Interpreting the Omics 'era' Data

Georgios A. Pavlopoulos, Ernesto Iacucci, Ioannis Iliopoulos and Pantelis Bagos

Abstract The analysis and the interpretation of the complex and dynamic biological systems has become a major bottleneck nowadays. The latest high-throughput "omics" approaches, such as genomics, proteomics and transcriptomics and the available data repositories hosting information concerning bioentities and their properties grow exponentially in size over time. Therefore, to better understand biological systems as a whole and at a higher level, visualization is a necessity as clear and meaningful views and intuitive layouts can give a better insight into coping with data complexity. The implementation of tools to maximize user friendliness, portability and provide intuitive views is a difficult task and still remains a hurdle to overcome. In this chapter, we present a variety of significant visualization tools as they specialize in different topics covering different areas of the broad biological spectrum varying from visualization of molecular structures to phylogenies, pathways, gene expression, networks, and next generation sequencing. We emphasize their functionality, the latest research findings, and insights into how these tools could be further developed both in terms of visualization but also in the direction of data integration and information sharing.

G. A. Pavlopoulos (✉) · E. Iacucci
ESAT-SCD/iMinds-KU.Leuven Future Health Department Katholieke,
Universiteit Leuven, Kasteelpark Arenberg 10, box 2446 3001 Leuven, Belgium
e-mail: Georgios.Pavlopoulos@esat.kuleuven.be

E. Iacucci
e-mail: Ernesto.Iacucci@esat.kuleuven.be

I. Iliopoulos
Division of Basic Sciences, University of Crete Medical School, 71110 Heraklion, Greece
e-mail: iliopj@med.uoc.gr

P. Bagos
Department of Computer Science and Biomedical Informatics,
University of Central Greece, 35100 Lamia, Greece
e-mail: pbagos@ucg.gr

G. A. Tsihrintzis et al. (eds.), *Multimedia Services in Intelligent Environments*,
Smart Innovation, Systems and Technologies 25, DOI: 10.1007/978-3-319-00375-7_6,

1 Molecular Structures

Starting with proteins, we introduce the four levels of protein structures. Thus, the *primary* structure refers to the amino acid sequence of the polypeptide chain, the *secondary* structure to highly regular local sub-structures such as α-helices, β - strands or β-sheets, the *tertiary* structure to the 3D structure of a single protein molecule and the *quaternary* structure to the assemblies of several protein molecules or polypeptide chains, usually called subunits in this context. As the 3D structure defines the functionality of a protein, much effort has been made in the past years in order to precisely detect it, mainly using experimental techniques such as X-Ray crystallography, NMR and electron microscopy. Simultaneously, many computational methods try to accurately predict the 3D tertiary structure of a protein given the amino-acid sequence. Today over 60 K solved protein structures are hosted in wwPDB [1] whereas ∼86 % of the structures are derived from X-ray crystallography, ∼13 % from NMR spectroscopy and less than ∼1 % from electron microscopy [2]. Typical resolutions vary from 1.2 to 4 **Å**. Similarly to proteins, ∼4 K solved RNA 3D structures are hosted in NDB [3], whereas 8 % of them correspond to PDB entries [2]. While a great variety of reviews that comment on the visualization approaches for such cases exists [4–6], here we give an overview of what is the status of the cutting-edge research in the field.

Most of the available visualization tools currently try to picture the chemistry of the biomolecules, the atoms and the bonds among them. Different representations include ribbons, space-filling atoms, ball-and-stick and others. Coloring schemes are used in order to highlight important parts of a protein such as binding sites, atoms with certain physiochemical characteristics, SNPs, active sites, different chains, exon boundaries or whole domains. Despite it is beyond the scope of this chapter to analyze all of the available visualization tools below we give some example of representative tools that are widely used and we try to categorize them according to their functionality (Table 1). Despite presenting the tools as a non-redundant list, many of them share functionalities and characteristics. For example, PyMol [7], Jmol [8], KiNG or Mage [9] offer typical views and can be incorporated in a web page. Others such as Chimera [10], SRS 3D [11], STRAP [12], Cn3D [13–15] or PdbViewer [16, 17] are able to combine the 3D structural visualization in space with the linear amino acid sequence (Fig. 1). They are highly interactive and therefore users can select regions in any of the two views and highlight the corresponding area in the other view. For example, when a sequence region is predicted to be functional or when a part of it is aligned to another sequence of interest, the 3D structural components are highlighted. This way one can look at the region of interest either from a linear or a structural point of view. Tools such as Molscript [18], PMV [19], VMD [20], ICM-Browser [21] and plusRaster3D [22] export images at a high dpi quality to be used for scientific publications. In order to superimpose two proteins and compare them directly in 3D space, tools such as MOLMOL [23], MOE, VMD [20] or PyMol [7] are suitable. In cases where computationally expensive superimposition is required,

external CPU intensive packages such as STAMP [24], STRAP [12] or THESEUS [25] are recommended. Cases that require advanced computational power exist when one wants to superimpose very large regions (high size complexity) or sequences with low sequence similarity (many possible combinations). Looking at other characteristics such as hydrophobicity, electrostatics, residue conservation or connolly surfaces, MSMS software [26] is the most widely used. In order to show annotations from databases that are related to a certain part of the structure, tools such as ProSAT2 [27], JenaLib [28], PDBsum [29], SYBYL, Swiss-PdbViewer [17] or WHAT IF [30] can be used. Tools like Relibase [31, 32] and Superligands [33] can directly compare smaller molecules such as ligands between each other simultaneously. Notably, while tools such as tCONCOORD [34] and FIRST/FRODA [35] are able to picture conformational changes, Moviemaker [36] and Yale Morph [37] server applications can show two different transition stages of the same molecule. NOMAD-ref [38] and ANM [39] are can combine many transition stages but only for low frequency events. Despite the fact that few of the aforementioned tools such as PyMol [7] are also suitable for RNA structure visualization, specialized tools such as S2S Assemble [40] are implemented for RNAs.

Despite the fact that visualization of macromolecular structures is today very mature compared to other areas in biology, the current rendering techniques still lack the computational capacity to process more complex systems such as protein complexes or protein interactions at very high resolutions. In addition, molecular dynamics, simulations and motion are difficult to picture at such levels of detail, as the current tools are CPU greedy for more advanced analysis when visualization of more than one molecule per time is required. In order to come closer to a physical model and combine the chemistry-based visualization with real images from electron or cryo-microscopy great effort should still to be done in that direction towards the generation of real and more informative prototypes. In terms of data integration, tools are still available as standalone applications but a great variety of them can run as a part of a web page and come with standardized file formats and services to increase portability and data exchange.

Table 1 Software tools in the area of proteomics

Software	URL
Chimera	http://www.cgl.ucsf.edu/chimera/
FirstGlance	http://firstglance.jmol.org/
ICM-Browser	http://tinyurl.com/icm-browser/
JenaLib	http://tinyurl.com/JenaLib/
Jmol	http://www.jmol.org/
KiNG	http://tinyurl.com/KiNGapp/
Mage	http://tinyurl.com/kinemage/
MOE	http://www.chemcomp.com/
MOLMOL	http://tinyurl.com/molmol1/
Molscript	http://www.avatar.se/molscript/
NDB	http://ndbserver.rutgers.edu/

(continued)

Table 1 (continued)

Software	URL
PDBe	http://www.ebi.ac.uk/pdbe/
PDBsum	http://www.ebi.ac.uk/pdbsum/
PMV	http://tinyurl.com/PMV-MGL/
ProSAT	http://tinyurl.com/ProSAT2/
PyMOL	http://www.pymol.org/
RasMol	http://www.rasmol.org/
Raster3D	http://tinyurl.com/raster3d/
Relibase	http://tinyurl.com/relibase/
RSCB PDB	http://www.pdb.org/
SRS 3D	http://SRS3D.org/
STRAP	http://tinyurl.com/STRAP1/
Swiss-Model	http://swissmodel.expasy.org/
Swiss-PdbViewer	http://spdbv.vital-it.ch/
SYBYL	http://tinyurl.com/triposSYBYL/
VMD	http://tinyurl.com/VMD-viewer/
Chimera	http://www.cgl.ucsf.edu/chimera/
FirstGlance	http://firstglance.jmol.org/

2 Tree Hierarchies

Tree data structures and representations are widely used in biological studies in order to show hierarchies of data [41]. These include for example the Gene Ontologies (GO) [42] to describe functional annotation of genes via a hierarchically organized set of terms or the Unified Medical Language System (UMLS) [43] which serves a similar function for biomedical notions.

Another very important area of biology raises the topic of investigating and visualizing the evolution between the species. Thus, evolutionary studies try to reveal and understand how different species evolved over time and whether two different species have a common ancestor and at which time point. To picture these evolutionary transitions, phylogenetic trees are mainly used. A prime example of such tree representations is the so-called tree of life [44] which displays such evolutionary relationships between species and how they have separated and over millennia. From about $\sim$1.7 million identified species, only $\sim$80,000 of them have been analyzed for evolutionary relationships and have been assigned into a hierarchy [45] (Fig. 2).

Other areas in biology that involve high-throughput technologies such as Chip–Chip arrays, microarrays, next generation sequencing or proteomics often use tree-based clustering algorithms to interpret and visualize their results. In the case of microarrays [46–48] for example, genes are clustered according to their expression patterns in order to see which of them are correlated or anti-correlated. When one compares a healthy with a non-healthy tissue, the purpose is to find which of them

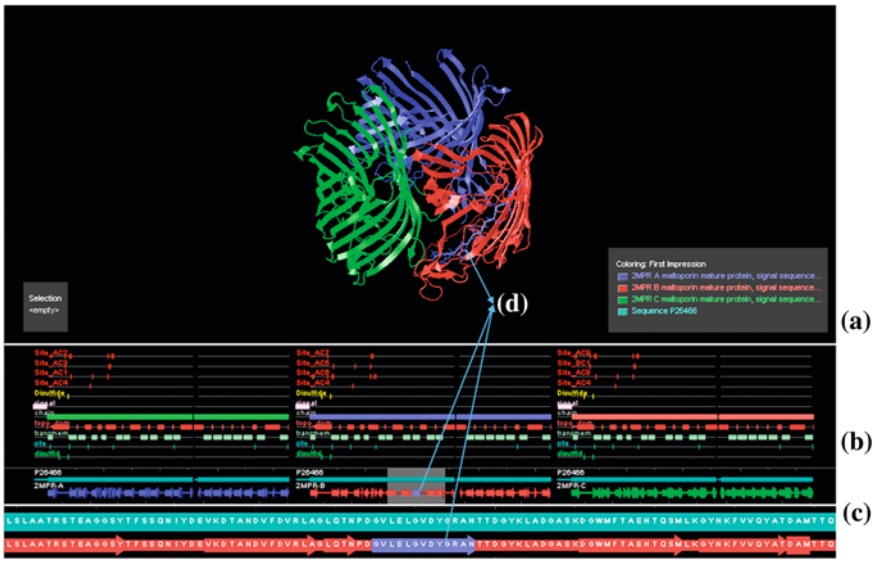

Fig. 1 P04637 (P53_HUMAN) tumor suppressor protein visualized by SRS3D application. **a** 3D structure representation of the three chains of P04637 as ribbons using three different colors. **b** Three different columns show the domains of the three different chains from different databases individually. **c** The sequence of the protein in a linear form. **d** Interactivity enables the highlighting of a chosen domain in every view (sequence and 3D structure). Switching between different representations, the 3D chemical structure of the specific domain is highlighted and visualized as a coil

are up- or down-regulated. Algorithms that are widely used, include the single linkage, average linkage, complete linkage [49], UPGMA [50], Neighbor Joining [51, 52] etc. (Figure 2).

In addition, in the field of sequence analysis, biologists try to determine the similarity between two protein or nucleotide sequences. For a given set of sequences, often a multiple alignment or an all-against-all pairwise alignment is performed constructing a distance matrix that hosts the similarity scores between every pair of genes. Notably, widely used applications that perform such analyses include the Clustal W [53], MUSCLE [54], BLAST [55], and the T-Coffee suite [56]. In order to classify these sequences in families, a clustering algorithm is applied based on the constructed similarity matrix by bringing together those sequences that are closely related to each other. The post-clustering results are visualized using a tree hierarchy (Fig. 2).

While a variety of computer readable formats exist, most phylogenetic trees are described using either the New Hampshire/Newick [57], the NHX extended Newick file format or the Nexus [58] file format. In terms of tree annotation and information sharing across repositories, Markup languages such as phyloXML [59] and NeXML are of demand.

Although the most common representation of hierarchies is a tree representation based on a 2D Euclidean drawing [60], treemaps [61] which present a tree hierarchy as nested rectangles serve as an alternative as they are often best suited for classifications rather than phylogenies [62]. While tree visualization is today a mature area, the growth of taxa is still a limiting factor as the space to represent such huge hierarchies on a single screen is insufficient. Traditional viewers that have been in use for many years such ATV [63] or TreeView [64] are nowadays weak for displaying huge taxonomies with thousands of data such as [65]. To overcome this problem, several approaches have been proposed (Table 2). One approach is the implementation of efficient zooming. Thus, as users zoom in or out, nodes collapse or expand respectively. Tools that try to compress the information into a given smaller canvas include DOI trees [66], space trees [67] and expand-ahead browsers [68]. Another approach that tools such as HyperTree [69] follow, is to project data on hyperbolic space [70]. While, this idea is very efficient in terms of visualization, in practice users find these views difficult to navigate [61]. Preferred tree visualization on the other hand involve radial layouts like those found in iTOL [71], TreeDyn [72], TreeVector [73], or Dendroscope [74]. A third

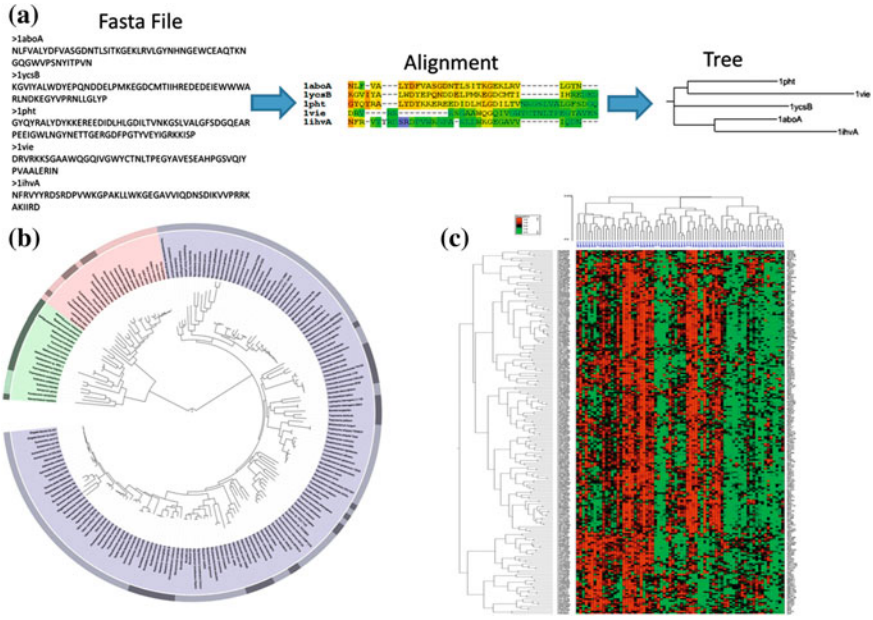

Fig. 2 Examples of tree hierarchies in Biology **a** 5 protein sequences were aligned with TCoffee and clustered according to their sequence similarity. The clustering results are shown as a tree hierarchy. **b** The Tree of Life presented in [44]. **c** Example of a gene expression heatmap. The expression levels of several genes (*tree hierarchy on the left*) were measured across several conditions (*tree hierarchy on the bottom*) using the Expander software. Genes and conditions were clustered using the average linkage hierarchical clustering. Dense red or green areas show the correlations between the genes and the experimental conditions

approach that tools such as Paloverde [75] and the Wellcome Trust Tree of Life follow, involves the utilization of 3D space. Despite the fact that such an approach is less disorientated compared to hyperbolic viewers it is still not preferred to single 2D visualization with an exception in the visualization of geophylogenies where geographical and phylogenetic information is combined towards the implementation of geographic information systems [76]. Based on such approaches, in Biology, georeferenced barcode DNA sequences are likely to become more widely used [77].

To directly compare two trees between each other so far methods such as tanglegram alignment have been proposed [78]. According to this methodology, two trees are mirrored against each other and lines connect the leaves that are equivalent to each other. Alternatively, color schemes can highlight the taxonomies that are different between each other. As tree hierarchies can vary in size and host overloaded information of thousands of taxa, direct comparison, navigation and exploration still remain a problem as the aforementioned approaches succeed in efficiently organizing the data but often fail to visually deliver them to the user in efficient ways. An example is the visualization of the tree of life versus the visualization of the forest of life [79]. While image tiling [80] methods to generate large images and then break them into smaller pieces at different resolutions (Google Earth) and recombine them could be of a solution, further opportunities for further investigation are still available in this respect.

Table 2 Tools to represent hierarchies

Software	URL
ATV	http://phylogeny.lirmm.fr/phylo_cgi/
Dendroscope	http://ab.inf.uni-tuebingen.de/software/dendroscope
Hypertree	http://kinase.com/tools/HyperTree.html
iTOL	http://itol.embl.de/
Paloverde	http://loco.biosci.arizona.edu/paloverde/paloverde.html
PhyloExplorer	http://www.ncbi.orthomam.univ-montp2.fr/phyloexplorer/
TreeDyn	http://www.treedyn.org/
TreeVector	http://supfam.cs.bris.ac.uk/TreeVector/
TreeView	http://taxonomy.zoology.gla.ac.uk/rod/treeview.html

3 Next Generation Sequencing

Recent technological improvements have led to great steps towards the understanding of the genome, its genes, their expression and their function. While the Human Genome Project (1990–2003) allowed the release of the first human reference genome by determining the sequence of ~3 billion base pairs and identifying the approximately ~25,000 human genes [81–83], current technologies allow the sequencing of a whole exome in a few days and at a very low cost. The

first generation sequencing technique was discovered back in 1977 and is known as the Sanger (dideoxy) [24] technique. High-throughput second generation technologies have already been developed by Illumina [84], Roche/454 [85] and Biosystems/SOLiD [86] while Helicos BioSciences [87], Pacific Biosciences [88], Oxford Nanopore [89] and Complete Genomics [90] belong to the third generation of sequencing techniques. Similarly to DNA sequencing, RNA Sequencing [91, 92] which allows today the simultaneous gene expression measures in a cell and ChIP-Sequencing which uses immunoprecipitation with massively parallel DNA sequencing to mainly identify DNA regions that are binding sites for proteins such as transcription factors [93] are now more feasible and more accurate due to the rapid technological advantages as the aforementioned. Projects like the *1000 Genomes Project* (started in 2008) to sequence a large number of human genomes and provide a comprehensive resource for human genetic variation [94] and the *International HapMap Project* [95–99] to identify common genetic variations among people from different countries show the broad spectrum of the application of such technologies and the scale of the data that they can process.

Advances in high throughput next generation sequencing techniques allow the production of vast amounts of data in different formats that currently cannot be analyzed in a non-automated way. Visualization approaches are today called to cope with huge amounts of data, efficiently analyze them and deliver the knowledge to the user in a visual, easier to grasp, way. User friendliness, pattern recognition and knowledge extraction are the main targets that an optimal visualization tool should excel in. Issues such as de novo genome assemblies, SNP identification, visualization of structural variations, whole genome alignment, alignment of short reads, comparisons between several genomes simultaneously, alignment of unfinished genomes, intra/inter chromosome rearrangements, identification of functional elements and display of sequencing data and genome annotations are still open fields for visualization. Therefore, tasks like handling the overload of information, displaying data at different resolutions, fast searching or smoother scaling and navigation are not trivial when the information to be visualized consists of millions of elements and reaches an enormously high complexity. Modern libraries, able to scale millions of data points smoothly and visualize them using different resolutions are essential. While established genome browsers (Fig. 3) such as Ensembl [100, 101], UCSC Genome Browser [102] and IGV [103] are able to partially address some of the aforementioned challenges, visualization of genomic data in this respect is still an underdeveloped field.

4 Network Biology

In Systems and Integrative Biology, often bioentities are interconnected with each other and are represented as networks where nodes (bioentities) are linked with edges. Several categories of different biological networks already exist [104] such as protein–protein interactions networks, signal transduction networks, pathways, knowledge and integration networks (where bioentities are found to be related in

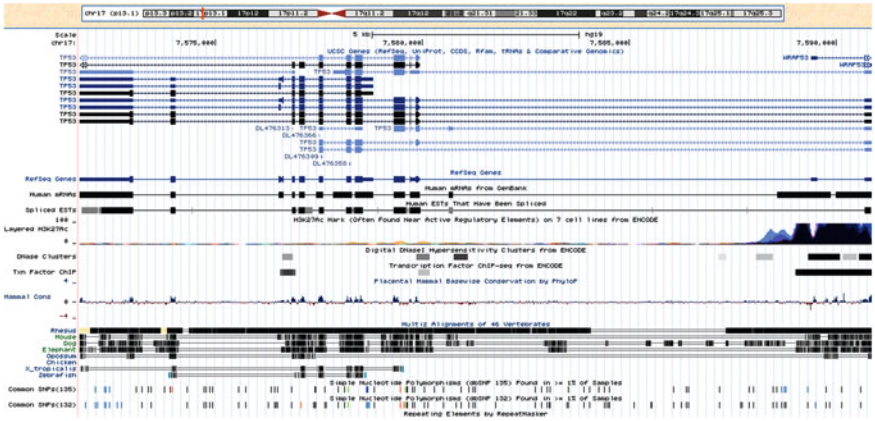

Fig. 3 P04637 (P53_HUMAN) tumor suppressor protein is found in Chromosome 17 in positions: 6,375,874–878,791,773 and visualized by the UCSC Genome browser at the highest resolutions. While a red mark shows where TP53 is in the chromosome information about alignments, SNPs, mRNA coding regions and others are shown. Notably, one can interactively zoom in and out to see the information even at the lowest nucleotide level

literature or in records of public repositories), metabolic and biochemical networks, or gene regulatory networks which picture the factors that control gene expression.

While it is not within the scope of this section to present a thorough review of available repositories for each individual network category, we shortly mention experimental, computational and high throughput techniques to detect protein–protein interactions in order to give an overview of a few of the available repositories to demonstrate the size complexity of the available data and their heterogeneity. Thus, the most widely used experimental methods include pull down assays [105], tandem affinity purification (TAP) [106], yeast two hybrid systems (Y2H) [107], mass spectrometry [108], microarrays [109] and phage display [110]. Furthermore, computational methods such as MCODE [111], jClust [112], Clique [113], LCMA [114], DPClus [115], CMC [116], SCAN [117], Cfinder [118], GIBA [119] or PCP [120] are graph-based algorithms that use graph theory to detect highly connected subnetworks. DECAFF [121], SWEMODE [122] or STM [123] have been developed to predict protein complexes incorporating graph annotations, whereas others like DMSP [124], GFA [125] and MATISSE [126] also take the gene expression data into account. A very useful review article that describes and compares the aforementioned techniques can be found in [127].

Of course, such biological networks share common characteristics but they can differ significantly in their topology and properties such as for example the number of their highly connected nodes or regions, their average eccentricity, betweeness or other types of centralities, shortest paths or their clustering coefficient. Protein–protein interaction networks tend to have hubs as signal transduction networks do not. Today, there exists a wide variety of tools (Table 3) that are network specific

as reviewed in [104, 128, 129], but the field of network visualization is an active fields with many challenges to be addressed as the amount of data increases exponentially and the annotation databases expand continuously.

Currently the most widely used network representations include node-link diagrams where bioentities are represented as nodes and the interactions between them as edges forming a hairball or distance or similarity matrices which hold information about every pairwise relationship with size $N(N-1)/2$ and hybrid views that combine the two previous ones. While matrices are often preferred for larger scale networks, all of the aforementioned approaches suffer in terms of scaling when the size of the network consists of few thousands of nodes and edges. In order to make large scale biological networks more informative, several layout algorithms [130] try to reveal the properties of the network such as showing the hubs using a force-directed algorithm and simultaneously try to minimize the crossovers between the lines. Similar to node-link diagrams, various ordering algorithms try to efficiently order the columns and the rows of a distance matrix to

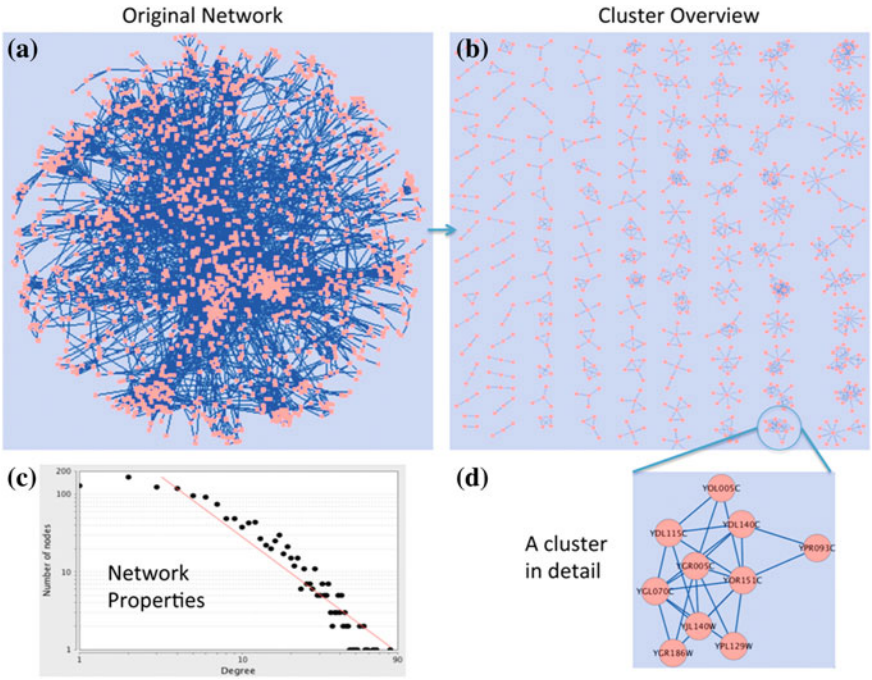

Fig. 4 A Yeast protein–protein interaction network [147] consisting of ~1600 proteins is analyzed by Cytoscape. **a** A force-directed layout algorithm is applied on the network. **b** The network was clustered according to MCL algorithm and ~240 clusters were produced. **c** The zooming functionality enables the user to the cluster and the node's labels in detail. **d** Connectivity degree versus number of nodes is plotted to show some network characteristics. Notably, different combinations of node properties can also be plotted such as the clustering coefficient (0.42 for the specific network) versus network centralization (0.053 for this network)

make highly connected regions more visible. While established tools such as Ondex [131], Pajek [132], Cytoscape [133], Medusa [134], VisANT [135] for real world-like networks or iPath [136], PATIKA [137], PathVisio [138] for pathways or EXPANDER [139], HCE [140], ExpressionProfiler [141] for expression data implement such layouts most of them try to project data a 2D plane using advanced navigation techniques to make data exploration easier. Other tools such as Arena3D [142, 143] or BioLayout Express 3D [144] take advantage of 3D space to show data in a universe. While still very few of the aforementioned tools try to fill the gap between analysis and visualization, efforts have been made the past years. ClusterMaker [145] Cytoscape's plugin and jClust [112] applications for example try to cluster the data within the application without the help of an external application (Fig. 4). Similarly, CentiBiN application [146] tries to compute and visualize different vertex and graph centrality measures.

While network visualization is a developing area, there is much space for improvements as for example visualization of time series data, network evolution and dynamics are still important features to be visually represented. Similarly, node aggregation, edge bundling, faster and more efficient layout algorithms and their extension into 3D space, multi-dimensional data visualization, semantic zooming, interactivity and data integration still remain open problems in network biology.

Table 3 Tools in network biology

Software	Type	URL
Arena3D	Network	http://www.arena3d.org/
BioLayout Express 3D	Network	http://www.biolayout.org/
Cytoscape	Network	http://www.cytoscape.org/
Medusa	Network	https://sites.google.com/site/medusa3visualization/
Ondex	Network	http://www.ondex.org/
Osprey	Network	http://tinyurl.com/osprey1/
Pajek	Network	http://pajek.imfm.si/
STITCH	Network	http://stitch.embl.de/
VisANT	Network	http://visant.bu.edu/
iPath	Pathways	http://pathways.embl.de/
Patika	Pathways	http://www.patika.org/
PathVisio	Pathways	http://www.pathvisio.org/
EXPANDER	Expression	http://acgt.cs.tau.ac.il/expander/
ExpressionProfiler	Expression	http://tinyurl.com/expespro/
HCE	Expression	http://tinyurl.com/HCExplorer/

5 Visualization in Biology—the Present and the Future

In the aforementioned sections we widely discussed visualization tools which may be applied on different biological areas of the "omics" spectrum. These mainly include software for genome analysis, microarrays, molecular structures,

phylogenies, alignments and network biology. Despite the tremendous efforts for the development of better, more efficient, more interactive and user friendlier visualization tools which has been going on over approximately the past 20 years [148] and despite the fact that all of these tools share common characteristics, the future challenges partially overlap and many difficulties still need to be addressed.

So far, there is a tendency to produce tools that mainly run as standalone applications being able to read their own file formats. While this has slowly changed over time, it is still a limiting factor, as integration needs to come to the foreground. Thus, visualization tools should share common human and computer readable file formats in order to easier exchange information. Such a demand for integration can be partially solved whenever each tool is launched with its own API or by implementing specific web services for data exchange. In addition, it is highly recommended to make tools directly available through a web interface (i.e. Flash, JavaFX, Processing.org, Applets) or directly make them downloadable through other technologies such as JNLP (Java web start) in the case of a Java implementation. Such an effort for integration would greatly help to further bridge the gap between analysis and visualization as visualization tools often use external packages to perform a typical analysis that is not embedded in the tool. A visible example of such a gap can be observed whenever one works with network biology where the nodes usually represent bioentities and the edges the connections between those. As such networks can increase in size and complexity, clustering analysis to categorize data and investigate the clusters individually is often in demand. Unfortunately, today very limited number of visualization tools hosts such functionality to cluster data within a visual application. Another example can be given for genomic data analysis where tasks such as SNP and variation calling, genome assembly or genome alignments should initially be performed individually and sequentially, the results of the analysis should be visualized by different tools after reformatting them to the tool-specific input format. In conclusion, it is expected in the future, the visualization tools will follow golden standards both in terms of data storage, analysis and integration (as to manually merge software packages and combine their functionalities requires some expertise, something that is tedious and time consuming). A first step would provide tools with a pluggable architecture where users can implement their own plugin for a tool based on their own expertise.

During the past 10 years, a progress has been made to move away from static images and cope with the increasing size complexity by handling biological data interactively. This includes operations such as efficient zooming, panning and navigation. Noticeably, multi-touch screens today encourage more modern and less conservative interfaces to handle multiple events simultaneously to increase interactivity. Similarly, vibrations could potentially be used to get the attention of the user when a property or a characteristic of the system changes. A characteristic example is the MacOS systems where icons start to vibrate in order to indicate that the corresponding process is running. Apart from these operations, in biology, often data need to be explored at multiple-scales and at different resolutions. Similarly to GoogleEarth application, which can be used to explore maps from

different heights, one could imagine the biological world as a universe that can be observed from the organism to a cell or to an atomic level. In the case of genome browsers for example, a genome can be explored at a chromosome, at a gene or even at a single nucleotide resolution. Similarly, in biological networks, node aggregation or edge bundling methodologies could be applied on the network while exploring it at different levels. In order to explore multi-scaled data, pre-processing and pre-indexing is normally required, as the enormous amount of data does not allow such calculations on the fly. GoogleEarth application is a great approach to be followed as real-world images that refer to a specific resolution are indexed and stored in a database and get loaded on the fly upon users request.

Besides user interface challenges, progress in biological data management has been made over the past years. Current technologies, infrastructures and architectures allow the parallel processing of information at significantly lower costs. Unfortunately, not many visualization tools for biology today are engaged to these technologies with an exception being the tools that are implemented for biological image analysis as in the case of microscopy. Taking advantage of libraries like CUDA, which allows parallel calculations at multiple Graphical Processing Units (GPUs), protocols like Message Passing Interface (MPI) to distribute computational tasks to computers over the network or other multi-core supercomputers with multiple CPUs are ways to significantly reduce the processing time and the running time complexity of huge-scale data. Similar to architectures, display hardware such as large screens, tiled arrays or virtual reality environments, which take advantage of a very large space to project data, should be taken into consideration by programmers and designers as they become more and more affordable over time. A great advantage of such technologies is that they allow the representation of the dataset as a whole without the need of algorithms to project data at lower dimensions, something that can lead to information loss.

As visualization in biology evolves rapidly, a great variety of new visualization concepts and representations appear. While this is encouraging and it can become a source of inspiration for other fields such as economics, physics, environmental or social studies, golden standards concerning the design, the interactivity and the prototyping should be strictly defined, aiming to maximize human–computer interaction. In addition, as visualization tools are designed for a broad range of users, prototypes should take into consideration rare cases like the careful choice of color schemes as 10 % of the population suffers from color-blindness.

As biological systems are highly dynamic, visualization tools to capture the behavior and property changes of such systems and how they evolve over time is a necessity. Approaches that picture how the properties of a system change, currently include parallel coordinates, 3D representations using multi-layered graphs or animations. The effectiveness of the animation-approach is however often very low and limited by human perception capabilities, mainly due to changes in the user's mental map of the structure. More efficient approaches should be implemented to tackle this problem, as time-series visualization for biology is still a very immature field.

Finally, an ideal visualization system of the future should be able to track users preferences and learn users behavior while he or she explores specific data types. After training, such a system could guess and suggest possible solutions that anticipate the users preference, something that would minimize the time–cost to solve a problem. SVMs, SOMs, neural networks and other approaches have significantly evolved and can be used as initial steps for such user profiling. Concerning data parameterization, today visual analytics approaches that require human judgment are followed as data properties and results can vary significantly as one changes the parameters of a software or workflow. Automation of such procedures like optimal parameterization finding and profiling still remain a bottleneck.

Acknowledgments We would like to acknowledge Dr. Theodoros Soldatos, Dr. Charalampos Moschopoulos and Mgr. Izabella Januszewska for their valuable input. This work was supported by the Greek State Scholarship Foundation (I.K.Y—http://www.iky.gr/IKY/portal/en).

References

1. Berman, H., Henrick, K., Nakamura, H.: Announcing the world wide Protein Data Bank. Nat. Struct. Biol. **10**(12), 980 (2003)
2. O'Donoghue, S.I., Goodsell, D.S., Frangakis, A.S., Jossinet, F., Laskowski, R.A., Nilges, M., Saibil, H.R., Schafferhans, A., Wade, R.C., Westhof, E., Olson, A.J.: Visualization of macromolecular structures. Nat. Methods **7**(Suppl 3), S42–S55 (2010)
3. Berman, H.M., Olson, W.K., Beveridge, D.L., Westbrook, J., Gelbin, A., Demeny, T., Hsieh, S.H., Srinivasan, A.R., Schneider, B.: The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids. Biophys. J. **63**(3), 751–759 (1992)
4. Tate, J.: Molecular visualization. Methods Biochem. Anal. **44**, 135–158 (2003)
5. Olson, A.J., Pique, M.E.: Visualizing the future of molecular graphics. SAR QSAR Environ. Res. **8**(3–4), 233–247 (1998)
6. Goddard, T.D., Ferrin, T.E.: Visualization software for molecular assemblies. Curr. Opin. Struct. Biol. **17**(5), 587–595 (2007)
7. PyMol: http://www.pymol.org/
8. jMol: http://jmol.sourceforge.net/
9. Richardson, D.C., Richardson, J.S.: The kinemage: A tool for scientific communication. Protein Sci. **1**(1), 3–9 (1992)
10. Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., Ferrin, T.E.: UCSF Chimera–a visualization system for exploratory research and analysis. J. Comput. Chem. **25**(13), 1605–1612 (2004)
11. O'Donoghue, S.I., Meyer, J.E., Schafferhans, A., Fries, K.: The SRS 3D module: Integrating structures, sequences and features. Bioinformatics **20**(15), 2476–2478 (2004)
12. Gille, C., Frommel, C.: STRAP: Editor for STRuctural alignments of proteins. Bioinformatics **17**(4), 377–378 (2001)
13. Porter, S.G., Day, J., McCarty, R.E., Shearn, A., Shingles, R., Fletcher, L., Murphy, S., Pearlman, R.: Exploring DNA structure with Cn3D. CBE Life Sci. Educ. **6**(1), 65–73 (2007)
14. Wang, Y., Geer, L.Y., Chappey, C., Kans, J.A., Bryant, S.H.: Cn3D: Sequence and structure views for Entrez. Trends Biochem. Sci. **25**(6), 300–302 (2000)

15. Hogue, C.W.: Cn3D: A new generation of three-dimensional molecular structure viewer. Trends Biochem. Sci. **22**(8), 314–316 (1997)
16. Guex, N., Peitsch, M.C., Schwede, T.: Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: A historical perspective. Electrophoresis **30**(Suppl 1), S162–S173 (2009)
17. Guex, N., Peitsch, M.C.: SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. Electrophoresis **18**(15), 2714–2723 (1997)
18. Esnouf, R.M.: An extensively modified version of MolScript that includes greatly enhanced coloring capabilities. J. Mol. Graph. Model. **15**(2), 132–134, 112–133 (1997)
19. Sanner, M.F.: A component-based software environment for visualizing large macromolecular assemblies. Structure **13**(3), 447–462 (2005)
20. Humphrey, W., Dalke, A., Schulten, K.: VMD: Visual molecular dynamics. J. Mol. Graph. **14**(1), 33–38, 27–38 (1996)
21. ICM-Browser: http://www.molsoft.com/icm_browser.html
22. Merritt, E.A., Murphy, M.E.: Raster3D version 2.0. A program for photorealistic molecular graphics. Acta Crystallogr. D Biol. Crystallogr. **50**(Pt 6), 869–873 (1994)
23. Koradi, R., Billeter, M., Wuthrich, K.: MOLMOL: A program for display and analysis of macromolecular structures. J. Mol. Graph. **14**(1), 51–55, 29–32 (1996)
24. Russell, R.B., Barton, G.J.: Multiple protein sequence alignment from tertiary structure comparison: Assignment of global and residue confidence levels. Proteins **14**(2), 309–323 (1992)
25. Theobald, D.L., Wuttke, D.S.: THESEUS: Maximum likelihood superpositioning and analysis of macromolecular structures. Bioinformatics **22**(17), 2171–2172 (2006)
26. Sanner, M.F., Olson, A.J., Spehner, J.C.: Reduced surface: An efficient way to compute molecular surfaces. Biopolymers **38**(3), 305–320 (1996)
27. Gabdoulline, R.R., Ulbrich, S., Richter, S., Wade, R.C.: ProSAT2–protein structure annotation server. Nucleic Acids Res. **34**(Web Server issue), W79–W83 (2006)
28. Reichert, J., Suhnel, J.: The IMB Jena image library of biological macromolecules: 2002 update. Nucleic Acids Res. **30**(1), 253–254 (2002)
29. Laskowski, R.A.: Enhancing the functional annotation of PDB structures in PDBsum using key figures extracted from the literature. Bioinformatics **23**(14), 1824–1827 (2007)
30. Vriend, G.: WHAT IF: A molecular modeling and drug design program. J. Mol. Graph. **8**(1), 52–56, 29 (1990)
31. Gunther, J., Bergner, A., Hendlich, M., Klebe, G.: Utilising structural knowledge in drug design strategies: Applications using Relibase. J. Mol. Biol. **326**(2), 621–636 (2003)
32. Hendlich, M., Bergner, A., Gunther, J., Klebe, G.: Relibase: Design and development of a database for comprehensive analysis of protein-ligand interactions. J. Mol. Biol. **326**(2), 607–620 (2003)
33. Michalsky, E., Dunkel, M., Goede, A., Preissner, R.: SuperLigands—a database of ligand structures derived from the Protein Data Bank. BMC Bioinf. **6**, 122 (2005)
34. Seeliger, D., De Groot, B.L.: tCONCOORD-GUI: Visually supported conformational sampling of bioactive molecules. J. Comput. Chem. **30**(7), 1160–1166 (2009)
35. Thorpe, M.F., Lei, M., Rader, A.J., Jacobs, D.J., Kuhn, L.A.: Protein flexibility and dynamics using constraint theory. J. Mol. Graph. Model. **19**(1), 60–69 (2001)
36. Maiti, R., Van Domselaar, G.H., Wishart, D.S.: MovieMaker: A web server for rapid rendering of protein motions and interactions. Nucleic Acids Res. **33**(Web Server issue), W358–W362 (2005)
37. Flores, S., Echols, N., Milburn, D., Hespenheide, B., Keating, K., Lu, J., Wells, S., Yu, E.Z., Thorpe, M., Gerstein, M.: The Database of Macromolecular Motions: New features added at the decade mark. Nucleic Acids Res. **34**(Database issue), D296–D301 (2006)
38. Lindahl, E., Azuara, C., Koehl, P., Delarue, M.: NOMAD-Ref: Visualization, deformation and refinement of macromolecular structures based on all-atom normal mode analysis. Nucleic Acids Res. **34**(Web Server issue), W52–W56 (2006)

39. Eyal, E., Yang, L.W., Bahar, I.: Anisotropic network model: Systematic evaluation and a new web interface. Bioinformatics **22**(21), 2619–2627 (2006)
40. Jossinet, F., Westhof, E.: Sequence to structure (S2S): Display, manipulate and interconnect RNA data from sequence to structure. Bioinformatics **21**(15), 3320–3321 (2005)
41. Pavlopoulos, G.A., Soldatos, T.G., Barbosa-Silva, A., Schneider, R.: A reference guide for tree analysis and visualization. BioData Min. **3**(1), 1 (2010)
42. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G.: Gene ontology: Tool for the unification of biology. The gene ontology consortium. Nat. Genet. **25**(1), 25–29 (2000)
43. Bodenreider, O.: The unified medical language system (UMLS): Integrating biomedical terminology. Nucleic Acids Res. **32**(Database issue), D267–D270 (2004)
44. Ciccarelli, F.D., Doerks, T., von Mering, C., Creevey, C.J., Snel, B., Bork, P.: Toward automatic reconstruction of a highly resolved tree of life. Science **311**(5765), 1283–1287 (2006)
45. Pennisi, E.: Modernizing the tree of life. Science **300**(5626), 1692–1697 (2003)
46. Schmidt, M.T., Handschuh, L., Zyprych, J., Szabelska, A., Olejnik-Schmidt, A.K., Siatkowski, I., Figlerowicz, M.: Impact of DNA microarray data transformation on gene expression analysis—comparison of two normalization methods. Acta Biochim. Pol. **58**(4), 573–580 (2011)
47. Geller, S.C., Gregg, J.P., Hagerman, P., Rocke, D.M.: Transformation and normalization of oligonucleotide microarray data. Bioinformatics **19**(14), 1817–1823 (2003)
48. Quackenbush, J.: Microarray data normalization and transformation. Nat. Genet. **32**(Suppl), 496–501 (2002)
49. Johnson, S.C.: Hierarchical clustering schemes. Psychometrika **32**(3), 241–254 (1967)
50. Sneath, P.H.A., Sokal, R.R.: Unweighted pair group method with arithmetic mean. Numer. Taxonomy, 230–234 (1973)
51. Gascuel, O., Steel, M.: Neighbor-joining revealed. Mol. Biol. Evol. **23**(11), 1997–2000 (2006)
52. Saitou, N., Nei, M.: The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**(4), 406–425 (1987)
53. Thompson, J.D., Higgins, D.G., Gibson, T.J.: CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**(22), 4673–4680 (1994)
54. Edgar, R.C.: MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. **32**(5), 1792–1797 (2004)
55. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic local alignment search tool. J. Mol. Biol. **215**(3), 403–410 (1990)
56. Notredame, C., Higgins, D.G., Heringa, J.: T-Coffee: A novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. **302**(1), 205–217 (2000)
57. Cardona, G., Rossello, F., Valiente, G.: Extended Newick: It is time for a standard representation of phylogenetic networks. BMC Bioinf. **9**, 532 (2008)
58. Maddison, D.R., Swofford, D.L., Maddison, W.P.: NEXUS: An extensible file format for systematic information. Syst. Biol. **46**(4), 590–621 (1997)
59. Han, M.V., Zmasek, C.M.: phyloXML: XML for evolutionary biology and comparative genomics. BMC Bioinf. **10**, 356 (2009)
60. Procter, J.B., Thompson, J., Letunic, I., Creevey, C., Jossinet, F., Barton, G.J.: Visualization of multiple alignments, phylogenies and gene family evolution. Nat. Methods **7**(Suppl 3), S16–S25 (2010)
61. Schneiderman, B.: Tree visualization with tree-maps: 2-d space-filling approach. ACM Trans. Graph. **11**, 92–99 (1992)
62. Page, R.D.: Space, time, form: Viewing the tree of life. Trends Ecol. Evol. **27**(2), 113–120 (2012)

63. Zmasek, C.M., Eddy, S.R.: ATV: Display and manipulation of annotated phylogenetic trees. Bioinformatics **17**(4), 383–384 (2001)
64. Page, R.D.: TreeView: An application to display phylogenetic trees on personal computers. Comput. Appl. Biosci. **12**(4), 357–358 (1996)
65. Goloboff, P.A., Catalano, S.A., Mirande, J.M., Szumik, C.A., Arias, J.S., Källersjö, M., Farris, J.S.: Phylogenetic analysis of 73060 taxa corroborates major eukaryotic groups. Cladistics **25**(3), 211–230 (2009)
66. Heer, J., Card, S.K.: DOITrees revisited: Scalable, space-constrained visualization of hierarchical data. In: AVI '04 Proceedings of the Working Conference on Advanced Visual Interfaces (2004)
67. Parr, C.S., Lee, B., Campbell, D., Bederson, B.B.: Visualizations for taxonomic and phylogenetic trees. Bioinformatics **20**(17), 2997–3004 (2004)
68. Michael, J.M., Gord, D., Ravin, B.: Expand-ahead: A space-filling strategy for browsing trees. In: INFOVIS '04 Proceedings of the IEEE Symposium on Information Visualization 2004, pp. 119–126. (2004)
69. Bingham, J., Sudarsanam, S.: Visualizing large hierarchical clusters in hyperbolic space. Bioinformatics **16**(7), 660–661 (2000)
70. Hughes, T., Hyun, Y., Liberles, D.A.: Visualising very large phylogenetic trees in three dimensional hyperbolic space. BMC Bioinf. **5**, 48 (2004)
71. Letunic, I., Bork, P.: Interactive tree of life v2: Online annotation and display of phylogenetic trees made easy. Nucleic Acids Res. **39**(Web Server issue), W475–W478 (2011)
72. Chevenet, F., Brun, C., Banuls, A.L., Jacq, B., Christen, R.: TreeDyn: Towards dynamic graphics and annotations for analyses of trees. BMC Bioinf. **7**, 439 (2006)
73. Pethica, R., Barker, G., Kovacs, T., Gough, J.: TreeVector: Scalable, interactive, phylogenetic trees for the web. PLoS ONE **5**(1), e8934 (2010)
74. Huson, D.H., Richter, D.C., Rausch, C., Dezulian, T., Franz, M., Rupp, R.: Dendroscope: An interactive viewer for large phylogenetic trees. BMC Bioinf. **8**, 460 (2007)
75. Sanderson, M.J.: Paloverde: An OpenGL 3D phylogeny browser. Bioinformatics **22**(8), 1004–1006 (2006)
76. Kidd, D.M., Ritchie, M.G.: Phylogeographic information systems: Putting the geography into phylogeography. J. Biogeogr. **33**(11), 1851–1865 (2006)
77. Kidd, D.M.: Geophylogenies and the map of life. Syst. Biol. **59**(6), 741–752 (2010)
78. Buchin, K., Buchin, M., Byrka, J., Nöllenburg, M., Okamoto et al. : Drawing (complete) binary tanglegrams: Hardness, approximation, fixed-parameter tractability. 16th international symposium on graph drawing, Heraklion,Lecture Notes in Computer Science 5417 (2008)
79. Puigbo, P., Wolf, Y.I., Koonin, E.V.: Search for a 'Tree of Life' in the thicket of the phylogenetic forest. J. Biol. **8**(6), 59 (2009)
80. Tanner, C.C., Migdal, C.J., Jones, M.T.: The clipmap: A virtual mipmap. In: SIGGRAPH '98 Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques (1998)
81. Collins, F.S., et al.: Finishing the euchromatic sequence of the human genome. Nature **431**(7011), 931–945 (2004)
82. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W.,

McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R.A., Muzny, D.M., Scherer, S.E., Bouck, J.B., Sodergren, E.J., Worley, K.C., Rives, C.M., Gorrell, J.H., Metzker, M.L., Naylor, S.L., Kucherlapati, R.S., Nelson, D.L., Weinstock, G.M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D.R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H.M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R.W., Federspiel, N.A., Abola, A.P., Proctor, M.J., Myers, R.M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D.R., Olson, M.V., Kaul, R., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Roe, B.A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W.R., de la Bastide, M., Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J.A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D.G., Burge, C.B., Cerutti, L., Chen, H.C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S.R., Eichler, E.E., Furey, T.S., Galagan, J., Gilbert, J.G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L.S., Jones, T.A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E.V., Korf, I., Kulp, D., Lancet, D., Lowe, T.M., McLysaght, A., Mikkelsen, T., Moran, J.V., Mulder, N., Pollara, V.J., Ponting, C.P., Schuler, G., Schultz, J., Slater, G., Smit, A.F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y.I., Wolfe, K.H., Yang, S.P., Yeh, R.F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld, A., Wetterstrand, K.A., Patrinos, A., Morgan, M.J., de Jong, P., Catanese, J.J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y.J.: Initial sequencing and analysis of the human genome. Nature **409**(6822), 860–921 (2001)

83. Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., Gocayne, J.D., Amanatides, P., Ballew, R.M., Huson, D.H., Wortman, J.R., Zhang, Q., Kodira, C.D., Zheng, X.H., Chen, L., Skupski, M., Subramanian, G., Thomas, P.D., Zhang, J., Gabor Miklos, G.L., Nelson, C., Broder, S., Clark, A.G., Nadeau, J., McKusick, V.A., Zinder, N., Levine, A.J., Roberts, R.J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A.E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T.J., Higgins, M.E., Ji, R.R., Ke, Z., Ketchum, K.A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G.V., Milshina, N., Moore, H.M., Naik, A.K., Narayan, V.A., Neelam, B., Nusskern, D., Rusch, D.B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M.L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y.H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N.N., Tse, S., Vech, C., Wang, G.,

Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J.F., Guigo, R., Campbell, M.J., Sjolander, K.V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y.H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., Zhu, X.: The sequence of the human genome. Science **291**(5507), 1304–1351 (2001)

84. Bennett, S.: Solexa Ltd. Pharmacogenomics **5**(4), 433–438 (2004)
85. Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., Dewell, S.B., Du, L., Fierro, J.M., Gomes, X.V., Godwin, B.C., He, W., Helgesen, S., Ho, C.H., Irzyk, G.P., Jando, S.C., Alenquer, M.L., Jarvie, T.P., Jirage, K.B., Kim, J.B., Knight, J.R., Lanza, J.R., Leamon, J.H., Lefkowitz, S.M., Lei, M., Li, J., Lohman, K.L., Lu, H., Makhijani, V.B., McDade, K.E., McKenna, M.P., Myers, E.W., Nickerson, E., Nobile, J.R., Plant, R., Puc, B.P., Ronan, M.T., Roth, G.T., Sarkis, G.J., Simons, J.F., Simpson, J.W., Srinivasan, M., Tartaro, K.R., Tomasz, A., Vogt, K.A., Volkmer, G.A., Wang, S.H., Wang, Y., Weiner, M.P., Yu, P., Begley, R.F., Rothberg, J.M.: Genome sequencing in microfabricated high-density picolitre reactors. Nature **437**(7057), 376–380 (2005)
86. Barabási, A.-L., Gulbahce, N., Loscalzo, J.: Network medicine: A network-based approach to human disease. Nat. Rev. Genet. **12**, 56–68 (2011)
87. Xie, C., Tammi, M.T.: CNV-seq, a new method to detect copy number variation using high-throughput sequencing. BMC Bioinf. **10**, 80 (2009)
88. Keravala, A., Lee, S., Thyagarajan, B., Olivares, E.C., Gabrovsky, V.E., Woodard, L.E., Calos, M.P.: Mutational derivatives of PhiC31 integrase with increased efficiency and specificity. Mol. Ther. **17**(1), 112–120 (2009)
89. Chiang, D.Y., Getz, G., Jaffe, D.B., O'Kelly, M.J., Zhao, X., Carter, S.L., Russ, C., Nusbaum, C., Meyerson, M., Lander, E.S.: High-resolution mapping of copy-number alterations with massively parallel sequencing. Nat. Methods **6**(1), 99–103 (2009)
90. Kim, T.M., Luquette, L.J., Xi, R., Park, P.J.: rSW-seq: Algorithm for detection of copy number alterations in deep sequencing data. BMC Bioinf. **11**, 432 (2010)
91. Wang, Z., Gerstein, M., Snyder, M.: RNA-Seq: A revolutionary tool for transcriptomics. Nat. Rev. Genet. **10**(1), 57–63 (2009)
92. Morin, R., Bainbridge, M., Fejes, A., Hirst, M., Krzywinski, M., Pugh, T., McDonald, H., Varhol, R., Jones, S., Marra, M.: Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. Biotechniques **45**(1), 81–94 (2008)
93. Johnson, D.S., Mortazavi, A., Myers, R.M., Wold, B.: Genome-wide mapping of in vivo protein-DNA interactions. Science **316**(5830), 1497–1502 (2007)
94. Durbin, R.M., Abecasis, G.R., Altshuler, D.L., Auton, A., Brooks, L.D., Gibbs, R.A., Hurles, M.E., McVean, G.A.: A map of human genome variation from population-scale sequencing. Nature **467**(7319), 1061–1073 (2010)
95. Buchanan, C.C., Torstenson, E.S., Bush, W.S., Ritchie, M.D.: A comparison of cataloged variation between International HapMap consortium and 1000 genomes project data. J. Am. Med. Inform. Assoc. **19**(2), 289–294 (2012)
96. Tanaka, T.: International HapMap project. Nihon Rinsho **63**(Suppl 12), 29–34 (2005)
97. Thorisson, G.A., Smith, A.V., Krishnan, L., Stein, L.D.: The international HapMap project web site. Genome Res. **15**(11), 1592–1593 (2005)

98. Integrating ethics and science in the international HapMap project. Nat. Rev. Genet. **5**(6), 467–475 (2004)
99. The international HapMap project. Nature **426**(6968), 789–796 (2003)
100. Stalker, J., Gibbins, B., Meidl, P., Smith, J., Spooner, W., Hotz, H.R., Cox, A.V.: The Ensembl web site: Mechanics of a genome browser. Genome Res. **14**(5), 951–955 (2004)
101. Birney, E., Bateman, A., Clamp, M.E., Hubbard, T.J.: Mining the draft human genome. Nature **409**(6822), 827–828 (2001)
102. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., Haussler, D.: The human genome browser at UCSC. Genome Res. **12**(6), 996–1006 (2002)
103. Integrative Genomics Viewer: http://www.broadinstitute.org/software/igv/
104. Pavlopoulos, G.A., Secrier, M., Moschopoulos, C.N., Soldatos, T.G., Kossida, S., Aerts, J., Schneider, R., Bagos, P.G.: Using graph theory to analyze biological networks. BioData Min. **4**, 10 (2011)
105. Vikis, H.G., Guan, K.L.: Glutathione-S-transferase-fusion based assays for studying protein–protein interactions. Methods Mol. Biol. **261**, 175–186 (2004)
106. Puig, O., Caspary, F., Rigaut, G., Rutz, B., Bouveret, E., Bragado-Nilsson, E., Wilm, M., Seraphin, B.: The tandem affinity purification (TAP) method: A general procedure of protein complex purification. Methods **24**(3), 218–229 (2001)
107. Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., Sakaki, Y.: A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc. Natl. Acad. Sci. U S A **98**(8), 4569–4574 (2001)
108. Gavin, A.C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.M., Cruciat, C.M., Remor, M., Hofert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M.A., Copley, R.R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G., Superti-Furga, G.: Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature **415**(6868), 141–147 (2002)
109. Stoll, D., Templin, M.F., Bachmann, J., Joos, T.O.: Protein microarrays: Applications and future challenges. Curr. Opin. Drug Discov. Devel. **8**(2), 239–252 (2005)
110. Willats, W.G.: Phage display: Practicalities and prospects. Plant Mol. Biol. **50**(6), 837–854 (2002)
111. Bader, G.D., Hogue, C.W.: An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinf. **4**, 2 (2003)
112. Pavlopoulos, G.A., Moschopoulos, C.N., Hooper, S.D., Schneider, R., Kossida, S.: jClust: A clustering and visualization toolbox. Bioinformatics **25**(15), 1994–1996 (2009)
113. Spirin, V., Mirny, L.A.: Protein complexes and functional modules in molecular networks. Proc. Natl. Acad. Sci. U S A **100**(21), 12123–12128 (2003)
114. Li, X.L., Tan, S.H., Foo, C.S., Ng, S.K.: Interaction graph mining for protein complexes using local clique merging. Genome Inf. **16**(2), 260–269 (2005)
115. Altaf-Ul-Amin, M., Shinbo, Y., Mihara, K., Kurokawa, K., Kanaya, S.: Development and implementation of an algorithm for detection of protein complexes in large interaction networks. BMC Bioinf. **7**, 207 (2006)
116. Liu, G., Wong, L., Chua, H.N.: Complex discovery from weighted PPI networks. Bioinformatics **25**(15), 1891–1897 (2009)
117. Mete, M., Tang, F., Xu, X., Yuruk, N.: A structural approach for finding functional modules from large biological networks. BMC Bioinf. **9**(Suppl 9), S19 (2008)
118. Adamcsek, B., Palla, G., Farkas, I.J., Derenyi, I., Vicsek, T.: CFinder: Locating cliques and overlapping modules in biological networks. Bioinformatics **22**(8), 1021–1023 (2006)
119. Moschopoulos, C.N., Pavlopoulos, G.A., Schneider, R., Likothanassis, S.D., Kossida, S.: GIBA: A clustering tool for detecting protein complexes. BMC Bioinf. **10**(Suppl 6), S11 (2009)
120. Chua, H.N., Ning, K., Sung, W.K., Leong, H.W., Wong, L.: Using indirect protein–protein interactions for protein complex prediction. J. Bioinf. Comput. Biol. **6**(3), 435–466 (2008)

121. Li, X.L., Foo, C.S., Ng, S.K.: Discovering protein complexes in dense reliable neighborhoods of protein interaction networks. Comput. Syst. Bioinf. Conf. **6**, 157–168 (2007)
122. Lubovac, Z., Gamalielsson, J., Olsson, B.: Combining functional and topological properties to identify core modules in protein interaction networks. Proteins **64**(4), 948–959 (2006)
123. Cho, Y.R., Hwang, W., Ramanathan, M., Zhang, A.: Semantic integration to identify overlapping functional modules in protein interaction networks. BMC Bioinf. **8**, 265 (2007)
124. Maraziotis, I.A., Dimitrakopoulou, K., Bezerianos, A.: Growing functional modules from a seed protein via integration of protein interaction and gene expression data. BMC Bioinf. **8**, 408 (2007)
125. Feng, J., Jiang, R., Jiang, T.: A max-flow-based approach to the identification of protein complexes using protein interaction and microarray data. IEEE/ACM Trans. Comput. Biol. Bioinf. **8**(3), 621–634 (2011)
126. Ulitsky, I., Shamir, R.: Identification of functional modules using network topology and high-throughput data. BMC Syst. Biol. **1**, 8 (2007)
127. Li, X., Wu, M., Kwoh, C.K., Ng, S.K.: Computational approaches for detecting protein complexes from protein interaction networks: A survey. BMC Genomics **11**(Suppl 1), S3 (2010)
128. Pavlopoulos, G.A., Wegener, A.L., Schneider, R.: A survey of visualization tools for biological network analysis. BioData Min. **1**, 12 (2008)
129. Gehlenborg, N., O'Donoghue, S.I., Baliga, N.S., Goesmann, A., Hibbs, M.A., Kitano, H., Kohlbacher, O., Neuweger, H., Schneider, R., Tenenbaum, D., Gavin, A.C.: Visualization of omics data for systems biology. Nat. Methods **7**(Suppl 3), S56–S68 (2010)
130. Suderman, M., Hallett, M.: Tools for visually exploring biological networks. Bioinformatics **23**(20), 2651–2659 (2007)
131. Kohler, J., Baumbach, J., Taubert, J., Specht, M., Skusa, A., Ruegg, A., Rawlings, C., Verrier, P., Philippi, S.: Graph-based analysis and visualization of experimental results with ONDEX. Bioinformatics **22**(11), 1383–1390 (2006)
132. Breitkreutz. BJ, Stark. C, M, T.: Pajek—program for large network analysis. Connections **21**, 47–57 (1998)
133. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., Ideker, T.: Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Res. **13**(11), 2498–2504 (2003)
134. Pavlopoulos, G.A., Hooper, S.D., Sifrim, A., Schneider, R., Aerts, J.: Medusa: A tool for exploring and clustering biological networks. BMC Res. Notes **4**(1), 384 (2011)
135. Hu, Z., Hung, J.H., Wang, Y., Chang, Y.C., Huang, C.L., Huyck, M., DeLisi, C.: VisANT 3.5: Multi-scale network visualization, analysis and inference based on the gene ontology. Nucleic Acids Res. **37**(Web Server issue), W115–W121 (2009)
136. Letunic, I., Yamada, T., Kanehisa, M., Bork, P.: iPath: Interactive exploration of biochemical pathways and networks. Trends Biochem. Sci. **33**(3), 101–103 (2008)
137. Dogrusoz, U., Erson, E.Z., Giral, E., Demir, E., Babur, O., Cetintas, A., Colak, R.: PATIKAweb: A web interface for analyzing biological pathways through advanced querying and visualization. Bioinformatics **22**(3), 374–375 (2006)
138. van Iersel, M.P., Kelder, T., Pico, A.R., Hanspers, K., Coort, S., Conklin, B.R., Evelo, C.: Presenting and exploring biological pathways with PathVisio. BMC Bioinf. **9**, 399 (2008)
139. Shamir, R., Maron-Katz, A., Tanay, A., Linhart, C., Steinfeld, I., Sharan, R., Shiloh, Y., Elkon, R.: EXPANDER–an integrative program suite for microarray data analysis. BMC Bioinf. **6**, 232 (2005)
140. Seo, J., Gordish-Dressman, H., Hoffman, E.P.: An interactive power analysis tool for microarray hypothesis testing and generation. Bioinformatics **22**(7), 808–814 (2006)
141. Kapushesky, M., Kemmeren, P., Culhane, A.C., Durinck, S., Ihmels, J., Korner, C., Kull, M., Torrente, A., Sarkans, U., Vilo, J., Brazma, A.: Expression Profiler: Next generation–an online platform for analysis of microarray data. Nucleic Acids Res. **32**(Web Server issue), W465–W470 (2004)

142. Secrier, M., Pavlopoulos, G.A., Aerts, J., Schneider, R.: Arena3D: Visualizing time-driven phenotypic differences in biological systems. BMC Bioinf. **13**(1), 45 (2012)
143. Pavlopoulos, G.A., O'Donoghue, S.I., Satagopam, V.P., Soldatos, T.G., Pafilis, E., Schneider, R.: Arena3D: Visualization of biological networks in 3D. BMC Syst. Biol. **2**, 104 (2008)
144. Freeman, T.C., Goldovsky, L., Brosch, M., van Dongen, S., Maziere, P., Grocock, R.J., Freilich, S., Thornton, J., Enright, A.J.: Construction, visualisation, and clustering of transcription networks from microarray expression data. PLoS Comput. Biol. **3**(10), 2032–2042 (2007)
145. Morris, J.H., Apeltsin, L., Newman, A.M., Baumbach, J., Wittkop, T., Su, G., Bader, G.D., Ferrin, T.E.: clusterMaker: A multi-algorithm clustering plugin for Cytoscape. BMC Bioinf. **12**, 436 (2011)
146. Gavin, A.C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L.J., Bastuck, S., Dumpelfeld, B., Edelmann, A., Heurtier, M.A., Hoffman, V., Hoefert, C., Klein, K., Hudak, M., Michon, A.M., Schelder, M., Schirle, M., Remor, M., Rudi, T., Hooper, S., Bauer, A., Bouwmeester, T., Casari, G., Drewes, G., Neubauer, G., Rick, J.M., Kuster, B., Bork, P., Russell, R.B., Superti-Furga, G.: Proteome survey reveals modularity of the yeast cell machinery. Nature **440**(7084), 631–636 (2006)
147. Junker, B.H., Koschutzki, D., Schreiber, F.: Exploration of biological network centralities with CentiBiN. BMC Bioinf. **7**, 219 (2006)
148. O'Donoghue, S.I., Gavin, A.C., Gehlenborg, N., Goodsell, D.S., Heriche, J.K., Nielsen, C.B., North, C., Olson, A.J., Procter, J.B., Shattuck, D.W., Walter, T., Wong, B.: Visualizing biological data-now and in the future. Nat. Methods **7**(Suppl 3), S2–S4 (2010)