J. Pestic. Sci., **31**(3), 289–292 (2006)

Commentary

(Special Topic)

The VANTED software system for transcriptomics, proteomics and metabolomics analysis

Christian KLUKAS, Björn H. JUNKER† and Falk SCHREIBER*

Leibniz Institute of Plant Genetics and Crop Plant Research, Corrensstrasse 3, 06466 Gatersleben, Germany
† *Brookhaven National Laboratory, Biology Department, 50 Bell Avenue, Upton, NY 11973, USA*

(Received May 24, 2006)

Current research in biology generates data sets of increasing size that are very difficult to manage and analyze manually. Bioinformatics tools are necessary to facilitate statistical analysis and visualization of the data. While multiple tools exist for this purpose, they are often limited to specific kinds of data or allow only certain types of analyses. Recently, we have reported on the development of VANTED, a software system that allows mapping of multi-dimensional data sets onto relevant biological networks. VANTED provides a variety of functions for network editing, data mapping and processing, statistical analysis, and visualization. This review summarizes the main features of VANTED. © Pesticide Science Society of Japan

Keywords: genomics, transcriptomics, proteomics, metabolomics, biological networks, data analysis, visualization.

Introduction

Knowledge-generation in the life sciences is a cyclic process. To understand a biological phenomenon, experiments are carried out which produce large amounts of -omics data (*e.g.* transcriptomics, proteomics and metabolomics). This data has to be analyzed and is then used to build or improve models which represent the biological system. Based on the knowledge obtained by data analysis and modeling, hypotheses can be created. These usually lead to new experiments, which in turn produce new data, which can be used to refine the model, and so on. All steps of this cycle result in new knowledge about the biological system under investigation, and all steps involve methods and tools from Bioinformatics.

This review considers the first part of the cycle: the analysis of data obtained by biological experiments. During the last years the methodology of the biochemical research has greatly changed. Nowadays large amounts of experimental data can be measured simultaneously using modern massive-parallel techniques: for example, automated enzyme-assays,[1] transcript-[2] and metabolite[3]-profiling determine up to some thousands of data points from a single biological sample. The resulting data basis enables the user to gain a comprehensive view of the biochemistry of an organism. These methods are typically used to compare a wild type with different transgenic organisms, or to analyze the effect of different environmental conditions such as environmental stress by pesticides. Measurements are typically repeated several times to improve the reliability of the analysis. Also in certain experimental settings time series data needs to be analyzed. These factors increase the number of measured data points. Ideally this data should not be analyzed independently; instead, the context of the measured substances given by related biological processes should be considered during data analysis. The interpretation of the increasing amount of data is a challenging task, as time and effort for data analysis increases, and traditional visualization methods turn out to be insufficient for bigger data sets. Therefore, new analysis and visualization methods become more and more important. These methods aim at bringing the data in a form that, on the one hand gives an overview about the overall system, and on the other hand provides sufficient detail.

Several systems have been developed to support the outlined analysis and visualization tasks, for example, MapMan,[4] KaPPA-View,[5] PathwayExplorer[6] and the Omics Viewer incorporated into MetaCyc-related databases.[7] However, most of these tools do not support the direct comparison of more than two data sets with each other. Further, they are

* To whom correspondence should be addressed.
E-mail: schreibe@ipk-gatersleben.de

restricted to the analysis of expression data or they rely on static maps, that is, fixed pictures which cannot dynamically changed depending on the users' needs. Therefore, we developed a new system which supports the visualization and analysis of complex datasets (including different -omics areas, time series-data and data for the comparison of different transgenic plant lines) in the context of underlying biological processes or networks.

VANTED—A System for Data Analysis in the Context of Biological Networks

VANTED (visualization and analysis of networks containing experimental data) is a platform-independent software system which enables researchers to evaluate extensive biochemical data in an easy way. A screenshot of the system is shown in Fig. 1. VANTED supports the integrated analysis of data for different growth conditions or transgenic lines from optionally different time points. It uses the KEGG Pathway database,[8] which includes a comprehensive set of pathway maps representing knowledge about metabolism, and the Gene Ontology,[9] which contains a hierarchy of controlled vocabulary to describe gene and gene products in organisms. Such networks and hierarchies, either imported, modified or newly created by the user, serve as the basis for different methods for mapping, visualization and analysis of data, which are described in more detail in the following sections.

Not all imaginable use-cases may be supported by a software system, and therefore the possibility of enhancing an application with newly developed analysis, visualization and data exchange methods is of great importance for many researchers. A Java and Ruby script-interface and interpreter allows the user to dynamically extend the VANTED system for such purposes.

Network Generation and Data Mapping

For the analysis of experimental data it is sensible to apply an integrated view of the measured data and its related background information, such as metabolic pathways or regulative processes. This approach corresponds to the general idea of systems biology, where a biological system is analyzed not only by studying a single phenomenon, but by considering a broader view which includes all elements of a biological system.

To achieve this, three aspects were important for the design of the VANTED software system:

(1) In contrast to many other systems, VANTED supports dynamic networks. Networks can be imported from databases (*e.g.* KEGG) or may be loaded from files in different formats (GML,[10] SBML,[11] Pajek-.NET[12]). It is also possible to create networks by hand with an integrated graphical editor. A big advantage of dynamic networks is the possibility of customizing them easily for different requirements. For instance, networks can be easily extended when more substances are measured.

(2) The integration of measured data and relevant network elements is supported. An automatic mapping of data onto relevant network elements occurs if the measured data and the network nodes have common identifiers. Also during this mapping procedure synonyms are used as long as they are included in one of the supported databases (*e.g.* the SIB En-

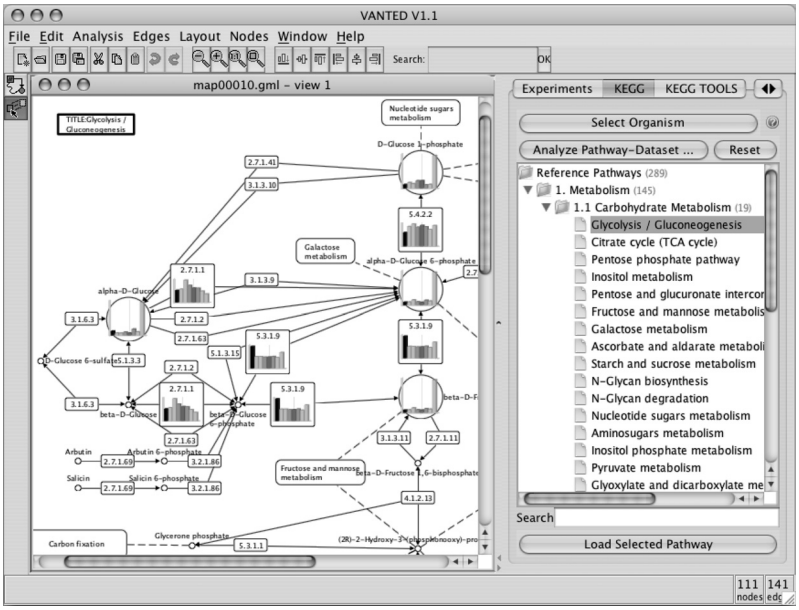

Fig. 1. Mapping of metabolite and enzyme data onto the KEGG pathway Glycolysis/Gluconeogenesis. Each circular node represents a metabolite, each rectangular node represents an enzyme. The bars in each diagram representing different plant lines showing from left to right the values from wild-type potato tubers, and tubers expressing a yeast invertase either in an inducible or constitutive manner.[16]

zyme nomenclature database from the Swiss Institute of Bioinformatics[13]). If an automated integration is not possible, a new graph node is generated, which is then used for data mapping. Additionally, data may be assigned manually to user-given network elements.

(3) VANTED supports the display of multiple values on a single network element. While current approaches often support only the coloring of network elements based on single values (*e.g.* directly measured data or a computed factor, such as comparison of two different datasets), the inclusion of diagrams in the network representation allows the visualization of more complicated data. An additional advantage which arises from the use of line charts or bar charts is the easy interpretation of such a representation.

Statistical Tests

The measured data of a sample varies around a mean value because of measuring inaccuracies and biological variability. When a wild-type of a plant is compared to different other lines, or the plant is exposed to environmental stress, it is of interest whether the sample means differ significantly or not. For normally distributed data two variations of the *t*-test can be applied. Depending on the assumption of equality of variances, Student's unpaired *t*-test or the Welch-Satterthwaite *t*-test can be carried out. Whether a sample is normally distributed can be checked within VANTED with the built-in David-quick test. The measurements which do not fulfill this criterion are marked and can then be examined separately. As an alternative to the *t*-test, the *U*-test is provided, which may also be used for not normally distributed data. Another phenomenon is outliers in the dataset which can be identified in VANTED based on the Grubbs test.

Correlation Coefficients

Relations between different measured substances can be recognized with XY diagrams. Here VANTED allows the selection of a number of substance nodes from the network view. These substances are pair-wise related to each other and displayed in an array of XY diagrams. The correlation factor between each two substances is computed and visualized using different colors of the diagram borders. In a similar way, the correlation of a user-selected substance to all other substances can be determined. A positive or negative correlation between the selected substance and another substance becomes immediately visible by different node background colors. Furthermore it is possible to determine statistically significant correlations between all measured substances which can be visualized by new graph edges connecting significantly correlated graph nodes.

Automated Data Clustering

To recognize typical patterns in the temporal courses of the substance concentrations, VANTED includes a neuronal network algorithm, the Self-Organizing Map (SOM).[14] In the first phase of this algorithm a given number of typical profiles of substance concentrations over time is determined. For instance, the substance concentration may increase in one group of substances during the time and decrease in another. In the second step every measured substance is assigned to the best suitable pattern. Each substance afterwards belongs to a specific group. In the graphical network view the grouped data sets can then be separated and individually analyzed.

Summary

VANTED is a helpful system for the visualization and analysis of -omics data in the context of relevant biological networks. As a Java Webstart application it runs on all common computer platforms such as Linux, Windows and Mac OS. It is available free of charge at http://vanted.ipk-gatersleben.de.

VANTED has already been shown to be useful for the analysis of metabolite data.[15] It can be also used for the analysis of transcript and protein data (an example is shown in Fig. 1). With the support of a network integrated analysis of data from different -omics areas and the access to the KEGG Pathway database, the Gene Ontology hierarchy and SBML import functionality, VANTED provides a unique combination of features, not available in other systems.

References

1) Y. Gibon, O. E. Blaesing, J. Hannemann, P. Carillo, M. Hohne, J. H. M. Hendriks, N. Palacios, J. Cross, J. Selbig and M. Stitt: *Plant Cell* **16**, 3304–3325 (2004).
2) J. L. DeRisi, V. R. Iyer and P. O. Brown: *Science* **278**, 680–686 (1997).
3) O. Fiehn, J. Kopka, P. Dormann, T. Altmann, R. N. Trethewey and L. Willmitzer: *Nat. Biotechnol.* **18**, 1157–1161 (2000).
4) B. Usadel, A. Nagel, O. Thimm, H. Redestig, O. E. Blaesing, N. Palacios-Rojas, J. Selbig, J. Hannemann, M. C. Piques, D. Steinhauser, W. R. Scheible, Y. Gibon, R. Morcuende, D. Weicht, S. Meyer and M. Stitt: *Plant Physiol.* **138**, 1195–1204 (2005).
5) T. Tokimatsu, N. Sakurai, H. Suzuki, H. Ohta, K. Nishitani, T. Koyama, T. Umezawa, N. Misawa, K. Saito and D. Shibatanenell: *Plant Physiol.* **138**, 1289–1300 (2005).
6) B. Mlecnik, M. Scheideler, H. Hackl, J. Hartler, F. Sanchez-Cabo and Z. Trajanoski: *Nucleic Acids Res.* **33**, W633–W637 (2005).
7) C. J. Krieger, P. Zhang, L. A. Müller, A. Wang, S. Paley, M. Arnaud, J. Pick, S. Y. Rhee and P. D. Karp: *Nucleic Acids Res.* **32**, 438–442 (2004).
8) M. Kanehisa, S. Goto, M. Hattori, K. F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki and M. Hirakawa: *Nucleic Acids Res.* **34**, D354–357 (2006).
9) The Gene Ontology Consortium: *Nat. Genet.* **25**, 25–29 (2000).
10) M. Himsolt: *Softw. Pract. Exp.* **30**, 1303–1324 (2000).
11) A. Finney and M. Hucka: *Biochem. Soc. Trans.* **31**, 1472–1473 (2003).
12) V. Batagelj and A. Mrvar: 'Pajek—analysis and visualization of large networks,' in *Graph Drawing Software*, ed. by M. Jünger and P. Mutzel, Springer, pp. 77–103, 2004.
13) A. Bairoch: *Nucleic Acids Res.* **28**, 304–305 (2000).

14) T. Kohonen: *Proc. IEEE* **78**, 1464–1480 (1990).
15) H. Rolletschek, R. Radchuk, C. Klukas, F. Schreiber and L. Borisjuk: *New Phytol.* **167**, 777–786 (2005).
16) B. H. Junker, R. Wuttke, A. Tiessen, P. Geigenberger, U. Sonnewald, L. Willmitzer and A. R. Fernie: *Plant Mol. Biol.* **56**, 91–110 (2004).