Visual Representation of Cellular Networks

Alexander Mazein

Doctor of Philosophy

Centre for Intelligent Systems and their Applications

School of Informatics

University of Edinburgh

2011

To my children, Ilya and Masha

Abstract

Development of advanced techniques for biological network visualisation is crucial for successful progress in the areas of systems-level biology and data-intensive bioinformatics.

However, current techniques for biological network visualisation fall short of expectations for representing extensive biological networks. In order to provide really useful network visualisation tools, new approaches have to be proposed and applied alongside with those most powerful features of current visualisation systems. The resulting representation techniques have to be tested by applying to large-scale examples that would include metabolic, signaling and gene expression events. User survey should also be carried out to further prove the advantages of the new techniques.

The present thesis describes an attempt to achieve the above objectives, by performing the following steps: 1) existing approaches in the area of network representation were analyzed and their shortcomings and advantages were defined; 2) new notation has been developed, in which, the defined best features of the existing systems were integrated with newly introduced potent features such as compact visualization, 'functional gate' and 'identity gate', 4) new framework was developed that allows managing large-scale networks that are represented on different levels of details and different levels of constrains, while keeping each diagram semantically unambiguous, 5) extensive examples, including genome-scaled human metabolic network and TNF-alpha receptor signalling network, were used to prove that the designed notation and the framework can be applied efficiently, and, finally, 6) a notation survey has been carried out to validate the advantages of the newly developed notation over the existing ones.

Declaration

I declare that this thesis has been composed by myself and that it has not been submitted, either in whole or in part, in any previous application for a degree. Except where otherwise acknowledged, the work presented is entirely my own.

Alexander Mazein

Contents

Acknowledgments

This research would not be possible without constant support and advice from Professor Igor Goryanin. I have been very fortunate to receive his guidance and help beginning with our first meeting in the Edinburgh airport in September 2005.

I am also most grateful to my friend and colleague Anatoly Sorokin for many helpful and inspiring discussions regarding my work.

I would like to thank my colleagues Stuart Moody and Hongwu Ma for their provocative questions and discussions on the topics related to this research.

The financial support for this research was provided by the Dorothy Hodgkin Postgraduate Awards (sponsored by ERSRC and GlaxoSmithKline) and the Centre for Systems Biology at Edinburgh, and I would like to express my gratitude for meeting my tuition fees, a maintenance allowance and research cost.

I would like to thank the team of the Edinburgh Pathway Editor developers, and especially Richard Adams and Shakir Ali for providing new functionalities of the EPE that were required for this research.

My deepest appreciation goes to my friends Sveta Milyaeva and Nestor Milyaev for their invaluable help in preparing the text of the thesis.

Since I am not a native English speaker, I would like to thank my colleague Richard Adams for his help in revising the text of the thesis.

List of Figures

List of tables

List of abbreviations

EC	Enzyme Commission
EGFR	Epidermal growth factor receptor
EHMN	Edinburgh human metabolic network
EPE	Edinburgh Pathway Editor
ERK	Extracellular signal-regulated kinase
JNK	c-Jun N-terminal kinase
HCI	Human Computer Interaction
IFN-gamma	Interferon gamma
KEGG	Kyoto Encyclopedia of Genes and Genomes
MIM	Molecular Interaction Maps
NF-kappaB	Nuclear factor of kappa light polypeptide gene enhancer in B-cells
NA	Not applicable
NR	Network Representation framework
p38-MAPK	Mitogen-activated protein kinase p38
PDK 1	Pyruvate dehydrogenase kinase isozyme 1
PFK2/F2,6BPase	6-Phosphofructo-2-kinase / fructose-2,6-biphosphatase 2
RAS	RAS protein subfamily (an abbreviation of RAt Sarcoma)
RSK2	Ribosomal S6 kinase 2
SBGN	Systems Biology Graphical Notation
SBGN AF	Systems Biology Graphical Notation Activity Flow
SBGN ER	Systems Biology Graphical Notation Entity Relationship
SBGN PD	Systems Biology Graphical Notation Process Description
SBML	Systems Biology Markup Language
TLR	Toll-like receptor
TNF-alpha	Tumour necrosis factor alpha
UN	Unified Notation

Publications

1. Ma H, Sorokin A, Mazein A, Selkov A, Selkov E, Demin O, Goryanin I (2007) The Edinburgh human metabolic network reconstruction and its functional analysis. Mol Syst Biol 3: 135

Comment: My contribution to this paper was the approach used for subdividing the extensive Edinburgh human metabolic network into smaller metabolic pathways and practical implementation of this approach for preparing maps of Edinburgh human metabolic network. This experience helped developing principles used in the Network Representation framework. Also during this very practical task it became clear that it was necessary to address these questions: 1) what to do when generic and specific metabolites appear on the same map? 2) how connections between different maps could be shown? 3) how to show incomplete or missing information? 4) what is the best way to prepare a system that could be used for navigation and easier understanding of the network structure?

2. Mazein A, Sorokin A, Adams R, Moodie S, Golyanin I (2011) Visual Knowledge Management. International Journal of Knowledge Engineering and Data Mining (IJKEDM), (in press)

List of materials included in the CD

1. A copy of the thesis.
2. Portable version of EPE with preinstalled database that contains human metabolic outlines and TNF receptor signalling maps.
3. Chapter 2 figures in high resolution. The TLR example in the UN and the process description notation.
4. Chapter 3 figures in high resolution. EGFR example.
5. Chapter 4 figures in high resolution:
    a) The Edinburgh human metabolic network pathway diagrams in the UN;
    b) The system of outlines for the Edinburgh human metabolic network.
6. Chapter 5 figures in high resolution:
    a) The TNFR signalling detailed maps in the UN;
    b) Outlines for the TNFR signalling network;
    c) Regulation of Glycolysis diagram;
    d) EGFR alteration in glioblastoma examples.
7. The Graphical Notation Survey materials:
    a) Pre-survey;
    b) Final version of the survey;
    c) Survey results in Excel format
8. The SBGN proposal materials.

Introduction

Motivation. Any ambitious project in the area of biological network visualisation requires a highly advanced environment in which biological knowledge could be comprehensively represented and stored in such way that information could be easily shared, updated and re-used. Despite noticeable progress in the area of Systems Biology, there is still a lack of effective approaches that would allow us to amalgamate all available information on biological networks in one powerful system. Furthermore, to demonstrate their effectiveness these approaches would have to be not only well-thought-through, but also supported by comprehensive and wide-ranging examples. While developing these extensive test diagrams, the approaches should be modified if required.

Objectives. The main objectives of this research are:

1. To design a new system of symbols for unambiguous biological network visualisation. In order to do so a large number of the existing notations' features have to be combined with newly introduced entities into a new notation. This notation is introduced in this work as the Unified Notation (UN). The diagrams in this notation have to be both human- and machine-readable. The Unified Notation should be able to represent all types of process description maps: metabolic pathways, signalling pathways, gene expression events, and maps combining these pathways.

2. To develop a framework where a large scale network could be represented efficiently. For this purpose the thesis suggest (A) subdividing a large network into

comparatively small, manageable and easy-updatable diagrams that can be used and reused as elements for assembling large networks for a particular cell type and a particular organism; (B) developing an approach for creating a system that could represent a network on different levels of granularity; (C) developing a method that unites multiple detailed maps into a single virtual map.

3. To test the newly developed notation and the framework using large-scale biological networks. This will ensure the system's consistency and its ability to visualise many different types of networks.

Software. Although the approach is developed to be used in many different editors thus accommodating all or some of the features proposed, to test the new features it is suggested using a tool (editor) that would meet the following requirements:

- information on a map has to be both human and machine-readable, therefore properties of each object have to be stored in a database the software creates, so it could be potentially exported, imported, transformed into other formats or translated into other graphical languages if needed; in this way the software can be used as a data-sharing environment
- it should be able to create/change/delete user-defined objects with user-defined properties, with the aim that the entities of new notation could be tested, added or removed;
- it should be possible to use hyperlinks so that one map could be connected to another;
- possibility to work in collaboration with software developers so new features could be introduced if necessary.

All these requirements are met by the Edinburgh Pathway Editor (EPE) (Sorokin, 2006) (EPE, http://epe.sourceforge.net/SourceForge/EPE.html) that facilitated the development of all the diagrams. Please see Section 2.6 for more details on the Edinburgh Pathway Editor.

It is of crucial importance that the research approaches were developed and tested relying not on one particular editor, but introduced as general methods that can be used for efficient biological network visualisation with any appropriate software.

Development in the field since this research started. The work started in 2006 and since then there have been a number of significant developments in the area of graphical biological network representation. The most important of all was the initiation of the SBGN project, which was supported by scientists from many different fields . In many ways the notation part of this research overlaps with the efforts related to developing the SBGN PD language therefore a detailed comparison of the two notations is provided in Chapter 2. The experience of using the Unified Notation to represent metabolic and metabolism regulation pathways were utilised in the proposal presented at the COMBINE meeting in 2010 (COmputational Modeling in BIology Network, http://sbml.org/Events/Forums/COMBINE_2010, fore more details please see Chapter 9).

Structure of the thesis

Chapter 1 gives an account of the existing research in the field.

Chapter 2 introduces the Unified Notation (UN).

Chapter 3 describes the Network Representation Framework (NRF).

Chapters 4 and 5 provide examples of using the novel approaches. Chapter 4: the Edinburgh Human Metabolic Network representation. Chapter 5: TNF-alpha receptor network representation.

Chapter 6 summarizes results of the Graphical Notation Survey.

Chapter 7 describes the features of the Unified Notation aiming at improving the SBGN PD language

All the necessary additional information is provided in the appendices and on the CD enclosed with each copy of the thesis. Given that the thesis contains some figures that can not be viewed in A4 format, the CD also contains high resolution versions of the figures.

Chapter 1

Review of related work

1.1 Human-Computer Interaction

The size of current datasets and the scale of analyses required in biological science increasingly depend on computer-aided support. On the other hand, it is important that the computer-based tools provide adequate level of functionality while being convenient to use. In order to represent vast amounts of information in a meaningful and convenient way, the computer-based tools inevitably should use bleeding-edge features and paradigms (Keefe, 2010, Pettifer, 2009). The discipline that 'concerned with the design, evaluation and implementation of interactive computing systems for human use and with the study of major phenomena surrounding them' (ACM SIGCHI Curricula for Human-Computer Interaction, http://old.sigchi.org/cdg/ cdg2.html#2_1) is the area of Human Computer Interaction (HCI).

Unfortunately there is no unified theory of HCI that currently could be presented (Dix et al, 2004; Sears, 2008; Zaphiris, 2009). In many ways this area is as much science as it is a craft and involves a lot of experimentation on the part of software developers, experts from linguistics, graphic designers and scientists from many different fields (Beaudouin-Lafon, 1993; Dix et al, 2004).

This section reviews the HCI conceptions that are applicable to the challenges related to biological networks visualisation.

'Usability' is one of the main conceptions in HCI and initially was defined as 'ease to use', 'user-friendliness' and 'easy to learn' (Karat, 2003). This understanding is still widely accepted (Dix et al, 2004). The 'ease of use' is often one of the crucial parameters that would determine user decision whether to use a system or not.

The following set of usability criteria is identified based on the analyses by Nielsen and Shackel (Folmer, 2004):

- Effectiveness,
- Learnability,
- Ease-to-remember,
- Flexibility
- Attitude/satisfaction (subjective component).
- Reliability

Shneiderman proposed usability as research agenda (http://universalusability.org/researchagenda) that includes the three issues listed below (Shneiderman, 2000):

- Technology variety ("supporting a broad range of hardware, software, and network access" (http://universalusability.org/technologyvariety));
- User diversity ("accommodating users with different skills, knowledge, age, gender, disabilities, disabling conditions (mobility, sunlight, noise), literacy, culture, income, etc." (http://universalusability.org/userdiversity));
- Gaps in user knowledge ("bridging the gap between what users know and what they need to know" (http://universalusability.org/gapsuserknowledge)).

In this research we define main criteria for usability as effectiveness, ease to use, ease to maintain, ease-to-learn and ability to bridge gaps in user knowledge.

The newly developed notation satisfies the principle of maintainability by using comparatively small and easy to maintain low-level maps that form a library of 'building blocks'. These small maps are then used for building representations of large cellular networks (see Chapters 4 and 5). When underlying mini-maps are updated because the underlying metabolic process gets rectified, these changes are reflected in the large maps. Therefore, updating a mini-map in only one place saves great amount of work on maintaining such maps.

Ease-to-learn criteria is satisfied by using a notation with minimal number of symbols that are easy to remember (see Section 8.7.11). Many of these symbols are similar to those used in other notations, such as SBGN, which makes cross-learning much easier. At the same time the used set of symbols remains being sufficient for unambiguous cellular network representation (effectiveness criteria) (see Sections 6.5.9).

Human limitations in processing of information have important implication for design (Dix et al, 2004; Olson, 2003; Sears, 2008). One of the most currently widely spread design philosophies is the user-centered design (Zaphiris, 2009) that makes users the most important element of design process. Wickens and co-authors (Wickens, 2004) have developed a list of principles for display design that take into account human's limitations. These principles are subdivided into 4 categories: perceptual principles, mental model principles, principles based on attention and memory principles. For example, one of the memory principles is the principle of consistency: new information is easier to understand if it is designed in a consistent manner from one display to another.

In the current research the principle of consistency of representation was used in providing translation mechanism between levels in the Network Representation Framework. According to our approach, moving between different levels of details the user stays within a comfortable familiar environment. That was possible due to using the same set of objects at all levels. Only small amount of new information is added when moving to a more detailed level as described in Sections 4.3 and 5.2. Another application of this principle is used in the design of enzyme function representation: ‘functional gate’ entity represents enzyme function consistently on different types of diagrams (see Section 6.5.3) and, in fact, provides consistent representation when events from different domains, such as metabolic and signalling pathways, have to be shown on a single diagram, for example for representing metabolism regulation (see Section 5.3).

Progress of HCI research in the area of display design and information visualisation is necessary for further development of biological science and a number of researches and software developers address this concern (for example, Kang, 2008, Lee, 2006; Lieberman, 2011; Keefe, 2010; Perer, 2009; Plaisant, 2008; Shneiderman, 2006; also see Section 2.6 on the Edinburgh Pathway Editor).

In this work we proposed to use a combination of existing paradigms together with the newly developed ones and test these on large real-life networks. Additionally we have performed a comparison of the usability of the new techniques and paradigms against the existing ones in a survey (Chapter 6). We hope that the outcomes of the current research would allow to design a better biological network representation system.

The next section discusses existing approaches in systems biology in relation to biological network visulisation.

1.2 Network visualisation in systems biology

According to Kitano (2002) the life cycle of systems biology research comprises the following stages: identification of a problematic fact, development of formalised research model, formulation of a hypothesis, simulation and analysis of the model, experimental design, experimental validation and data analysis and identification of a next iteration target.

As highlighted in the survey done by Klipp et al. (2007), visualisation and drawing techniques have always been considered as part of the dissemination process rather than the development process. At the same time, in such a cross-disciplinary area as systems biology, the dissemination process becomes an essential part of the research life cycle. It is worth noting that all the steps above are conducted by the same team of researchers or become a part of the same project. Nevertheless, all six steps of the research should have an overarching goal thus and inform/be informed by knowledge generated both experimentally and computationally at each stage of the research. For the most part in systems biology experimentally oriented groups embark on stage one, three and five, while theoretical groups perform steps one, two, three and five. Let us consider what types of diagrams representing different types of knowledge are required at different stages of the development process.

At the beginning of the modelling stage (Stage 2), reserachers identify the biological scope of a problem to solve, therefore define a context for further model development. Here 'at a glance' or 'bird's eye' types of diagrams are typically used. A good example of the 'at a glance' diagram is the PI3K/Akt Signalling pathway (http://www.cellsignal.com/pathways/akt-signalling.jsp) provided by CellSignal (CellSignal, http://www.cellsignal.com).

Further development at this stage requires more detailed information about what is known and what are the 'known unknowns' about the system. Pathways collections from Biocarta (BIOCARTA: Charting pathways of life**,** http://www.biocarta.com), ProteinLounge (ProteinLounge, http://www.proteinlounge.com), KEGG (Kanehisa et al., 2002), BioCyc (Karp et al., 2005), Edinburgh Human reconstruction (Ma et al., 2007) and diagrams in MIM (Molecular Interaction Maps) (Kohn, 2001; Kohn and Aladjem, 2006), CellDesigner (Kitano et al., 2005), or Cytoscape (Shannon et al., 2003) and other drawing tools are used for this purpose. The recently proposed SBGN (Systems Biology Graphical Notation, www.sbgn.org) notation is developing to standardize this type of diagrams.

At the third step (Formulation of a hypothesis. Simulation and analysis of the model), the 'known unknowns' are replaced by the assumptions and biological hypothesis and, at this stage, some initial mathematical models can be formulated. The visualisation notations at this point can vary and depend upon the modelling platform. Models could be represented as SBML layout diagrams (Deckard et al., 2006; Gauges et al., 2006). Petri nets (Pinney et al., 2003) and LiveStateCharts (Kam et al., 2001) have their own visual notations.

The choice of visual notation for the fifth step (Experimental validation and data analysis) generally depends upon the target audience. Usually it combines 'model' or 'pathway' diagrams with charts, histograms, scatter plots from experimental data and results of data analysis.

At the last step (Identification of a next iteration target) we will use 'at a glance' diagram again to visualise results and unresolved issues, and define a context for the next iteration of the research.

To sum up, all described diagrams, used in a project life cycle, could be clustered into three main categories according to their purpose and the biological question under consideration:

1. Outline and biological context visualisation. These help answering the question: What is the context or biological domain and scope for modelling?
2. Existing experimental knowledge detailed representation. These help answering the question: What are the known facts?

3. Mathematical (*in silico*) model structure representation. These help answering the question: What is the detailed hypothesis for experimental computational check?

Despite their importance, diagrams currently used in bioinformatics research are often ambiguous. On the one hand, the ‘at a glance’ diagrams, used at the steps one and five of the systems biology research cycle, by necessity are very ambiguous. This type of diagram requires textual description or explanation attached to it and should not be treated as an independent source of knowledge.

On the other hand, the detailed diagrams usually do not contain enough evidence and references to experimental data for hypothesis formulation, and they should be upgraded by human annotators by providing references to original experiments and/or establishing casual relationships using intelligent guessing or assumptions. Ambiguity caused by missing knowledge or the lack of detailed information could be resolved by the ability to visually highlight such ‘grey areas’. In this case, strong evidence and corroborating data should be assigned to each element on the diagram. ‘Grey areas’ could be considered as target for modelling or experimental analysis.

1.3. Notations for unambiguous representation of cellular networks

According to Hiroaki Kitano and co-authors ‘a successful diagram scheme must: (i) allow representation of diverse biological objects and interactions, (ii) be semantically and visually unambiguous, (iii) be able to incorporate notations, (iv) allow software tools to convert a graphically represented model into mathematical formulas for analysis and simulation, (v) have software support to draw the diagrams, and (vi) ensure that the community can freely use the notation scheme’ (Kitano et al., 2005).

Currently three notations are discussed in detail: the Molecular Interaction Map (MIM) notation, the Process Diagram Notation and SBGN (Systems Biology Graphical Notation).

Below we offer a brief overview and cross-comparison of each of these notations.

The Molecular Interaction Map (MIM) notation. Kurt Kohn was the first to introduce a well-defined notation that was able to visualise relationships between molecules in biological networks. To compare to the process diagram notation maps, the MIM diagrams differ in two ways: first, they ‘show each named molecular specie only once on a map’ and, second, ‘they do not specify particular event sequences, but instead show all of the interactions that can occur if potentially interacting species are in the same place at the same time’ (Kohn and Aladjem, 2006).

The Process Diagram Notation (Kitano et al., 2005) is probably the most used today. Its main advantage is in allowing one to show cellular, metabolic and signalling events in an unambiguous way. Large examples have been prepared (Oda and Kitano, 2006; Oda et al., 2005) and can be used as a source of knowledge about the biological system. Moreover, a special tool (CellDesigner: a modeling tool of biochemical networks, http://www.celldesigner.org/) is available for building diagrams in the Process Diagram Notation and storing them using the Systems Biology Markup Language (SBML).

On the basis of MIM notation and the Process Diagram Notation two languages of SBGN are being developed: Process Description (PD) language and Entity Relationships (ER) language (Systems Biology Graphical Notation, www.sbgn.org).

Systems Biology Graphical Notation (SBGN). Systems Biology Graphical Notation project was initiated as a community effort by Kitano in 2005 (Kitano et al., 2005). Currently SBGN (Systems Biology Graphical Notation, www.sbgn.org) consists of three notations: Process Diagram (SBGN PD), Entity Relationship Diagram (SBGN ER), and Activity Flow Diagram (SBGN AF) (Le Novere et al., 2009). They aim to represent different slants of the same biological information. SBDG PD notation inherits many graphical elements and features of the process diagram notation. According to the original definition, this language describes 'the causal sequences of molecular processes and their result' (Systems Biology Graphical Notation: Process Diagram Level 1, www.sbgn.org, p.3). Entity Relationship language represents 'interactions between entities irrespective of sequence' (Systems Biology Graphical Notation: Process Diagram Level 1, www.sbgn.org, p.3). Finally, the Activity Flow language is used to describe: 'the flux of information going from one entity to another (Systems Biology Graphical Notation: Process Diagram Level 1, www.sbgn.org, p.3).

While the first two languages in many ways correspond to the original Process Diagram Notation and the Molecular Interaction Map (MIM) notation, SBGN Activity Flow notation addresses the requirement to represent a network in a compact way showing only activity flow between entities.

The symbols of the process diagram notation and the SBGN PD will be reviewed in more detail in sections 2.4 and 2.5.

1.4. Genome-scale metabolic network representation

There have been many attempts to visualise genome-scale metabolic network as a wall chart or as a compact outline.

The first example of wall chart diagrams were hand-drawn by Michal and called the Metabolic pathway wall charts (Michal, 1968).

Another example is the IUBMB-Sigma-Nicholson Metabolic Pathways Chart (IUBMB-Nicholson: Metabolic Pathways Chart, http://www.iubmb-nicholson.org/chart.html). The last version - the 22nd edition, 2003 - consists of about 600 reactions that are identified by the IUBMB Enzyme Commission numbers (EC) (Deckard et al.). Although the chart contains only the 'textbook' pathways it provides a useful overview of metabolic pathways and can be successfully used for education purposes. On the other hand, as the chart is not linked to a constantly updated pathway database, that makes it less useful for complete network visualisation. Also the readability of such large maps is questionable - even with such a limited number of reactions it is difficult sometimes to follow links from one compound to another when they are far from each other on the chart.

One more example of the wall chart diagrams is the Roche Applied Science 'Biochemical Pathways' wall chart (Access to the digitized version of the Roche Applied Science 'Biochemical Pathways' wall chart via ExPASy, http://www.expasy.ch/cgi-bin/show_thumbnails.pl). Despite the obvious usefulness of such diagrams the questions remain: How easy it is to read the diagram? How easily can the diagram be updated? Can the compromise between a space limit and the necessity to show a great number of biological events be avoided?

Comparatively small 'at a glance' diagrams present an overview of a metabolic network. Often the same scheme is reused when placing a particular phenomenon in the focus.

Simplified diagrams appear to be also useful for a network overview. For instance, a colourful map uploaded to Wikipedia (Wikipedia: Metabolic pathway, http://en.wikipedia.org/wiki/Metabolic_pathway) represents metabolic network as key metabolites and links between them. A coloured background shows major parts of a metabolic network or a particular metabolic pathway, for example the 'Fatty acid metabolism', 'Amino sugar metabolism' or 'Urea cycle'.

Recently, KEGG database introduced the KEGG Atlas (Okuda et al., 2008). The KEGG Atlas (KEGG Atlas, www.genome.jp/kegg/atlas/) has been developed as a graphical interface to the KEGG pathways. KEGG metabolism map (www.genome.jp/kegg/atlas/metabolism/2/) by manually combining 120 metabolic pathway maps. It is available online and is proposed to be used as a reference map to detailed KEGG pathways. In the KEGG metabolism maps, each node (circle) is a chemical compound identified by the C number. Each line (curved or straight) connecting two nodes is manually defined as a segment lacking branches in the existing maps, named NetElement, and identified by the N number. Each NetElement corresponds to one to several KO's in the reference pathway view, or one to several KO's in the reference pathway view, or one to several genes in an organism-specific view. Different parts of a metabolic network are shown in a different colour, key compound are shown as nodes and detailed map names are shown in corresponding places. Although compound names are not shown the main structure of the network is shown by nodes and lines.

Other network visualisation projects have applied the concept of hierarchical levels of detail. The Reactome project (Reactome, http://www.reactome.org; Joshi-Tope et al, 2005) offers interactive maps for navigation through their pathways, which include metabolic, signalling and gene expression pathways. Detailed pathways at different levels of detail share a simple common visual notation but use text descriptions to add semantic detail.

Our experience in representing metabolic networks can be summarized as follows: 1) successful diagrams are manually prepared, 2) small diagrams are easier to read, 3) despite their size, large wall charts are not able to include all cellular events, 4) interactive diagrams can be used for navigation through a metabolic pathway database.

1.5. Semantic zooming approach

In order to deal with the huge amount of biological data, new technologies have to be considered for systems biology studies. Taking into account that detailed biological diagrams can be compared to geographical maps, the approach used in such popular software as Google Earth (Google Earth, http://earth.google.com), for example, might be taken as prototype for biological networks visualisation.

Semantic zooming is being discussed as a perspective approach for graphical data representation in systems biology (Hu et al., 2007). It is important that in addition to scale enlargement semantic zooming also introduces new properties. Although objects remain the same, their number and their description can differ from layer to layer depending on context and the system of representation.

An example of such expansion of a network while zooming is found in the work on the KEGG database hierarchy. Hu and co-authors has detailed the metabolic network on four levels (Hu et al., 2007). Metabolism is represented by 1 module on the first level. Then it is branched into 8 modules on the second level. The third level consists of 161 modules and fourth level contains 810 elements.

The summary of the advantages of the semantic zooming are: 1) semantic zooming handles any large network as soon as objects in it are connected; 2) the data structure dictates how the hierarchy in each case must be developed; 3) if detailed diagrams are stored using an unambiguous notation, the semantic zooming approach can be used to make higher levels of network representation semantically unambiguous.

Chapter 2

The Unified Notation

2.1. Introduction to the Unified Notation

The Unified Notation (UN) has been developed for representing cellular pathway diagrams. It incorporates the most advantageous features of existing notations. Similarly to KEGG representation (KEGG, www.genome.jp/kegg), rather than using ‘label inside a shape’ visualisation, ‘label near the symbol’ is used in the UN. This type of representation is based on the assumption that it will make the diagrams more compact and easy to understand. In the UN, signalling networks are represented as a set of reactions/state transitions and activation/inhibition links from activators/inhibitors to corresponding processes, similarly to the process diagram notation (Kitano et al., 2005). Symbols for Compound, Monomer, Homodimer, Heterodimer and Complex were adopted from GenNet maps (Ananko et al., 2002). Logic Gates and the concept of Multistate Protein/Complex representations were derived from the Edinburgh Pathway Notation (EPN) (Moodie et al., 2006). Furthermore, to address some shortcomings of the existing notations, three new symbols - the multistate entity, the identity gate and the functional gate - are added. The next section provides the detailed description of the UN.

2.2. System of symbols for unambiguous representation of cellular networks

By entity here we understand any element/symbol of the notation that can be singled out from other elements/symbols on a diagram.

2.2.1. Unknown/proposed entity (Table 2.1, A) symbol is used when an entity structure is unknown. It is represented by a rounded rectangle. For instance, a Toll-like receptor ligand can be a compound or a protein and it can be shown as a generic entity with unspecified structure.

2.2.2. Compound (Table 2.1, B) is an entity identical to the 'simple molecule' in the process diagram notation (Kitano et al., 2005) and is represented by a square box.

2.2.3. Monomer (Table 2.1, C) is an entity identical to the 'protein' in the process diagram notation (Kitano et al., 2005) and is represented by an ellipse.

2.2.4. Complex in the UN can be shown as **Heterodimer** (Table 2.1, E), **Homodimer** (Table 2.1, D) and **Complex** (Table 2.1, F). A heterodimer is represented by an ellipse and a circle. A homodimer is represented by two ellipses. A complex is represented by three circles. **Complex** is used if there are more than two macromolecules involved. A Complex composition is expressed in the complex name. Each element is shown on its own line in a list near the symbol. Multimers are shown as (PROTEIN NAME) n where n is cardinality. A complex state is shown near the symbol as a list of states of its component entities. The symbols for Heterodimer, Homodimer and Complex have been adopted from GenNet maps (Ananko et al., 2002).

2.2.5. Gene (Table 2.1, G) - any DNA species are represented by a rectangle. A type of DNA species can be specified inside a rectangle or in an entity name.

2.2.6. RNA (Table 2.1, H) - any RNA species including mRNA, tRNA, miRNA etc. A type of RNA species can be specified inside RNA shape.

2.2.7. State (Table 2.1, I) of an entity is shown as a label near a corresponding symbol. As in the process diagram notation (Kitano et al., 2005) the symbol '@' is used to separate a modifier from a modified site. State naming rules for a monomer and complex are described in Table 2.2. If detailed information is not available, the state can be described in general way, for example: 'unknown state', or 'active'. Unmodified states remain unlabelled.

2.2.8. Multistate entity (Table 2.1, J). Anatoly Sorokin (personal communication, 2007) suggested including all states of the same species into a special container to make the diagram easier to read. As a result, the multistate entity unites several states of the same specie into one container, thus saving space given that the name of the species is shown only once as a name of container 'multistate entity'.

2.2.9. Process (Table 2.1, K) entity includes reaction, association, dissociation and state transition, represented by a small circle. If these are identified processes, the Process symbol is in black colour. A circle remains uncoloured for proposed processes. This entity has at least two connectors: input and output.

2.2.10. Translocation (Table 2.1, L) symbol shows the transport of species between different compartments, and is represented by a circle with the letter 'T' inside. This entity has at least two connectors: input and output.

2.2.11. Gene expression (Table 2.1, M) symbol depicts translation of genetic information from DNA to RNA and from RNA to protein, and is represented by a small square. On a diagram this symbol has to be at least twice as small as a square box used for representing a compound so the two symbols could be distinguished easily.

2.2.12. Omitted processes (Table 2.1, N) symbol is used to mark hidden known or unknown steps. Similarly to the process diagram notation, the entity is represented by two lines.

2.2.13. Unknown/proposed process (Table 2.1, O) symbol is used in cases where the evidence for a process is unconfirmed, and is represented by a circle with a question mark inside it.

2.2.14. Activation (Table 2.1, P) link shows the stimulating influence of an entity on a process. It links the entity to the 'process' symbol. Represented by a link from an activator (protein or complex) to corresponding 'process' symbol with an arrow at the end (via a logic gate if more than one protein/complex involved).

2.2.15. Functional gate (catalysis) (Table 2.1, Q) was initially introduced to keep the KEGG-like appearance of metabolic maps where EC numbers are usually shown. It also marks out catalysis and clarifies the function of the enzymes both in metabolic and signalling networks. The functional gate is represented by a rectangle with EC or TC number inside it. In case of a metabolic pathway the functional gate is not visually connected to any entity but is positioned near a corresponding 'process' symbol. If particular proteins/complexes with such enzyme activity are on a diagram, they are connected to the process via the functional gate.

2.2.16. Inhibition (Table 2.1, R) shows negative influence on a process and connects an entity to the 'process' symbol. The inhibition is represented by a link from an activator (protein or complex) to a corresponding 'process' symbol with a bar at the end (via a logic gate if more than one protein/complex involved).

2.2.17. Logic gates (Table 2.1, S) originate from the Edinburgh Pathway Notation (EPN) (Moodie et al., 2006). **AND gate** (Table 2.1, S) is used when all of the entities are required to affect a process. **OR gate** (Table 2.1, T) is used when any of the entities linked to the OR gate can affect a process. **NOT gate** (Table 2.1, U) is used when absence of an entity is required for activation/inhibition of a process. The AND logic gate is represented by a circle with the symbol '&' inside it. The OR logic gate is represented by a circle with ‘OR’ inside it. The NOT logic gate is represented by a circle with ‘NOT’ inside it.

2.2.18. Identity gate (Table 2.1, V) is a newly introduced solution for visualisation of connection between generic and specific entities. The Identity gate is represented by a circle with the symbol ‘≡’ inside it.

2.2.19. Compartment (Table 2.1, W) is a container that is used as a background for appropriate entities.

2.2.20. Link (Table 2.1, X) marks connections to other maps. Represented by underlined text or by a circle with the ‘>‘ symbol inside it.

2.2.21. Degradation (Table 2.1, Y) is an element that is used to show the removal of an entity when details of the degradation process are not important or unknown.

2.2.22. Text (Table 2.1, Z) fields can be used for information that cannot be shown by using any of the other elements of the notation.

The multistate entity (Table 2.1, J) is designed to reduce redundant diagram components in cases where successive state transitions alter a small number of the subunits of the complex only chemically. This results in the compaction of a diagram so that all states of the same proteins/complexes are located close to each other. Sections 2.4 and 2.5 give examples of the use of this symbol, compared to the SBGN Process Diagram notation (SBGN PD) (SBGN Process Diagram Level 1 specification, http://sbgn.org/Documents/Specifications) and to the process diagram notation (Oda and Kitano, 2006).

The identity gate (Table 2.1, V) is a solution for the generic/specific compound problem. For example, quite often a subtype of a fatty acid, be it an amino acid, acyl-CoA or D-glucose (which is alpha-D-glucose or/and beta-D-glucose), could be an appropriate reactant. However, there are occasions where the specific compound is required. To make our network consistent we need to show that the generic compound 'amino acid' is understood as a collection of all particular amino acids. In these cases, the ability to link generic and particular compounds (or a protein/complex in a particular state with the same protein/complex in unspecified state) via the Identity Gate is important, especially if both the generic and specific compound names are shown on the same map. Diagrams can be created to demonstrate parent-child relationships for a particular generic compound. Figure 2.1 shows such a diagram for a particular type of the generic compound, namely the 'fatty acid'. The diagram displays the parent-child relationships between the generic compound 'fatty acid' (on the left), the intermediately specified compounds (in the middle), and the specific fatty acids (on the right). If a specific fatty acid is shown on the map of fatty acid biosynthesis, we should use the generic 'fatty acid' on another map in order to visualise the reaction of acyl-CoA formation (EC 2.3.1.85).

The functional gate (Table 2.1, Q) is the way to use one of the most successful representations of metabolic pathways (KEGG metabolic diagrams) and, at the same time, be able to show signalling events on the same diagram without changing elements of metabolic pathway representation.

Despite the fact that the EC number does not uniquely identify proteins/complexes involved, it is still very valuable information for a reader and is expressed in a concise way. Using EC makes it much easier to find a particular protein connected to the reaction.

Here it should be said that EC on a signal transduction diagram is not that helpful as EC on a metabolic diagram, because the phosphorylation is the most common way to activate a protein, but not many ECs are connected to phosphorylation. Therefore, in case of signalling pathway it is reasonable to avoid using EC and names of the proteins can be used as identifiers instead.

On the other hand, it is very common that a metabolic reaction is linked to many different proteins/complexes, and an attempt to show all of them on a diagram is most likely to make it very difficult to read, as well as requiring a lot of additional unnecessary work.

Table 2.1. The Unified Notation's system of symbols.

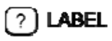
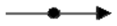
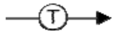
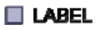
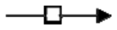
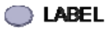
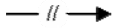
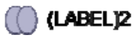
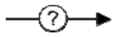
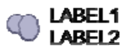
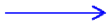
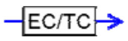
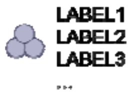
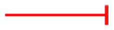
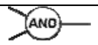
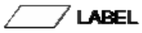
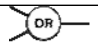
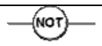
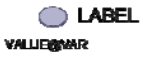
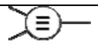
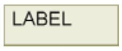
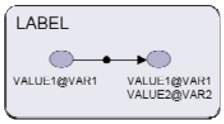
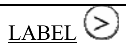
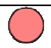
Symbol	Entity	Symbol	Entity
	A. Unknown/proposed entity		K. Process
			L. Translocation
	B. Compound		M. Gene expression
	C. Monomer		N. Omitted processes
	D. Homodimer		O. Unknown/proposed process
	E. Heterodimer		P. Activation
			Q. Functional gate (catalysis)
	F. Complex		R. Inhibition
			S. AND gate
	H. RNA		T. OR gate
			U. NOT gate
	J. State		V. Identity gate
			W. Compartment
	K. Multistate entity		X. Link
			Y. Degradation
		LABEL	Z. Text

Table 2.2. State naming rules.

State description	Comments
MONOMER STATE	
active	General description is allowed if detailed information is not available
MODIFIER @ SITE	One site of a protein is modified
MODIFIER 1 @ SITE1 MODIFIER 2 @ SITE2 ...	Several sites are modified simultaneously
MODIFIER @ ?	Site is unknown
? @ SITE	The type of modification is unknown
n MODIFIER @	The same modification for several sites. Only for cases when it is not essential (or unknown) which sites are modified
COMPLEX STATE	
PROTEIN1 MODIFIER1 @ SITE1 PROTEIN2 MODIFIER2 @ SITE2 ...	Complex state is described similarly to a monomer state description except a protein name is shown first

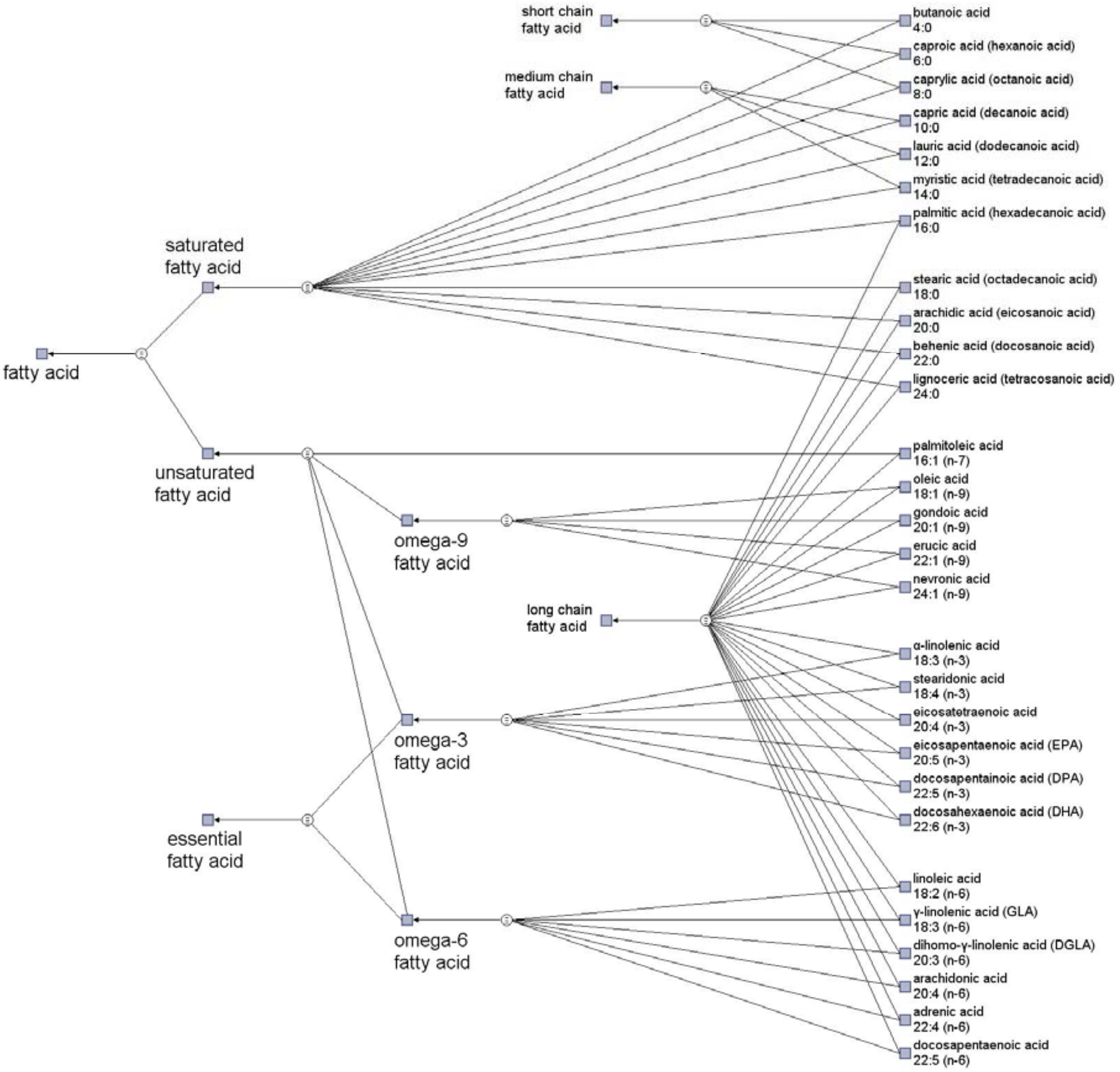

short chain fatty acid
medium chain fatty acid
saturated fatty acid
fatty acid
unsaturated fatty acid
omega-9 fatty acid
long chain fatty acid
omega-3 fatty acid
essential fatty acid
omega-6 fatty acid
butanoic acid 4:0
caproic acid (hexanoic acid) 6:0
caprylic acid (octanoic acid) 8:0
capric acid (decanoic acid) 10:0
lauric acid (dodecanoic acid) 12:0
myristic acid (tetradecanoic acid) 14:0
palmitic acid (hexadecanoic acid) 16:0
stearic acid (octadecanoic acid) 18:0
arachidic acid (eicosanoic acid) 20:0
behenic acid (docosanoic acid) 22:0
lignoceric acid (tetracosanoic acid) 24:0
palmitoleic acid 16:1 (n-7)
oleic acid 18:1 (n-9)
gondoic acid 20:1 (n-9)
erucic acid 22:1 (n-9)
nevronic acid 24:1 (n-9)
α-linolenic acid 18:3 (n-3)
stearidonic acid 18:4 (n-3)
eicosatetraenoic acid 20:4 (n-3)
eicosapentaenoic acid (EPA) 20:5 (n-3)
docosapentainoic acid (DPA) 22:5 (n-3)
docosahexaenoic acid (DHA) 22:6 (n-3)
linoleic acid 18:2 (n-6)
γ-linolenic acid (GLA) 18:3 (n-6)
dihomo-γ-linolenic acid (DGLA) 20:3 (n-6)
arachidonic acid 20:4 (n-6)
adrenic acid 22:4 (n-6)
docosapentaenoic acid 22:5 (n-6)

Figure 2.1. Fatty acid generic chart.

2.3. The UN grammar rules

The UN has been developed as both a human-readable and a machine-readable graphical language so UN diagrams could be easily transformed into a model. For that all the entities must be linked properly. Tables 2.3 and 2.4 describe the connectivity rules in the UN.

Table 2.5 summarizes the rules of containment for the ‘multistate entity’ and ‘compartment’ entities.

Labels on a diagram have to be presented in such way so it would be clear what object each label describes. In other words, a label should be positioned closer to the corresponding entity to exclude the possibility of any visual ambiguity.

Similarly, entities consumed in a process have to be shown in such way so they would be visually clearly distinguished from entities produced as a result of that process.

Table 2.3. Connectivity rules. Part 1.

Connector \ Entity	Unspecified entity	Compound	Homodimer	Heterodimer	Complex	Gene	RNA
Process input	Source	Source	Source	Source	Source	Source	Source
Process output	Target	Target	Target	Target	Target	Target	Target
Activation	Source	Source	Source	Source	Source	Source	Source
Inhibition	Source	Source	Source	Source	Source	Source	Source
Logical gate input	Source	Source	Source	Source	Source	Source	Source
Logical gate output	NA	NA	NA	NA	NA	NA	NA
Identity gate input	Source	Source	Source	Source	Source	Source	Source
Identity gate output	NA	NA	NA	NA	NA	NA	NA
Functional gate input	Source	Source	Source	Source	Source	Source	Source
Functional gate output	NA	NA	NA	NA	NA	NA	NA

Table 2.4. Connectivity rules. Part 2.

Connector \ Entity	Unknown process	Process	Translocation	Gene expression	AND gate	OR gate	NOT gate	Identity gate	Functional gate
Process input	Target	Target	Target	Target	NA	NA	NA	NA	NA
Process output	Source	Source	Source	Source	NA	NA	NA	NA	NA
Activation	Target	Target	Target	Target	NA	NA	NA	NA	NA
Inhibition	Target	Target	Target	Target	NA	NA	NA	NA	NA
Logical gate input	NA	NA	NA	NA	Target	Target	Target	NA	NA
Logical gate output	Target	Target	Target	Target	Source	Source	Source	NA	NA
Identity gate input	NA	NA	NA	NA	NA	NA	NA	Target	NA
Identity gate output	NA	NA	NA	NA	NA	NA	NA	Source	NA
Functional gate input	NA	NA	NA	NA	NA	NA	NA	NA	Target
Functional gate output	Target	Target	Target	Target	NA	NA	NA	NA	Source

Table 2.5. Containers definition.

Entity \ Container	Multistate entity	Compartment
Unknown entity	Possible	Possible
Compound	Possible	Possible
Monomer	Possible	Possible
Homodimer	Possible	Possible
Heterodimer	Possible	Possible
Complex	Possible	Possible
Multistate entity	NA	Possible
Compartment	NA	NA
Process	Possible	Possible
Unknown process	Possible	Possible
Translocation	NA	Possible
Omitted process	Possible	Possible
Degradation	NA	Possible
AND gate	NA	Possible
OR gate	NA	Possible
NOT gate	NA	Possible
Functional gate	NA	Possible
Identity gate	NA	Possible

2.4. Comparison between the UN and the SBGN

The entities of the UN and the SBGN Process Diagram notation (SBGN PD) (SBGN Process Diagram Level 1 specification, http://sbgn.org/Documents/Specifications) are compared in detail in tables 2.6 - 2.10. Table 2.11 provides four examples of the notations.

The main disadvantage of the SBGN in comparison with the UN is that it has been developed mainly for representing signal transduction networks and is less suitable for metabolic pathway visualisation. Two issues have to be mentioned here.

First, the most useful metabolic pathway representations include EC numbers (KEGG database, for example). It is the most compact way to give information about the enzymes involved. In the SBGN the usual way to show EC is not available. In fact, it is possible to display a 'protein'/'complex' with a corresponding 'unit of information', but that would mean one is forced to annotate additionally all related enzymes. Another option would be to use an 'unspecified entity' with a corresponding 'unit of information'. In both cases the resulting diagram would be very cluttered.

The second issue is the way SBGN shows labels inside of a shape. In the case of metabolic pathway labels, they would be shown partly outside of a 'small molecule' glyph, and the SBGN allows doing that as soon as a label is positioned in the center of a glyph. As compounds tend to have long names, sometimes comprising few lines, it makes difficult using horizontal links. In most cases instead of having straight line one needs to use several bend points for a link to avoid overlapping with a compound name. An alternative way would be to use larger shapes and smaller fonts so that a compound name would not appear outside of the 'compound' shape. In that case there are two options arise. The first option is to use different size for each 'compound' symbol fitting it in to a compound name. However that would not make a diagram look consistent. The second option is to use the same size for all 'compound' shapes on a map, but that requires more space as the size of all 'compound' glyphs on a map depends on the longest compound name.

However, the UN overcomes these difficulties using EC as a functional gate thus allowing positioning labels outside of the 'compound' symbol.

Other significant differences between the two notations are listed below.

1. The UN labels entity pool nodes by placing a label next to the symbol, while the SBGN places labels inside the shape. As a result, maps in UN notation are more compact. Although it could be suggested that a label area plus a shape area should be at least as big as a single shape with a text inside, the actual examples prove otherwise (Table 2.8, Figures 2.5 and 2.6). That is due to the fact that SBGN requires adding new shapes to display a protein state and that affects the size of a 'macromolecule' glyph significantly.

2. 'Complex' entities are shown differently. The SBGN PD complex is shown as a container with particular entities inside it, whilst the UN uses a single symbol and lists entities involved near the symbol (Table 2.11, A). Positioning of elements in a complex container in the SBGN suggests certain structure of a complex that can not be fully described in this notation. Another difficulty could pose the cases where 'complex-in-complex' visualisation is used because of several possible expressions available (Figure 2.2 A and B). In contrast, the UN shows the complex composition as a list of elements involved (Figure 2.2 C).

3. While SBGN uses transition, association and dissociation as separate elements to show processes, UN uses a single symbol for all of them. At the same time, UN contains elements representing transport and gene transduction. The gene transduction (Table 2.8, D) is shown in UN by special symbols to avoid ambiguity, because it might be confused with a black box or child pathway. When appropriate visualisation is developed and information is available, translating information from gene to protein might be shown as a child pathway.

4. UN introduces the multistate entity symbol (Table 2.11, B, C), a container that puts different states of the same protein/complex in one place. This serves several purposes. Firstly, it saves space by listing a complex composition only once as a name of a container. Secondly, it facilitates the placement of all states of an entity on a map.

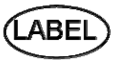
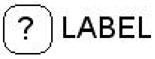
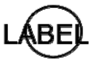
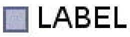
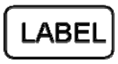
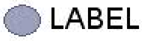
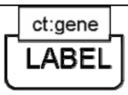
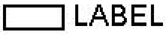
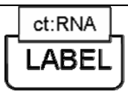
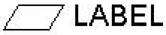
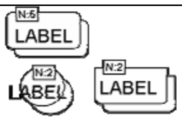
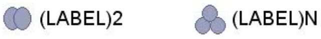
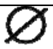
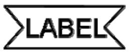
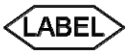
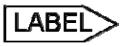
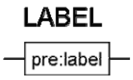
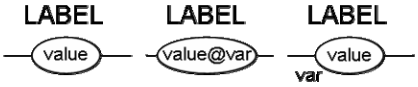
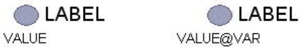
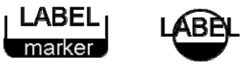
Thirdly, it reduces the number of state combinations that have to be displayed, as it visually encapsulates all internal state transitions.

5. UN introduces a new element - the ‘identity gate’ - to link generic and specific compounds. There are many cases when it is necessary to show generic and specific entities on the same map. For example, the Toll-Like Receptor(TLR) map (Oda and Kitano, 2006) shows connections between several particular ligands and generic compounds such as ‘TLR4 ligand’.

6. Several connecting arcs that SBGN uses are not available in the UN, namely catalysis, modulation and trigger (absolute stimulation). Catalysis is represented in UN by functional gates (Figure 2.1, Q). Such elements as modulation and absolute stimulation are being considered for the next version of the UN.

7. The UN notation has been designed to visualise both the signalling and metabolic pathways. ‘Compound’ symbol in UN is similar to the KEGG visual solution in metabolic pathway representation and seems to be more appropriate for showing entities that often have long names. Showing compound names inside an ellipse in SBGN PD takes more space. The UN proposes to keep the most efficient representation of a metabolic network by displaying labels outside of a shape, and also by using the ‘functional gates’ which appears as EC/TC on metabolic diagrams.

Table 2.6. Comparison between the UN and the SBGN. Entity pool nodes.

Entity	SBGN	UN
A. Unspecified entity		
B. Simple chemical		
C. Macromolecule		
D. Gene		
E. RNA		
F. Multimer		
G. Source/sink		
H. Perturbation		Not available. Free text comments are used instead.
I. Observable		Not available. Free text comments are used instead.
J. Tag		No available. Direct link to another diagram is used instead
K. Unit of information		Not available. The entity type is shown by a symbol; complex composition is shown as a text.
L. State variable	 LABEL value LABEL value@var LABEL value var	
M. Clone marker		Not available. Even if it is necessary to show the same entity

		several times on the same diagram no markers are used. All object have identifiers in EPE and similar entities are easily recognised.

Table 2.7. Comparison between the UN and the SBGN. Container nodes.

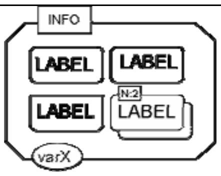
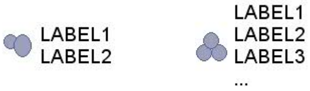
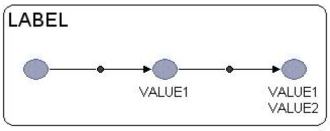
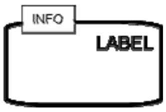
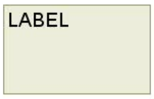
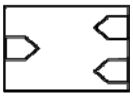
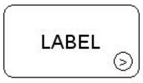
Entity	SBGN	UN
A. Complex		
B. Multistate entity	Not available. Entities with the same name are not grouped in any specific way	
C. Compartment		
D. Submap		

Table 2.8. Comparison between the UN and the SBGN. Process nodes.

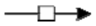
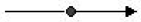
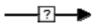
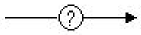
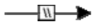
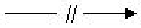
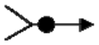
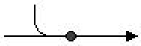
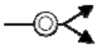
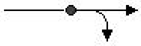
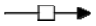
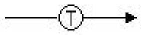
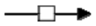
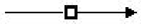
Entity	SBGN	UN
A. Transition		
B. Uncertain process		
C. Omitted process		
D. Association		
E. Dissociation		
F. Translocation		
G. Gene expression		

Table 2.9. Comparison between the UN and the SBGN. Connecting arcs.

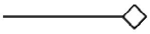
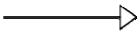
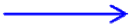
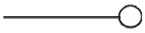
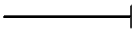
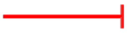
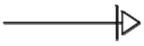
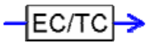
Entity	SBGN	UN
A. Modulation		Not available. Introducing in the next version of the UN is being discussed.
B. Stimulation		
C. Catalysis (particular case of stimulation)		Not available. Functional gates are used in case of catalysis
D. Inhibition		
E. Trigger (absolute stimulation)		Not available. Introducing in the next version of the UN is being discussed.
F. Functional gate	Not available. It is possible to show EC only as a name of a macromolecule (see chapter 9 for details). Introducing in the next version of the SBGN PD language is being discussed by the SBGN community	

Table 2.10. Comparison between the UN an the SBGN. Logical operators.

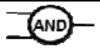
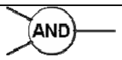
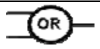
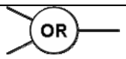
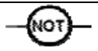
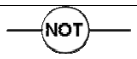
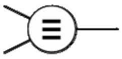
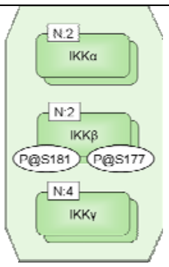
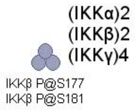
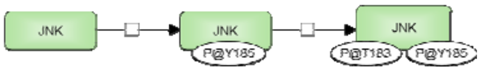
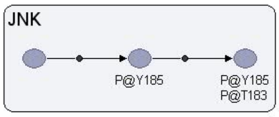
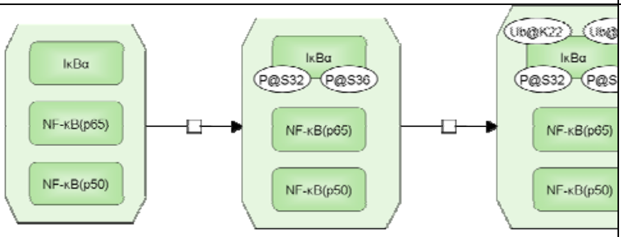
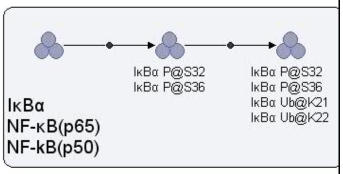
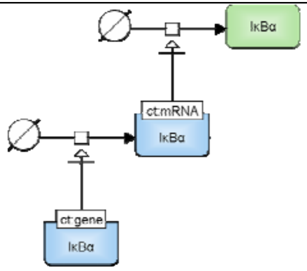
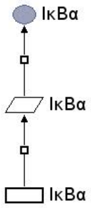
Logical operator	SBGN	UN
A. AND gate		
B. OR gate		
C. NOT gate		
D. Identity gate	Not available. Introducing in the next version of the SBGN PD language is being discussed by the SBGN community	

Table 2.11. Comparison between the UN and the SBGN. Examples.

Example	SBGN	UN
A. Complex	 N:2 IKKα N:2 IKKβ P@S181 P@S177 N:4 IKKγ	
B. Different states of the same protein	 	
C. Different states of the same complex	 IκBα NF-κB(p65) NF-κB(p50) IκBα P@S32 P@S36 NF-κB(p65) NF-κB(p50) Ub@K22 Ub@ IκBα P@S32 P@S NF-κB(p65) NF-κB(p50)	 IκBα P@S32 IκBα P@S36 IκBα P@S32 IκBα P@S36 IκBα Ub@K21 IκBα Ub@K22 IκBα NF-κB(p65) NF-kB(p50)
D. Gene expression	 	

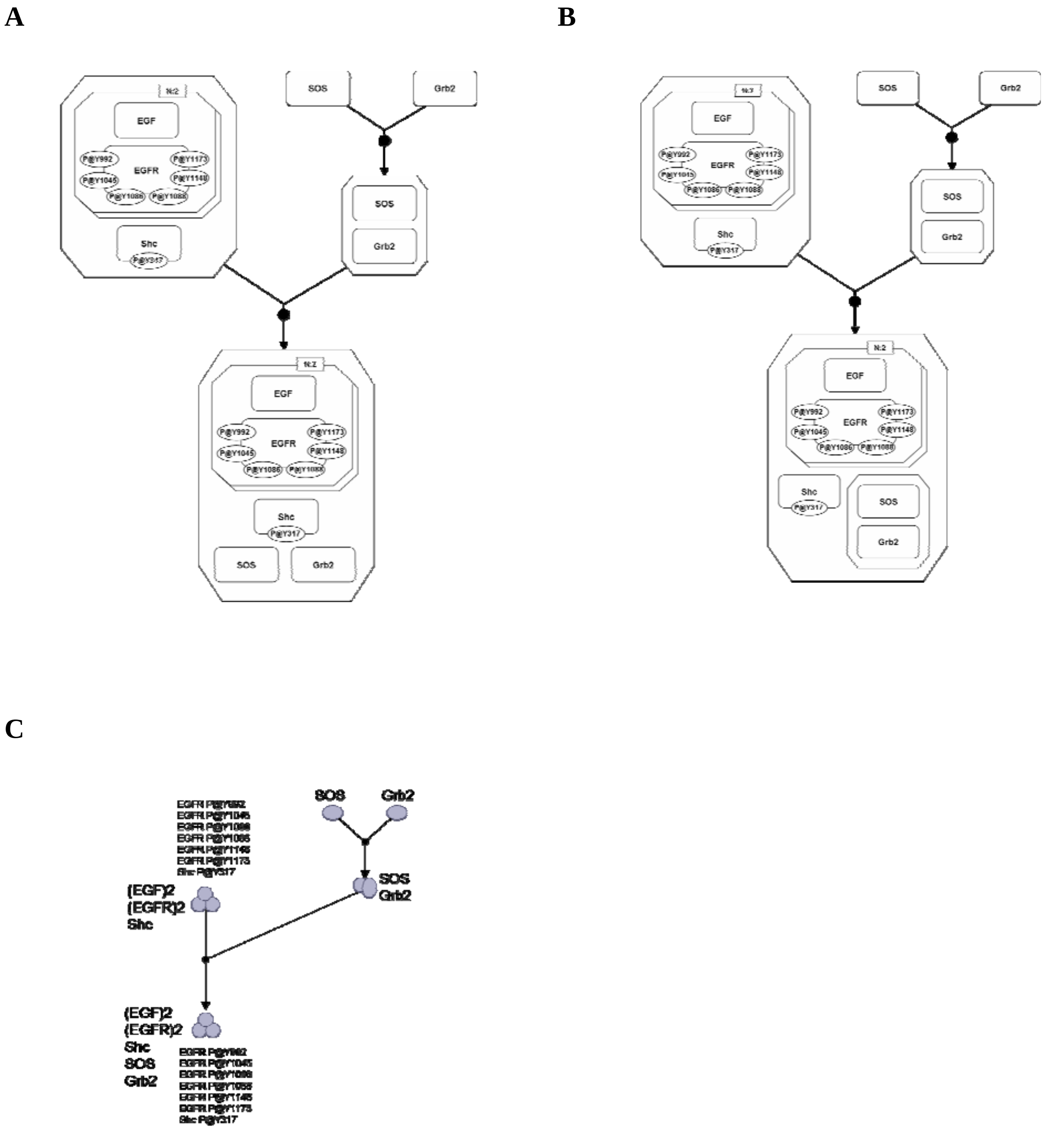


Figure 2.2. Complex-in-complex case in the SBGN. Two possibilities to describe the resulting complex are available: A. SOS and Grb2 are shown as single proteins in the complex; B. SOS and Grb2 are shown as a complex-in-complex; C. The same events in the UN.

2.5. Comparison between the UN and the process diagram notation

In order to demonstrate the advantages of using the UN and compare it to the process diagram notation, 'a comprehensive map of Toll-like receptor signalling' (Oda and Kitano, 2006) (Figure 2.3) has been transformed into a UN notation map (Figure 2.4). Without going into comparing the two notations symbol by symbol as it is done with SBGN PD notation, let us identify the most noticeable differences using examples from the Toll-receptor map in the process diagram notations and in the UN.

First of all, taking into account the smallest font size, the UN has been proved to be more compact than the process diagram notation. The Toll-like receptor signalling map (Oda and Kitano, 2006) converted into the current version of the Unified Notation is at least twice as compact as the original map shown in the process diagram notation. This is mainly due to using the 'label near a shape' instead of the 'label in a shape' visualisation (Figure 2.5). Also using the 'multistate entity' helps to minimize space for several states of the same protein complex (Figure 2.6). Figure 2.7 shows another example of a complex representation. Figure 2.8 demonstrates the use of the 'identity gate' symbol in the UN. Figure 2.9 includes both versions of the same part of the network so that readability and compactness of the two notations could be compared.

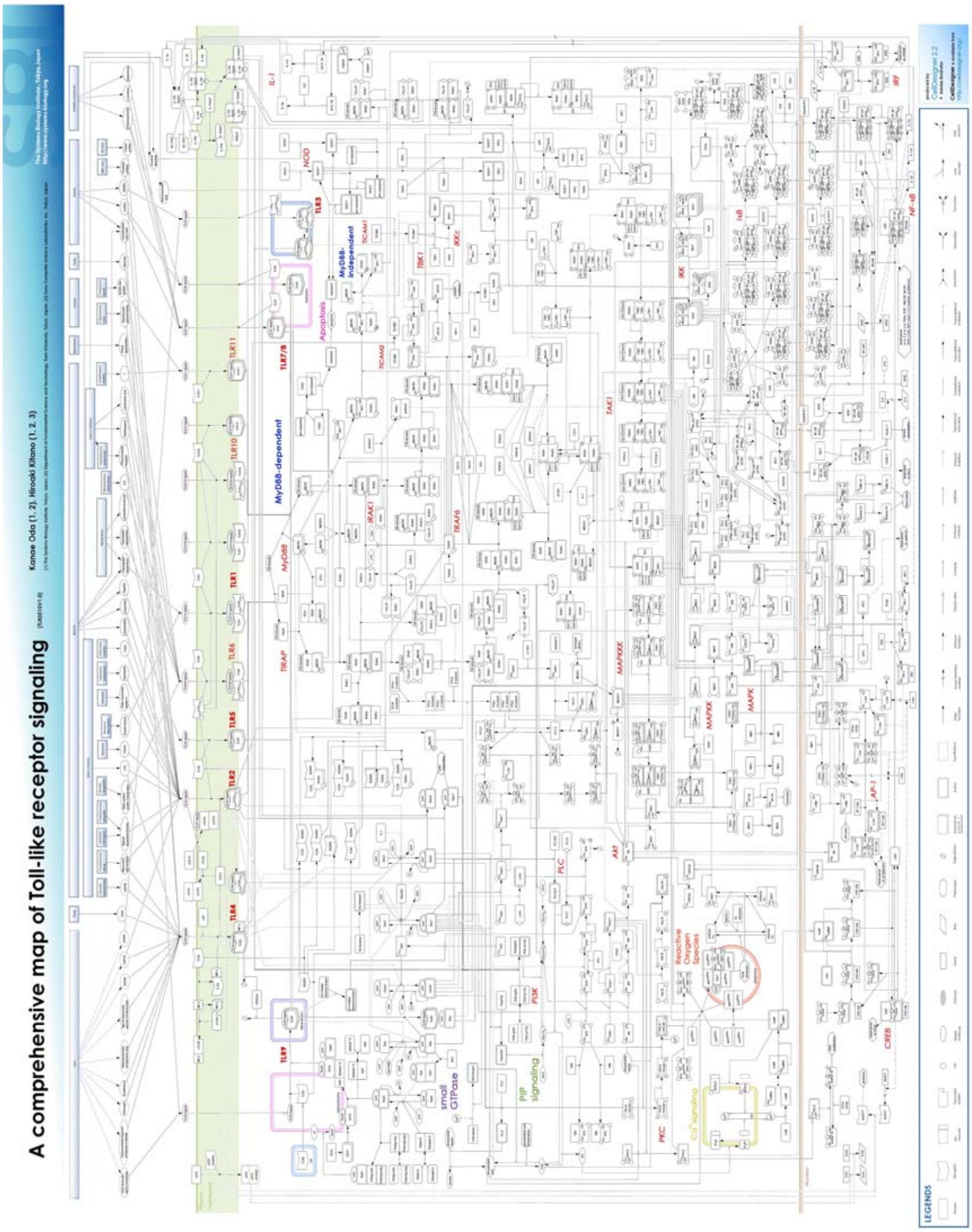


Figure 2.3. TLR signalling in the process diagram notation (Oda and Kitano, 2006).

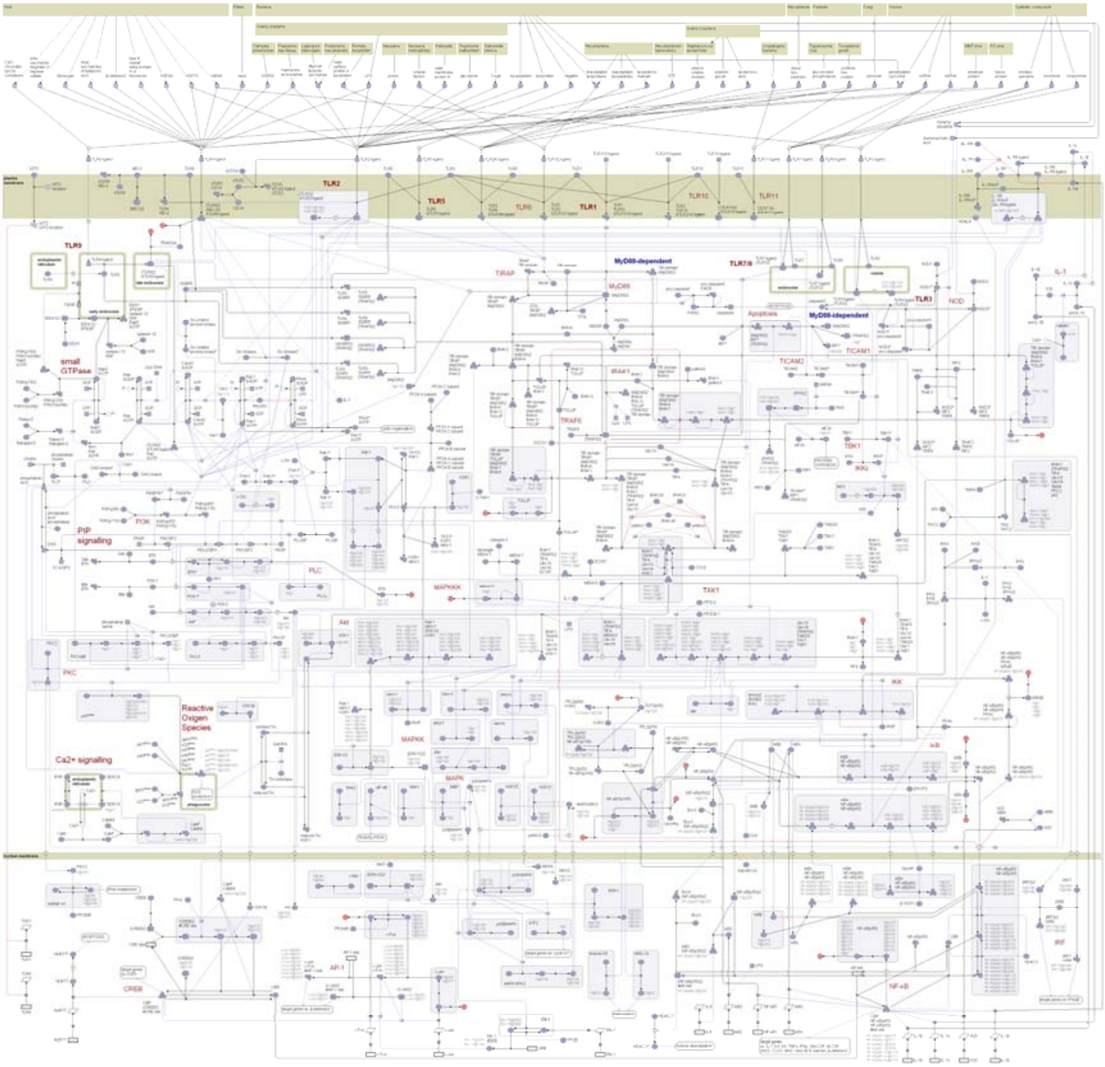


Figure 2.4. TLR signalling map (Figure 2.3) represented in the Unified Notation. Please use the CD provided to see a high resolution version of the image.

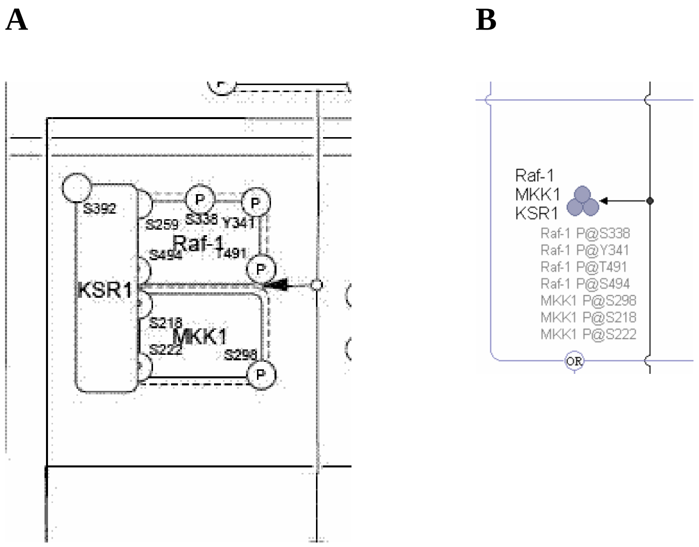


Figure 2.5. ‘Name in the shape’ to compare to ‘name near the shape’. A. Raf1-MKK1-KSR1 complex in the process diagram notation. B. Raf1-MKK1-KSR1 complex in the UN. Smallest font size on both diagrams is approximately the same.

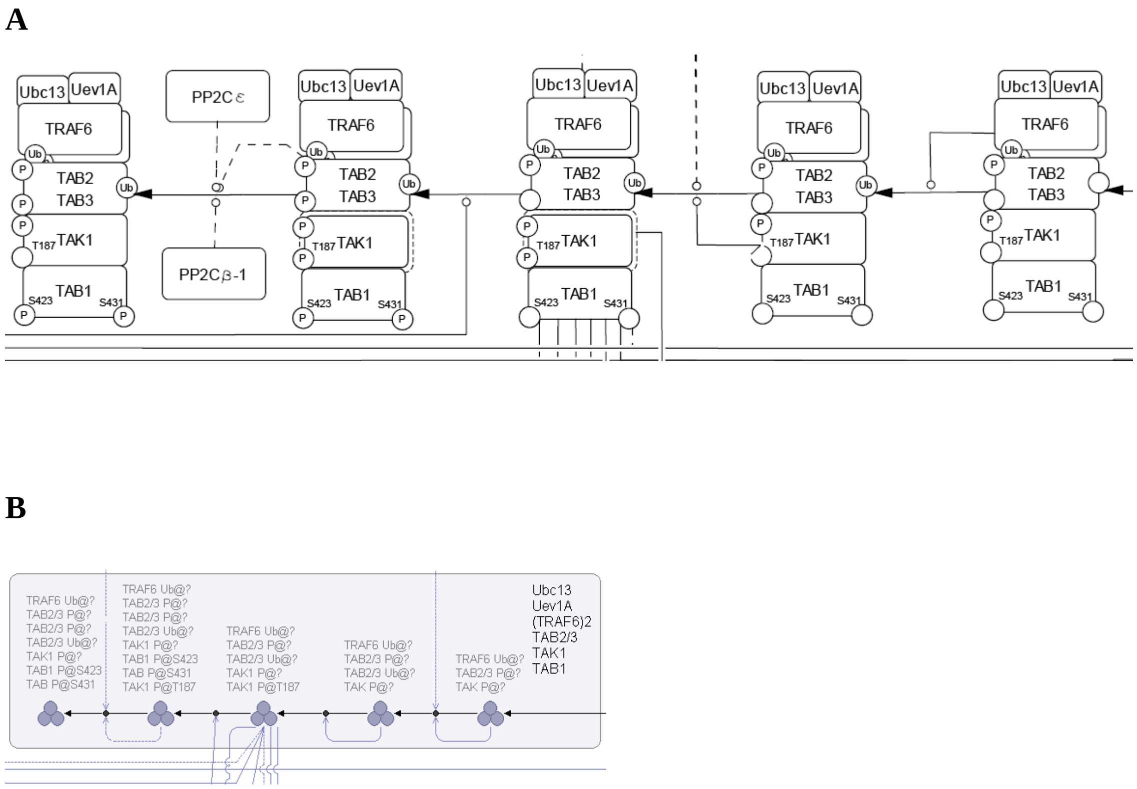


Figure 2.6. Comparison of the same events shown in the UN and in the process diagram notation. A. A part of the toll-like receptor network diagram (Oda and Kitano, 2006). B. The same events shown in the UN. Names of the proteins in the complex are shown only once. The state in each case is shown near the symbol. Smallest font size on both diagrams is approximately the same.

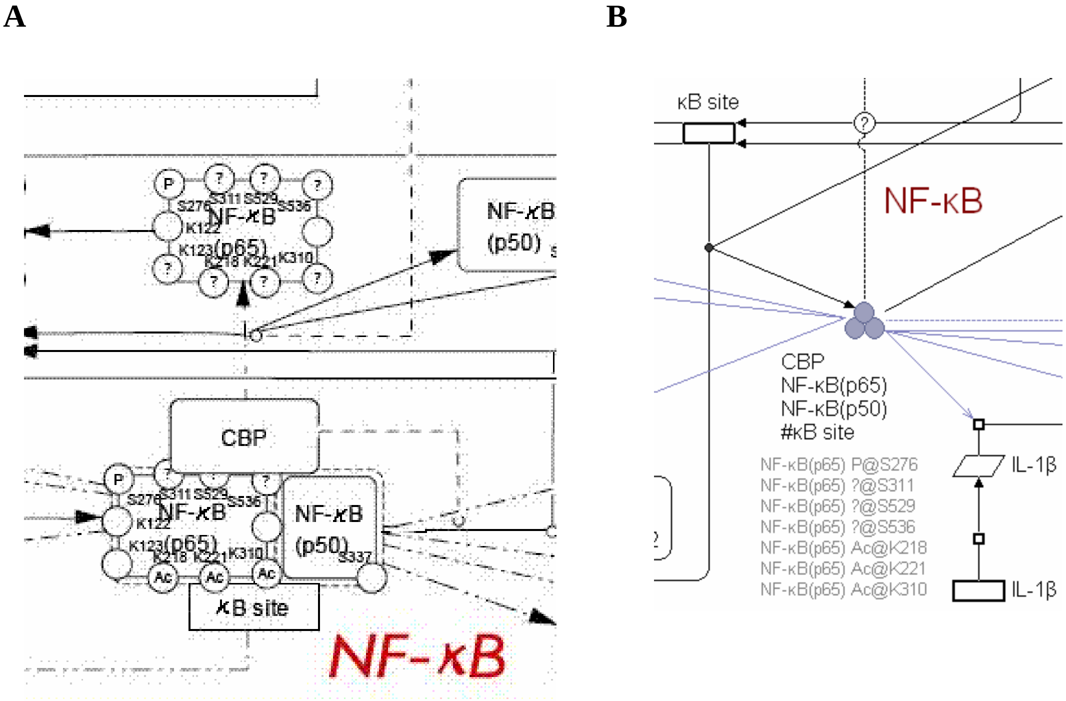


Figure 2.7. A complex representation. A. In the process diagram notation. B. In the UN. Smallest font size on both diagrams is approximately the same.

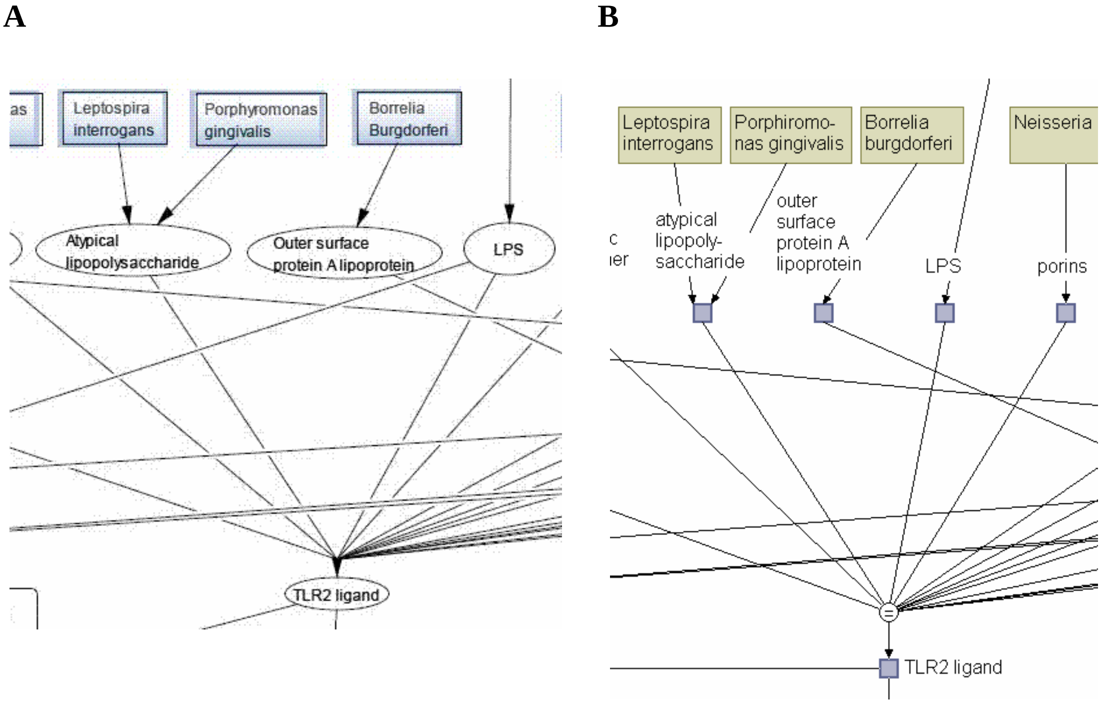


Figure 2.8. Using identity gates in the UN. The same detail from Toll-like receptor signalling map is shown in the process diagram notation (A) and in UN (B).

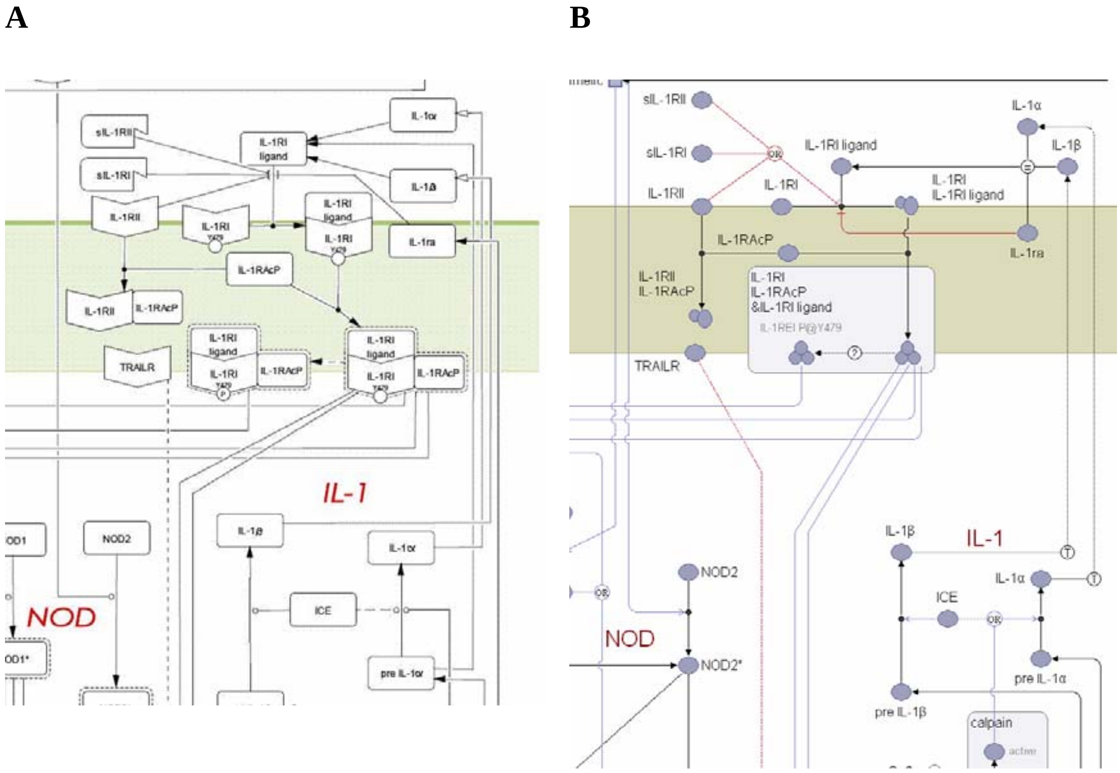


Figure 2.9. Readability and compactness. The same detail from Toll-like receptor signalling map is shown in the process diagram notation (A) and in UN (B). The same number of objects can be seen on both fragments. The font size is larger on the diagram in UN.

2.6. The Edinburgh Pathway Editor (EPE) and the UN implementation in EPE.

This section briefly describes the Edinburgh Pathway Editor features relevant to the research, explains why the editor was chosen, and provides context definition (EPE concept of ‘context’ is introduced below) for the Unified Notation.

The Edinburgh Pathway Editor (EPE) (Sorokin, 2006) is a Java-based and therefore platform-independent software. It is distributed under the Eclipse open-source application platform license (Sorokin, 2006).

The editor uses a set of basic objects: shapes, processes, links and labels. These objects illustrate concepts of a biological network. Shapes represent biological objects such as compounds, macromolecules, complexes and subsystems. Processes show a sequence of events and allow a user to describe reaction, state transition etc. Links are reserved to show any relationships between two objects: shape-shape, shape-process and process-process. The levels represent textual information and hyperlinks to other maps or external resources (Sorokin, 2006).

The information about maps is stored in the editor in a relational format. The Apache Derby database is supported by the tool and is used as local data storage (Sorokin, 2006).

The most important EPE feature for the current research is a concept of ‘context’. A ‘context’ in EPE is a user-defined set of objects, their properties and default values that form a drawing palette for using a particular graphical language (Sorokin, 2006).

The EPE allows creating hyperlinks and provides environment for developing hierarchical representation (Sorokin, 2006).

Other pathway editors are currently available: CellDesigner (CellDesigner: A Modeling Tool of Biochemical Networks, http://celldesigner.org/), TERANODE (Design Automation for Life Sciences, http://www.teranode.com/), Bio Sketch Pad (Bio Sketch Pad, http://www.cis.upenn.edu/biocomp/new_html/biosketch.php3), Systems Biology Workbench JDesigner (SBW, http://www.sys-bio.org/), BioUML (BioUML Framework for Systems Biology, http://www.biouml.org/), BioTapestry (BioTapestry, http://labs.systemsbiology.net/bolouri/software/BioTapestry/), Pathway Builder (Pathway Builder Tool, http://www.proteinlounge.com/pathwaybuilder.asp), NetBuilder (NetBuilder Home, http://strc.herts.ac.uk/bio/maria/NetBuilder/), PathwayLab (PathwayLab, http://innetics.com/pathwaylab_overview.htm), VitaPad (Holford, 2005), PATIKA (Demir, 2002) and PathwayStudio (PathwayStudio, http://www.ariadnegenomics.com/products/pscentral/). Detailed comparison of the listed above applications and the EPE is provided by Sorokin and coauthors (Sorokin, 2006). The most distinctive advantage of the EPE is the combination of the following features: ability to represent networks of different types (metabolic, signalling, gene expression); possibility to add annotations; possibility to change visual and annotation properties; availability of import end export in SBML format; possibility to represent hierarchy of diagrams (Sorokin, 2006).

The main reasons why the Edinburgh Pathway Editor was chosen rather than one of the listed above pathway editors:

1) the EPE is highly flexibile in allowing to create, edit and annotate objects, which is essential for developing a new graphical notation;

2) the editor supports definition of syntax and semantics of user-defined notation/context; and, finally;

3) the possibility to work in close collaboration with the software developers' team. The latter was possible because the group of the EPE software developers was part of the same Computational Systems Biology group I have been doing my research in and therefore new functionalities could be introduced when required for this research.

The Unified Notation is defined in the Edinburgh Pathway Editor in the context 'Biological'. The list of the context objects (shapes, processes, labels and connectors) defined is provided bellow.

2.6.1. MAP

Map name: <TEXT>

Grid: On

Grid Size. Height: 32.

Grid Size. Width: 32

Corresponding EMP map ID(s): <SIMPLE DATA>

Corresponding KEGG map ID(s): <SIMPLE DATA>

Organism: <TEXT>

Tissue/cell type: <TEXT>

Last updated: <DD.MM.YYYY>

2.6.2. SHAPES

2.6.2.1. Unknown entity

Shape: Rounded rectangle

Fit to text: Off

Size: Height: 30. Width: 30

Background colour: RBG 255, 255, 255

Foreground colour: RBG 92, 99, 143

Line style: solid

Line width: 1

Snap options. Snap location: On. Snap size: On. Snap Type: Center

Name: <TEXT>

Synonyms: < TEXT COLLECTIONS>

UniProt ID: <TEXT>

Fixed: <Yes / No>

Text: ?

Hyperlink: <TEXT>

2.6.2.2. Compound

Entity ID: <SIMPLE DATA>

KEGG compound ID: <SIMPLE DATA>

EMP compound ID: <SIMPLE DATA>

Shape: Rectangle

Fit to text: Off

Size: Height: 16. Width: 16

Background colour: RBG 157, 159, 189

Foreground colour: RBG 92, 99, 143

Line style: solid

Line width: 2

Name: <TEXT>

Snap options. Snap location: On. Snap size: On. Snap Type: Center

Synonyms: < TEXT COLLECTIONS>

Fixed: <Yes / No>

Hyperlink: <TEXT>

2.6.2.3. Protein

UniProt ID: <TEXT>

Shape: Ellipse

Fit to text: Off

Size: Height: 20. Width: 24

Background colour: RBG 157, 159, 189

Foreground colour: RBG 92, 99, 143

Line style: solid

Line width: 1

Snap options. Snap location: On. Snap size: On. Snap Type: Center

Name: <TEXT>

Synonyms: < TEXT COLLECTIONS>

Protein state: <TEXT COLLECTION>

State: $protein state@\n$

Fixed: <Yes / No>

Hyperlink: <TEXT>

2.6.2.3. Homodimer

Shape: Homo dimer

Fit to text: Off

Size: Height: 25. Width: 35

Background colour: RBG 157, 159, 189

Foreground colour: RBG 92, 99, 143

Snap options. Snap location: On. Snap size: On. Snap Type: Center

Line style: solid

Line width: 1

Protein name: <TEXT >

Name: ($protein name$)2

Synonyms: <EXT COLLECTIONS>

Complex state: <TEXT COLLECTION>

State: $protein state@\n$

Fixed: <Yes / No>

Hyperlink: <TEXT>

2.6.2.4. Heterodimer

Shape: Hetero dimer

Fit to text: Off

Size: Height: 25. Width: 35

Background colour: RBG 157, 159, 189

Foreground colour: RBG 92, 99, 143

Snap options. Snap location: On. Snap size: On. Snap Type: Center

Line style: solid

Line width: 1

Complex composition: <TEXT COLLECTION >

Name: $complex composition@\n$

Synonyms: <TEXT COLLECTIONS>

Complex state: <TEXT COLLECTION>

State: $protein state@\n$

Fixed: <Yes / No>

Hyperlink: <TEXT>

2.6.2.5. COMPLEX

Shape: Complex

Fit to text: Off

Size: Height: 30. Width: 31

Background colour: RBG 157, 159, 189

Foreground colour: RBG 92, 99, 143

Snap options. Snap location: On. Snap size: On. Snap Type: Center

Line style: solid

Line width: 1

Complex composition: <TEXT COLLECTION >

Name: $complex composition@\n$

Synonyms: <TEXT COLLECTIONS>

Complex state: <TEXT COLLECTION>

State: $protein state@\n$

Fixed: <Yes / No>

Hyperlink: <TEXT>

2.6.2.6. Gene

Gene ID: <SIMPLE DATA>

Shape: Rectangle

Fit to text: Off

Size: Height: 16. Width: 40

Background colour: RBG 255, 255, 255

Foreground colour: RBG 0, 0, 0

Snap options. Snap location: On. Snap size: On. Snap Type: Center

Line style: solid

Line width: 2

Name: <TEXT>

Fixed: <Yes / No>

Hyperlink: <TEXT>

2.6.2.7. RNA

Shape: LParallelogramm

Fit to text: Off

Size: Height: 16. Width: 40

Background colour: RBG 255, 255, 255

Foreground colour: RBG 0, 0, 0

Snap options. Snap location: On. Snap size: On. Snap Type: Center

Line style: solid

Line width: 1

Name: <TEXT>

Fixed: <Yes / No>

Hyperlink: <TEXT>

2.6.2.8 Multistate entity

Shape: Rounded rectangle

Fit to text: Off

Size: Height: 80. Width: 80

Background colour: RBG 242, 242, 247

Foreground colour: RBG 128, 128, 128

Snap options. Snap location: On. Snap size: On. Snap Type: Edge

Line style: solid

Line width: 1

Name: <TEXT>

Fixed: <Yes / No>

Hyperlink: <TEXT>

Text: $name$

2.6.2.9. Degradation

Shape: Degradation

Fit to text: Off

Size: Height: 21. Width: 21

Background colour: RBG 255, 255, 255

Foreground colour: RBG 0, 0, 0

Snap options. Snap location: On. Snap size: On. Snap Type: Center

Line style: solid

Line width: 1

Name: <TEXT>

Fixed: <Yes / No>

Hyperlink: <TEXT>

2.6.2.10. Compartment

Shape: Rectangle

Fit to text: Off

Size: Height: 64. Width: 64

Background colour: RBG 237, 237, 220

Foreground colour: RBG 147, 146, 98

Snap options. Snap location: On. Snap size: On. Snap Type: Edge

Line style: solid

Line width: 1

Name: <TEXT>

Fixed: <Yes / No>

Hyperlink: <TEXT>

2.6.2.11. Text

Shape: Rectangle

Fit to text: Off

Size: Height: 32. Width: 64

Background colour: RBG 255, 255, 255

Foreground colour: RBG 255, 255, 255

Snap options. Snap location: On. Snap size: On. Snap Type: Edge

Line style: solid

Line width: 1

Name: <TEXT>

Fixed: <Yes / No>

Hyperlink: <TEXT>

Text: $name$

Font: 14

2.6.2.12. Link to another map

Shape: Rounded rectangle

Fit to text: Off

Size: Height: 64. Width: 128

Background colour: RBG 255, 255, 255

Foreground colour: RBG 0, 0, 0

Snap options. Snap location: On. Snap size: On. Snap Type: Edge

Line style: solid

Line width: 1

Name: <TEXT>

Fixed: <Yes / No>

Hyperlink: <TEXT>

Text: $name$

Font: 14

2.6.2.13. Map name

Shape: Rectangle

Fit to text: On

Size: Height: 72. Width: 720

Background colour: RBG 255, 255, 255

Foreground colour: RBG 255, 255, 255

Snap options. Snap location: On. Snap size: On. Snap Type: Edge

Line style: solid

Line width: 1

Name: <TEXT>

Fixed: <Yes / No>

Hyperlink: <TEXT>

Text: $name$

Font: 36

Text colour: RGB 128,128,128

2.6.3. LABELS

2.6.3.1. Name label

Shape: Rectangle

Fit to text: On

Background colour: RBG 255, 255, 255

Foreground colour: RBG 255, 255, 255

Line style: solid

Line width: 0

Label position: Top-right

Text: $leader.name$

Font size: 14

2.6.3.2. State label

Shape: Rectangle

Fit to text: On

Background colour: RBG 255, 255, 255

Foreground colour: RBG 255, 255, 255

Line style: solid

Line width: 0

Label position: Top-right

Text: $leader.state$

Font size: 10

2.6.4. PROCESSES

2.6.4.1. Process

Shape: Ellipse

Fit to text: Off

Size: Height: 7. Width: 7

Background colour: RBG 0, 0, 0

Foreground colour: RBG 0, 0, 0

Snap options. Snap location: On. Snap size: On. Snap Type: Center

Line style: solid

Line width: 1

Fixed: <Yes / No>

Hyperlink: <TEXT>

Reaction ID: (SIMPE COLLECTION)

KEGG reaction ID: (SIMPE COLLECTION)

Reaction equation: (TEXT COLLECTION)

Reversible: (SIMPE COLLECTION)

EC: (SIMPE COLLECTION)

Subpathway: (TEXT COLLECTION)

2.6.4.2. Translocation

Shape: Ellipse

Fit to text: Off

Size: Height: 17. Width: 17

Background colour: RBG 255, 255, 255

Foreground colour: RBG 0, 0, 0

Snap options. Snap location: On. Snap size: On. Snap Type: Center

Line style: solid

Line width: 1

Fixed: <Yes / No>

Hyperlink: <TEXT>

Text: T

Reaction ID: (SIMPE COLLECTION)

KEGG reaction ID: (SIMPE COLLECTION)

Reaction equation: (TEXT COLLECTION)

Reversible: (SIMPE COLLECTION)

EC/TC: (SIMPE COLLECTION)

Subpathway: (TEXT COLLECTION)

2.6.4.3. Gene expression

Shape: Rectangle

Fit to text: Off

Size: Height: 11. Width: 11

Background colour: RBG 255, 255, 255

Foreground colour: RBG 0, 0, 0

Snap options. Snap location: On. Snap size: On. Snap Type: Center

Line style: solid

Line width: 2

Fixed: <Yes / No>

Hyperlink: <TEXT>

Reversible: (SIMPE COLLECTION)

Subpathway: (TEXT COLLECTION)

2.6.4.4. Omitted process

Shape: Ellipse

Fit to text: Off

Size: Height: 17. Width: 17

Background colour: RBG 255, 255, 255

Foreground colour: RBG 0, 0, 0

Snap options. Snap location: On. Snap size: On. Snap Type: Center

Line style: solid

Line width: 1

Fixed: <Yes / No>

Hyperlink: <TEXT>

Reversible: (SIMPE COLLECTION)

Subpathway: (TEXT COLLECTION)

Text: //

2.6.4.5. Unknown/proposed process

Shape: Ellipse

Fit to text: Off

Size: Height: 17. Width: 17

Background colour: RBG 255, 255, 255

Foreground colour: RBG 0, 0, 0

Snap options. Snap location: On. Snap size: On. Snap Type: Center

Line style: solid

Line width: 1

Fixed: <Yes / No>

Hyperlink: <TEXT>

Reversible: (SIMPE COLLECTION)

Subpathway: (TEXT COLLECTION)

Text: ?

2.6.4.6. OR gate

Shape: Ellipse

Fit to text: Off

Size: Height: 19. Width: 19

Background colour: RBG 255, 255, 255

Foreground colour: RBG 0, 0, 0

Snap options. Snap location: On. Snap size: On. Snap Type: Center

Line style: solid

Line width: 1

Fixed: <Yes / No>

Hyperlink: <TEXT>

Reversible: (SIMPE COLLECTION)

Subpathway: (TEXT COLLECTION)

Text: OR

2.6.4.7. AND gate

Shape: Ellipse

Fit to text: Off

Size: Height: 19. Width: 19

Background colour: RBG 255, 255, 255

Foreground colour: RBG 0, 0, 0

Snap options. Snap location: On. Snap size: On. Snap Type: Center

Line style: solid

Line width: 1

Fixed: <Yes / No>

Hyperlink: <TEXT>

Reversible: (SIMPE COLLECTION)

Subpathway: (TEXT COLLECTION)

Text: &

2.6.4.8. NOT gate

Shape: Ellipse

Fit to text: Off

Size: Height: 19. Width: 19

Background colour: RBG 255, 255, 255

Foreground colour: RBG 0, 0, 0

Snap options. Snap location: On. Snap size: On. Snap Type: Center

Line style: solid

Line width: 1

Fixed: <Yes / No>

Hyperlink: <TEXT>

Reversible: (SIMPE COLLECTION)

Subpathway: (TEXT COLLECTION)

Text: N

2.6.4.9. Identity gate

Shape: Ellipse

Fit to text: Off

Size: Height: 19. Width: 19

Background colour: RBG 255, 255, 255

Foreground colour: RBG 0, 0, 0

Snap options. Snap location: On. Snap size: On. Snap Type: Center

Line style: solid

Line width: 1

Fixed: <Yes / No>

Hyperlink: <TEXT>

Reversible: (SIMPE COLLECTION)

Subpathway: (TEXT COLLECTION)

Text: ≡

2.6.4.10. Functional gate

EC: <SIMPLE DATA>

Shape: Rectangle

Fit to text: Off

Size: Height: 24. Width: 300

Background colour: RBG 255, 255, 255

Foreground colour: RBG 0, 0, 0

Snap options. Snap location: On. Snap size: On. Snap Type: Center

Line style: solid

Line width: 1

Fixed: <Yes / No>

Hyperlink: <TEXT>

Reversible: (SIMPE COLLECTION)

Subpathway: (TEXT COLLECTION)

Text: EC

2.6.5. CONNECTORS

2.6.5.1. Process input

Foreground colour: RGB 0, 0, 0

Line style: solid

Line width: 1

Router: Manual

Source decorator: None

Target decorator: Triangle

2.6.5.2. Process output

Foreground colour: RGB 0, 0, 0

Line style: solid

Line width: 1

Router: Manual

Source decorator: None

Target decorator: None

2.6.5.3. Activation

Foreground colour: RGB 0, 0, 0

Line style: solid

Line width: 1

Router: Manual

Source decorator: None

Target decorator: Arrow

2.6.5.4. Inhibition

Foreground colour: RGB 0, 0, 0

Line style: solid

Line width: 1

Router: Manual

Source decorator: None

Target decorator: Bar

2.6.5.5. Gate input

Foreground colour: RGB 0, 0, 0

Line style: solid

Line width: 1

Router: Manual

Source decorator: None

Target decorator: None

Chapter 3

The Network Representation Framework

3.1. Introduction to the Network Representation Framework

The previous chapter described the system of symbols that allows one to show signalling and metabolic events unambiguously.

Nevertheless, the system does not resolve all the difficulties in biological network visualisation. Even if individual phenomena can be shown comprehensively, the problem of representing large biological networks still remains. Although some approaches look promising (Sections1.4 and 1.5), so far there has been no effective system introduced and exemplified for that task.

The Network Representation introduced in this thesis uses a system of rules that allows us representing genome-scale networks successfully, and review them on different levels of detail (Sorokin A, personal communication, 2006). This system is based on the use of comparatively small and detailed maps that are used as bricks for developing large networks. The system based on these maps has the following properties: 1) three categories of maps are used in order to view a biological system from different perspectives (detailed below); 2) the maps are organised hierarchically;

3) this hierarchic structure is used not only for visualisation of a network, but also provides the means for assembling individual detailed maps into larger virtual maps.

3.2. Organization of pathway diagrams

It could be quite challenging to put everything on to a single, all-encompassing map. The larger the diagram, the greater the difficulty in reading and making changes to it; moreover, it makes the task of updating the diagram very difficult.

One can compare creating pathway diagrams with organizing words in sentences while writing a book. There are many words, but their number is limited. Similarly, the number of cellular processes in biology is limited as well. By arranging these elementary processes in to a structured whole, we are creating pathway maps that have the following properties. First, a map has to be meaningful. Second, a size of map should not be too large or too small.

During our research on human metabolic pathways and the macrophage signalling network it has been estimated that the most convenient number of processes on a single diagram is about 40-80. That gives printer-friendly, readable, maintainable maps and provides a basis for the hierarchical organization of the diagrams. The characteristics of the different categories of diagrams are shown in Table 3.1. Several levels of diagrams can be shown if desired.

3.3. Three categories of diagrams

All the diagrams we use for visualisation and analysis of biological networks can be organised in to three categories (Table 3.1).

PHENOMENOLOGICAL diagrams (or outlines) express our knowledge about a system at different levels of abstraction, or underline some particular aspect.

BIOLOGICAL diagrams show all known facts unambiguously. A well-defined notation is very important here.

KINETIC diagrams are created on the basis of biological diagrams and unknown facts are replaced by assumptions so the full kinetic mechanism could be represented. The kinetic model diagrams can be converted into SBML, ODE or other dynamic models for further simulation, analysis and hypothesis generation.

Phenomenological category of diagrams: This category includes schematic representations of biological phenomena, for example an outline view that helps understanding of a network structure. Here knowledge representation requires additional textual description.

In the Network Representation a phenomenological diagram is made semantically clear by being linked to appropriate biological diagrams. In this sense the key element on a phenomenological map is linked to sub-level diagrams. A link could be a special symbol (such as the 'link glyph, table 1X) or underlined text.

Our example of metabolic network representation (Chapter 4) shows that it is convenient to develop several sub-levels inside the phenomenological layer. This could be done to unite related diagrams into a single network; furthermore, by using an outline of outlines it could be formed into an even larger network. This way phenomenological diagrams serve several purposes. Firstly, they provide an easily readable and navigable representation of biological phenomena. Secondly, phenomenological diagrams could be assembled into a single system or a virtual network environment. Finally, a separate phenomenological diagram can be used to reveal a particular aspect of a biological system, or to describe a large kinetic model without all unnecessary details.

Biological category is an unambiguous representation of an experiment-based knowledge. Corresponding references to the papers with experimental evidence must be provided. All related facts have to be shown in a well-defined notation in a format that is both human and machine readable. The United Notation is used for that in the Network Representation, but other formal notations could be used as well, for example the process diagram notation or the SBGN PD notation.

Kinetic category is an exact kinetic model structure representation. Kinetic category diagrams use the same visual elements as the biological category diagrams. However, while both biological and kinetic diagrams show facts that are known, the latter also contain assumed/unconfirmed events, replacing unknown information with assumptions to make them directly convertible to SBML, or other mathematical representation.

Depending on the size of a model, it could be connected to one or several biological diagrams. Similarly to a biological layer, the kinetic model can consist of many kinetic diagrams connected to each other. Again, an outline (or phenomenological diagram) can be used to assemble several kinetic diagrams into one virtual environment.

Figures 3.2 - 3.4 demonstrate how the same events of EGFR signalling can be represented as phenomenological, biological and kinetic diagrams. A PHENOMENOLOGICAL diagram (Figure 3.2) in this example represents EGFR signalling as a network of protein-protein interactions. The corresponding BIOLOGICAL diagram (Figure 3.3) visualises detailed information based on experimental data from the literature. The KINETIC diagram (Figure 3.4) of the EGFR signalling by Kholodenko et al. (1999) does not repeat biological diagrams as several assumptions are added and resuling kinetic model is shown.

Taken together, the collection of these categories is defined as the Network Representation framework (NR).

Table 3.1. The organisation of the levels of the Network Representation Framework.

	DIAGRAM		
	PHENOMENOLOGICAL (OUTLINE)	BIOLOGICAL (DETAILED DIAGRAM)	KINETIC MODEL
Hierarchy of constrains	No constraint. No standard	Constrained	Constraint and rigorous representation. Should be converted to kinetic models.
Standard to support	None	SBGN	SBML, SBGN
Hierarchy of details	Not detailed. Represents biological phenomena	Detailed. Omissions allowed	All states and entities should be explicitly stated
Functions	Generalisation, reviewing, navigation and network managing	Visualisation of detailed evidence-based information	Assumptions
Notation	Notation is not required usually	Well-defined notation is required	Well-defined notation and special set of rules
Ambiguity	Ambiguous; but could be semantically unambiguous	Unambiguous	Unambiguous
Machine-readability	Human-readable; can contain machine-readable elements (hyperlinks)	Machine-readable; human-readable	Machine-readable. Fully automated conversion to kinetic models

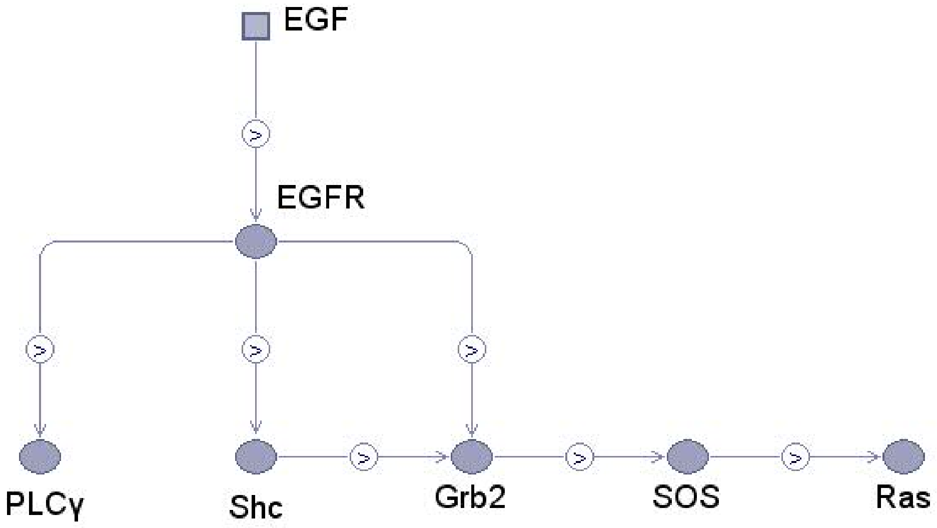


Figure 3.2. Early events of EGFR signalling. PHENOMENOLOGICAL MODEL. An outline for the detailed EGFR signalling map (Figure 3.3) represents the events as protein-protein interactions. Each link from these diagrams corresponds to a more complex and detailed representation on biological layer. On this diagram a link to another diagram is represented by circled symbol '>'. Each macromolecule from the phenomenological layer corresponds to the same entity or corresponding entity included in a complex on a biological diagram.

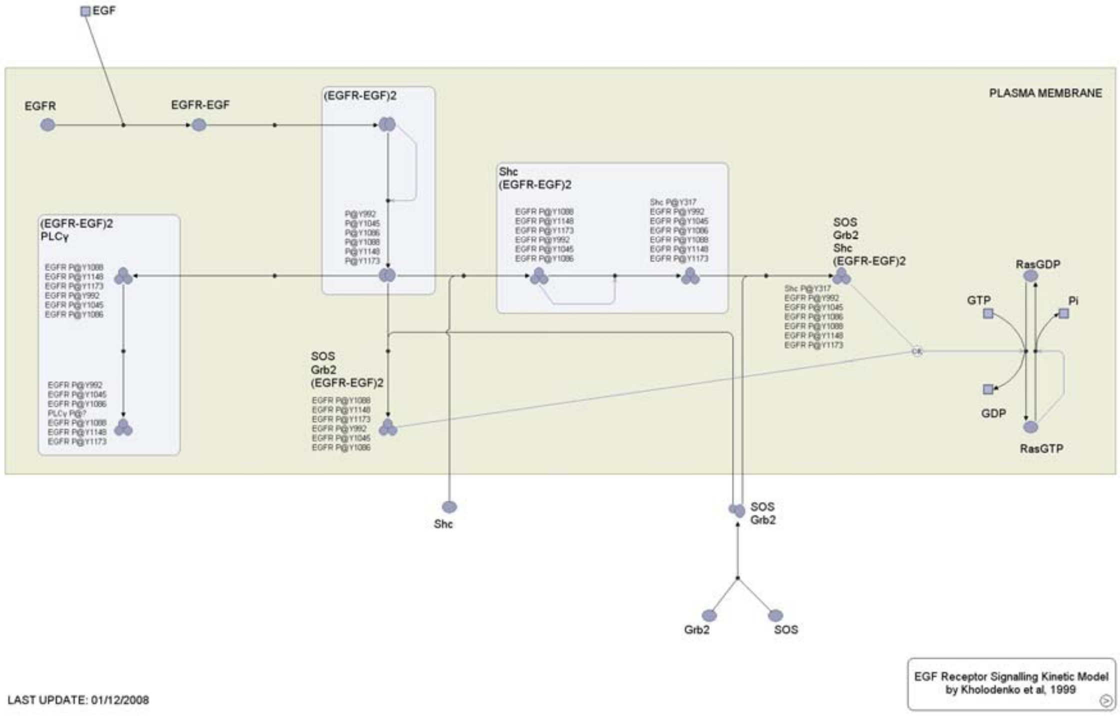


Figure 3.3. Early events of EGFR signalling. BIOLOGICAL MODEL. Events on a biological diagram has to be represented in a well-defined notation. In this case it is done in the UN, but using other graphical language is possible too. For example, including SBGN PD language as an alternative to the UN in the Edinburgh Pathway Editor is discussed.

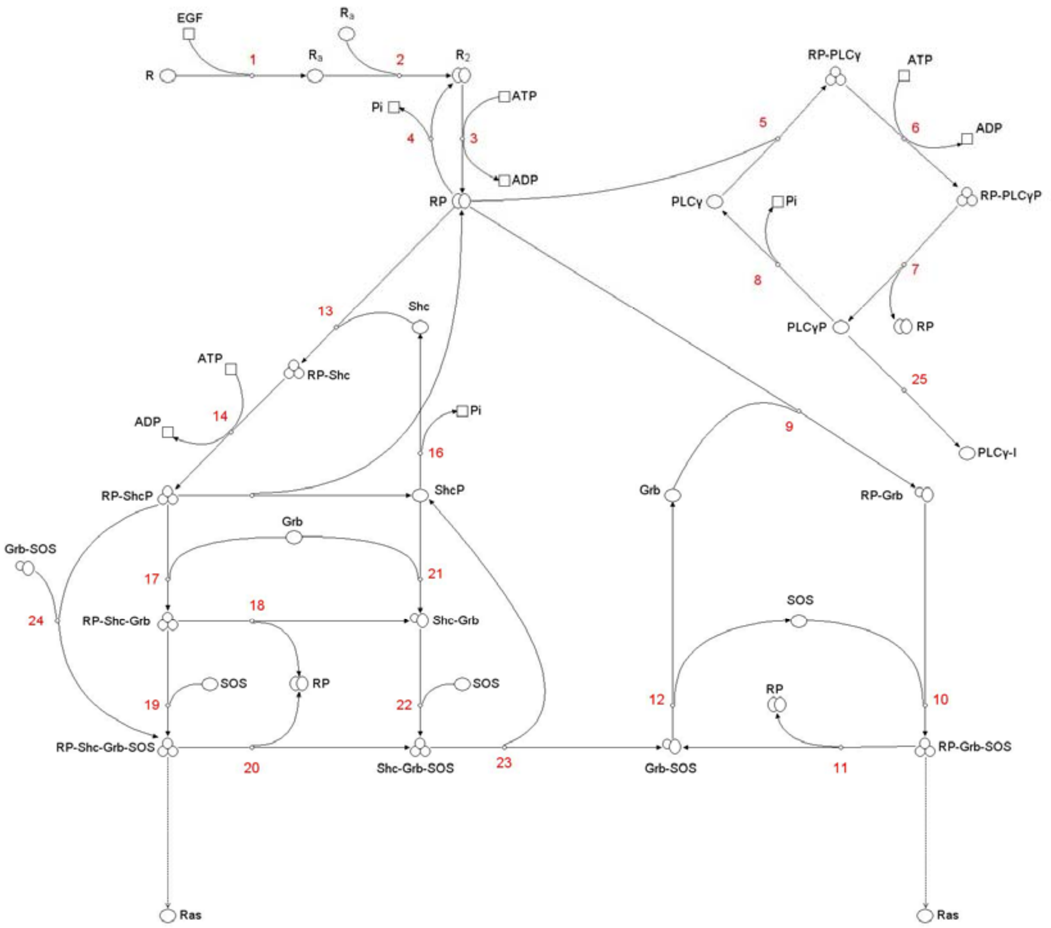


Figure 3.4. Early events of EGFR signalling. KINETIC MODEL by Kholodenko (Kholodenko et al., 1999) represented in the NR. Each macromolecule from the kinetic diagram corresponds to the same or similar (due to possible differences in complex representation on these two layers) entity on the biological level (Figure 3.3). It is possible that the system is represented here with more details, less details or even both when some events are shown in more details and some are omitted. On this diagram the receptor complex is shown in more details to compare to the biological one. An example of simplified representation is the events that include Ras protein. Names could be modified if it helps to make a kinetic scheme easier to understand or simply because the names correspond to the name used in equations etc. Numbers used here correspond to the equation of this model (Kholodenko et al., 1999).

3.4. The hierarchically organized system of diagrams

All three categories of diagrams are connected in the Network Representation framework. The connection is facilitated by three groups of links.

1. Links from one biological diagram to another biological diagram via shared elements (for example, shared compounds, proteins or a set of events).

2. Links from an outline (phenomenological model) to a set of detailed diagrams (biological model). A set of detailed diagrams via shared elements can be assembled into a single virtual map. For example, three biological diagrams of the EGFR signalling network (Figures 3.5 – 3.7) can be joined by an outline (Figure 3.8) into a single virtual map. This virtual map can be visualised if necessary (Figure 3.9).

One biological diagram can be linked to many outlines and be included in multiple virtual maps.

3. Kinetic diagrams are linked to a corresponding detailed map or to an outline (to a large virtual detailed map).

The most important advantages of the system are:

1) There is no need to draw large maps in order to describe a large system. As maps can be comparatively small, it is easy to review, check and update them.

2) A single framework is able to visualise a large system if necessary, due to the hierarchical organization of different types of diagrams. Outlines can be created at as many levels of generalization as necessary.

3) The diagrams that are created based on experimental data are clearly separated from biological and mathematical diagrams, i.e. those in which assumptions or proposed events are added.

4) As the NR is adapted to using different languages (from a graphically ambiguous scheme to kinetic model visualisation), it can be used as a data-sharing and model-sharing environment by both biologists and mathematicians.

In the Network Representation framework outlines play an essential role as they are used for network management. In reality, there exists no single metabolic or signalling network, given that different cell types display varying subsets of possible pathways at different developmental stages or in varying environmental conditions. The framework described here enables the construction of cell- and developmental stage-specific virtual maps. This is done by combining appropriate subsets of detailed maps. For example, for the liver-specific metabolic network steroid hormone metabolism is not going to be included.

The next two chapters show how the framework can be used for visualising large networks.

There are several ways to prepare outlines on phenomenological layer. Some of the possible ways to represent diagrams are listed bellow.

1. A shape/box on an outline corresponds to a map on the lower level;
2. A shape/box on an outline corresponds to a map fragment on the lower level;
3. A connector on an outline could signify one or several events on the lower level (Figures 3.2);
4. An object from an outline (for example, a macromolecule, a complex or a metabolite) could correspond to an object with the same name or several objects with the same name but different states (Figure 3.3).

It is important to clarify that designing the way information from biological layer is presented on an outline, or information from lower level outline(s) is presented on a higher level outline, is a task that is more complex than simple linkage of diagrams from different levels. Diagrams from a lower level can be represented on an outline in a very straightforward way: any object (simple box) can be used as a link. The challenge is to show information in such a way that it could be used for better understanding of a network structure.

This multi-level linkage approach was used in developing systems of outlines for the Edinburgh Human Metabolic Network (Chapter 4) and outlines of the TNF-alpha receptor signalling network (Chapter 5). The aim was to provide a better understanding of a network using simplified text-book-like representations where each outline diagram refers to lower level diagrams and contains selected objects from those diagrams to provide consistency in representation. In other words, the same entity (macromolecule or metabolite) is presented on several outline/detailed levels, but represents more details. Sections 4.3 and 5.2 describe the way a detailed diagram is 'translated' into an outline in further details. Additionally, it should be said that all the outlines have been created manually. However, the generation of this type of outlines could be also made automatic or semiautomatic.

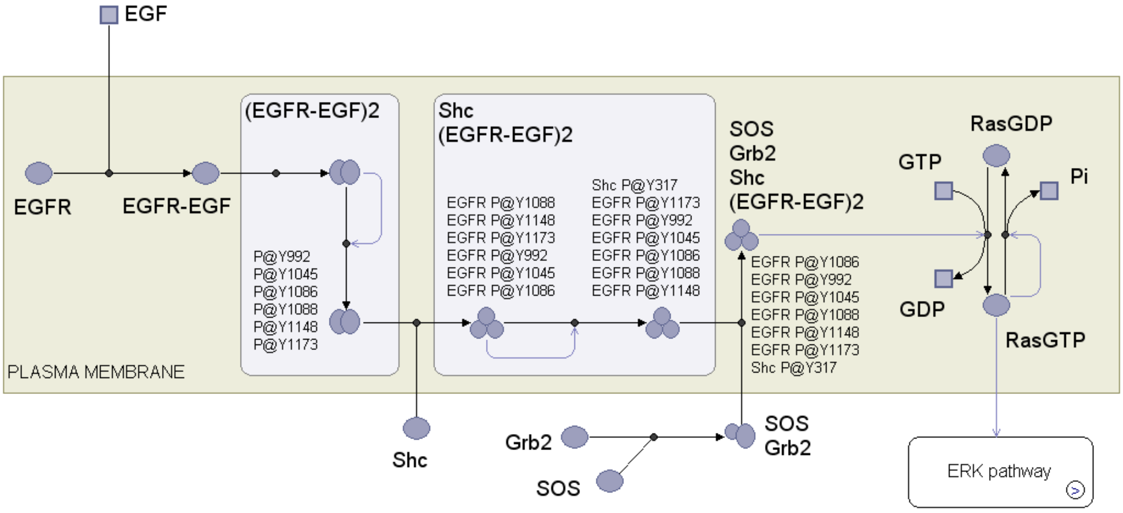


Figure 3.5. EGFR signalling network. BIOLOGICAL. Early events.

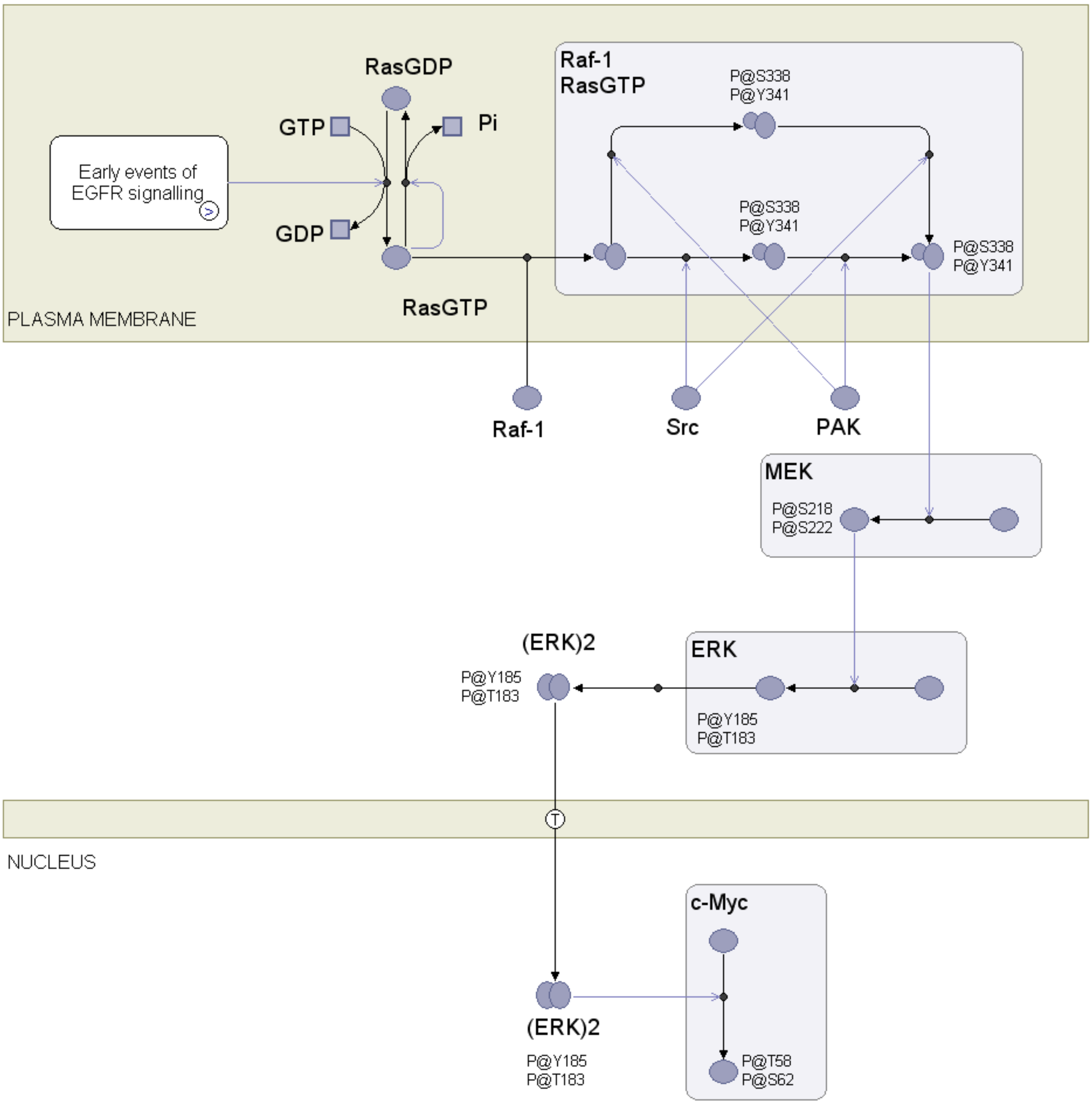


Figure 3.6. EGFR signalling network. BIOLOGICAL. RAS/ERK signaling.

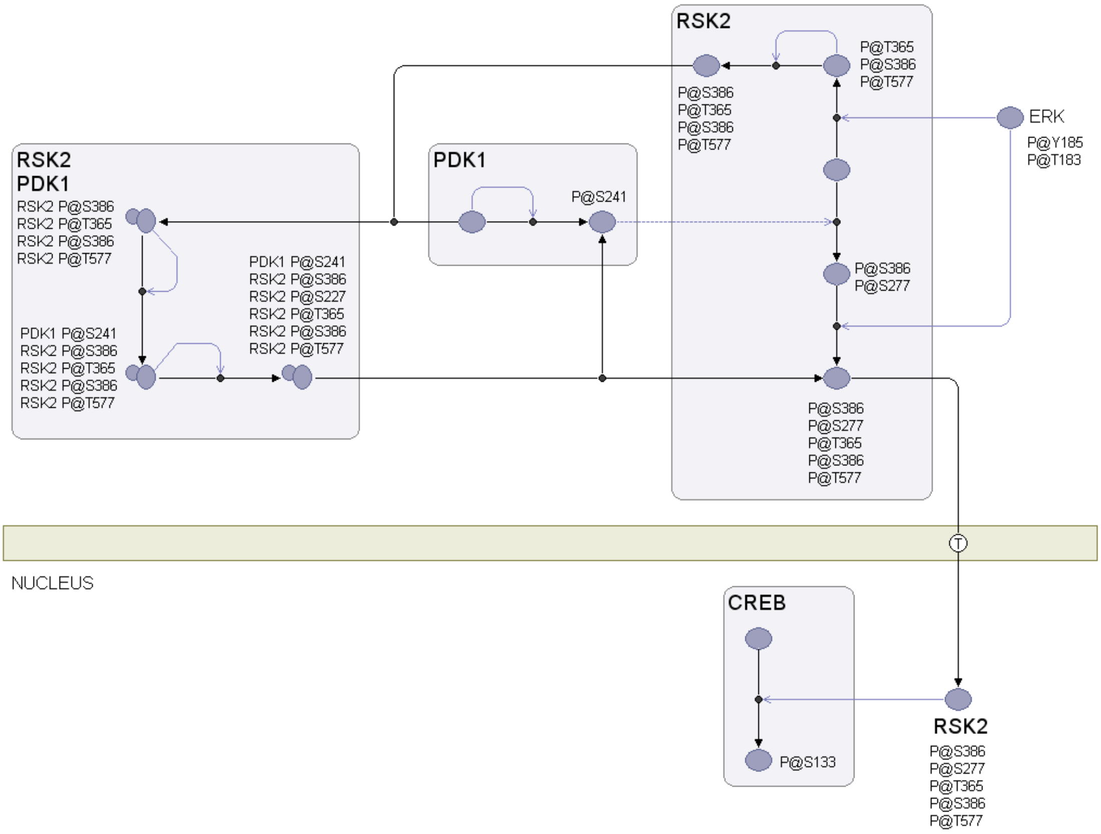


Figure 3.7. EGFR signalling network. BIOLOGICAL. RSK2-PDK1 signalling.

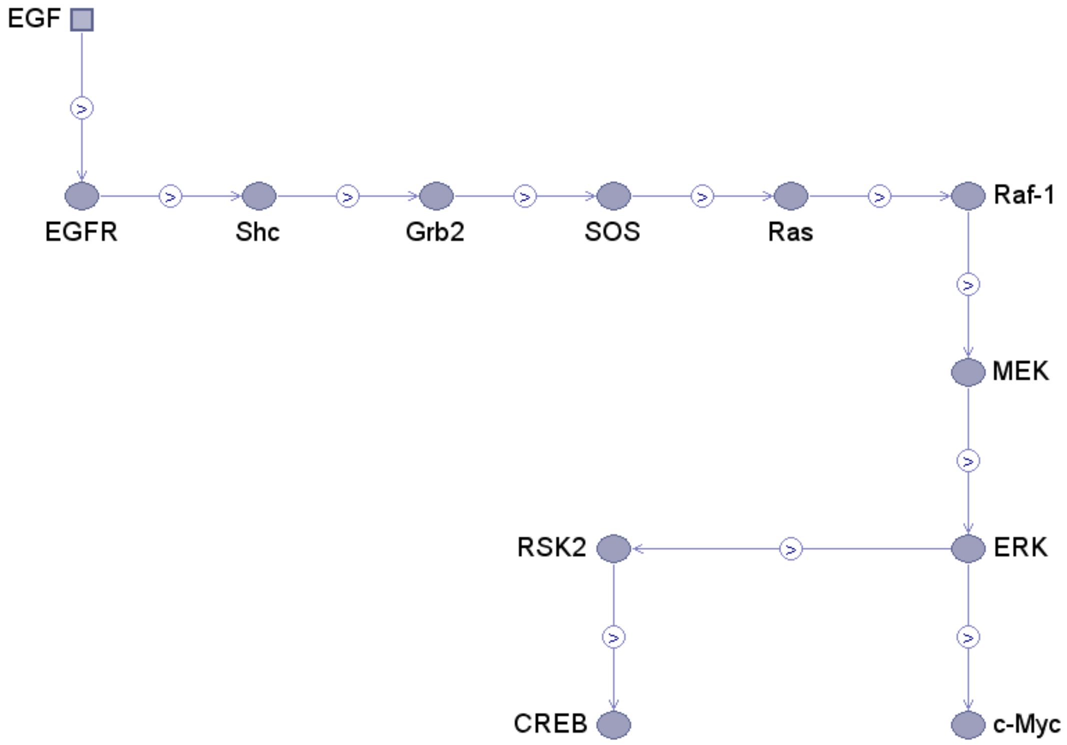


Figure 3.8. EGFR signalling network. PHENOMENOLOGICAL MODEL (OUTLINE). Similarly to the outline for EGFR pathway (Figure 3.2) the main macromolecules involved in the pathway are shown. The connectors represent protein-protein interaction shown as biochemical events on the detailed diagrams in the UN (Figures 3.5 – 3.7). The symbols (>) from the connectors lead to corresponding biological diagram.

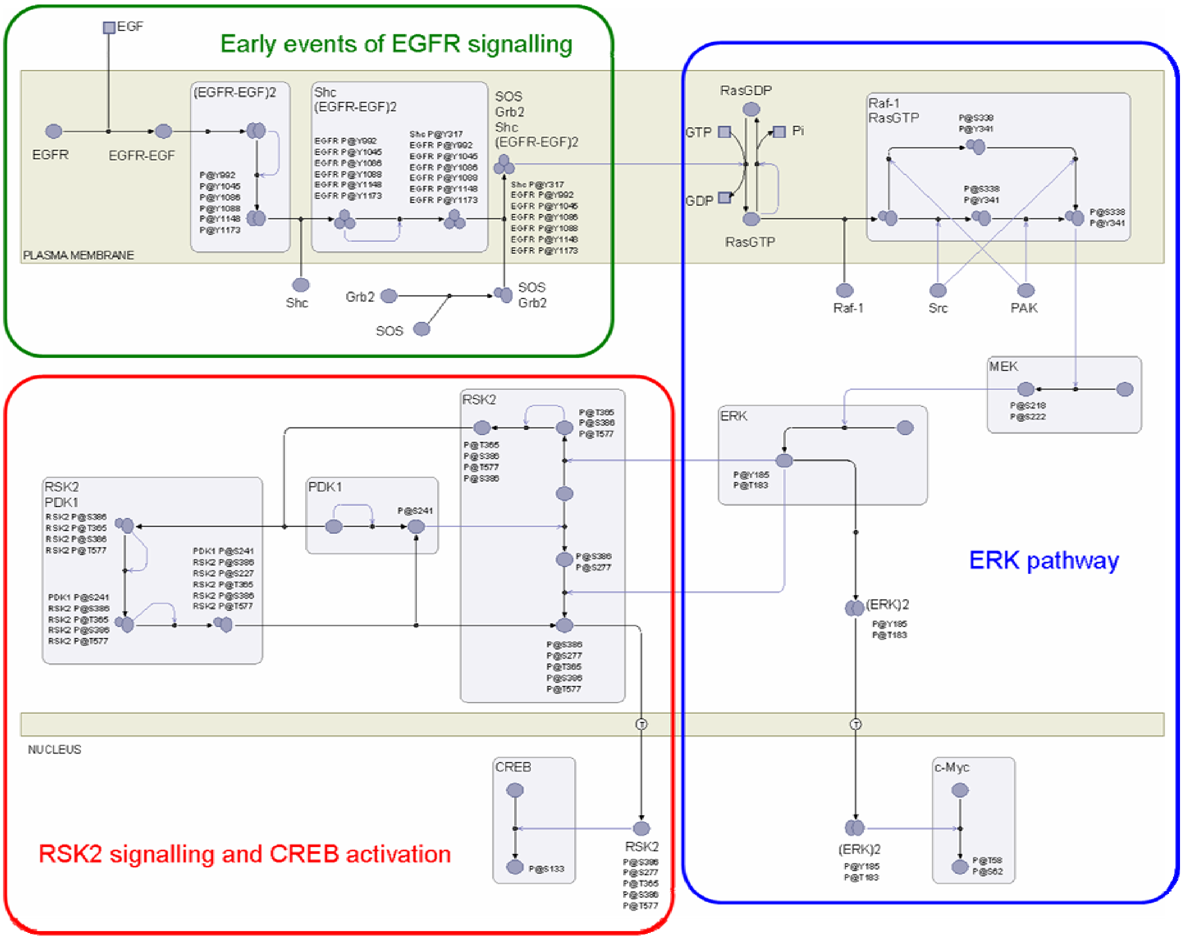


Figure 3.9. EGFR signalling network. BIOLOGICAL. Three diagrams united into a single map. This diagram shows resulting network that created based on the outline (Figure 3.8). Thee biological maps are united into a single map. Shared on primal diagrams elements are used to properly connect separate pathway into one. It is important that smaller diagrams have these shared elements (Figures 3.5 - 3.7).

Chapter 4

Metabolic network representation

The Network Representation Framework (NRF) described in the previous chapter requires a large biological network to prove its abilities. This chapter introduces an extensive example of the NRF application. In the Network Representation Framework an extensive metabolic network is represented at several levels of granularity. Representing a network at several levels of granularity allows one to keep a network representation compact yet providing sufficient level of details. In our example, each diagram included fits within a reasonable printer-friendly size.

Representation of the human metabolic network was developed in 2 steps that are described in the next two sections.

4.1. Step 1. Visualisation of detailed metabolic pathways

During our work on the Edinburgh human metabolic network that contains approximately 3000 reactions (Ma et al., 2007) I participated in preparing the dataset and was responsible for the pathways visual representation. The goal was to compare reactions from the two initial datasets from KEGG and EMPProject in order to avoid repetitions and exclude as many mistakes as possible. On the next stage of the project my responsibility was to subdivide the reactions into different pathways. For that the KEGG pathway has been compared to EMPProject pathways, all the reactions in the dataset have been subdivided into subpathways and a new set of pathways has been created bearing in mind a goal that one pathway should contain no more than 100 reaction to have a printer-friendly size. All the pathways have been visualised in the UN using the Edinburgh Pathway Editor (Sorokin, 2006). Visualisation of the dataset has appeared to be important not only because it is easier to read a graphical representation rather than textual description, but also because in the process unconnected parts of the pathway have been discovered and then missing reactions have been added.

The results of this work were published in 2007 (Ma et al., 2007).

In order to make transformation from textual representation of the pathways (Excel file) to graphical diagrams in the UN, special EPE plugins have been developed by the Computational Systems Biology’s team of software developers that are working on EPE, particularly by Richard Adams and Shakir Ali. My role was to develop corresponding context in EPE (‘Metabolic’) and specify what data have to be uploaded to what field of the context. Next, automatic layout plugin has been prepared by Richard Adams. The initial idea for the plugin was to combine existing algorithms of graphical layouts and take into account the grid size of a map. That puts objects on a map in certain pattern which makes it easier to read a map.

4.2. Step 2. Developing a system of outlines for human metabolic network representation

Applying the NRF to the Edinburgh human metabolic network, a system of outlines has been developed to show networks at different levels of detail. This system of outlines has at least three important functions:

1) providing navigation through the whole network;

2) outlines (phenomenological diagrams) provide an intelligent overview;

3) each outline unites lower-level diagrams into one large virtual map.

While developing the outlines many different examples of metabolic network representation (Section 1.4) have been taken into account. The main idea of the proposed outline system was to provide a user with 1) useful text-book oriented version of the main outline where only the most known metabolites and pathways (for example, glycolysis) are represented and then 2) show more details on the next level where a set of outlines would be prepared and each of them would correspond to the major area of metabolism. Finally, 3) detailed metabolic diagrams in the UN could be made available through navigation via the two layers of the outlines described above.

The result of this work is illustrated in Figures 4.1 – 4.14.

The main outline provides an overview of the whole network and underlines the connection between major parts of the network (Figure 4.1). Links connect the main outline to eleven outlines of the lower level (Figures 4.2 – 4.12). The lower level outlines in their turn are linked to the detailed diagrams.

Each major part of metabolism is represented by a separate outline and each of them is linked to the main outline (Figure 4.1): Carbohydrate metabolism, Amino acid metabolism, Essential fatty acid metabolism, Eicosanoid metabolism, Sphingolipid metabolism, Glycerophospholipid metabolism, Steroid metabolism, Vitamins and Porphyrin metabolism (Figures 4.2 – 4.12). Links from each of the outlines on this level lead to another more detailed representation of metabolic pathway on a biological layer. Figure 4.13 shows one of such detailed diagrams. In the Edinburgh Pathway Editor (EPE) it is possible to navigate through the network using hyperlinks (Figure 4.14). The enclosed copy of the CD contains a portable version of EPE with pre-installed database that contains all the examples including the EHMN visualisation.

The approach used for preparing the system of outlines is described in the next section as a proposed algorithm for semi-automatic outline generation so this work could be reproduced and applied to other metabolic networks.

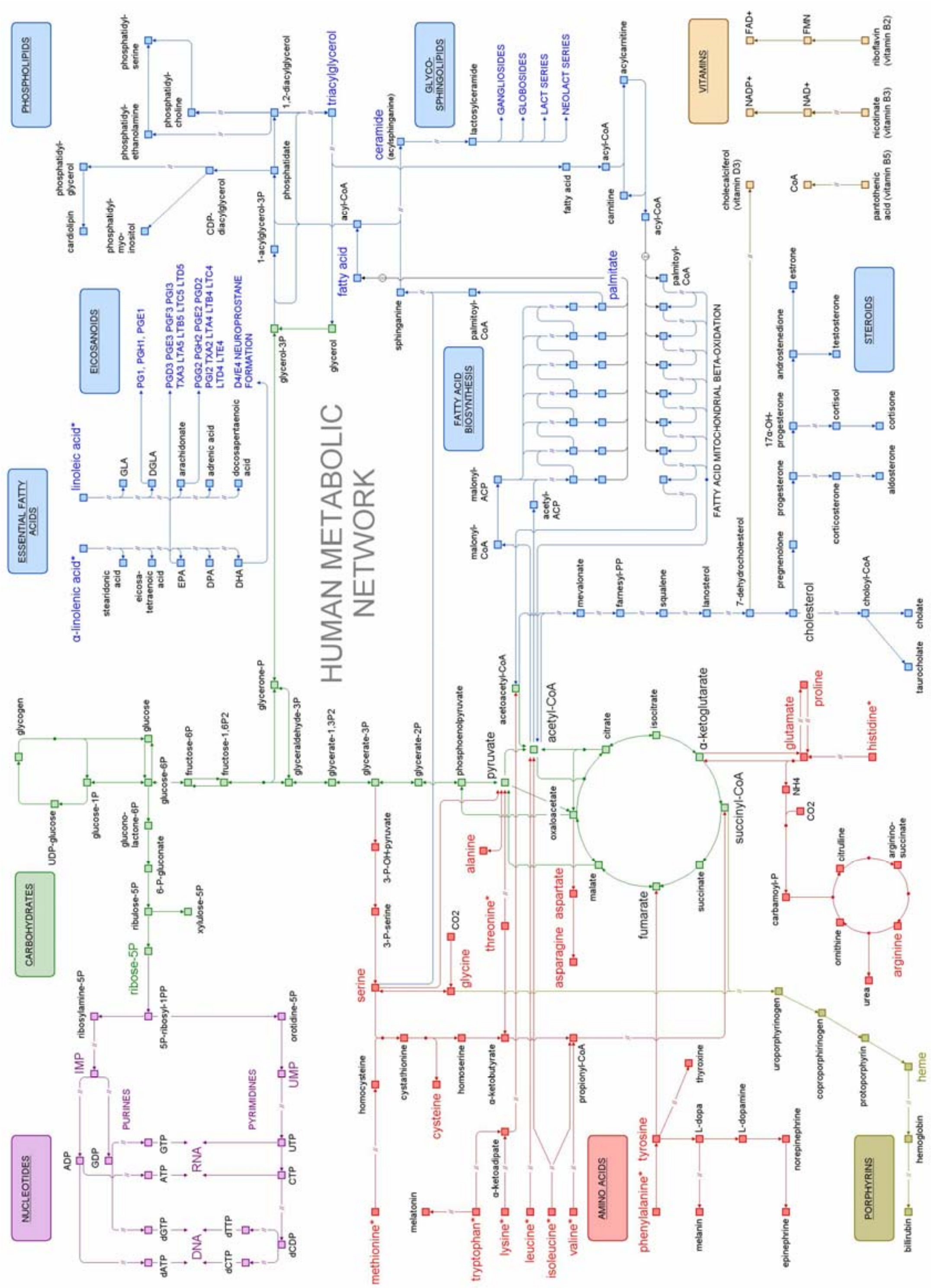


Figure 4.1. The EHMN outlines. Human metabolic network (main outline). Please use the CD provided to see a high resolution version of the image.

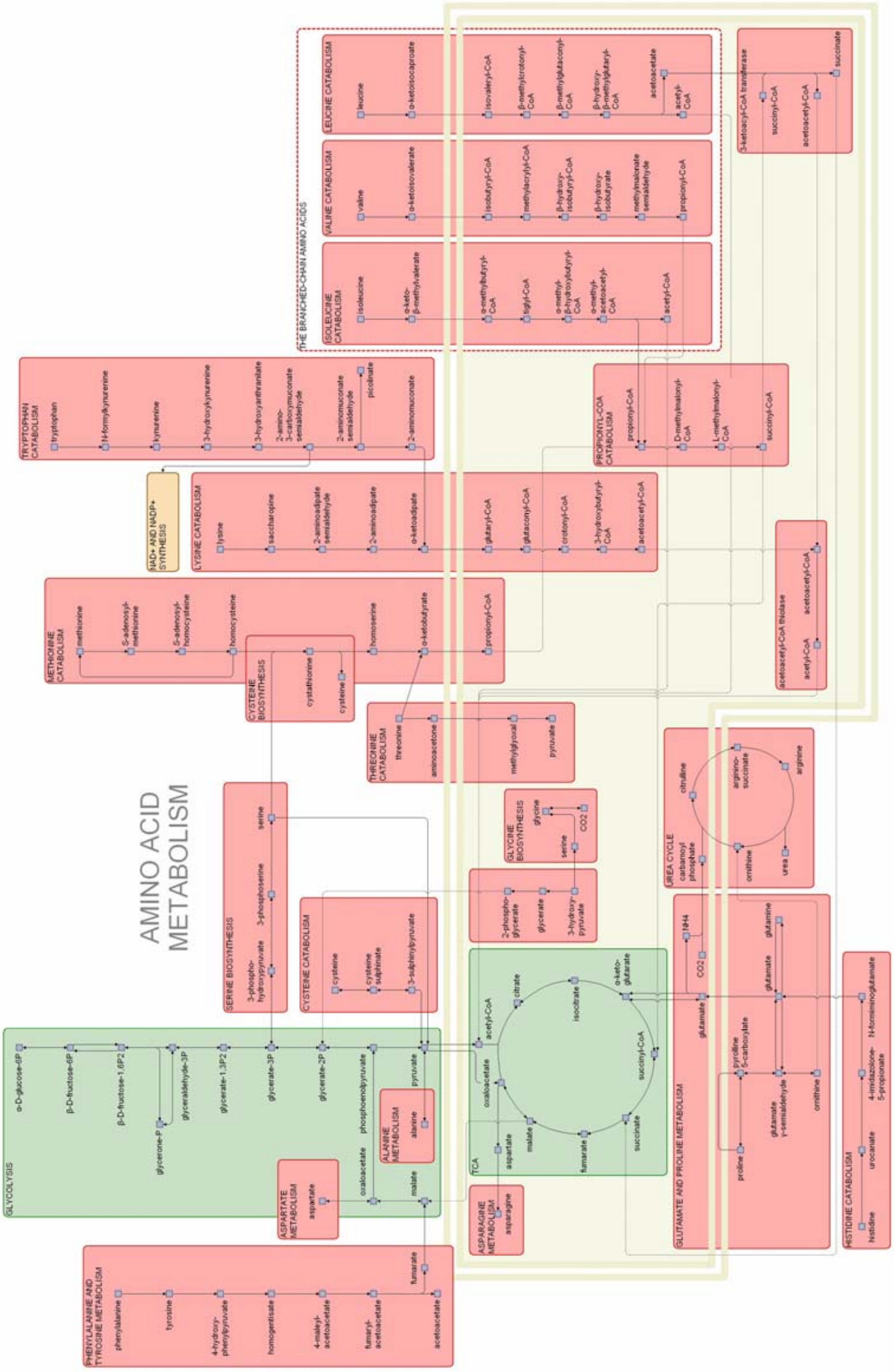


Figure 4.2. The EHMN outlines. Amino acid metabolism. Please use the CD provided to see a high resolution version of the image.

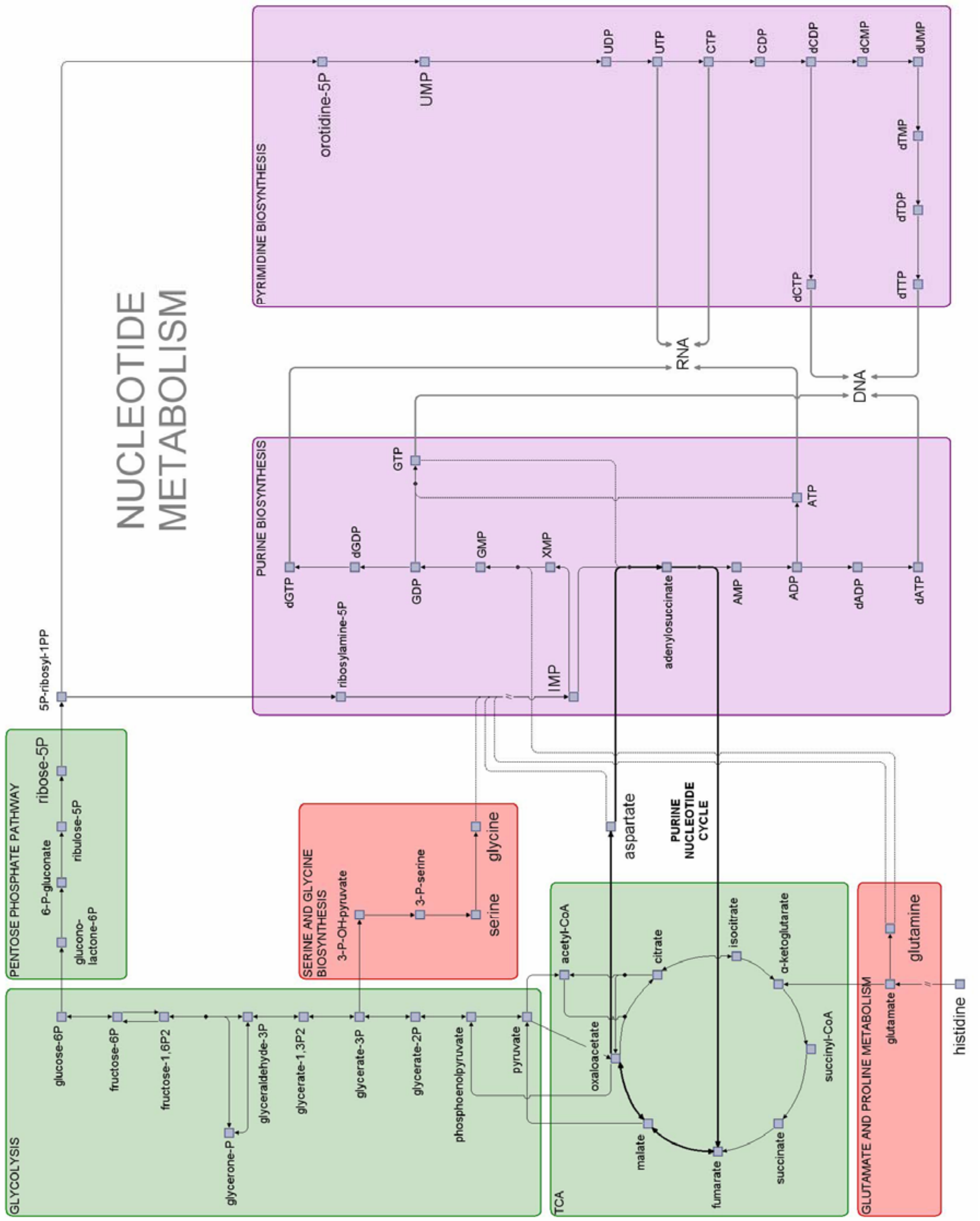


Figure 4.3. The EHMN outlines. Nucleotide metabolism. Please use the CD provided to see a high resolution version of the image.

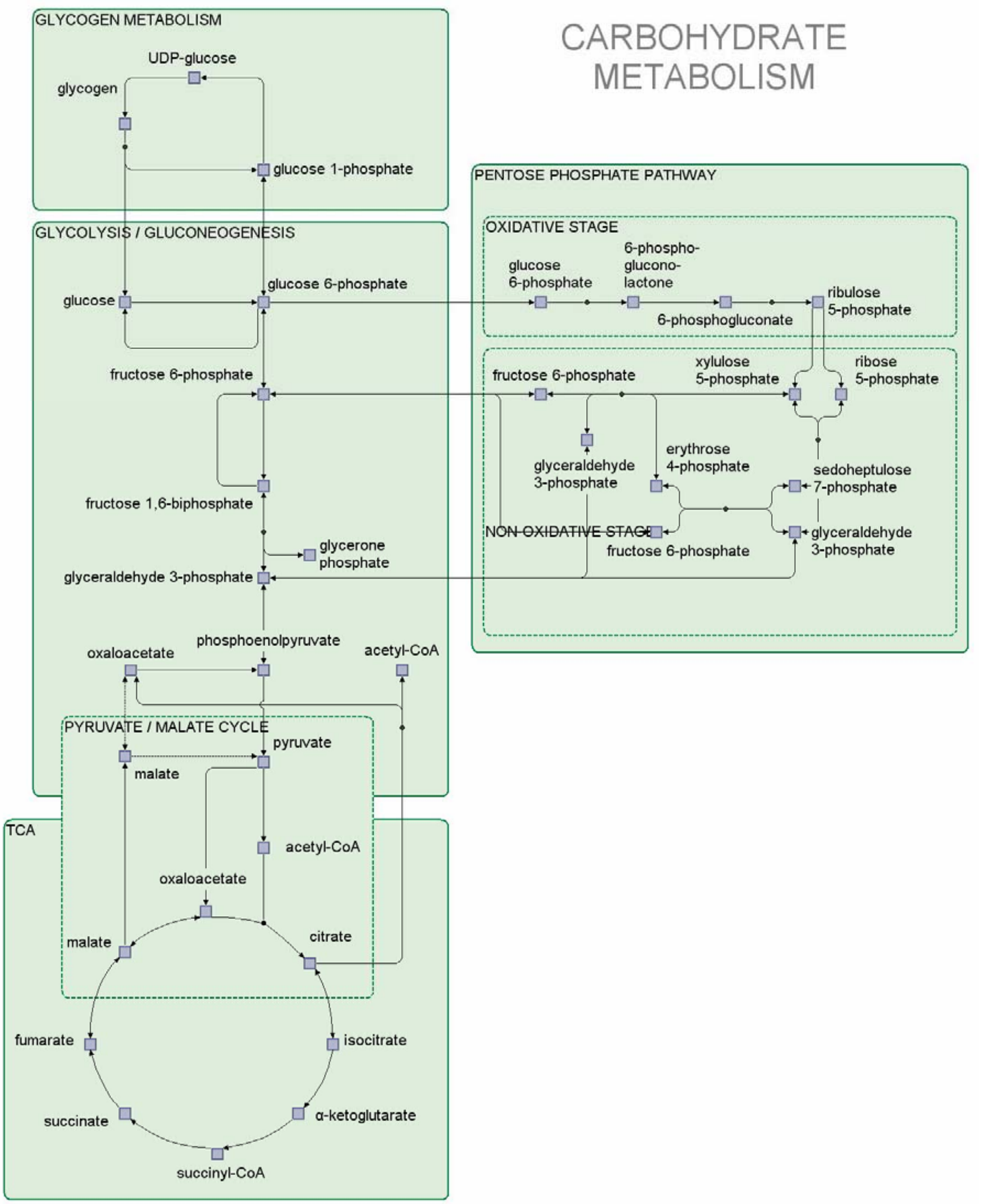


Figure 4.4. The EHMN outlines. Carbohydrate metabolism. Please use the CD provided to see a high resolution version of the image.

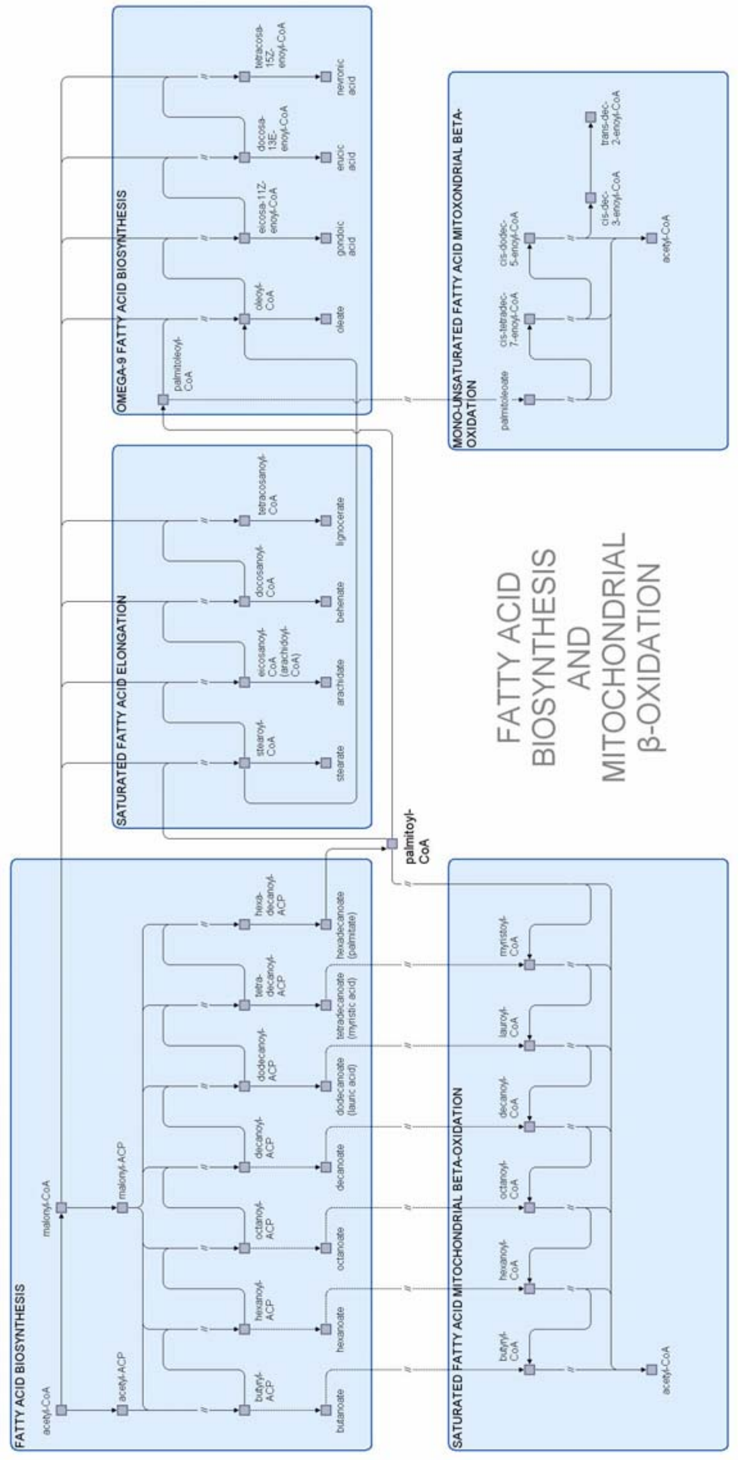


Figure 4.5. The EHMN outlines. Fatty acid biosynthesis. Please use the CD provided to see a high resolution version of the image.

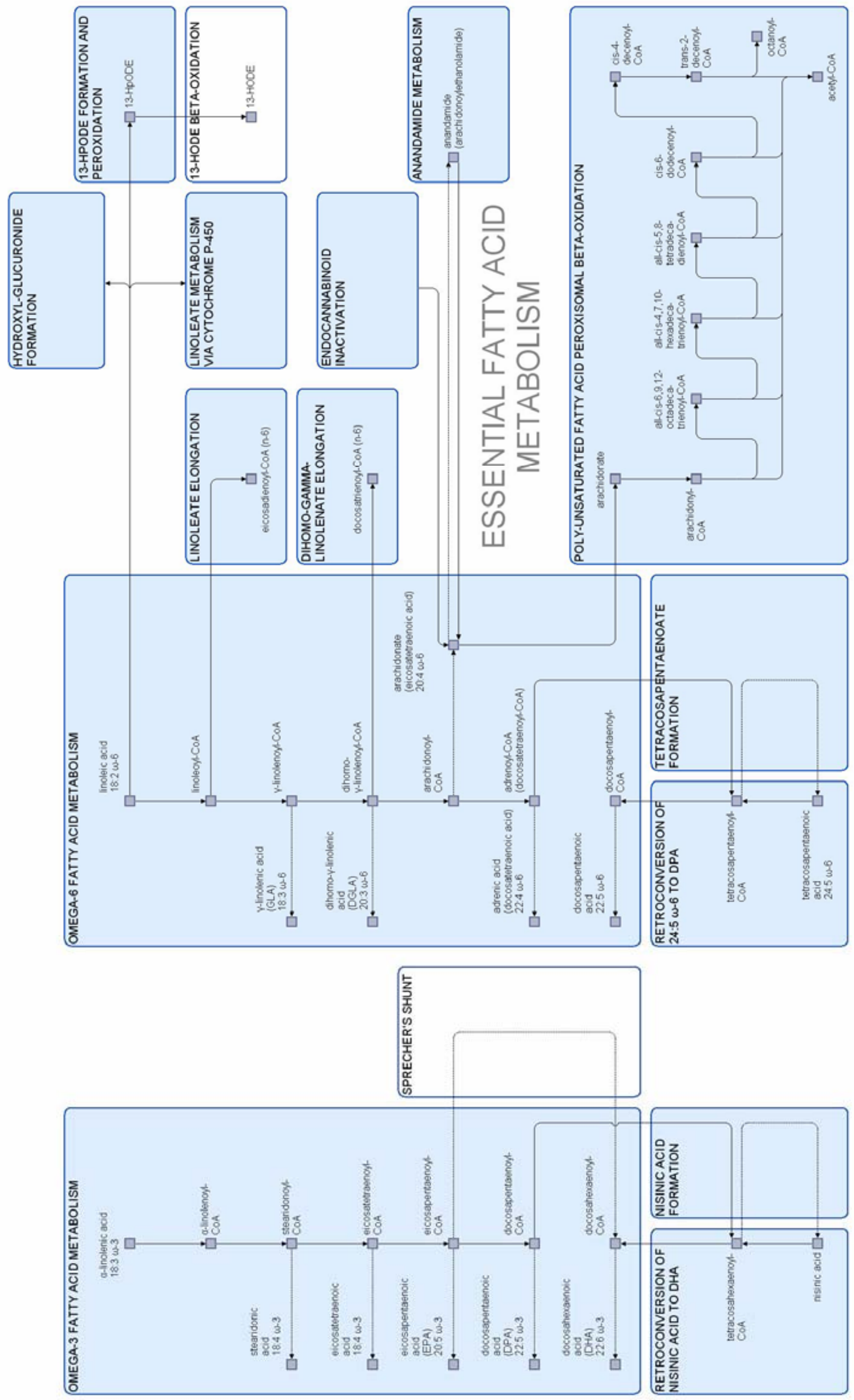

ESSENTIAL FATTY ACID METABOLISM
HYDROXYL-GLUCURONIDE FORMATION
13-HPODE FORMATION AND PEROXIDATION
13-HpODE
13-HODE BETA-OXIDATION
13-HODE
LINOLEATE METABOLISM VIA CYTOCHROME P-450
ENDOCANNABINOID INACTIVATION
ANANDAMIDE METABOLISM
anandamide (arachidonoylethanolamide)
LINOLEATE ELONGATION
eicosadienoyl-CoA (n-6)
DIHOMO-GAMMA-LINOLENATE ELONGATION
docosatrienoyl-CoA (n-6)
POLY-UNSATURATED FATTY ACID PEROXISOMAL BETA-OXIDATION
arachidonate
arachidonyl-CoA
all-cis-6,9,12-octadeca-trienoyl-CoA
all-cis-4,7,10-hexadeca-trienoyl-CoA
all-cis-5,8-tetradeca-dienoyl-CoA
cis-6-dodecenoyl-CoA
cis-4-decenoyl-CoA
trans-2-decenoyl-CoA
octanoyl-CoA
acetyl-CoA
OMEGA-6 FATTY ACID METABOLISM
linoleic acid 18:2 ω-6
linoleoyl-CoA
γ-linolenoyl-CoA
γ-linolenic acid (GLA) 18:3 ω-6
dihomo-γ-linolenoyl-CoA
dihomo-γ-linolenic acid (DGLA) 20:3 ω-6
arachidonoyl-CoA
arachidonate (eicosatetraenoic acid) 20:4 ω-6
adrenoyl-CoA (docosatetraenoyl-CoA)
adrenic acid (docosatetraenoic acid) 22:4 ω-6
docosapentaenoyl-CoA
docosapentaenoic acid 22:5 ω-6
TETRACOSAPENTAENOATE FORMATION
RETROCONVERSION OF 24:5 ω-6 TO DPA
tetracosapentaenoyl-CoA
tetracosapentaenoic acid 24:5 ω-6
SPRECHER'S SHUNT
OMEGA-3 FATTY ACID METABOLISM
α-linolenic acid 18:3 ω-3
α-linolenoyl-CoA
stearidonoyl-CoA
stearidonic acid 18:4 ω-3
eicosatetraenoyl-CoA
eicosatetraenoic acid 18:4 ω-3
eicosapentaenoyl-CoA
eicosapentaenoic acid (EPA) 20:5 ω-3
docosapentaenoyl-CoA
docosapentaenoic acid (DPA) 22:5 ω-3
docosahexaenoyl-CoA
docosahexaenoic acid (DHA) 22:6 ω-3
NISINIC ACID FORMATION
RETROCONVERSION OF NISINIC ACID TO DHA
tetracosahexaenoyl-CoA
nisinic acid

Figure 4.6. The EHMN outlines. Essential fatty acid metabolism. Please use the CD provided to see a high resolution version of the image.

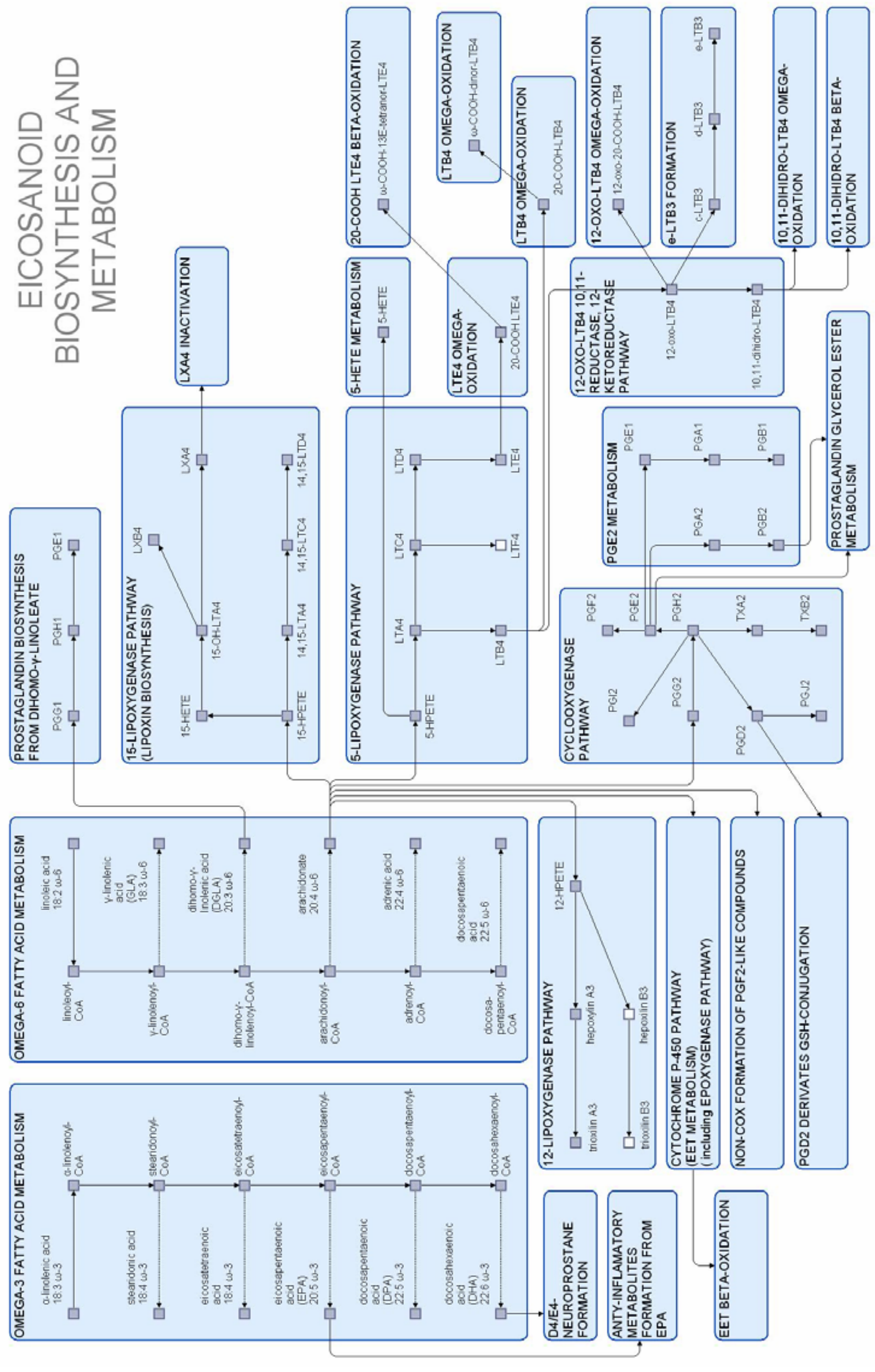

EICOSANOID BIOSYNTHESIS AND METABOLISM
OMEGA-3 FATTY ACID METABOLISM
α-linolenic acid 18:3 ω-3
α-linolenoyl-CoA
steaidonic acid 18:4 ω-3
steaidonoyl-CoA
eicosatetraenoic acid 18:4 ω-3
eicosatetraenoyl-CoA
eicosapentaenoic acid (EPA) 20:5 ω-3
eicosapentaenoyl-CoA
docosapentaenoic acid (DPA) 22:5 ω-3
docosapentaenoyl-CoA
docosahexaenoic acid (DHA) 22:6 ω-3
docosahexaenoyl-CoA
OMEGA-6 FATTY ACID METABOLISM
linoleic acid 18:2 ω-6
linoleoyl-CoA
γ-linolenic acid (GLA) 18:3 ω-6
γ-linolenoyl-CoA
dihomo-γ-linolenic acid (DGLA) 20:3 ω-6
dihomo-γ-linolenoyl-CoA
arachidonate 20:4 ω-6
arachidonoyl-CoA
adrenic acid 22:4 ω-6
adrenoyl-CoA
docosapentaenoic acid 22:5 ω-6
docosa-pentaenoyl-CoA
PROSTAGLANDIN BIOSYNTHESIS FROM DIHOMO-γ-LINOLEATE
PGG1
PGH1
PGE1
15-LIPOXYGENASE PATHWAY (LIPOXIN BIOSYNTHESIS)
15-HETE
15-OH-LTA4
LXB4
LXA4
15-HPETE
14,15-LTA4
14,15-LTC4
14,15-LTD4
LXA4 INACTIVATION
5-LIPOXYGENASE PATHWAY
5-HPETE
LTA4
LTC4
LTD4
LTB4
LTF4
LTE4
5-HETE METABOLISM
5-HETE
LTE4 OMEGA-OXIDATION
20-COOH LTE4
20-COOH LTE4 BETA-OXIDATION
ω-COOH-13E-tetranor-LTE4
LTB4 OMEGA-OXIDATION
ω-COOH-dinor-LTB4
LTB4 OMEGA-OXIDATION
20-COOH-LTB4
12-OXO-LTB4 OMEGA-OXIDATION
12-oxo-20-COOH-LTB4
e-LTB3 FORMATION
c-LTB3
d-LTB3
e-LTB3
10,11-DIHIDRO-LTB4 OMEGA-OXIDATION
10,11-DIHIDRO-LTB4 BETA-OXIDATION
12-OXO-LTB4 10,11-REDUCTASE, 12-KETOREDUCTASE PATHWAY
12-oxo-LTB4
10,11-dihidro-LTB4
PGE2 METABOLISM
PGE1
PGA1
PGB1
PGA2
PGB2
PROSTAGLANDIN GLYCEROL ESTER METABOLISM
CYCLOOXYGENASE PATHWAY
PGF2
PGE2
PGH2
TXA2
TXB2
PGI2
PGG2
PGD2
PGJ2
12-LIPOXYGENASE PATHWAY
12-HPETE
hepoxilin A3
hepoxilin B3
trioxilin A3
trioxilin B3
CYTOCHROME P-450 PATHWAY (EET METABOLISM) (including EPOXYGENASE PATHWAY)
NON-COX FORMATION OF PGF2-LIKE COMPOUNDS
PGD2 DERIVATES GSH-CONJUGATION
D4/E4-NEUROPROSTANE FORMATION
ANTY-INFLAMATORY METABOLITES FORMATION FROM EPA
EET BETA-OXIDATION

Figure 4.7. The EHMN outlines. Eicosanoid metabolism. Please use the CD provided to see a high resolution version of the image.

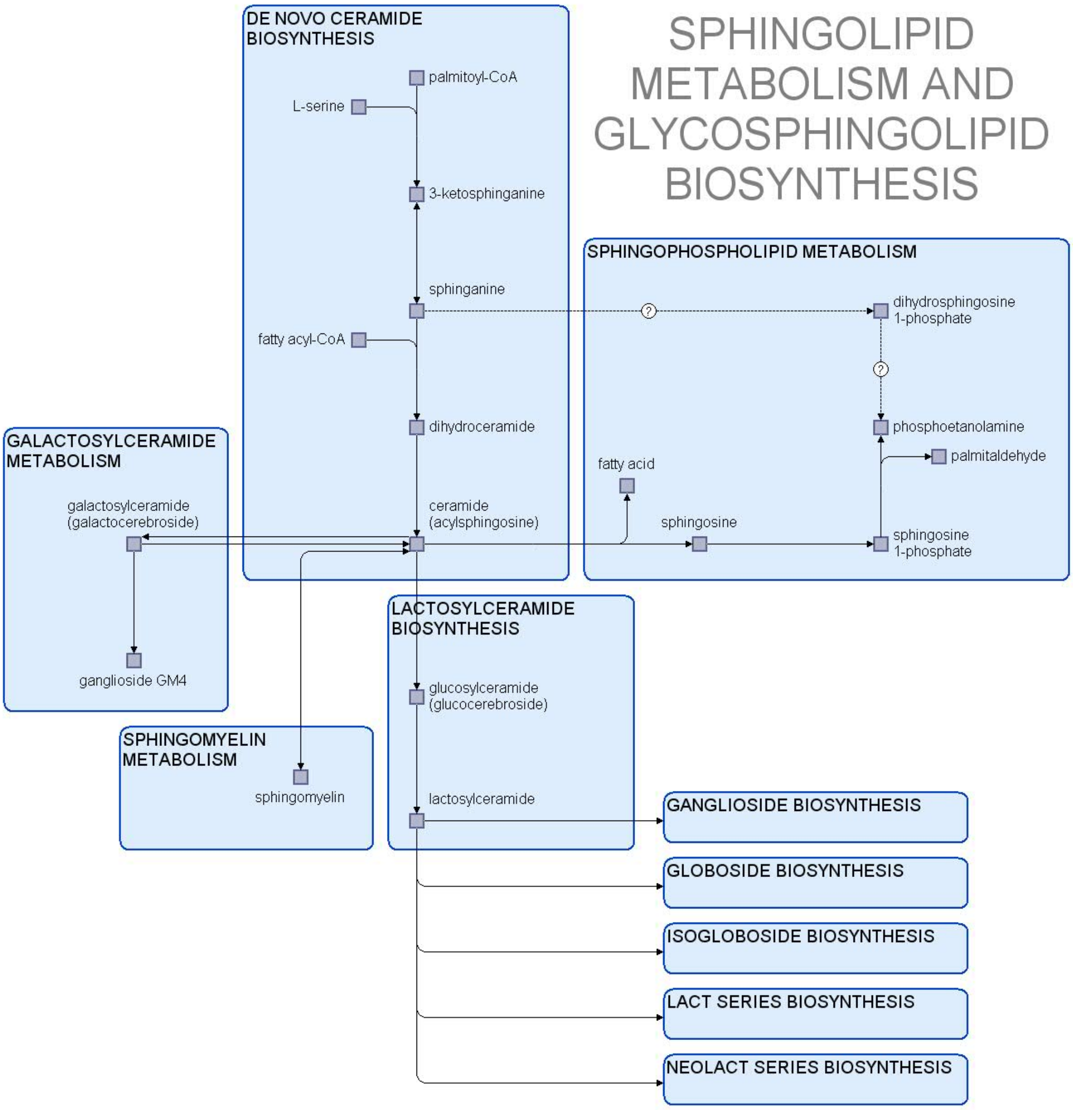

SPHINGOLIPID METABOLISM AND GLYCOSPHINGOLIPID BIOSYNTHESIS
DE NOVO CERAMIDE BIOSYNTHESIS
palmitoyl-CoA
L-serine
3-ketosphinganine
sphinganine
fatty acyl-CoA
dihydroceramide
ceramide (acylsphingosine)
SPHINGOPHOSPHOLIPID METABOLISM
dihydrosphingosine 1-phosphate
?
?
phosphoetanolamine
palmitaldehyde
fatty acid
sphingosine
sphingosine 1-phosphate
GALACTOSYLCERAMIDE METABOLISM
galactosylceramide (galactocerebroside)
ganglioside GM4
LACTOSYLCERAMIDE BIOSYNTHESIS
glucosylceramide (glucocerebroside)
lactosylceramide
SPHINGOMYELIN METABOLISM
sphingomyelin
GANGLIOSIDE BIOSYNTHESIS
GLOBOSIDE BIOSYNTHESIS
ISOGLOBOSIDE BIOSYNTHESIS
LACT SERIES BIOSYNTHESIS
NEOLACT SERIES BIOSYNTHESIS

Figure 4.8. The EHMN outlines. Sphingolipid metabolism. Please use the CD provided to see a high resolution version of the image.

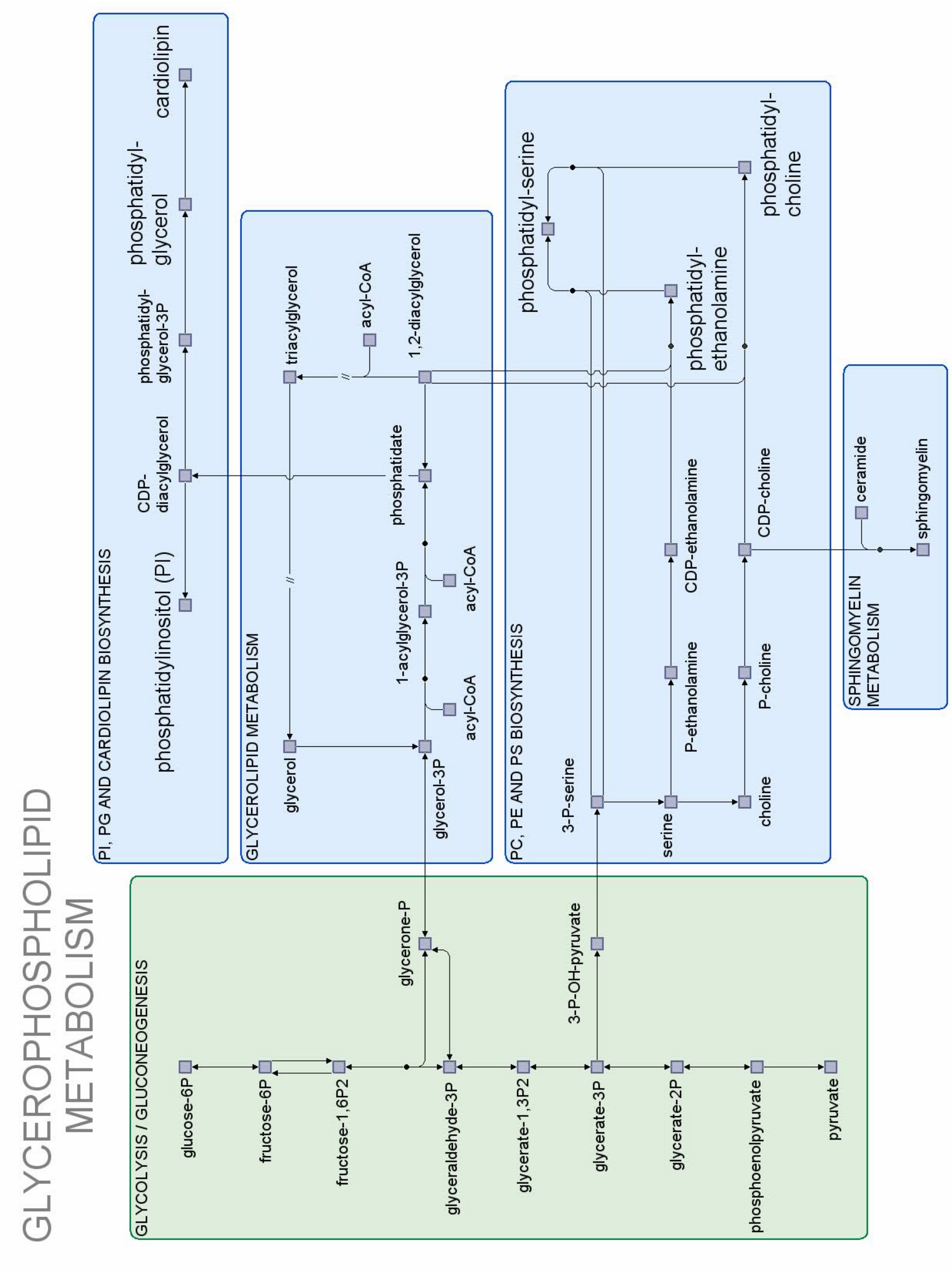


Figure 4.9. The EHMN outlines. Glycerophospholipid metabolism. Please use the CD provided to see a high resolution version of the image.

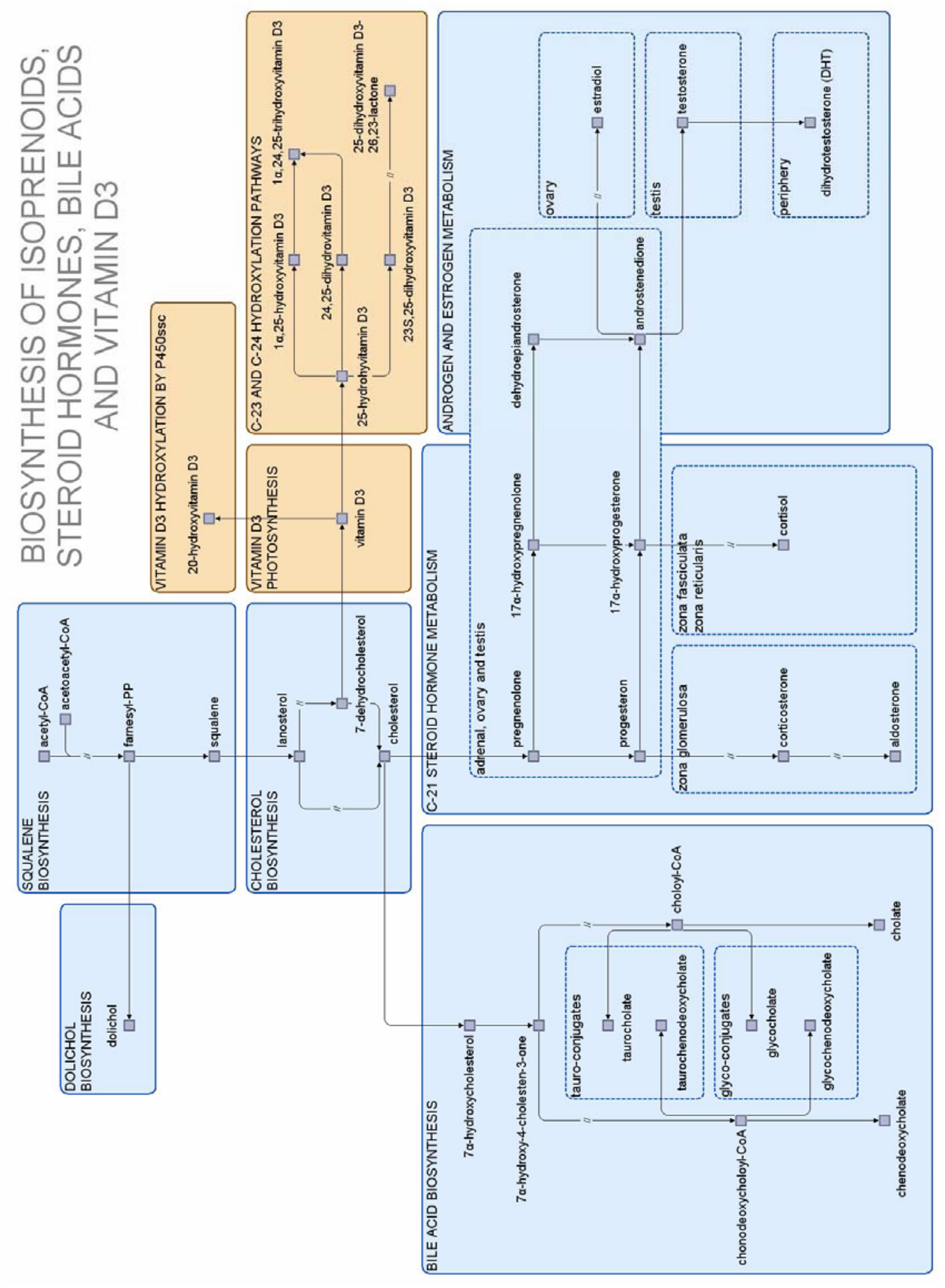
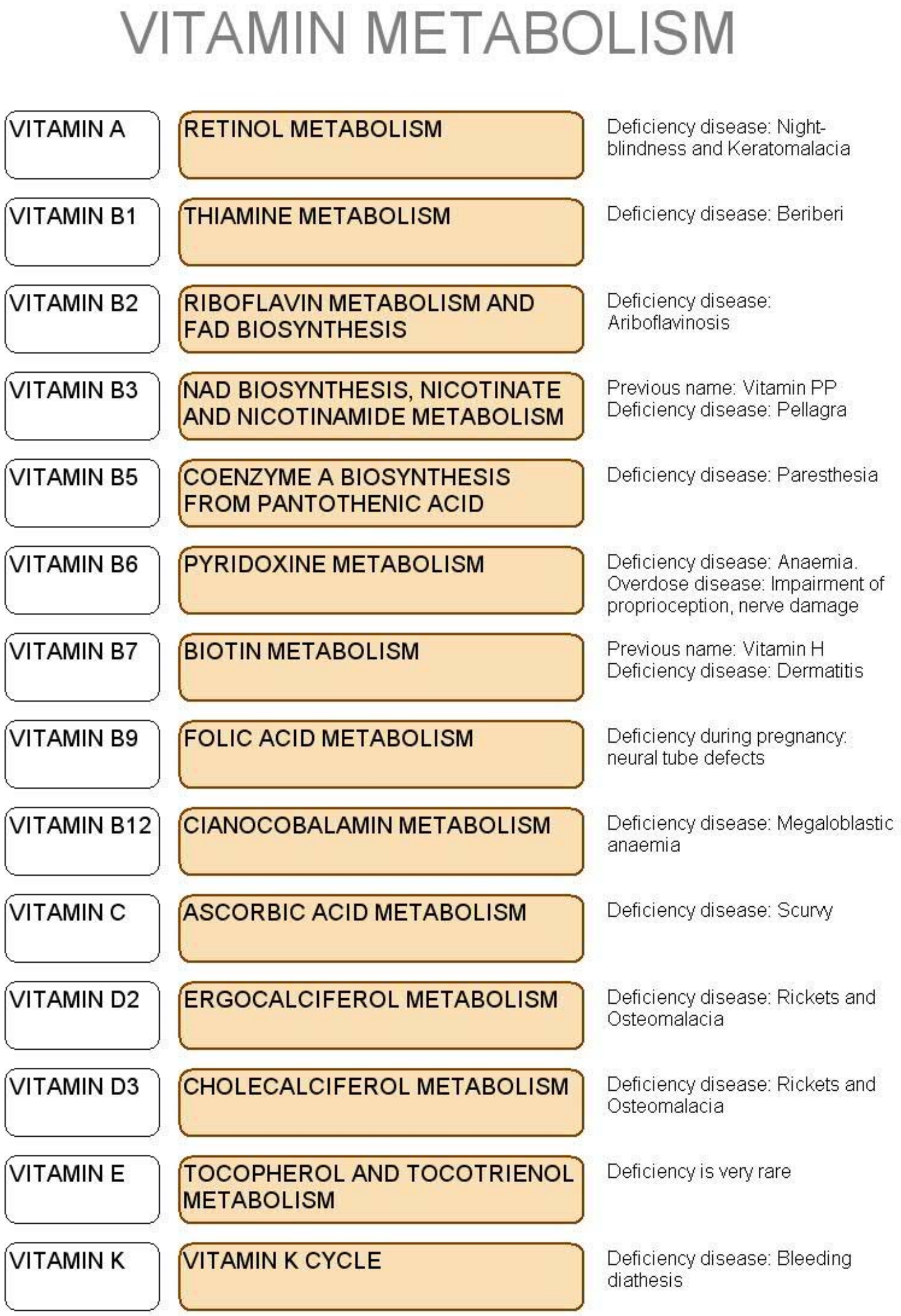


Figure 4.10. The EHMN outlines. Steroid metabolism. Please use the CD provided to see a high resolution version of the image.



VITAMIN METABOLISM

VITAMIN A	RETINOL METABOLISM	Deficiency disease: Night-blindness and Keratomalacia
VITAMIN B1	THIAMINE METABOLISM	Deficiency disease: Beriberi
VITAMIN B2	RIBOFLAVIN METABOLISM AND FAD BIOSYNTHESIS	Deficiency disease: Ariboflavinosis
VITAMIN B3	NAD BIOSYNTHESIS, NICOTINATE AND NICOTINAMIDE METABOLISM	Previous name: Vitamin PP Deficiency disease: Pellagra
VITAMIN B5	COENZYME A BIOSYNTHESIS FROM PANTOTHENIC ACID	Deficiency disease: Paresthesia
VITAMIN B6	PYRIDOXINE METABOLISM	Deficiency disease: Anaemia. Overdose disease: Impairment of proprioception, nerve damage
VITAMIN B7	BIOTIN METABOLISM	Previous name: Vitamin H Deficiency disease: Dermatitis
VITAMIN B9	FOLIC ACID METABOLISM	Deficiency during pregnancy: neural tube defects
VITAMIN B12	CIANOCOBALAMIN METABOLISM	Deficiency disease: Megaloblastic anaemia
VITAMIN C	ASCORBIC ACID METABOLISM	Deficiency disease: Scurvy
VITAMIN D2	ERGOCALCIFEROL METABOLISM	Deficiency disease: Rickets and Osteomalacia
VITAMIN D3	CHOLECALCIFEROL METABOLISM	Deficiency disease: Rickets and Osteomalacia
VITAMIN E	TOCOPHEROL AND TOCOTRIENOL METABOLISM	Deficiency is very rare
VITAMIN K	VITAMIN K CYCLE	Deficiency disease: Bleeding diathesis

Figure 4.11. The EHMN outlines. Vitamin metabolism.

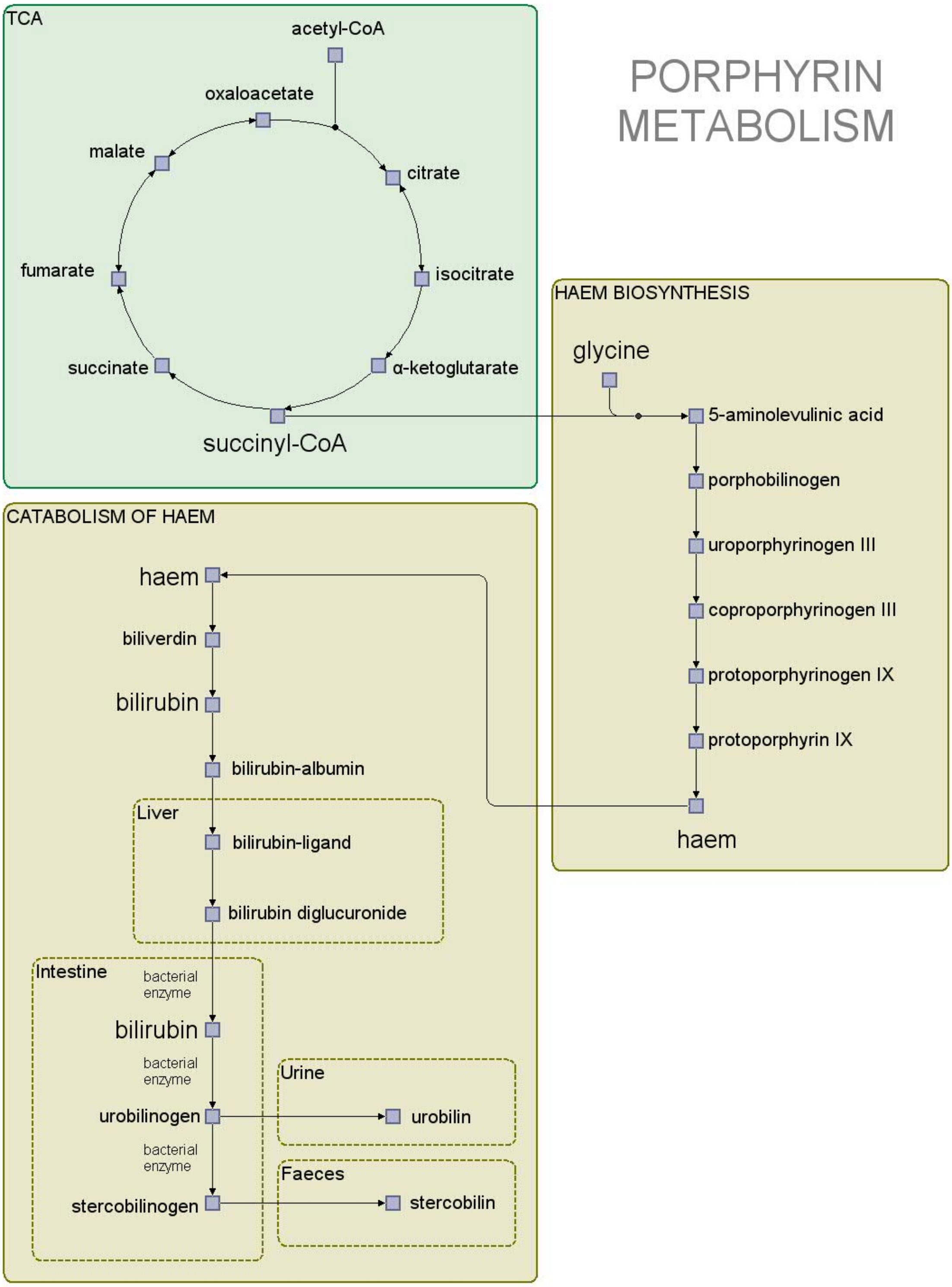


Figure 4.12. The EHMN outlines. Porphyrin metabolism. Please use the CD provided to see a high resolution version of the image.

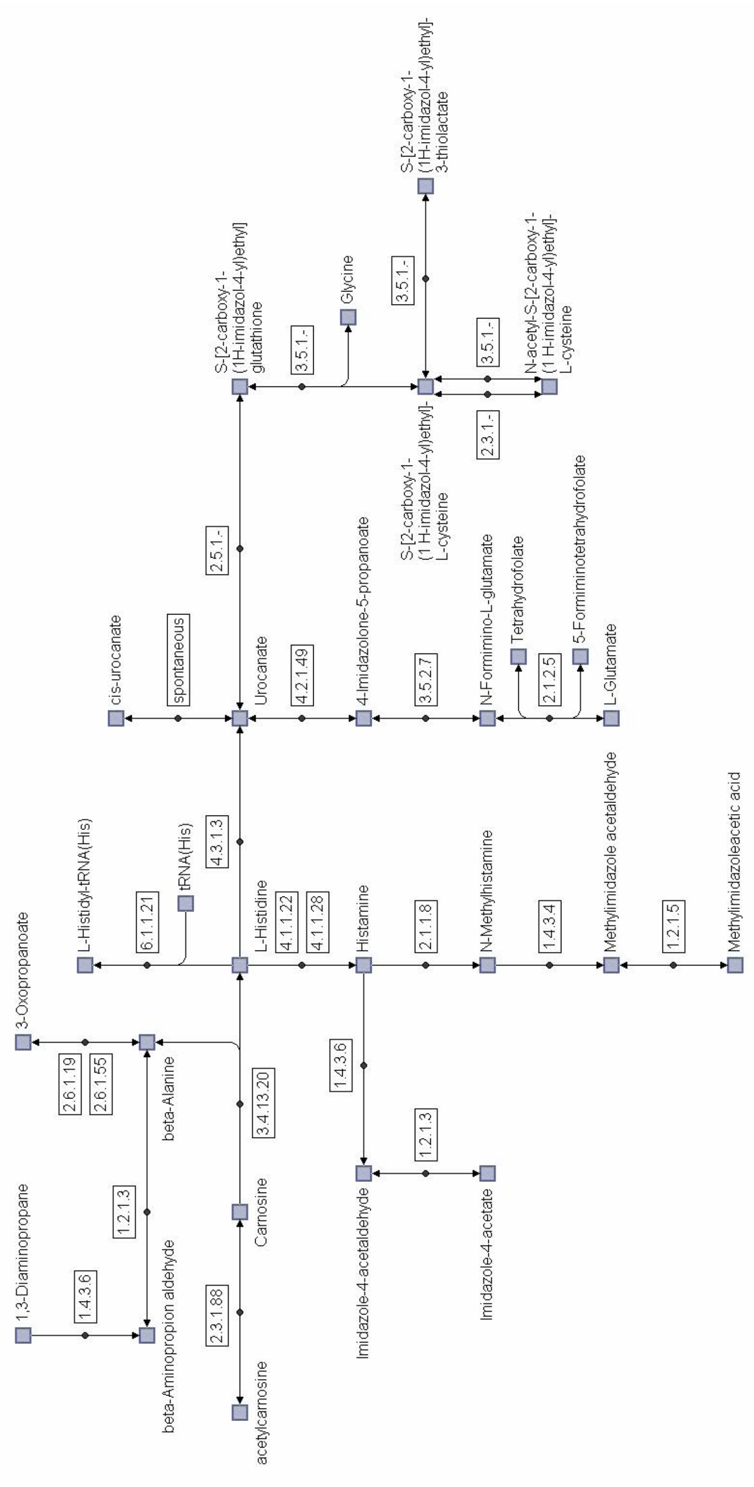


Figure 4.13 Histidine metabolism. Biological layer.

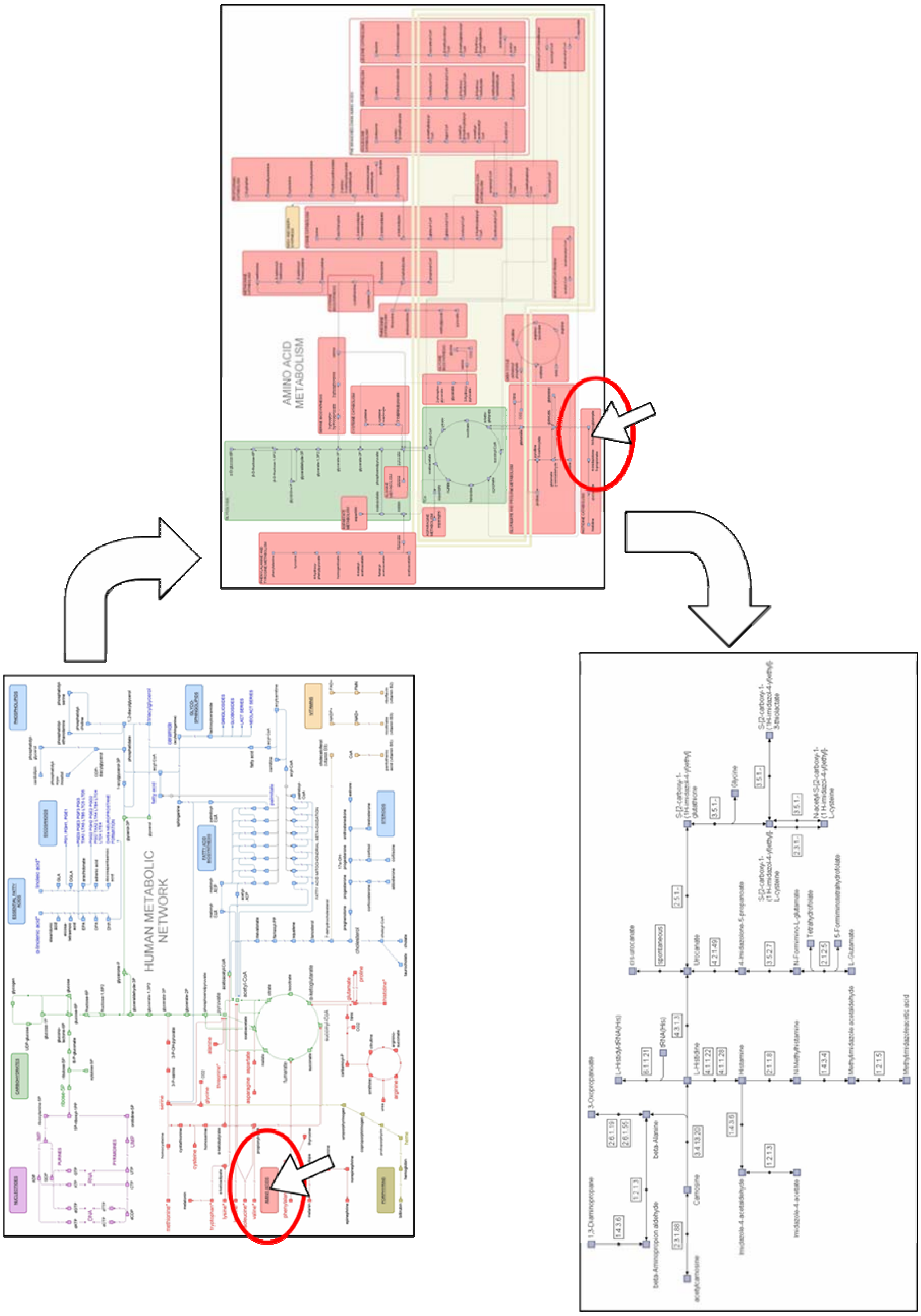


Figure 4.14 Navigation through the network. Navigation is possible due to hyperlinks that lead from an outline to a lower level outline and from a lower level outline to a detailed map.

4.3. Informal description of proposed method for semi-automatic outline generation: metabolic network.

Before the algorithm is described, it is necessary to introduce the elements/symbols that are used on metabolic outlines. Then we can use this vocabulary in the algorithm description.

The term ‘outline’ is used as a synonym of phenomenological model (Chapter 3). We use two levels of outlines: ‘lower level outlines’ (first, more detailed level of outlines) and ‘higher level outlines’ (second, less detailed level of outlines). Potentially more levels of abstraction could be added: third, fourth levels of outlines and so on. For the purpose of this example only two levels are used. The term ‘link’ refers to a hyperlink that connects one diagram (map) to another diagram**.** Links from the higher level outlines lead only to the lower level outlines. Links from the lower level outlines lead only to the detailed diagrams (‘biological diagram’ or ‘biological map’) that correspond to biological model (Chapter 3). The terms ‘editor’ or ‘tool’ refer to any editor that is able to support such functionalities as data import, automatic layout, semi-automatic outline generation. We consider the process of creating outlines being done in a semi-automatic fashion because a list of entities (in this case compounds) for each level of outlines has to be prepared partly via automatic analysis of a network and partly manually selected by an expect.

Symbols that are used on a metabolic outline (please see Figures 4.1 and 4.10 for examples):

A. ‘Compounds’ on an outline are the same entities that are used on biological diagrams.

B. Connectors link two compounds and replace ‘process’ entity from biological map.

C. In cases when two or more reactions on a biological diagram are represented by a single step on an outline, the symbol ‘omitted reaction’ of the UN is used on an outline. This way even on an outline it is clear if two compounds have one or more reaction between them.

D. ‘Pathway container’ is a rectangular shape with a map name in it. ‘Pathway container’ is used as a hyperlink (‘link’). In theory any element on a map can be used as a hyperlink including compounds and connectors. In this example only ‘pathway container’ element is used for that. The lower level of outlines shows compounds using ‘pathway container’ as a background. In that case each ‘pathway container’ includes only the compound that belongs to the pathway that the ‘pathway container’ refers to. The higher level of outlines shows ‘pathway container’ near the metabolites that correspond to the pathway that this ‘pathway container’ refers to.

The algorithm for outline generation description follows below. There are 3 main parts: biological pathways visualisation (steps 1-3), the lower level outlines visualisation (steps 4-6) and the higher level outlines visualisation (step 7).

1. Biological pathway visualisation. Step 1. A set of reaction of a metabolic network has to be defined and stored in excel file, SBML file or any other format that could be imported by a tool that enables automatic layout.

2. Biological pathway visualisation. Step 2. The reactions have to be subdivided into comparatively small pathways according to the part of metabolism they belong to. Existing reaction distribution in pathways can be used as an example, for example Kyoto Encyclopaedia of Genes and Genomes (KEGG, www.genome.jp/kegg) and the Edinburgh human metabolic network (Ma, 2007). The number of reaction in each pathway should not exceed approximately 100 reactions so the diagram could be kept at reasonable printer-friendly size. One reaction can be assigned to several metabolic pathways if necessary.

3. Biological pathway visualisation. Step 3.The reaction in excel, SBML or any other appropriate format should be uploaded into an editor that enables automatic layout. The pathways has to be visualized and automatically/semi-automatically layed out. Alternatively, the pathways could be visualised manually.

4. The lower level of outlines. Step 1. The pathways have to be subdivided into

several groups according to the major area of metabolism they belong to. One pathway can be assigned to several groups if necessary. In other words, the reactions that have been assigned to the pathways (biological maps) and then to larger groups (correspond to the set of the lower level of outlines). The outlines on the lower level outlines correspond to these larger groups such as Carbohydrate metabolism, Amino acid metabolism, Fatty acid metabolism etc.

5. The lower level of outlines. Step 2. A list of compounds that is going to be shown on the lower level outlines has to be prepared. The compounds that are proposed to be included in the lower level outlines:

- inputs and outputs of the pathway;
- ‘important’/textbook compounds (all the amino acids, all the fatty acids, all the glycolysis pathway compounds, vitamins etc);
- compounds that are in cross-talk between several pathway ‘roads’;
- compounds that appear on several pathways.

6. The lower level of outlines. Step 3. Ideally, a special plugin should be used so the tool could import the list of compounds and using information on how the compounds are linked on the maps to automatically generate an outline diagram. Pathway containers that symbolise the detailed pathways should be shown automatically and corresponding metabolites and links between them have to be inside those containers. The containers are allowed to be overlapping. Pathway containers should be linked to corresponding pathways. All of the pathways that belong to each particular larger group have to be presented on an outline that corresponds to this group by at least one of pathway containers. Additional manual correction should be made if necessary.

7. The higher level of outlines. Similar approach is used to create the higher level of outlines. A list of compounds for the outline has to be generated. The main rule is that all of the compounds that appear on the second (less detailed) level outlines have to be from the list of the compounds selected for the lower level outlines. In other words, gradual hiding the details has to be performed when moving from one level (more detailed) to another (less detailed). Moving from a less detailed overview diagram to a more detailed one a user should be able to see the same elements (compounds) plus additional information (more compounds and links between them).

Chapter 5

Signalling network representation

5.1. Macrophage activation and TNFalpha signalling network

After successful application of the Network Representation Framework (NRF) to the Edinburgh human metabolic network, the NRF has been also applied to a signal transduction network. This section describes macrophage signalling network represented using the NRF. The resulting representation consists of detailed diagrams and outlines that are united into a single system similarly to the previous example from Chapter 4.

It is not easy to decide how signal transduction events could be subdivided into separate outlines as cellular signalling pathways are so intensively interconnected. Despite this complexity it is possible to mark out several subnetworks according to the most important protein(s) involved in each case. Figure 5.1, the highest level outline (PHENOMENOLOGICAL LAYER, LEVEL 3) for the macrophage signalling network shows how macrophage is involved into cytokine network. The Macrophage signalling network, shown in Figure 5.2 (PHENOMENOLOGICAL LAYER, LEVEL 2) consists of several subnetworks (according to the receptor involved) such as the

TLR, TNF-alpha signalling and IFN-gamma signalling pathways. These in turn can be represented in a second biological level of outlines. Figure 5.3 (PHENOMENOLOGICAL LAYER, LEVEL 1) shows a phenomenological diagram of TNF signalling that unites comparatively small biological diagrams (Figures 5.4 - 5.7) comprising the BIOLOGICAL LAYER into a large virtual map. List of references for TNF signalling network maps is provided in the references section.

During the preparation of different signal transduction diagrams, including reproducing the Toll-like receptor diagram (Figure 2.4) (Oda and Kitano, 2006) in the UN, it was noticed that a particular signalling pathway contains a limited number of proteins that are involved in a great number of processes as different states of the same protein or as part of complexes. In order to make it easier to deal with signalling diagrams in EPE special search plugin has been developed by Richard Adams. The basis of the plugin is the fact that if the names of particular proteins/compound/genes are specified, all the pathway events could be marked out by using 'one-step' algorithm, which means that together with all the entities that contain specified names all the other entities that are connected to these group via the 'process', 'activation link', 'logical gates' or 'functional gates' have to be marked out too.

Another rule in representing signalling pathways in comparison to metabolic pathways is to avoid using functional gates (EC numbers). In metabolic pathways, we routinely show EC numbers near reactions. On a signalling pathway however, this is not so useful. Since phosphorylation is the most common way to change protein/complex state, showing EC numbers for signalling event makes the diagram unnecessarily complex. In our example diagrams (Figures 5.4–5.7), between 61% and 100% of reactions are phosphorylation. Therefore in signalling pathways we display the EC number only for relationships of type 'enzyme - metabolic reaction'.

The modular approach to the organisation of networks has the further advantage of promoting the reuse of individual components in different pathways. For example, the NF-kB pathway can be used in many other cases besides TNF signalling in macrophages.

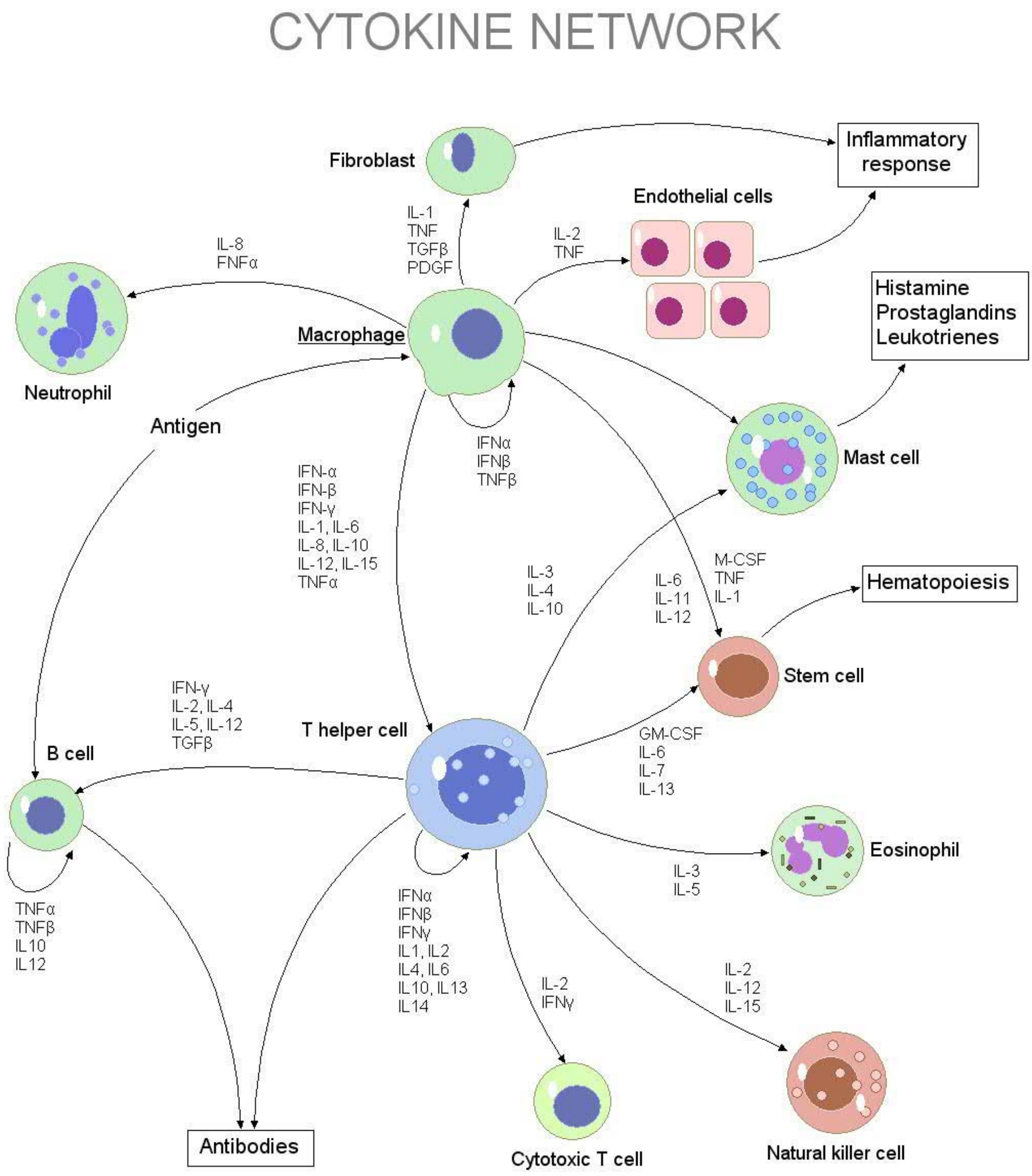


Figure 5.1. Cytokine network. PHENOMENOLOGICAL LAYER, LEVEL 3. Please use the CD provided to see a high resolution version of the image.

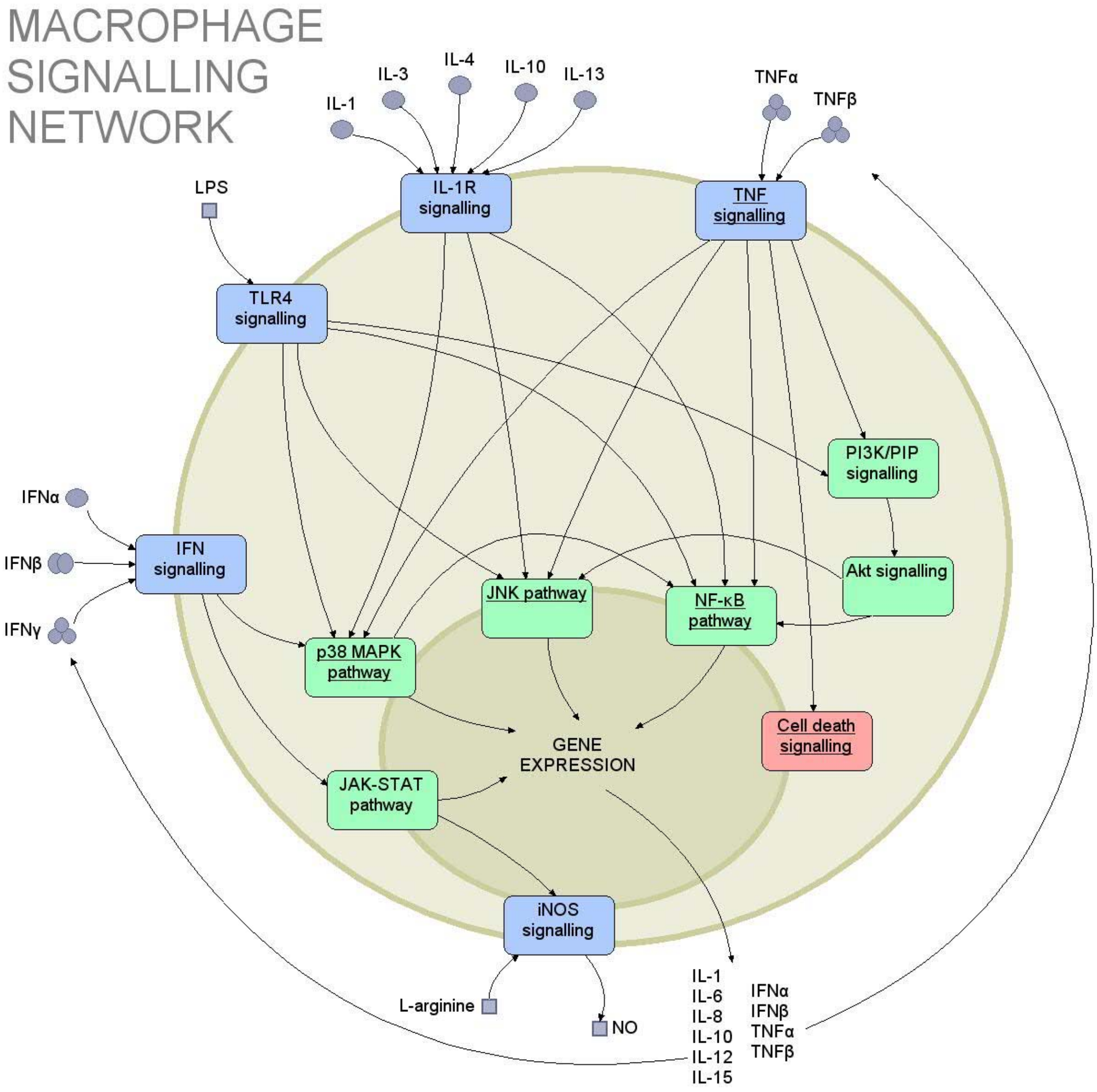


Figure 5.2. Macrophage signalling network. PHENOMENOLOGICAL LAYER, LEVEL 2. Please use the CD provided to see a high resolution version of the image.

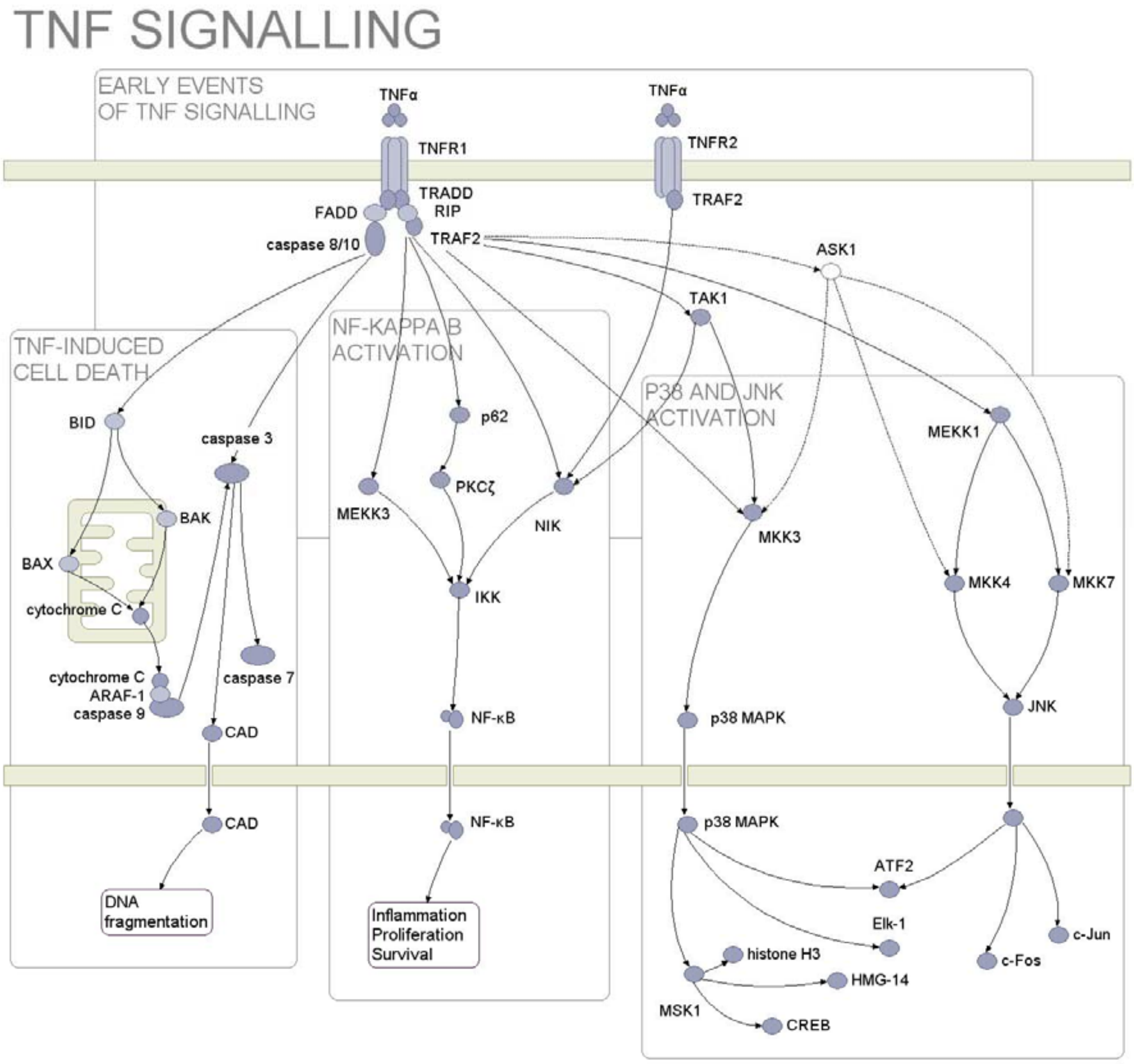


Figure 5.3. TNF signalling network. PHENOMENOLOGICAL LAYER, LEVEL 1. Please use the CD provided to see a high resolution version of the image.

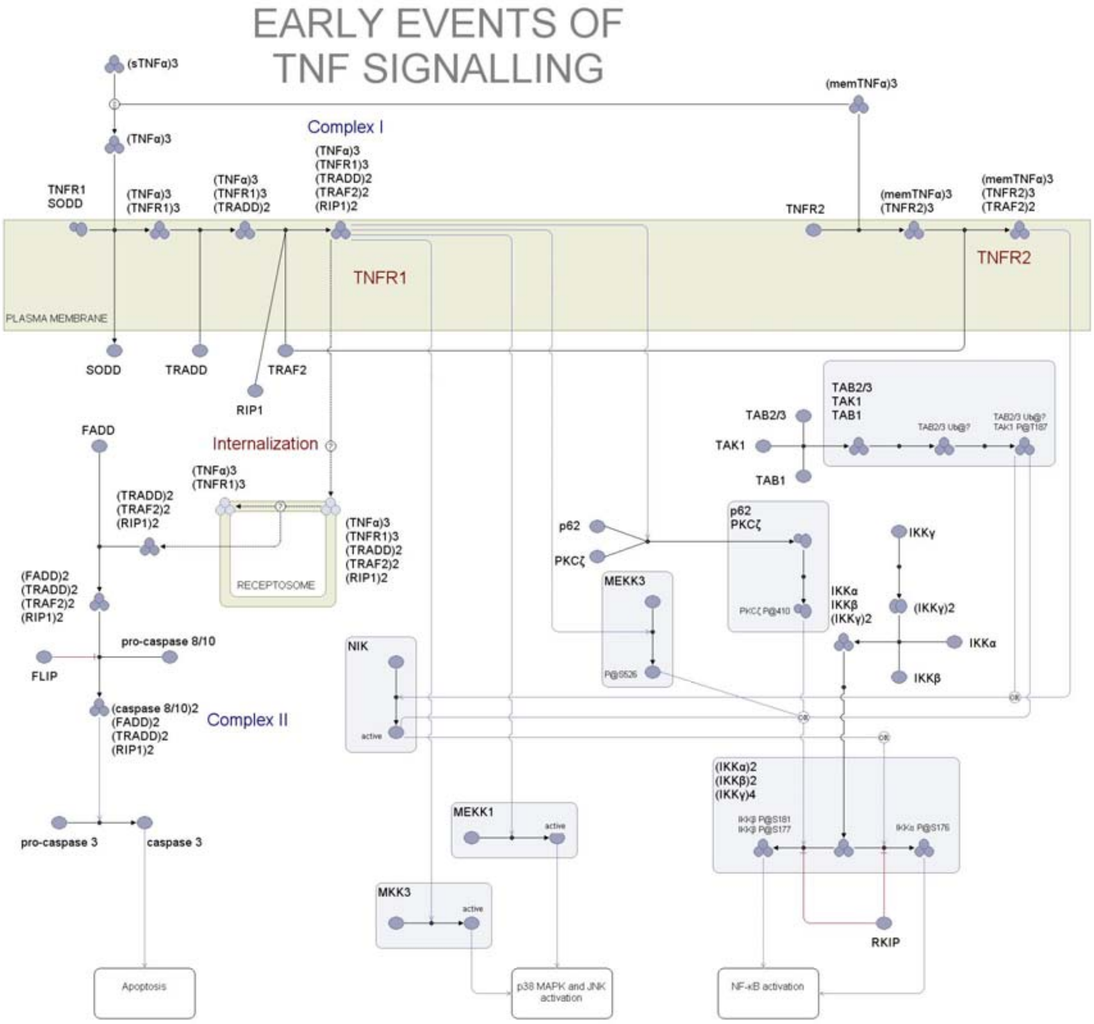


Figure 5.4. TNF receptor signalling network. Early events. BIOLOGICAL LAYER. Please use the CD provided to see a high resolution version of the image.

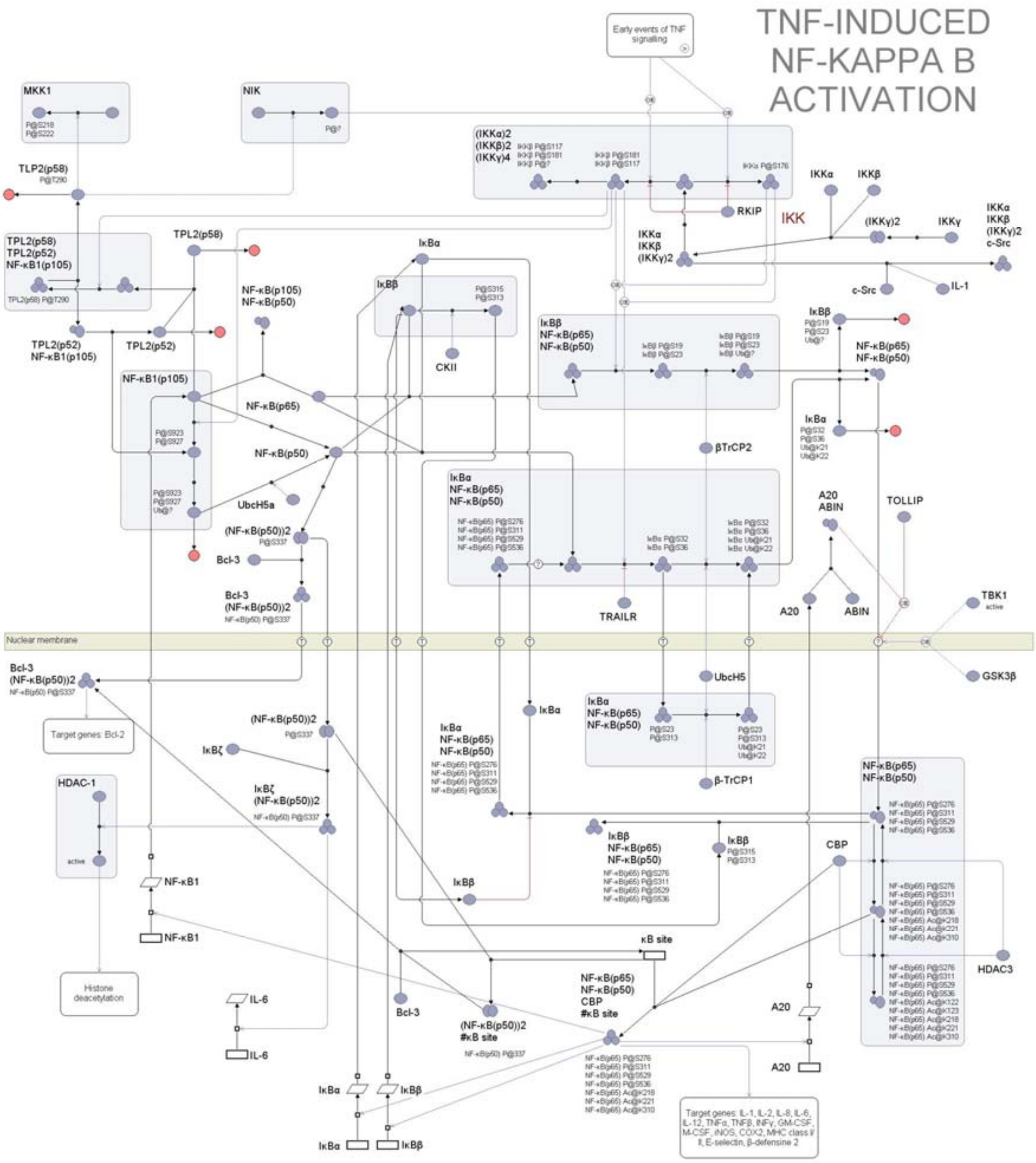


Figure 5.5 TNF receptor signalling network. NF-kappa B activation. BIOLOGICAL LAYER. Please use the CD provided to see a high resolution version of the image.

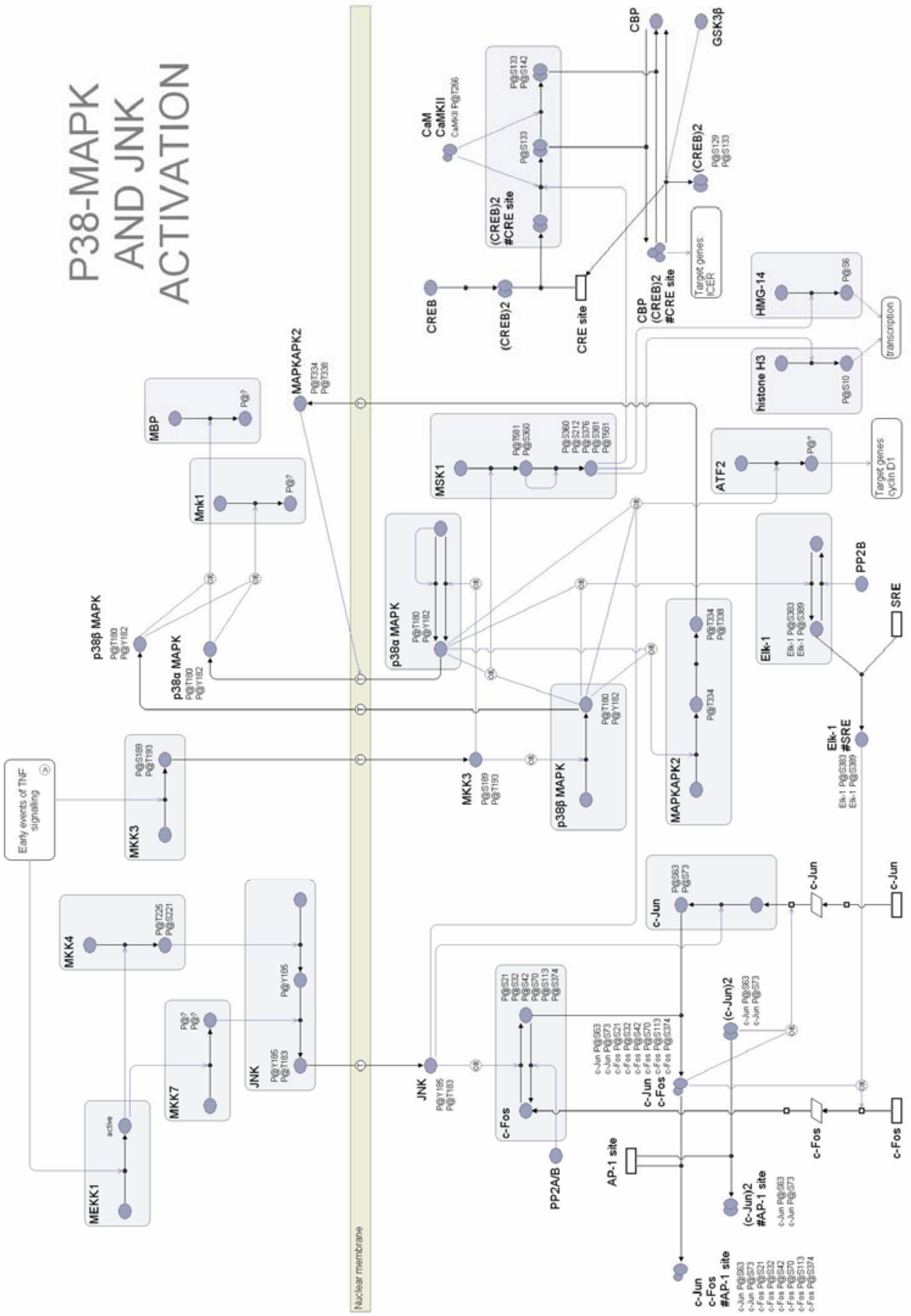


Figure 5.6 TNF receptor signalling network. P38-MAPK and JNK activation. BIOLOGICAL LAYER. Please use the CD provided to see a high resolution version of the image.

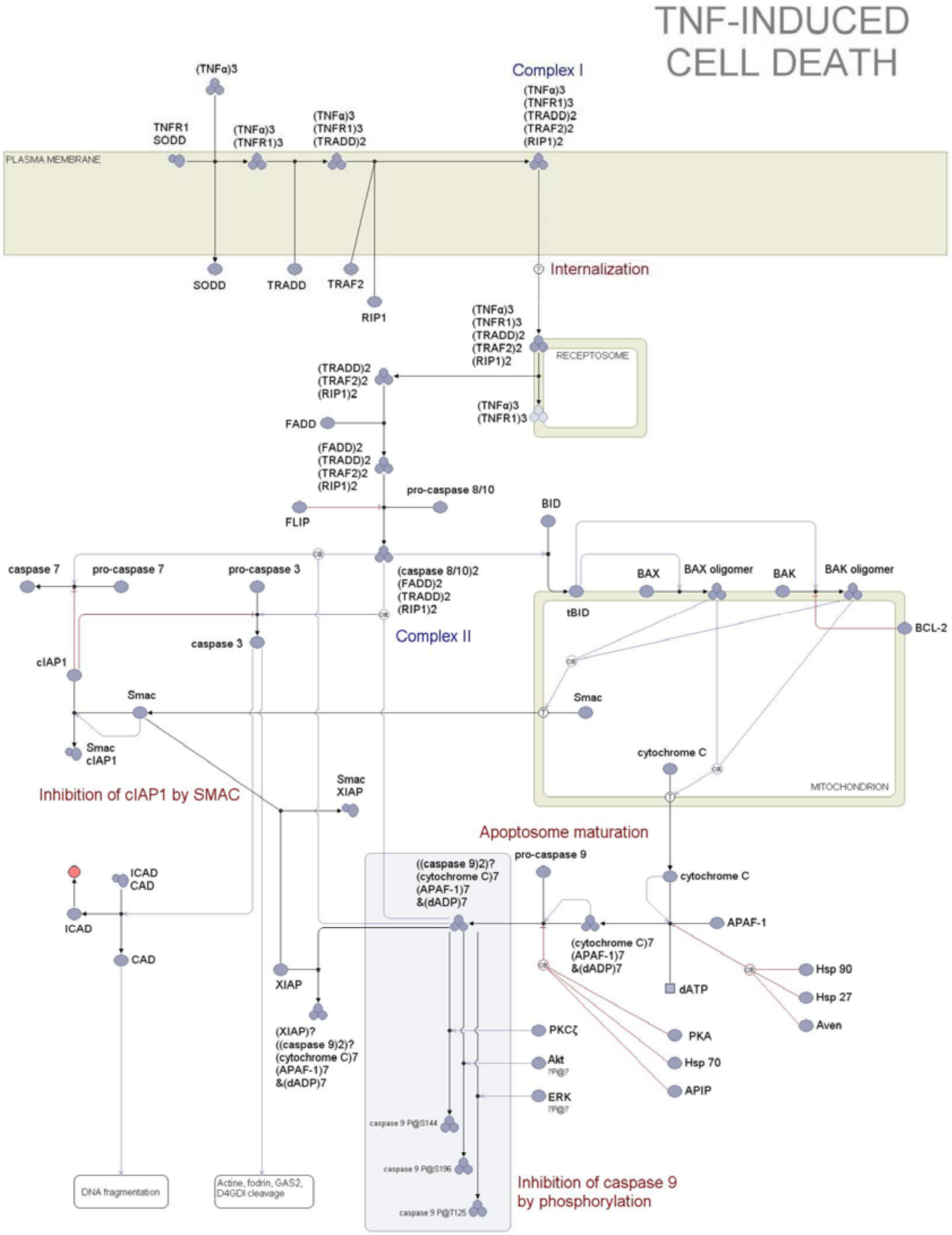


Figure 5.7 TNF receptor signalling network. Cell death. BIOLOGICAL LAYER. Please use the CD provided to see a high resolution version of the image.

5.2. Informal description of proposed method for semi-automatic outline generation: signalling and gene expression events.

The process of semi-automatic outline generation proposed here is similar to the one described in section 4.3 for the metabolic network.

In order to adequately represent the signalling events it was necessary to introduce the following additional elements: ‘macromolecule’ (correspond to ‘protein’ or ‘complex’ on biological diagram), ‘compound’ (correspond to ‘compound’ on biological diagram), ‘RNA’ (correspond to ‘RNA’ on biological diagram) and gene (correspond to ‘gene’ on biological diagram).

1. Preparing the lower level of outlines. Step 1. The lower level of outlines represents signalling network as protein-protein interactions. For example, on a detailed (biological) diagram (Figure 5.6) unphosphorylated MKK4 is being phosphorylated and becomes a new entity – phosphorylated MKK4. MEKK1 activates this process. 7 different objects represent this event on a detailed diagram: MEKK1, unphosphorylated MKK4, phosphorylated MKK4, process glyph, process consumption connector, process production connector and activation connector. On an outline (Figure 5.3) this step is represented as only 3 objects: MEKK, MKK4 and one connector. This way biological diagram can be simplified into protein-protein interaction graph. A list of proposed entities for the outlines can be generated automatically. The entities that have the same name but different state will be represented as one entity on an outline. In the example above unphosphorylated MKK4 and phosphorylated MKK4 are split into one entity – MKK4.

2. Preparing the lower level of outlines. Step 2. On this type of signalling outline a single event is presented as a link between two macromolecules. On this level of

abstraction it is irrelevant if any of the proteins are actually part of a complex or if there are several states of this protein exist.

In order to generate an outline from a detailed diagram it is advised to mark the macromolecules (proteins, genes, RNAs) that form the main activity flow on the detailed map. The list of entities generated during step 1 can be used to simplify this process. Then these entities and the relations between them can be picked up by a designed for that tool and shown as an outline. Additional manual layout might be required at the final stage.

Similarly to the proposed metabolic outline generation process (Section 4.3), pathway containers can be used to represent the pathway and overlapping is allowed in this case too (for example, Figure 5.3).

3. Preparing the higher level of outlines. Each lower level outline should be represented by a single shape with corresponding map name (‘pathway container’), compounds/proteins/genes/RNAs are shown only as pathway inputs and outputs and are linked directly to ‘pathway container’ shape and not to other compounds/proteins/genes/RNAs. Connections between different pathways and pathway inputs/outputs can be derived automatically based on the information about shared elements (compounds/proteins/genes/RNAs or processes) (for example, Figure 5.2).

Recently a specification for SBGN Activity Flow language (SBGN AF) was published (http://www.sbgn.org/Documents/Specifications). We are considering the possibility to enable NRF to include SBGN PD and SBGN AF languages. In this case SBGN PD will represent the biological layer and SBGN AF will be used as lower level outlines. For now it is not clear if automatic translation of SBGN PD language into SBGN AF language would be possible.

5.3. Metabolism regulation: linking signalling and metabolic pathways

Many signalling pathways control metabolic fluxes. On the other hand, the inverse control (from metabolic state to signalling pathway) plays a critical role in many cellular processes, for instance in the calcium signalling and the inositol signalling (Oda and Kitano, 2006).

However, most metabolic and signalling databases provide a set of maps that focus on signalling or metabolic events only. Cases for simultaneous visualisation of both hardly exist (e.g. AMPK signalling and Insulin Receptor Signalling (Cell Signalling Technology Pathways, http://www.cellsignal.com/pathways/glucose-metabolism.jsp)). In order to unite the metabolic and signalling networks into a single cellular network system, we decided to generate a set of combined maps. The Unified Notation allowed us to create the mixed maps and avoid the conflict between different styles of metabolic and signalling pathways' visual representation.

The use of EC numbers as functional gates in these maps ensures an adequate and unambiguous representation even in the cases where visualisation is usually confusing. For example, Figure 10 demonstrates the regulation of glycolysis. In this pathway, phosphofructokinase/fructose-2,6-bisphosphatase is a protein that has two different activities depending on its phosphorylation state. The phosphorylated form of the protein acts as a fructose-2,6-bisphosphatase (EC 3.1.3.46), while the non-phosphorylated form acts as a phosphofructokinase (EC 2.7.1.105). The diagram clearly reflects the multifunctional nature of this protein.

Figure 5.8 shows regulation of glycolysis by glucagon via the phosphofructokinase 2/ fructose-2,6-bisphosphatase (PFK2/F2,6BFase).

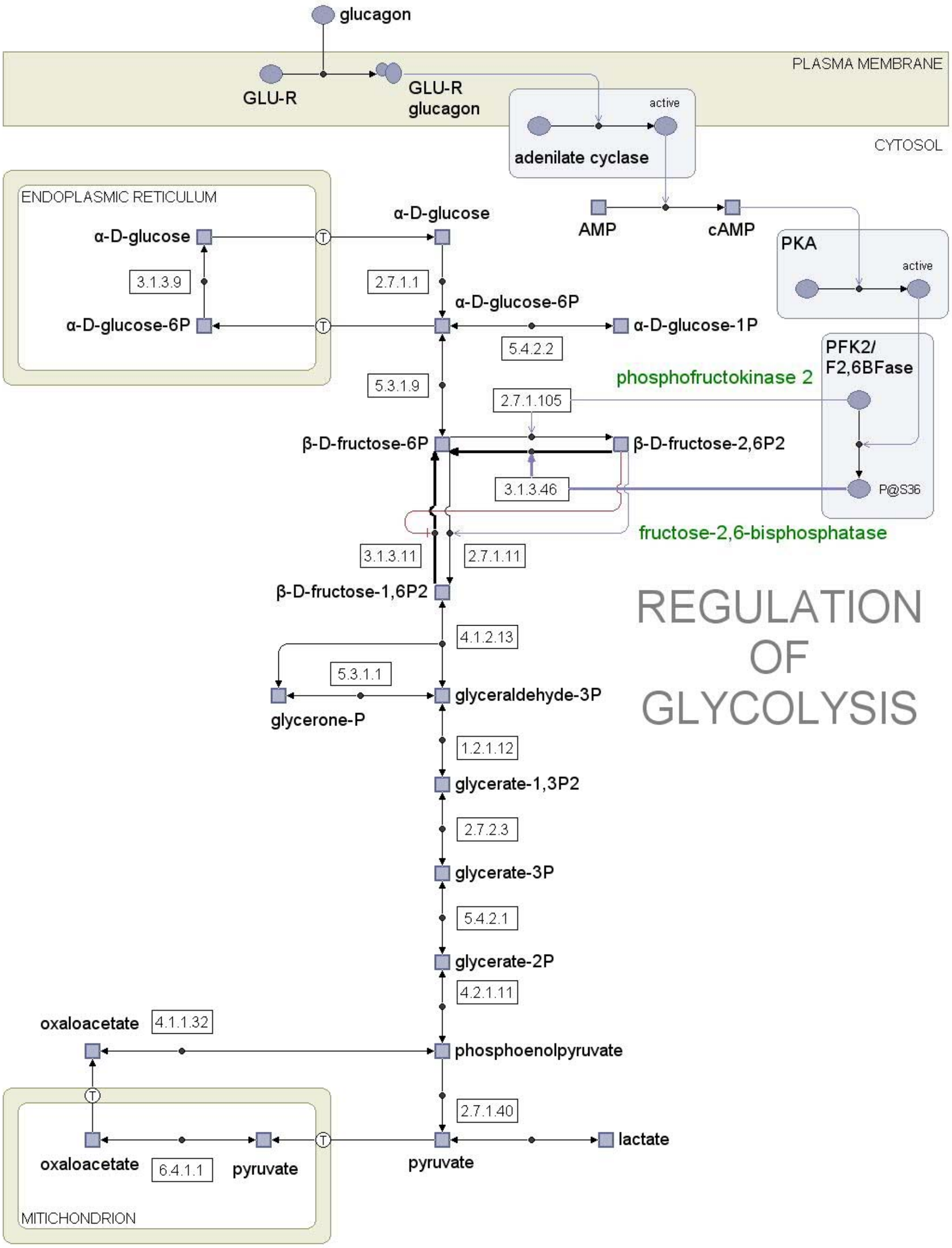


Figure 5.8. Regulation of glycolysis by glucagon via phosphofructokinase 2 (PFK2) / fructose-2,6-bisphosphatase (F2,6BPase). Please use the CD provided to see a high resolution version of the image.

5.4. Additional information visualisation. A disease influence on a biological network

By adding certain minor changes to basic diagrams we are able to visualise information about changes in biological system related to a disease or other influences. To exemplify the result the Cancer Genome Atlas project (Chin, 2008) (Pathway analysis of genetic alterations in glioblastoma, http://cbio.mskcc.org/cancergenomics/gbm/pathways/) has been used, specifically the signalling pathway alterations in glyoblastoma based on mutations and copy number changes in 91 samples (Supplementary information, http://www.nature.com/nature/journal/v455/n7216/suppinfo/nature07385.html). Figure 5.9 demonstrates how the authors' (Chin, 2008) original data visualisation could be shown using the symbols of the UN. The data is mapped by adding numeric expressions or/and using a designated colour scheme.

Similar representation could be applied to other notations, for example to the SBGN PD notation. Corresponding diagrams with SBGN PD notation symbols have been produced to illustrate this (Figures 5.10). The 'identity gate' symbol has been added to the SBGN PD set of symbols to enable the notation to visualise this kind of information.

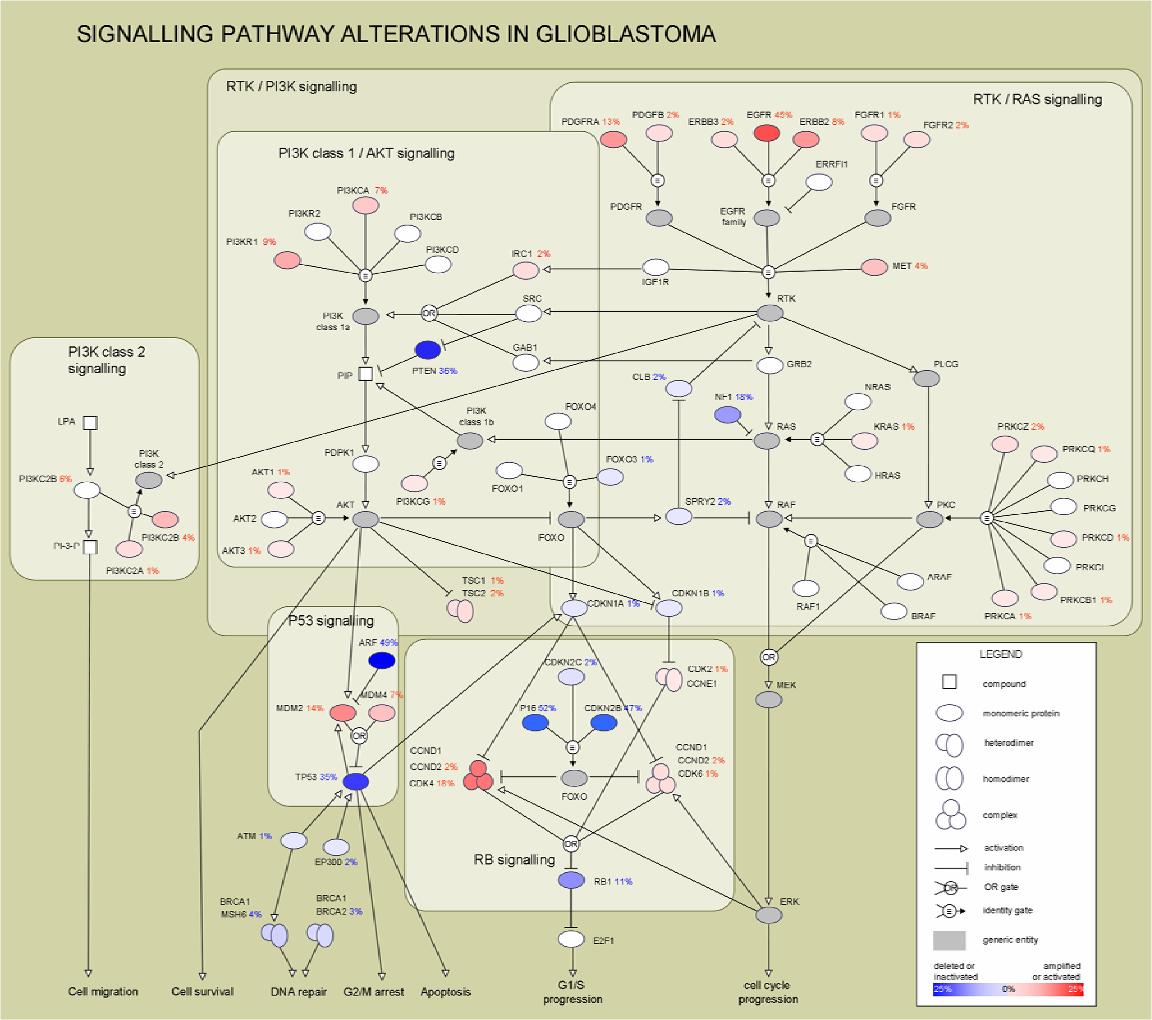


Figure 5.9. Signalling pathway alterations in Glyoblastoma. PHENOMENOLOGICAL (OUTLINE). The UN symbols are used on the outline. Please use the CD provided to see a high resolution version of the image.

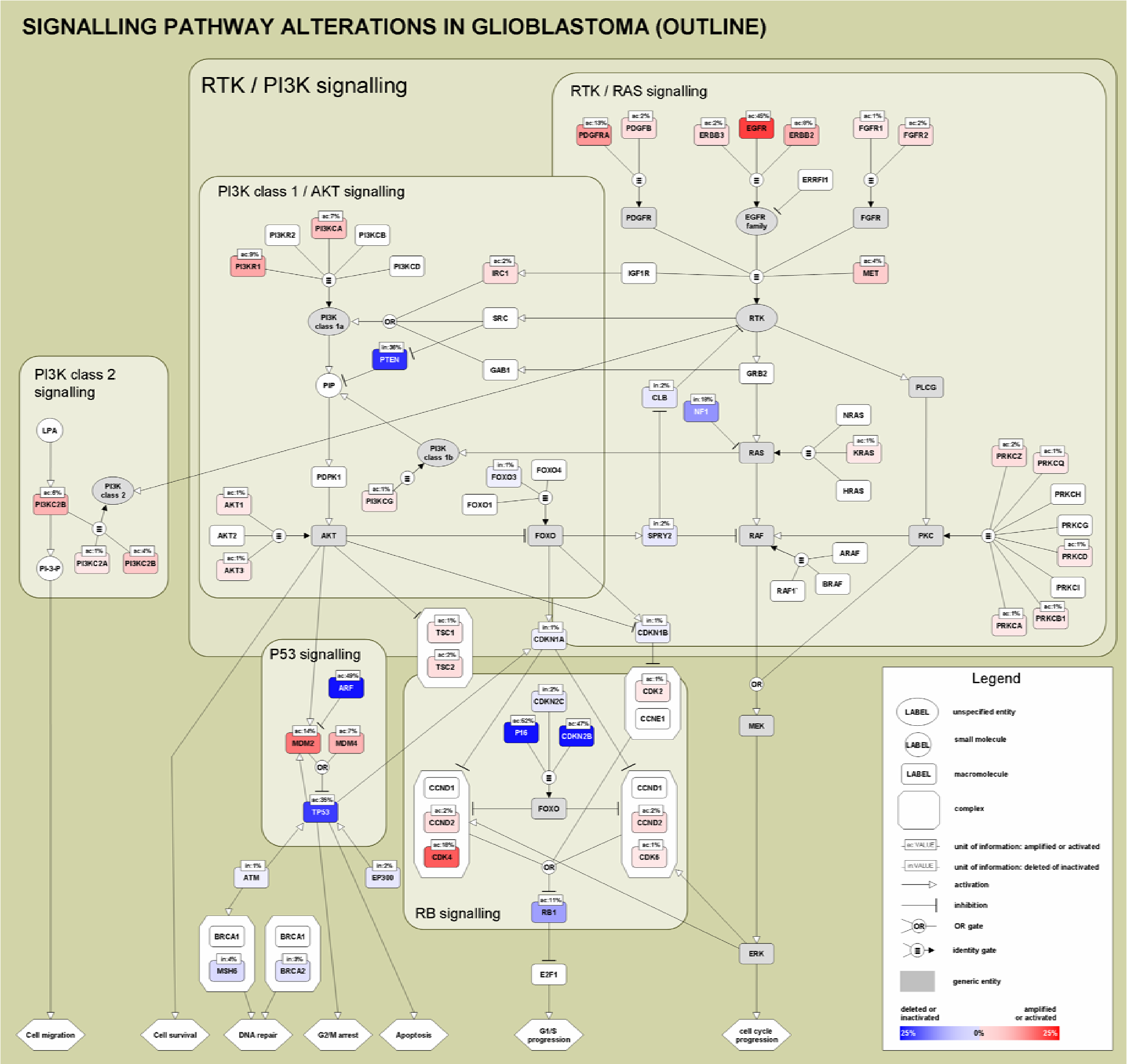


Figure 5.10. Signalling pathway alterations in Glyoblastoma. PHENOMENOLOGICAL (OUTLINE). The SBGN PD symbols are used on the outline. Please use the CD provided to see a high resolution version of the image.

Chapter 6

Graphical notation survey

6.1. The objectives of the survey

The main objective of the survey is to show if the notation used in this research has any advantages compared to other similar notation used in the field. At the same time it would be interesting to compare different notations and see if some features are more important for the users than others. Such output could be used later for improving graphical notations for biological networks representation.

The features of the notations to be evaluated in the course of the survey were:

1) adequacy for metabolism representation,

2) adequacy for enzyme function,

3) adequacy for metabolism regulation representation,

4) adequacy for signalling representation,

5) readability (how it is easy to read/understand information presented on the diagram?)

6) compactness,

7) adequacy for complex representation,

8) adequacy for gene expression representation,

9) the number of symbols used (is it sufficient? too many symbols?),

10) ability to show incomplete and omitted information,

11) how it is easy to learn the system of symbols.

Because it was not the priority of this research to evaluate a particular graphical editor and taking into account that the additional notations chosen for the survey are not supported by EPE, it was decided to prepare a paper version of the survey rather than EPE-based version.

The survey had to evaluate the merits of the several notations. The most natural way to do so was to prepare several pathway diagrams where the same information would be presented in different graphical languages for comparative analysis. Because the aim of the survey was to receive a feedback on different aspects of the notations, the questionnaire should have included examples of different types of biological pathways.

6.2 The choice of the notations for the survey

In this survey, three different notations were used: the SBGN PD language (Le Novere, 2009), the mEPN (Freeman, 2010) and the UN (Mazein, in press). The choice of the notations to which UN would be compared to was justified according to the following criteria:

1) a notation is developed for describing processes on detailed level;

2) a notation has to be unambiguous;

3) a notation has to be well-described and some examples have to be available;

4) the visual languages should be visually easily distinguishable (use different set of shapes and ways to represent information);

It has been decided compare the UN with only two more notations so it would not be too difficult to compare them and would not require too much time to complete the questionnaire.

On the basis of the first criterion, for example, MIM notation (Kohn, 2001; Kohn, 2006; Kohn, 2006) and SBGN ER language (Le Novere, 2009) were excluded from the list because they describe relationships between the entities and can not specify the order of events on a diagram. SBGN AF language also could not be used because it provides ambiguous representation without particular details (Le Novere, 2009). Process description language (Kitano, 2005) was not included because its visual elements are very similar to the ones of SBGN PD language (criterion 4).

As a result of this selection process the survey has been restricted to comparing the above three notations.

To provide enough examples for comparison several diagrams were prepared in UN and then translated into the other two languages.

6.3 The questionnaire

In order to prepare a better questionnaire the tips on conducting a survey and developing a questionnaire were taken into account (for example, Hints for Designing Effective Questionnaires, http://pareonline.net/ getvn.asp?v=5&n=3; Questionnaire Design, www.cc.gatech.edu/classes/cs6751_97_winterTopics/quest-design/; Survey and Questionnaire Design, www.statpac.com/surveys/; How to Write a Survey or Questionnaire, www.ehow.com/ how_16596_write-survey-questionnaire.html).

The final version of the questionnaire is available in Appendix II.

In order to elicit as much data as possible from the survey while making it as unobtrusive as possible to the participants, the following constraints were put in place.

The first section of the questionnaire describes the purpose of the survey and introduces the three notations that are being compared. The references for each notation are provided. Next section provides brief instructions that explain the format of the questions.

Some of the questions that could be considered leading are put at the end of the questionnaire. This is done, for example, with the question about the importance of the ability to show EC on metabolic maps.

In order to rate a notation the questions had to be as detailed as possible. On the other hand, it would not be reasonable to expect each participant be familiar with all the aspect and difficulties related to biological information visualisation or would be able to learn all three notation quickly enough so they could, for example, actually use them to draw a diagram. To overcome this problem, different types of diagrams were prepared in all three notations and a respondents were asked to rate how the same information is shown in three notations in each of the following cases: metabolic pathway representation, signalling pathway representation and so on. More difficult to answer questions are placed at the end of the questionnaire. This way the participants do not have to learn the notations but still will be able to rate their features. The detailed description of the questionnaire is provided in section 8.6.

To ensure the objectiveness and quality of the results, we restricted the survey in the following respects :.

1. All the participants of the final version of the survey were from outside the research group. More details can be found in section 6.5.

2. Developing and everyday use of different graphical languages for representing biological knowledge is a particular area of expertise. We could not expect the participants to be familiar with all the graphical languages used in the survey. At the same time, it was important to ensure that the participants can actually read and understand biological information represented on the questionnaire’s diagrams.

That is why it was decided to include a simple test that would demonstrate the participant’s ability to understand the diagrams. The questionnaires in which the test was not done correctly were excluded from the dataset.

3. To make the comparison easier and impartial, colours were excluded from the diagrams.

4. In all cases diagrams represent exactly the same biological events. If there were any differences in the information presented that relates only to an ability of a particular notation to show certain type of data.

6.4 The survey pre-test

The first version of the survey can be found on the CD enclosed. 12 students of the Systems Biology Course 2009/2010 (University of Edinburgh, School of Informatics) were offered to complete the questionnaire after they had learned how to use different graphical languages including those used in the survey.

This version of the survey consists of two different parts. The first part contains several pathways represented in the 3 notations and responders had to write their comments on each diagram. The second part offered to evaluate each notation listing their features such as readability, complex representation, metabolism representation, signalling representation, gene expression representation, number of symbols used and the ability to represent incomplete knowledge. At the end the students had to add three more features they considered to be important and rate them for each notation. The final task was to list strong points and weaknesses for each of the notations.

Later the content of this version of the survey was discussed with the students in order to see if there were any unclear or misleading questions or suggestions on how the questionnaire could be improved.

The main concern was about the fact that it generally took too long for the students to complete the questionnaire: 45-90 minutes. This was mainly because in the first part the participants were asked to provide feedback on each notation and each diagram in free form and compare as many features as possible.

During the analysis of the results of pre-survey it became clear that it is difficult to deal with the feedback which is provided in different formats for different questions. For most of the questions it was necessary to provide rating using the scale from 1 to 10 and for some questions 4 options of answers were offered and the participants had to choose one. As the answers were given in different format it would be necessary to prepare different types of excel file templates and different types of graphs for visualising the results and that could cause unnecessary complications both in analysis and representing.

On the next stage another version of the questionnaire was prepared in the way so it would take approximately 20-30 minutes to complete it.

In the final version all the ratings were done on the scale from 1 to 10, where 1 is the most negative mark and 10 is the most positive assessment of a notation feature.

Open-ended questions were excluded from the final version of the questionnaire to make it easier to analyze the results.

The survey pre-test helped to significantly improve the quality of the questionnaire and made the final version much easier to complete without noticeable decrease in information value of the survey.

.

6.5 The survey results

The participants of the survey were: M.Sc. students of biological and informatics courses at the University of Edinburgh (61 completed the questionnaire, 15 with incorrect test), Ph.D. students in biology and in bioinformatics at the University of Edinburgh (25 completed the questionnaire, 4 with incorrect tests), participants of the COmputational Modeling in BIology Network meeting in Edinburgh in October 2010 (COMBINE 2010, http://sbml.org/Events/Forums/COMBINE_2010) (18 completed the questionnaire, in all cases the test was correct) and attendants of the 11th International Conference on Systems Biology in Edinburgh in October 2010 (ISCB

2010, http://www.icsb2010.org.uk) (8 completed the questionnaire, 1 with incorrect test).

Total number of the returned completed questionnaires was 112. In 20 cases the test was incorrect and those questionnaires were excluded from the dataset. These answers are considered to be correct in the test (Appendix II):

Protein A activates protein B	FALSE
Protein B activates protein C	FALSE
Protein A activates protein C	TRUTH
Protein D activates protein C	TRUTH
Protein A inhibits protein D	FALSE or NOT SURE

The results described in this section are based on the 92 completed questionnaires with correct test.

Data from the questionnaire sheets were combined into an Excel file that can be found on the CD (\Survey\survey.results.xls). Average value and standard deviation value was calculated for each rating. For that such Excel functions as CALCIF, AVEGARE and STDEV were used. On the diagrams bellow (Figures 6.1-6.11) the average value is represented by a column and the standard deviation value is shown for each column.

A plot with a distribution of the answers for each notation has been created (Appendix III). Small extra peak around ratings 5 and 6 has been considered insignificant because ratings 5 and 6 were reserved for answer ‘not sure’ and the peak can signify the number of participant that were not sure how to answer the question.

Although it was not possible to quantify the comments, in cases when the major number of comments for a particular rating point was given toward the same issue, this issue is discussed in the section that corresponds to that rating.

The results are provided below for each of the questions.

6.5.1. Metabolism representation

The average rating for the mEPN (8.5) in category ‘metabolism representation’ is noticeably higher than the average ratings for the SBGN PD (6.8) and the UN(7.2). Since for this category only metabolites and processes are shown on the example diagram, and because the process representation is very similar in all the three cases we can conclude that mEPN provides better representation for ‘simple chemical’ entity.

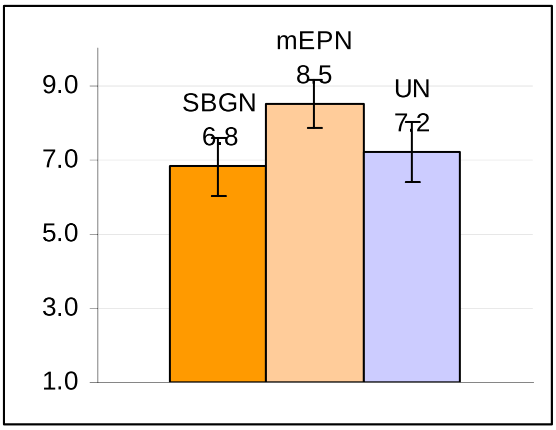


Figure 6.1. The average ratings for metabolism representation.

According to the results, the mEPN provides the best representation of metabolic pathway diagram.

6.5.2. Enzyme function representation

The UN has significantly higher rating for this feature (Figure 6.2) compared to the other two notations. We assume that this is because ECs can be shown in the UN in familiar, similar to KEGG (Kyoto Encyclopaedia of Genes and Genomes, www.genome.jp/kegg) representation. The newly introduced entity ‘functional gate’ allows one to show the EC and protein name without any difficulties simply because they are shown as different entities. In the UN an enzyme function is clearly separated form ‘protein’ of ‘complex’ entity. Using ‘functional gate’ it is easy to assign one EC to several entities, or show several EC to visualise multiple functions of a single protein or complex. The example ‘Phosphatidylinositol metabolism’ offered for comparison in the survey reflects those cases (Appendix II).

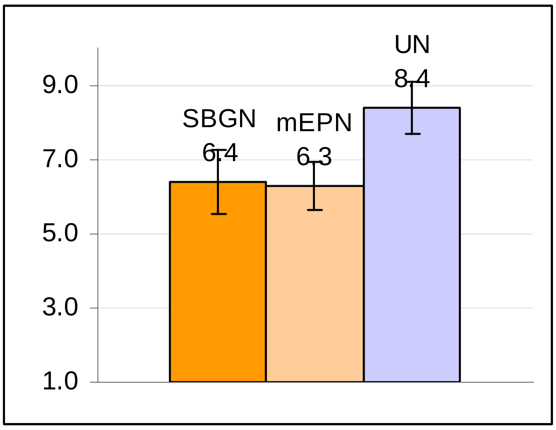


Figure 6.2. The average ratings for enzyme function representation in different notations.

6.5.3. Metabolism regulation representation

The UN has higher average rating in this category (8.4 in contrast to 6.6 and 6.4 for the other two notations) (Figure 6.3). The 'functional gate' entity offers the most consistent representation. In order to add signalling events to a metabolic diagram (see 'Glycolysis regulation' example in Appendix II) a user does not need to change metabolic diagram but only needs to add new entities to describe regulation.

For example in SBGN PD the enzyme function (EC) can be shown on a metabolic diagram as proteins but then the EC has to be replaced by the protein name when singnalling events are added. That makes enzyme function representation inconsistent. It is also impossible to show multiple enzyme functions in cases when there are several enzymological activities assigned to one protein or complex.

Similarly to SBGN PD mEPN can display EC only as an alternative name for the protein or complex (please see example diagrams in Appendix II).

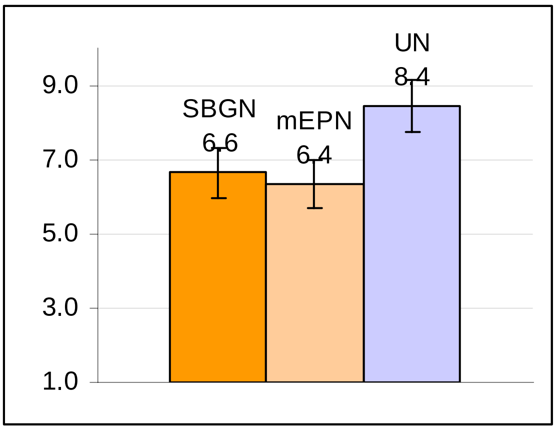


Figure 6.3. The average ratings for representation of metabolism regulation.

6.5.4. Signalling representation

The example for this category, MAPK cascade, was chosen so that it would be possible to evaluate the newly introduced ‘multistate’ entity of the UN. In contrast with our expectation most of the users gave the UN representation comparatively low rating (average rating 6.2) (Figure 6.4). According to the answers’ distribution (Appendix III, Part 4, Figure Q4C) most of the responders could not rate this feature for the UN in a definitive way. Most of them rated the UN 5 and 6 (‘not sure’). We conclude that the responders found this representation confusing and less convenient compared to the representations offered by the other two languages.

According to this survey the most convenient representation of signalling events with very high rating 9.0 was the one offered by SBGN PD language (Figure 6.2). The way the SBGN PD represents macromolecules and their states is different from the other two notations and we assume that is what makes this language successful in this category.

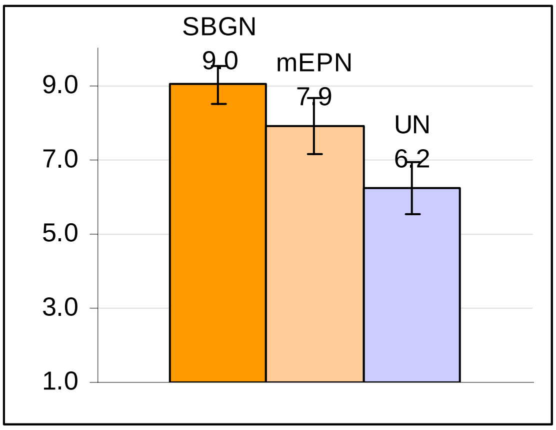


Figure 6.4. The average ratings for signalling representation.

6.7.5. Readability

The SBGN PD provides the best readability according to the Figure 6.5. Comparatively low average value for the mEPN probably is due to the number of different types of processes that are reserved in this notation (please see the mEPN reference card in Appendix II).

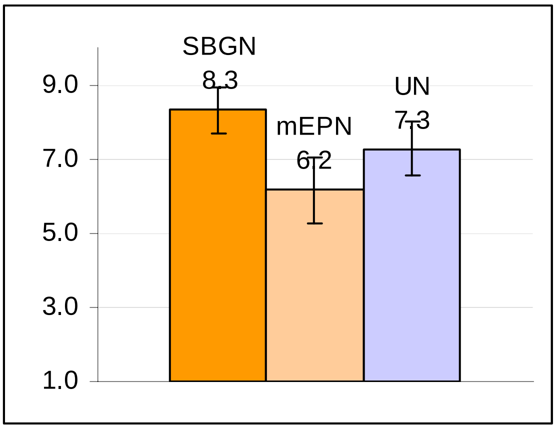


Figure 6.5. The average ratings for readability.

6.5.6. Compactness

None of the language has been singled out in this category (Figure 6.6) That is either because the notations offer similar level of compactness, or the size of the diagrams offered for comparison for evaluating this feature was not large enough for the differences to be considered distinctive.

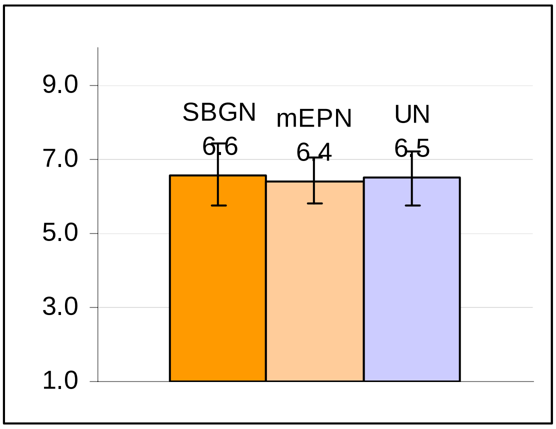


Figure 6.6. The average ratings for compactness.

6.5.7. Complex representation

The SBGN PD language was chosen as the most convenient notation for complex representation (Figure 6.7). All three language show rating 6.8 and higher which signifies that all of them offer sufficient representation capabilities in this category.

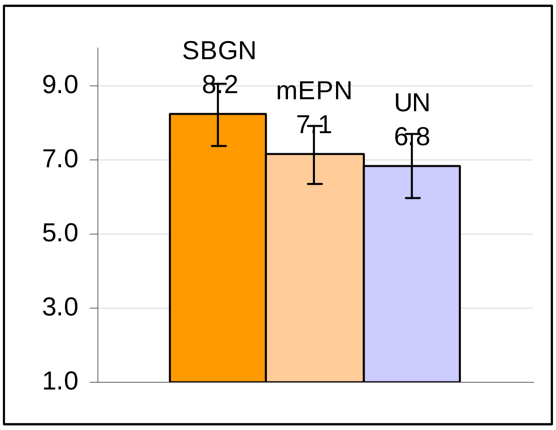


Figure 6.7. The average ratings for complex representation.

6.5.8. Gene expression representation

The result for gene expression show that all the three notations have comparatively similar ability for successful representation of gene expression events (average rating was more than 7 for all of them) (Figure 6.8).

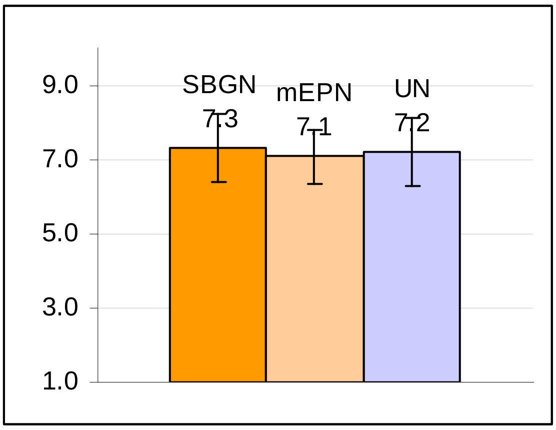


Figure 6.8. The average ratings for gene expression representation.

6.5.9. Number of symbols

While the SBGN PD and the UN were evaluated as notations with sufficient number of symbols, the mEPN was marked out as a notation 'with too many or too few' number of symbols (average rating lower than 5) (Figure 6.9).

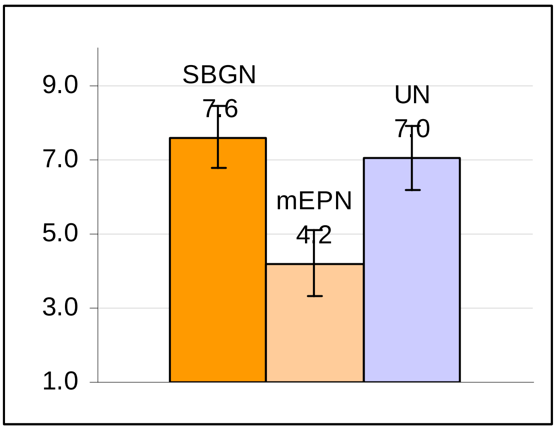


Figure 6.9. The average ratings for the number of symbols.

6.5.10. Ability to represent incomplete or omitted information

The results in this category are inconclusive. None of the notations have been singled out (Figure 6.10). The average rating is higher that 5. It is interesting that a significant number of responders found it difficult to answer this question definitively (see Part Q10 in Appendix III).

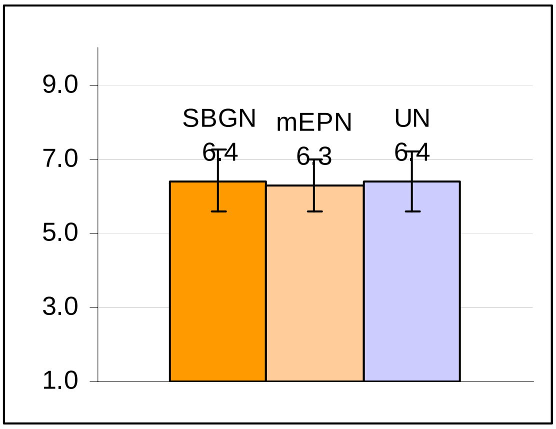


Figure 6.10. The average ratings for the ability to represent incomplete/omitted information.

8.7.11. Is it easy to learn/remember the system of symbols?

The users described the UN as the most easy-to-learn language and the mEPE as the most difficult to learn language (average rating is lower than 5) (Figure 6.11).

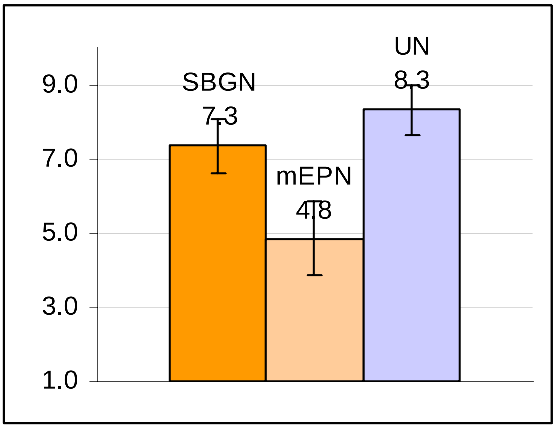


Figure 6.11. The average ratings of how it is easy to learn a notation

6.5.12. Is compactness an important feature of a graphical language?

A significant number if responders did not find the ‘compactness’ feature to be important, 25% were not sure and 58% believed this feature is important (Figure 6.12). According to several responders’ comments, even despite it might be useful to be able to represent a network in compact way, this feature can not be treated as the most important one.

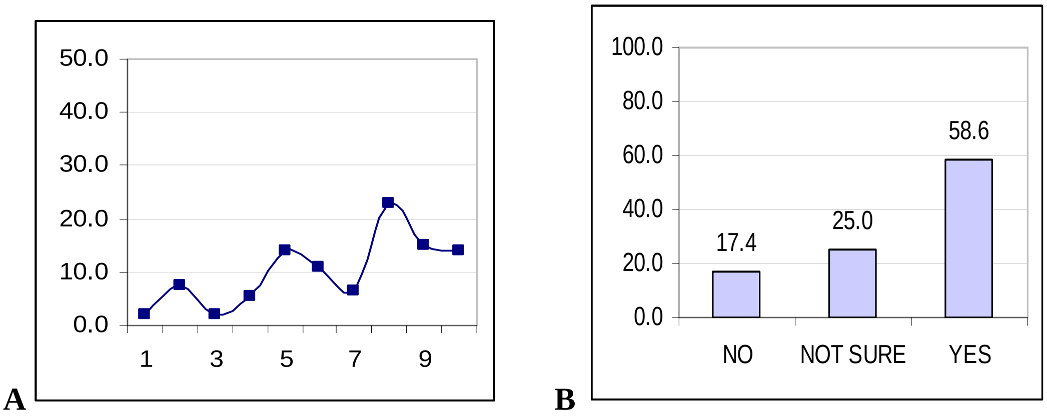


Figure 6.12. A. The distribution of the answers for the question about the importance of the compactness feature. B. The discrete distribution of the answers for the question about the importance of the compactness feature. Ratings from 1 to 4 are aggregated together under the first column (NO), ratings 5 and 6 are – under the second column (NOT SURE) and ratings from 7 to 10 – were united under the third column (YES).

6.5.13. Is it important to be able to show the ECs on a metabolic diagram?

67.5% of the responders have answered to this question positively. A significant number of answers falls into ‘not sure’ category (Figure 6.13). According to their comments, some of the respondents want to be able to show not only the EC but also the enzyme name or corresponding gene name. Several comments pointed out that it is not always easy to determine a EC number even though the question was about the ability to show EC which not necessarily meant that a user has to determine a EC in each case. Probably the questions should have made this aspect more clear.

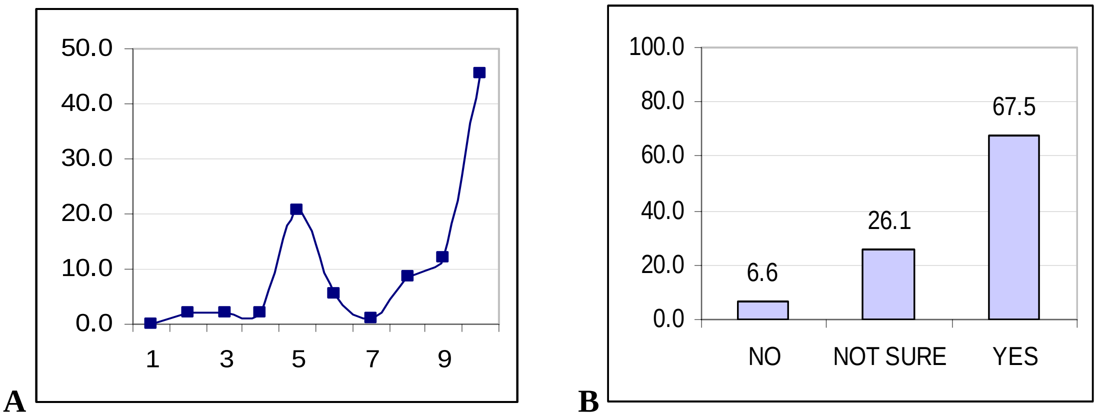


Figure 6.13. A. The distribution of the answers for the question about the importance of the ability to show EC on a metabolic diagram. B. The discrete distribution of the answers. Ratings from 1 to 4 are united and represented by the first column (NO), ratings 5 and 6 are – by the second column (NOT SURE) and ratings from 7 to 10 – by the third column (YES).

6.5.14. Is it important to be able to represent generic-specific relationships on a diagram?

57.6% of the participants think it is important to be able to represent generic-specific relationships on a diagram. 39.1 % were not sure how to answer (Figure 6.14), probably because they were not familiar with this problem and were not sure about the importance of this issue.

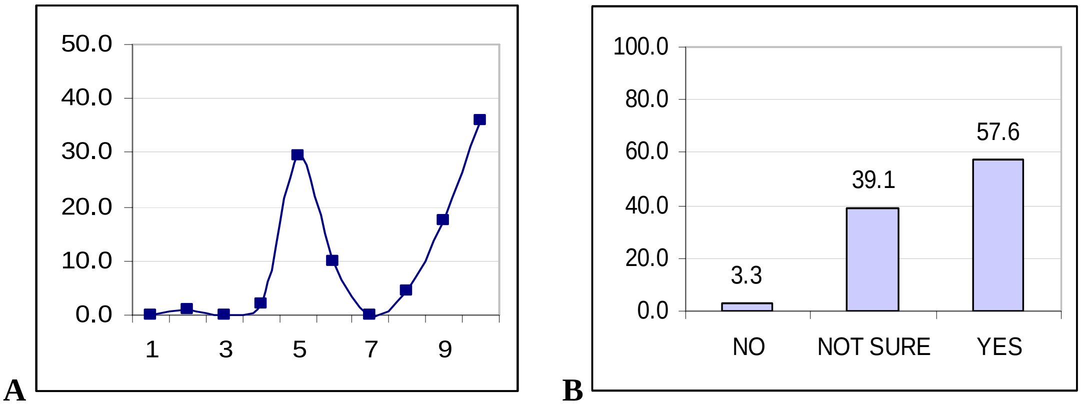


Figure 6.14. The distribution of the answers to the question about the importance of the ability to show EC on a metabolic diagram. B. The discrete distribution of the answers. Ratings from 1 to 4 are summarized and represented by the first column (NO), ratings 5 and 6 are – by the second column (NOT SURE) and ratings from 7 to 10 – by the third column (YES).

6.5.15. ‘Label outside the shape’ or ‘label inside the shape’?

Most of the responders (82%) have decided in favour of ‘label inside the shape’ representation (Figure 6.15).

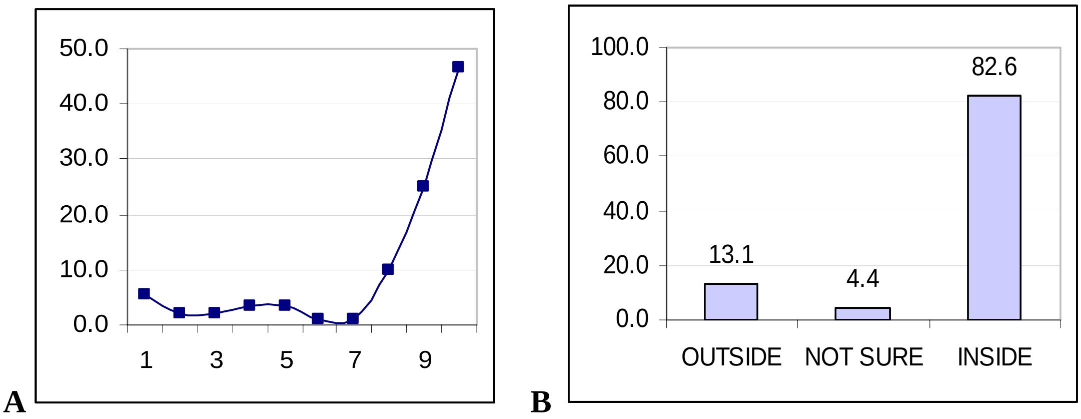


Figure 6.15. A. The distribution of the answers on the question whether the label should be shown inside of outside the shape. B. The discrete distribution of the answers for the question about the importance of the ability to show EC on a metabolic diagram. Ratings from 1 to 4 are summarized and represented by the first column (OUTSIDE), ratings 5 and 6 are – by the second column (NOT SURE) and ratings from 7 to 10 – by the third column (INSIDE).

The table 6.1 summarizes the most important results of the survey and provides brief interpretation of the results:

- In the mEPN ‘simple biochemical’ entity is well represented. As the approach to the label representation in the mEPN is similar to the SBGN PD, the ‘label inside the shape’, new version of the SBGN PD language could benefit from changing the shape from a circle to a more convenient form.
- The UN provides a consistent and convenient way to represent enzyme function. The consistency is ensured using the ‘functional gate’ entity which allows visualising EC in the same way on different types of maps. For example, in current version of the SBGN PD it is possible to show EC as a name of the ‘macromolecule’ entity, but in case when actual protein name has to be represented, it is impossible to show EC any more.
- The new entity ‘functional gate’ ensures the most convenient way to link signalling and metabolic events on metabolism regulation diagrams.
- The new entity ‘multistate entity’ does not provide a better representation. Removing this entity from the UN’s set of symbols has to be considered.
- The SBGN PD provides the most easy-to-read representation. mEPN has too many symbols and that makes it difficult to read the diagrams in this notation.
- The SBGN PD provides the most convenient complex representation.
- The SBGN PD provides the most convenient signalling events representation (macromolecule representation, state variable representation).
- The number of symbols used in the mEPN has to be reduced in order to make it more successful notation. That would also make the mEPN system of symbols easier to learn.
- Compactness is an important feature of a graphical language. On the other hand, according to some of the respondents’ comments, it is not the most important feature.
- Most of the users would like to be able to show EC on metabolic diagrams.
- It is important to be able to represent ‘generic-specific’ relationships. The UN is the only notation that enables a user to do that.

‘Label outside the shape’ approach used in the UN has to be reconsidered as most users find the ‘label inside the shape’ visualisation more convenient.

Table 6.1. Interpretation of the survey results

QUESTION	KEY WORDS	SBGN PD	MEPN	UN	INTERPRETATION
Q1	metabolic	6.8	**8.5**	7.2	mEPN: well-represented simple chemical New shape for compounds should be introduced in the SBGN PD
Q2	EC	6.4	6.3	**8.4**	UN: consistent way to represent enzyme function
Q3	regulation	6.6	6.4	**8.4**	UN: functional gate to link metabolism and signalling
Q4	signalling	**9.0**	7.9	**6.2**	UN: confusing representation (multistate entity) SBGN PD: well-represented signalling events
Q5	readability	8.3	**6.2**	7.3	mEPN: too many types of the ‘process nodes’
Q6	compactness	6.6	6.4	6.5	Results are inconclusive
Q7	complex	**8.2**	7.1	6.8	SBGN PD: well-represented complex
Q8	gene expression	7.3	7.1	7.2	Results are inconclusive
Q9	number of symbols	7.6	**4.2**	7.0	mEPN: too many symbols
Q10	incomplete knowledge	6.4	6.3	6.4	Results are inconclusive
Q11	learning	7.3	**4.8**	**8.3**	mEPN: difficult to learn. UN: easy to learn
Q12	compactness issue	no 17%, not sure 25%, yes 59%			Compactness is an important feature of a notation
Q13	EC visualisation issue	no 7%, not sure 26%, yes 68%			It is important to be able to show EC
Q14	‘generic-specific’ issue	no 3%, not sure 39%, yes 58%			Representing generic-specific is important
Q15	label position issue	outside 13%, not sure 4%, inside 82%			“Label outside the shape” approach in UN should be reconsidered

One of the important outcomes of the survey was the evaluation of the features of different graphical languages that allowed us to determine those most usable (from user's point of view). The list of approved/desired features that were identified as the result of comparable analysis of the three notations is given below:

1. Simple chemical representation by the mEPN,
2. Complex representation by the SBGN PD,
3. Well-represented signalling events in the SBGN PD,
4. A consistent way to represent enzyme function in the UN,
5. Minimal efficient set of symbols in the UN,
6. Easy-to-learn features of the UN,
7. Generic-specific entities representation provided by the UN.

The result of the survey showed that the UN has new features currently not available in other notation that allow improving usability of graphical representation for cellular networks.

The described in this chapter survey addresses only one of the two main subjects of this research, namely, the system of symbols designed for unambiguous network representation. It would be also desirable to evaluate the usability of the Network Representation Framework. In order to make meaningful evaluation of this type, the framework should actually be used by an expert in context of a graphical network editor for a particular task and would require significant amount of work. Unfortunately, this kind of resource for usability evaluation is not currently available.

The next section discusses the attempt to put together the useful users'-approved features of the mEPN, the UN and SBGN PD on the basis of SBGN PD language.

Chapter 7

Proposed new entities for SBGN PD language

As part of the SBGN community efforts to develop SBGN languages, a workshop series were initiated in 2006. SBGN 6 meeting took place in Edinburgh in 2010 and was part of the COMBINE meeting (COmputational Modeling in BIology Network, http://sbml.org/Events/Forums/COMBINE_2010).

Reviewing SBGN PD language features and preparing diagrams in SBGN PD language I came to conclusion that despite the SBGN PD notation is very well developed, it has some disadvantages, mainly in the area of metabolism representation. I felt that the SBGN PD language could benefit from adopting those features of the UN that are related to metabolism representation, in particular, ‘functional gate’ and ‘identity gate’. Corresponding proposal was prepared and introduced as an oral presentation during the COMBINE meeting (Mazein A. Metabolic Network Representation in SBGN PD: EC and Identity Gate, COMBINE 2010, 06 October 2010, http://precedings.nature.com/documents/4974/version/1). The proposal was further discussed after the COMBINE meeting as part of SBNG discussion (Nicolas LeNovere, personal communication, 2010). The proposal is described in the next three sections.

7.1 The new shape for metabolites.

Most metabolic compounds have long names compared to the names of proteins where mainly abbreviations are used. It is not easy to fit a long name into a circle shape reserved for the “small molecule” entity in SBGN PD (SBGN, www.sbgn.org).

There are several possible solutions:

- applying the approach used in the UN “name-outside-the-shape” (this is not allowed in SBGN PD and would make the notation inconsistent);
- making circles large enough so the name would fit the shape (this solution would require a lot more space);
- replacing long compound names by short names (this solution would require changing thousands of names in case of genome-scale metabolic network visualisation; a list of abbreviations would require to be enclosed with each metabolic map);
- using comparatively small circles and allowing part of the name to be outside the shape (this solution was used by Falk Schreiber, the winners of the SBGN annual competition in 2010 (http://www.sbgn.org/Competition);
- changing the shape from circle to a more convenient shape (this solution is proposed here).

The second best solution is using comparatively small circles and allowing a part of the name be outside the shape (Figure 7.1, A). In this case no changes in SBGN PD language would require. The problem with this solution is that having a part of the name outside a compound shape makes it extremely difficult to use horizontal connectors as in SBGN PD the link have to go to the edge of a shape and a horizontal links are not allowed to cross the name (Figure 7.1, B) according to the SBGN PD rules (SBGN PD Level 1 Version 1.2 specification, http://www.sbgn.org/Documents/Specifications). Then the only solution is using several bend points (Figure 7.1, C). This problem becomes even more difficult to avoid in cases of a very busy network (Figure 7.2).

Proposed here solution is to change the shape for the “small molecules” entity. An example with a new shape is shown in Figure 7.3. A new shape is used instead the current SBGN PD circle shape for compound.

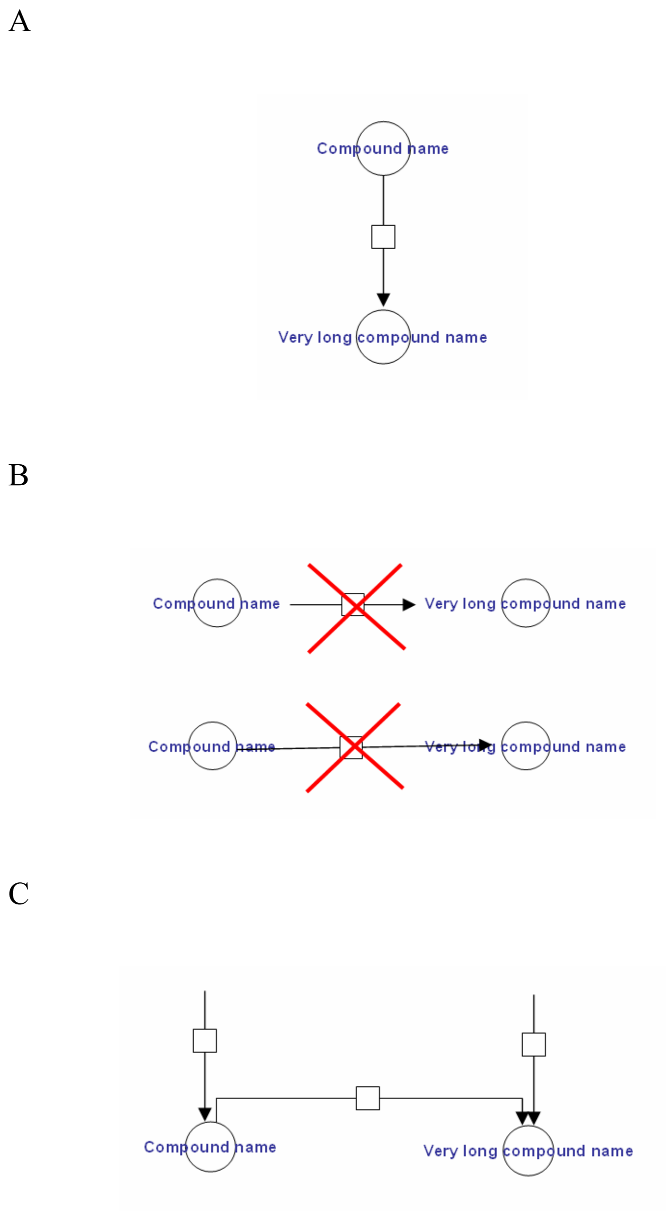


Figure 7.1. Metabolic reaction representation in SBGN PD. A. Using comparatively small circles and allowing a part of the name be outside the shape. B. Forbidden representation in SBGN PD. C. Possible solution for representing horizontal links in the current version of SBGN PD.

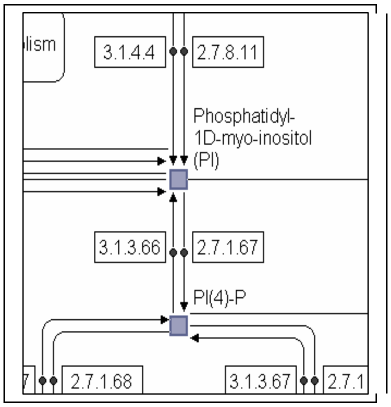


Figure 7.2. A fragment of Phosphatidylinositol pathway map in the UN. In this example a compound is connected to other compounds by 10 links, and 5 of them are horizontal. In this case that would be very difficult to show it in the current version of SBGN PD language.

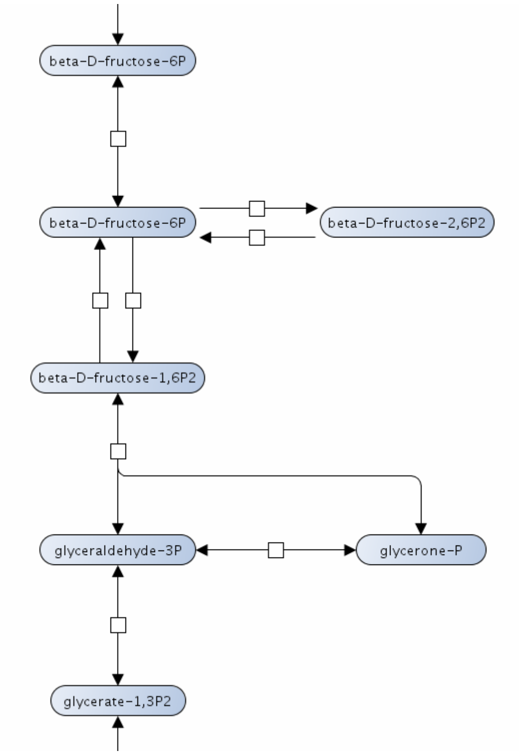


Figure 7.3. A new shape for the “small molecule” entity of SBGN PD language. The shape accommodates the name of the entity. This way horizontal links can be used without any difficulty.

7.2 Using EC as a functional gate

EC is the most compact way to show enzymological function on a metabolic/signalling/metabolism regulation diagram. In cases when several EC numbers are assigned to one reaction, or one protein/complex has several activities (ECs), it is impossible to show the enzyme function without using an additional symbol.

The currently used representation in SBGN PD uses ‘macromolecule’ entity with EC as a name. For example the glycolysis diagram available from SBGN website (SBGN PD examples, http://sbgn.org/Documents/PD_L1_Examples). EC is shown as a generic protein that could actually be not only a protein but also a multimer or a complex.

Another problem is inconsistency in representation. As soon as an actual protein name has to be shown the EC can not be shown any more (Figure 7.4).

The same example is shown on Figure 7.5 with the proposed ‘functional gate’ entity.

‘Functional gate’ entity is represented by a special glyph which has to be placed on the link between a protein/complex and a reaction. One or more outputs are possible. No input, one or many inputs are allowed.

One of the important features of the new entity is that it can be used on its own on maps were only metabolites and EC need to be shown; and it can be linked to a particular proteins on more complicated maps (metabolism regulation, signalling).

Another example represents the cases when one protein can catalyze two reactions and when one reaction is being catalyzed by several proteins with the same enzymological activities (Figure 7.6).

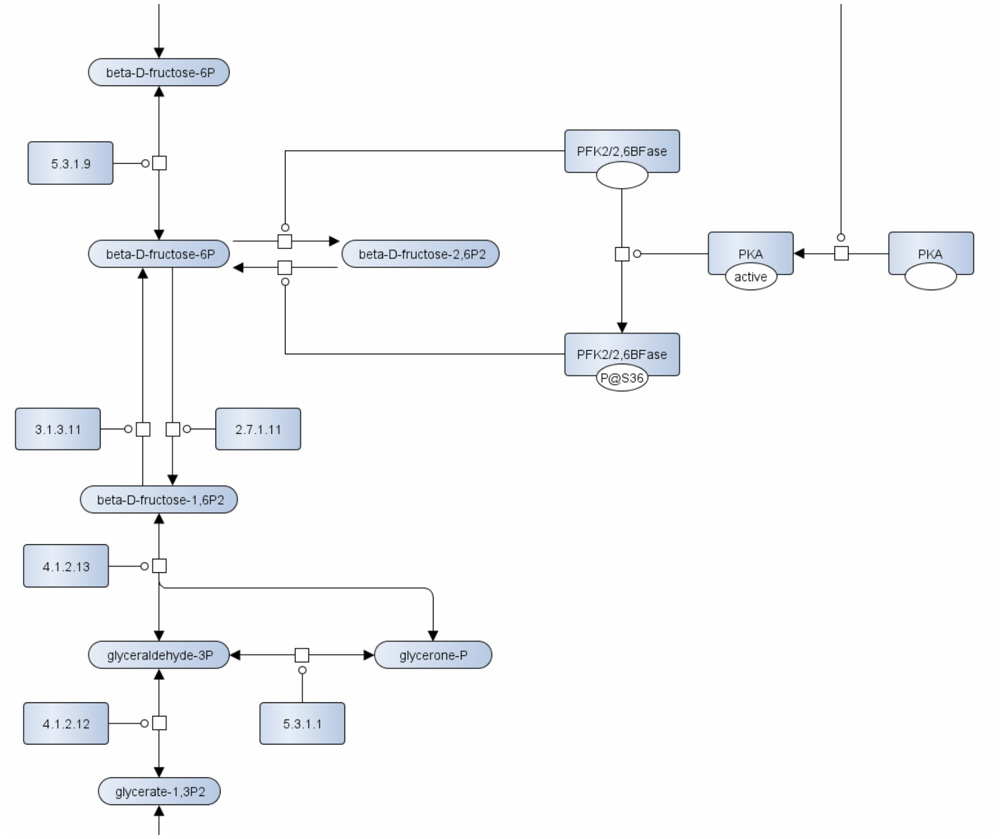


Figure 7.4. Inconsistent EC representation in SBGN PD. A fragment of Glycolysis pathway. Most of the enzymes are shown as generic protein entity. Bifunctional enzyme PFK2/2,6BFase visualized without its two enzymological functions shown as EC. Please compare this representation to the proposed visualisation (Figure 7.5).

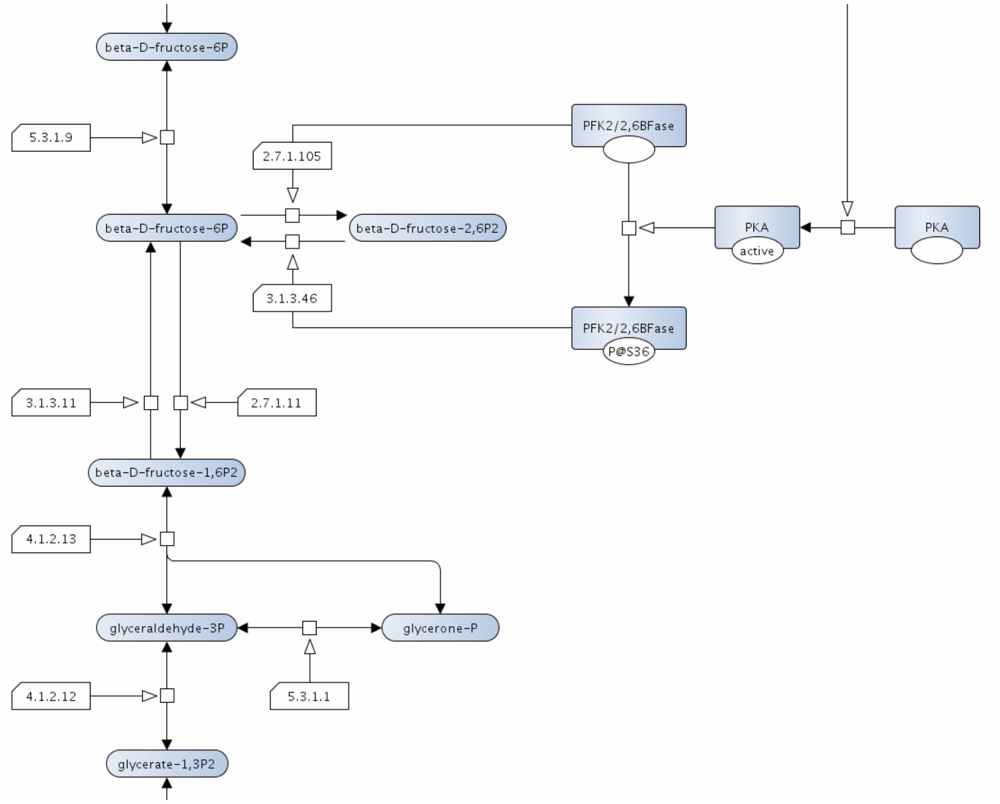


Figure 7.5. Using EC as a special glyph ‘functional gate’. A fragment of Glycolysis pathway. Please compare to Figure 9.4. Both the enzyme functions (2.7.1.105 and 3.1.3.46) of the protein PFK2/2,6BFase are shown without any difficulties.

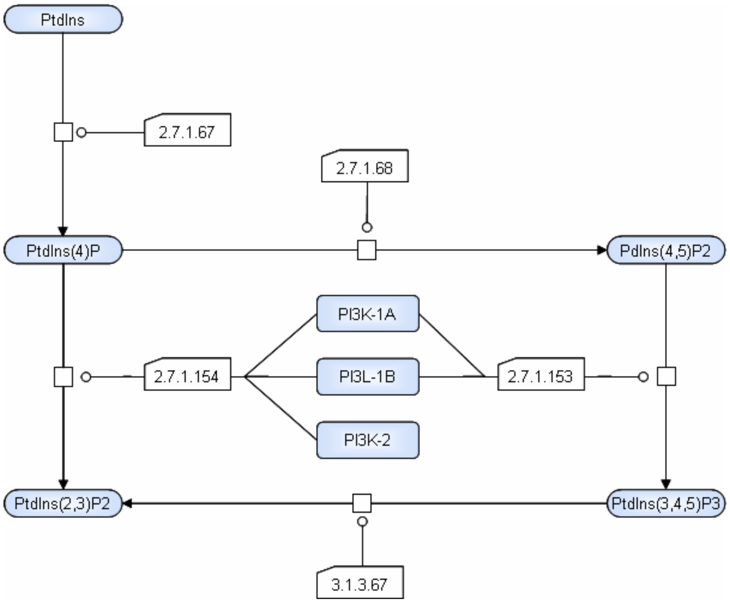


Figure 7.6. A fragment of Phosphatidylinositol pathway map shown in the modified SBGN PD language. ‘Functional gate’ entity is used to show EC.

7.3 Identity gate

In the current version of SBGN PD there is no specific way to connect generic entity to particular specific entities on a diagram (SBGN PD Level 1 Version 1.2 specification, http://www.sbgn.org/Documents/Specifications).

The “identity gate” glyph is proposed to connect a generic and corresponding specific entities. A circle with symbol “≡“ in it can be used for this entity. Only one output is allowed, multiple inputs are possible.

Figure 9.7 shows an example where ‘identity gate’ is used to show the relationships between soluble and membrane forms of TNFalpha. Both soluble and membrane forms can activate TNFR1, but only the membrane form can activate TNFR2. Using “identity gate” is the only way to show such a case on a single diagram. Generic TNFalpha would be used as TNFR1 activator, mTNFalpha – as TNFR2 activator; and it would be possible to show relationships between sTNFalpha, mTNFalpha and generic TNFalpha.

Using identity gates potentially might cause ambiguity on a diagram. An example of such case is shown on Figure 7.8.

To avoid ambiguity until this issue is resolved I would like to propose a set of rules that would allow using ‘identity gates’ in limited number of cases.

Rule 1. It is safe if there is only one ‘identity gate’ one a map (Figure 9.7).

Rule 2. It is safe if the map does not include any processes but only ‘identity gates’ and related entities are shown (Figure 9.9).

Rule 3. It is safe if several generic entities (with corresponding ‘identity gates’) are used only as inputs in the processes and are not used as outputs. In other words, it is safe if two generics (with corresponding identity gates) are not connected via one or several processes in the way that one is an input and another is an output (Figure 9.10).

The currently proposed entity is discussed as ‘identity operator’ and it is not clear if it is going be a part of entity pool node set in SBGN PD (‘identity operator’), a part of the process node set of SDGN PD (‘identity process’) or both (Nicolas LeNovere, personal communication, 2010).

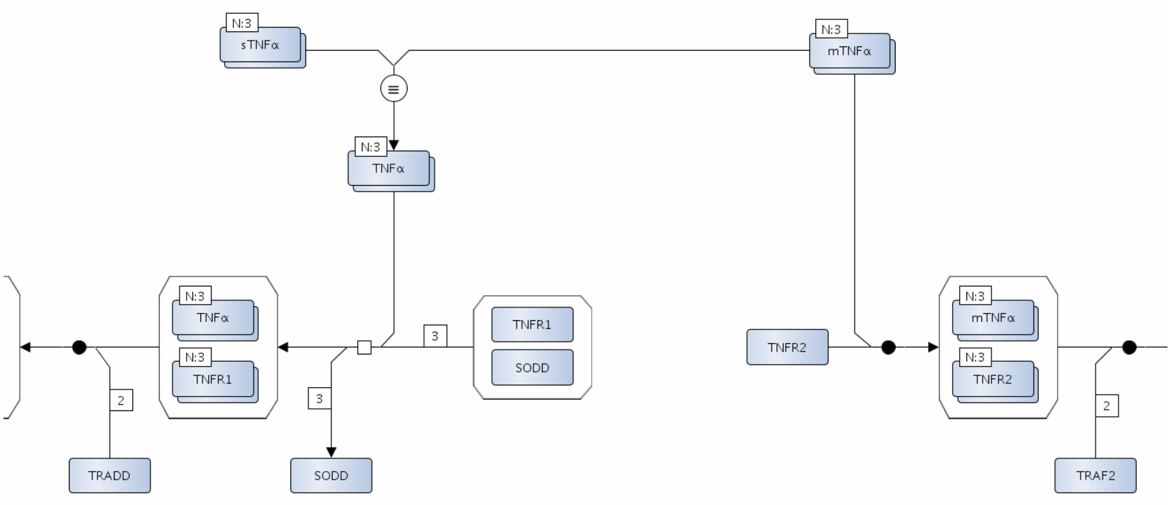


Figure 7.7. Using proposed 'identity gate' entity. Both soluble sTNF-alpha and membrane mTNF-alpha forms can activate TNFR1. Generic entity TNF-alpha is used to show that. Only membrane form can activate TNFR2.

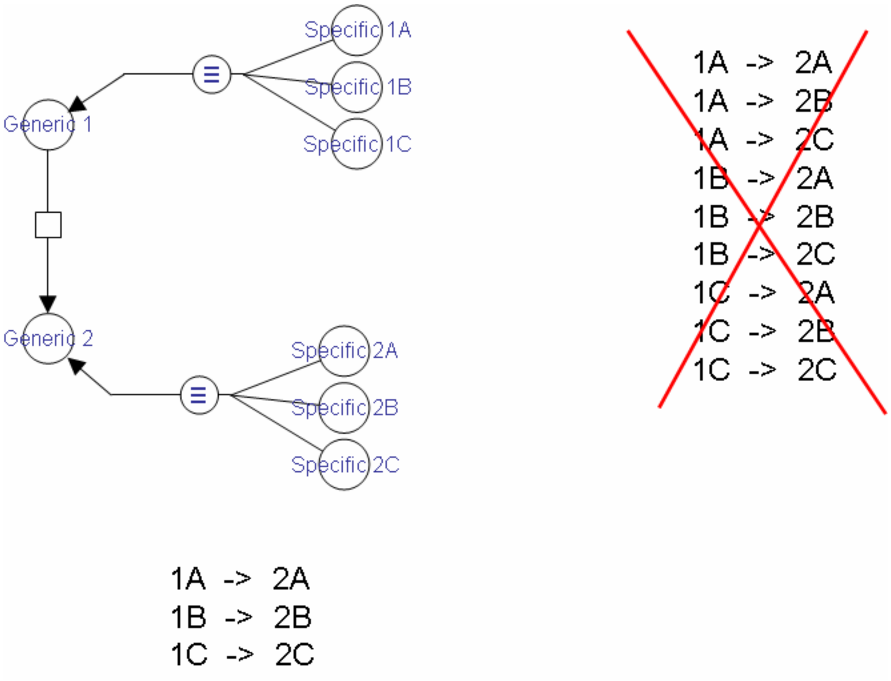


Figure 7.8. Incorrect using of the ‘functional gate’ entity. On the diagram an attempt is made to show that 1A specific entity transforms into 2A specific entity, but not to 2B or 2C. On the other had the reader can assume that any of the transformations are possible. In other words, using ‘identity gate’ in this case makes the representation ambiguous.

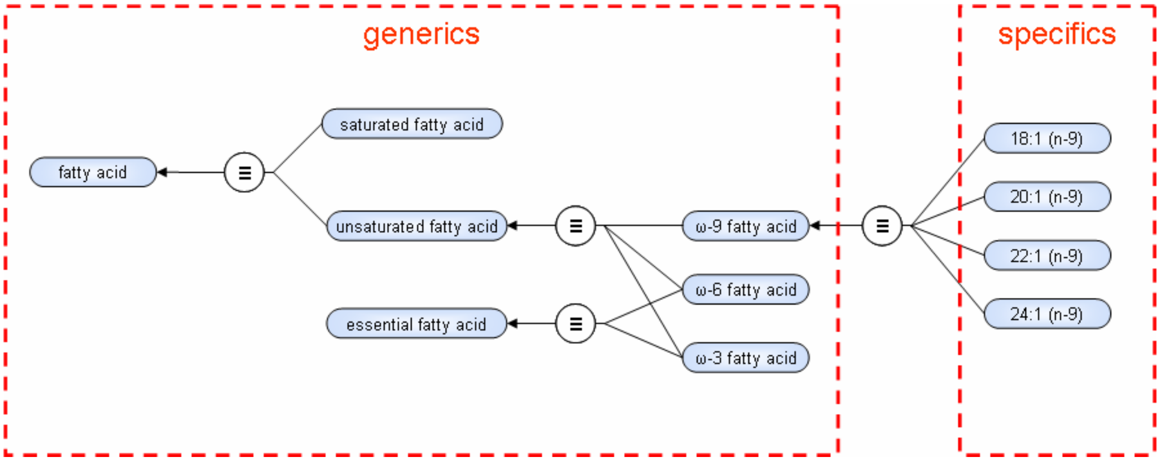


Figure 7.9. Fatty acid generic chart. The scheme represents relationships between generic and generic, generic and specific entities. No processes are shown on the map.

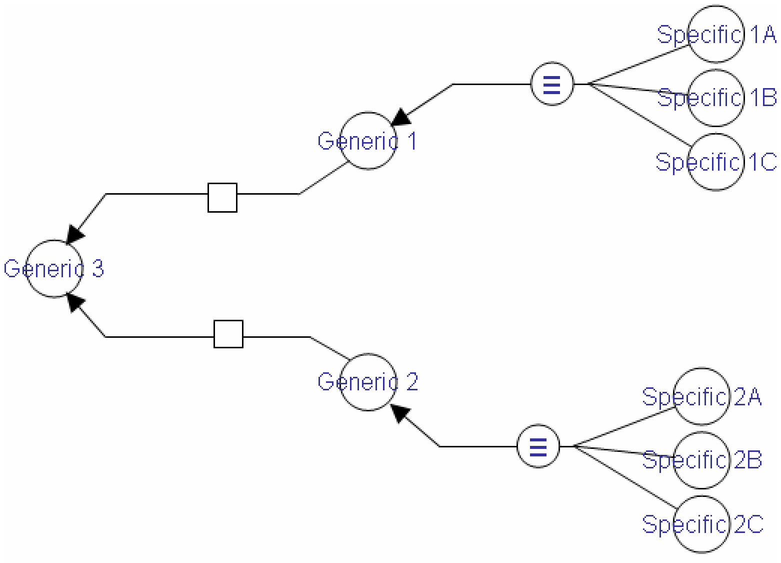


Figure 7.10. Illustration of using rule 3 of representing ‘identity gate’. Generics 1 and 2 that are connected each to corresponding ‘identity gate’. Generics 1 and 2 are used only as inputs on the map and none of them is used as an output.

Conclusion

Summary

In this research the most powerful features of the existing approaches have been considered alongside with newly proposed techniques. Two extensive examples show how a complex network can be efficiently represented at several levels of detail. A new notation, the UN, has been evaluated against the SBGN PD language and mEPN during the user survey. The results have suggested us to consider the exclusion of the unsuccessful features, while keeping the most useful ones.

The survey results have also indicated that it is desired that a consensus visual language should comply with the following two criteria in order to be both user-friendly and able to represent cellular network unambiguously: minimal number of easy-to-learn symbols and consistency in representation between different levels and aspects of network representation.

Individual points and results of the current research are listed below:

1. Existing techniques of biological network representation have been reviewed and the strongest features were selected for inclusion when developing a new representation system.

2. New graphical language has been developed and is introduced here as the Unified Notation. The most distinctive features of the UN are: minimal number of symbols sufficient for representing cellular networks, the ‘label outside the shape’ visualisation, compactness of network representation, ‘functional gate’ entity, ‘multistate’ entity and ‘identity gate’ entity. The last three entities are used only in the UN and were introduced in order to improve the notation usability.

3. The new representation system, the Network Representation Framework, enables managing large cellular networks by organising the diagrams on multiple levels of details and multiple levels of constraints. This thesis proposes using comparatively small ‘basic’ diagrams represented in unambiguous notation. The outline diagrams unite these basic diagrams into a single large-scale virtual map.

4. In order to test the capabilities of the new features, large-scaled examples have been developed:

   a) the Edinburgh Human Metabolic Network representation;

   b) the TNF-alpha receptor network representation

   The ‘Glycolysis regulation’ example demonstrates the ability of the UN to represent events from different domains (metabolic and signalling) on a single diagram.

5. The Graphical notation survey has been carried out that allowed to obtain user feedback on the features of new visualisation language. The UN has been compared with the SBGN PD and mEPN languages. The results of the survey have showed that while using ‘label outside the shape’ approach and the ‘multistate’ entity should be reconsidered, such features of the UN as ‘functional gate’ and ‘identity gate’ entities were deemed a welcome addition.

6. The new features of the UN ‘functional gate’ and ‘identity gate’ have been adopted for SBGN language and alongside with proposed new shape for ‘simple chemical’ entity offered for consideration for the new version of SBGN PD language.

Future work

I would like to continue working on improving the techniques used for representing cellular networks. A new version of the UN is going to be developed and its features will be changed according to the results of the survey in order to provide better usability and network representation.

If most adventurous features of the UN such as the ‘functional gate’ and ‘identity gate’ are accepted for the next version of SBGN PD language, in addition to the UN we are planning to use the SBGN PD on the detailed level of the NR representation and propose a algorithm for translation between the SBGN PD and SBGN AF languages in order to make them a part of the Network Representation.

References

Ananko, E.A., Podkolodny, N.L., Stepanenko, I.L., Ignatieva, E.V., Podkolodnaya, O.A. and Kolchanov, N.A. (2002) GeneNet: a database on structure and functional organisation of gene networks. *Nucleic Acids Res*, 30, 398-401.

Beaudouin-Lafon, M. (1993) An overview of human-computer interaction. *Biochimie*, 75, 321-329.

Deckard, A., Bergmann, F.T. and Sauro, H.M. (2006) Supporting the SBML layout extension. *Bioinformatics*, 22, 2966-2967.

Dix, A.J. (2004) *Human-computer interaction*. Pearson, Harlow.

Fayard, E., Tintignac, L.A., Baudry, A. and Hemmings, B.A. (2005) Protein kinase B/Akt at a glance. *J Cell Sci*, 118, 5675-5678.

Folmer, E. and Bosch, J. (2004) Architecting for usability: a survey. The Journal of Systems and Software, 70, 61–78.

Gauges, R., Rost, U., Sahle, S. and Wegner, K. (2006) A model diagram layout extension for SBML. *Bioinformatics*, 22, 1879-1885.

Hu, Z., Mellor, J., Wu, J., Kanehisa, M., Stuart, J.M. and DeLisi, C. (2007) Towards zoomable multidimensional maps of the cell. *Nat Biotechnol*, 25, 547-554.

Kam, N., Cohen, I.R. and Harel, D. (2001) The immune system as a reactive system: modeling T cell activation with statecharts. *Human-Centric Computing Languages and Environments, 2001. Proceedings IEEE Symposia on*, pp. 15-22.

Kanehisa, M., Goto, S., Kawashima, S. and Nakaya, A. (2002) The KEGG databases at GenomeNet. *Nucleic Acids Res*, 30, 42-46.

Kang, H., Getoor, L., Shneiderman, B., Bilgic, M. and Licamele, L. (2008) Interactive entity resolution in relational data: a visual analytic tool and its evaluation. *IEEE Trans Vis Comput Graph*, 14, 999-1014.

Karat, J. Karat, C.M. (2003) The Evolution of User-centered Focus in the Human Computer Interaction Field. *IBM Systems Journal*, 42, 532–541.

Karp, P.D., Ouzounis, C.A., Moore-Kochlacs, C., Goldovsky, L., Kaipa, P., Ahren, D., Tsoka, S., Darzentas, N., Kunin, V. and Lopez-Bigas, N. (2005) Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res*, 33, 6083-6089.

Keefe, D.F. (2010) Integrating visualization and interaction research to improve scientific workflows. *IEEE Comput Graph Appl*, 30, 8-13.

Kholodenko, B.N., Demin, O.V., Moehren, G. and Hoek, J.B. (1999) Quantification of short term signalling by the epidermal growth factor receptor. *J Biol Chem*, 274, 30169-30181.

Kitano, H. (2002) Systems biology: A brief overview. *Science*, 295, 1662-1664.

Kitano, H., Funahashi, A., Matsuoka, Y. and Oda, K. (2005) Using process diagrams for the graphical representation of biological networks. *Nat Biotechnol*, 23, 961-966.

Klipp, E., Liebermeister, W., Helbig, A., Kowald, A. and Schaber, J. (2007) Systems biology standards - the community speaks. *Nat Biotech*, 25, 390-391.

Kohn, K.W. (2001) Molecular interaction maps as information organizers and simulation guides. *Chaos*, 11, 84-97.

Kohn, K.W. and Aladjem, M.I. (2006) Circuit diagrams for biological networks. *Mol Syst Biol*, 2, 2006 0002.

Le Novere, N., Hucka, M., Mi, H., Moodie, S., Schreiber, F., Sorokin, A., Demir, E., Wegner, K., Aladjem, M.I., Wimalaratne, S.M., Bergman, F.T., Gauges, R., Ghazal, P., Kawaji, H., Li, L., Matsuoka, Y., Villeger, A., Boyd, S.E., Calzone, L., Courtot, M., Dogrusoz, U., Freeman, T.C., 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. and Kitano, H. (2009) The Systems Biology Graphical Notation. *Nat Biotechnol*, 27, 735-741.

Lee, B., Parr, C.S., Plaisant, C., Bederson, B.B., Veksler, V.D., Gray, W.D. and Kotfila, C. (2006) TreePlus: interactive exploration of networks with enhanced tree layouts. *IEEE Trans Vis Comput Graph*, 12, 1414-1426.

Lieberman, M.D., Taheri, S., Guo, H., Mirrashed, F., Yahav, I., Aris, A. and Shneiderman, B. (2011) Visual Exploration across Biomedical Databases. *IEEE/ACM Trans Comput Biol Bioinform*, 8, 536-550.

Ma, H., Sorokin, A., Mazein, A., Selkov, A., Selkov, E., Demin, O. and Goryanin, I. (2007) The Edinburgh human metabolic network reconstruction and its functional analysis. *Mol Syst Biol*, 3, 135.

Michal G. (1968) Biochemical Pathways wall chart. Mannheim, Germany, Boehringer Mannheim.

Moodie, S.L., Sorokin, A.A., Goryanin, I. and Ghazal, P. (2006) Graphica Notation to describe the Logical Interactions of Biological Pathways. *Journal of Integrative Bioinformatics*, 3, 36.

Oda, K. and Kitano, H. (2006) A comprehensive map of the toll-like receptor signalling network. *Mol Syst Biol*, 2, 2006 0015.

Oda, K., Matsuoka, Y., Funahashi, A. and Kitano, H. (2005) A comprehensive pathway map of epidermal growth factor receptor signalling. *Mol Syst Biol*, 1, 2005 0010.

Okuda, S., Yamada, T., Hamajima, M., Itoh, M., Katayama, T., Bork, P., Goto, S. and Kanehisa, M. (2008) KEGG Atlas mapping for global analysis of metabolic pathways. *Nucleic Acids Res*.

Olson, G.M. and Olson, J.S. (2003) Human-computer interaction: psychological aspects of the human use of computing. *Annu Rev Psychol*, 54, 491-516.

Perer, A. and Shneiderman, B. (2009) Integrating statistics and visualization for exploratory power: from long-term case studies to design guidelines. *IEEE Comput Graph Appl*, 29, 39-51.

Pettifer, S. Thorne, D., McDermott, P., Marsh, J., Villeger, A., Kell, D.B. and Attwood, T.K. (2009) Visualising biological data: a semantic approach to tool and database integration. *BMC Bioinformatics*, 10, S19.

Pinney, J.W., Westhead, D.R. and McConkey, G.A. (2003) Petri Net representations in systems biology. *Biochem Soc Trans*, 31, 1513-1515.

Plaisant, C., Fekete, J.D. and Grinstein, G. (2008) Promoting insight-based evaluation of visualizations: from contest to benchmark repository. *IEEE Trans Vis Comput Graph*, 14, 120-134.

Salway, J.G. (1994) *Metabolism at a glance*. Blackwell Science Ltd, Oxford.

Sears, A. and Jacko, J.A. (2008) *The human-computer interaction handbook : fundamentals, evolving technologies, and emerging applications*. Lawrence Erlbaum Associates, New York; London.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 13, 2498-2504.

Shneiderman, B. (2000) Universal Usability. Communications of The ACM May, 200, 43(5).

Shneiderman, B. and Aris, A. (2006) Network visualization by semantic substrates. *IEEE Trans Vis Comput Graph*, 12, 733-740.

Wickens, Christopher D., John D. Lee, Yili Liu, and Sallie E. Gordon Becker. (2004) An Introduction to Human Factors Engineering. Second ed. Upper Saddle River, NJ: Pearson Prentice Hall, 185–193.

Zaphiris, P. and Ang, C.S. (2009) *Human computer interaction: concepts, methodologies, tools, and applications*. Information Science Reference, Hershey, PA ; London.

Appendix I.

References for TNF signalling network

Acehan D, Jiang X, Morgan DG, Heuser JE, Wang X, Akey CW: Three-dimensional structure of the apoptosome: implications for assembly, procaspase-9 binding, and activation. *Mol Cell* 2002, 9(2):423-432.

Andersen PL, Zhou H, Pastushok L, Moraes T, McKenna S, Ziola B, Ellison MJ, Dixit VM, Xiao W: Distinct regulation of Ubc13 functions by the two ubiquitin-conjugating enzyme variants Mms2 and Uev1A. *J Cell Biol* 2005, 170(5):745-755.

Bao Q, Shi Y: Apoptosome: a platform for the activation of initiator caspases. *Cell Death Differ* 2007, 14(1):56-65.

Bayir H, Fadeel B, Palladino MJ, Witasp E, Kurnikov IV, Tyurina YY, Tyurin VA, Amoscato AA, Jiang J, Kochanek PM *et al*: Apoptotic interactions of cytochrome c: redox flirting with anionic phospholipids within and outside of mitochondria. *Biochim Biophys Acta* 2006, 1757(5-6):648-659.

Besse A, Lamothe B, Campos AD, Webster WK, Maddineni U, Lin SC, Wu H, Darnay BG: TAK1-dependent signaling requires functional interaction with TAB2/TAB3. *J Biol Chem* 2007, 282(6):3918-3928.

Blonska M, Shambharkar PB, Kobayashi M, Zhang D, Sakurai H, Su B, Lin X: TAK1 is recruited to the tumor necrosis factor-alpha (TNF-alpha) receptor 1 complex in a receptor-interacting protein (RIP)-dependent manner and cooperates with MEKK3 leading to NF-kappaB activation. *J Biol Chem* 2005, 280(52):43056-43063.

Blonska M, You Y, Geleziunas R, Lin X: Restoration of NF-kappaB activation by tumor necrosis factor alpha receptor complex-targeted MEKK3 in receptor-interacting protein-deficient cells. *Mol Cell Biol* 2004, 24(24):10757-10765.

Bratton SB, Walker G, Roberts DL, Cain K, Cohen GM: Caspase-3 cleaves Apaf-1 into an approximately 30 kDa fragment that associates with an inappropriately oligomerized and biologically inactive approximately 1.4 MDa apoptosome complex. *Cell Death Differ* 2001, 8(4):425-433.

Cain K, Bratton SB, Langlais C, Walker G, Brown DG, Sun XM, Cohen GM: Apaf-1 oligomerizes into biologically active approximately 700-kDa and inactive approximately 1.4-MDa apoptosome complexes. *J Biol Chem* 2000, 275(9):6067-6070.

Cain K, Brown DG, Langlais C, Cohen GM: Caspase activation involves the formation of the aposome, a large (approximately 700 kDa) caspase-activating complex. *J Biol Chem* 1999, 274(32):22686-22692.

Chen D, Li X, Zhai Z, Shu HB: A novel zinc finger protein interacts with receptor-interacting protein (RIP) and inhibits tumor necrosis factor (TNF)- and IL1-induced NF-kappa B activation. *J Biol Chem* 2002, 277(18):15985-15991.

Ea CK, Deng L, Xia ZP, Pineda G, Chen ZJ: Activation of IKK by TNFalpha requires site-specific ubiquitination of RIP1 and polyubiquitin binding by NEMO. *Mol Cell* 2006, 22(2):245-257.

Feng Y, Longmore GD: The LIM protein Ajuba influences interleukin-1-induced NF-kappaB activation by affecting the assembly and activity of the protein kinase Czeta/p62/TRAF6 signaling complex. *Mol Cell Biol* 2005, 25(10):4010-4022.

Fritz A, Brayer KJ, McCormick N, Adams DG, Wadzinski BE, Vaillancourt RR: Phosphorylation of serine 526 is required for MEKK3 activity, and association with 14-3-3 blocks dephosphorylation. *J Biol Chem* 2006, 281(10):6236-6245.

Garrido C, Galluzzi L, Brunet M, Puig PE, Didelot C, Kroemer G: Mechanisms of cytochrome c release from mitochondria. *Cell Death Differ* 2006, 13(9):1423-1433.

Gogvadze V, Orrenius S, Zhivotovsky B: Multiple pathways of cytochrome c release from mitochondria in apoptosis. *Biochim Biophys Acta* 2006, 1757(5-6):639-647.

Guicciardi ME, Gores GJ: AIP1: a new player in TNF signaling. *J Clin Invest* 2003, 111(12):1813-1815.

Habelhah H, Takahashi S, Cho SG, Kadoya T, Watanabe T, Ronai Z: Ubiquitination and translocation of TRAF2 is required for activation of JNK but not of p38 or NF-kappaB. *Embo J* 2004, 23(2):322-332.

He KL, Ting AT: A20 inhibits tumor necrosis factor (TNF) alpha-induced apoptosis by disrupting recruitment of TRADD and RIP to the TNF receptor 1 complex in Jurkat T cells. *Mol Cell Biol* 2002, 22(17):6034-6045.

Heyninck K, De Valck D, Vanden Berghe W, Van Criekinge W, Contreras R, Fiers W, Haegeman G, Beyaert R: The zinc finger protein A20 inhibits TNF-induced NF-kappaB-dependent gene expression by interfering with an RIP- or TRAF2-mediated transactivation signal and directly binds to a novel NF-kappaB-inhibiting protein ABIN. *J Cell Biol* 1999, 145(7):1471-1482.

Hsu H, Shu HB, Pan MG, Goeddel DV: TRADD-TRAF2 and TRADD-FADD interactions define two distinct TNF receptor 1 signal transduction pathways. *Cell* 1996, 84(2):299-308.

Hu Y, Benedict MA, Ding L, Nunez G: Role of cytochrome c and dATP/ATP hydrolysis in Apaf-1-mediated caspase-9 activation and apoptosis. *Embo J* 1999, 18(13):3586-3595.

Izadi H, Motameni AT, Bates TC, Olivera ER, Villar-Suarez V, Joshi I, Garg R, Osborne BA, Davis RJ, Rincon M *et al*: c-Jun N-terminal kinase 1 is required for Toll-like receptor 1 gene expression in macrophages. *Infect Immun* 2007, 75(10):5027-5034.

Kario E, Marmor MD, Adamsky K, Citri A, Amit I, Amariglio N, Rechavi G, Yarden Y: Suppressors of cytokine signaling 4 and 5 regulate epidermal growth factor receptor signaling. *J Biol Chem* 2005, 280(8):7038-7048.

Lamkanfi M, Festjens N, Declercq W, Vanden Berghe T, Vandenabeele P: Caspases in cell survival, proliferation and differentiation. *Cell Death Differ* 2007, 14(1):44-55.

Lamothe B, Besse A, Campos AD, Webster WK, Wu H, Darnay BG: Site-specific Lys-63-linked tumor necrosis factor receptor-associated factor 6 auto-ubiquitination is a critical determinant of I kappa B kinase activation. *J Biol Chem* 2007, 282(6):4102-4112.

Lee TH, Shank J, Cusson N, Kelliher MA: The kinase activity of Rip1 is not required for tumor necrosis factor-alpha-induced IkappaB kinase or p38 MAP kinase activation or for the ubiquitination of Rip1 by Traf2. *J Biol Chem* 2004, 279(32):33185-33191.

Li H, Kobayashi M, Blonska M, You Y, Lin X: Ubiquitination of RIP is required for tumor necrosis factor alpha-induced NF-kappaB activation. *J Biol Chem* 2006, 281(19):13636-13643.

Lie TJ, Wood GE, Leigh JA: Regulation of nif expression in Methanococcus maripaludis: roles of the euryarchaeal repressor NrpR, 2-oxoglutarate, and two operators. *J Biol Chem* 2005, 280(7):5236-5241.

Ling L, Cao Z, Goeddel DV: NF-kappaB-inducing kinase activates IKK-alpha by phosphorylation of Ser-176. *Proc Natl Acad Sci U S A* 1998, 95(7):3792-3797.

Malinin NL, Boldin MP, Kovalenko AV, Wallach D: MAP3K-related kinase involved in NF-kappaB induction by TNF, CD95 and IL-1. *Nature* 1997, 385(6616):540-544.

Means JC, Hays R: Mitochondrial membrane depolarization in Drosophila apoptosis. *Cell Death Differ* 2007, 14(2):383-385.

Moscat J, Diaz-Meco MT: The atypical PKC scaffold protein P62 is a novel target for anti-inflammatory and anti-cancer therapies. *Adv Enzyme Regul* 2002, 42:173-179.

Ninomiya-Tsuji J, Kishimoto K, Hiyama A, Inoue J, Cao Z, Matsumoto K: The kinase TAK1 can activate the NIK-I kappaB as well as the MAP kinase cascade in the IL-1 signalling pathway. *Nature* 1999, 398(6724):252-256.

Pop C, Timmer J, Sperandio S, Salvesen GS: The apoptosome activates caspase-9 by dimerization. *Mol Cell* 2006, 22(2):269-275.

Ruffolo SC, Shore GC: BCL-2 selectively interacts with the BID-induced open conformer of BAK, inhibiting BAK auto-oligomerization. *J Biol Chem* 2003, 278(27):25039-25045.

Sakahira H, Enari M, Nagata S: Cleavage of CAD inhibitor in CAD activation and DNA degradation during apoptosis. *Nature* 1998, 391(6662):96-99.

Samuel T, Welsh K, Lober T, Togo SH, Zapata JM, Reed JC: Distinct BIR domains of cIAP1 mediate binding to and ubiquitination of tumor necrosis factor receptor-associated factor 2 and second mitochondrial activator of caspases. *J Biol Chem* 2006, 281(2):1080-1090.

Sanz L, Sanchez P, Lallena MJ, Diaz-Meco MT, Moscat J: The interaction of p62 with RIP links the atypical PKCs to NF-kappaB activation. *Embo J* 1999, 18(11):3044-3053.

Schafer ZT, Kornbluth S: The apoptosome: physiological, developmental, and pathological modes of regulation. *Dev Cell* 2006, 10(5):549-561.

Sgorbissa A, Benetti R, Marzinotto S, Schneider C, Brancolini C: Caspase-3 and caspase-7 but not caspase-6 cleave Gas2 in vitro: implications for microfilament reorganization during apoptosis. *J Cell Sci* 1999, 112 (Pt 23):4475-4482.

Shi CS, Kehrl JH: Tumor necrosis factor (TNF)-induced germinal center kinase-related (GCKR) and stress-activated protein kinase (SAPK) activation depends upon the E2/E3 complex Ubc13-Uev1A/TNF receptor-associated factor 2 (TRAF2). *J Biol Chem* 2003, 278(17):15429-15434.

Shi Y: Mechanical aspects of apoptosome assembly. *Curr Opin Cell Biol* 2006, 18(6):677-684.

Shim JH, Xiao C, Paschal AE, Bailey ST, Rao P, Hayden MS, Lee KY, Bussey C, Steckel M, Tanaka N *et al*: TAK1, but not TAB1 or TAB2, plays an essential role in multiple signaling pathways in vivo. *Genes Dev* 2005, 19(22):2668-2681.

Shimokawa N, Qiu CH, Seki T, Dikic I, Koibuchi N: Phosphorylation of JNK is involved in regulation of H(+)-induced c-Jun expression. *Cell Signal* 2004, 16(6):723-729.

Singhirunnusorn P, Suzuki S, Kawasaki N, Saiki I, Sakurai H: Critical roles of threonine 187 phosphorylation in cellular stress-induced rapid and transient activation of transforming growth factor-beta-activated kinase 1 (TAK1) in a signaling complex containing TAK1-binding protein TAB1 and TAB2. *J Biol Chem* 2005, 280(8):7359-7368.

Spierings D, McStay G, Saleh M, Bender C, Chipuk J, Maurer U, Green DR: Connected to death: the (unexpurgated) mitochondrial pathway of apoptosis. *Science* 2005, 310(5745):66-67.

Srinivasula SM, Hegde R, Saleh A, Datta P, Shiozaki E, Chai J, Lee RA, Robbins PD, Fernandes-Alnemri T, Shi Y *et al*: A conserved XIAP-interaction motif in caspase-9 and Smac/DIABLO regulates caspase activity and apoptosis. *Nature* 2001, 410(6824):112-116.

Tang ED, Wang CY, Xiong Y, Guan KL: A role for NF-kappaB essential modifier/IkappaB kinase-gamma (NEMO/IKKgamma) ubiquitination in the activation of the IkappaB kinase complex by tumor necrosis factor-alpha. *J Biol Chem* 2003, 278(39):37297-37305.

Twiddy D, Cohen GM, Macfarlane M, Cain K: Caspase-7 is directly activated by the approximately 700-kDa apoptosome complex and is released as a stable XIAP-caspase-7 approximately 200-kDa complex. *J Biol Chem* 2006, 281(7):3876-3888.

Wajant H, Pfizenmaier K, Scheurich P: Tumor necrosis factor signaling. *Cell Death Differ* 2003, 10(1):45-65.

Wang Y, Wu TR, Cai S, Welte T, Chin YE: Stat1 as a component of tumor necrosis factor alpha receptor 1-TRADD signaling complex to inhibit NF-kappaB activation. *Mol Cell Biol* 2000, 20(13):4505-4512.

Wu W, Pew T, Zou M, Pang D, Conzen SD: Glucocorticoid receptor-induced MAPK phosphatase-1 (MPK-1) expression inhibits paclitaxel-associated MAPK activation and contributes to breast cancer cell survival. *J Biol Chem* 2005, 280(6):4117-4124.

Yamaguchi R, Andreyev A, Murphy AN, Perkins GA, Ellisman MH, Newmeyer DD: Mitochondria frozen with trehalose retain a number of biological functions and preserve outer membrane integrity. *Cell Death Differ* 2007, 14(3):616-624.

Yamamoto M, Okamoto T, Takeda K, Sato S, Sanjo H, Uematsu S, Saitoh T, Yamamoto N, Sakurai H, Ishii KJ *et al*: Key function for the Ubc13 E2 ubiquitin-conjugating enzyme in immune receptor signaling. *Nat Immunol* 2006, 7(9):962-970.

Zhang D, Facchinetti V, Wang X, Huang Q, Qin J, Su B: Identification of MEKK2/3 serine phosphorylation site targeted by the Toll-like receptor and stress pathways. *Embo J* 2006, 25(1):97-107.

Zhang H, Zhang H, Lin Y, Li J, Pober JS, Min W: RIP1-mediated AIP1 phosphorylation at a 14-3-3-binding site is critical for tumor necrosis factor-induced ASK1-JNK/p38 activation. *J Biol Chem* 2007, 282(20):14788-14796.

Zhang H, Zhang R, Luo Y, D'Alessio A, Pober JS, Min W: AIP1/DAB2IP, a novel member of the Ras-GAP family, transduces TRAF2-induced ASK1-JNK activation. *J Biol Chem* 2004, 279(43):44955-44965.

Zhao Y, Conze DB, Hanover JA, Ashwell JD: Tumor necrosis factor receptor 2 signaling induces selective c-IAP1-dependent ASK1 ubiquitination and terminates mitogen-activated protein kinase signaling. *J Biol Chem* 2007, 282(11):7777-7782.

Zou H, Li Y, Liu X, Wang X: An APAF-1.cytochrome c multimeric complex is a functional apoptosome that activates procaspase-9. *J Biol Chem* 1999, 274(17):11549-11556.

Appendix II. Graphical Notation Survey

Graphical Notations Survey

Reasons to complete this questionnaire:

- the questionnaire includes useful examples of metabolic, signalling and gene expression events visualization
- make yourself familiar with several visual languages at the same time
- the result of these survey will help developing and evaluating graphical languages for unambiguous network representation

This survey compares three graphical languages:

Systems Biology Graphical Notation Process Description (SBGN PD) (Le Novere et al, 2009)

Modified Edinburgh Pathway Notation (mEPN) (Freeman et al, 2010)

Unified Notation (UN) (Mazein et al, 2011, in press)

Depends on the level of expertise the questionnaire will take **from 20 to 30 min**

References:

Le Novere N, Hucka M, Mi H, Moodie S, Schreiber F, Sorokin A, Demir E, Wegner K, Aladjem MI, Wimalaratne SM, Bergman FT, Gauges R, Ghazal P, Kawaji H, Li L, Matsuoka Y, Villeger A, Boyd SE, Calzone L, Courtot M et al (2009) The Systems Biology Graphical Notation. *Nat Biotechnol* 27: 735-741

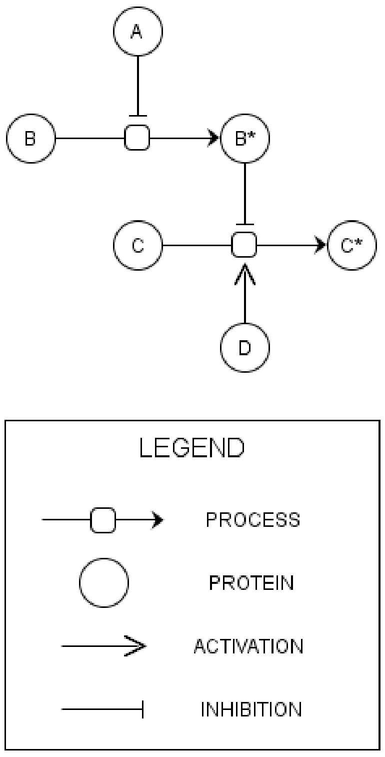
Freeman TC, Raza S, Theocharidis A, Ghazal P. (2010) The mEPN scheme: an intuitive and flexible graphical system for rendering biological pathways. *BMC Syst Biol*, Vol. 4, p. 65.

Mazein A, Sorokin A, Adams R, Moodie S, Golyanin I (2011) Visual Knowledge Management. International Journal of Knowledge Engineering and Data Mining (IJKEDM), (in press)

Test

So that the results of this questionnaire can be included into the dataset please complete the test bellow:

Please circle the correct answers that correspond to the diagram.



Protein A activates protein B	TRUTH	FALSE	NOT SURE
Protein B activates protein C	TRUTH	FALSE	NOT SURE
Protein A activates protein C	TRUTH	FALSE	NOT SURE
Protein D activates protein C	TRUTH	FALSE	NOT SURE
Protein A inhibits protein D	TRUTH	FALSE	NOT SURE

OPTIONAL COMMENTS:

Instructions

While answering the questions in the survey please refer to the legends on the last page.

In each section you will find that same events are represented in three different graphical notations. Please first take a look at all three diagrams, compare them and after that rate the features of the notations.

Please rate your answers as follows:

Strongly disagree 1 2 Disagree 3 4 5 6 Not sure 7 8 Agree 9 10 Strongly agree

For example, if strongly disagree please circle 1 and if strongly agree please circle 10, if not sure circle 5 or 6.

You should spend approximately 2 minutes on each of the next 4 pages and approximately 4-6 minutes on each of the last 3 pages.

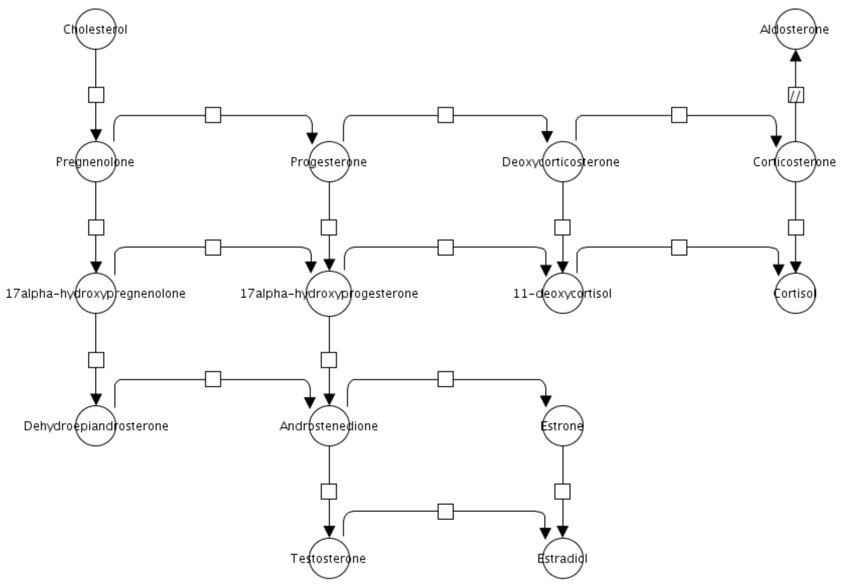
If you notice that any of the languages is misrepresented or there are mistakes in using the notation, please comment on the corresponding diagram.

Please return the questionnaire by 1st November 2010.

Please send the complete questionnaire by email or post using contact information bellow.

Email: Alexander Mazein a.mazein@sms.ed.ac.uk
Address: Alexander Mazein, University of Edinburgh, Informatics Forum - Room 1.43,
10 Crichton Street, Edinburgh EH8 9AB, United Kingdom

A SBGN PD

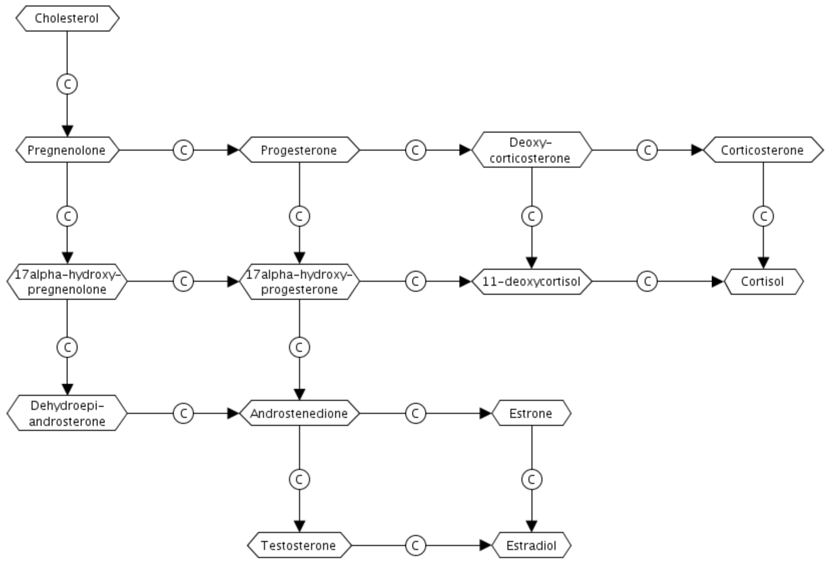


How would you rate METABOLIC PATHWAY representation in this notation?

Extremely poor 1 2 3 4 5 6 7 8 9 10 Excellent

COMMENTS

B mEPN

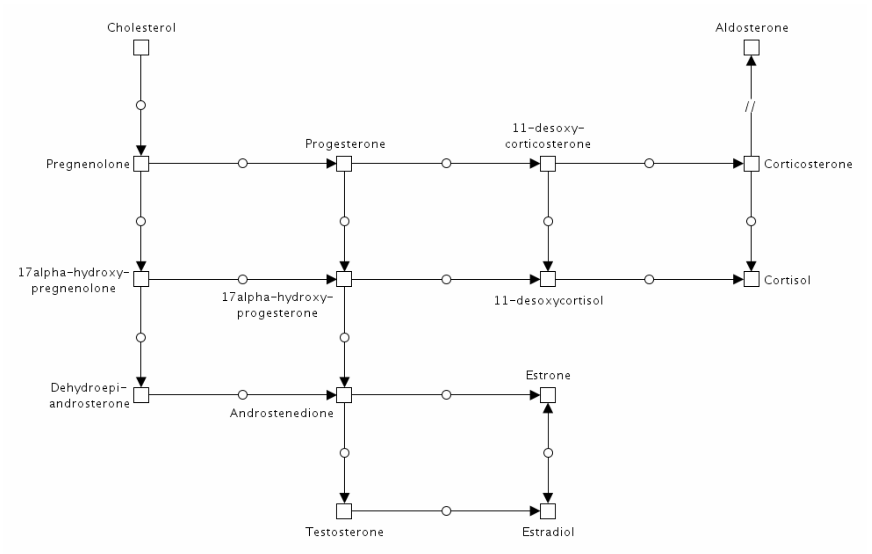


How would you rate METABOLIC PATHWAY representation in this notation?

Extremely poor 1 2 3 4 5 6 7 8 9 10 Excellent

COMMENTS

C UN

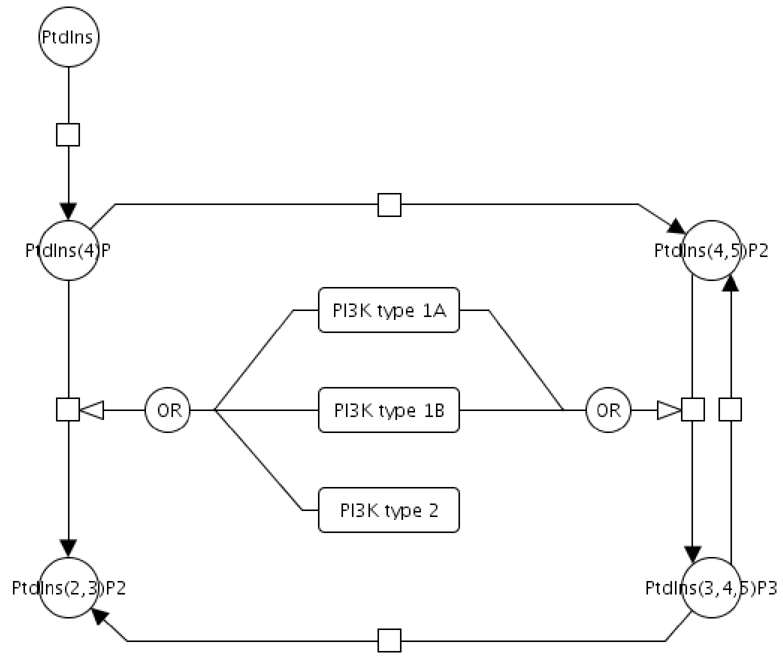


How would you rate METABOLIC PATHWAY representation in this notation?

Extremely poor 1 2 3 4 5 6 7 8 9 10 Excellent

COMMENTS

Ⓐ SBGN PD

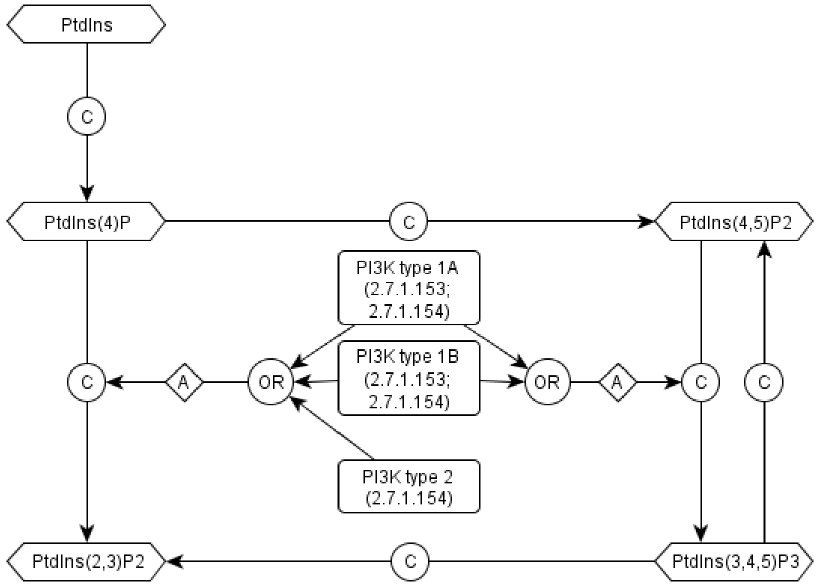


How would you rate ENZYME FUNCTION representation in this notation?

Extremely poor 1 2 3 4 5 6 7 8 9 10 Excellent

COMMENTS:

Ⓑ mEPN

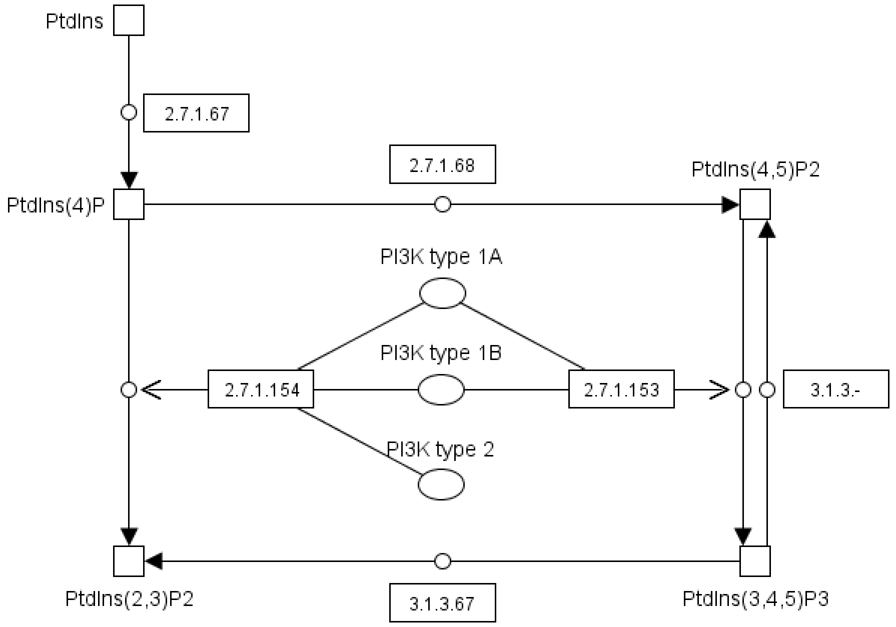


How would you rate ENZYME FUNCTION representation in this notation?

Extremely poor 1 2 3 4 5 6 7 8 9 10 Excellent

COMMENTS:

Ⓒ UN

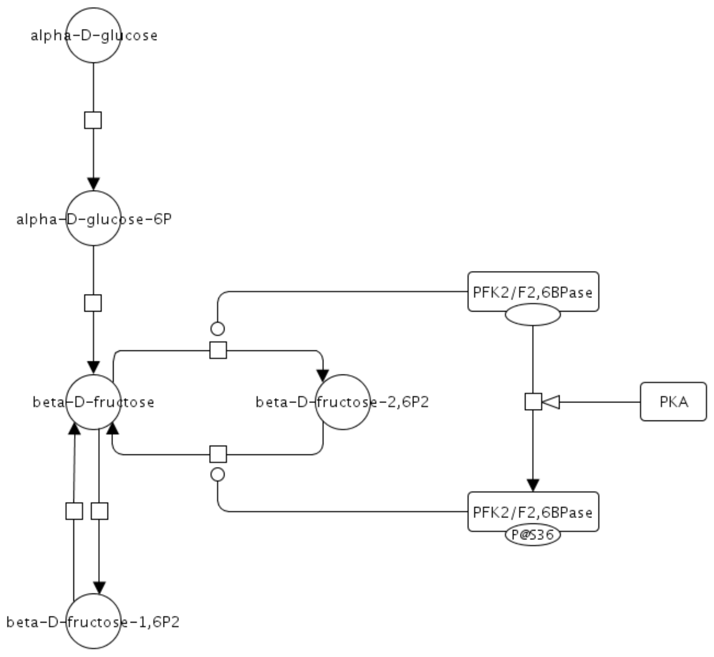


How would you rate ENZYME FUNCTION representation in this notation?

Extremely poor 1 2 3 4 5 6 7 8 9 10 Excellent

COMMENTS:

(A) SBGN PD

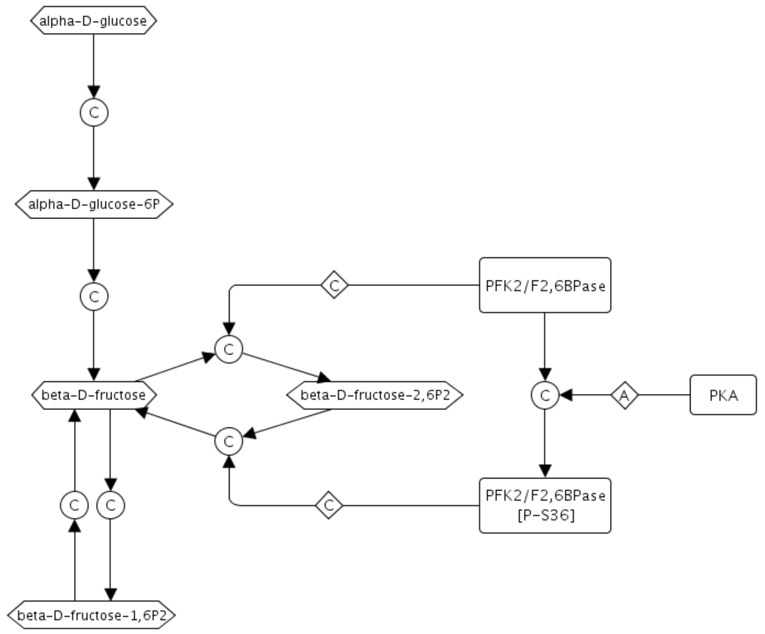


How would you rate the representation of METABOLISM REGULATION in this notation?

Extremely poor 1 2 3 4 5 6 7 8 9 10 Excellent

COMMENTS:

(B) mEPN

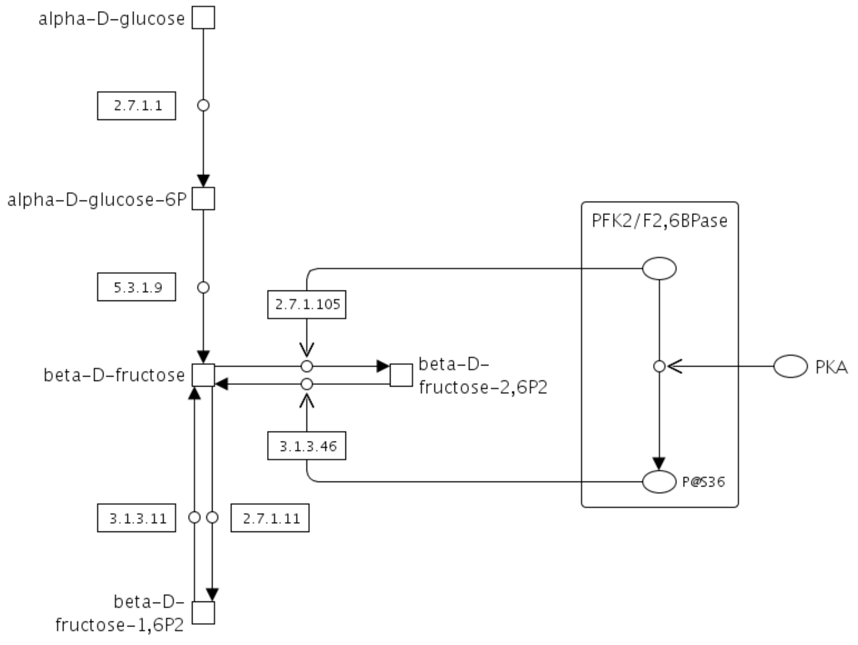


How would you rate the representation of METABOLISM REGULATION in this notation?

Extremely poor 1 2 3 4 5 6 7 8 9 10 Excellent

COMMENTS:

(C) UN

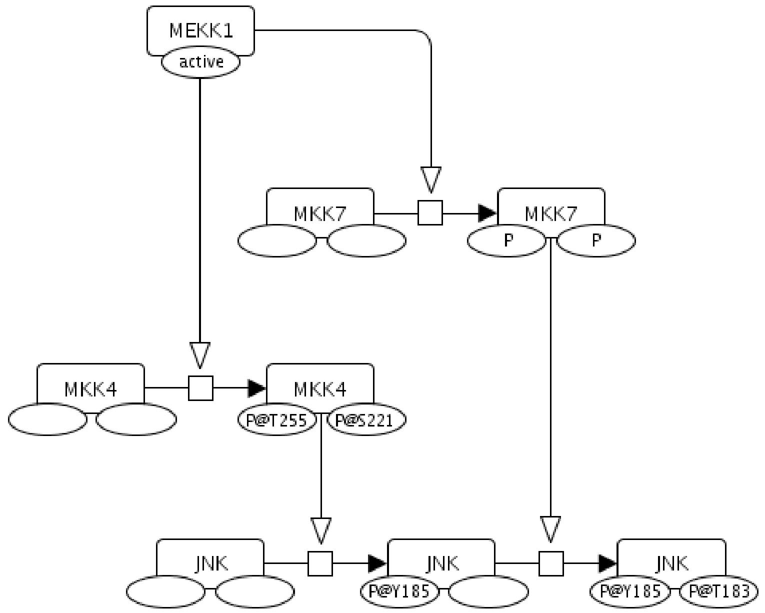


How would you rate the representation of METABOLISM REGULATION in this notation?

Extremely poor 1 2 3 4 5 6 7 8 9 10 Excellent

COMMENTS:

A SBGN PD

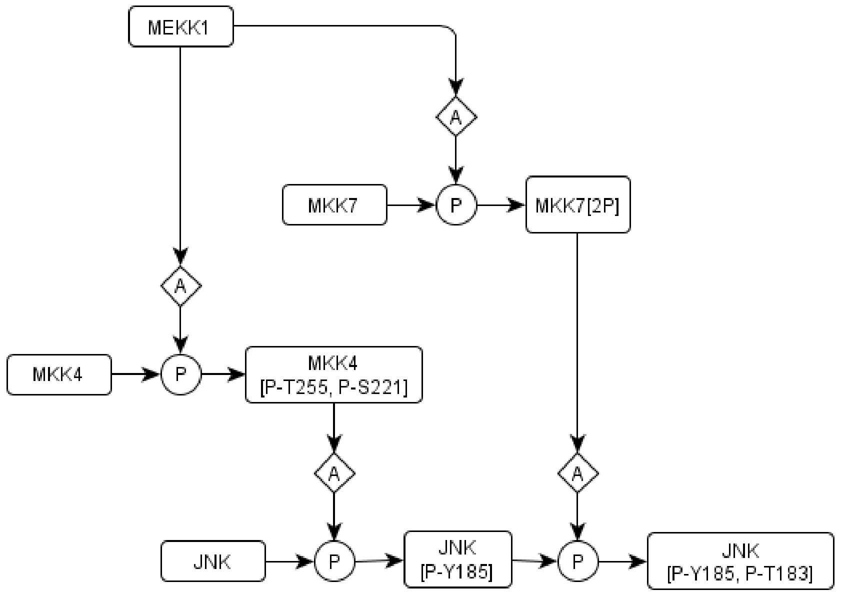


How would you rate SIGNALLING representation in this notation?

Extremely poor 1 2 3 4 5 6 7 8 9 10 Excellent

COMMENTS:

B mEPN

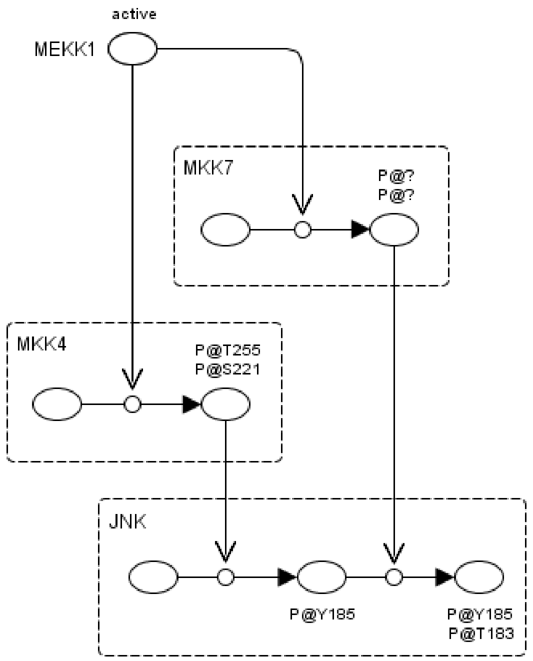


How would you rate SIGNALLING representation in this notation?

Extremely poor 1 2 3 4 5 6 7 8 9 10 Excellent

COMMENTS:

C UN

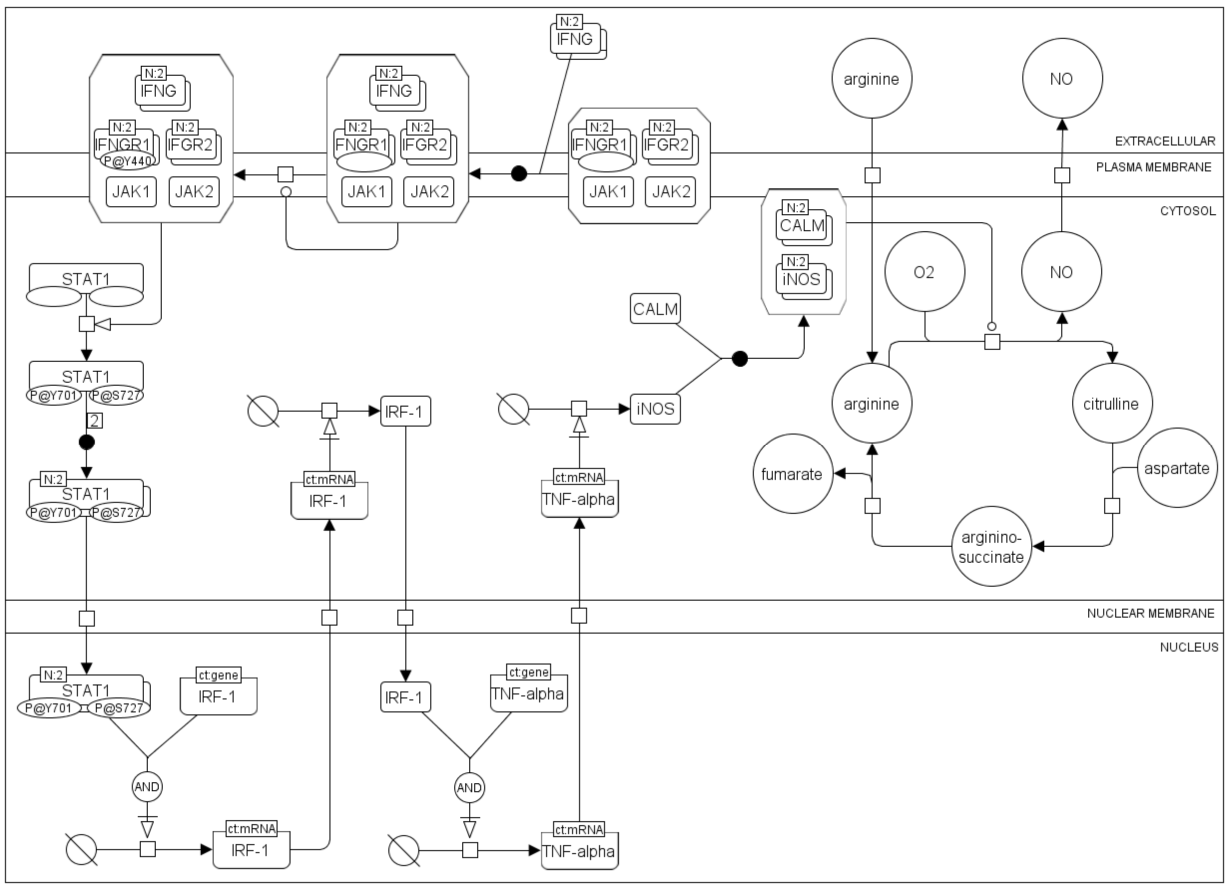


How would you rate SIGNALLING representation in this notation?

Extremely poor 1 2 3 4 5 6 7 8 9 10 Excellent

COMMENTS:

A SBGN PD



Is it easy to read and understand the diagram?

Very difficult 1 2 3 4 5 6 7 8 9 10 Very easy

How would you rate COMPACTNESS of this notation?

Extremely poor 1 2 3 4 5 6 7 8 9 10 Excellent

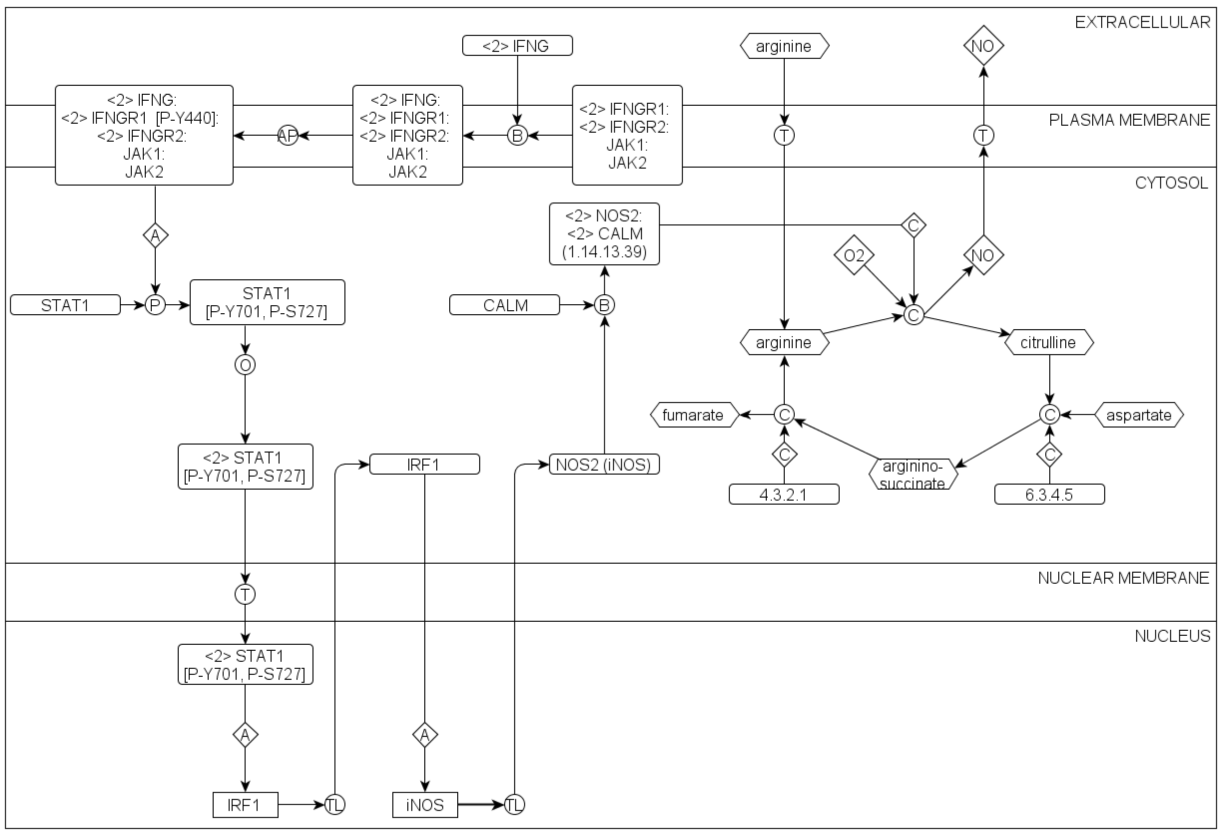
How would you rate COMPLEX representation in this notation?

Extremely poor 1 2 3 4 5 6 7 8 9 10 Excellent

How would you rate GENE EXPRESSION representation in this notation?

Extremely poor 1 2 3 4 5 6 7 8 9 10 Excellent

B mEPN



Is it easy to read and understand the diagram?

Very difficult 1 2 3 4 5 6 7 8 9 10 Very easy

How would you rate COMPACTNESS of this notation?

Extremely poor 1 2 3 4 5 6 7 8 9 10 Excellent

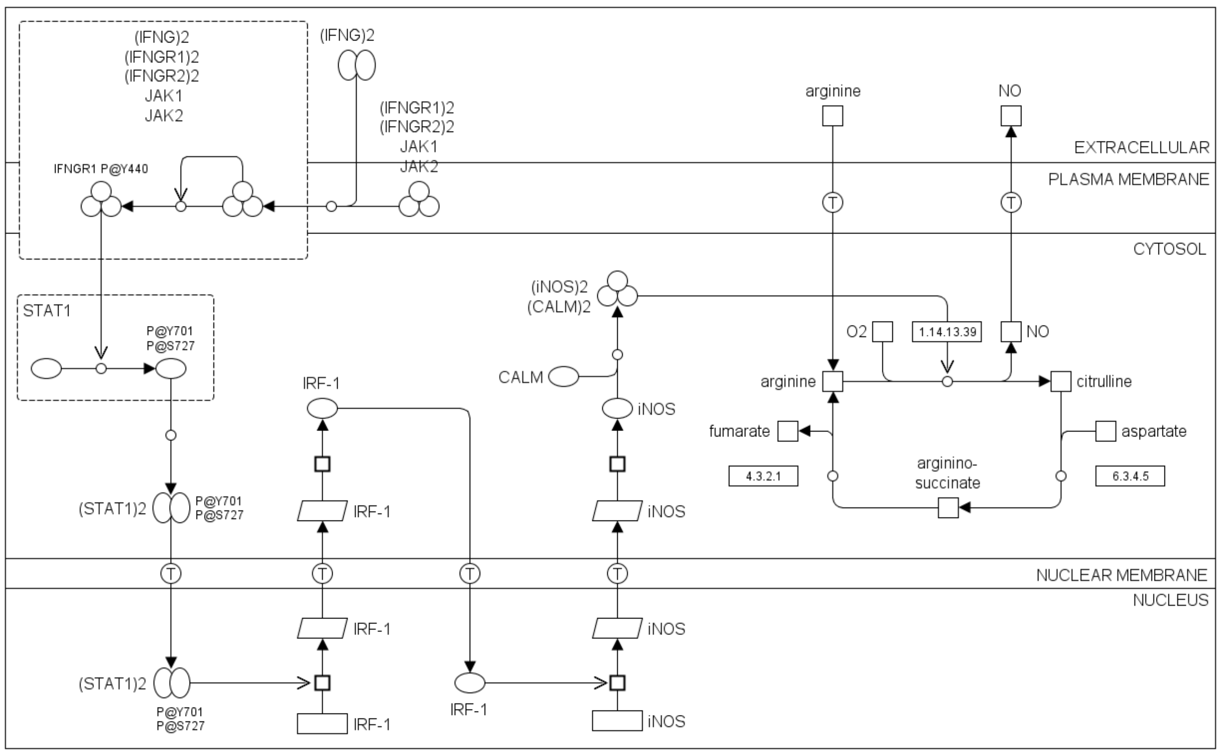
How would you rate COMPLEX representation in this notation?

Extremely poor 1 2 3 4 5 6 7 8 9 10 Excellent

How would you rate GENE EXPRESSION representation in this notation?

Extremely poor 1 2 3 4 5 6 7 8 9 10 Excellent

C UN



Is it easy to read and understand the diagram?

Very difficult 1 2 3 4 5 6 7 8 9 10 Very easy

How would you rate COMPACTNESS of this notation?

Extremely poor 1 2 3 4 5 6 7 8 9 10 Excellent

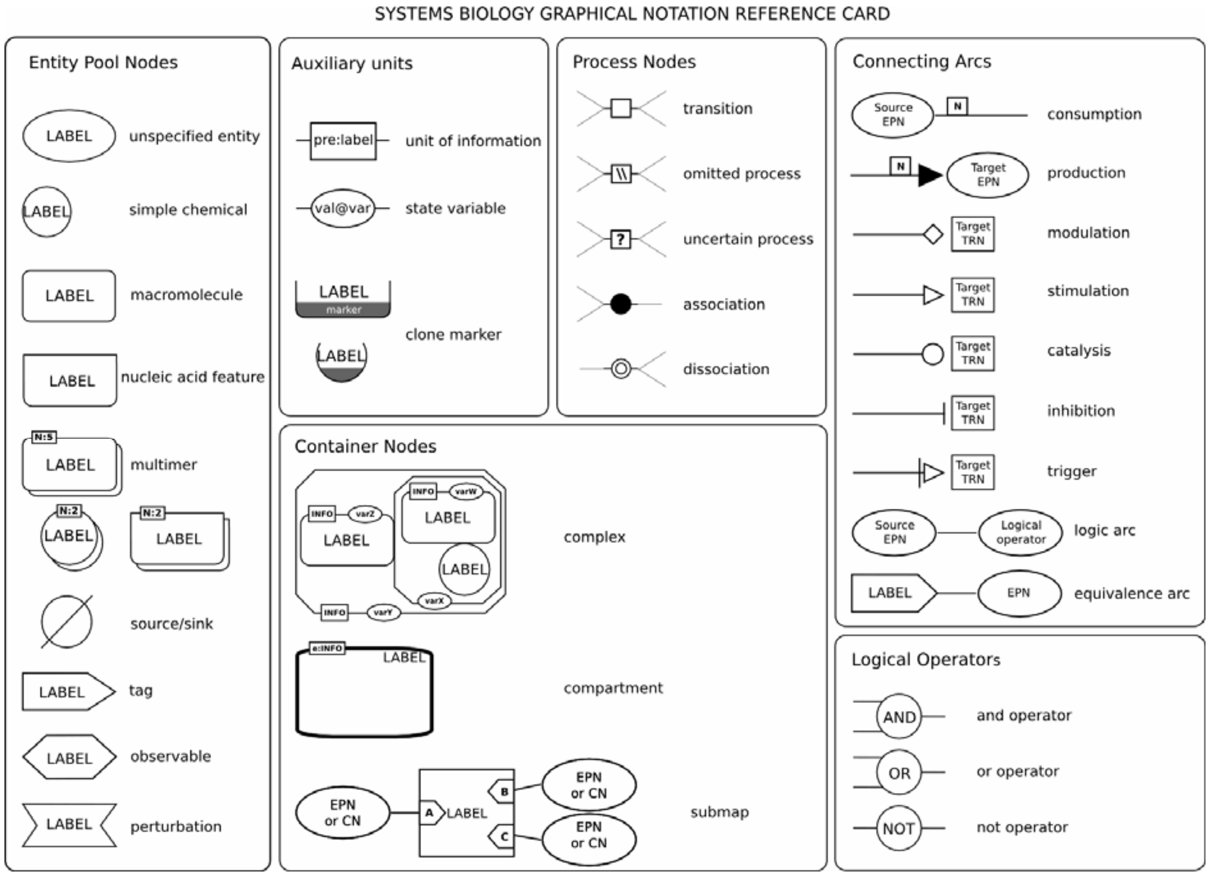
How would you rate COMPLEX representation in this notation?

Extremely poor 1 2 3 4 5 6 7 8 9 10 Excellent

How would you rate GENE EXPRESSION representation in this notation?

Extremely poor 1 2 3 4 5 6 7 8 9 10 Excellent

A SBGN PD



How would you rate the NUMBER OF SYMBOLS used in this notation?

Too few
or too many 1 2 3 4 5 6 7 8 9 10 Sufficient

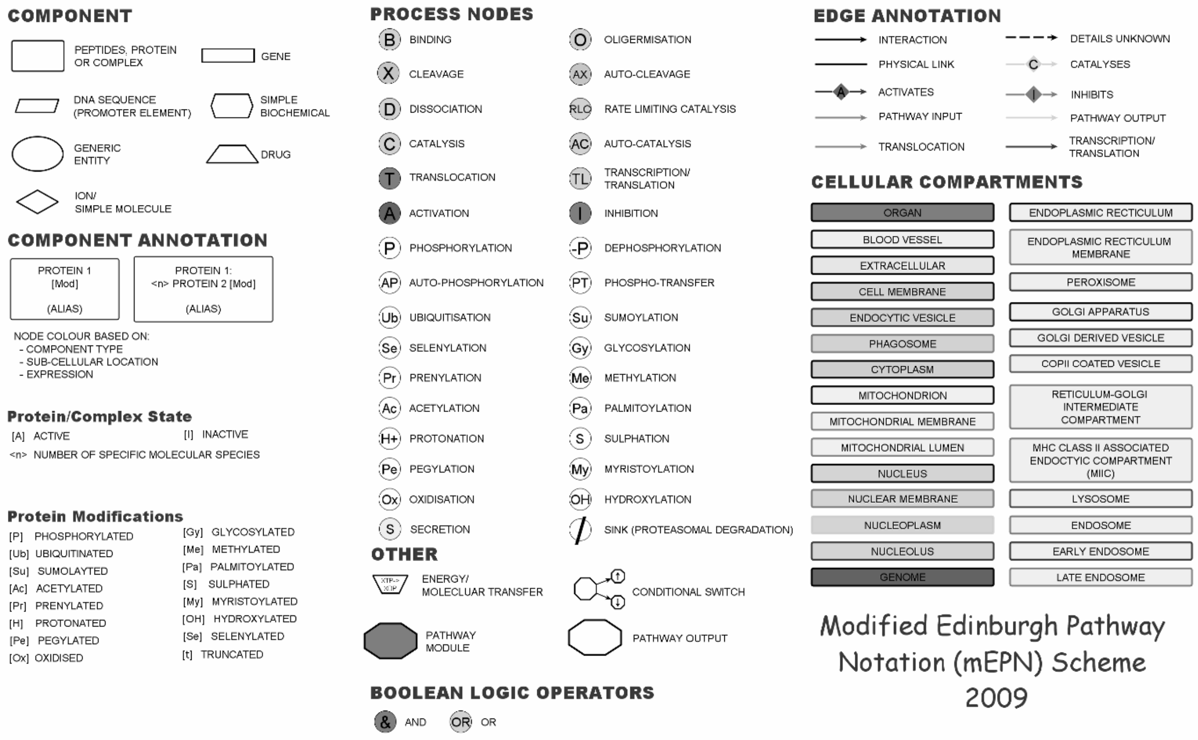
How would you rate ability of this notation to represent INCOMPLETE/OMITTED information?

Too few
or too many 1 2 3 4 5 6 7 8 9 10 Sufficient

Would you agree that it is fairly easy to LEARN/REMEMBER this system of symbols?

Strongly
disagree 1 2 3 4 5 6 7 8 9 10 Strongly agree

B mEPN



How would you rate the NUMBER OF SYMBOLS used in this notation?

Too few
or too many 1 2 3 4 5 6 7 8 9 10 Sufficient

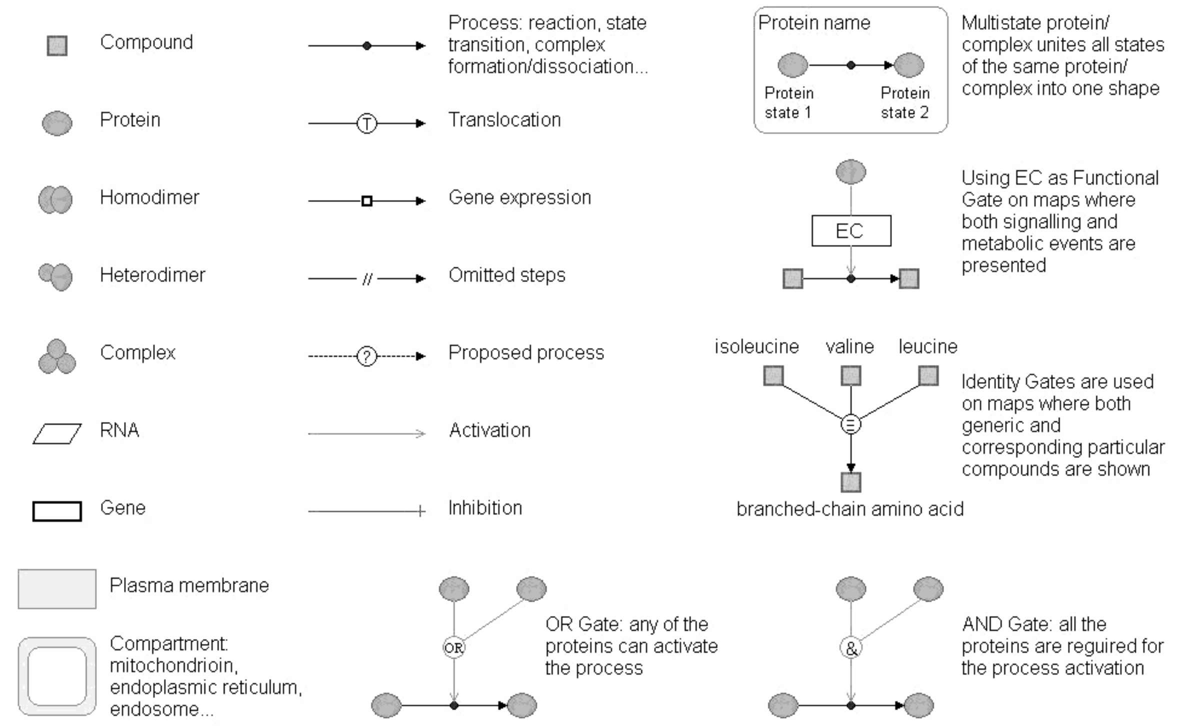
How would you rate ability of this notation to represent INCOMPLETE/OMITTED information?

Too few
or too many 1 2 3 4 5 6 7 8 9 10 Sufficient

Would you agree that it is fairly easy to LEARN/REMEMBER this system of symbols?

Strongly
disagree 1 2 3 4 5 6 7 8 9 10 Strongly agree

C UN


Compound
Protein
Homodimer
Heterodimer
Complex
RNA
Gene
Plasma membrane
Compartment: mitochondrioin, endoplasmic reticulum, endosome...
Process: reaction, state transition, complex formation/dissociation...
T
Translocation
Gene expression
//
Omitted steps
?
Proposed process
Activation
Inhibition
Protein name
Protein state 1
Protein state 2
Multistate protein/ complex unites all states of the same protein/ complex into one shape
EC
Using EC as Functional Gate on maps where both signalling and metabolic events are presented
isoleucine
valine
leucine
=
branched-chain amino acid
Identity Gates are used on maps where both generic and corresponding particular compounds are shown
OR
OR Gate: any of the proteins can activate the process
&
AND Gate: all the proteins are required for the process activation

How would you rate the NUMBER OF SYMBOLS used in this notation?

Too few
or too many 1 2 3 4 5 6 7 8 9 10 Sufficient

How would you rate ability of this notation to represent INCOMPLETE/OMITTED information?

Too few
or too many 1 2 3 4 5 6 7 8 9 10 Sufficient

Would you agree that it is fairly easy to LEARN/REMEMBER this system of symbols?

Strongly
disagree 1 2 3 4 5 6 7 8 9 10 Strongly agree

Would you agree that the compactness is an important feature of a graphical language?

Strongly disagree 1 2 Disagree 3 4 5 Not sure 6 7 8 Agree 9 10 Strongly agree

OPTIONAL COMMENTS:

Would you agree that it is important to be able to show EC numbers on a metabolic or metabolism regulation diagram?

Strongly disagree 1 2 Disagree 3 4 5 Not sure 6 7 8 Agree 9 10 Strongly agree

OPTIONAL COMMENTS:

Would you agree that it is important to be able to represent GENERIC-SPECIFIC RELATIONSHIPS on a diagram? Example of generic entity: “amino acid”. Corresponding specific entities are particular amino acids, for example alanine, serine etc.

Strongly disagree 1 2 Disagree 3 4 5 Not sure 6 7 8 Agree 9 10 Strongly agree

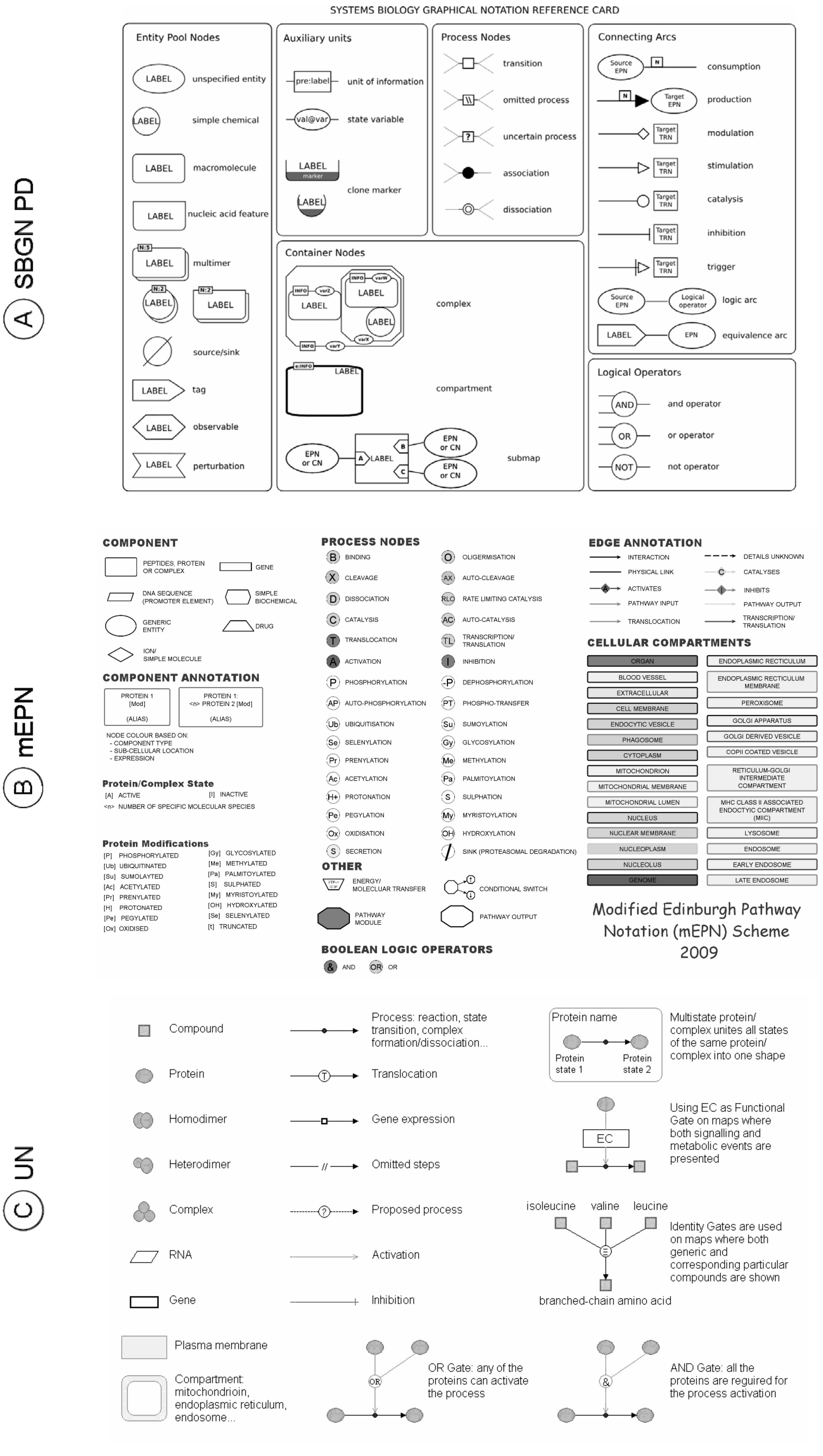
OPTIONAL COMMENTS:

What type of representation do you prefer: “label outside the shape” or “label inside the shape”?

Label outside the shape 1 2 3 4 5 Not sure 6 7 8 9 10 Label inside the shape

OPTIONAL COMMENTS:

Thank you
for participating in the survey!


A SBGN PD
SYSTEMS BIOLOGY GRAPHICAL NOTATION REFERENCE CARD
Entity Pool Nodes
LABEL unspecified entity
LABEL simple chemical
LABEL macromolecule
LABEL nucleic acid feature
N:5 LABEL multimer
N:2 LABEL
N:2 LABEL
source/sink
LABEL tag
LABEL observable
LABEL perturbation
Auxiliary units
pre:label unit of information
val@var state variable
LABEL marker
LABEL
clone marker
Process Nodes
transition
omitted process
? uncertain process
association
dissociation
Container Nodes
INFO var2 LABEL
INFO varW LABEL
LABEL
INFO varY varX
complex
e:INFO LABEL
compartment
EPN or CN
A LABEL B C
EPN or CN
EPN or CN
submap
Connecting Arcs
Source EPN N consumption
N Target EPN production
Target TRN modulation
Target TRN stimulation
Target TRN catalysis
Target TRN inhibition
Target TRN trigger
Source EPN Logical operator logic arc
LABEL EPN equivalence arc
Logical Operators
AND and operator
OR or operator
NOT not operator
B mEPN
COMPONENT
PEPTIDES, PROTEIN OR COMPLEX
GENE
DNA SEQUENCE (PROMOTER ELEMENT)
SIMPLE BIOCHEMICAL
GENERIC ENTITY
DRUG
ION/ SIMPLE MOLECULE
COMPONENT ANNOTATION
PROTEIN 1 [Mod] (ALIAS)
PROTEIN 1: <n> PROTEIN 2 [Mod] (ALIAS)
NODE COLOUR BASED ON:
- COMPONENT TYPE
- SUB-CELLULAR LOCATION
- EXPRESSION
Protein/Complex State
[A] ACTIVE
[I] INACTIVE
<n> NUMBER OF SPECIFIC MOLECULAR SPECIES
Protein Modifications
[P] PHOSPHORYLATED
[Ub] UBIQUITINATED
[Su] SUMOLAYTED
[Ac] ACETYLATED
[Pr] PRENYLATED
[H] PROTONATED
[Pe] PEGYLATED
[Ox] OXIDISED
[Gy] GLYCOSYLATED
[Me] METHYLATED
[Pa] PALMITOYLATED
[S] SULPHATED
[My] MYRISTOYLATED
[OH] HYDROXYLATED
[Se] SELENYLATED
[t] TRUNCATED
PROCESS NODES
B BINDING
X CLEAVAGE
D DISSOCIATION
C CATALYSIS
T TRANSLOCATION
A ACTIVATION
P PHOSPHORYLATION
AP AUTO-PHOSPHORYLATION
Ub UBIQUITISATION
Se SELENYLATION
Pr PRENYLATION
Ac ACETYLATION
H+ PROTONATION
Pe PEGYLATION
Ox OXIDISATION
S SECRETION
O OLIGERMISATION
AX AUTO-CLEAVAGE
RLC RATE LIMITING CATALYSIS
AC AUTO-CATALYSIS
TL TRANSCRIPTION/ TRANSLATION
I INHIBITION
-P DEPHOSPHORYLATION
PT PHOSPHO-TRANSFER
Su SUMOYLATION
Gy GLYCOSYLATION
Me METHYLATION
Pa PALMITOYLATION
S SULPHATION
My MYRISTOYLATION
OH HYDROXYLATION
SINK (PROTEASOMAL DEGRADATION)
OTHER
XTP-> XDP ENERGY/ MOLECLUAR TRANSFER
CONDITIONAL SWITCH
PATHWAY MODULE
PATHWAY OUTPUT
BOOLEAN LOGIC OPERATORS
& AND
OR OR
EDGE ANNOTATION
INTERACTION
PHYSICAL LINK
A ACTIVATES
PATHWAY INPUT
TRANSLOCATION
DETAILS UNKNOWN
C CATALYSES
I INHIBITS
PATHWAY OUTPUT
TRANSCRIPTION/ TRANSLATION
CELLULAR COMPARTMENTS
ORGAN
BLOOD VESSEL
EXTRACELLULAR
CELL MEMBRANE
ENDOCYTIC VESICLE
PHAGOSOME
CYTOPLASM
MITOCHONDRION
MITOCHONDRIAL MEMBRANE
MITOCHONDRIAL LUMEN
NUCLEUS
NUCLEAR MEMBRANE
NUCLEOPLASM
NUCLEOLUS
GENOME
ENDOPLASMIC RECTICULUM
ENDOPLASMIC RECTICULUM MEMBRANE
PEROXISOME
GOLGI APPARATUS
GOLGI DERIVED VESICLE
COPII COATED VESICLE
RETICULUM-GOLGI INTERMEDIATE COMPARTMENT
MHC CLASS II ASSOCIATED ENDOCTYIC COMPARTMENT (MIIC)
LYSOSOME
ENDOSOME
EARLY ENDOSOME
LATE ENDOSOME
Modified Edinburgh Pathway Notation (mEPN) Scheme 2009
C UN
Compound
Protein
Homodimer
Heterodimer
Complex
RNA
Gene
Plasma membrane
Compartment: mitochondrioin, endoplasmic reticulum, endosome...
Process: reaction, state transition, complex formation/dissociation...
T Translocation
Gene expression
// Omitted steps
? Proposed process
Activation
Inhibition
Protein name
Protein state 1
Protein state 2
Multistate protein/ complex unites all states of the same protein/ complex into one shape
EC
Using EC as Functional Gate on maps where both signalling and metabolic events are presented
isoleucine
valine
leucine
=
branched-chain amino acid
Identity Gates are used on maps where both generic and corresponding particular compounds are shown
OR
OR Gate: any of the proteins can activate the process
&
AND Gate: all the proteins are reguired for the process activation

Appendix III

Part Q1

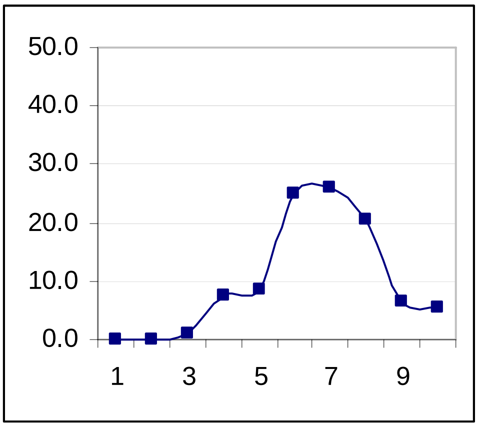


Figure Q1A. A plot of the SBGN PD rating distribution

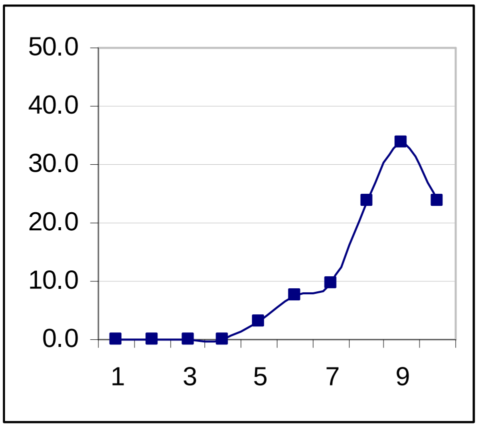


Figure Q1B. A plot of the mEPN rating distribution

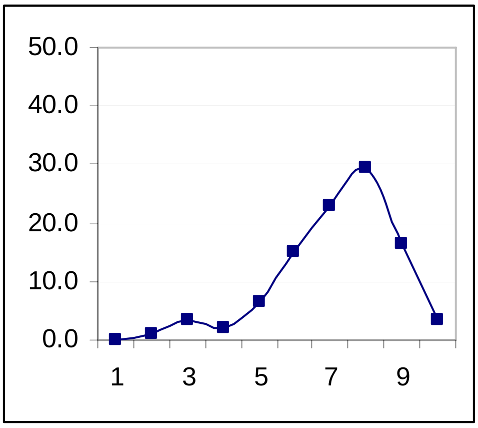


Figure Q1C. A plot of the UN rating distribution

Part Q2

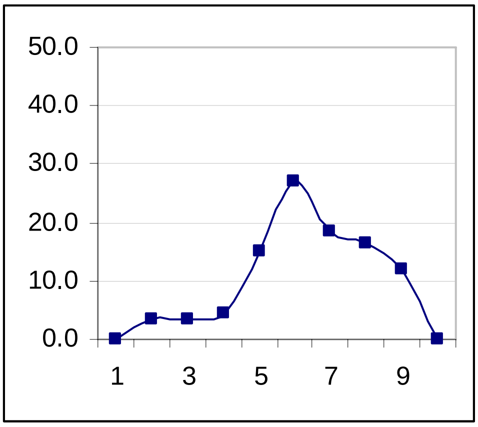


Figure Q2A. A plot of the SBGN PD rating distribution

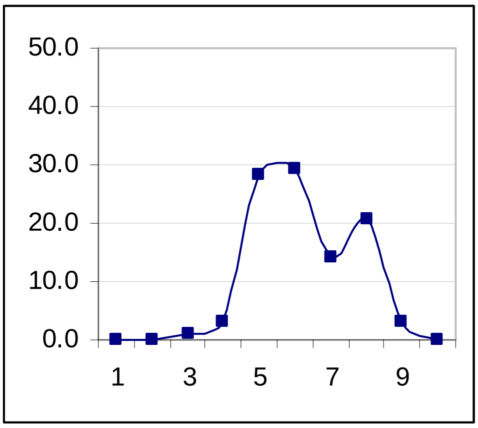


Figure Q2B. A plot of the mEPN rating distribution

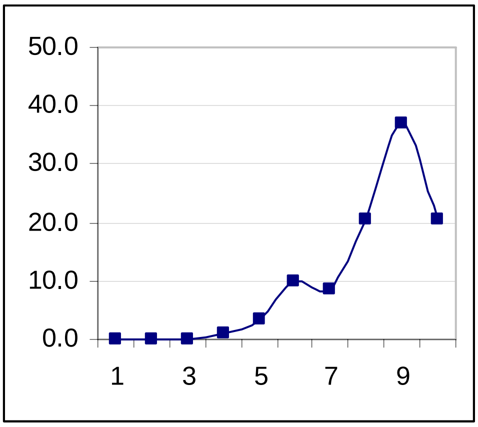


Figure Q2C. A plot of the UN rating distribution

Part Q3

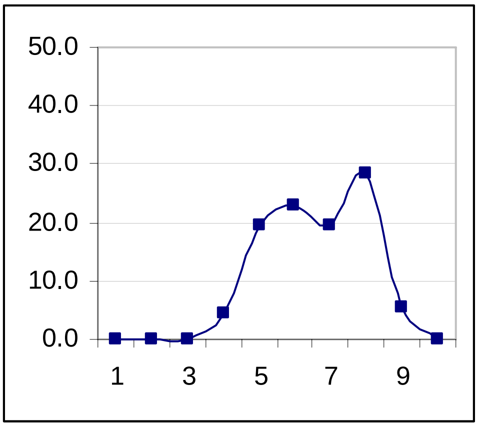


Figure Q3A. A plot of the SBGN PD rating distribution

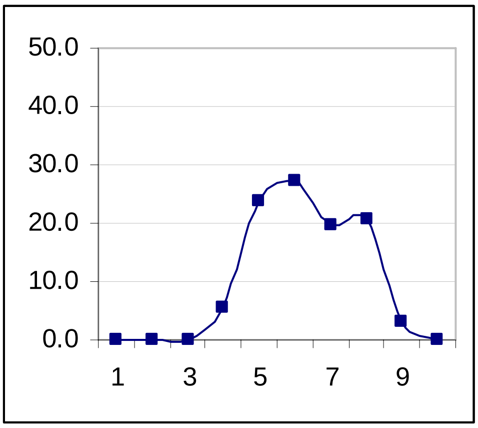


Figure Q3B. A plot of the mEPN rating distribution

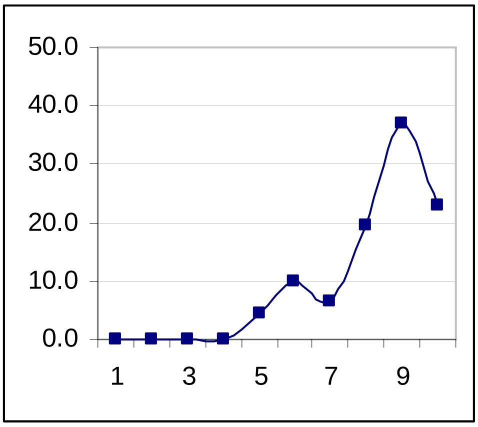


Figure Q3C. A plot of the UN rating distribution

Part Q4

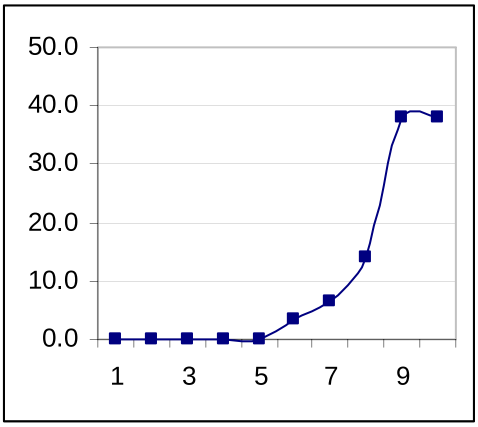


Figure Q4A. A plot of the SBGN PD rating distribution

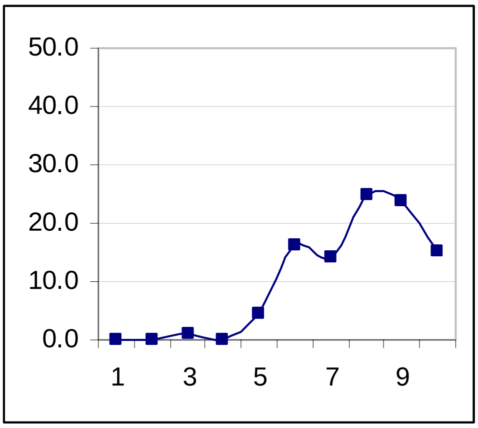


Figure Q4B. A plot of the mEPN rating distribution

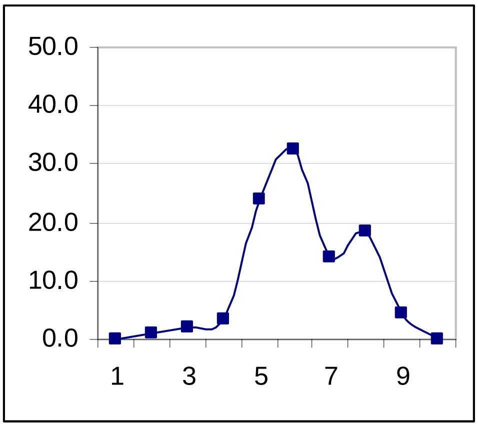


Figure Q4C. A plot of the UN rating distribution

Part Q5

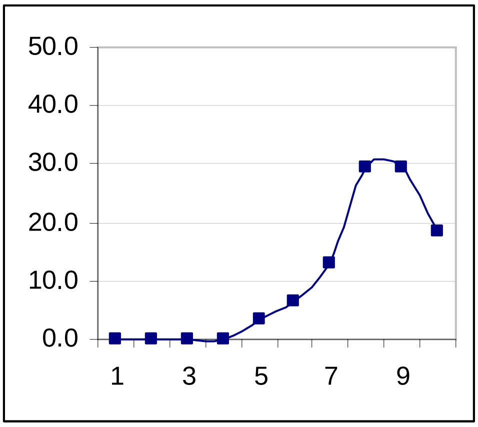


Figure Q5A. A plot of the SBGN PD rating distribution

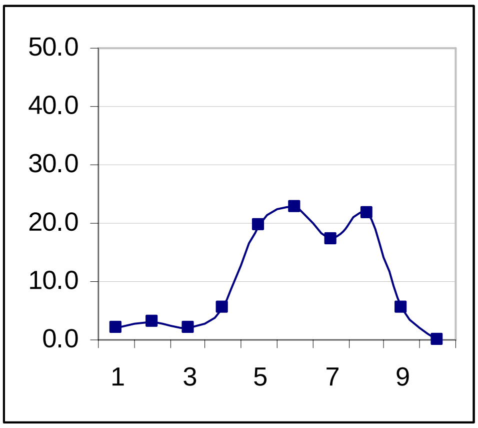


Figure Q5B. A plot of the mEPN rating distribution

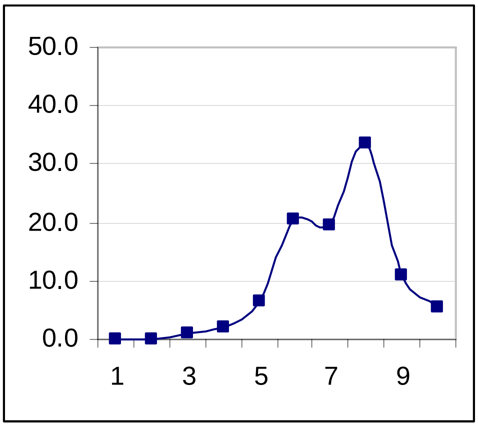


Figure Q5C. A plot of the UN rating distribution

Part Q6

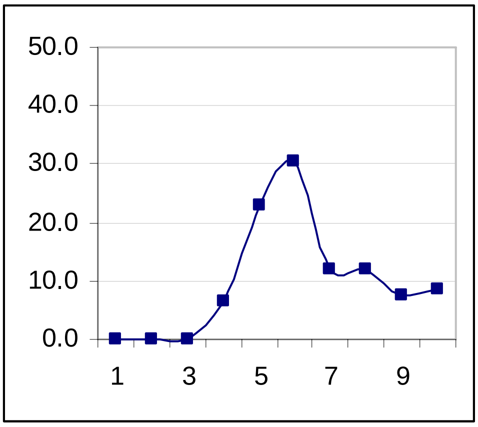


Figure Q6A. A plot of the SBGN PD rating distribution

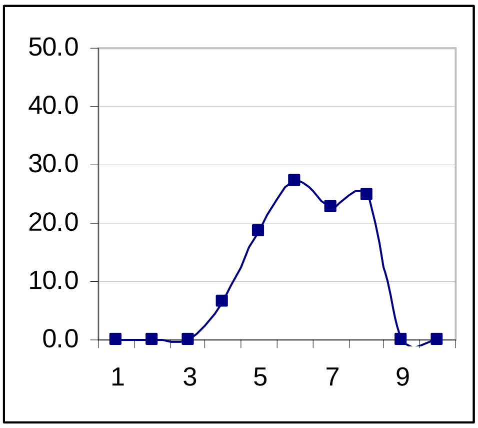


Figure Q6B. A plot of the mEPN rating distribution

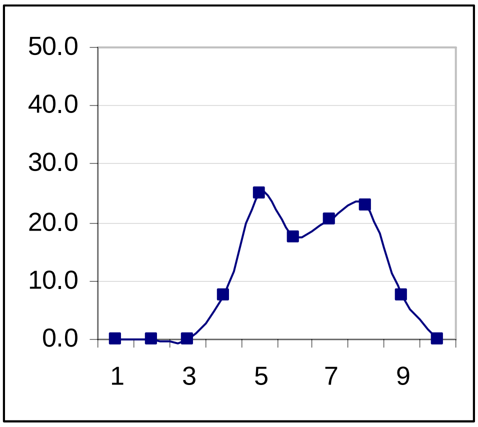


Figure Q6C. A plot of the UN rating distribution

Part Q7

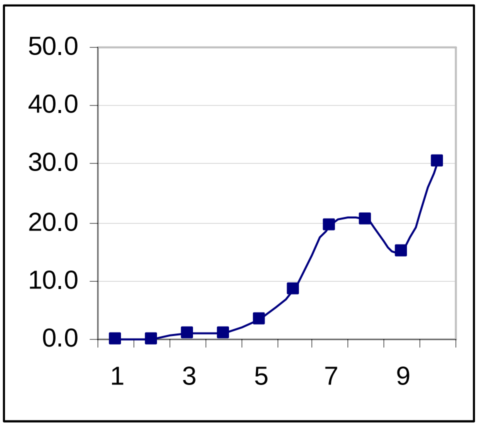


Figure Q7A. A plot of the SBGN PD rating distribution

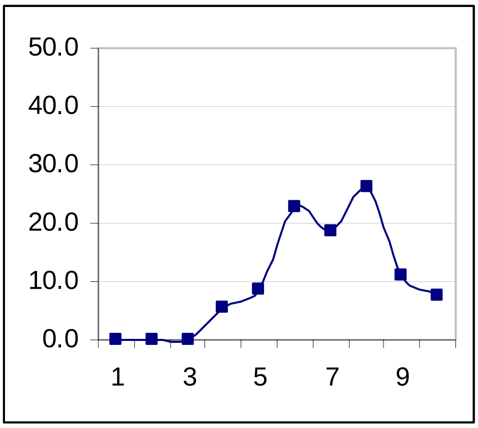


Figure Q7B. A plot of the mEPN rating distribution

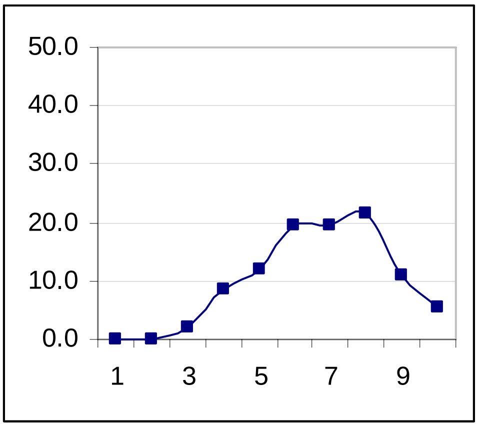


Figure Q7C. A plot of the UN rating distribution

Part Q8

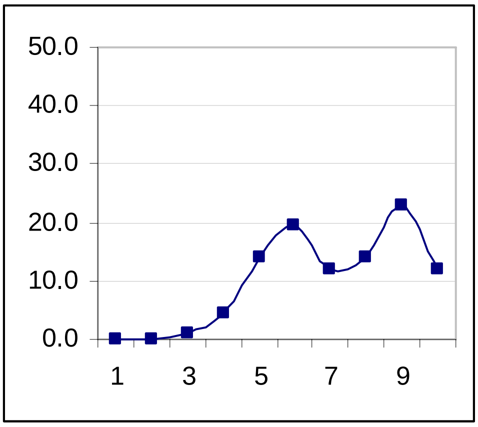


Figure Q8A. A plot of the SBGN PD rating distribution

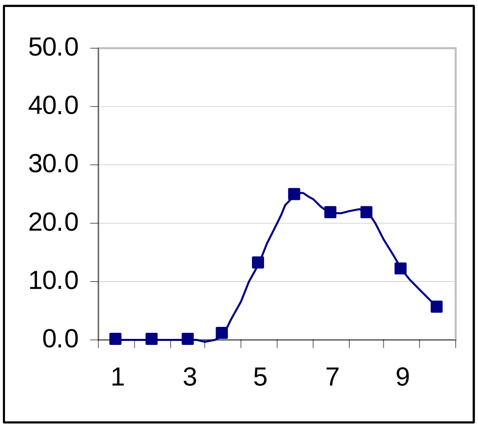


Figure Q8B. A plot of the mEPN rating distribution

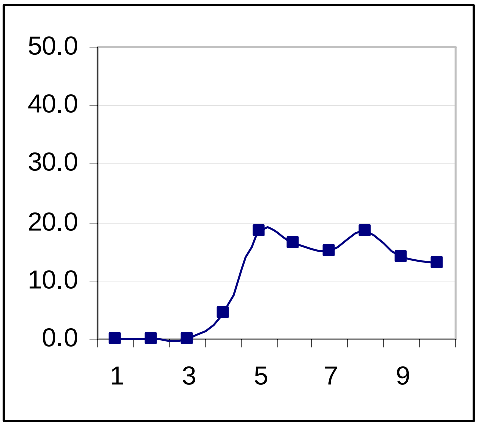


Figure Q8C. A plot of the UN rating distribution

Part Q9

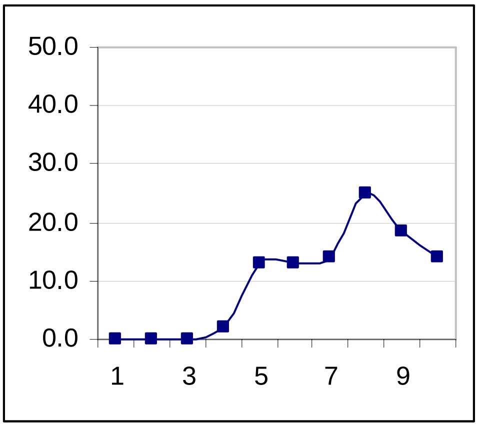


Figure Q9A. A plot of the SBGN PD rating distribution

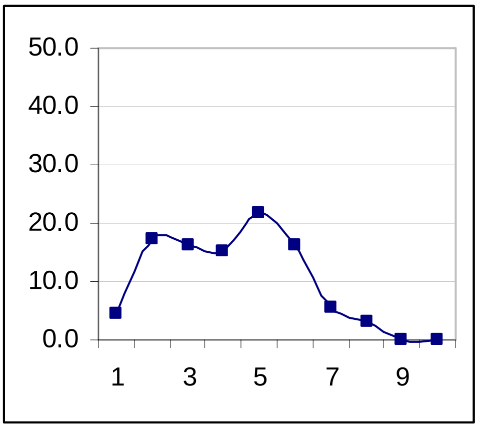


Figure Q9B. A plot of the mEPN rating distribution

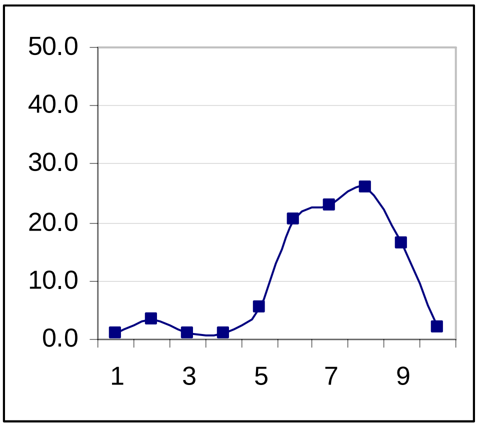


Figure Q9C. A plot of the UN rating distribution

Part Q10

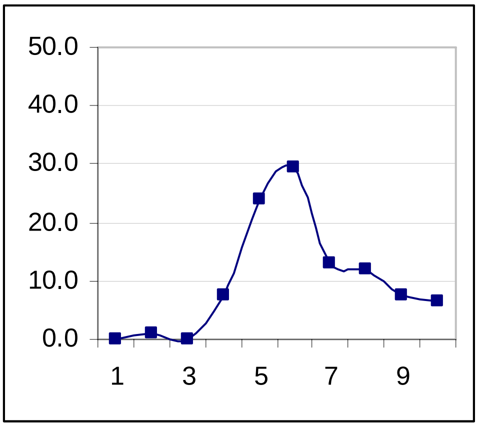


Figure Q10A. A plot of the SBGN PD rating distribution

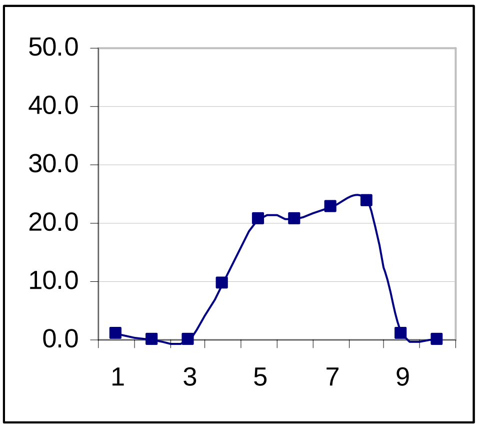


Figure Q10B. A plot of the mEPN rating distribution

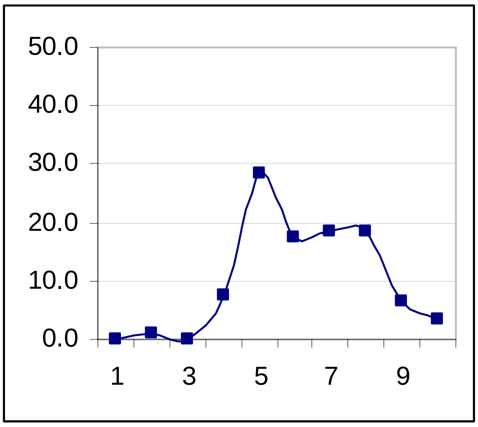


Figure Q10C. A plot of the UN rating distribution

Part Q11

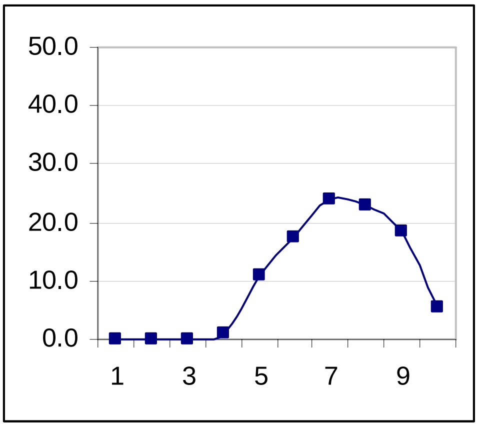


Figure Q11A. A plot of the SBGN PD rating distribution

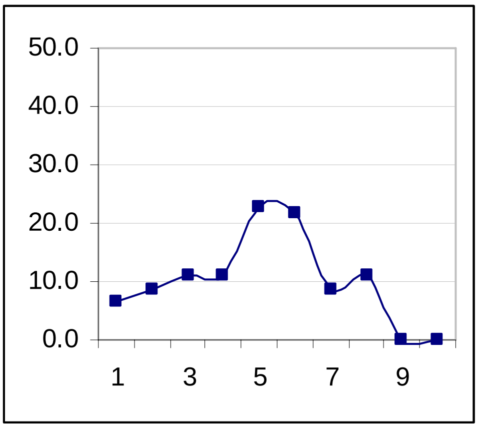


Figure Q11B. A plot of the mEPN rating distribution

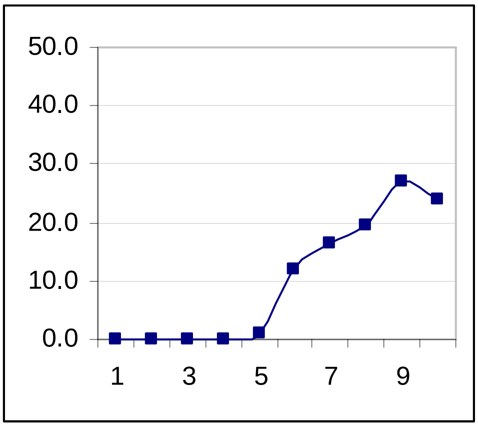


Figure Q11C. A plot of the UN rating distribution