Available online at www.sciencedirect.com

ScienceDirect

Electronic Notes in Theoretical Computer Science 299 (2013) 5–13

Electronic Notes in Theoretical Computer Science

www.elsevier.com/locate/entcs

Representing and Comparing Metabolic Pathways as Petri Nets with MPath2PN and CoMeta

Paolo Baldan [1]

Dipartimento di Matematica
Università di Padova,
Padova, Italy

Nicoletta Cocco and Marta Simeoni [2]

Dipartimento di Scienze Ambientali, Informatica e Statistica,
Università Ca' Foscari di Venezia
Venezia, Italy

Abstract

We present two tools, *MPath2PN* and *CoMeta*, which are part of an ongoing project for representing and comparing metabolic pathways as Petri Nets. *MPath2PN* is intended to support an automatic translation of metabolic pathways from the major biological databases into corresponding Petri nets expressed in the input formalisms for Petri net tools. *CoMeta* is devised to compare metabolic pathways of different organisms through their Petri net representation produced by *MPath2PN*. *CoMeta* automatically takes the data from the KEGG database and, in the comparison, it considers both homology of reactions and behavioural aspects of the pathways as expressed by the T-invariants of the underlying Petri nets.

Keywords: Metabolic pathways, Petri nets, translation, comparison

1 Introduction

The metabolism is the chemical system which generates the amino acids, sugars, lipids, nucleic acids and energy necessary to the life of any organism. Metabolic pathways are subsystems dealing with specific functions. Their comparison in different species yields interesting information on evolution and it may help in understanding the associated functions. This is important for studying diseases and drugs

[1] Email: baldan@math.unipd.it

[2] Email: cocco@unive.it, simeoni@unive.it

http://dx.doi.org/10.1016/j.entcs.2013.11.002

design and for industrial applications. We present two tools which are part of an ongoing project for the representation and comparison of metabolic pathways as basic Petri nets. Preliminary prototypes of such tools have been presented respectively in [20] and in [21].

A metabolic pathway consists of a network of chemical reactions, catalysed by one or more enzymes, where some molecules (reactants or substrates) are transformed into others (products). Enzymes are not consumed in a reaction, even if they are necessary while the reaction takes place. The product of a reaction is the substrate of the next one. Quantitative relations can be represented through a stoichiometric matrix, where rows represent molecular species and columns represent reactions. An element of the matrix, a stoichiometric coefficient n_{ij}, represents the degree to which the i-th chemical species participates in the j-th reaction. The kinetic of a pathway is determined by the rate associated to each reaction. Information on metabolic pathways are collected in several databases. In particular the KEGG PATHWAY database (Kyoto Encyclopedia of Genes and Genomes) [8,25] contains metabolic, regulatory and genetic pathways for different species whose data are derived by genome sequencing. It integrates genomic, chemical and systemic functional information represented as maps which are linked to additional information on reactions, proteins and genes, possibly stored in other databases. KEGG pathways are coded using KGML (KEGG Markup Language) [7] based on XML. A web service for querying the system from users programs is available. KEGG provides a uniform view of the same pathway in different organisms which is particularly useful for comparison. There are other important free access repositories on metabolic pathways, e.g., the BioModels Database [2], whose biological models are coded in SBML (Systems Biology Markup Language) [15], and MetaCyc [10], which is part of BioCyc Database Collection [1].

Petri nets (PNs) [28,22] are a well studied formalism offering many tools for visualisation, simulation and analysis (see *The Petri net World site* [12]). They are well suited for metabolic pathways because of the natural correspondence between PNs and biochemical networks [33,32,23] and indeed they have been largely used for qualitative and quantitative modelling and analysis of metabolic pathways (see e.g. [19] for a survey). In a Petri net places represent molecular species, such as metabolites or enzymes; transitions correspond to chemical reactions; input places represent the substrate or reactants and output places reaction products. The incidence matrix of the PN is identical to the stoichiometric matrix, the number of tokens in each place indicates the amount of substance associated with that place, the flux modes and the conservation relations for metabolites correspond to specific properties of PNs. In particular minimal semi-positive T-invariants correspond to elementary flux modes [34] of a metabolic pathway, i.e., minimal sets of reactions that can operate at a steady state. Minimal T-invariants form a basis for the set of semi-positive T-invariant (Hilbert basis) which is unique and characteristic of the PN.

The rest of the paper is devoted to the presentation of the tools. In Section 2 we describe *MPath2PN*, a tool which supplies an automatic translation of the data on a metabolic pathway into a PN model. It is written in Java and it is conceived to deal with different databases in input, such as KEGG and the BioModels Database, and different PN tools formats in output. In Section 3 we describe *CoMeta*, a tool which automatically takes the data from the KEGG database and it compares metabolic pathways of different organisms relying on their PN representation. The comparison is based on a similarity measure which considers both homology of reactions and functional aspects of the pathways. The latter are captured by the semi-positive T-invariants of the PNs which correspond to potential fluxes at steady state. Some conclusions are drawn in Section 4.

The presented tools, *Mpath2PN* and *CoMeta* are freely available at the address `http://www.dsi.unive.it/~biolab`.

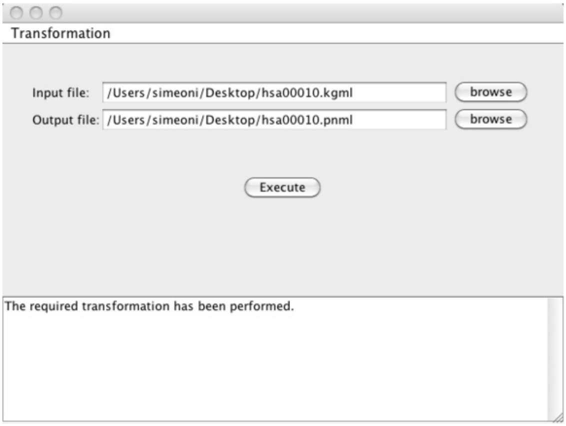
2 The Tool *MPath2PN*

The availability of different sources of data on metabolic pathways and of many simulation and analysis tools for PNs paradoxically is a problem, because of the variety of database formats and of input formats for PNs tools. In the literature, in order to cope with this problem, we find proposals for

- a standard format for metabolic data, such as SBML [15], the format of the BioModels Database, or BioPAX [3], and a standard format for PN tools, such as PNML (Petri Net Markup Language) [11];
- unification or integration of different databases such as in [31] or [24], and translations between different data formats, such as in KEGGtranslator [9] or KGML2SBML and KGML2BioPAX [26].

Automatic translators from metabolic pathways to PN models are offered as specific functionalities in some modelling and analysis tools based on PNs. Snoopy [16,27], a tool for the design and animation of PNs to verify technical systems and validate biosystems, provides translations from SBML to basic PNs, continuous PNs and extended stochastic PNs in a Snoopy-specific format based on XML. Cell Illustrator (CI)[4,30], a tool for modelling and simulating complex biological systems, allows the user to import models in SBML and BioPAX and translates them into CSML, the XML-based format of CI models, which are Petri nets extended with continuous and generic processes and quantities.

We developed the tool *MPath2PN* to support the automatic translation of metabolic pathways into PN models, coping with different input and output formats. *MPath2PN* is written in Java and it has a modular structure which is designed to facilitate the integration of different translations. Since most of the formats for metabolic pathways and for PNs are based on XML, translations are implemented by using XSLT (eXtensible Stylesheet Language Transformation [6]) in the Saxon open source version [14]. Each translation requires the definition of an appropriate style sheet XSL [5] which specifies the translation rules to be applied. For integrating different sources of information the translation process in *MPath2PN* is organised into three phases: pre-treatment, XSLT translation and post-treatment.



Input file: /Users/simeoni/Desktop/hsa00010.kgml

Output file: /Users/simeoni/Desktop/hsa00010.pnml

The required transformation has been performed.

Fig. 1. The MPath2PN graphical user interface

In the pre- and post-treatment Java classes can be developed for modifying respectively the input and the output files. For example when KEGG data are in input, the KGML file representing a KEGG pathway does not specify the stoichiometric values of reactions. The post-treatment phase can retrieve the missing information from the KEGG web service and add them to the PN model.

The user interface of *MPath2PN*, shown in Figure 1, is very simple. The different translations offered by the tool may be chosen through the "*Transformation*" menu. It is sufficient to specify the input and output file names and press the "*Execute*" button. Information messages on the translation status are displayed in the lower part of the window.

At present *MPath2PN* provides various translations from KEGG metabolic pathways expressed in KGML to PNML, the standard format for PNs tools. Such translations differ on the adopted modelling choices with respect to the following issues.

- *Representation of ubiquitous substances*: Once assumed to be constant, ubiquitous substances, such as ATP, ADP and H_2O, can be omitted and the resulting model is greatly simplified, but it could become disconnected. *MPath2PN* translations either include ubiquitous substances or ignore them.
- *Representation of reactions*: *MPath2PN* translations are either *reaction-based* or *enzyme-based*. In a reaction-based translation, each reaction, possibly catalysed by various enzymes, is represented by a single transition in the PN. In an enzyme-based translation, each reaction is represented by as many transitions as the number of enzymes catalysing it.
- *Representation of boundaries*: A PN corresponding to a metabolic pathway of an organism can be considered either in isolation, focusing on its internal behaviour, or as an open interactive subsystem of the full metabolic network. *MPath2PN* offers the option of producing either *isolated* or *open* PN models. In the open case, compounds which connect the pathway to the rest of the metabolic network are represented as open places, i.e. places where the environment can freely put/remove substances through corresponding input/output transitions. Com-

pounds which are only substrates/only products (the source and the target substances of the pathway) are represented as open places where the environment can freely put/remove substances.

Most of the reactions in a pathway are reversible. *MPath2PN* represents a reversible reaction by using two transitions, a forward and a backward one. Such pairs of transitions are recognisable by their identifiers, so that the corresponding cyclic behaviour can be filtered out, if desired, in the PN model analysis.

A first prototype version of *MPath2PN* was presented in [20], it provided two translations from KEGG to the input language of PIPE2 [13], an open source platform independent tool for creating and analysing PNs. Both translations produce reaction-based, isolated PNs, differing for the treatment of the ubiquitous substances.

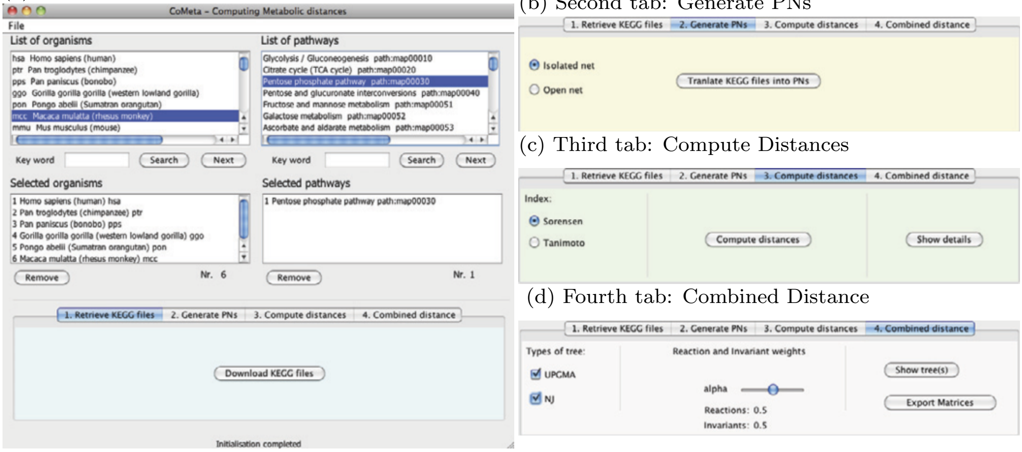
3 The tool *CoMeta*

In the literature many proposals for comparing metabolic pathways, or even whole metabolic networks, in different organisms can be found. They are based on two ingredients: a simplified representation of a metabolic pathway, viewed as a set of reactions and/or compounds or as a set of sequences of reactions or as a graph, and a notion of similarity score (or distance measure) between two pathways which depends on the chosen representation. In [21] various approaches to the comparison of metabolic pathways are discussed, pointing out that they are eminently static and ignore the flow of metabolites in the pathways. Our tool *CoMeta* (COmparing METAbolic pathways) compares metabolic pathways of different organisms by considering also their behaviour since it relies on a PN representation of the pathways. It computes and combines two distances, a “static” distance d_R, taking into account the reactions in the pathways, and a “behavioural” distance d_I, taking into account potential fluxes in the pathways at steady state, as expressed by the T-invariants of the corresponding PNs. Given two pathways represented as PNs, P_1 and P_2, each distance is derived from a corresponding similarity score: $d_X(P_1, P_2) = 1 - score_X(P_1, P_2)$, with $X \in \{R, I\}$.

When computing d_R, $score_R(P_1, P_2)$ represents the similarity between the reactions in P_1 and those in P_2. Concretely, each reaction is represented by the enzymes which catalyse it and, in turn, each enzyme is identified by its EC number [40], so that a pathway is reduced to a multiset of EC numbers (a multiset, rather than a set, since the same enzyme may have multiple occurrences in a pathway). Hence $score_R(P_1, P_2)$ measures the similarity between the multisets of the EC numbers associated to P_1 and P_2, respectively. The similarity between two enzymes is simply the identity of their EC numbers, but finer similarity measures could be easily accommodated in this setting. For comparing multisets, the present version of *CoMeta* allows to use either the Sørensen [37] or the Tanimoto [39] indexes extended to multisets.

When computing d_I, the set of minimal T-invariants (Hilbert basis) $\mathcal{B}(P_1)$ and $\mathcal{B}(P_2)$ of the two nets are compared. Each invariant is represented as a multiset

(a) CoMeta main window



CoMeta – Computing Metabolic distances

List of organisms

hsa Homo sapiens (human)
ptr Pan troglodytes (chimpanzee)
pps Pan paniscus (bonobo)
ggo Gorilla gorilla gorilla (western lowland gorilla)
pon Pongo abelii (Sumatran orangutan)
mcc Macaca mulatta (rhesus monkey)
mmu Mus musculus (mouse)

List of pathways

Glycolysis / Gluconeogenesis path:map00010
Citrate cycle (TCA cycle) path:map00020
Pentose phosphate pathway path:map00030
Pentose and glucuronate interconversions path:map00040
Fructose and mannose metabolism path:map00051
Galactose metabolism path:map00052
Ascorbate and aldarate metabolism path:map00053

Selected organisms

1 Homo sapiens (human) hsa
2 Pan troglodytes (chimpanzee) ptr
3 Pan paniscus (bonobo) pps
4 Gorilla gorilla gorilla (western lowland gorilla) ggo
5 Pongo abelii (Sumatran orangutan) pon
6 Macaca mulatta (rhesus monkey) mcc

Nr. 6

Selected pathways

1 Pentose phosphate pathway path:map00030

Nr. 1

Initialisation completed

(b) Second tab: Generate PNs

● Isolated net
○ Open net

(c) Third tab: Compute Distances

Index:
● Sorensen
○ Tanimoto

(d) Fourth tab: Combined Distance

Types of tree:
☑ UPGMA
☑ NJ

Reaction and Invariant weights
alpha
Reactions: 0.5
Invariants: 0.5

Fig. 2. The CoMeta graphical user interface

of EC numbers, corresponding to the reactions in the invariant, and the similarity between two invariants is given, as before, by their similarity index. The similarity score between the two Hilbert bases is computed through a heuristic match and it represents the similarity of the matching pairs of invariants. In *CoMeta* the two distances may be combined: $d_C(P_1, P_2) = \alpha\ d_R(P_1, P_2) + (1 - \alpha)\ d_I(P_1, P_2)$, with $\alpha \in [0, 1]$, to move the focus between reactions and functional components, and two organisms can be compared on n metabolic pathways $P_1, \ldots, P_n$ by considering the average distance over the n pathways.

In [21] a detailed explanation of the distances is given and a prototype version of *CoMeta* is presented along with some experiments for validating the proposal. This early version of the tool was running under Windows and Linux and used INA (Integrated Tool Analyser) [38] as an external tool for computing the Hilbert basis. Few tuning possibilities were offered, the pathways were modelled as isolated networks and only the Sørensen index was used in computing the distances.

The current version of *CoMeta* is a user-friendly tool written in Java and running under Linux and Mac. For computing Hilbert bases it resorts to 4ti2 [18], a very efficient tool offered in a software package for algebraic, geometric and combinatorial problems on linear spaces. The tool *MPath2PN* is used for transforming a metabolic pathway into a corresponding PN. The translation is from KGML to PNML, the standard format for PNs tools, without ubiquitous substances and enzyme-based, and the user can choose either the isolated or the open PN models of the pathways. *CoMeta* offers a set of integrated functionalities through a graphical user interface as shown in Figure 2(a). In the upper part of the window the desired KEGG organisms and pathways can be selected. In the lower part a tabbed panel offers the commands to be performed. The first tab of the panel is shown in the main window, while the others are shown in Figure 2(b), 2(c), and 2(d), respectively. The main functionalities of the tool are the following:

- *Select organisms and pathways* (Figure 2(a), upper part): *CoMeta* proposes the lists of all KEGG organisms and pathways and allows the user to select those to be compared by double-clicking them.

- *Retrieve KEGG information* (Figure 2(a), lower part): *CoMeta* downloads from the KEGG database the selected organisms and pathways.
- *Translate into PNs* (Figure 2(b)): *CoMeta* translates the selected organisms and pathways into corresponding PNs, by using *MPath2PN*. The user can choose to generate either isolated or open PN models of the pathway. CoMeta produces the stoichiometric matrix of the net in a text file which is the input of 4ti2.
- *Compute Distances* (Figure 2(c)): d_R and d_I are computed as described above. The user can select either the Sørensen or the Tanimoto index for multiset comparison. The user can inspect the details of the comparison between any pair of organisms (e.g., T-invariants bases, invariants matches, reactions and invariants scores).
- *Show Phylogenetic trees* (Figure 2(d)): *CoMeta* computes the distance which combines d_R and d_I according to a weight parameter α specified by the user. Such a distance is used to produce and visualise corresponding phylogenetic trees. The user can specify the method to be used for the generation of the phylogenetic trees. Currently *CoMeta* offers the UPGMA [36,35] and Neighbour Joining methods [29,35]. The matrices for d_R, d_I and the combined distance can be exported as text files for further analyses.

4 Conclusions

We presented two tools for the representation and the analysis of metabolic pathways through Petri nets. *MPath2PN* is a tool for automatically transforming a metabolic pathway, expressed in one of the various existing databases formats (like KGML or SBML), into a corresponding PN, expressed in one of the formats for PN tools. At present it offers several translations corresponding to different modelling choices, it gets metabolic pathways from KEGG and it generates the corresponding PNs encoded in PNML. *CoMeta* is an interactive tool for comparing metabolic pathways of different organisms. The comparison is based on a distance which may consider both homology of reactions and potential fluxes at steady state.

We are conducting experiments to validate *CoMeta* and our distance. In [21] we already showed that it produces valid phylogenetic classifications and that other measures in the literature, based on more sophisticated representations of a pathway (e.g., using graphs rather than multisets, or considering compounds besides enzymes), do not necessarily give better results than our measure.

This is an ongoing project and we plan to extend both *MPath2PN* and *CoMeta*. We are working on new translations for *MPath2PN*, i.e. from SBML to extended deterministic stochastic PNs (eDSPNs) and coloured stochastic PNs (SCPNs), expressed in the input language of the tool TimeNET [17]. For *CoMeta* we are working in various directions. We are performing extensive explorations of the KEGG database to analyse the significance of our distance. We are experimenting with isolated and open PN models to study the effects of the different modelling choices.

Besides, we intend to improve our distance by refining the similarity measures on enzymes, by refining the simple greedy algorithm for matching invariants bases and by associating weights to the pathways when considering sets of pathways in the comparison.

Acknowledgement

We are grateful to A. Roncato, P. Besenzon and S. Alaimo for their contribution to the implementation of *CoMeta*. We thank moreover F. De Nes, M. Llabrés Segura and A. Marin for their contribution to the first version of *MPath2PN*.

References

[1] BioCyc: database collection. http://BioCyc.org.

[2] Biomodels Database. http://www.ebi.ac.uk/biomodels.

[3] BioPAX: Biological Pathway Exchange. http://www.biopax.org/index.php.

[4] Cell Illustrator 5.0: A software tool that enables biologists to draw, model, elucidate and simulate complex biological processes and systems. http://www.cellillustrator.com/new_version.

[5] Extensible Stylesheet Language. http://www.w3.org/Style/XSL/.

[6] Extensible Stylesheet Language Transformations. http://www.w3.org/TR/xslt.

[7] Kegg Markup Language manual. http://www.genome.ad.jp/kegg/docs/xml.

[8] KEGG pathway database - Kyoto University Bioinformatics Centre. http://www.genome.jp/kegg/pathway.html.

[9] KEGGtranslator. http://www.ra.cs.uni-tuebingen.de/software/KEGGtranslator/.

[10] MetaCyc Encyclopedia of Metabolic Pathways. http://metacyc.org.

[11] Petri Net Markup Language. http://www.pnml.org.

[12] Petri net tools. http://www.informatik.uni-hamburg.de/TGI/PetriNets/tools.

[13] Platform Independent Petri net Editor 2. http://pipe2.sourceforge.net/.

[14] SAXON: the XSLT and XQuery processor. http://saxon.sourceforge.net/.

[15] SBML: Systems Biology Markup Language. http://sbml.org.

[16] SNOOPY: a software tool to design and animate hierarchical graphs. http://www-dssz.informatik.tu-cottbus.de/DSSZ/Software/Snoopy.

[17] TimeNET. http://www.tu-ilmenau.de/fakia/TimeNet.timenet.0.html?&L=1.

[18] 4ti2 team. 4ti2—a software package for algebraic, geometric and combinatorial problems on linear spaces. Available at www.4ti2.de.

[19] P. Baldan, N. Cocco, A. Marin, and M Simeoni. Petri nets for modelling metabolic pathways: a survey. *Natural Computing*, 9(4):955–989, 2010.

[20] P. Baldan, N. Cocco, F. De Nes, M. Llabrés Segura, and M. Simeoni. MPath2PN - Translating metabolic pathways into Petri nets. In M. Heiner and H. Matsuno, editors, *BioPPN2011 Int. Workshop on Biological Processes and Petri Nets*, CEUR Workshop Proceedings - Vol-724, pages 102–116, 2011. online: http://ceur-ws.org/Vol-724.

[21] P. Baldan, N. Cocco, and M Simeoni. Comparison of metabolic pathways by considering potential fluxes. In *BioPPN2012 - 3rd International Workshop on Biological Processes and Petri Nets*, CEUR Workshop Proceedings - Vol-852, pages 2–17, 2012.

[22] J. Esparza and M. Nielsen. Decidability issues for Petri Nets - a survey. *Journal Inform. Process. Cybernet. EIK*, 30(3):143–160, 1994.

[23] R. Hofestädt. A Petri net application of metabolic processes. *Journal of System Analysis, Modelling and Simulation*, 16:113–122, 1994.

[24] I. Kanaris, K. Moutselos, A. Chatziioannou, I. Maglogiannis, and F.N. Kolisis. Building in-silico pathway SBML models from heterogeneous sources. In *BioInformatics and BioEngineering (BIBE2008)*, pages 1–6. IEEE, 2008.

[25] M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, and Y. Yamanishi. KEGG for linking genomes to life and the environment. *Nucleic Acids Research*, pages 480–484, 2008.

[26] K.E. Lee, M.H. Jang, A. Rhie, C.T. Thong, S. Yang, and H.S. Park. Java DOM parsers to convert KGML into SBML and BioPAX common exchange formats. *Genomics & Informatics*, 8(2):94–96, 2010.

[27] W. Marwan, C. Rohr, and M. Heiner. Petri nets in Snoopy: A unifying framework for the graphical display, computational modelling, and simulation of bacterial regulatory networks. volume 804 of *Bacterial Molecular Networks*, pages 409–437. Springer New York, 2012.

[28] T. Murata. Petri Nets: Properties, Analysis, and Applications. *Proceedings of IEEE*, 77(4):541–580, 1989.

[29] Saitou N. and Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425, 1987.

[30] M. Nagasaki, A. Saito, E. Jeong, C. Li, K. Kojima, E. Ikeda, and S. Miyano. Cell Illustrator 4.0: A computational platform for systems biology. *In Silico Biology*, 10(1):5–26, 2010.

[31] Harsha K. Rajasimha. PathMeld: A methodology for the unification of metabolic pathway databases. Master's thesis, Virginia Polytechnic Institute and State University, 2004. `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.96.1278`.

[32] V. N. Reddy, M.N. Liebman, and M.L. Mavrovouniotis. Qualitative Analysis of Biochemical Reaction Systems. *Computers in Biology and Medicine*, 26(1):9–24, 1996.

[33] V. N. Reddy, M. L. Mavrovouniotis, and M. N. Liebman. Petri net representations in metabolic pathways. In *ISMB93: First Int. Conf. on Intelligent Systems for Molecular Biology*, pages 328–336. AAAI press, 1993.

[34] S. Schuster and C. Hilgetag. On elementary flux modes in biochemical reaction systems at steady state. *Journal of Biological Systems*, 2:165–182, 1994.

[35] P. Sestoft. Programs for biosequence analysis. `http://www.itu.dk/people/sestoft/bsa.html`.

[36] R. Sokal and Michener C. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 38:14091438, 1958.

[37] T. Sørensen. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on danish commons. *Biologiske Skrifter / Kongelige Danske Videnskabernes Selskabg*, 5(4):1–34, 1948.

[38] P.H. Starke and S. Roch. The Integrated Net Analyzer. *Humbolt University Berlin*, 1999. www.informatik.hu-berlin.de/ starke/ina.html.

[39] T.T. Tanimoto. Technical report, IBM Internal Report, November 17 1957.

[40] E. C. Webb. *Enzyme nomenclature 1992: recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes.* San Diego: Published for the International Union of Biochemistry and Molecular Biology by Academic Press, 1992.