4

Multivariate Networks in the Life Sciences

Oliver Kohlbacher, Falk Schreiber, and Matthew O. Ward

Data in the life sciences is being obtained at a steadily increasing speed. Modern technology enables observing many of the fundamental building blocks of a cell such as genes and their activity or metabolites and their concentration, as well as many phenotypical parameters on a macroscopic level, such as shape, volume or tissue composition. The sequencing of a large number of genomes—the blueprints of life—enabled so-called post-genomics methods. The suffix '-omics' indicates the generation of data on a large, comprehensive scale. Genomics thus studies all genes and proteomics all proteins in a cell or a tissue. Recent developments have led to a staggering list of these omics technologies. Some of the more popular omics technologies and the data associated with them include:

- *Genomics:* DNA sequence and genes
- *Transcriptomics:* mRNA sequence and expression levels
- *Proteomics:* protein sequence and expression levels
- *Metabolomics:* metabolite concentrations
- *Interactomics:* protein-protein interactions

Each of these data types requires different technologies for its generation. In genomics, DNA is extracted and fragmented into a library of small segments that are each sequenced in parallel. These sequence reads are then reassembled and annotated to derive genes. In transcriptomics, sample RNA is extracted and amplified. The expression level of each mRNA can then be estimated by next-generation sequencing (RNA-Seq) or by hybridization to oligonucleotide probes (microarrays). The key technology in proteomics and metabolomics is currently mass spectrometry, where peptides (derived from proteins by enzymatic digestion) or metabolites are separated by chromatographic techniques and then detected in a high-resolution mass spectrometer. The resulting datasets of most of these technologies are huge (up to terabytes per sample) and often extremely complex.

A. Kerren et al. (Eds.): Multivariate Network Visualization, LNCS 8380, pp. 61–73, 2014.

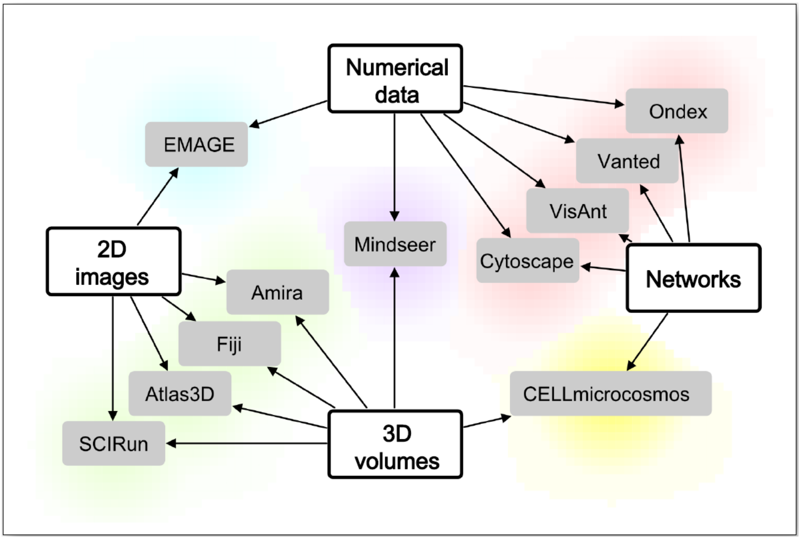


Fig. 4.1. Major data types in the life sciences and some bioinformatics tools that integrate more than one data type into the analysis process, see [17] for details

Several other types of data and information can also be integrated into the analysis process, including images, volumes, and text documents. Figure 4.1 shows a sampling of bioinformatics tools that integrate multiple forms of data.

4.1 Characteristics of Data and Tasks

Depending on the application, the data sources, and the questions under investigation, the resulting multivariate graphs can be very different. They will differ both in the semantics of the nodes and edges[1] (the type of the network) as well as in the data attached to nodes and edges.

4.1.1 Types of Biological Networks

Omics data is characterized as high-throughput and high-dimensional, as many parameters are measured at once. It is often of limited accuracy, as much noise exists in the process of extracting the data. Finally, the analysis can be quite complex, drawing from techniques found in statistics, data mining, machine learning, pattern recognition, as well as visualization.

While networks have been used for biological visualization for a long time (e. g., phylogenetic trees have been used since the early 1800s), the availability

[1] *Nodes* are also called *vertices* and *edges* are called *links*, respectively.

of high-throughput data resulted in network data on an unprecedented scale. This gave rise to the idea of 'network biology', understanding biology in terms of networks [3].

Omics data in the life sciences either represents a network (e. g., interactomics or regulomics) or can be interpreted in the context of a network (e. g., proteomics, transcriptomics, and metabolomics). Analysts may study these networks in many ways. They may focus just on a single network or part of a network, they may be interested in the interconnection between different networks, or they may want to compare multiple networks at once. In addition, they may wish to project a wide range of different data onto the networks, either on the nodes or the links, which is why the development of visualization techniques for multivariate networks is so important.

Biological networks can be organized into a hierarchy based on the entities represented by nodes and edges (see Fig. 4.2). From metabolic processes happening on an atomic scale to ecological and evolutionary networks taking place on planet-wide scales these networks cover a wide range of scales with respect to time and space. The networks differ mainly in the type of biological entities or processes represented by their nodes and edges:

- *Molecular graphs:* nodes are atoms, links are bonds.
- *Metabolic networks:* nodes are metabolites, links are reactions.
- *Interaction networks:* nodes are proteins, links are interactions.
- *Regulatory networks:* nodes are proteins, links are actions (activation, repression etc.).
- *Ecological networks:* nodes are species, links are interactions.
- *Evolutionary networks:* nodes are species, links indicate evolution.

This list is neither complete nor uniquely defined. Multiple representations are possible for many of these networks. The entities present in one network type (or layer) often have equivalents in other network types. A reaction node in a metabolic network represents an enzyme, which can interact with other proteins and is thus also represented by a node in an interaction network, or can be regulated by other genes or gene products (see Fig. 4.3).

Layouts can be either overlapping or non-overlapping. Nesting of nodes is possible to show hierarchical relationships. Additional marks and symbols can be incorporated to convey direction of relationships, locations within a cell or organism, and other types of meta-data.

4.1.2 Data Mapping and Multivariate Networks

The choice of networks underlying the data depends on the application and on the available data. In most cases, the structure of the networks is more or less fixed and the network data is taken from curated databases (such as KEGG [7], Reactome [13], BIND [2], and DIP [18]). This reflects the fact that within a given species, the structure of most networks shows little

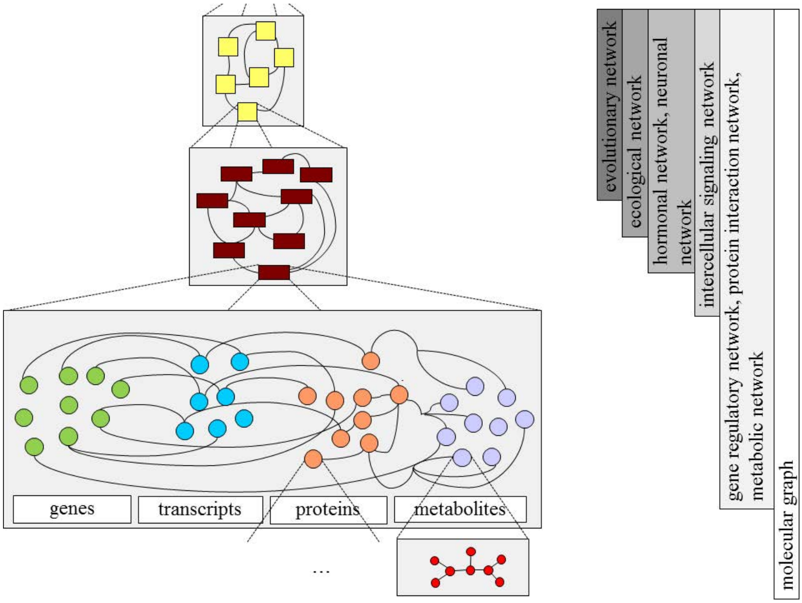

genes
transcripts
proteins
metabolites
...
evolutionary network
ecological network
hormonal network, neuronal network
intercellular signaling network
gene regulatory network, protein interaction network, metabolic network
molecular graph

Fig. 4.2. A hierarchy of biological networks

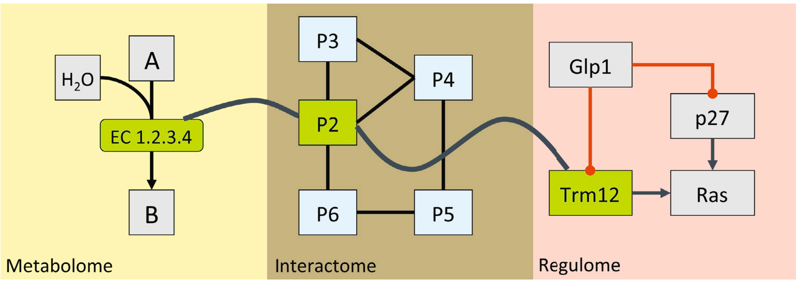
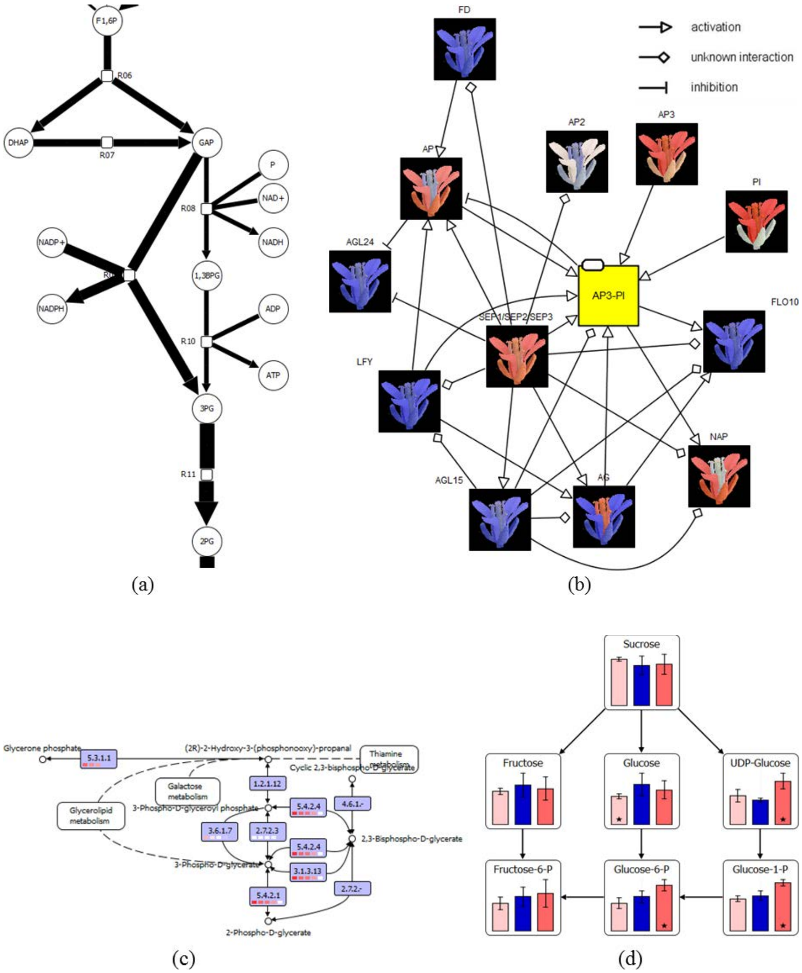


Fig. 4.3. Different levels of networks are connected through shared entities

variation[2]. What changes, though, is the state of the network, such as the concentrations of metabolites as a function of time or the expression level of genes as a function of the tissue.

The purpose of network visualization is thus, more often than not, to show the omics data in the context of these networks. Due to the size of the underlying networks, it is usually not meaningful to visualize the whole network. In most cases, only parts of the whole network are relevant and these can be identified by statistical means. For example, so-called enrichment analyses

[2] Although it should be noted that the network data itself is often incomplete, and therefore, the networks change over time due to increasing knowledge.



(a) (b)

(c) (d)

Fig. 4.4. Examples of multivariate data in biological networks, (a) flux data in metabolic networks [15], (b) spatial resolution of gene expression in a gene regulatory networks of Arabidopsis [6], (c) expression data mapped on a KEGG pathway [7], and (d) metabolite concentrations under three different conditions

can identify subnetworks that show statistically significant changes in expression levels [19]. The visualization can thus be focused onto the relevant parts of the network only, omitting the unchanged parts. In Fig. 4.4, as well as in Sect. 4.2, we give some examples of how these network and omics datasets can be represented. Typical graphical attributes used on nodes and edges to convey information are:

- *Nodes:* text labels, shape, size, color, diagrams, etc.
- *Edges:* text labels, line style, thickness, color, etc.

4.2 Use Cases

Here we discusses some use cases that show a variety of networks and ways in which multivariate biological network data has been visualized in the past.

4.2.1 Signaling

Signaling in cells can be conveyed via different mechanisms. One of the best-studied of these mechanisms is the chemical modification (phosphorylation) of certain amino acids of a protein (serine, threonine, tyrosin). This modification is reversible and is usually catalyzed by specific enzymes (kinases, phosphatases). By modifying amino acid sites in a protein very specifically, the activity of these proteins can be modulated – they can be activated or deactivated. If kinases or phosphatases themselves are activated or deactivated, they can in turn change the phosphorylation of other enzymes/proteins. In this way, a signal can be transmitted from one protein to another. This information flow follows well-defined *signaling pathways* and these pathways are part of large *signaling networks*. Signaling itself plays a key role in many biological processes and proteomics provides a time-resolved view of these signaling events. In order to unravel these networks, i. e., to figure out which protein activates which other protein at what timepoint, the visualization of these datasets in a larger context is quite helpful. In the example above, we visualized the phosphorylation patterns as a function of time (Fig. 4.5) for those nodes of the network for which (phospho-)proteomics could determine the phosphorylation patterns. Analysis of these patterns can be used to understand the dynamic behavior of signaling networks.

4.2.2 Genetic Linkage

Genetic linkage analysis is focused on the tendencies of genes that are close to each other on a chromosome to be inherited together during meiosis (cell division necessary for sexual reproduction in eukaryotes). A set of genes or gene markers undergo pairwise comparison to ascertain how frequently they undergo recombination during crossover of homologous chromosomes. This linkage score reflects the frequency of recombination between two markers or genes, which is an indication of their genetic distance (as well as physical distance). CheckMatrix (`http://cgpdb.ucdavis.edu/XLinkage`) is a visualization tool for analyzing and validating genetic maps. It uses a set of genetic markers (x and y axes in matrix) and recombination/linkage data for all possible pairs of markers computed via a variety of algorithms to create a matrix, where the color of each cell is based on the linkage score (see Fig. 4.6). Along the right border are the names for the markers and their positions in the sequence. Allele composition is shown along the bottom.

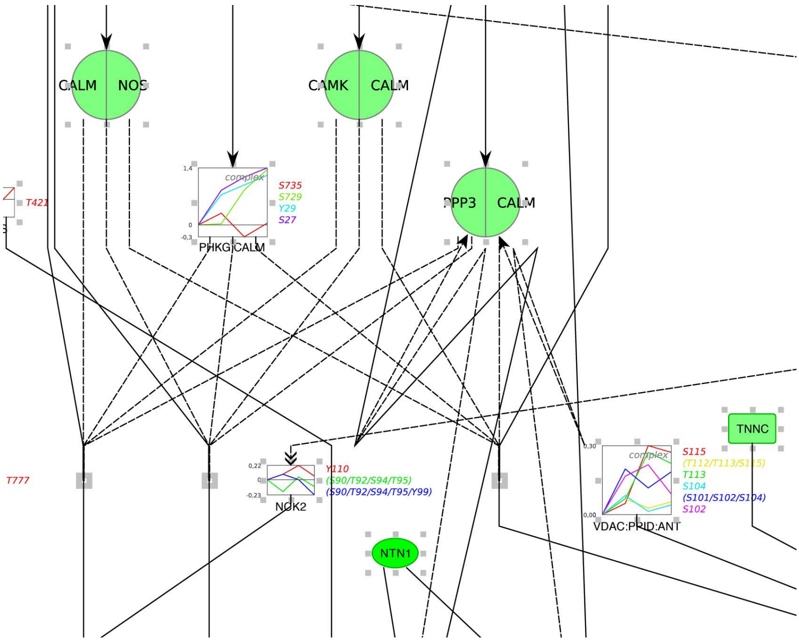


Fig. 4.5. Visualization of a signaling network showing timeseries data of the phosphorylation patterns of selected proteins. The different curves in each box represent different phosphorylation sites in the same protein.

4.2.3 Relationship Discovery Based on Document Analysis

While much biological data visualization is focused on the analysis of data sets containing sequences, numbers, and images, there is a growing interest in harvesting information from large document repositories such as PubMed (`http://www.ncbi.nlm.nih.gov/pubmed`). Chilibot (CHIp LIterature roBOT) is a tool that accepts a user's set of input keywords and gene symbols and mines PubMed abstracts for relations between the supplied terms [4]. It first augments the list with synonyms compiled from several databases (users can add to this table) and then does sophisticated natural language processing on each sentence of a collection of retrieved abstracts to find not only co-occurrences, but also types of relationships (stimulatory, inhibitory, neutral, parallel, and simple co-occurrence). The visualization represents query terms as boxes and relations as lines. Box colors are set based on degree of up/down regulation from experimental data, while line color is based on whether the relationship is stimulatory (green), inhibitory (red), or both. Grey lines are neutral. Each edge also can have a circled number indicating how many abstracts contained information about the relationship. Mousing over an edge or node provides a text annotation of the relationship or term extracted from the abstracts. Finally, arrows are added if the abstract

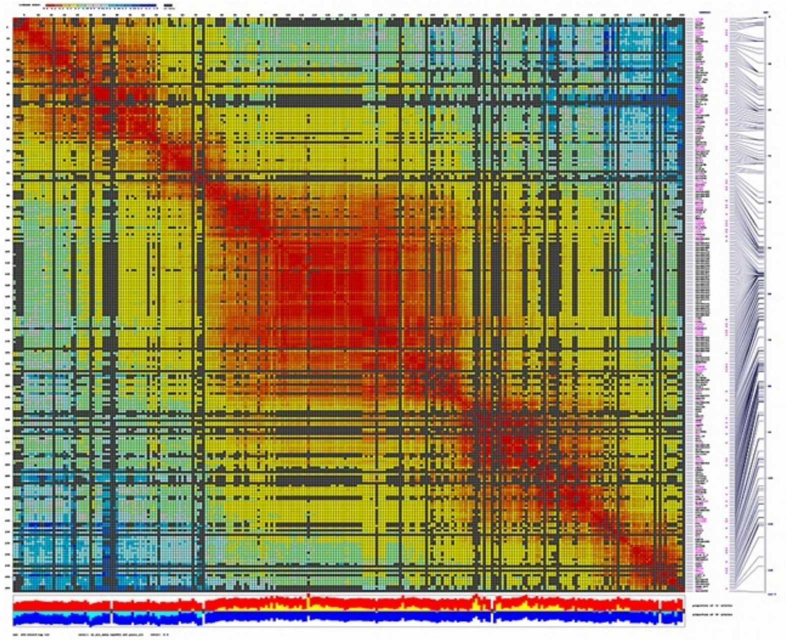


Fig. 4.6. A transcript-based genetic map generated by CheckMatrix, showing linkage information for a set of genetic markers from chromosome 4 from the plant Arabidopsis. Python_MadMapper BIT scores are mapped to color [20].

indicated directionality of the relationship. Grey diamonds indicate only a co-occurrence relationship exists. See Fig. 4.7 for an example application.

4.2.4 Gene Regulation and Transcriptome Data

Gene regulation is a complex process commonly represented by gene regulatory networks. Both the static structure of the network as well as the dynamics of regulatory events are important to understand gene regulation. The static structure of a gene regulatory network is often used to investigate functional building blocks derived from network motifs [14] or central regulatory nodes based on network centrality analysis [11]. Dynamic changes, such as organ development and morphological characteristics of higher organisms, can be traced back to gene regulatory events, which are shown by changes in the expression level of genes. The steadily increasing temporal and spatial resolution of transcriptome datasets (measuring the expression levels of genes) requires a set of analysis methods including exploration and visualization to provide insights into developmental processes.

An example is shown in Fig. 4.4(b), where we consider the visualization and exploration of tissue-specific gene expression data for master regulators of

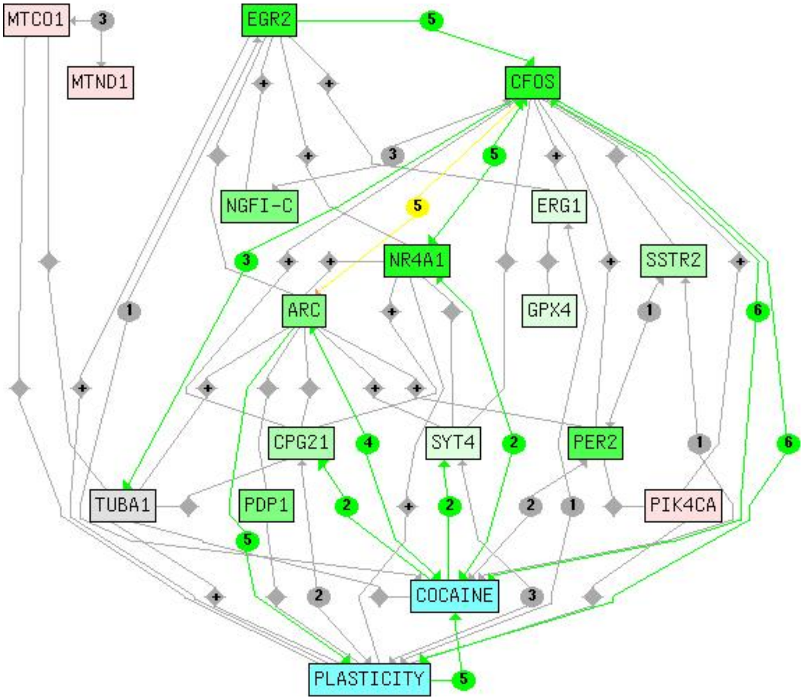


Fig. 4.7. Using Chilibot to study the relationships between genes reported to be regulated by cocaine. The network is formed automatically based on discovered relations [4].

Arabidopsis thaliana flower development in the context of the corresponding gene regulatory network [6]. In the network in Fig. 4.4(b), nodes represent genes and different types of links (represented by different arrow heads) are used to represent information about activation and inhibition. The nodes contain color-coded images which show the expression levels of the genes represented by the network node in different floral organs of the plant *Arabidopsis.*

The combination of network, omics data, and spatial information provides a fast visual exploration not only of regulatory events, but also of similar and different expression of a specific gene in the context of different tissues or organs (spatial context). Such representations can support the comparative analysis of genes with specific transcript patterns, thereby helping in extracting functional relationships.

4.3 Challenges

High-throughput data is rapidly growing in popularity in all areas of research in the life sciences. This implies that more and more non-experts get in contact with this type of data and are forced to tackle the complexity of analyzing complex multi-omics data sets. Further background information concerning the interactive visual analysis of biological networks (in particular information visualization, visual analytics, and automatic layout of networks) is given in [9]. Although there are many tools available for biological network visualization (for overviews and comparisons see, for example, [5, 10, 16]), there are still many challenges to be met [1]. The challenges arise partially from the growing amount of available high-throughput data, partially from novel applications, partially from the integration of different networks, and partially from the increasing need of more user-friendly visual analytics tools.

Currently, the key challenges concern *scale*, *uncertainty/ambiguity*, *heterogeneity*, *interactivity* and *standardization*. We will discuss each of these challenges separately in the following.

4.3.1 Scale

For some biological processes the complete networks have to be taken into consideration and thus need to be visualized. Currently networks range from a few dozens to a few thousand nodes and up to several thousand edges (for example, protein interaction or whole-genome metabolic networks), and networks with hundreds of nodes and thousands of links are in common use. This likely will expand by at least one order of magnitude in the near future. So far, tools commonly lack good methods to navigate through such large networks.

In addition, the amount and complexity of multivariate date (especially omics data, but also images, volumes, texts and so on) is steadily increasing. To make sense out of the data their integration into cellular processes and biological networks is often required. This also has implications for interactivity, exploration, and visualization. See Chap. 10 for scalability considerations for multivariate graph visualization.

4.3.2 Uncertainty/Ambiguity

Unlike in some domains, relations and values in bioinformatics are never one hundred percent certain. Concerning the structure of the networks, generally there is evidence to support a relationship, but it could be a very weak correlation that may, as more evidence is analyzed, prove to be incorrect. Also the data mapped onto the networks is often uncertain. Both the uncertainty of the network structure (and thereby the reliability of the underlying network data) as well as the uncertainty of the different related data has to be shown to a user.

Typical examples are measurement errors, missing data, multiple solutions produced by algorithms (such as in the process of finding mappings from one sequence to another, most search algorithms will report only the best match found, but in reality there may be multiple matches for the same subsequence of comparable quality), and ambiguous mappings between elements of different domains.

4.3.3 Heterogeneity

While most multivariate network visualizations incorporate a single data type, it is increasingly important to tie different data types within the analysis process. The result are heterogeneous networks with a complex structure: different types of nodes, edges, hyper edges, and hierarchical relationships (see Fig. 4.2).

Two major challenges are: (1) the compilation of heterogeneous networks which requires the identification of the biological entities and the interconnection between networks. The interconnection is especially difficult to obtain as identifiers for biological entities are often only unique in the context of one data source, for example, a database or an ontology, and identifier mapping mechanisms have to be established. (2) the visualization and interactive exploration of heterogeneous networks, which so far has not been sufficiently solved. See Chap. 9 for discussions of heterogeneous networks at multiple levels.

4.3.4 Interactivity

The scale and complexity of the data implies that discovery of new biological insights requires large-scale data exploration. Network visualization is thus more and more tied into visual analytics workflows [8]. In order to make such tools usable, interactive response times, mental map preserving animations, and easy to use interfaces are required to achieve acceptance in the user base. See Chap. 6 for discussion of interaction in the visualization of multivariate networks.

4.3.5 Standardization

Standardized glyphs for different node and link types are common in other areas of science, such as electrical engineering. Such uniform systems of nomenclature that describe the components of networks and are based on a well-defined set of symbols greatly facilitates communication efficiency and clarity. Although many visualizations in biology still do not follow uniform rules, graphical standards such as the Systems Biology Graphical Notation [12] have been established and should be obeyed to foster better understanding of network visualizations in biology.

4.4 Summary and Conclusions

In this chapter, we have described the broad range of biological data that is being routinely collected and analyzed, ranging from the atomic to the planetary scale. Data is not only available in the form of genetic sequences and numeric tables, but also in the form of images, volumes, text, and relational information. This relational information, whether explicit in the data or implicitly derived, is then the focus of multivariate network visualization. We then briefly described the typical mappings of such data to networks and presented a number of case studies showing their use in performing a variety of bioinformatics tasks. Finally, we concluded with our views on some of the key challenges facing the field of network visualization in bioinformatics.

In the future, we expect that visualization and interactive exploration will play an increasingly important role in the study of biological data and processes. This will lead to not only increased understanding of how living organisms develop, but also their relationships to other organisms. It will also be a key factor in expanding our understanding of diseases and lead to improved methods for their treatment. As mentioned in Sect. 4.3, there are many challenges that will need to be overcome in order to achieve these goals. We expect that new biological data types, as well as increased needs to integrate these types into the analytics process, will provide a wealth of opportunities for visualization researchers for many years to come.

References

1. Albrecht, M., Kerren, A., Klein, K., Kohlbacher, O., Mutzel, P., Paul, W., Schreiber, F., Wybrow, M.: On open problems in biological network visualization. In: Eppstein, D., Gansner, E.R. (eds.) GD 2009. LNCS, vol. 5849, pp. 256–267. Springer, Heidelberg (2010)
2. Bader, G.D., Betel, D., Hogue, C.W.: BIND: the biomolecular interaction network database. Nucleic Acids Research 31(1), 248–250 (2003)
3. Barabasi, A.L., Oltvai, Z.N.: Network biology: understanding the cell's functional organization. Nature Reviews Genetics 5(2), 101–113 (2004)
4. Chen, H., Sharp, B.M.: Content-rich biological network constructed by mining PubMed abstracts. BMC Bioinformatics 5(1), 147 (2004)
5. 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. Nature Methods 7, S56–S68 (2010)
6. Junker, A., Rohn, H., Schreiber, F.: Visual analysis of transcriptome data in the context of anatomical structures and biological networks. Frontiers in Plant Science 3, 252 (2012)
7. Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., Hirakawa, M.: From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Research 34, D354–D357 (2006)

8. Kerren, A., Schreiber, F.: Toward the role of interaction in visual analytics. In: Rose, O., Uhrmacher, A.M. (eds.) Proceedings of the Winter Simulation Conference (WSC 2012). pp. 420:1–420:13 (2012)
9. Kerren, A., Schreiber, F.: Network visualization for integrative bioinformatics. In: Approaches in Integrative Bioinformatics: Towards the Virtual Cell, pp. 173–202. Springer (2014)
10. Kono, N., Arakawa, K., Ogawa, R., Kido, N., Oshita, K., Ikegami, K., Tamaki, S., Tomita, M.: Pathway projector: Web-based zoomable pathway browser using KEGG atlas and Google maps API. PLoS One 4(11), e7710 (2009)
11. Koschützki, D.: Network centralities. In: Junker, B.H., Schreiber, F. (eds.) Analysis of Biological Networks. Wiley Series on Bioinformatics, Computational Techniques and Engineering, pp. 65–84. Wiley (2008)
12. Le Novère, N., Hucka, M., Mi, H., Moodie, S., Schreiber, F., Sorokin, A., Demir, E., Wegner, K., Aladjem, M., Wimalaratne, S.M., Bergman, F.T., Gauges, R., Ghazal, P., Kawaji, H., Li, L., Matsuoka, Y., Villéger, A., Boyd, S.E., Calzone, L., Courtot, M., Dogrusoz, U., Freeman, T., Funahashi, A., Ghosh, S., Jouraku, A., Kim, S., Kolpakov, F., Luna, A., Sahle, S., Schmidt, E., Watterson, S., Wu, G., Goryanin, I., Kell, D.B., Sander, C., Sauro, H., Snoep, J.L., Kohn, K., Kitano, H.: The systems biology graphical notation. Nature Biotechnology 27, 735–741 (2009)
13. Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., de Bono, B., Garapati, P., Hemish, J., Hermjakob, H., Jassal, B., Kanapin, A., Lewis, S., Mahajan, S., May, B., Schmidt, E., Vastrik, I., Wu, G., Birney, E., Stein, L., D'Eustachio, P.: Reactome knowledgebase of human biological pathways and processes. Nucleic Acids Research 37(1), D619–D622 (2009)
14. Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., Alon, U.: Network motifs: Simple building blocks of complex networks. Science 298(5594), 824–827 (2002)
15. Rohn, H., Hartmann, A., Junker, A., Junker, B.H., Schreiber, F.: FluxMap: a Vanted add-on for the visual exploration of flux distributions in biological networks. BMC Systems Biology 6, 33 (2012)
16. Rohn, H., Junker, A., Hartmann, A., Grafahrend-Belau, E., Treutler, H., Klapperstck, M., Czauderna, T., Klukas, C., Schreiber, F.: VANTED v2: a framework for systems biology applications. BMC Systems Biology 6(139) (2012)
17. Rohn, H., Klukas, C., Schreiber, F.: Creating views on integrated multidomain data. Bioinformatics 27(13), 1839–1845 (2011)
18. Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U., Eisenberg, D.: The database of interacting proteins: 2004 update. Nucleic Acids Research 32(1), 449–451 (2004)
19. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P.: Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America 102(43), 15545–15550 (2005)
20. Truco, M.J., Ashrafi, H., Kozik, A., van Leeuwen, H., Bowers, J., Wo, S.R.C., Stoffel, K., Xu, H., Hill, T., Van Deynze, A., et al.: An ultra-high-density, transcript-based, genetic map of lettuce. G3: Genes— Genomes— Genetics 3(4), 617–631 (2013)