Organization and integration of biomedical knowledge with concept maps for key peroxisomal pathways

A.M. Willemsen\textsuperscript{1}, G.A. Jansen\textsuperscript{1}, J.C. Komen\textsuperscript{1}, S. van Hooff\textsuperscript{1}, H.R. Waterham\textsuperscript{2}, P.M.T. Brites\textsuperscript{2}, R.J.A. Wanders\textsuperscript{2} and A.H.C. van Kampen\textsuperscript{1,3,4}  

\textsuperscript{1}Bioinformatics Laboratory, Clinical Epidemiology, Biostatistics and Bioinformatics, \textsuperscript{2}Laboratory Genetic Metabolic Disease, Department of Clinical Chemistry and Pediatrics, Academic Medical Center, University of Amsterdam, Meibergdreef 9 1105 AZ Amsterdam, The Netherlands and \textsuperscript{3}Biosystems Data Analysis, Swammerdam Institute for Life Science, University of Amsterdam, Nieuwe Achtergracht 166, 1018 WV Amsterdam, The Netherlands

ABSTRACT

Motivation: One important area of clinical genomics research involves the elucidation of molecular mechanisms underlying (complex) disorders which eventually may lead to new diagnostic or drug targets. To further advance this area of clinical genomics one of the main challenges is the acquisition and integration of data, information and expert knowledge for specific biomedical domains and diseases. Currently the required information is not very well organized but scattered over biological and biomedical databases, basic text books, scientific literature and experts’ minds and may be highly specific, heterogeneous, complex and voluminous.

Results: We present a new framework to construct knowledge bases with concept maps for presentation of information and the web ontology language OWL for the representation of information. We demonstrate this framework through the construction of a peroxisomal knowledge base, which focuses on four key peroxisomal pathways and several related genetic disorders. All 155 concept maps in our knowledge base are linked to at least one other concept map, which allows the visualization of one big network of related pieces of information.

Availability: The peroxisome knowledge base is available from www.bioinformaticalaboratory.nl (Support \textrightarrow Web applications).

Supplementary information: Supplementary data is available from www.bioinformaticalaboratory.nl (Research \textrightarrow Output \textrightarrow Publications \textrightarrow KB_SuppInfo).

1 INTRODUCTION

Biomedical scientists typically focus on a specific and small part of a larger scientific challenge such as the full characterization of the peroxisome. However, the need for scientists that think and work in a multidisciplinary and integrative manner is increasingly required to keep pace with scientific developments as in fields such as clinical genomics and systems biology. One important area of clinical genomics research involves the elucidation of molecular mechanisms underlying (complex) disorders to enhance our basic understanding of these disorders, which eventually may lead to new diagnostic or drug targets. To further advance this area of clinical genomics, one challenge is the acquisition and integration of data, information and expert knowledge for specific biomedical domains or diseases. The distinction between data, information and knowledge may be subtle and sometimes artificial but we intuitively use it to denote different levels of understanding of the underlying domain, which generally implies an increased level of complexity and connectedness between the underlying entities. In the remainder of this article we will use these terms indistinguishably.

Information available for a specific biological system (pathway or organelle), a specific human disease and their interrelation are not always easily accessible or applicable for researchers who are not familiar with the biological system or disorder under consideration. Even domain experts may experience difficulties when retrieving and relating all pieces of information. Progress in the organization and integration of biomedical information and the development of methodologies to use this information will therefore dictate the pace of implementation of clinical genomics. Information about specific biological systems includes (public) data about genes, proteins, biological pathways and disorders to information on cell organelles, cells, tissues, organ (systems) and organisms. However, the acquisition, organization and integration of information are challenging and time consuming tasks but increasingly important given the ever growing amount of (literature) information that becomes available. In particular, the acquisition of knowledge, which is not yet explicitly defined, from experts is generally a difficult task (Payne \textit{et al}, 2007).

Typically biomedical information is not very well organized in a unified and standardized format but scattered over many biological and biomedical databases, student text books, scientific literature and experts minds and may be highly specific, heterogeneous, complex and voluminous. Given these difficulties, a lively research community emerged aiming to develop novel approaches for knowledge representation (Barriot \textit{et al}, 2004; Cañas \textit{et al}, 2005; Hasegawa \textit{et al}, 2006; Hayes \textit{et al}, 2005; Lee \textit{et al}, 2004; Luciano and Stevens, 2007; Smith \textit{et al}, 2007, http://www.biopax.org/), the development of vocabularies and ontologies (Cote \textit{et al}, 2006; Harris \textit{et al}, 2004; Rosse and Mejino, 2003; Whetzel \textit{et al}, 2006) and the development of knowledge bases for the biomedical (Barbulescu \textit{et al}, 2007; Hervold \textit{et al}, 2007; Samarghitean \textit{et al}, 2007; Schlüter \textit{et al}, 2007; Seebah \textit{et al}, 2007; Shoot \textit{et al}, 2005; Vastrik \textit{et al}, 2007) and clinical (Al-Busaidi \textit{et al}, 2006) domains. In general, these knowledge bases provide access to information that is manually curated by domain experts but also provide links to primary or secondary databases.

In this article, we present a new approach towards the construction of a peroxisome knowledge base. We use a network of OWL (web ontology language)-based (Hayes \textit{et al}, 2005; Lacy, 2005,
These patients suffer from a range of neurological, ocular and bone related disorders. We constructed a generic OWL knowledge model to ensure connectivity and consistency of the concept maps. The OWL representation of the concept maps and knowledge model not only enables powerful querying and (graph-based) navigation, but also provides a first step towards future applications such as the visualization of experimental and clinical data, and the integration with computational models.

We constructed a peroxisome knowledge base with a focus on key metabolic pathways and related disorders to demonstrate the knowledge base framework and as basis for further improvement and extension of this framework. Knowledge on peroxisomes has expanded enormously in recent years, which makes it hard for people, not directly involved in peroxisome research, to keep track. This is especially relevant for clinicians, confronted with patient, suspected to suffer from a peroxosomal disease. Peroxisomes belong to the microbody family, a class of ubiquitous and essential cell organelles, which also includes glyoxysomes and glycosomes (Wanders and Waterham, 2006a). It is now clear that fatty acid β-oxidation is a general feature of peroxisomes. In addition, peroxisomes in higher eukaryotes, including humans, have a number of functions not shared by peroxisomes in lower eukaryotes, including ether phospholipid biosynthesis, fatty acid ω-oxidation and glyoxylate detoxification.

The list of functions attributed to peroxisomes continues to grow, emphasizing the role of the peroxisome as multipurpose organelle. Peroxosomal disorders are a group of inherited diseases in human and are usually classified in two groups including: (1) Peroxisomal Biogenesis Disorders (PBDs) (Steinberg et al., 2006; Wanders and Waterham, 2005) and (2) the single Peroxisomal Enzyme Deficiencies (PEDs) (Wanders and Waterham, 2006b) in which a protein involved in ether phospholipid (plasmalogen) biosynthesis, peroxisomal fatty acid β-oxidation, fatty acid ω-oxidation, glyoxylate detoxification or H2O2 metabolism is altered. Peroxisomes play an indispensable role in human physiology as can be concluded from the devastating consequences of a deficiency of peroxisomes as observed in Zellweger patients. These patients suffer from a range of neurological, ocular and bone defects and most infants do not survive past the first 6 months as a result of respiratory distress, gastrointestinal bleeding or liver failure. The aim of our project is to develop a general framework to construct biomedical knowledge bases. In this article we present a first implementation of this framework: the peroxisome knowledge base. This knowledge base includes information on the four key peroxosomal pathways (other phospholipid biosynthesis, fatty acid β-oxidation, peroxosomal ω-oxidation, glyoxylate detoxification). Currently, the peroxisome knowledge base contains 155 concept maps represented different aspects of these metabolic pathways and related disorders.

2 SYSTEMS AND METHODS
2.1 Visual knowledge presentation

The core of our knowledge base consists of so-called concept maps (Cañas et al., 2005; Hayes et al., 2005; Novak and Giowin, 1984), which provide a graphical representation of knowledge that can easily be understood by others. Concept maps are graphical 2D displays of knowledge that are comprised of concepts connected by directed arcs encoding relationships between the concept pairs. Figure 1 shows one specific concept map that provides information on the four key peroxosomal pathways and which is part of our knowledge base. Concept maps can be linked to establish relationships between pieces of knowledge, which enables the navigation of one concept map to another. So far, the links between the concept maps were determined by the domain experts. However, it is realized that in the near future we need to set up clear rules for this as part of the curation process. Additionally, individual concepts can be linked to other types of resources (pdf documents and public biological databases) that help to explain and complement the information represented by the concerning concepts. A main advantage of using concept maps in the knowledge acquisition process is the graphical presentation of knowledge, which allows the domain experts to quickly evaluate the represented parts of the knowledge. This method was also successfully applied in a number of other projects (Castro et al., 2006; Hayes et al., 2005).

The construction of the concept maps (i.e. knowledge acquisition) in the peroxisome knowledge base was done in tight collaboration with domain experts, who provided content and specified the layout of every concept map. More importantly, every concept map was curated by a domain expert, which ensures the quality and validity of such piece of knowledge. Our set of concept maps present different types of information with different levels of detail and for different groups of users (e.g. physicians, researchers or students). We have included several concept maps with information from student biochemistry textbooks which provide elementary information on, for example, the chemical structure of fatty acids (Supplementary Fig. 1) or the difference between β-oxidation of fatty acids in peroxisome and mitochondrion (Supplementary Fig. 2). Other concept maps provide more details on specific pathways (Supplementary Fig. 3), or were constructed to show the underlying molecular defects of certain disorders (Supplementary Fig. 4) and related symptoms (Supplementary Fig. 5). Currently our knowledge base contains 155 concept maps.

Metadata tags were defined to annotate each concept map, allowing the user to establish the overall content and status of each map. These metadata tags provide basic information on type and level of the presented knowledge (Supplementary Fig. 6) and information on curation by domain experts and from which the information was derived (i.e. PubMed reference). In addition, the metadata hold the history of each concept map such that changes in a map can easily be tracked.

As discussed by Cañas et al. (2005), it is important for concept maps to stay coherent and extensible. As concept maps are being enriched it is important to ensure the coherence of the story that is being captured (i.e. knowledge on the peroxisomes). Furthermore, there is a risk that concept maps will constantly gain information resulting in complex maps that are difficult to interpret. Extending knowledge on peroxosomal pathways involves not only addition of more details to the existing concept maps or generating new maps, but also grouping of concepts into higher level abstractions and validation of the latter by domain experts. An example of such higher level abstraction and hierarchy is shown in Figure 7 (Supplementary Material) in which one concept map explains the differences between ω- and β-oxidation of fatty acids while the two linked (child-) concept maps provide more detailed information on these two pathways in the peroxisome. The final hierarchy of concept maps allows the user to easily gain insight in the complex peroxosomal domain while still keeping a good overview.

2.2 Knowledge representation by OWL

Although our concept maps provide a visual presentation of pieces of information, the underlying representation format is provided by the web ontology language OWL, which implies that effectively the knowledge base is organized as an ontology. OWL is designed for use by applications that need to process the information instead of just presenting information to humans. For an overview on functions and applications of biomedical ontologies see the review of Rubin et al. (2008). We use the OWL format since this allows us to store, navigate, query and integrate information. One of
Organization and integration of biomedical knowledge

Fig. 1. Example of a concept map showing the four key peroxisomal pathways in our knowledge base. Icons associated with specific concepts provide links to other concept maps or information resources (pdf documents).

Fig. 2. Example of the relation between generic knowledge model and concept maps. The proposition (or property) ‘encodes’ models the relation between the entities (or classes) ‘Gene’ and ‘Protein’. The concepts ‘PHYH’ and ‘phytanoyl-CoA hydroxylase’ could be seen as instances of those classes.

The advantages of OWL is the ability to import and thus extend existing ontologies. Through this mechanism we import our generic knowledge model (class model) into all other OWL files. We might use the same mechanism for the future extension of existing ontologies. In a later stage the OWL constructs may become useful when we integrate experimental data and explore the usability of automatic reasoning for the analysis and interpretation of these data.

2.3 The generic knowledge model

The construction of concept maps is not subjected to restrictions with respect to concepts and relations between them. Therefore there is a clear risk to produce inconsistent concept maps (e.g. describing similar relations between the concepts with different terminology), or maps that are difficult or even impossible to relate and use without domain knowledge, because the semantics of the concepts are not defined (e.g. it is impossible to deduce that the concept ‘PEX7’ represents a gene). To ensure consistency and connectivity of the concept maps we constructed a generic OWL-based knowledge model (Fig. 2) that provides a foundation on which basis the concept maps are designed. The model represents a hierarchy of basic entities (classes) and (inheriting) relationships (properties) between these entities. All concepts that appear in the concept maps are instances of these classes.

The development of our model is driven by the content added to the knowledge base. The advantage of this approach is that the model is implicitly defined by the domain experts and is iteratively refined. It will serve as a ‘reference ontology’ (Burgun, 2006) to link and integrate existing ontologies and vocabularies that are available for specific domains. A major disadvantage of this development method is the risk of creating a model that is suboptimal, or that may not be able to represent new pieces of information without changing the fundamentals of the model and, consequently, requires redesigning all concept maps.

2.4 Navigation and querying

All concept maps in our knowledge base are linked to at least one other concept map, which allows the visualization of one big network of related pieces of information. We used the Prefuse toolkit (http://prefuse.org) to visualize radial graphs (Fig. 3), which proofs an powerful way to provide overview of the concepts maps in the knowledge base and, thereby, of the peroxisome domain. In this graph the ‘active’ concept map is displayed together with direct neighbors and more distant maps. The graph enables the user to navigate from one map to another along a logical route in which known information can be ignored while other information can be inspected in more detail. Each node (concept map) in the graph can be activated after which the graph is automatically redrawn such that the activated node becomes central in the graph. To further increase the utility of the graph as navigation mechanism, the appearance of the nodes and the size of the graph can be adjusted. The nodes can be displayed as little circles with (partial or full map titles) or without annotation and the radius of the graph can be changed to display only those concept maps that are likely to focus on the piece of information under investigation. Finally, the view can be zoomed in and out and moved in a google maps like manner. In addition to graph navigation, the graph is also used during querying and filtering of knowledge base information. Each term (concept label) occurring in a concept map can be used to search the
A network of concept maps provides the user the necessary overview of the pieces of information and their relation in the knowledge base. Each node (concept map) in the graph can be activated after which the graph is automatically redrawn. The activated concept map is centralized and its direct neighbors are indicated. Together with search and filter functionality this enables the user to navigate through the concept maps in a relevant order.

When a term or the beginning characters are found in a map the corresponding node is highlighted. The same principle is used to filter the graph on metadata tags. The metadata tag type is used to indicate so-called entry maps, which are maps that provide good starting points for further navigation. Examples of such maps are the key peroxisomal pathways (Fig. 1) or the concept map that provides an overview of all known peroxisomal disorders. The metadata tag level is used to distinguish between elementary, intermediate and advanced maps, indicating the level of the information displayed in the maps. So will elementary maps contain student textbook like background information such as the explanation of the structure and function of phospholipids, whereas advanced maps contain detailed information extracted from scientific literature or obtained from domain experts.

As aforementioned, all individual concepts present in the concept maps are classified to one of the classes in the generic knowledge model (Fig. 4). The relations between the concepts correspond to the property relations of the knowledge model. Similarly to the concept map graph, the generic knowledge model can be presented and navigated as a star graph with a central class and its relation to all neighboring classes. The red arrows in this visualization indicate sub-superclass relations of the class hierarchy, like Enzyme are Protein. The arrowhead in this type of relation is pointing to the superclass. Properties of the superclass are inherited by the subclass and are indicated by light-grey arrows and pink class boxes and can be filtered out from within the application. Dark-gray arrows with purple class boxes indicate properties introduced for the central class. All instances (e.g. all gene names and symbols) of a class (e.g. gene) can be retrieved through
Organization and integration of biomedical knowledge

Fig. 4. Presentation of the knowledge model, which can be navigated by clicking a peripheral class after which the star will automatically be redrawn and the selected class becomes the central node. Red arrows (left) indicate the class hierarchy and are drawn from subclass to superclass (‘are’ relations). Dark-gray arrows, pointing to purple boxes, indicate property relations between classes. The properties inherited from the superclasses are indicated by light-gray arrows, pointing to pink boxes. Part of the model (right) is filled with knowledge derived from concept maps. The red arrow, drawn from the central node, points to the represented class.

3 IMPLEMENTATION

We implemented our knowledge base as web application. Struts (http://struts.apache.org/) and Java Server Pages (http://java.sun.com/products/jsp/) were used as web framework to increase maintainability and modularity of the software code. For the construction of concept maps we used CmapTools Ontology Editor (COE) (Hayes et al., 2005). The concept maps were exported to HTML, OWL and scalable vector graphics (SVG). The HTML pages were converted to Java Server Pages to fit our application architecture. The graphics were used as preview, showing an overview of the whole concept map and the OWL files were used for querying and navigation. We developed an applet that reads OWL and displays navigation graphs. To validate the exported OWL files, we loaded them into the OWL editor Protégé (Noy et al., 2003), which was also used to design the generic knowledge model. Jena (http://jena.sourceforge.net/) was used to parse OWL. For graph visualization the graph drawing libraries Jgraph (http://www.jgraph.com/) and Prefuse (radial graph view algorithm, http://prefuse.org/) were used.

4 DISCUSSION

In this article we presented a new framework for visualizing structured knowledge in a way that is easily understandable and easy to navigate. The main novelty of our approach is that we represent all information in our knowledge base as instances of an OWL reference ontology (the generic knowledge model). This approach results in a semantic definition and integration of information. This enables an easier integration and use of this data in future applications. A second novelty is the graphical presentation of information to the user. We also present a first implementation of this framework: the peroxisome knowledge base. Currently it contains information on the four key peroxisomal pathways and several related genetic disorders. The concept maps are a systematic way to organize and present information on (1) metabolism and function of peroxisomes for people wanting to know about these aspects in a concise and lucid way and (2) in addition describe the molecular mechanisms underlying peroxisomal disorders. This approach intends to give biomedical researchers and physicians further insight in peroxisomal disorders, to help them select the right laboratory tests, and to gain easy access to the extensive knowledge on peroxisomes, which is relevant in a clinical setting.

The acquisition of information for the knowledge base (i.e. the construction of the concept maps) is time consuming and making knowledge explicit is a difficult task. In addition, because the amount
of information and knowledge on peroxisomes is enormous, it is important to clearly define the scope of the knowledge base. These choices were dictated by domain experts (from the Laboratory of Genetic Metabolic Disease, AMC) and the applications of this knowledge base. Knowledge about peroxisomes is in continual flux, the peroxisome knowledge domain is relatively large, complex, and therefore cannot be modeled in one single effort. Hence, we continue to further develop and complete the current knowledge base. As part of this we will integrate our concept maps with (non-curated) information from public biological databases.

In our experience, concept maps provide a very convenient approach for knowledge capturing and curation, especially because the knowledge is visually presented. However, the further completion of the peroxisome knowledge base will likely provide additional challenges. First of all, the current content of individual concept maps and the division of knowledge over concept maps was determined by consulting domain experts from a single research group and with a specific application in mind. However, other domain experts may disagree on pieces of knowledge or on the design of certain concept maps. Therefore, we will develop mechanisms to deal with overlapping or conflicting, yet curated, information in the knowledge base. Second, it is important to avoid constructions in which a semantic unit is described using more than one proposition (concept → proposition → concept) for example, the constructions: 'RCDP type 1 = is_caused_by mutations in → PEX7' and 'Refsum disease = is_caused_by mutations in → PHYH'. When merging these two ontology statements the result ('RCDP type 1 + Refsum disease = is_caused_by mutations in → PEX7 + PHYH') loses the information on the specific mutation that caused the disease. A better construction would be 'RCDP type 1 = is_caused_by_mutations_in → PEX7'. However this construction makes the proposition very specific. Furthermore, it is obvious that domain experts cannot foresee this kind of problems and should be warned when using such a construction. Third, the knowledge base will finally serve different user groups such as basic scientists, physicians, patients (and maybe even patients’ organizations), teachers and students. These user groups may require additional or different sets of concept maps to represent knowledge at different levels of detail or in different forms. For all these reasons, the captured knowledge will constantly evolve and concept maps may change over time. This will also require the definition of additional metadata that allows selecting information at different levels of detail or for different user groups. Finally, it is of vital importance but also an enormous challenge to keep the knowledge base up-to-date, or more general, to keep collecting new and more information over time. This will not only require long-term commitment of domain experts but also practical maintenance procedures, which we are currently working on. The curation and up-dating of the knowledge base is clearly not applicable to broad biological domains (e.g. complete metabolism) but can be applied to focused domains such as specific organelles, pathways or disorders. This will result in very specialized but valuable expert databases.

Currently, the genetic knowledge model and the concept maps do not make use of existing ontologies or standardized vocabularies as, for example, provided by Gene Ontology (Harris et al., 2004), BioPax (http://www.biopax.org), OBO (Smith et al., 2007), SBML (Hucka et al., 2003), UMLS (Bodenreider, 2004) or BRENDA (Barthelmes et al., 2007). So far, we did not use these ontologies since we required a single evolving ontology (knowledge model) that represents a broad part of the biomedical domain, necessary to serve as a foundation for the concept maps. Other ontologies focus on specific subdomains and, therefore, do not meet this requirement. Moreover, it is not possible to integrate existing ontologies in a model that will serve our needs nor do we require all details provided by these individual ontologies. Several ontologies have been designed for completely other purposes like data exchange (BioPax). Whereas other ontologies like Gene Ontology is too large for our application and therefore too much time is needed for loading when this ontology is imported in our ontology. However, as part of ongoing work we will integrate with existing ontologies to stay compatible with current efforts in the field of biomedical ontologies.

As part of ongoing work we will extend current functionalities of the knowledge base framework and the peroxisome knowledge base in particular. The peroxisome knowledge base was designed to organize, integrate, visualize and navigate information on the peroxisome and related disorders. To promote the role of the knowledge base as molecular information interface for physicians, it may be beneficial to integrate the knowledge base with electronic patient files and to integrate concept maps with clinical information from electronic health records, clinical trial reports, hospital databases, biobanks and other clinical repositories. For example, this integration could be established through concept maps like the one shown in Figure 5 (Supplementary Material) and would possibly result in further overview and insights in (patho) physiological conditions of patients suffering from peroxisomal disorders.

Another application of the knowledge base will be the guidance of researchers in the analysis and interpretation of high-throughput genomics data, such as produced in the fields of genomics (i.e. microarray data) and metabolomics (i.e. metabolite level measurements), in the context of biological pathways. Several systems have been developed for the visualization and analysis of experimental data in the context of pathways. Well-known examples include KEGG, GenMapp and Cytoscape. KEGG and GenMapp, for example, provide graphical pathway maps that are used to map and analyze genomics data. However, the information presented by these maps, or the underlying database, is not always sufficient when applied for interpreting experimental data. For example, consider the map for fatty acid β-oxidation (KEGG identifier: 00071). This map shows the breakdown of fatty acids from palmitoyl-CoA (hexadecanoyl-CoA) to acetyl-CoA through the iteration of four enzymatic steps. This map, however, does not show that β-oxidation of other fatty acids, such as 2-methyl branched fatty acids, very long chain fatty acids and dicarboxylic acid, takes partly place in the peroxisome, while the remainder of the breakdown, depending on the length of the carbon chain (C < 11), will be performed by the mitochondrial β-oxidation machinery. Furthermore, it is not clear from this map that, depending on the substrate, different enzymes are used for the same pathway step, as is explicitly shown by our concept maps (Fig. 8 Supplementary Material). Depending on the experimental study, this might be important information for the interpretation of the data. Pubmed provides 1516 references with ‘fatty acid oxidation’ in the title and 70 918 references with ‘fatty acid’ in the abstract. Clearly, the amount of information not only provided by KEGG but also by our knowledge base is still limited. Our concept maps will complement traditional pathway maps, as provided by KEGG, to supply biological knowledge required for detailed analysis and interpretation of experimental data, not only
with knowledge on pathways and related domains but also basic biochemical textbook information. Of course, as part of ongoing work, the content of our knowledge base should be extended for this purpose. We also will integrate the knowledge base with pathway visualization databases such as KEGG and visualize information from those databases in relation to our concept maps.

Conflict of Interest: none declared.

REFERENCES


