Towards a Semantic Web for Bioinformatics

Abstract

With the explosion of online accessible bioinformatics data and tools, systems integration has become very important for further progress. Currently, bioinformatics relies heavily on the Web. But the Web is geared towards human interaction rather than automated processing. The vision of a Semantic Web facilitates this automation by annotating web content and by providing adequate reasoning languages.

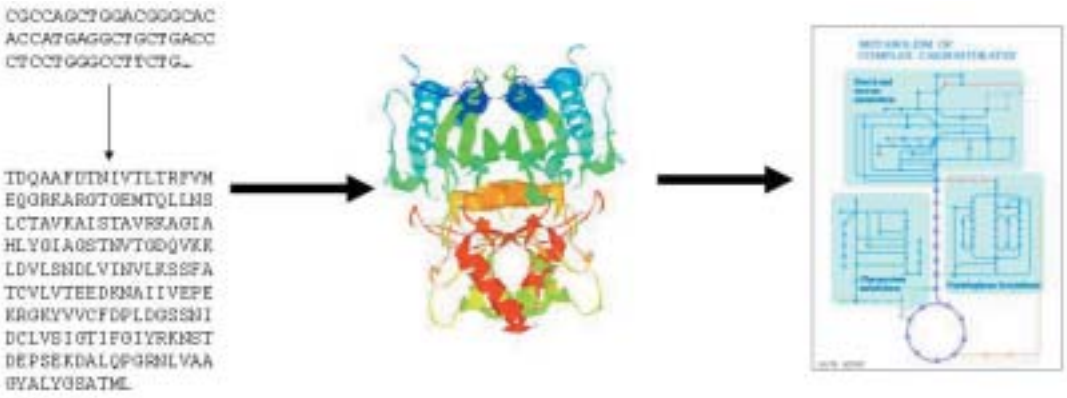

Fig 1: Sequence-Structure-Function. Biologist wish to go from DNA and protein sequences to structural information to understand the proteins' function

Mission

Biologists wish to understand protein function from their sequences and structure (Fig. 1). The bioinformatics group of REWERSE brings 9 European groups together, which have a strong background in the use of rules and reasoning for biological data on sequence, structure, and function.

They deploy rules and reasoning for ontologies and text mining, gene expression data analysis, metabolic pathways, structure prediction and protein interaction. Enabling these tools on the Web sets the foundation for a Semantic Web for bioinformatics.

Use Scenarios

Consider a biologist working on osteoporosis, a major bone disease affecting millions of people. To target osteoporosis, understanding the balance between cells, which produce bone substance, and cells called osteoclasts (Fig. 2), which consume it, is important. Imagine that a scientist has measured the gene expression levels of osteoclasts during cell differentiation. Dozens of genes appear interesting and require further annotation to reveal any details. Numerous PhD students spend days in front of the computer using the internet to collect information on these genes. They wish to learn whether the proteins resulting from the expressed genes are similar to any known proteins with known function, which domains the proteins consist of, which interaction partners these proteins have, which metabolic pathways they are involved in, which cellular locations the proteins mostly occur in, references to the most suitable articles on these proteins, etc.. All this information taken together, may lead to insights into which proteins may be suitable targets to treat bone-related diseases.

Much of the needed information and analysis tools are accessible over the Web. However, they are designed for low-throughput human use and not for high-throughput automated use. The vision of a Semantic Web for bioinformatics transparently integrates some of these resources through the use of mark-up languages, ontologies, and rules. In such a bioinformatics Semantic Web, the biologist provides the identifiers of the experimentally determined genes to a rule-based workflow, which integrates various information sources.

A service for protein threading will determine the structural domains for the genes. Among others this service determines a number of overexpressed small GTPases, which act as molecular switches in signalling. A service for protein interactions determines potential interaction partners of the GTPases.

>

More information available at

http://rewerse.net/a2

It highlights for example the proteins de/activating the GTPases. Some of experimentally proteins play a role in the MAPK signalling pathway. A rule-based service allows the biologist to ask queries over the signalling pathway to determine whether the expressed proteins are essential or not. Another service determines the relevant literature for the overexpressed genes and uses an ontology to annotate the function, processes, and cellular locations of the genes. The biologist uses rules to reason over these ontologies. He or she uses the ontology to retrieve scientific articles, which are relevant for osteoporosis, but not postmenopausal osteoporosis, and which mentions any positive regulation of protein kinase activity such as MAPK or JUNK.

Description of Research

The bioinformatics group of REWERSE will build prototype applications to demonstrate the idea of a rule-based Web for bioinformatics. It brings groups together which have used a host of of rule-based techniques such as constraints for structure prediction (Jena, Lisbon), logic programming for systems integration, to reason over metabolic pathways and to classify expression data (Dresden, Paris, Bucarest, Edinburgh), and ontologies to annotate expression data and for text-mining (Dresden, Bucarest, Edinburgh, Manchester, Linköping). The working group will build on some of these existing tools and integrate them with rule-based Web technologies to demonstrate the idea of a Semantic Web for bioinformatics.

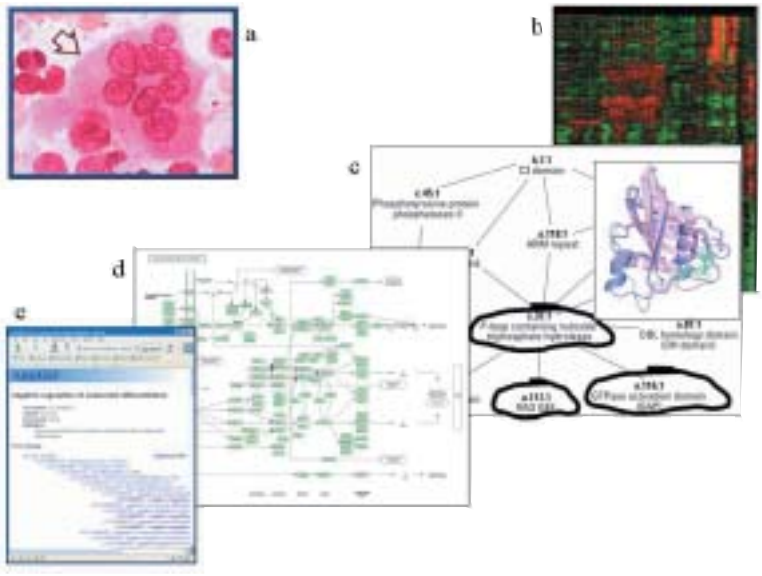

Fig2: Osteoclasts are cells which digest bone substance and are therefore important for bone-related diseases such as Osteoporosis. The figure a shows a large cell with separately identifiable, multiple nuclei. To understand osteoclasts various experimental and in-silico techniques are used:
Gene expression data captures the activity of genes. Figure b shows the result of a screen of many genes over a period of time. Red indicates over expression. The genes are grouped according to the similarity of their expression profiles. For the overexpressed gene products structural domains and their interaction partners are found (figure c). Among others small GTPases, which act as molecular switches, are found. The inset in figure c shows the superposition of two GTPases, one of which is switched on, the other off. The difference can be seen in the different conformations of the chains in the front. Some of the overexpressed gene products are relevant for human MAPK signalling pathway according to Kegg (www.genome.jp/kegg/) depicted in figure d. Finally, ontologies are used to annotate the data. Figure e shows the definition of "negative regulation of osteclast differentiation" according to GeneOntology,
www.geneontology.org

Tools & Technologies

All of the groups involved in the bioinformatics group have developed rule- and Web-based bioinformatics tools. Dresden has developed an ontology-based literature search engine (www.gopubmed.org) and a database to determine protein structure interactions. Jena has developed tools using constraints for structure prediction. Lisbon has used constraints for structure prediction and has developed the docking package Bigger. Paris has developed a rule-based reasoning engine for metabolic pathways
http://contraintes.inria.fr/BIOCHAM/
Edinburgh has developed an ontology for tissues and linked it to in-situ gene expression data. Bucarest has used rules and inductive logic programming for characterizing differentially expressed genes in microarray data. Manchester has developed a host of tools to analyse, maintain and deploy biomedical ontologies. Linköping has developed tools to merge and maintain ontologies.

Contact Person

Dr. Michael Schröder, Professor
Biotec/Dept. of Computing
TU Dresden

Tatzberg 47-51
01307 Dresden, DE

Phone: +49 351 463 40062
Email: ms@mpi-cbg.de

www.biotec.tu-dresden.de

Members

Michael Schröder (Dresden);
Rolf Backofen (Jena);
François Fages (Paris);
Werner Nutt (Edinburgh);
Carole Goble (Manchester);
Pedro Barahona (Lisbon);
Patrick Lambrix (Linköping);
Liviu Badea (Bucharest);
Mikael Berndtsson (Skövde)

Impressum

webXcerpt Software GmbH
REWERSE Technology Transfer
Aurbacherstr. 2, D-81541 Munich
http://rewerse.net

Contact: Andrea Kulas
ak@webxcerpt.com
Phone: +49 89 54 80 88 48

Responsible for the content:
Prof. Michael Schröder
Dept. of Computing
TU Dresden
Tatzberg 47-51, D-01307 Dresden
ms@mpi-cbg.de
Phone: +49 351 463 40060

2004.11-300