Automatic extraction of gene annotations from data-rich HTML pages

Giovanni E. Carrara [(1)], Andrea Stella [(1)], Francesco Pinciroli [(1)], Myriam Alcalay [(2)], Marco Masseroli [(1)]

[(1)] Dipartimento di Bioingegneria, Politecnico di Milano, piazza Leonardo da Vinci 32, 20133 Milano, Italy
masseroli@biomed.polimi.it
[(2)] IFOM - FIRC Institute of Molecular Oncology, I-20139 Milano, Italy

Keywords. Information extraction, Web-based biomolecular databanks, Wrapper software, Genomic annotations.

Introduction

High-throughput technologies create the necessity to integrate the resulting gene expression data with information mined from large amounts of gene annotations within several different biomolecular databanks. Most of these databanks [1] can be queried only via web, for a single gene at a time, and query results are generally available in HTML format. Although some databanks provide batch retrieval of data via FTP, this requires expertise and resources for locally re-implementing the databank. Web wrappers can automate extraction of the information of numerous genes from different web-based databanks. As the content of a dynamic web page can change from one query to another (e.g. tables with extra rows or missing fields), such wrappers should be able to locate and extract data of interest inside different HTML pages. Unfortunately, HTML tags describe the visual formatting of data, not their semantics. Thus, human-readability and machine-readability are often not equivalent. Wrapper generation tools help creating a wrapper for a specific source, i.e. a web-based biomolecular databank with its own HTML layout. First, the user is invited via a Graphic User Interface to select data of interest inside one or more sample HTML pages. Then, the system saves this information as an extraction template for that specific source. The long term goal is to generate wrappers that scale well with the number of processed web pages.

Materials and Methods

Using Java programming language, we developed *GeneWebEx* (freely available for non-profit use at *http://www.medinfopoli.polimi.it/GeneWebEx/*), a tool aimed at researchers without high informatics skills or resources. *GeneWebEx* enables to generate user-defined templates identifying specific information of interest inside HTML pages, exploits the generated templates to easily mine selected annotations, and aggregates extraction results in tab delimited text files and in a local database. We tested our tool against three of the main freeware tools to create general-purpose wrappers for web pages (W4F: *http://cheops.cis.upenn.edu/W4F/*, XWRAP: *http://www.cc.gatech.edu/projects/disl/XWRAPElite/*, DEByE: *http://www.lbd.dcc.ufmg.br/~debye/*). Our case study was extracting all annotations of a group of genes from different web-based biomolecular databanks. We evaluated *GeneWebEx* using a set of 729 genes resulting from the analysis of microarray experiments aimed at identifying genes that are differentially expressed in U937 cells after 4 hours of treatment with 10^{-6} M Retinoic Acid (RA). *GeneWebEx* mined 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. Table 1 shows an example of extraction results: annotations mined with *GeneWebEx* from different databanks for some of the identified differentially expressed genes.

Results and Discussion

To evaluate correctness and efficacy of the implemented mining method, the automatically mined annotations

Table 1. Example of *GeneWebEx* mining results from Swiss-Prot and SOURCE databanks [1]. Name and title of gene, and symbol, name, molecular weight, and length of its protein product for two of the identified genes with decreased (D) and increased (I) expression after 4 hours Retinoic Acid treatment (RA 4h).

RA 4h	*GenBank (or RefSeq) AN*[a]	*Swiss-Prot AN*[a]	*Gene Name*	*Protein Symbol*	*Gene Title*	*Protein Name*	*Molecular Weight*	*Length*
D	NM_003921	O95999	BCL10 or CIPER or CLAP	BCLA_HUMAN	B-cell CLL/lymphoma 10	B cell lymphoma / leukemia 10	26251 Da	233 AA
I	NM_002198	P10914	IRF1	IRF1_HUMAN	interferon regulatory factor 1	Interferon regulatory factor 1	36502 Da	325 AA

[a] *Accession Number.*

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). Therefore, *GeneWebEx* proved efficient in specifically and rapidly mining the requested information for virtually all of the genes for which such information was actually available. Moreover, within *GeneWebEx* we provide all modules necessary for the extraction and further analysis of biomolecular data from data-rich HTML pages. An example MS-Access implementation of the local database where storing mined data is also provided, although advanced users can easily utilize other relational DBMS).
We found general-purpose wrapper tools lack of essential requirements for our case study. First of all, life science researchers are not specialized in computer programming. All of the examined general-purpose tools are GUI-based, but they often require to write code to refine wrappers. Secondly, 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. In *GeneWebEx* we implemented a software agent module to update periodically the mined data in batch mode. General-purpose tools do not have a batch mode for extractions. Finally, most general-purpose tools export mined data in XML format but this requires to develop a custom module to perform additional analysis.
On the other hand, we found general-purpose tools show good flexibility in locating the data of interest when the HTML page structure changes respect to the sample page used. These tools exploit Wrapper Induction [2] and other heuristic pattern matching algorithms to give wrappers the ability to adapt to slight changes in the HTML tag structure. A Wrapper Induction system scales well with the number of processed web pages, requiring less interaction with the user to refine created wrappers. These features are not yet implemented in *GeneWebEx*.

Conclusions and Future Work

GeneWebEx constitutes a powerful and user-friendly tool for mining several annotations (e.g. genomic, proteomic, cytogenetic, phylogenetic, expression, structural, functional, disease) of multiple genes, allowing their integration to expression profiling results. *GeneWebEx* is aimed at researchers without extended informatics knowledge and with limited supporting resources, and provides the functionalities necessary to easily and quickly exploit relevant information sparsely stored. However, to provide *GeneWebEx* with more flexibility and scalability, we plan to implement in our tool some of the efficient heuristics present in some of the general-purpose wrapper tools we tested.

References

[1] R. T. Walker, D. Söll and A. S. Jones (eds.), Database issue. *Nucleic Acids Res*., 32, 2004

[2] N. Kushmerick, Wrapper induction: efficiency and expressiveness. *Artificial Intelligence Journal*, 118:15-68, 2000.