Published in final edited form as:

Drug Discov Today Biosilico. 2004 November ; 2(6): 237–245. doi:10.1016/S1741-8364(04)02423-0.

Information Visualization Techniques in Bioinformatics during the Postgenomic Era

Ying Tao, Yang Liu, Carol Friedman, and Yves A. Lussier

Department of Biomedical Informatics, Columbia University, New York, NY 10032, USA

Ying Tao: Ying.tao@dbmi.columbia.edu; Yang Liu: Yang.liu@dbmi.columbia.edu; Carol Friedman: Friedman@dbmi.columbia.edu; Yves A. Lussier: Lussier@dbmi.columbia.edu

Abstract

Information visualization techniques, which take advantage of the bandwidth of human vision, are powerful tools for organizing and analyzing a large amount of data. In the postgenomic era, information visualization tools are indispensable for biomedical research. This paper aims to present an overview of current applications of information visualization techniques in bioinformatics for visualizing different types of biological data, such as from genomics, proteomics, expression profiling and structural studies. Finally, we discuss the challenges of information visualization in bioinformatics related to dealing with more complex biological information in the emerging fields of systems biology and systems medicine.

Keywords

Information visualization; Bioinformatics

Introduction

The quantity of molecular biological information has been expanding exponentially in the last few years, especially after the completion of human genome project. However, due to the limitation of human cognition and perception, the quantity of information a user can examine and analyze at a moment is very limited. Therefore the quality of scientific knowledge a user can retrieve from the information for his (her) specific purpose will be affected. During this post-genome era, biomedical scientists are facing the difficulty of exploring and analyzing a huge amount of biological data.

Consequently, Information Visualization (IV), which takes advantage of the bandwidth of the human vision, provides a possible solution to the problem of managing and analyzing an overwhelming amount of biological information. The purpose of this paper is to introduce the field of IV in bioinformatics and present an overview of various areas in bioinformatics where IV techniques have been used. We will discuss the implementation steps, challenges and future perspectives of IV in bioinformatics as well. In addition, we will discuss possible new fields within bioinformatics that could benefit from IV techniques.

Correspondence to: Yves A. Lussier, Lussier@dbmi.columbia.edu.

Teaser: This paper provides an overview of information visualization techniques that are used in various areas of bioinformatics and a discussion of future challenges and perspectives.

Information Visualization

IV techniques are generally accepted as computerized methods that involve selecting, transforming and representing data in a visual form that facilitates human interaction for exploring and understanding the data. IV techniques depend on two major functions of a human visual system. First, a human visual system has a very broad bandwidth that can deal with a large amount of information at one time. Human retina can accept millions of pixels' information in a moment, which is much larger than the regular size of human working memory. Second, a human visual system contains innate ability to discern trends and patterns within visual fields, such as location, shape, size, and color of objects.

Based on these two facts, the use of IV techniques is aimed at two major goals. First, it helps to view a large amount of information at a time which human can not perceive easily otherwise. Second, it helps to retrieve useful knowledge from massive information by recognizing patterns and trends.

There is a rich spectrum of IV techniques. Several different taxonomies have been developed according to different perspectives. For example, Shneiderman provided a taxonomy for IV techniques based on data types and tasks [1]. Shneiderman's taxonomy includes seven data types (1-D Linear Data, 2-D planar or map, 3-D data, temporal data, multidimensional data, tree data and network data) and seven tasks (overview, zoom, filter, details-on-demand, relate, history, and extract). Ankerst and Keim [2] classified IV techniques into six categories according to data visualization techniques, namely, geometric techniques, icon-based techniques, pixel-oriented techniques, hierarchical techniques, graph-based techniques and hybrid techniques. Besides data visualization techniques, Ankerst and Keim [2] also mentioned other dimensions that could be used in IV taxonomy, i.e., data preprocessing techniques, distortion techniques and dynamic/interaction techniques. Chi [3] provided a taxonomy for IV according to a "data state reference model", which describes four stages of data state in IV and three transformation operators between every two adjacent stages. Recently, Pfitzner et al [4] has attempted to provide a unified taxonomic framework for IV techniques in the aspect of IV system designers. They added more aspects, such as data relationships, display dimensions, user's skill level and context factors etc. Table 1 summarizes the previously described dimensions in existing taxonomies for IV. The combination of these dimensions constitutes a variety of IV techniques. Of significance for this review, from table 1, we can observe that no conventional taxonomies of IV techniques document the use of external knowledge bases to constrain the visualized information (e.g. as would ontologies or the output of a natural language understanding system).

Purposes of using IV in bioinformatics

The purposes of using IV in bioinformatics follow the two general goals of IV techniques mentioned in the previous section, 1) to visualize large amount of information; 2) facilitate analysis and data mining by aiding recognition of patterns and trends.

1) Visualization of large amount of complex information

Bioinformatics is the application of computer technology to biological information, which is well-known for its vast amount of data. IV techniques are extremely useful in bioinformatics for studying large amount of biological data. First, IV summarizes the data and provides users with views at various levels so that users not only get an overview of the entire data but also zoom in to any details of their interest. For example, to give summarized views of genomes, many genome browsers provide the function of zooming so that the whole genome can be displayed in different level of details [5,6] including chromosome level, locus level, gene level, exon/intron level, and nucleic acid level. Second, IV techniques also include viewing complex

data that are difficult to understand in their original form. For example, visualization systems of 3-D structure of a biomolecule construct a virtual 3-D image that users can easily understand based on the atom coordinates of the molecule from experimental results [7], while the original experimental information is too complex for users to construct a 3-D virtual image in their mind. Visualization of large molecular interaction networks to facilitate human understanding is also for this purpose [8-14].

2) Facilitating data mining and analysis

IV can facilitate the discovery of patterns and relations of biological data. For example, in the analysis of high-dimensional microarray data, certain preprocessing techniques can reduce the dimensions of microarray data to two dimensions so that each microarray profile can be mapped to a point in a two-dimensional plane [15-17]. Thus, similar gene expression profiles are represented as points closer to each other, and dissimilar gene expression profiles are represented as points far from each other. Therefore, clustering patterns can be easily perceived by visualization.

It should be noted that IV systems could serve multiple purposes in bioinformatics. The common theme in all the applications is that the recognition of patterns becomes easier when a large amount of data is visualized by users.

Implementation strategy

IV in bioinformatics follows the three main steps for implementing a general IV system as summarized by Lang *et al*, which include data acquisition, data mapping, and rendering [18]. The first step is to acquire data and apply preprocessing operations so that the data is in the format ready to be mapped to visual spaces. A good example in bioinformatics is to use dimension reduction algorithms to reduce the high-dimensional microarray data to two dimensions. The second step is to map preprocessed data from the first step to visualizable geometrical shapes with appropriate attributes, such as location, color, and size. In our example, the preprocessed two-dimensional data could be mapped to points in a plane. Different groups could be encoded using different size and color. Since this step decides what geometrical shapes should be used in the final graphical views, different mapping schemas in this step determine the diverse graphical appearances of IV techniques. The last step is to render final output images to a visible graphical view. In this example, the microarray data are drawn on a computer screen so that users can see the data and find patterns.

Use of IV techniques in bioinformatics

IV techniques have been used in many areas of bioinformatics. In this section, we will investigate each of these areas and the techniques that have been used.

Biomolecular structures

Because biomolecules are too small to be seen via conventional microscope, the structural information has to be obtained through other indirect measures, such as X-ray crystallography, NMR, electron microscopy, or prediction algorithms. IV techniques are used to regenerate virtual images of biological molecules based on that indirect information which is difficult for human to understand otherwise. The visualized biomolecular structures involve the secondary, tertiary and higher structures of proteins, nucleic acids and other bio-compounds.

Many tools have been developed for biomolecular visualization. These tools can be divided into two groups, 2-D tools and 3-D tools. 2-D tools visualize secondary structures of nucleic acid or protein in a 2-D space. For example, HAN K. et al [19] introduced an algorithm for

automatically drawing RNA secondary structure and the tool allows the user to generate and compare structure models for different RNA molecules. There are also tools specialized in visualizing particular structures, such as pseudoknots within RNA secondary structure [20]. In contrast to the above static tools, the conformation change of RNA's secondary structure can be visualized in movies [21]. In this way changing parts of RNA can be observed moving and attract users' attentions for closer scrutiny. 3-D tools visualize tertiary and higher structures of biomolecules in 3-D visual space. The usual method of achieving a 3-D perception through a planar computer screen is to show the molecule's rotation so that human vision can rebuild the molecule's 3-D structure by integrating the rotational information. This 3-D visualization techniques have been used in most of 3-D tools [22,23]. Conformational change and structural flexibility of proteins could also been visualized in 3-D movies to help understand how conformation changes affect a protein's function [24]. Another use of 3-D visualization of biomolecules is for comparing 3-D structures of multiple molecules [25,26].

Expression profile

High-throughput analysis methods, such as DNA microarray and 2D gel electrophoresis, have made it possible to obtain a snapshot of a gene expression profile in nearly a genome. Today, IV techniques have become indispensable for analyzing the high-dimensional data generated by these methods.

IV techniques have been extensively used to assist clustering expression profiles obtained from microarray. A commonly used method is to visualize the result of hierarchical clustering using colored mosaic and dendrogram [16]. Colored mosaic encodes the measured fluorescence ratio using different colors and arranges them in some meaningful orders, such as time and chromosome position. Dendrogram could be used to cluster these profiles based on hierarchical clustering algorithm, in which the most similar profiles are clustered together first and then less similar profiles are added to already formed clusters such that a tree reflecting clustering structure is constructed in the end (Figure 1). Different from dendrogram which uses a tree to show clusters, another group of clustering tools use 2-D scatter plot. These tools usually employ dimension reduction algorithms to reduce high-dimensional microarray data to two dimensions so that each expression profile could be mapped to a point within a 2-D plane. The arrangement of data points makes similar profiles located close to each other. Thus, clusters of expression profiles can be easily observed by human eyes. Various dimension reduction algorithms have been investigated, including principal component analysis (PCA) [27], multidimensional scaling (MDS) [28], self-organizing map (SOM) [29,30] and discrete Fourier transform (DFT) [15].

Another wide use of IV in analyzing expression profiles is to map gene expression profiles to gene annotated functions so that functional profiles could be obtained according to gene expression profiles. Most of these tools map gene expression to the generally accepted annotation, Gene Ontology (GO), based on GO's hierarchical structure [31-36]. IV techniques have also been used to map gene expression data to gene regulation pathways [37] and chromosome locations [38,39] to illustrate the gene expression profiles in the context of gene regulation pathways and gene chromosome locations.

Genome and sequence annotation

The huge volumes and complexity of genomes of any organisms make it difficult to maintain a global view of entire genomes and to obtain interesting details at the same time. Highly dynamic IV techniques with various interactive functionalities have been developed to meet this requirement. Most of the genomes are considered as linear data with other information along the DNA sequence, such as exon, intron, gene, cytogenetic band and other annotations. Genomes are usually directly mapped to linear representations in one dimension of visual

space, and different types of annotation information are represented as parallel tracks along DNA sequence [5,6,40-42]. Then various glyphs can be designed to give visual hints about the property of information on the tracks. For example, thick lines could be used for coding exons and thin lines for introns. The key IV technique in these systems is dynamic semantic zooming, which automatically change the semantics of the displayed content based on different levels of zooming, such as zooming in to display sequence data and zooming out to display cytogenetic bands or chromosome. Sometimes genomes can also be represented as circles in order to display circular genomes [43]. When linear chromosomes are aligned around a circle, circular arrangement could also be used to more easily show relationships between different chromosomes by lines connecting specific chromosome positions [44]. Additionally, some of genome viewers are designed for genome comparisons [45]. An interesting IV method for comparing genomes, called Z curve [46], maps genome sequences to a 3-D visual space such that the coordinates of a point on the curve are determined by the cumulative composition information of the four types of bases before that point. Thus different genomes form different 3-D curves and differences between genomes can be easily observed.

Due to the length of DNA sequence, IV techniques are also suitable for displaying sequence annotations. These tools are similar to genome viewers except that they mark annotation information along the displayed genomes [47,48].

Sequence analysis

Sequence analysis is another domain in bioinformatics where IV techniques have been significantly used. The visualization of sequence alignment could be as simple as aligning sequences side by side as lines of text with highlighted consensus segments [49]. A more sophisticated way of visualizing sequence alignment is to add one more plot about the similarity measure along the aligned sequences [50,51]. Alternative ways of comparing two sequences include dot plot [52] and percent identity plot (pip) [53]. Dot plot places two compared sequences along x and y axes and uses dots within the coordinate to represent the base pairs that are identical to both sequences. Thus the plot of more similar sequences is closer to a straight line and any dissimilar segments will cause deviations. The idea of pip is to plot a sequence in horizontal axis and the position in vertical axis reflects the degree of similarity in percentage to other sequences. There are also tools for specific tasks of sequence analysis, such as tools for detecting and visualizing repetitive patterns in genomes [54].

Molecular pathway

Because of the complexity of molecular pathways, such as metabolic pathways [8-10], gene regulation pathways [11] and signal transduction networks [12-14], using IV techniques to show pathway information becomes a necessity. Although various types of pathways have been visualized, the techniques for these visualizations are quite similar, i.e., a network structure in 2-D visual space with different nodes and edges representing different molecules and relations respectively. The researches on visualization of molecular pathways have mainly concentrated on performance improvement, better layout algorithms, and new ways of displaying large networks. For example, 3-D displays have been used to visualize larger networks and more intricate topologies [55]. Some 2-D innovative methods are also developed [56,57], in which algorithms are used to cluster close nodes into cliques so that the node and edge numbers of a network will be greatly reduced.

Ontology, taxonomy and phylogeny

Ontology, taxonomy and phylogeny all have a hierarchical structure. Therefore, their visualization techniques are also similar. Taxonomy and phylogenies are usually organized as trees, where items are organized in a hierarchy and one item can have only one parent, such as a phylogenetic tree of viruses [58]. However GO, the major ontology in bioinformatics, is

organized as a Directed Acyclic Graph (DAG) in which nodes are concepts and relationships between nodes is a directed hierarchy where each node is allowed to have more than one parent. A DAG can be transformed in a tree if duplicate redundant nodes are allowed. The simplest visualization of a tree is to show child nodes under their parents in wedge-like formations. We refer to this simple representation as a “graphical tree”. But representing trees in a graphical tree severely limits both the depth and breadth of a tree in a limited space. Another way to visualize large trees is to use expandable nodes to represent tree nodes, such as AmiGO [59]. Users can unfold wanted nodes and fold undesired ones so that visual space is saved. But it does not solve the problem of visualizing large numbers of nodes at one time. Many new visualization techniques have been invented to address this problem, such as treemap [32], TreeWiz [60], 3D tree [61] and 3-D hyperbolic tree [62]. For example, Treemap (figure 2) is a space-filling visualization technique. It is recursively drawn as a box and each of its children are drawn as boxes within it. In this way, a tree can be rendered in a predefined space no matter how large the tree is. The TreeWiz employs a layout algorithm to hide nodes below certain levels and still preserve the overall structure of the tree. 3D tree uses 3D virtual space and more interactive methods to display more complicated tree structures.

Recently, the visualization of GO hierarchical structure has been frequently used to facilitate the functional interpretation of microarray data. Using external knowledge from the ontology, researchers have modified the visualization of conventional techniques, which leads to novel hybrid techniques not documented in the conventional IV methods described in Table 1. For example, the purpose of using a visual approach is that the functional profiles derived from microarray data are easily perceived in the context of GO structure. The gene expression profiles obtained from microarray data are mapped to GO terms with calculated statistical scores, such as p-value. Then the resulting metrics are represented in the hierarchical structure of GO. In this way, the patterns of overexpressions and underexpressions of genes in microarray data are translated to the functional patterns expressed as GO terms and visualized in GO's hierarchical structure. The ontology-anchored techniques mentioned in the previous paragraph are fundamentally a group of preprocessing techniques (Table 1) which can be used for visualization, such as graphical tree [33,35], expandable tree [34-36], treemap [31,32]. The advantage of using treemap, a relatively new IV technique, over other methods is that it allows visualizing summaries of gene expression profiles while obtaining meaningful details. This is very important for microarray analysis because microarray data contain a large number of genes and the whole GO hierarchy is often too large for conventional IV techniques. A snapshot of treemap displaying a gene expression profile is shown in figure 2. Treeview is an ontology-anchored clustering technique that can be contrasted with its conventional form (Figure 1), which is ontology-free.

The discussion in this section provides an overview of the broad use of IV techniques in bioinformatics. Due to the large number of related tools and the limitation of space, it is by no means a complete list. Table 2 summarizes the major biomedical IV techniques we described and provides examples and references.

Challenges and perspectives of IV in bioinformatics

With the advancement of biomedical research, challenges of IV in bioinformatics reside in the large size and the high complexity of biological information.

Although IV techniques have been successfully used in many biological domains, such as structure visualization, expression profile analysis, sequence analysis, visualization of genome, pathway and hierarchical data, the large size of biological information is still challenging. For example, the need to visualize large molecular networks and large hierarchical data requires more innovative IV techniques, such as more interactive 3-D visualization and more distortion

techniques, for end users to easily understand and analyze the data. High-dimensional gene expression data is challenging dimension-reduction algorithms for analysis and visualization in the aspects of accuracy and performance.

The complexity of biological information will impose more serious challenges on IV. It should be noted that most of the areas where IV techniques have been successfully used are at molecular level. At present, we enter an era of systems biology, in which a system approach to a biological system often generates complex systems data. Biological information at the systems level, such as biological processes at cellular, phenotypic, and behavior level, is increasingly complicated and dynamic. IV techniques will play an important role in understanding these intricate biological processes and their models. For example, in cancer biology and medicine, IV tools can provide dynamic visualization of gene regulation networks, signal transduction pathways and cellular activities of tumor cells, and generate different models based on the effects of potential medicines in the treatment of the cancer cells.

To meet the challenge of visualizing complex information of biological process, new IV researches are emerging. Recently, we have seen the emerging of new IV researches on visualizing the following fields of biology: cellular processes [63], phenotypic information [64] and dynamic simulation of gene regulation [65,66]. Nonetheless, due to the limitation of our knowledge and complexity of life phenomena, visualizing more complex biological information at system level still remain challenging. Yet, there could be other areas of bioinformatics that could benefit more from IV techniques. For example, text visualization methods have been studied in biomedical literature to identify major language patterns in biology text [67]. Moreover, pioneer research using visualizing complex semantic networks from biological and medical ontologies and from literature using natural language processing is undergoing to meet the need of systems biology. In the mean time, the application of new IV techniques will also speed up the development of bioinformatics. For example, New IV techniques, such as treemap [31,32] (Figure 2), have helped to overcome the difficulties of existing IV methods in analyzing gene expression profiles. And more interactive IV techniques, such as virtual reality, have been applied more and more in bioinformatics recently [55,68].

Conclusion

IV techniques are becoming increasingly important for managing and mining biological information. They have been successfully used in many domains of bioinformatics, such as biomolecular structure visualization, expression profile analysis, sequence analysis and annotation, genome visualization, molecular pathways and hierarchical biological data. Creating and applying innovative IV techniques, such as virtual reality and novel ontology-anchored ones, are poised to expedite the advance of biomedical research in the post-genome era. There are still many areas in bioinformatics where IV techniques have yet to be fully exploited, such as visualization of biological process simulation, text visualization and visualizing the increased complexity of biological data in systems biology. Additionally, as we have shown in this review, conventional IV techniques and semantic relationships arising from ontologies or natural language processing can produce innovative approaches to IV and are likely to evolve well beyond the current approaches focused on GO (e.g. SNOMED, Tissue Ontology, etc.). Challenges derived from the massive volume and extreme complexity of biological information make the development of IV techniques in bioinformatics a critical path for discovery science.

Acknowledgments

The authors thank Spotfire Company and Human-Computer Interaction Laboratory at University of Maryland for providing software to produce the figures in this paper. We also thank Baehrecke Laboratory for sharing microarray data on the web. This study is partially supported by the National Institute for Allergy and Infectious Disease Grant

#1U54 AI 57159-01, by the National Library of medicine Grant ## R01 LM007659-01, by NLM grant #1K22 LM008308-01 and by the NYSTAR grant # 5-67674.

References

1. Shneiderman, B. The eyes have it: a task by data type taxonomy for information visualizations. Proceedings 1996 IEEE Symposium on Visual Languages; p. 336-43.
2. Keim, DA.; Ankerst, M. Visual Data Mining and Exploration of Large Databases. Tutorial at PKDD'2001; 2001; Freiburg, Germany.
3. Chi, EH. A taxonomy of visualization techniques using the data state reference model. Proceedings of IEEE Visualization; 2000. p. 69-75.
4. Pfitzner, D., et al. A Unified Taxonomic Framework for Information Visualization. Conferences in Research and Practice in Information Technology; Adelaide, Australia. 2003. p. 57-66.
5. Helt GA, et al. BioViews: Java-Based Tools for Genomic Data Visualization. Genome Res 1998;8(3): 291-305. [PubMed: 9521932]
6. Pook S, et al. Zomit: biological data visualization and browsing. Bioinformatics 1998;14(9):807-14. [PubMed: 9918951]
7. Can T, et al. FPV: fast protein visualization using Java 3D. Bioinformatics 2003;19(8):913-22. [PubMed: 12761052]
8. Krieger CJ, et al. MetaCyc: a multiorganism database of metabolic pathways and enzymes. Nucleic Acids Res 2004;32(Database issue):D438-42. [PubMed: 14681452]
9. Goesmann A, et al. PathFinder: reconstruction and dynamic visualization of metabolic pathways. Bioinformatics 2002;18(1):124-9. [PubMed: 11836220]
10. Karp PD. Pathway Databases: A Case Study in Computational Symbolic Theories. Science 2001;293 (5537):2040-2044. [PubMed: 11557880]
11. Kolpakov F, et al. GeneNet: a gene network database and its automated visualization. Bioinformatics 1998;14(6):529-537. [PubMed: 9694992]
12. Steffen M, et al. Automated modelling of signal transduction networks. BMC Bioinformatics 2002;3 (1):34. [PubMed: 12413400]
13. Bhalla US. The chemical organization of signaling interactions. Bioinformatics 2002;18(6):855-63. [PubMed: 12075021]
14. Schacherer F, et al. The TRANSPATH signal transduction database: a knowledge base on signal transduction networks. Bioinformatics 2001;17(11):1053-7. [PubMed: 11724734]
15. Zhang L, et al. VizStruct: exploratory visualization for gene expression profiling. Bioinformatics 2004;20(1):85-92. [PubMed: 14693813]
16. Eisen MB, et al. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 1998;95(25):14863-8. [PubMed: 9843981]
17. Nikkila J, et al. Analysis and visualization of gene expression data using self-organizing maps. Neural Netw 2002;15(8-9):953-66. [PubMed: 12416686]
18. Lang, U., et al. Visualization-related metadata. Data Issues for Data Visualization IEEE Visualization '95 Workshop; p. 26-34.
19. Han K, et al. A vector-based method for drawing RNA secondary structure. Bioinformatics 1999;15 (4):286-297. [PubMed: 10320396]
20. Han K, et al. PseudoViewer: automatic visualization of RNA pseudoknots. Bioinformatics 2002;18 (90001):321S-328.
21. Evers D, Giegerich R. RNA movies: visualizing RNA secondary structure spaces. Bioinformatics 1999;15(1):32-37. [PubMed: 10068690]
22. Wang Y, et al. Cn3D: sequence and structure views for Entrez. Trends Biochem Sci 2000;25(6):300-2. [PubMed: 10838572]
23. Can T, et al. FPV: fast protein visualization using Java 3DTM. Bioinformatics 2003;19(8):913-922. [PubMed: 12761052]
24. Echols N, et al. MolMovDB: analysis and visualization of conformational change and structural flexibility. Nucl Acids Res 2003;31(1):478-482. [PubMed: 12520056]

25. Holm L, Sander C. Touring protein fold space with Dali/FSSP. *Nucleic Acids Res* 1998;26(1):316–9. [PubMed: 9399863]
26. Catherinot V, Labesse G. ViTO, a tool for refinement of protein sequence-structure alignments. *Bioinformatics* 2004:bth429.
27. Yeung KY, Ruzzo WL. Principal component analysis for clustering gene expression data. *Bioinformatics* 2001;17(9):763–774. [PubMed: 11590094]
28. Ewing RM, Cherry JM. Visualization of expression clusters using Sammon's non-linear mapping. *Bioinformatics* 2001;17(7):658–659. [PubMed: 11448886]
29. Wang J, et al. Clustering of the SOM easily reveals distinct gene expression patterns: results of a reanalysis of lymphoma study. *BMC Bioinformatics* 2002;3(1):36. [PubMed: 12445336]
30. Hautaniemi S, et al. Analysis and visualization of gene expression microarray data in human cancer using self-organizing maps. *Machine Learning* 2003;52
31. Baehrecke EH, et al. Visualization and analysis of microarray and gene ontology data with treemaps. *BMC Bioinformatics* 2004;5(1):84. [PubMed: 15222902]
32. McConnell P, et al. Applications of Tree-Maps to hierarchical biological data. *Bioinformatics* 2002;18(9):1278–9. [PubMed: 12217926]
33. Cheng J, et al. NetAffx Gene Ontology Mining Tool: a visual approach for microarray data analysis. *Bioinformatics* 2004;20(9):1462–1463. [PubMed: 14962933]
34. Zhang B, et al. GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. *BMC Bioinformatics* 2004;5(1):16. [PubMed: 14975175]
35. Zeeberg BR, et al. GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol* 2003;4(4):R28. [PubMed: 12702209]
36. Doniger SW, et al. MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biol* 2003;4(1):R7. [PubMed: 12540299]
37. Bouton CMLS, Pevsner J. DRAGON View: information visualization for annotated microarray data. *Bioinformatics* 2002;18(2):323–324. [PubMed: 11847082]
38. Herrero J, et al. New challenges in gene expression data analysis and the extended GEPAS. *Nucleic Acids Res* 2004;32(Web Server issue):W485–91. [PubMed: 15215434]
39. Kim J, et al. ChromoViz: multimodal visualization of gene expression data onto chromosomes using scalable vector graphics. *Bioinformatics* 2004;20(7):1191–1192. [PubMed: 14764551]
40. Loraine AE, Helt GA. Visualizing the genome: techniques for presenting human genome data and annotations. *BMC Bioinformatics* 2002;3(1):19. [PubMed: 12149135]
41. Vernikos GS, et al. GeneViTo: Visualizing gene-product functional and structural features in genomic datasets. *BMC Bioinformatics* 2003;4(1):53. [PubMed: 14594459]
42. Barillot E, et al. The HuGeMap Database: interconnection and visualization of human genome maps. *Nucleic Acids Res* 1999;27(1):119–22. [PubMed: 9847155]
43. Kerkhoven R, et al. Visualization for genomics: the Microbial Genome Viewer. *Bioinformatics* 2004;20(11):1812–1814. [PubMed: 14988111]
44. Ekdahl S, Sonnhammer ELL. ChromoWheel: a new spin on eukaryotic chromosome visualization. *Bioinformatics* 2004;20(4):576–577. [PubMed: 14990454]
45. Kozik A, et al. GenomePixelizer--a visualization program for comparative genomics within and between species. *Bioinformatics* 2002;18(2):335–336. [PubMed: 11847088]
46. Zhang CT, et al. The Z curve database: a graphic representation of genome sequences. *Bioinformatics* 2003;19(5):593–599. [PubMed: 12651717]
47. Lewis SE, et al. Apollo: a sequence annotation editor. *Genome Biol* 2002;3(12):RESEARCH0082. [PubMed: 12537571]
48. Sun H, Davuluri RV. Java-based application framework for visualization of gene regulatory region annotations. *Bioinformatics* 2004;20(5):727–734. [PubMed: 14751988]
49. Gouet P, et al. ESPript: analysis of multiple sequence alignments in PostScript. *Bioinformatics* 1999;15(4):305–308. [PubMed: 10320398]
50. Mayor C, et al. VISTA : visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics* 2000;16(11):1046–1047. [PubMed: 11159318]

51. Shah N, et al. Phylo-VISTA: interactive visualization of multiple DNA sequence alignments. *Bioinformatics* 2004;20(5):636–643. [PubMed: 15033870]
52. Sonnhammer EL, Durbin R. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* 1995;167(1-2):GC1–10. [PubMed: 8566757]
53. Schwartz S, et al. PipMaker--a web server for aligning two genomic DNA sequences. *Genome Res* 2000;10(4):577–86. [PubMed: 10779500]
54. Taneda A. Adplot: detection and visualization of repetitive patterns in complete genomes. *Bioinformatics* 2004;20(5):701–708. [PubMed: 14764564]
55. Rojdestvenski I. Metabolic pathways in three dimensions. *Bioinformatics* 2003;19(18):2436–2441. [PubMed: 14668228]
56. Batada NN. CNplot: visualizing pre-clustered networks. *Bioinformatics* 2004;20(9):1455–1456. [PubMed: 14871859]
57. Kose F, et al. Visualizing plant metabolomic correlation networks using clique-metabolite matrices. *Bioinformatics* 2001;17(12):1198–1208. [PubMed: 11751228]
58. Jackson AP, Charleston MA. A Cophylogenetic Perspective of RNA-Virus Evolution. *Mol Biol Evol* 2004;21(1):45–57. [PubMed: 12949128]
59. AmiGO. <http://www.godatabase.org/cgi-bin/go.cgi>
60. Rost U, Bornberg-Bauer E. TreeWiz: interactive exploration of huge trees. *Bioinformatics* 2002;18(1):109–114. [PubMed: 11836218]
61. Ruths DA, et al. Arbor 3D: an interactive environment for examining phylogenetic and taxonomic trees in multiple dimensions. *Bioinformatics* 2000;16(11):1003–1009. [PubMed: 11159311]
62. Hughes T, et al. Visualising very large phylogenetic trees in three dimensional hyperbolic space. *BMC Bioinformatics* 2004;5(1):48. [PubMed: 15117420]
63. Demir E, et al. PATIKA: an integrated visual environment for collaborative construction and analysis of cellular pathways. *Bioinformatics* 2002;18(7):996–1003. [PubMed: 12117798]
64. Toyoda T, Wada A. Omic space: coordinate-based integration and analysis of genomic phenomic interactions. *Bioinformatics* 2004;20(11):1759–1765. [PubMed: 15037505]
65. Baker C, et al. GeneVis: simulation and visualization of genetic networks. *Information Visualization* 2003;2
66. De Jong H, et al. Genetic Network Analyzer: qualitative simulation of genetic regulatory networks. *Bioinformatics* 2003;19(3):336–344. [PubMed: 12584118]
67. Grimes, AE.; Futrelle, RP. Text pattern visualization for analysis of biology full text and captions. *Proceedings of the 2003 IEEE Bioinformatics Conference. CSB2003*; p. 648-51.
68. Nagata H, et al. Concept and prototype of protein-ligand docking simulator with force feedback technology. *Bioinformatics* 2002;18(1):140–146. [PubMed: 11836222]
69. Shoop E, et al. Data exploration tools for the gene ontology database. *Bioinformatics*. 2004

Figure 1. Microarray gene expression profiles visualized as colored mosaic and clustered by dendrogram. Similar profiles are arranged close to one another and are linked together in a tree structure. However, the cluster is solely based on mathematical similarity of gene expressions and no biological knowledge is used. Produced using Spotfire software (<http://www.spotfire.com/>).

Microarray data visualized in Treemap in the context of Gene Ontology. This novel hybrid system is representative of a ‘semantic enhancement’ trend in visualization: it consists of a visualization method exploiting ontology-anchoring (preprocessing and organizing data with GO) followed by a conventional clustering (analogous to the one presented in Figure 1). Larger boxes of molecular functions representing higher level GO terms contain smaller ones representing lower level terms. Microarray expression data are represented as colored boxes with gene identities distributed in this hierarchy with their sizes representing expression levels and their colors representing p-values. Thus the overview of expression pattern is presented in the context of GO within only one screen. Corresponding biological interpretations can be easily perceived by identifying the GO boxes that contain brighter and larger gene boxes. Produced using TreeMap 4.1 [68] (<http://www.cs.umd.edu/hcil/treemap/>) and data from [69] (<http://www.umbi.umd.edu/~cbr/baehrecke/sdata2.htm>).

Table 1
Summary of major dimensions in taxonomies of IV techniques

	Dimensions	Dimension values
Techniques	Task[1]	Overview, zoom, filter, details-on-demand, relate, history, extract
	Dynamic/interaction[2]	Data-to-visualization mapping, projection, filter (selection, querying, MagicLens), linking & brushing, zoom, semantic zoom, details-on-demand, virtual reality...
	Data visualization[2]	Geometric techniques, icon-based techniques, pixel-oriented techniques, hierarchical techniques, graph-based techniques and hybrid techniques
	Distortion[2]	Perspective wall, bifocal displays, fisheye views, hyperbolic tree, hyper-box...
	Preprocessing[2]	Techniques for dimension reduction: principal component analysis (PCA), factor analysis, multidimensional scaling (MDS), Self-Organizing Maps (SOM), discrete Fourier transform (DFT), fastmap... Subsetting Techniques: sampling, querying Segmentation techniques: based on attribute values or attribute ranges Aggregation techniques: sum, count, min, max
Data	Data type[1]	1-D data, 2-D data, 3-D data, temporal data, multidimensional data, tree data, network data
	Data relationship[4]	Linear, circular, tree, graph

Table 2
Summary of the major uses of IV techniques in bioinformatics

Fields of bioinformatics	Description	Data	Example systems	Major IV techniques used in examples
Molecular structure	Visualizing 2-D or 3-D structures of DNA, RNA and proteins.	2-D, 3-D image	[19-26]	Task (zoom, filter, details-on-demand)
				Dynamic/interaction (virtual reality)
Expression profile analysis (microarray)	Using IV to cluster microarray data, analysis of gene expression data in the context of chromosome and Gene Ontology	Multidimensional data	[15,16,27-39]	Task (Overview) Geometric techniques (colored mosaic, scatterplot) Hierarchical techniques (graphical tree including dendrogram, expandable tree, treemap) Techniques for dimension reduction (PCA, MDS, SOM, DFT)
Genome and sequence annotation	Provide overview and various levels of details of genomes and annotations. Genome comparison.	1-D linear data, 1-D circular data	[5,6,40-48]	Task (overview, zoom, filter, detail-on-demand, relate, extract) Dynamic/interaction (semantic zoom, magic lens, brushing)
Sequence Analysis	Compare DNA sequences and detect special patterns of sequences	1-D linear data, 1-D circular data	[49-54]	Geometric techniques (line chart, bar chart, Dot plot, percent identity plot)
Molecular pathway	Visualizing metabolic pathways, gene regulation pathways and signal transduction networks.	Network data	[8-14,55-57]	Task (Overview, zoom, filter, detail-on-demand, relate) Dynamic/interaction (virtual reality)
				Graph-based techniques (2-D graph, pre-clustered 2-D graph, 3-D graph)
Ontology and taxonomy, phylogeny	Visualizing hierarchical biological data	Tree data, or Graph data (DAG)	[31-36,58-62,69]	Task (Overview, zoom, filter, detail-on-demand, relate, extract) Hierarchical techniques (graphical tree including dendrogram, expandable tree, TreeWiz, treemap and 3D tree) Distortion techniques (hyperbolic tree)