BACHELOR PROJECT 1

Web-based visualization of high throughput experimental molecular data.
Bioinformatics group, Delft University of Technology, Faculty of Electrical Engineering Mathematics and Computer Science, Dept. Mediamatics
url: http://ict.ewi.tudelft.nl/index.php?option=com_sections&id=71&Itemid=138
Contact person: Prof.dr.ir. M.J.T. Reinders

Background: Through advances in genomics research a wealth of information is now being generated concerning molecules present in a cell. Especially high throughput techniques like 1) microarrays, that enable the measurement of gene activities through the messenger RNA that they produce (wikipedia: microarray); or 2) mass spectrometry that enables the measurement of protein and metabolite availability (wikipedia: mass spectrometry). As a consequence genomics research groups start acquiring genomic data at several molecular levels (sequence, RNA, protein and metabolites) in well-defined growth conditions that are varied over a complex set of parameters (e.g. oxygen availability or a variety of nutrient limitations). Analyzing this data will shed light on the molecular mechanisms in the cell. The TUD bioinformatics group is developing a database that stores these high throughput data. To be able to interrogate the data we would like to develop a web-based visualization of the high throughput data present in the database.

Project objectives: The visualization should make it possible to interrogate the data in different ways. Expression profiles of choice need to be displayed over the different conditions including appropriate visualization of the status of each of the conditions. Sequential ordering of these profiles should be possible according to the data in chosen conditions. The visualization should allow for different overlays of expression profiles. For example, overlaying profiles of expression at different molecular levels (e.g. RNA profiles with protein profiles) or overlaying profiles of groups of molecules (e.g. RNA profiles of all genes having a certain annotated function). This should include also overlaying groups of molecules that result after a simple analysis of the profiles (e.g. protein profiles of proteins that show a similar expression over the different conditions). The visualization should be enriched with annotation data (such as gene names or functional groups) and links to general databases such as ensemble (www.ensemble.org).

BACHELOR PROJECT 2

Web interface for querying and visualizing complex interrelated biological data.
Bioinformatics group, Delft University of Technology, Faculty of Electrical Engineering Mathematics and Computer Science, Dept. Mediamatics
url: http://ict.ewi.tudelft.nl/index.php?option=com_sections&id=71&Itemid=138
Contact person: Prof.dr.ir. M.J.T. Reinders

Background: In the last few years, biological research has resulted in more and more large datasets, describing the cell on different levels. For example, information about DNA sequences (wikipedia: DNA), mRNA expression levels (wikipedia: DNA microarray) or protein interactions (wikipedia: protein interaction). Currently, large efforts are being made to integrate the data on these different levels. However, easy access to data as well as the results of these integration algorithms is still an issue as data is separated over a large number of databases (~1000), each using its own unique keys and names (wikipedia: biological databases). In the TUD bioinformatics group we conduct many data integration studies. For that purpose we are now working on a database that can store and link these different data and result sets. This database is made accessible by a command line language, enabling us to easily extract specifically constrained datasets and also visualize them using a linked visualization tool (www.cytoscape.org). To make such a tool accessible to other researchers (bioinformaticians as well as biologists) we like to develop a web interface which can be used instead of the command line interface.

Project objectives: The web interface will consist of several subsystems for browsing the data sets. One of those subsystems is an interface for creating complex queries. An example of such a query is: *all genes which are linked by their protein products to genes which have transporter functionality*. The difficulties here are especially in the generation of correct database queries, as well as creating an easy but flexible interface for the end-user. A second subsystem will be a graph browser utility which visualizes relations within the data sets (e.g. between genes) to improve accessibility of the data. Technical difficulties here will be the drawing and layout of

this graph within a web interface. A third subsystem could be a module which enables researchers to import or export data for further analyses, using different file formats. Difficulties here are within the mapping of the data to existing objects in the databases as well as security measures to prevent misuse.

BACHELOR PROJECT 3

Advanced visualization of large biological networks
Bioinformatics group, Delft University of Technology, Faculty of Electrical Engineering Mathematics and Computer Science, Dept. Mediamatics
url: http://ict.ewi.tudelft.nl/index.php?option=com_sections&id=71&Itemid=138
Contact person: Prof.dr.ir. M.J.T. Reinders

Background: With the increasing availability of biological data (wikipedia: biological databases), biological research is becoming more focused on the research of complete systems as well as the interrelations between systems/modules in living cells/organisms (wikipedia: systems biology). To that end, various visualization tools have been developed. However, a problem with current tools is that the large networks which you find in reality have too many relationships, resulting graphs which are completely covered by edges and in which no structure can be found anymore (google image: protein network). Two methods which could be used to improve this are:

- Node summarization: Summarize nodes based on common properties. This is often done using clustering algorithms. However, currently, the meta-nodes resulting from this clustering do not represent any information on the cluster (which could be hub-like, or complete linked, etc.). This can be improved by using a hybrid approach where meta-nodes are represented by matrices indicating the link structure within a cluster.
- Edge summarization: Bundle edges to create a nicer view. For example, if one cluster is tightly connected with another cluster, the edges between those two clusters can be bundled (using a technique called hierarchical edge bundles, see http://www.win.tue.nl/~dholten/papers/bundles_infovis.pdf) to improve the visualization.

An existing visualization tool within the biological research community is Cytoscape (www.cytoscape.org). It is becoming more and more a standard due to its open structure. We would like to develop the aforementioned summarizations into this software tool.

Project objectives: Development of a plug-in to the open-source Cytoscape software. The main difficulties will be in:

- Understanding the current structure of the Cytoscape software before beginning the development
- Writing meta-nodes drawing code as well as determining/developing suitable layout algorithms for meta-nodes expansion to node groups within an existing layout.
- Implementing the hierarchical edge bundles using a specified hierarchical clustering of the nodes.
- Creating an open interface for using the new functionality.