Genes Nutr (2008) 3:139–142
DOI 10.1007/s12263-008-0098-x

MINI-REVIEW

The role of bioinformatics in pathway curation

A. S. Waagmeester · T. Kelder · C. T. A. Evelo

Received: 8 October 2008 / Accepted: 12 November 2008 / Published online: 26 November 2008

Abstract Diagrams and models of biological pathways are useful tools in biology. Pathway diagrams are mainly used for illustrative purposes for instance in textbooks and in presentations. Pathway models are used in the analysis of genomic data. Bridging the gap between diagrams and models allows not only the analysis of genomics data and interactions but also the visualisation of the results in a variety of different ways. The knowledge needed for pathway creation and curation is available from three distinct sources: databases, literature and experts. We describe the role of bioinformatics in facilitating the creation and curation of pathway.

Keywords Bioinformatics · Biological models · Community curation · Data analysis · Biological pathways

Introduction

Biological pathway diagrams are used to describe molecular biology processes in a graphical way. A pathway is a set of related reactions in a given context, i.e. glycolysis, Krebs cycle or apoptosis [4]. Traditionally, pathway diagrams are used as representations of knowledge, such as used in textbooks or in discussions among scientists.

Recently, pathway representations gained momentum as a research instrument. High-throughput genomics experiments, such as DNA microarrays, present the researcher with volumes of research data that are often too large for manual assessment. Mathematical methods like clustering or principal component analysis can be used to structure the large data volumes [2], but do not normally lead to increased understanding. Pathway diagrams can be used to present the outcome of such mathematical methods or genomics data directly. Biologically relevant changes are more visible when projected on pathway diagrams than when presented as large sets of tabular data.

This new role of the pathway representations in analysis creates specific requirements for their creation and curation. When a pathway diagram is used as an illustration, desktop publishing tools can be used to draw the diagram (Adobe Photoshop, Paintshop Pro, etc.). In this role, pathway entities and relations between entities have not to be made explicit, as long as the diagram can be interpreted by human assessment. When pathway diagrams are used as a research tool, they should be available in a computer readable form. Not only every visible aspect of a pathway diagram needs to be made explicit, but also all relationships needed for analysis. For instance, visible genes products and metabolites need to be connected to a database entry that can be used to link them to experimental data and reactions. Reactions between metabolites or gene products need to be treated as edges in an interaction network to allow network analysis.

The research field of bioinformatics has an active role in this part of pathway modelling and creation. In this paper we emphasise on pathway models, which can be used in research tools in bioinformatics. We will distinguish between aesthetical pathway diagrams and technical pathway models. We will also show why it could be advantageous to be able to combine both aesthetical pathway diagrams and computer readable pathway models.

A. S. Waagmeester (✉) · T. Kelder · C. T. A. Evelo
Department of Bioinformatics, BiGCaT, Maastricht University,
P.O. Box 616, 6200 MD Maastricht, The Netherlands
e-mail: andra.waagmeester@bigcat.unimaas.nl

Pathway diagram formats

The visual style and information of pathway diagrams varies between different resources. Figure 1 contains four examples of pathway diagrams covering knowledge about the same topic. They are extracted from KEGG [9], WikiPathways [10], Biocarta (http://www.biocarta.com) and Metacore (http://www.genego.com/metacore.php). When one wants to project research data onto a pathway, the diagrams shown in Fig. 1 will not suffice. In order to be suitable for data analysis, pathway diagrams need to be stored in a computer readable format, such as xml. All the four resources mentioned above have pathway repositories that contain both a graphical representation and at least logically structured data on the genes appearing in those pathways. WikiPathways and KEGG do this in an easily computer readable and extendable xml format. The utilisation of xml alone is not sufficient to allow computational analysis in a biological context. Scalable vector graphics (SVG), for example, is an xml format that only describes the graphical elements of an image, but not the biological meaning of these elements. A model of at least the biological entity types in a pathway is also required.

Biopax [8] is an initiative started at the ISMB'02 Conference that aims at developing an exchange standard for facilitating the integration of pathway knowledge from various sources. Biopax could be seen as the opposite pathway type of graphical pathway diagrams. In contrast to the graphically oriented pathway diagrams, BioPAX focuses on capturing the pathway information in a non-graphical, highly structured form. However, the BioPAX model lacks support for storing any graphical information. Although it is a deliberate design choice, it limits the practical usability of Biopax for visual interpretation of genomics data.

To be able to both capture logical knowledge and graphical information we started the development of GPML (Genmapp, Pathway Markup Language) in collaboration with the Conklin Group at the University of California, San Francisco [11]. GPML is an xml reimplementation of the older Genmapp data format extended with explicit interactions and the possibility to add Biopax elements like literature references.

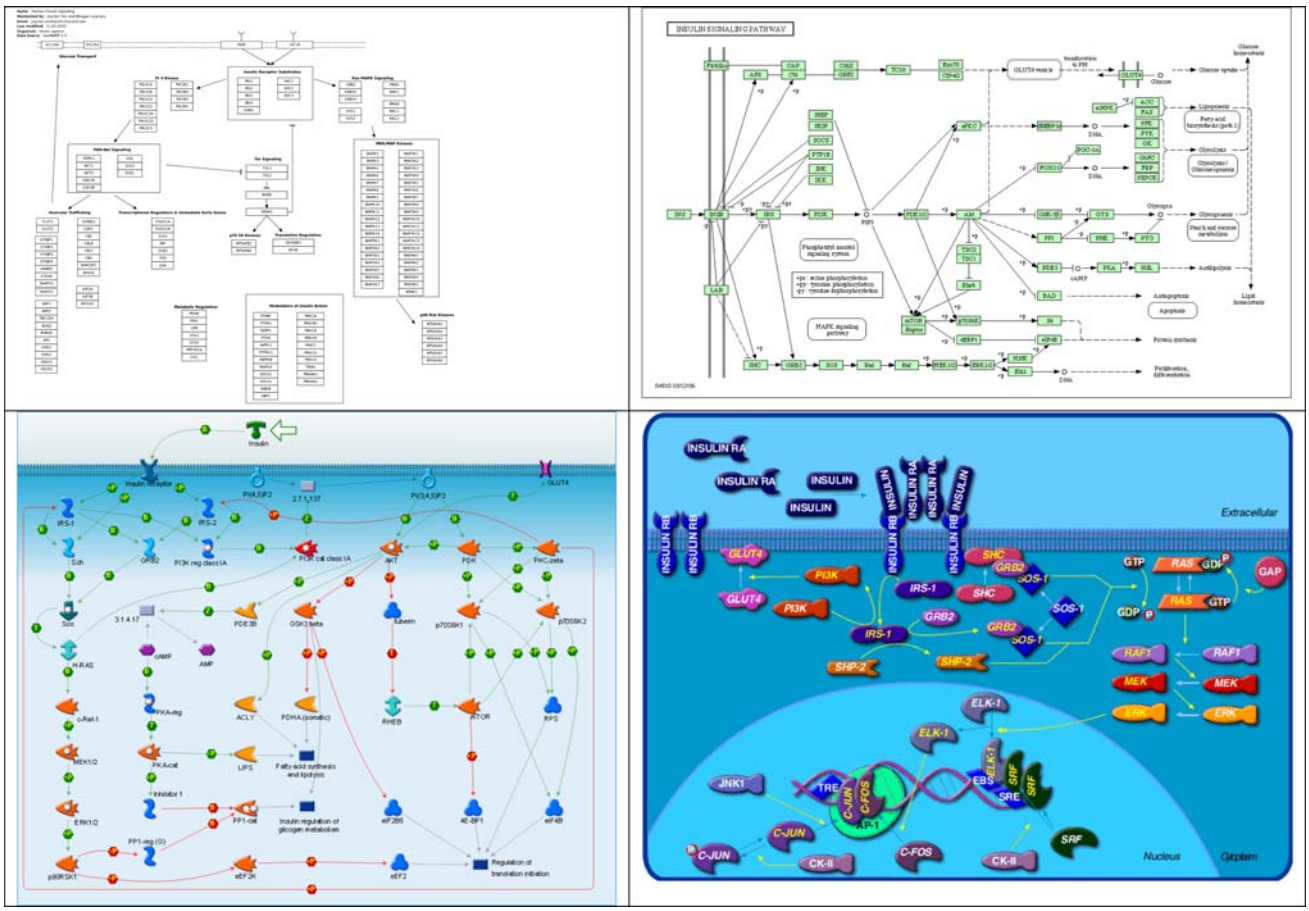

Fig. 1 Four examples of pathway diagrams from different pathway resources, all capturing knowledge from a similar topic. From *left to right*, *top to bottom*: *1* WikiPathways, *2* KEGG, *3* Metacore, and *4* Biocarta

The process of pathway curation

The application of pathway diagrams for data analysis requires the integration of knowledge from a wide variety of sources. Typically, a pathway diagram is created by integrating knowledge from biological databases, the scientific literature and intrinsic knowledge from domain experts [1]. Each of these three sources presents with specific challenges to extract and integrate the knowledge into a pathway diagram. Part of this challenge is the exponentially growing amount of available information that is scattered over a wide variety of sources and knowledge domains. Pathguide (http://www.pathguide.org), a website that lists pathway resources, already lists 261 different databases (October 2008) containing pathway knowledge [3]. As Cary [4] points out, there is the need to be aware of potential biases when integrating data from biological databases. Biological databases are often constructed around a specific species or with a specific set of research questions in mind.

Presenting knowledge as easily readable text often means that computers have a problem interpreting it. An author would for instance try to avoid the utilisation of the same word more than once in the same paragraph, even when it denotes the same entity. This means that the actual lines containing crucial information have to be read and interpreted by humans before it can be available in a computer readable form. As it is no longer possible to keep up-to-date with the exponentially growing amounts of literature, we need "text mining approaches" to find the important parts [7]. This provides specific challenges to the (bio) informatics community. New approaches need to be developed to automatically deal with text primarily intended for human interpretation. The requirement for text mining is not exclusive to the biology community; other disciplines in science are also facing an increase in scientific literature. Although each discipline could benefit from general text mining solutions, the specific characteristics of the way literature is stored requires specific solutions for each discipline. One benefit biomedical sciences have is the existence of Pubmed (www.ncbi.nlm.nih.gov/pubmed/), which provides us with a standard in the representation of scientific literature. Another specific characteristic for the current biology field is the vast amount of experimental data that is created in genomics. This directs the questions we want to see answered from text mining efforts. For instance, we do want to find relationships between genes influencing each other's regulation.

Integrating knowledge from different domain experts is probably the most challenging task, especially if more than one domain expert is involved. This process requires commitment from the domain experts and an extensive social network and social skills by the pathway diagram curator. In the next chapter, we will address this point more in detail.

The role of bioinformatics in pathway curation

These different challenges provide the opportunity for bioinformatics to have an active role in pathway curation.

An illustration of how bioinformatics can assist pathway curation is WikiPathways [10] (http://www.wikipathways.org). WikiPathways builds upon the concept of community-based curation of biological knowledge using a wiki [5, 10, 12]. A wiki is a website the content of which can be edited by users. A well-known example of a wiki is Wikipedia, an online encyclopaedia where all content is created and maintained by its users. No formal approval of human editor is needed, still recent studies showed that the quality of Wikipedia is similar to editor-based encyclopaedias [6].

WikiPathways applies a similar approach to biological pathway information. Pathway diagrams on WikiPathways can be created and curated by every member of the scientific community. As mentioned earlier, integrating expert knowledge requires an extensive social network and commitment from the community. Still it requires time-intensive procedure to collect knowledge from different domain experts involved. Using WikiPathways, domain experts from all over the world can directly collaborate on improving specific pathway diagrams. Research groups focusing on a specific research area can adopt a set of pathways and identify themselves as a community by creating a portal page. This makes it easier to capture and integrate information from different domains and increases the commitment of the community in pathway curation. Specific research communities can adopt a set of pathways, which will then be arranged into a portal.

Where aesthetic pathway diagrams meet pathway knowledge models

We can define a spectrum of pathway knowledge types with complete graphical pathway diagrams on one side, and Biopax pathway models on the other tail. In between these extremes, we find a spectrum of resources capturing pathway knowledge in both graphical form and in a computer-readable format. Additionally, computational methods based on pathway diagrams can extract relevant parts out of large experimental dataset and improve statistical power.

As mentioned before, the different applications pose different requirements on the pathway representations. As an example, a pathway diagram needs to be primarily

aesthetic and in a human readable format. In this format, we can use colours and other graphical features to distinguish various entity type such cell-structures, genes, proteins, etc. Indeed pathways need to be "beautiful". By applying colours or other aesthetical features, researchers understand knowledge better. However, pathways diagrams, which are used in data analysis, need to be simple; they should allow an overlay of the experimental data that we want to show. Very colourful and detailed pathway representations like we see them for instance in Metacore, actually make it harder to understand data representations.

One would expect that typically bioinformatics have a role in the curation of pathway models for use in data analysis. For complete graphical pathways diagram any standard imaging program would be sufficient. We would advocate for bridging the two application areas by providing hybrid pathway representations. Having such hybrid pathway diagrams would eliminate redundancy. Such pathway diagrams would contain enough structured, computer-readable information to be used in data analysis with bioinformatics tools, as well as a clear and flexible graphical representation to aid human interpretation.

Conclusion

As bioinformaticians, we have taken an active role in facilitating community-based pathway curation. Data analysis on results from genomic studies requires accurate and complete pathway models and the corresponding diagrams need to be interpretable by scientists. Initiatives such as WikiPathways aim to collect and present pathway information that meets both requirements using a community-based curation approach and a graphically oriented pathway format, while at the same time facilitate the integration of knowledge from other sources such as databases and scientific literature.

Acknowledgments This work is part of the BioRange programme of the Netherlands Bioinformatics Centre (NBIC), which is supported by a BSIK grant through the Netherlands Genomics Initiative (NGI). We are grateful for the valuable discussions surrounding this topic with Prof. Dr. H. J. van den Herik.

References

1. Adriaens ME, Jaillard M, Waagmeester A, Coort SL, Pico AR, Evelo CT (2008) The public road to high-quality curated biological pathways. Drug Discov Today 13:856–862
2. Allison DB, Cui X, Page GP, Sabripour M (2006) Microarray data analysis: from disarray to consolidation and consensus. Nat Rev Genet 7:55–65
3. Bader GD, Cary MP, Sander C (2006) Pathguide: a pathway resource list. Nucleic Acids Res 34:D504–D506
4. Cary M, Bader G, Sander C (2005) Pathway information for systems biology. FEBS Lett 579:1815–1820
5. Doerr A (2008) We the curators. Nat Methods 5:754–755
6. Giles J (2005) Internet encyclopaedias go head to head. Nature 438:900–901
7. Jensen LJ, Saric J, Bork P (2006) Literature mining for the biologist: from information retrieval to biological discovery. Nat Rev Genet 7:119–129
8. Luciano JS (2005) PAX of mind for pathway researchers. Drug Discov Today 10:937–942
9. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 27:29–34
10. Pico AR, Kelder T, van Iersel MP, Hanspers K, Conklin BR, Evelo C (2008) WikiPathways: pathway editing for the people. PLoS Biol 6:e184
11. van Iersel MP, Kelder T, Pico AR, Hanspers K, Coort S, Conklin BR, Evelo C (2008) Presenting and exploring biological pathways with PathVisio. BMC Bioinformatics 9:399
12. Waldrop M (2008) Big data: wikiomics. Nature 455:22–25