusLink (LL) and Swiss-Prot (SP) databanks. Functional analysis: few of the mined transcription related genes with decreased (D) and increased (I) expression after 4 h Retinoic Acid treatment (RA 4 h)

RA 4 h	Gene title	Gene symbol	LL Gene Ontology	LL RefSeq Summary	LL GeneRIF	SP Keywords	SP Function
D	Aryl hydrocarbon receptor	AHR	X	X	X	X	X
D	Homeo box A9	HOXA9		X		X	
D	Homeo box A10	HOXA10	X	X		X	X
D	Interferon regulatory factor 2	IRF2	X	X	X	X	
D	v-jun sarcoma virus 17 oncogene homolog (avian)	JUN	X		X	X	X
D	v-myb myeloblastosis viral oncogene homologue (avian)	MYB			X	X	X
I	B-cell CLL/lymphoma 3	BCL3	X	X		X	X
I	Homeo box A1	HOXA1	X	X		X	X
I	Homeo box A2	HOXA2	X	X		X	X
I	Interferon regulatory factor 1	IRF1	X	X	X	X	
I	MYB binding protein (P160) 1a	MYBBP1A	X				
I	Nuclear factor (erythroid-derived 2)	NFE2	X			X	
	Total transcription genes mined	83	43	32	37	38	28

involved in the early phases of the differentiation process. Thus, we searched the mined annotations for functions clearly related to the early phases of RA response, such as transcriptional regulation and control of differentiation/development processes. We used the keyword '*transcription*' to mine transcription related genes; the keywords '*differentiation*', '*development*', '*embryogenesis*', '*maturation*', '*hematopoiesis*', '*hemopoiesis*', '*retinoic*' and '*retinoid*' for genes related to RA-dependent differentiation/development. As partially presented in Table 3, the results show that of the 513 mined genes, 83 have functional annotations that link them directly to transcription, and 87 have annotations that link them to differentiation/development (for complete results, see the Mining Results section of MyWEST Web site). Furthermore, 35 genes (more than 40% of the genes present in each group) were common to both groups. Therefore, using MyWEST we were able to rapidly identify transcriptional regulators and differentiation genes that are differentially expressed upon RA treatment in U937 cells.

DISCUSSION

Over the past years, several efforts have been made to effectively exploit the increasing amount of information sparsely contained inside many heterogeneous biomolecular databanks accessible through Web servers (Etzold *et al*., 1996; Davidson *et al*., 1997; Haas *et al*., 2001; Freier *et al*., 2002; Rebhan *et al*., 1998; Diehn *et al*., 2003). Nevertheless, the solutions proposed to extract the information contained in different databanks and to execute comparisons either require advanced informatics knowledge and significant resources (e.g. legacy systems such as SRS, BioKleisli, DiscoveryLink or BioDataServer), or only partially solve individual problems (e.g. integrational databanks such as GeneCards or SOURCE). The first one is adequate for big research centres but not always for the needs of small research laboratories or individual researchers. The second, represented by centrally curated and publicly accessible resources, are useful for retrieving information only from the databanks they curate. In fact, they often either prevent users from performing batch queries on multiple nucleotide or amino acid sequences simultaneously, or enable aggregation of some limited information only. Furthermore, in almost all cases the retrieved data are published inside HTML pages, i.e. in a format not suitable to store and structure them for further mining and analysis. On the other hand, direct access to annotations in their structured form is provided by some databanks only, and usually only through FTP. This requires adequate expertise and resources for locally re-implementing and maintaining several biomolecular databanks in order to comprehensively browse and mine data for proper analysis of high-throughput experiment results. Direct access to data in the XML format, provided by Web services, would be the best option. Unfortunately, at present, very few providers offer proper data access by a Web services model. Thus, currently the option of extracting data from databank Web interfaces remains attractive for researchers without extended informatics knowledge and with limited supporting resources.

Wrapper and screen scraping software have been developed and implemented to extract the data contained in a HTML page and to organize them in other formats (Muslea *et al*., 2001; Sahuguet and Azavant, 2001; Lacroix, 2002; Laender *et al*., 2002). Nevertheless, these software solutions are based

on script programs that are generally not easily applicable to HTML pages with a complex structure, such as those of some integrational databanks, and whose script code needs to be modified when the page structure or simply the extraction needs change. Therefore, they are suitable for central bioinformatics facilities rather than for customizations directly performed by the end user.

Nowadays, rapid improvements in processor speed, RAM memory and hard disk storage capacity have made it possible for a desktop PC to perform most of the functions that were traditionally done by a server. We think that an effective and personalized use of the vast amount of publicly available information will be possible only when individual researchers have all the instruments to easily modify mining criteria and comparison rules without the need to ask bioinformaticians for constant supervision. Rather, bioinformaticians must develop tools with a friendly and intuitive interface adequate also for non-informatics oriented people. The goal should be to move bioinformatics from server side to client side, so that small laboratories and individual researchers can also manage and use software tools with ease and at low cost.

Mining method and software prototype

With the above aims we created MyWEST, which is freely available for academic and non-profit use at http://www.medinfopoli.polimi.it/MyWEST/. Its characteristics make the implemented prototype a flexible and adaptable instrument that does not require program code modifications. Users with basic informatics knowledge can easily handle the prototype, create and configure templates according to specific data mining needs, use them to automatically extract data from different HTML pages also of different biomolecular databanks, and comprehensively query the retrieved mined data. If the HTML page structure of the considered databanks changes or if mining requirements vary, templates can be easily and quickly recreated. More expert users can modify template configuration parameters, involving knowledge of HTML language, to optimize the mining performance in relation to the specific extraction.

As the performed tests demonstrated, the controls implemented in the mining method do not allow irrelevant information to enter in the mined data database, and give users the chance of carefully revising those HTML pages that could have presented problems during the automatic extraction, i.e. possibly containing annotations of interest that were not automatically mined.

The high-temporal variability of the data contained in many biomolecular databanks requires an equally high-updating frequency of the extracted annotations to prevent the latter from rapidly becoming obsolete. The software agent developed for updating the mined data achieves this goal autonomously and intelligently, and provides the aggregated and structured data inside the MyWEST database with the fundamental characteristic of being up to date.

All these features make MyWEST prototype a tool especially adequate for the needs of small and medium research laboratories, which often do not have the expertise and resources to manage instruments that are more sophisticated but also more expensive and complex to use.

Mining validation and applications

The validation performed using MyWEST in a model system of RA-induced differentiation demonstrated the efficiency and versatility of the proposed mining method. It also showed the utility and potential of the implemented software to help interpreting results from differential gene expression experiments. In fact, MyWEST enriches a list of differentially expressed genes with annotations mined from different biomolecular databanks, freely chosen by the user according to the requirements of specific experimental objectives. This allows the creation of tables that enhance significant characteristics of the considered set of genes by integrating and organizing the mined annotations. Examples are the tables of protein similarities in different organisms, protein structures and functions (including domain, subcellular location and pathway), and related phenotypes and diseases (Tables 1 and 2 and the Mining Results section of MyWEST Web site). At present, to our knowledge, these data are not easily obtainable in such an aggregated way with any other publicly available resource.

Furthermore, because MyWEST enables mining the same type of annotations from various sources, it allows either comparison of equivalent annotations, or integration of similar information to increase the number of annotated genes (Table 3). This can be very useful especially for functional or disease annotations that are still scarce and not homogeneously represented in the diverse sources. In annotating our chosen dataset, we coherently classified a significant number of genes as involved in the relevant functional pathways by interrogating different sources. Moreover, the number of functionally annotated genes increase in ~16% by considering different functional descriptions besides the GO controlled vocabulary. All annotations extracted with MyWEST for a set of RA-regulated genes were in agreement with the results of a gene-by-gene literature search independently performed on the same gene list. However, using MyWEST we were able to extract relevant information for our selected dataset within a few minutes.

CONCLUSIONS

MyWEST constitutes a powerful and user-friendly tool for mining several annotations (e.g. genomic, proteomic, cytogenetic, phylogenetic, expression, structural, functional and disease) of multiple genes from distinct user-selected biomolecular databanks accessible through Web

interfaces, allowing their integration to expression profiling results. MyWEST is aimed at researchers without extended informatics knowledge and with limited supporting resources, and provides the functionalities necessary to easily exploit relevant information sparsely stored, without requiring local re-implementation of biomolecular databanks.

ACKNOWLEDGEMENTS

We thank Heiko Muller for suggesting the initial idea from which we started the development of MyWEST and the method it implements illustrated here.

REFERENCES

Ben Miled,Z., Li,N., Kellett,G.M., Sipes,B. and Bukhres,O. (2002) Complex life science multidatabase queries. *Proc. IEEE*, **90**, 1754–1763.

Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I., Pilbout,S. and Schneider,M. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.

Davidson,S.B., Overton,C., Tanen,V. and Wong,L. (1997) BioKleisli: a digital library for biomedical researchers. *Int. J. Digit. Libr.*, **1**, 36–53.

Diehn,M., Sherlock,G., Binkley,G., Jin,H., Matese,J.C., Hernandez-Boussard,T., Rees,C.A., Cherry,J.M., Botstein,D., Brown,P.O. and Alizadeh,A.A. (2003) SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data. *Nucleic Acids Res.*, **31**, 219–223.

Etzold,T., Ulyanov,A. and Argos,P. (1996) SRS: information retrieval system for molecular biology data banks. *Methods Enzymol.*, **266**, 114–128.

Freier,A., Hofestädt,R., Lange,M., Scholz,U. and Stephanik,A. (2002) BioDataServer: a SQL-based service for the online integration of life science data. *In Silico Biol.*, **2**, 0005.

Galperin,M.Y. (2004) The molecular biology database collection: 2004 update. *Nucleic Acids Res.*, **32**, D3–D22.

Grignani,F.R., Ferrucci,P.F., Testa,U., Talamo,G., Fagioli,M., Alcalay,M., Mencarelli,A., Grignani,F., Peschle,C., Nicoletti,I. and Pelicci,P.G. (1993) The acute promyelocytic leukaemia-specific PML/RARα fusion protein inhibits differentiation and promotes survival of myeloid precursor cells. *Cell*, **74**, 423–431.

Haas,L.M., Rice,J.E., Schwarz,P.M., Swops,W.C., Kodali,P. and Kotlar,E. (2001) DiscoveryLink: a system for integrated access to life sciences data sources. *IBM Sys. J.*, **40**, 489–511.

Lacroix,Z. (2002) Biological data integration: wrapping data and tools. *IEEE Trans. Inform. Technol. Biomed.*, **6**, 123–128.

Laender,A.H.F., Ribeiro-Neto,B. and da Silva,A.S. (2002) DEByE—data extraction by example. *Data Knowl. Eng.*, **40**, 121–154.

Lenhard,B., Hayes,W.S. and Wasserman,W.W. (2001) GeneLynx: a gene-centric portal to the human genome. *Genome Res.*, **11**, 2151–2157.

Muslea,I., Minton,S. and Knoblock,C.A. (2001) Hierarchical wrapper induction for semistructured information sources. *Auton. Agent. Multi Agent. Syst.*, **4**, 93–114.

Pruitt,K., Tatusov,T. and Maglott,D. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.*, **29**, 137–140.

Rebhan,M., Chalifa-Caspi,V., Prilusky,J. and Lancet,D. (1998) GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. *Bioinformatics*, **14**, 656–664.

Sahuguet,A. and Azavant,F. (2001) Building intelligent Web applications using lightweight wrappers. *Data Knowl. Eng.*, **36**, 283–316.

Schuler,G.D. (1997) Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J. Mol. Med.*, **75**, 694–698.

Siepel,A., Farmer,A., Tolopko,A., Zhuang,M., Mendes,P., Beavis,W. and Sobral,B. (2001) ISYS: a decentralized, component-based approach to the integration of heterogeneous bioinformatics resources. *Bioinformatics*, **17**, 83–94.

Stevens,R., Baker,P., Bechhofer,S., Ng,G., Jacoby,A., Paton,N.W., Goble,C.A. and Brass,A. (2000) TAMBIS: transparent access to multiple bioinformatics information sources. *Bioinformatics*, **16**, 184–185.

Tatusova,T.A., Karsch-Mizrachi,I. and Ostell,J.A. (1999) Complete genomes in WWW Entrez: data representation and analysis. *Bioinformatics*, **15**, 536–543.