Date of Acceptance Grade

Instructor

# Analysis and Visualization Software in Context of Biological Networks

Jyrki Kankaanpää

Helsinki, 18.4.2007

Seminar Report

UNIVERSITY OF HELSINKI

Department of Computer Science

HELSINGIN YLIOPISTO – HELSINGFORS UNIVERSITET – UNIVERSITY OF HELSINKI

| | |
|---|---|
| Tiedekunta/Osasto – Fakultet/Sektion – Faculty/Section<br>Faculty of Science | Laitos – Institution – Department<br>Department of Computer Science |

Tekijä – Författare – Author

Jyrki Kankaanpää

Työn nimi – Arbetets titel – Title

Analysis and Visualization Software in Context of Biological Networks

Oppiaine – Läroämne – Subject

Bioinformatics

| | | |
|---|---|---|
| Työn laji – Arbetets art – Level<br>Seminar report | Aika – Datum – Month and year<br>18.04.2007 | Sivumäärä – Sidoantal – Number of pages<br>11 |

Tiivistelmä – Referat – Abstract

New technical research methods and cyclical research strategies are producing data in increasing amounts for analysis purposes in life sciences. The automatic data analysis and visualization tools are, however, not at the level with the amount of produced data. Therefore new bioinformatics tools are needed and especially in system biological dynamic network research. This seminar report describes an implemented research software tool for the automatic data analysis and visualization purposes of biological networks.

ACM Computing Classification System (CCS)

D. [Software]
G.3 [Probability and Statistics]
I.5 [Pattern recognition]
J.3 [Life and Medical Sciences]

Avainsanat – Nyckelord – Keywords

Bioinformatics, biological networks, data analysis, visualization

Säilytyspaikka – Förvaringställe – Where deposited

Muita tietoja – Övriga uppgifter – Additional information

# Contents

# 1 Introduction

Life science research has come to a situation where the nature of research and novel technical research methods are producing massive amounts of raw data for interpretation purposes. Also, there is a general paradigm shift to create understanding from a top-down viewpoint instead of a more traditional reductionist bottom-up approach. This novel multidisciplinary procedure, systems biology, tries to model the structure and function of whole organisms concentrating especially on dynamic biological networks and their components.

Biological experiments are conducted repeatedly to test hypotheses and models from different viewpoints. These experiments produce almost, but not exactly, similar data sets to be fitted in existing theoretic models. This difference in data sets is quite obvious, because even genetically identical organisms (i.e. clones) grown in identical environmental conditions differ by their chemical structures. In addition, experimental equipment are a source of noisy data. In other words, because of the biological variation and technical limitations of laboratory equipment, the experimental data even from seemingly similar samples is identical only within certain statistical limits.

New automatic analysis tools are needed because massive amounts of raw data are slow and unpractical to interpret by labor methods. To be more specific the problem is that many technical laboratory research equipment and their software produce data in complex numerical table forms. These tables are difficult to interpret because data is neither analyzed nor visualized meaningfully in biological context. Therefore, a novel approach in bioinformatics is to develop software tools which are able to analyze and visualize experimental data and especially in the context of biological networks. However, the abilities of these tools are still limited [JKS06]. Most of them, for example, do not allow the comparison of more than two data sets. Also, many of them use static pathway pictures from research databases as an only source of structural network models.

This seminar report is based on an article [JKS06] which originally presents a software tool for automatic data analysis and visualization purposes in context of biological networks. The structure of this report is as follows. In Section two, the software system is described. In Section three, the computational solutions of the system are discussed. Finally, Section four concludes.

# 2 Analysis and visualization software for biological networks

VANTED is a freeware software application for the visualization and analysis of networks with related experimental data in context of biological networks [JKS06]. It is implemented with Java and usable over internet as a Java Web Start application or as a stand alone installation. The graphical user interface of the system is mimicking the general user interface layouts of software applications. Details of these issues can be found here [Van07].

VANTED is an extension of a prototypic data exploration module of DBE-information system. DBE (D*ata analysis and visualisation system for* B*iological* E*xperiments*) is built to visualize experimental data in the context of metabolic networks. The system consists of five subsystems: 1) database, 2) data import application, 3) web-based user interface, 4) application for the up- or download of image files for network models and 5) a network analysis and graph visualization system. DBE is presented in more details in [Bor04]. The main new implemented features in VANTED are statistical analysis and automated data clustering tools. Further, VANTED is built to be dynamically extensible by the user by Java scripts [Bea07, GiM07] and Ruby Scripts [JRu07]. With these scripts, the user can add into the system new algorithms for analysis, graph layout, data exchange and other functionalities [JKS06].

The software system is an analysis tool of large-scale biochemical data sets. It supports automatic visualization and mapping of data into dynamic biological networks (i.e. pathways). Dynamic networks are in this context graphs which describe biological networks and which are locally customizable by the individual users of the system. These graphs are derived from global research databases via Internet or specifically given by the user. Hence, dynamic networks are an extension of static maps (i.e. network pictures) which are generally used in biochemical research and which are downloadable from research databases like KEGG [KaG00] and UniProt [Uni07].

With VANTED, the data can be presented in the context of corresponding biological network elements like protein levels in protein-protein interaction networks or transcriptomic data in gene regulatory networks [JKS06]. The system allows the import of any type of biochemical data from different genotypes, environmental conditions and time points, network loading and editing and the mapping of the data on the corresponding dynamic networks. Though there is no actual limitation on the type of networks or data, the imported data sets should contain up to a few hundred items (metabolite, enzymes) from up to a few dozen conditions (genotypes, time points). Hence, it is not practical to visualize genome wide data sets.

As a standard data import procedure the system utilizes MS Excel format which is most commonly used data export format in life-science laboratory software packages [JKS06]. Accordingly, it is the one and only importing method of experimental data into the system database. In practice the user copy pastes source data into the template form from which a specific importer application utilizes information into the system database [Bor04]. Data handling is further simplified by including in the import template all relevant general experimental information related to a specific conducted experiment, e.g. a replicate number, sample time, a measuring tool, genotype information, environmental conditions, etc.

Network information can be imported into the system in different file formats (e.g. GML, SBML or Pajek-.NET) or from databases (e.g. SOAP, KGML). Graph Modelling Language (GML) is a proposal for a platform independent graph file exchange format [HiM07]. Systems Biology Mark up Language (SBML) is a standard exchange format designed for computational models of biochemical networks [FiH03]. Pajek-.NET is a file format for large network analysis [BaM04]. Information can be exported as a standard JPEG or PNG format image files or as a GML format graph file.

The networks can be derived from the network data of research databases or if there are no corresponding networks, defined by the user. Further the structure of the networks can change locally depending on the set of substances analysed in a particular experiment [Bor04]. The system utilizes mainly Kyoto Encyclopaedia of Genes and Genomes (KEGG) database for basic network information [JKS06, KaG00]. KEGG is a combined effort to standardize gene annotations and to create a knowledge base where genomic information is linked to higher order functional information in organisms. KEGG consists of three databases: PATHWAY for the representation of higher order functions in terms of biological networks, GENES for the collection of gene catalogues of all completely sequenced genomes and LIGAND for the collection of chemical compounds in a cell. While utilizing KEGG databases the user can choose between a top-down or a bottom-up approach starting from a pathway or a genomic information while searching or browsing the database. For example, the search is connected with corresponding names, i.e. annotations, of the genes or pathways and browsing to the taxonomic structures of organisms (Eucaryotic, Procaryotic or Archaea). If the name of a pathway is known, the user can search the network model information which is saved in

KEGG pathway database. If a specific enzyme gene structure is known for example, the user can search this gene and with this information corresponding links to network information. After a relevant network information is found it can be downloaded from the Pathway database to the system database as an image file.

The system is able to import any kind of omics-data (e.g. transcript, protein and metabolite) from different growth conditions and time points [Bor04, JKS06]. Imported data can be mapped on to a corresponding and editable dynamic network structure once inside the system. Assignment of data from experiments to network elements is automated but this data mapping procedure can also be refined by the user. In an automatic procedure, the system compares the identifiers of imported biochemical information with the identifiers of the nodes of the loaded network. If the name of the network target node is the same as the name of the measured substance in input form, the system maps imported data to the target node. The target nodes (i.e. network elements) can be metabolites, transcripts or enzymes. The data mapping procedure checks additionally any synonyms or other identifiers defined in KEGG Ligand Database or in the Swiss Institute of Bioinformatics (SIB) database. In case that a mapping with KEGG pathways is desired, the substance names for the measurements should be Enzyme Commission (EC) numbers or Compound Identifications (CID) [JKS06, BNC07, Pub07]. If the system can not find any suitable target node for automatic integration, the user can create a new network module or map data subsets manually to a specific target node. Also VANTED supports an automatic creation of new nodes for all measured data subsets which do not map to the loaded network. The data mapping is visualized as a diagram in each graph node for which specific experimental data is available, as in Figure 1.

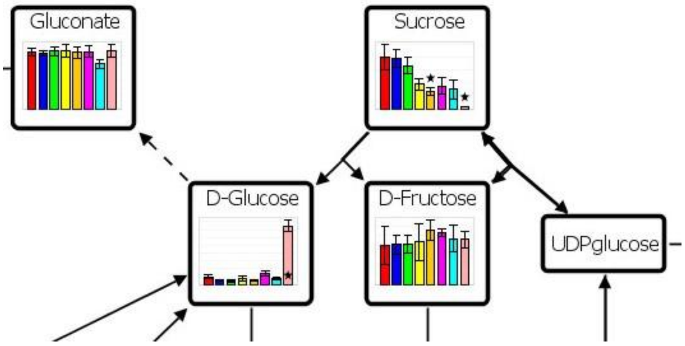

Figure 1 Substance graph nodes with mapped statistical information from multiple experiments. There is no corresponding experimental data to map into the empty node [JKS06].

The graphical visualization module is based on an extensible graph library and editor Gravisto [Bor04, JKS06, Bac04]. Gravisto is a graph visualization toolkit which allows basic routines for graph display and layout. New experimental data can be added to graph views with the help of additional view plug-ins and JFreeChart library [GiM07, JKS06]. General graph editing functions are implemented, like node/edge selection, modification and deletion. Also an algorithm for the removal of node overlaps [DMS05] and algorithms for circular, tree-shaped and force-directed layouts are available.

The actual visualization of data within the network modules is implemented by line or bar charts [JKS06]. Data is mapped into a specific target node and charts are inserted into the image of the node accordingly. Data from multiple experiments can thus be seen in a single diagram inside each substance node, as in Figure 1. If different genotypes or plants are to be visualized, the data can be shown in separate diagrams as separate data sets in a single node image. Replicated and merged measurement values can be shown as an error bar of standard deviation (SD) or the standard error of the mean (SEM) in both kinds of diagrams. The bars describe reference and experimental mean values and the star describes statistically significant values compared with reference values. Section three discusses more of these statistical issues. In line charts the variability of the data can be visualized as a polygon around the diagram line. The style of the diagrams may be modified with parameters of colours, line widths and range/category labels.

The visualization is enhanced by different levels of detail by a zooming tool [Bor05]. It enables a robust and general top-down view of large networks without interfering details. The system drops out extra details (e.g. legends, labels and captions) of the network which are not appropriate at a particular zooming level. However, whenever the user thinks that the details are relevant, it is possible to zoom down to an individual network element.

Figure 2 presents a normal working session with VANTED. The procedure can be split into three phases. First data is imported with a MS Excel template form. Second a suitable network file is searched from the system database. If there is none, the system creates the network, imports it from KEGG Pathway database or imports it as a GML-file. Third phase includes data mapping, analysis and visualization. After the experimental data is mapped to the network the final graph image is created as a modifiable and exportable graph file. The choice, whether to use data analysis or data clustering tools, is left to the user.

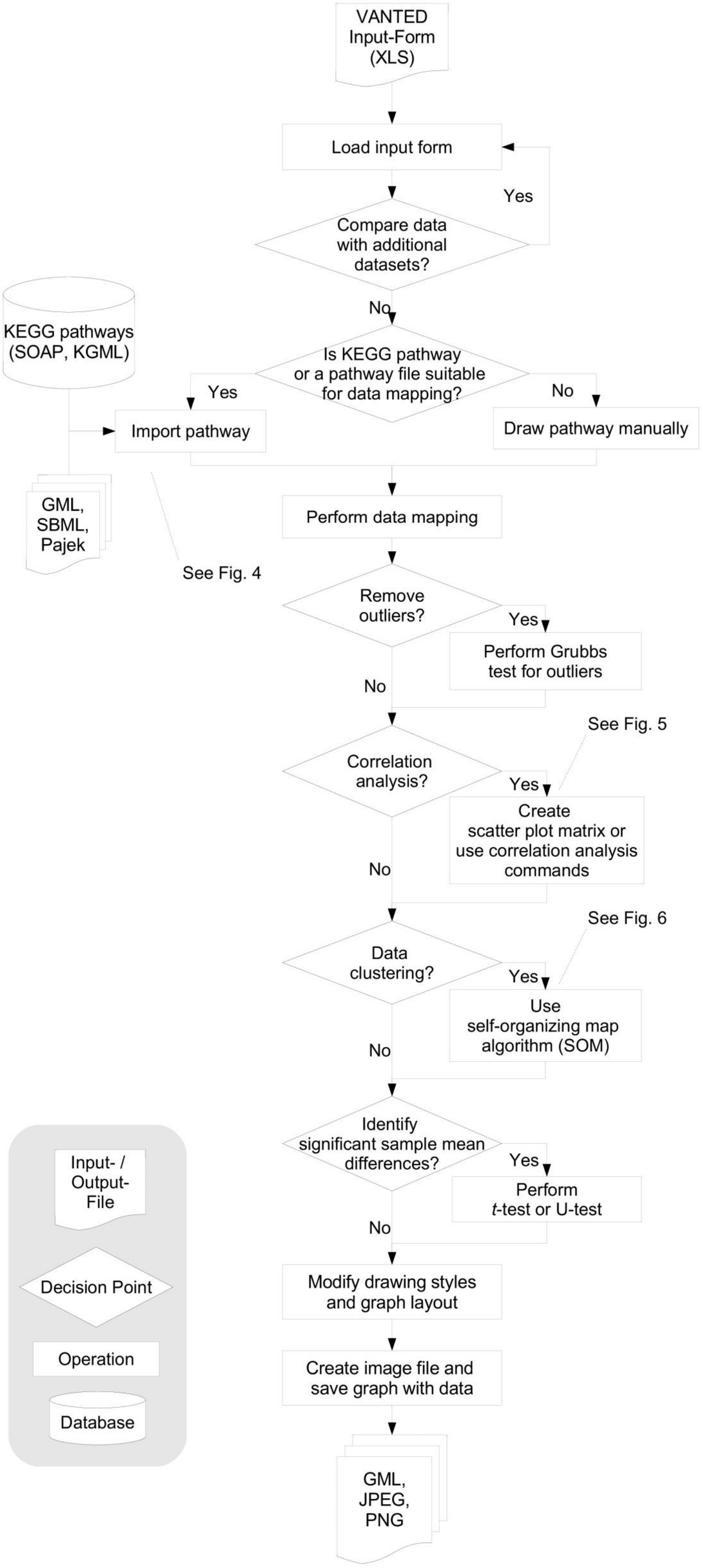
VANTED Input-Form (XLS)
Load input form
Yes
Compare data with additional datasets?
No
KEGG pathways (SOAP, KGML)
Is KEGG pathway or a pathway file suitable for data mapping?
Yes
No
Import pathway
Draw pathway manually
GML, SBML, Pajek
Perform data mapping
See Fig. 4
Remove outliers?
Yes
Perform Grubbs test for outliers
No
See Fig. 5
Correlation analysis?
Yes
Create scatter plot matrix or use correlation analysis commands
No
See Fig. 6
Data clustering?
Yes
Use self-organizing map algorithm (SOM)
No
Identify significant sample mean differences?
Yes
Perform t-test or U-test
No
Input- / Output-File
Decision Point
Operation
Database
Modify drawing styles and graph layout
Create image file and save graph with data
GML, JPEG, PNG

Figure 2 Work flow of a typical working session with VANTED [JKS06].

# 3 Computational data-analysis methods

In biological research statistical analysis is in a vital role while interpreting the results of conducted experiments. The analysis and comparison of resulting data sets from different genetic lines or developmental stages is feasible only by data analysis methods. Therefore, this chapter presents the main data analysis methods which are utilized by VANTED. However, the basic statistical tests (i.e. functions) and their underlying assumptions, as well as the pattern recognition method, are described only briefly.

## 3.1 Implemented statistical methods

Both parametric (Student's t, Welch-Satterthwaite t, Pearson's correlation) and nonparametric (Mann-Whitney U, Wilcoxon U, Spearman's correlation) standard tests of statistical significance and correlation are implemented into the system [Est07, EwG05]. The difference between nonparametric and parametric test is that with parametric tests at least one population parameter is estimated from the sample. To do this it is assumed that all measured variables are normally distributed into the population to which we plan to generalize our findings. Therefore, the normality of the data must be checked before any parametric tests are used. Such distribution assumption is not needed with nonparametric tests. If the data is non-normal and/or sample size is small, we should use an analogous nonparametric test and consider its limitations.

To test whether the measured data is normally distributed the system uses a test called *the David quick test for normal distribution*. It seems that there are no generally known references available in statistical literature to this test. However, it is generally known that the normality (i.e. mean, variance, skewness and kurtosis) of the data can be easily tested by known statistical tests, e.g. Shapiro-Wilk. The authors of the paper [JKS06] are not motivating why they have implemented a specific test though there are more generally known tests for normality available.

Grubb's test (maximum normalized residual test) is used to test outliers (distant values) in an univariate data set. The test assumes normality from the distribution. Grubb's test pinpoints one outlier at a time so it has to be repeated until all necessary outliers are eliminated from the data set. The elimination of outliers is important because they create excess skewness in value distributions. However, the test should not be used with non normal small samples. There is a risk of data manipulation if relevant outliers are dropped out to fit the data into the model.

The comparison of data from different experiments can be done with the *t*-test. Different samples of plants or genotypes can be compared to find out significant differences in their mean values. The Student’s *T*-test, is used in VANTED to examine whether two groups differ significantly from each other by their mean values (or from zero). The test assumes that the variables of the sample populations are normally distributed and that the samples have equal variances. The *T*-test is used for small samples though theoretically measured populations asymptotically approach the normal curve. Hence, the *t*-test is useful in such designs in which experimental and control group are compared. In addition an unpaired (independent samples) t-test is also implemented. This Welch-Satterthwaite *t*-test is an alternative to the Student's *t*-test and is used when an assumption of an equal variance of two populations is not reasonable.

Mann-Whitney U-test is the non-parametric analog of the t-test and it is also used for pair wise comparisons of significance. The samples do not need to be normally distributed and it does not utilize mean or median values. The samples should, however, be independent. If this is not the case, one may use Wilcoxon U-test for this purpose.

Pearson's correlation coefficient r measures the strength of the relationship between two variables. It ranges from +1 through 0 to -1 with a perfect positive relationship, no relationship or perfect negative relationship. A perfect correlation produces a perfect linear relationship and, if visualized by scatter diagrams, an ascending or descending straight line in XY co-ordinates. The variables should be measured in interval scales at least, whereas Spearman's correlation is also usable on the ordinal scale and is as a nonparametric test more robust against outliers. Moreover, the relationship between two variables does not need to be linear [JKS06].

The correlation information is utilized by VANTED in two ways [JKS06]. Scatter plot matrices or correlation networks are created to visualize relationships found of the data. In scatter plot matrices, the substances for pair-wise relation can be chosen by the user. A value pair is formed if three annotations correspond: 1) the plant/genotype name, 2) time value and 3) replicate number [JKS06]. After the correlation factor is computed different colors are used to describe positive (blue) or negative (red) correlation. Also in case of significant correlation scatter plot matrices borders are highlighted by an increased border width. A correlation network can be created from a number of selected substance nodes and correlations are calculated between all possible pairs of nodes. A new correlation edge is created between nodes if there is a significant correlation. An example of correlation networks and scatter plot matrices can be seen in Figures 3 and 4.

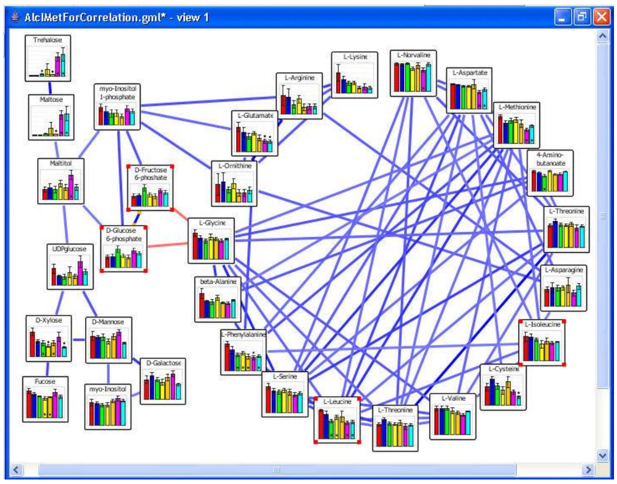

Figure 3 Correlation network [JKS06].

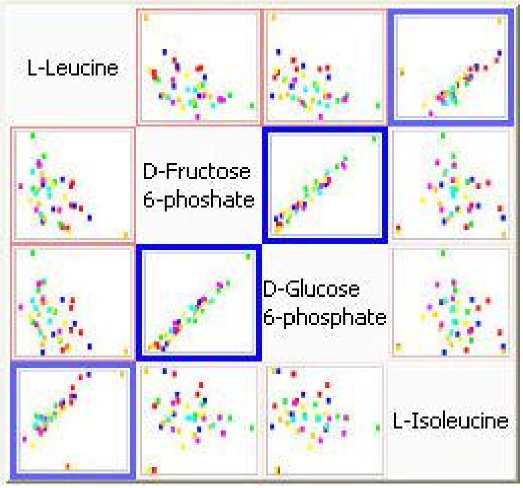

Figure 4 Correlation scatter plot matrix [JKS06].

## 3.2 Pattern recognition with Self Organizing Map

Self-Organizing Map (SOM) [Koh90] is a sheet like artificial neural network which is constructed to imitate biological neuronal networks functionally. The artificial cells of the network become specifically tuned to the various input signal patterns or classes of patterns through an unsupervised learning process. This enables a visualization of complex and noisy data in a more humanly understandable form in a visual (2-dimensional) pattern map where similar data sets get grouped on the same areas of the map. There is practically no limitation for the information types which SOM networks can handle, making it a very versatile tool for data analysis purposes and for visualization and classification tasks.

In VANTED SOM is used to extract common measurement patterns over time [JKS06]. First the system is trained with normalized input vectors, which are created from an ordered set of average sample values. Practically this means that a given number of the typical profiles of substance concentrations (clusters) over time are entered into the SOM. During this machine learning phase SOM adapts itself to these common input patterns. When real experimental data is entered to a trained system all substances are grouped and visualized according to their similarity to trained SOM nodes, i.e. target clusters are determined by the minimum distance of target vectors from model vectors. Colouring, filter and layout operations are possible for further visualization purposes [JKS06].

# 4 Discussion

There are several other implemented software systems for data analysis and visualization purposes, for example KaPPa-View, Omics Viewer and Pathway Explorer [JKS06]. Restrictions with these tools compared with VANTED are many. They do not support a direct comparison of more than two data sets. They use static network pictures which can not be edited to reflect single user's needs. Also, many of these tools are only for limited data visualization purposes without support for statistical analysis. In that sense, VANTED seems to widen the features of existing analysis software in a way as the implementers claim. By utilizing statistical methods, the system can map multiple data sets and visualize them inside individual network elements. The system is able to utilize static network pictures as a source of local dynamic pathway structures. This makes VANTED more versatile than many other analysis and visualization software.

The systems most important benefit for biologists might be the automatic statistical visualization of experimental data of any biological network. Usually those biologists who are not mathematically strong prefer visual results compared with any numeric results. In that sense, because the results of several experiments can be visualized in one picture, the system shows more clearly and more quickly any tendency in the results in contrast to a comparison of separate data sets. Hence, the automatization brings quickness to the research which is enhanced by the ability to modify network structures locally. This modification feature gives the possibility to test the biological models locally while the actual implications of the collected results are still somewhat unclear.

While used in biological research, the system has produced results which are in accordance with previous biological studies [JKS06]. For example selected metabolic data (amino acids, sugars, and sugar derivatives) of potato tubers expressing a yeast invertase has been mapped into a corresponding metabolic network structure. The system has given the correlation results of different substances which are analogous to the results of previous studies. With the visualization of correlation (Figure 3 and 4) was seen that some carbohydrate concentrations are massively increased by the constitutive expression form of yeast invertase while there is no increase with an inducible expression form. To find, for example, these results from large numeric tables would have been a challenging task for several researchers [JKS06]. However though the system has been used in biological research, there usually seems to be a connection to the implementers with corresponding research articles. It seems that there is yet no independent and systematic analysis of the implementation available. Hence, the real and long lasting benefits of VANTED for life-science research are to be seen.

It is difficult to see the real benefits of specifically implemented statistical tools in VANTED software system described in this report. Why not to use common statistical programs for data analysis purposes as usually well established tools have implemented statistical tests correctly. Moreover and for example, it is possible to use SPSS statistical software through Python programming language interface [SPS07].

# References

Bac04 Bachmeier, C., et. al., Gravisto: Graph Visualization Toolkit. In *Proceedings of the 12th International Symposium on Graph Drawing, LNCS,* Vol. 3383, Springer, 2004, 502-503.

BaM04 Batagelji, V., Mrvar, A., Pajek – Analysis and Visualization of Large Networks. In *Graph Drawing Software*, Springer, 2004, 77-103.

Bea07 BeanShell: Lightweight Scripting for Java. *http://www.beanshell.org*, 03.04.2007.

BNC07 Biochemical Nomenclature Committees. *http://www.chem.qmul.ac.uk/iubmb/nomenclature/*, 26.3.2007.

Bor04 Borisjuk, L., et. al., Integrating data from biological experiments into metabolic networks with the DBE information system. *In Silico Biology, 5,* 2004.

DMS05 Dwyer T., Marriot, K., Stuckey, P., Fast Node Overlap Removal. In *Proceedings of the 13th International Symposium on Graph Drawing*, LNCS, Springer, 2004, 77-103.

Est07 Electronic Statistics Textbook. *http://www.statsoft.com/textbook/stathome.html*, 03.04.2007

EwG05 Ewens, W., J., Grant, G. R., *Statistical Methods in Bioinformatics: an introduction*. Springer, New York, 2005.

FiH03 Finney, A., Huchka, M., Systems Biology Mark Up Language: Level 2 and Beyond. *Biochemical Society Transactions*, 2003, 31, 1472-1473.

GiM07 Gilbert, D., Morgner, T., JfreeChart, a free Java class library for generating charts. *http://www.jfree.org*, 03.04.2007.

HiM07 Himsolt, M., GML: a portable Graph File Format. *http://www.infosun.fim.uni-passau.de/Graphlet/GML/gmr-tr.html*, 03.04.2007.

JKS06 Junker, B., H., Klukas, C., Schreiber, F., VANTED: A system for advanced data analysis and visualization in the context of biological networks. *BMC Bioinformatics*, *7, 109*, 2006.

JRu07 JRuby: A Ruby interpreter written in pure Java. *http://www.jruby.sourceforge.net*, 03.04.2007.

KaG00 Kanehisa, M., Goto, S., KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acid Research*, 2000, 28, 27-30.

Koh90 Kohonen, T., The Self-Organizing Map. *Proceedings of the IEEE*, 1990, 78, 1464-1480.

Pub07 PubChem. *http://pubchem.ncbi.nlm.nih.gov/*, 03.04.2007.

SPS07 SPSS, *http://www.spss.com/*, 03.04.2007.

Uni07 The Universal Protein Resource, *http://www.ebi.uniprot.org*, 03.04.2007

Van07 VANTED homepage, *http://vanted.ipk-gatersleben.de*, 03.04.2007