Computational Approaches for Analyzing Metabolic Pathways

Dissertation

der Fakultät für Informations- und Kognitionswissenschaften
der Eberhard-Karls-Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von

Dipl.-Inform. Torsten Blum
aus Leipzig

Tübingen
2009

Tag der mündlichen Qualifikation:	15.7. 2009
Dekan:	Prof. Dr. Oliver Kohlbacher
1. Berichterstatter:	Prof. Dr. Oliver Kohlbacher
2. Berichterstatter:	Prof. Dr. Hans-Peter Lenhof (Universität des Saarlandes)

Erklärung

Hiermit erkläre ich, dass ich diese Schrift selbständig und nur mit den angegebenen Hilfsmitteln angefertigt habe und dass alle Stellen, die im Wortlaut oder dem Sinne nach anderen Werken entnommen sind, durch Angaben der Quellen kenntlich gemacht sind.

Tübingen, Juli 2009 *Torsten Blum*

Zusammenfassung

Der Metabolismus lebender Organismen besteht aus einem komplexen Netzwerk chemischer Reaktionen, welche kleine Moleküle transformieren, um Energie und Biomasse aus Nährstoffen zu gewinnen. In solch einem Netzwerk repräsentieren Stoffwechselwege regulierte funktionelle Einheiten zur Konversion bestimmter Quellmetaboliten in Produktmoleküle durch eine Sequenz von Reaktionen. Jedoch ist das Wissen über Stoffwechselwege, vor allem in neu sequenzierten Organismen, unvollständig oder benötigt noch experimentelle Bestätigung. Das mögliche Vorkommen neuartiger oder alternativer Pfade muss bei der Erforschung der metabolischen Fähigkeiten von Organismen berücksichtigt werden. In diesem Zusammenhang bietet die rechnergestützte Herleitung biologisch bedeutsamer Pfade eine attraktive Ergänzung zu experimentellen Studien und besitzt zahlreiche Anwendungen in der Systembiologie.

Diese Arbeit prsentiert mehrere neuer rechnergestützter Methoden zur Analyse von Stoffwechselwegen in genomweiten Netzwerken. Entwickelt wurde ein graphtheoretischer Ansatz, der das metabolische Netzwerk auf einen gewichteten Graphen abbildet und einen effizienten Pfadsuch-Algorithmus zur Berechnung relevanter Biotransformationsrouten verwendet. Der Ansatz wurde ergänzt durch die Integration weiterer relevanter Informationen, abgeleitet aus den biochemischen Entitäten (Metaboliten, Reaktionen und Enzyme), die das Netzwerk aufbauen. Aus diesem Grund wurde eine verbesserte Methode erzeugt, welche atomare Abbildungsregeln aus den chemischen Strukturen der Netzwerkverbindungen automatisch berechnet. Für eine gegebene Reaktion definiert eine atomare Abbildungsregel welches Atom einer Eduktverbindung auf welches Atom einer Produktverbindung transferiert wird. Die Anwendung dieser Regeln erlaubt es den Fluss von Atomen in der Pfadsuche zu verfolgen, was für die Erkennung biochemisch unzulässiger Routen hilfreich ist. Eine weitere Methode zur Abschätzung freier Reaktionsenthalpien (Gibbs-Energien) unter (biochemischen) Standardbedingungen wurde entwickelt und verwendet um die Pfadsuche zu verbessern. Die dritte Methode erweiterte die Pfadanalyse durch vorhergesagte Informationen über die subzelluläre Lokalisierung der beteiligten Enzyme.

Um die Nützlichkeit der entwickelten Methoden für metabolische Pfadanalysen zu demonstrieren, wurden experimentell bestätigte Biotransformationsrouten in den Netzwerken von *Escherichia coli* und *Arabidopsis thaliana* vorhergesagt.

Im letzten Teil dieser Arbeit wird ein benutzerfreundliches Web-Interface, genannt MetaRoute, zur Erkundung der metabolischen Netzwerke von hunderten von Organismen beschrieben.

Abstract

The metabolism of living organisms consists of a complex network of chemical reactions that transform small molecules to gain energy and biomass from nutrients. In such a network, metabolic pathways represent regulated functional units for converting particular source metabolites into product molecules by a sequence of reactions. However, knowledge about pathways, especially in newly sequenced genomes, is incomplete or remains to be experimentally verified. The potential presence of novel or alternative pathways has to be considered when investigating metabolic capabilities of organisms. In this context, computational inference of biologically meaningful pathways constitutes an attractive complement to experimental studies and has numerous applications in systems biology.

This thesis presents several novel computational approaches for analyzing metabolic pathways in genome-scale networks. A graph theoretical approach was developed that maps the metabolic network onto a weighted graph and uses an efficient path-finding algorithm to calculate relevant biotransformation routes. The approach was complemented by the integration of further relevant information derived from the biochemical entities (metabolites, reactions and enzymes) that build up the network. For this purpose, an improved method was created that automatically calculates atom mapping rules from chemical structures of the network compounds. Given a chemical reaction, an atom mapping rule defines which atom of an educt compound is transferred to which atom of a product compound. The application of these rules allows one to trace the flow of atoms in the path search, which is useful for detecting biochemically unfeasible routes. A further method for estimating Gibbs energy changes of reactions under (biochemical) standard conditions was developed and used to improve the path search. The third method extended the pathway analysis using predicted information about subcellular localizations of the enzymes involved.

To demonstrate the usefulness of the developed approaches for metabolic pathway analysis, experimentally verified biotransformation routes in the metabolic networks of *Escherichia coli* and *Arabidopsis thaliana* were predicted.

In the last part of this thesis a user-friendly web interface, called MetaRoute, for exploring the metabolic networks for hundreds of organisms, is described.

Acknowledgments

First of all, I want to thank my supervisor Prof. Dr. Oliver Kohlbacher for the opportunity to work on an exciting research topic and the comprehensive support during all phases of this thesis.

I would also like to thank all my colleagues for a pleasant time at the Division for Simulation of Biological Systems. Thanks to Nico Pfeiffer, Dr. Andreas Kämper and Sebastian Briesemeister for critical proofreading of my scientific writings; Muriel Quenzer and Jan Schulze for providing a running working environment; Sandra Gesing which is not just a colleague but also a good friend; Marc Sturm, Nora Toussaint, Lena Feldhahn, Andreas Bertsch, Nina Fischer, Sven Nahnsen, Marc Röttig, Marcel Schumann, Erhan Kenar, Claudia Walter and my former colleagues Dr. Annette Höglund, Dr. Pierre Dönnes and Dr. Andreas Kerzmann for a good and motivating working atmosphere.

Many thanks also to my cooperating partners: Scott Brady, Dr. Jan Küntzer, Andreas Gerasch, Jan Mitschke, Yin Lam, Dr. Hagit Shatkay, Dr. Stefan Rensing, Prof. Dr. Hans-Peter Lenhof and Prof. Dr. Michael Kaufmann.

Furthermore, I want to thank my parents which made my study and Ph.D. thesis possible.

Last but not least, I would like to thank Martina Leibig, the most important person for me, for her encouragements, inspiring discussions and sharing my interests in bioinformatics, biochemistry and molecular biology (among many other things).

In accordance with the standard scientific protocol, I will use the personal pronoun “we” to indicate the reader and the writer, or my scientific collaborators and myself.

Contents

Chapter 1

Introduction

1.1 Motivation

Cellular metabolism consists of a complex network of chemical reactions that transform small molecules and that operate together to convert nutrients into energy and biomass. In such a network, metabolic pathways represent functional units that are responsible for specific metabolic processes, for example, the degradation of carbon sources like glucose or the synthesis of amino acids. The metabolic pathways known from many biochemical textbooks have been discovered through painstaking work on specific model organisms. However, it has been shown that in reality metabolic networks are much more variable and more interconnected than the (mostly linear) textbook pathways [Cordwell, 1999]. In microbial genomes especially, even standard pathways of the core metabolism like glycolysis, the TCA cycle or the pentose phosphate pathway can vary widely even within a species (e.g., from strain to strain) due to missing or mutated enzymes. The existence of alternative pathways is the result of an organism's adaptation to its environment or niche. Therefore, knowledge of all feasible routes transforming a source metabolite into a target metabolite can help to understand the metabolism better or to decide whether particular enzymes or intermediates are essential in the process. However, experimental determination of pathways is laborious and time-consuming. So far, there is no high-throughput method for this task. Hence, there is a need to develop computational approaches for detecting plausible pathways in genome-scale metabolic networks. Applications can be found in systems biology related fields like metabolic engineering to support genetic modification of microorganisms in order to increase the yield of industrially important metabolites. The identification, based on computational tools, of (non-)essential enzymes in metabolic pathways is also useful

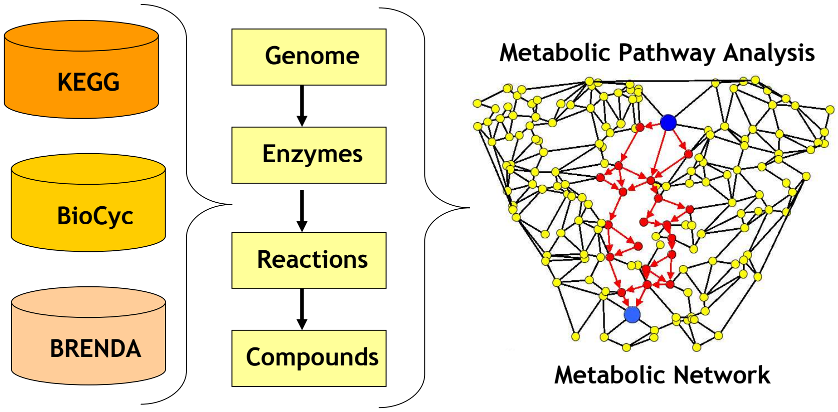


Figure 1.1: The reconstruction of metabolic networks based on data extracted from pathway databases enables computational analysis of metabolic pathways.

for detecting potential drug targets. Furthermore, the design of tracer or knock-out experiments is much easier knowing all alternative routes that are affected by enzymes under study or compounds marked by radioisotopes.

1.2 State of the art

With the availability of whole-genome data and functional annotation for a wide range of organisms, computational tools can now be applied to a much broader range of problems and model organisms. Starting from gene-enzyme relations, one can use enzyme-reaction as well as reaction-compound relations (extracted from pathway databases like KEGG [Kanehisa, 1996], EcoCyc [Keseler *et al.*, 2005], MetaCyc [Caspi *et al.*, 2006] and BRENDA [Schomburg *et al.*, 2002]) to reconstruct an organism-specific metabolic network (see also Fig. 1.1). The computational analysis of these networks, focusing on the detection of novel or alternative pathways that transform a particular source into a target compound, requires sophisticated approaches. A major problem in this context is the computational effort caused by the combinatorial explosion of the number of possible routes in large-scale metabolic networks. Searching for relevant pathways without information other than the connectivity, i.e., when two successive reactions are connected by a common metabolite, often delivers meaningless results. Küffner *et al.* [2000] applied such a naive or "blind" search to a metabolic network at genome-scale. An exhaustive enumeration algorithm was developed to analyze the glycolysis pathway, i.e., for the enumeration of all routes starting from glucose as source and ending in pyruvate. The authors found at least 500,000 different routes

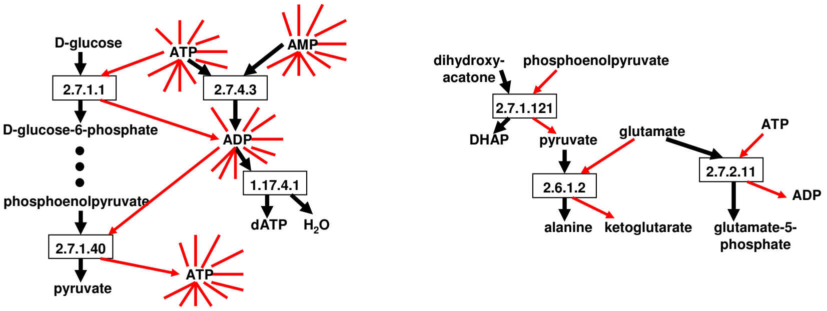


Figure 1.2: Finding relevant routes in metabolic networks is complicated by the fact that many metabolites can assume different roles (e.g. main educt/product, black arrows or side educt/product, red arrows) in different reactions depending on the context.

of at most nine reaction steps from glucose to pyruvate. Of course nearly all of them are biologically irrelevant. The large number of routes results from the presence of so-called pool metabolites like water, ATP or NADH that participate in many reactions (see also Fig. 1.2 for an illustration). For example, using ADP as an intermediate would lead to a very short but irrelevant route from glucose to pyruvate. A simple strategy to avoid irrelevant short cuts is the removal of pool metabolites. But ignoring these network hubs cannot be a satisfying solution since their choice is not always obvious. Removing compounds runs the risk of missing relevant routes and does not guarantee the retrieval of only relevant ones. The main problem is that even such a typical side metabolite like ADP acts as a real intermediate in several pathways. Further examples are glutamate or pyruvate where their role as a main or side metabolite is not unique or clear in all reactions [van Helden *et al.*, 2002].

The recent approaches to metabolic pathways described in the literature can be roughly divided into two main groups. Constraint-based methods [Schuster *et al.*, 1999; Schilling *et al.*, 2000] infer network-based, stoichiometrically balanced pathways defined as a metabolic subnetwork in which the net production and consumption of all compounds is zero. Excluded from this balance are the source and target compounds and a predefined set of so-called external compounds or pool metabolites. The metabolic network is represented as a stoichiometric matrix where rows and columns represent metabolites and reactions. The pathways are inferred using convex analysis [Rockafellar, 1970], a branch of mathematics for analyzing a set of linear equations under a given set of constraints. The advantage of the method is that it is mathematically well defined. Since stoichiometry is the driving concept, the comparison of alternative pathways with respect to biotechnological applications is well established. In this context, the analysis of pathways is focused on increasing the yield of industrially important compounds from specific source metabolites. Irrelevant routes, as described in the previous section, cannot be inferred by stoichiometric approaches. However, while the problem of enumerating all relevant pathways is avoided, the underlying calculation still represents a computationally hard problem. It seems to be intractable to use the approach for most genome-scale networks [Klamt *et al.*, 2002, 2003; Yeung *et al.*, 2007]. Furthermore, it is not clearly defined how to distinguish between internal and external (pool) metabolites, which is, as has already been described, a non-trivial task. In practice, the computational complexity is reduced by using networks of moderate size and irreversible reactions.

Graph theory-based methods search for linear biotransformation routes, simply defined as a linear sequence of chemical reactions in which a source compound is converted into a target compound step by step. The metabolic network is represented as a graph [Arita, 2000; Rahman *et al.*, 2005; Croes *et al.*, 2006]. An advantage is the availability of already established and efficient path-finding algorithms that have polynomial runtime. Therefore, these algorithms can be used for genome-scale network analysis [Aittokallio *et al.*, 2006]. An interactive navigation through metabolic networks is possible, simply by searching for (k-shortest) paths between a given source and target, without the need for user-defined constraints [van Helden *et al.*, 2002]. However, the main challenge for graph theory-based methods is to detect only relevant routes within the first calculated (k-shortest) paths as well as to filter out biologically meaningless routes. This step requires the definition of relevance or optimization criteria. For this purpose, consideration of the structural information of the metabolites is used in several methods. The PathwayHunter tool [Rahman *et al.*, 2005] uses chemical fingerprints to guide a shortest-path search between structurally similar metabolites. Chemical fingerprints are unique patterns that represent the presence and absence of a defined set of chemical groups or substructures. Another promising idea is to trace the flow of atoms in a shortest-path search using atom mapping rules [Arita, 2000, 2003]. Given a chemical reaction, an atom mapping rule defines which atom of an educt compound is transferred to which atom of a product compound. The sequential application of these rules is helpful for detecting biochemically irrelevant shortest paths in which no atom is transferred from the source to the target. The main problem is that, despite the atom trace, the shortest-path search tends to go through pool metabolites or network hubs that connect many different pathways across the whole metabolic network. The structural information, necessary for atom mapping calculation, is either not given or is incomplete for a fraction of compounds participating in reactions stored in pathway databases. Those compounds are often described only by a string name or represent general molecules like ‘an alcohol’. This is also a principal problem when using chemical fingerprints. Furthermore, the automated and efficient calculation of atom mapping rules, given thousands of reactions in a database like KEGG, is complicated and requires sophisticated algorithms. Although the consideration of structural information is doubtlessly very useful, its incorporation requires pre-calculation effort and makes the path-finding process more complicated. An alternative strategy was proposed by Croes *et al.* [2006], where the metabolic network is represented by a degree-weighted graph. In this graph each node is assigned a weight equal to its degree. Searching for the lightest path significantly reduces the probability of finding irrelevant routes containing pool metabolites as intermediates. An advantage is that structural information about the compounds is not needed and that the guided shortest-path search is replaced by the search for the lightest path based on an easy-to-use optimization criterion. However, the method fails for routes containing network hubs as intermediates or for routes passing through several pathways of the core metabolism such as glycolysis or the TCA cycle. Those pathways contain highly connected metabolites like pyruvate or acetyl-CoA and receive, therefore, high overall path weights.

Compared to constraint-based approaches, the main advantage of graph theory-based methods for metabolic pathways is that efficient (k-shortest) path-finding algorithms can be used to deal even with genome-scale networks. However, graph theory-based search requires the consideration of suitable optimization criteria to find relevant routes. Approaches described in the literature so far, like node-degrees or atom mapping rules, are not sophisticated enough and still produce irrelevant routes. The reason is that they ignore important biochemical and biological constraints like reaction energetics or subcellular localizations of enzymes. Furthermore, current state-of-the-art methods are still computationally demanding due to the high number of irrelevant routes and the necessary adaptation of path-finding algorithms. This also renders efficient interactive search difficult.

1.3 Contributions of this thesis

This thesis introduces novel approaches for analyzing metabolic pathways which improve or complement existing approaches. A graph theory-based approach for finding feasible biotransformation routes represents the basic framework. Our method ensures efficient calculation of relevant routes in metabolic networks at the genome scale without the need for pre-defining pool metabolites. To this end, we integrated atom mapping rules and the lightest path search into a joint method. The key component of the approach is a novel method for the fully automated and efficient calculation of atom mapping rules. In addition to the detection of biochemically unfeasible routes, the application of atom mapping rules allows to create graph representations that are specialized for carbon, nitrogen, sulfur or phosphorous metabolism. The advantage is a reduced network complexity. Compared to the degree-weighted graph approach [Croes *et al.*, 2006] we use more complex weighting schemes. The combined weighting distinguishes between weights assigned to the compounds and those assigned to the reactions in the network.

An important goal of this work was also the integration of further relevant biological or biochemical data into the weighting scheme. For this purpose, we transformed thermodynamic information (Gibbs reaction energy, ΔG_r) and the subcellular localization of the catalyzing enzymes into numerical weights, which improved the combined weighting.

The Gibbs energy represents the driving force for each biochemical reaction in the metabolic network. Reactions require a negative change in Gibbs energy to take place spontaneously. Those reactions that are associated with a positive change in Gibbs energy will not occur spontaneously. The biochemical meaning of using Gibbs energy information to select plausible pathways derives from the assumption that biological systems prefer to use the thermodynamically most favorable route among a set of alternative routes for converting a particular source into a target metabolite. The actual change in Gibbs energy of a reaction depends on the specific physiological conditions, the compounds involved (the educts and products) and their intracellular concentrations. Gibbs reaction energies can be determined experimentally under standard conditions. These standard Gibbs energies can provide valuable clues about the thermodynamic feasibility of metabolic pathways. However, this kind of thermodynamic information is available only for a very limited number of reactions stored in pathway databases. Computational approaches could be the solution by complementing reactions with standard Gibbs energies. To our knowledge, there is only one method that is specialized in the estimation of standard Gibbs energies of biochemical reactions [Mavrovouniotis, 1990, 1991]. Since this method requires a non-trivial decomposition of compounds into non-overlapping groups of atoms and ignores important biochemical effects like the ionic strength, the presence of metal ions and the dissociation of a compound into several ionic species in dilute aqueous solution, we decided to develop a novel approach. We applied quantitative structure-property relationship (QSPR) techniques. To this end, we calculated molecular descriptors for the educts of products of reactions with a known change in Gibbs energy under biochemical standard conditions. Multiple linear regression and stepwise feature selection were used to obtain a high-quality predictive model. Performance evaluation using an independent test procedure showed excellent performance of the model.

Another important restriction on the feasibility of a metabolic pathway is the presence of all its constituent enzymes in the same subcellular compartment or at least a small number of transitions between compartments along the pathway. Enzymes are built up from protein monomers, which in turn, are synthesized in the cytoplasm and need to be further transported into their destination compartment based on sorting signals in their amino acid sequence. Eukaryotic cells, in particular, are organized into different membrane-surrounded compartments. Each compartment is specialized for a specific set of cellular functions. This includes the spatial organization of enzymes in metabolic pathways [Hrazdina and Jensen, 1992]. Spatially distinct enzymes and metabolites enable a better fine-tuning of the metabolism. While transport between compartments is not uncommon, enzymes belonging to adjacent steps in a pathway are usually localized in the same compartment. This information can be exploited as well in order to recognize infeasible or less probable pathways. The underlying idea of considering the subcellular localizations of successive enzymes in the path search is, therefore, the assumption that a metabolic pathway is more efficient if its enzymes are co-localized. This was also proposed in a similar way by Gille *et al.* [2005] who describe the consideration of the cellular compartmentalization as a new dimension to the formulation of network models.

However, for newly sequenced genomes especially, experimentally determined subcellular localizations of enzymes are rarely available. This kind of information is desirable not only for enzymatic proteins but for the whole proteome. Hence, a variety of computational approaches for predicting the subcellular localizations of proteins have been developed in recent years. Some of them are based on the detection of sorting signal sequences. However, the whole protein sorting process is very complex and not completely understood. Many protein sequences lack clearly identifiable signals. Therefore, other approaches primarily rely on more indirect data like the presence of functional domains or associated textual information extracted from annotation databases like Swiss-Prot. Our contribution to the problem, within the scope of this thesis, is a novel approach based on support vector machines (SVM) that combines features that are directly involved in the protein sorting process and derived from the amino acid sequence with evolutionary information in the form of phylogenetic profiles and textual information in the form of Gene Ontology (GO) terms. Using independent datasets, our approach performed considerably better for most tested categories than current state-of-the art tools.

Having computational methods for predicting thermodynamic and subcellular localization information ready, we could extend the weighting scheme of our graph theory-based approach for finding relevant biotransformation routes. Different graph types and search strategies were analyzed in genome-scale studies with the intention to demonstrate that adding relevant information into the graph representation and path-finding algorithm would increase the search performance. Therefore, we tried to infer experimentally determined biotransformation routes in the metabolic networks of *Escherichia coli* and *Arabidopsis thaliana*. Besides the overall performance results, which are very promising, we also present detailed results for a selection of several selected interesting pathways.

Furthermore, we developed a user-friendly web interface called MetaRoute, which offers interactive navigation through genome-scale metabolic networks for hundreds of organisms, combined with an easy-to-use visualization of the search results. Given a source and target metabolite, the tool calculates up to 500 metabolic routes that can be merged into a local network. Cross-species comparison is possible by searching in the combined (meta-) network of multiple organisms.

1.4 Structure of this thesis

The biological and biochemical background required for this thesis is sketched in Chapter 2. Along with the basic concepts of cellular metabolism, the basic principles of thermodynamics, especially of biochemical thermodynamics, and the details of protein sorting are outlined. The computational background to this work, including related work, is presented in Chapter 3. Here, the focus is on graph theory-based metabolic pathway analysis, which represents the main topic of this thesis. Chapter 4 contains the key contributions of this thesis. It describes the details of our graph theory-based framework for finding relevant biotransformation routes in metabolic networks, the computational approaches that provide relevant information deduced from compounds, reactions and enzymes in the form of atom mapping rules, standard Gibbs energies and subcellular localizations. For each approach the results obtained are presented and discussed. Finally, the functionality and potential applications of the implemented web interface for network navigation and visualization are outlined. Concluding remarks, including suggestions for future research projects, complete this work in Chapter 5.

Chapter 2

Biological and biochemical background

This thesis was concerned with the problem of computing all biologically relevant pathways transforming a source into a target compound in a metabolic network of interest. Knowing these pathways supports biomedical and biotechnological applications that require a deep understanding about the interplay of metabolic processes.

The sections in this chapter contain the underlying biological and biochemical background of this work. The basic concepts of cellular metabolism followed by the basic principles of bioenergetics and protein sorting are presented.

2.1 Cellular metabolism

Cellular metabolism consists of a highly complex network of chemical reactions that transform small molecules, also called metabolites. Depending on the needs of the organism, the metabolic network stores or converts energy extracted from given nutrients. The energy is used to maintain the functioning of the organism and to renew its structure by synthesizing macromolecules like proteins, nucleic acids, polysaccharides and lipids. In principle, nutrients are of the same type as these macromolecules and have to be decomposed in a process called biological degradation in order to gain the necessary energy and chemical building blocks for the synthesis of other essential biomolecules. To enable both macromolecular synthesis and degradation, also called anabolism and catabolism, cellular metabolism is regulated as well as temporally and spatially organized.

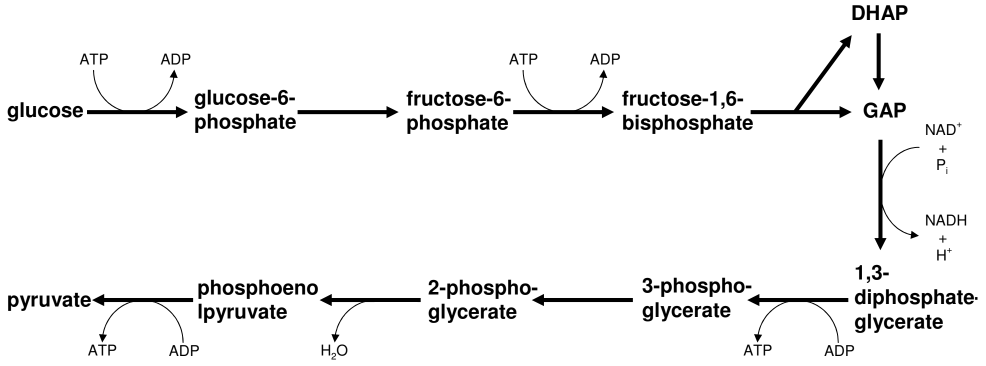


Figure 2.1: An illustration of the well-known glycolysis pathway. Glucose is degraded to pyruvate by a sequence of transforming reactions.

2.1.1 Metabolic pathways

Metabolic pathways represent functional units within a network. Depending on the particular purpose, one distinguishes between synthesis or degradation pathways. Typically, biochemistry defines a metabolic pathway as a sequence of chemical reactions, whereby a given source molecule is converted stepwise into some other molecule or molecules [Berg *et al.*, 2002]. For example, the glycolysis pathway (shown in Fig. 2.1) converts glucose into pyruvate. This is, however, a rather general definition. We will discuss more formal pathway definitions in Chapter 3.

The chemical reactions that constitute a pathway follow the law of mass conservation. A particular set of educt molecules is converted into a set of product molecules while the total mass of the educts remains equal to that of the products. Almost all metabolic reactions are controlled and catalyzed by enzymes which bind the reaction educts and release the products. Typically, enzymes are very specific for their substrates, which means that the majority of reactions can be catalyzed by only one enzyme which in turn catalyzes only one reaction. However, numerous exceptions exist. The International Union of Biochemistry and Molecular Biology (IUBMB) suggested a numerical classification scheme called the Enzyme Commission (EC) number. An EC number is assigned to each enzyme depending on its catalyzed reaction. Four hierarchical numbers describe the type of chemical conversion and the compounds involved. For example, the enzyme with EC number 2.7.1.2 (*glucokinase*) phosphorylates D-glucose to D-glucose-6-phosphate. Using these EC numbers, the set of all enzymatic reactions in metabolic networks (several thousand in all) can be subdivided into just six categories:

- Oxidoreductases (EC 1): oxidation/reduction reactions where electrons

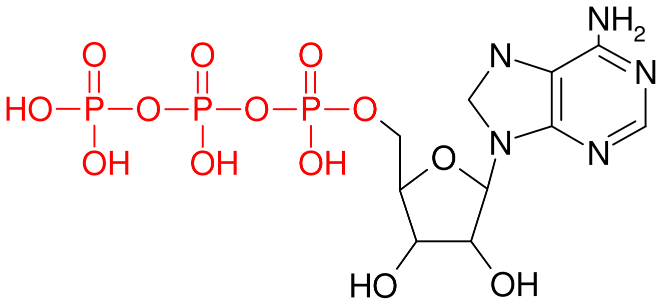


Figure 2.2: Adenosine triphosphate (ATP) an essential metabolite. The triphosphate group is highlighted in red.

are transferred between educts and products

- Transferases (EC 2): transfer of functional groups like methyl-, acyl-, amino- or phosphate groups between educts and products
- Hydrolases (EC 3): formation of two products from an educt by the cleavage of bonds and the addition of water
- Lyases (EC 4): cleavage of C-C, C-N, C-O or C-S bonds by the non-hydrolytic addition or removal of groups
- Isomerases (EC 5): structural changes within one molecule
- Ligases (EC 6): joining of two molecules by the parallel hydrolysis of the diphosphate bond in ATP or a similar triphosphate

Coenzymes and cofactors like NAD$^+$/NADH, CoA/acetyl-CoA, metal ions or vitamins support enzymes in the transfer of electrons, hydrogens or functional groups between the educts and products of the catalyzed reactions.

2.1.2 Pathway energetics

Metabolism is a highly dynamic process that permanently converts energy for a continuous degradation and synthesis of biomolecules. In principle, energy that is required by synthesis pathways to operate is gained by breaking down nutrients like glucose in degradation pathways. Typically, energy-producing and consuming processes are coupled via energy carriers. The most important and widely used carrier is adenosine triphosphate (ATP). Therefore, AMP and ADP are phosphorylated to ATP (shown in Fig. 2.2) to form high-energy phosphate bonds. Then ATP can drive energy-consuming reactions and pathways by the hydrolysis of ATP to ADP and inorganic phosphate. In Section 2.2 we will describe the basic principles of bioenergetics in detail.

2.1.3 Enzyme activity regulation

The biological activity of proteins, including enzymes has to be regulated to ensure efficient flow of metabolic pathways and quick adaptation of metabolism according to changing needs of the organism. The control of metabolic pathways via enzymatic regulation takes place in four ways:

1. **Allosteric control** (modulation) of enzymes through activators and inhibitors. Allosterically controlled enzymes have special regulatory sites that are sensitive to particular small signal molecules. For example, the product of a metabolic pathway sometimes inhibits the first reaction that is unique for that pathway. This feedback inhibition prevents the unnecessary accumulation of the product. The activation of an enzyme by a precursor of the substrate of that enzyme is called feed-forward activation. Feedback inhibition and feed-forward activation stabilize metabolic pathways and make them more efficient.

2. **Reversible covalent modification** like phosphorylation, acetylation or glycosylation of enzymes is controlled, for example, by hormones. Similar to allosteric control, the conformation of the enzyme and hence its activity is modified.

3. **Regulation of the amount** of enzymes available can be regulated through gene-transcription, translation and proteolytic degradation.

4. **Isoenzymes** are homologous enzymes with slightly different catalytic, structural and regulatory properties. These enzymes allow varying regulation of the same reaction at distinct tissues, subcellular localizations or times.

2.1.4 Compartmentalization

Eukaryotic cells are organized into different membrane-surrounded compartments also called subcellular locations where each location is specialized for a specific set of cellular functions. A consequence of this compartmentalization is spatially distinct sets of enzymes, metabolites and whole pathways which enable a better fine-tuning of the metabolism [Hrazdina and Jensen, 1992]. The spatial organization of several selected pathways is shown in Fig. 2.3. For example, the enzymes of glycolysis are localized in the cytoplasm, those of the

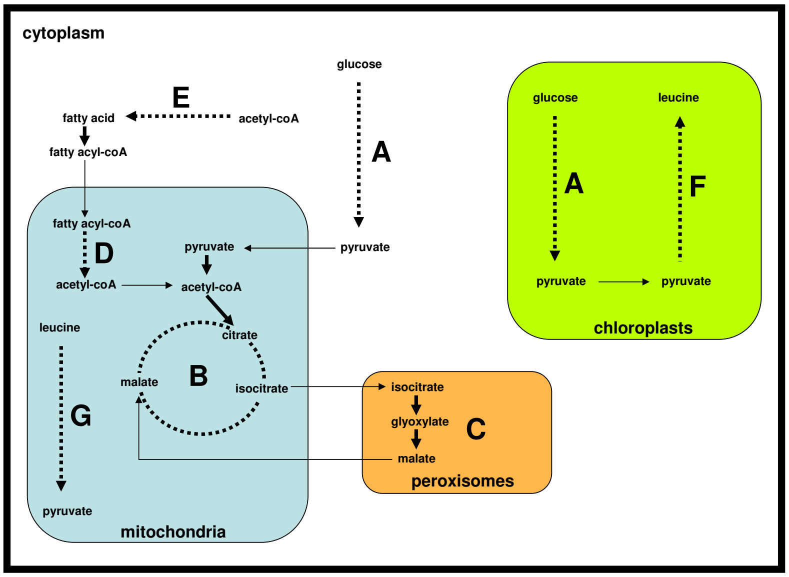


Figure 2.3: The spatial organization of several pathways: glycolysis (A), TCA-cycle (B), glyoxylate shunt (C), fatty acid degradation (D) and its synthesis (E) and leucine synthesis (F) and its degradation (G).

tricarboxylic acid cycle (TCA) in mitochondria and those of the glyoxylate-bypass in the peroxisomes (glyoxisomes). In plants, however, glycolysis also takes place in the chloroplasts where most of the amino acid biosynthesis pathways are also localized. Some of the TCA enzymes are also present in the peroxisomes in several organisms [Tolbert, 1981]. The presence of a pathway in several compartments is based on differently localized isoenzymes which are often regulated differently as already mentioned in the previous section. Control of the flux of metabolites from one compartment to another also regulates metabolism. Furthermore, a general principle of metabolism is that there are distinct biosynthesis and degradation pathways [Berg *et al.*, 2002]. Separate pathways are necessary for reasons of energetics and these support the control of metabolism. Control is further enhanced by pathway compartmentalization. For example, fatty acid degradation is localized in mitochondria and fatty acid synthesis in the cytoplasm. In *A. thaliana*, for example, leucine biosynthesis takes place in chloroplasts and degradation in mitochondria [Diebold *et al.*, 2002]. The basics of the protein sorting process in eukaryotic cells are described in Section 2.3.

2.2 Bioenergetics

2.2.1 Principles of thermodynamics

Cellular metabolism uses complex reaction cascades or networks to optimally exploit and transform the energy of nutrients or light. For example, plants use the process of photosynthesis to transform energy in the form of light into the chemical energy of ATP and other forms of energy. To understand metabolism better we have to keep in mind that chemical reactions follow the laws of thermodynamics.

The first law states the principle of energy conservation, which means that the total amount of energy in the universe is constant. In other words, it is not possible to create or destroy energy. However, energy can be converted from one form to another. The second law states that the disorder of the universe always increases. The discovery of the first and second laws of thermodynamics led to the definition of three thermodynamic quantities, which will be explained in the following sections. This kind of thermodynamic information can be used, for example, to calculate the equilibrium of chemical reactions and to predict whether a particular reaction can take place under given environmental conditions.

System: Thermodynamics defines a system as a partition of the space and the remaining part of the space as the environment of the system. Examples of systems are a whole living organism or a subcellular compartment.

Enthalpy: The enthalpy or heat (H) of a thermodynamic system, measured in kJ/mol, is defined by

$$H = U + PV$$

where U is the internal energy of the system, which depends on the temperature T, P is the pressure and V the volume of the system. The enthalpy of a reaction can be expressed as the difference in enthalpy between the products and educts (ΔH). If a reaction emits energy or heat, it is called exothermic ($\Delta H < 0$) and endothermic ($\Delta H > 0$) if heat has to be taken from the environment. At standard conditions ($T = 298.15$ K, $P = 10^5$ Pa, educts and products initially present at 1 mol/l concentrations), the reaction enthalpy is constant and expressed by ΔH^0.

Entropy: Entropy (S) can be described as a measure of the order of a system. The entropy of a system grows with a decrease of order. For example, entropy is the driving force for the diffusion of particles from a more highly concentrated solution towards one of lower concentration. Like enthalpy, the entropy of a system has a constant value under defined conditions and is measured in J/(K mol). Ludwig Boltzmann defined entropy by the relation

$$S = k_b \ln W$$

where k_b is the Boltzmann constant and W the total number of different states which can be captured by the particles of a system. Analogous to the change in enthalpy, there is also an entropy change (ΔS) in a chemical reaction. Under standard conditions, the reaction entropy is expressed by ΔS^0.

Gibbs energy: The Gibbs energy (G) is the thermodynamic measure of the driving force of a chemical reaction and is enhanced by both an enthalpy decrease ($\Delta H < 0$) and an entropy increase ($\Delta S > 0$) where the absolute entropic contribution also depends on temperature T. The change in Gibbs energy is defined as

$$\Delta G = \Delta H - T \Delta S.$$

The sign of ΔG indicates the favored direction of the reaction:

$\Delta G < 0$	the reaction runs spontaneously while releasing energy
$\Delta G = 0$	the reaction is at an equilibrium
$\Delta G > 0$	the reaction cannot run spontaneously and requires the supply of energy from the environment.

Reactions with $\Delta G < 0$ are called exergonic and with $\Delta G > 0$ endergonic. Since the Gibbs energy of reactions is additive, it is possible to couple an endergonic reaction with an exergonic one if the resulting overall reaction is exergonic. In this case, the exergonic (part-) reaction delivers the energy that is needed by the endergonic reaction. In cellular metabolism, reaction coupling is a frequently observed phenomenon. A well-known example is the releasing energy of ATP hydrolysis which is used to drive many endergonic reactions.

Whereas ΔG depends on the educt and product concentrations, ΔG^0 represents the change in Gibbs energy for a reaction under standard conditions ($T = 298.15$ K, $P = 10^5$ Pa, educts and products initially present at 1 mol/l concentrations). The actual Gibbs energy change ΔG as a function of the concentrations and the temperature T is defined by

$$\Delta G = \Delta G^0 + RT \ln(\prod [p_i] / \prod [e_j])$$

for a reaction $e_1 + e_2 + ... \rightleftharpoons p_1 + p_2 + ...$ where R is the gas constant. If the reaction is at chemical equilibrium, then $\Delta G = 0$ and

$$\Delta G^0 = -RT \ln K$$

where K is the equilibrium constant.

By convention, the Gibbs energy of all pure chemical elements is defined as null under standard conditions. Then the standard Gibbs energy of formation ($\Delta_f G^0$) for each non-elementary compound corresponds to the change in Gibbs energy in the formation of one mol of the compound from its elements under standard conditions. In the following, the term $\Delta_r G^0$ is used to distinguish the Gibbs reaction energy from the Gibbs energy of formation of compounds. If $\Delta_f G^0$ is known for all educts and products of a reaction, it is possible to calculate $\Delta_r G^0$ using the equation

$$\Delta G_r^0 = \sum \Delta G_f^0(products) - \sum \Delta G_f^0(educts).$$

2.2.2 Biochemical thermodynamics

The standard conditions for the study of biochemical reactions under "near physiological conditions" recommended by the IUPAC-IUBMB Joint Commission on Biochemical Nomenclature [Alberty, 1996] are $T = 298.15$ K (or $T = 310.15$ K), $P = 10^5$ Pa, pH 7.0, pMg 3.0, $I = 0.25$ mol/l where pMg is the free concentration of magnesium (or other metal) ions and I is the ionic strength of the dilute aqueous solution in which the biochemical reaction takes place. Consideration of the ionic strength I, the pH and pMg in addition to T and P requires the adjustment and transformation of the thermodynamic quantities (G, H and S) described in the previous section. In the following, we will restrict the discussion of these effects on the Gibbs energy G only.

Ionic strength: The ionic strength is a function of the concentrations of all ions in a solution and is defined by

$$I = 0.5 \sum_i c_i z_i^2$$

where the sum runs over the products of the molar concentration c_i with the squared charge number z_i for all ions i. The effects of the ionic strength on the thermodynamic quantities of ionic species in dilute aqueous solutions is significant. Based on the extended Debye-Hückel theory, the standard Gibbs energy of formation of species i at ionic strength I and 298.15 K is adjusted by

$$\Delta_f G_i^0(I) = \Delta_f G_i^0(I = 0) - 2.91482 z_i^2 I^{\frac{1}{2}} / (1 + B I^{\frac{1}{2}})$$

where z_i is the charge number of species i and $B = 1.6\,\mathrm{l}^{\frac{1}{2}}\mathrm{mol}^{-\frac{1}{2}}$ is an empirical constant that is taken to be independent of the temperature.

Transformed Gibbs energy: The Gibbs energy G provides the criterion for spontaneous chemical change and the corresponding equilibrium of reactions at specified T, P (and I). However, this is not the case for biochemical reactions if the pH is held constant. Furthermore, different equilibriums are obtained for different pH values. This fact led to the definition of the new thermodynamic quantity $G^{'}$ called the transformed Gibbs energy and the corresponding apparent equilibrium constant $K^{'}$. It is necessary to use $G^{'}$ and $K^{'}$ if the pH is a specified experimental condition. Alberty [1992a,b] applied a Legrendre transform to define $G^{'}$ as

$$G^{'} = G - n(H^{+})\mu(H^{+})$$

where $n(H)$ is the total amount of hydronium ions in the thermodynamic system and $\mu(H) = \Delta_f G^0(H^{+}) + RT\ln(10^{-\mathrm{pH}})$ is the chemical potential of H^{+}. The change in transformed Gibbs energy of a reaction under biochemical standard conditions can directly be inferred from the apparent equilibrium constant $K^{'}$ or the transformed Gibbs energies of formation of the educts and products using

$$\Delta_r G^{'0} = -RT\ln(K^{'}) = \sum_j (\Delta_f G_j^{'0} * p_j) - \sum_i (\Delta_f G_i^{'0} * e_i)$$

where e_i and p_j are the stoichiometric coefficients of the educts and products in the biochemical reaction. Applying Alberty's Legrendre transform, the standard transformed Gibbs energy of formation for the species i is calculated by

$$\Delta_f G_i^{'0} = \Delta_f G_i^0 - N_H(i)RT\ln(10^{-pH})$$

where $N_H(i)$ is the total amount of hydrogen atoms in species i and $\Delta_f G_i^0$ is the standard Gibbs energy of species i under (chemical) standard conditions with specified T, P and (I). Chemical equations are written in terms of species with balanced hydrogen atoms and electric charges. This is different to biochemical equations with specified pH written in terms of reactants, that are sums of species. The reason is that many educts and products of biochemical reactions are present as a mixture of different species in the neighborhood of pH 7. For example, ATP forms the species ATP^{4-}, HATP^{3-} and $\mathrm{H_2ATP}^{-}$. When a biochemical reactant consists of several species, the standard transformed Gibbs energy of formation for the reactant has to be calculated by combining the $\Delta_f G_i^{'0}$ values of its species using the following equation

$$\Delta_f G^{'0}(reactant) = -RT\ln(\sum_i \exp(\frac{-\Delta_f G_i^{'0}}{RT})).$$

Note that the apparent equilibrium constant $K^{'}$ is also expressed in terms of sums of species. We can use adenosine triphosphate hydrolysis as an example. The reaction can be represented by the chemical equation

$$\mathrm{ATP}^{4-} + \mathrm{H_2O} \rightleftharpoons \mathrm{ADP}^{3-} + \mathrm{H_2PO_4^{2-}} + \mathrm{H}^+$$

with balanced hydrogen atoms and electric charges. In a dilute solution the equilibrium constant depends on the temperature and pressure and is given by

$$K(T,P) = \frac{[\mathrm{ADP}^{3-}][\mathrm{H_2PO_4^{2-}}][\mathrm{H}^+]}{[\mathrm{ATP}^{4-}](c^0)^2}$$

where c^0 is the standard state concentration of 1 mol/l which makes the equilibrium constant dimensionless. The biochemical reaction equation at specified pH can be written as

$$\mathrm{ATP} + \mathrm{H_2O} \rightleftharpoons \mathrm{ADP} + \mathrm{P_i}$$

with the corresponding apparent equilibrium constant

$$K'(T,P,\mathrm{pH}) = \frac{[\mathrm{ADP}][\mathrm{P_i}]}{[\mathrm{ATP}]c^0}.$$

This representation is recommended by the IUPAC-IUBMB Panel on Biochemical Thermodynamics [Alberty, 1996] to distinguish biochemical equations from chemical ones.

Free concentration of metal ions: We have seen that the Gibbs energy G is not the criterion for spontaneous chemical change and equilibrium if the pH is specified as an additional independent variable. The same is true if the free concentration of magnesium ions pMg ($-\log_{10}([\mathrm{Mg}^{2+}]/c^0)$) or other metal ions is specified. In this case, the definition of the transformed Gibbs energy has to be extended to

$$G' = G - n'(H^+) * \mu(H^+) - n'(\mathrm{Mg}^{2+}) * \mu(\mathrm{Mg}^{2+})$$

where $n'(\mathrm{Mg}^{2+})$ is the total amount of magnesium ions in the system and $\mu(\mathrm{Mg}^{2+}) = \Delta_f G^0(\mathrm{Mg}^{2+}) + RT\ln(10^{-\mathrm{pMg}})$ is the chemical potential of magnesium. The treatment of metal ions like magnesium is important because these ions can be bound by phosphorylated reactants like ATP and change their thermodynamic properties. In the case of ATP, additional species may be present, e.g., MgATP^{2-}, MgHATP^{-} and $\mathrm{Mg_2ATP}$ depending on pH and pMg. ATP with bound magnesium forms a stable complex that supports efficient delivery of energy in metabolism.

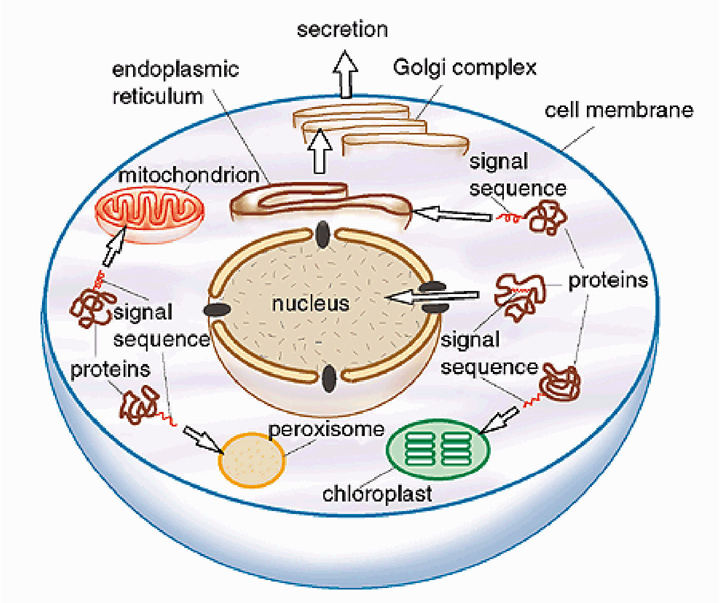


Figure 2.4: Cellular compartmentalization and protein sorting [The Nobel Assembly at Karolinska Institutet, 1999]

2.3 Protein sorting

2.3.1 Intracellular compartments and protein transport

A eukaryotic cell is organized into several different membrane-enclosed compartments (organelles) that are functionally specialized. Since proteins play an essential role in the functioning of a cell, it is important that they arrive at those subcellular localizations where their function is needed. Often proteins can fulfill their tasks only at a specific place because they require particular environmental conditions or interacting partners. However, organellar membranes (lipid bilayers) are impermeable to most proteins and hence, specific transport systems are required. This is necessary because nearly all proteins are synthesized at the ribosomes in the cytoplasm. Therefore, proteins that work outside the cytoplasm contain sorting signals in their amino acid sequence which are recognized by receptor molecules in transport machineries. Proteins without such sorting signals remain in the cytoplasm. See Fig. 2.4 for an illustration of cellular compartmentalization and protein sorting. The biological function of the nucleus is the storage of genetic information in the form of deoxyribonucleic acid (DNA), synthesis of mRNA (transcription) and assembly of the ribosomes. Proteins are transported from the cytoplasm to the nucleus via pore complexes in the nuclear membrane. The transport of proteins localized in the endoplasmic reticulum (ER), mitochondria, chloroplasts and peroxisomes are transported via membrane-bound translocators. Mitochondria and chloroplast (only in plants) compartments are specialized in the synthesis of ATP. In the peroxisomes the beta-oxidation of fatty acids and other oxidative reactions take place. The main function of the ER is the synthesis of nearly all cellular lipids as well as the synthesis and modification (glycosylation) of all transmembrane or soluble proteins of the organelles involved in the secretory and biosynthesis pathway. The subcellular compartments of the secretory and biosynthesis pathway communicate via particular transport vesicles. For example, proteins can be transported from the ER to the Golgi apparatus and then to the endosomes, lysosomes, plasma membrane or extracellular space. The Golgi apparatus receives lipids and proteins from the ER and distributes them after covalent modification to other localizations. Carbohydrates are also synthesized in the Golgi apparatus and often added to the lipids and proteins received from the ER. The intracellular digestion of proteins takes place in the lysosomes and, therefore, many acid hydrolases can be found there. The endosomes receive molecules for digestion and develop into lysosomes. In plants and fungi, there are no lysosomes. Intracellular digestion takes place in the vacuoles. Additionally, vacuoles maintain storage functions. Plant cells have especially large vacuoles and use them for storing nutrients, metabolites or waste products. Vacuoles can also be regarded as equivalent to the extracellular space of animals. The cellular plasma membrane encloses the cytoplasm, transduces external information and receives and releases metabolites.

2.3.2 Sorting signals

There are two main types of sorting signals: signal sequences and signal patches. Much more is known about signal sequences than about signal patches. In general, a signal patch is a specific thre-dimensional structure of residues, which arises from protein folding. The amino acids that take part in such a signal patch can be far apart in the linear amino acid sequence.

A signal sequence consists typically of 15 - 60 continuous amino acids. The sequences are frequently cleaved from the mature protein by a signal peptidase. Signal sequences can also be located elsewhere in the protein, but they are frequently located at the end of the polypeptide (N-terminal or C-terminal). Normally, chemical properties like hydrophobicity are more important for the signal recognition process than the exact amino acid order. Therefore, the signal sequences for one target organelle can vary in order and length.

2.3.3 Transport into the nucleus

Proteins destined for the nucleus contain a nuclear localization signal (NLS) somewhere in their amino acid sequence. There are two different types of NLSs. A monopartite NLS is a short sequence which is rich in positively charged amino acids like lysine and arginine and nearly always contains proline. If the monopartite NLS is split into two parts it becomes bipartite. Each part is between two and four amino acids long, connected by a spacer which is about 10 amino acids long. An NLS can be either a signal sequence or a signal patch. The exact location of an NLS in the protein is not important, but it must be exposed on the surface of the protein. The NLSs are bound by nuclear import receptors.

2.3.4 Transport into the peroxisomes

The exact import process of peroxisomal proteins is still not understood completely, but depends on signal sequences at the C- and N-termini. The best-known signal consists of three C-terminal-specific amino acids. This signal is also known as the SKL motif (-Ser-Lys-Leu-COO-).

2.3.5 Transport into the mitochondria

The transport of mitochondrial proteins from the cytoplasm depends on an N-terminal targeting sequence and on protein translocators, which are multiprotein complexes. The sorting signal is called the mitochondrial targeting peptide (mTP). A membrane-associated signal peptidase cleaves the mTP after import. Mitochondrial targeting peptides are normally between 25 and 45 amino acids long and prefer to fold as an amphiphilic alpha-helix with mainly positively charged amino acids on one side (particularly arginine) and mainly uncharged, hydrophobic ones on the opposite side. Negatively charged residues are not common in mTPs.

2.3.6 Transport into the chloroplasts

Only plant cells contain chloroplasts. Protein transport into chloroplasts is similar to the mitochondrial import machinery. The receptors of the chloroplasts and the mitochondria can distinguish between mTP and cTP (chloroplast targeting peptide). The cTPs have highly variable lengths (20 - 120 amino acids), are enriched for hydroxylated amino acids (particularly serine) and contain very few negatively charged residues.

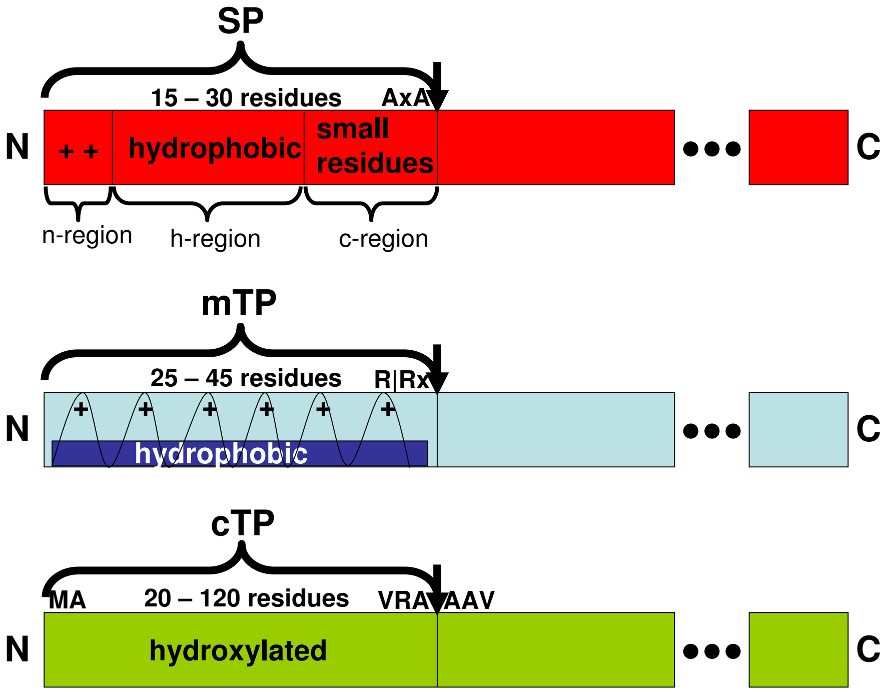


Figure 2.5: The three different kinds of N-terminal targeting sequences and their chemical properties: SP (top), mTP (middle) and cTP (bottom). Weakly conserved motifs around the cleavage sites are also shown.

2.3.7 Secretory pathway

The secretory pathway includes the ER, Golgi apparatus, lysosomes, and plasma membrane. Proteins destined for these organelles have an N-terminal targeting sequence called the signal peptide (SP) with a cleavage site for luminal proteins and another without a cleavage site for transmembrane proteins, called the signal anchor (SA). SPs are between 20 and 30 amino acids long and contain a short positively charged N-terminal segment, a central hydrophobic segment with eight or more non-polar residues and a more polar segment with mostly small residues. Frequently an alanine occurs at positions -1 and -3 before the cleavage site. In Fig. 2.5 the SP is compared to the other two kinds of N-terminal targeting sequences (mTP and cTP). All proteins of the secretory pathway are translocated into the ER at first, and from there further sorted to the other organelles or the extracellular space.

A special signal peptidase on the luminal site of the ER membrane cleaves off the SP after import and during the translation. The peptidase recognizes a cleavage site specifically. SPs without cleavage sites (SAs) serve as transmembrane segments. SAs are located more inside the protein and often have a larger hydrophobic segment. The imported proteins are automatically further transported to the other organelles or to the extracellular space by vesicles. ER-specific proteins contain a retention signal of four amino acids at the C-terminus (KDEL in one letter-code for luminal and KKXX for transmembrane proteins). Not all ER-specific proteins have such a retention signal. It is assumed that they remain in the ER because they form aggregations, which are too big to be packed into vesicles.

The Golgi apparatus is placed next to the ER and consists of batches of several cisternae. Proteins from the ER go through an ordered series of covalent modifications during their movement through the Golgi batches. Some of the modifications serve as markers for transport into other localizations. The proteins are partitioned into different kinds of packages for the plasma membrane, lysosomes or secretory vesicles. Since transport to the plasma membrane and the extracellular space occurs without a special signal, there are retention steps for Golgi-specific proteins. Membrane proteins and lipids are integrated into the plasma membrane and soluble proteins are released into the extracellular space.

Luminal proteins of the lysosomes have a mannose-6-phosphate (M6P) modification, which serves as a selection marker. Signal patches are responsible for selecting a protein to get the mannose-6-phosphate modification. It is known that lysosomal membrane proteins do not have this kind of modification, which means that there is probably an alternative transport path for them.

Chapter 3

Computational background and related work

Computational approaches to metabolic pathways can be roughly divided into two groups. The first group consists of constraint-based methods that apply convex analysis to the stoichiometric matrix of the network to calculate network-based or stoichiometrically balanced pathways. Methods of the second group are based on graph theory and apply shortest-path algorithms to a graph that represents the metabolic network. However, a major problem with recent methods is still the computational effort caused by the combinatorial explosion of the number of possible routes in a metabolic network at the genome scale.

In this chapter we present the computational background and related work with respect to methods for analyzing metabolic pathways. For a better understanding, we first introduce and discuss possible definitions of metabolic pathways. Then we give a historical overview of constraint-based and graph theory-based approaches including brief discussions about advantages and limitations of these methods. Since computational approaches require adequate network data as input, the last section presents the most relevant pathway databases that often serve as starting sources to build species-specific metabolic networks.

The focus in this chapter is on approaches computing metabolic pathways since this is the main topic of this thesis. Computational background and related work for the remaining subtopics including the calculation of atom mapping rules, the prediction of Gibbs reaction energies and the prediction of subcellular protein localization is presented in the corresponding subsections of Chapter 4.

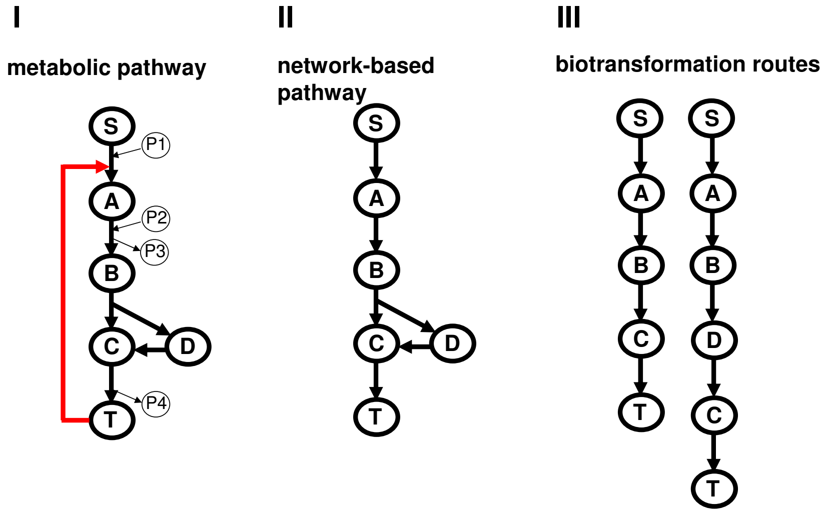


Figure 3.1: A standard metabolic pathway (I), the corresponding network-based pathway (II) and linear biotransformation routes (III).

3.1 Metabolic pathway definitions

In the literature, metabolic pathways are defined in different ways. More generic definitions are used in biochemical textbooks and pathway databases were the focus is on the presentation of pathways in the context of their historical discovery. In this context a metabolic pathway is often simply described as a connected set of enzymatic reactions that converts source metabolites into product or target molecules in a step-wise manner [Berg *et al.*, 2002].

The development of computational approaches for a systematic discovery of biologically meaningful pathways requires more formal definitions. Fig. 3.1(I) shows an example of a typical metabolic pathway that consists of five reactions, one source (S) as well as a target (T) compound, four intermediates (A, B, C and D), four side or pool compounds (P1, P2, P3 and P4) and a feedback inhibition of the initial reaction step (drawn in red). The pathway shown is branched because the third reaction splits its educt (B) into two products (C and D). The second product (D) is then further converted into the first product (C) by a subsequent reaction. Network-based metabolic pathway definitions [Schuster *et al.*, 2000] are the underlying concept of constraint-based approaches that compute extreme pathways (EPs) and elementary flux modes (EFMs) using convex analysis. A network-based or stoichiometrically balanced pathway represents a metabolic subnetwork in which the net production and consumption of the involved intermediates is zero. This stoichiometric constraint does not have to be fulfilled by the source and target compounds and a predefined set of pool metabolites. The corresponding network-based pathway in our example, computed by constraint-based approaches, is shown in Fig. 3.1(II). The network-based pathway is very similar to the real pathway. Regulatory aspects like feedback inhibition are not considered in either the pathway definition or computation. Side metabolites are also not part of the stoichiometric computation but always available from the predefined list of pool metabolites.

The pathway definitions discussed so far include pathways with branches and cycles. However, branching pathways are not directly considered by current graph theory-based approaches, which use path-finding concepts to compute linear biotransformation routes. In graph theory, a path is defined as a linear chain of nodes whereby each node is connected by an edge to the next node in the sequence. The path is called simple if it contains only distinct nodes. Cycles are produced if the last node in a path is also connected with the first node in the path. Based on this path concept, a linear biotransformation route is simply defined as an unbranched sequence of chemical reactions and metabolites where a source compound is converted into a target compound step by step. The biotransformation route is a cycle if source and target compounds are identical. As a consequence, graph theory-based approaches decompose our example pathway into two linear biotransformation routes which are shown in Fig. 3.1(III). The example pathway is indirectly available by merging these two routes. A linear definition in the context of pathway alignment was also formally introduced in a previous work [Chen and Hofestaedt, 2005].

Compared to network-based pathways, the main advantage of linear routes is that their computation is much easier. The use of efficient path-finding algorithms enables the detection of relevant pathways in metabolic networks at genome-scale. Despite these different network-based and linear definitions, we will often simply use the term pathway in the following.

3.2 Metabolic pathway analysis

3.2.1 Constraint-based approaches

Given a metabolic network and a set of external (pool) compounds, constraint-based approaches calculate all stoichiometrically balanced pathways (explained in Section 3.1) transforming a set of given source compounds into a set of sink (target) compounds. Depending on the underlying approach, these pathways are called extreme pathways or elementary flux modes.

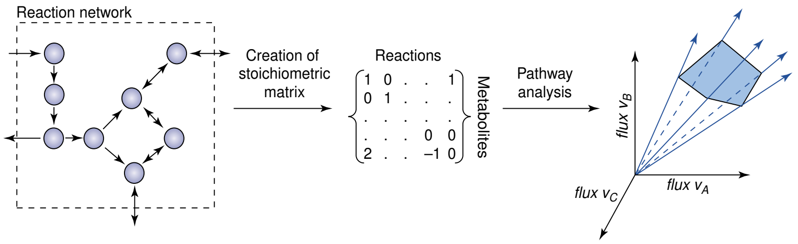


Figure 3.2: The basic concepts of steady-state network representation and analysis [Papin *et al.*, 2003].

Theoretical framework

A metabolic network can be represented by a stoichiometric $m \times n$ matrix $\mathbf{S}$. The m rows of $\mathbf{S}$ correspond to the metabolites and the n columns to the reactions in the network. The matrix element S_{ij} represents the stoichiometric coefficient of metabolite i in reaction j. Reaction educts receive negative and products positive values. A zero value is assigned to the matrix elements of metabolites not present in the corresponding reactions.

The change of compound concentrations in a metabolic network can be described by the dynamic mass balance equation

$$\frac{d\mathbf{x}}{dt} = \mathbf{S}\mathbf{v}.$$

The equation defines a system of ordinary differential equations where $\mathbf{x}$ is the concentration vector of all metabolites, $\mathbf{S}$ the stoichiometric matrix and $\mathbf{v}$ the vector of fluxes through the reactions. In steady-state, the concentrations of all metabolites are constant and the mass balance in the network can be represented by the equation $0 = \mathbf{S}\mathbf{v}$. This system of linear equations describes the so-called null space which is the set of all possible solutions for the flux vector $\mathbf{v}$ under steady-state conditions.

If all fluxes of the system are constrained to be non-negative ($v_i >= 0$), then the corresponding reactions are irreversible and the solution space is defined by a convex flux cone. Reversible reactions can be modeled by decomposing them into their respective forward and reverse directions. (See also Fig. 3.2 for a graphical illustration of the basic concepts.) Furthermore, the metabolites of the system under study have to be classified as internal or external according to whether or not they have to fulfill the steady-state condition. Typically, external metabolites represent pool metabolites, cofactors or by-products as well as the system sources and sinks. These compounds are called external because they also participate in additional reactions that are involved in external systems.

Extreme Pathways

Any point in the flux cone is a valid flux distribution and can be represented by a non-negative linear combination of the edge vectors which span the flux cone. These edges are also called the extreme pathways (EPs) of the network. The set of EPs is conceptually related to the concept of a basis in linear algebra and can be calculated from the stoichiometric matrix using convex analysis [Rockafellar, 1970], a branch of mathematics for analyzing a set of linear equations given a set of constraints. Algorithmic details of the approach can be found in Schilling *et al.* [2000].

Elementary Flux Modes

Elementary flux mode (EFM) analysis [Schuster *et al.*, 1999] is strongly related to the concept of extreme pathways. However, the algorithms for EFMs and EPs differ in their treatment of reversible and irreversible reactions. In EP analysis, each reversible reaction is decomposed into two separate reaction fluxes for the forward and backward directions. On the other hand, the calculation of EFMs is based on the split of the stoichiometric matrix into two parts, one for reversible and one for irreversible reactions. The computed set of EPs is always a subset of the resulting EFMs. The EPs represent a minimal set of EFMs and the remaining EFMs can be represented by a non-negative linear combination of the EPs. In other words, the set of EFMs represent, in contrast to the set of EPs, all feasible network-based pathways when reversible reactions are present. EFMs and EPs share the following two properties:

- The sets of EFMs and EPs are unique for a given network and a list of internal compounds.

- Each EFM or EP is non-decomposable, which means that it contains a minimum number of reactions in order to exist as a functional unit. The removal of any reaction does not allow the EFM or EP to operate as a functional unit.

In addition to these two properties, elementary flux modes represent the set of all network-based pathways within a metabolic network which are consistent with the second property. A further property specific to extreme pathways is that they are the systemically independent subset of the elementary flux modes. In other words, no extreme pathway can be represented as a non-negative linear combination of any other extreme pathway.

Advantages and limitations

Biotechnological and biomedical applications of extreme pathways and elementary flux modes involve the evaluation of network properties such as the optimal product yield, network robustness and pathway redundancy. Furthermore, the underlying analysis concept is mathematically well-defined. However, the enumeration of extreme pathways and elementary flux modes for a given stoichiometric matrix represents a computationally hard problem and can not be applied to genome-scale networks [Klamt *et al.*, 2002, 2003; Yeung *et al.*, 2007]. Genome-scale networks of microbes already contain approximately 1,000 or more reactions. Given a predefined distinction between internal and external compounds as well as a suitable number of irreversible reactions, constraint-based approaches are applied to networks with at most 100 reactions. The number of EPs and EFMs which have to be computed for a metabolic network increases drastically with the size and complexity of the network. For example, more than 500,000 different EFMs were calculated by Klamt *et al.* [2002] for a network with 110 reactions. In a second study by Yeung *et al.* [2007], the number of EPs in networks consisting of 904 and 3,311 reactions was estimated to be 3×10^{18} and 10^{29}, respectively.

3.2.2 Graph-theory approaches

Given a metabolic network, graph theory-based approaches use path-finding concepts to calculate linear biotransformation routes between given source and sink (target) compounds in a graph that represents the network.

Graph representation of metabolic networks

A mathematical graph $G = (V, E)$ is a data structure where V is a set of vertices (nodes) and E is a set of edges connecting pairs of nodes. The graph is directed/undirected if all edges $e = (v_1, v_2)$ are ordered/unordered node pairs. Furthermore, the graph is weighted if the edges are assigned weights according to a weighting function $w(e) : E \rightarrow \mathbb{R}$. Simple examples of the different graph types are shown in Fig. 3.3.

A metabolic network can be represented as a graph. Different graph representations for analyzing metabolic networks and pathways have been described in the literature. The most common graph types are summarized in this section (see also Fig. 3.4). More detailed information about different graph representations can be found in Deville *et al.* [2003].

- **Compound graphs:** In a compound graph, the nodes represent chemical compounds or metabolites and each edge connects compound E

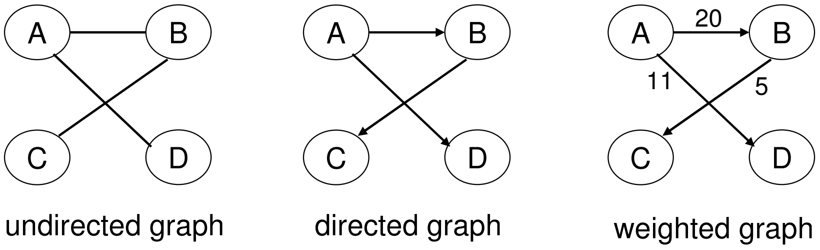


Figure 3.3: Examples of different graph types.

with compound P if E is an educt and P a product in the same reaction.

- **Reaction graphs:** In a reaction graph, the nodes correspond to the reactions in the network. Here, each edge connects reaction R_1 with R_2 if there is a compound that is a product of R_1 and an educt of R_2.
- **Bipartite graphs:** In a bipartite graph, there are two different types of nodes which represent the reactions and compounds of metabolic network. Edges between the nodes represent the educt/product relationships between compounds and reactions.
- **Hypergraphs:** Hypergraphs generalize compound graphs and can be seen as an equivalent representation of the bipartite graphs. Here, each hyperedge relates the set of educts of a reaction with the set of its products.

In all graph types, the reaction directions can be represented by directed edges. Reversible reactions can be modeled by multiple edges (with the opposite direction) between two nodes or by decomposing the reaction into two different nodes, one for the forward and one for the backward direction. If the reaction direction is irrelevant, undirected edges can also be used. Each graph representation has its advantages and disadvantages and the final choice depends on the available information, the purpose of the analysis and the graph algorithms used [van Helden *et al.*, 2002].

Path-finding concepts

The graph-theoretic representation of metabolic networks yields a well understood framework for searching pathways within large-scale networks by the application of efficient path-finding algorithms. In this section, we will present different path-finding concepts including a discussion about their usefulness for the detection of pathways in metabolic networks.

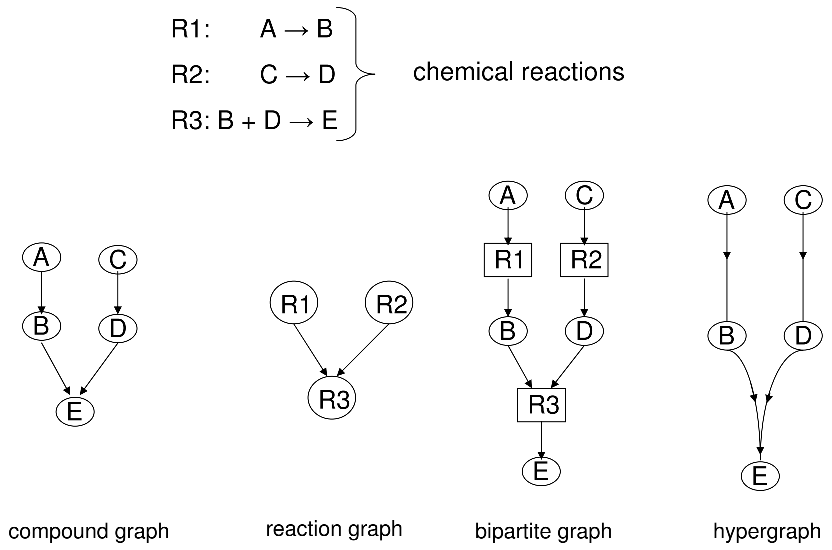


Figure 3.4: Common types of graph representations.

Path-finding algorithms can be used to compute one or more optimal paths connecting two different nodes in a graph. There is a huge number of real-world applications for path-finding including the detection of relevant routes in metabolic networks. The problem of finding the shortest path with a minimum number of nodes or with the minimum total weight in a weighted graph has received special interest. In the latter case, the shortest path is sometimes also called the lightest or cheapest path. We can distinguish a number of common variants of the shortest path problem:

- **single-pair shortest path:** The shortest path between two different nodes.
- **single-source shortest path:** The shortest paths between a given source node and all other nodes.
- **single-destination shortest path:** The shortest paths to a given destination node from all other nodes. The problem is also called the reverse single-source shortest path problem because it can be solved simply by reversing the edge directions in the graph.
- **all-pairs shortest path:** The shortest path between any two different nodes.

Several algorithms have been developed in the past for solving the shortest path problems. These algorithms differ in their potential range of applications and their underlying complexity. The simplest approach for finding the shortest path is **breadth-first search**. However, there are two main drawbacks to this method. The first is based on the fact that breadth-first search is an uninformed search because it traverses exhaustively the whole graph beginning with the source node but without considering the destination node until it is found. Furthermore, the standard breadth-first search algorithm requires an unweighted graph because the shortest path found always consists of a minimum number of steps, which is not the case for weighted graphs in general. Weighted graphs require improvements of the algorithm and make the search more complex. These difficulties are avoided by using more advanced search techniques. Algorithms based on **best-first search** find the shortest path from a source to a destination node using a heuristic evaluation function $F(v), v \in V$. This approach represents an informed search because a heuristic is used to guide the search and to speed up the path-finding. The heuristic function $F(v)$ can depend on any additional problem-specific information. In general, $F(v)$ utilizes information derived from the starting node to the current node v (the search up to node v) as well as from the current node v to the destination node. Best-first search examples are **Dijkstra's algorithm** [Dijkstra, 1959] and its generalization the **A* algorithm** [Hart *et al.*, 1968]. Both algorithms can be applied to directed graphs with non-negative edge weights. The **Bellman-Ford algorithm** [Bellman, 1958] is very similar to Dijkstra's algorithm but can also deal with negative edge weights. Using the **Floyd-Warshall algorithm** [Floyd, 1962] it is possible to efficiently solve the all-pairs shortest path problem in a weighted, directed graph. This can also be done using **Johnson's algorithm** [Johnson, 1977], which is, however, especially useful for sparse graphs.

When analyzing metabolic pathways the detection of alternative routes leading from a source to a target metabolite is of great importance. However, this cannot be achieved by simply computing the shortest path. A better strategy is to search for the k-shortest paths between two given nodes in a graph representing the metabolic network. The meaning of the k-parameter is to find, for example, the shortest path ($k = 1$), additionally the second shortest path ($k = 2$), the third shortest path ($k = 3$) etc. The four mentioned shortest path problems can also be extended to its k-shortest paths versions. Like the shortest path problem, the problem of finding the k-shortest paths also has a long history in computer science. One of the earliest discussions about the problem was published by Hoffman and Pavley [1959]. After this, numerous algorithms for many variations of the problem have been described. An exhaustive collection of these papers is also available online (*http://liinwww.ira.uka.de/bibliography/Theory/k-path.html*).

Eppstein's algorithm [Eppstein, 1998] represents a significant improvement in the field. The algorithm creates an implicit representation of the

k-shortest paths for a given source/destination node pair in a directed graph with n nodes and m edges in $\mathcal{O}(m+n\log n+k)$. Furthermore, the k-shortest paths to a given destination from every node in the graph can be computed in $\mathcal{O}(m+n\log n+nk)$ time. The paths themselves can be traversed from the implicit representation using breadth-first search.

The following sections describe the current state-of-the-art approaches to metabolic pathways that are based on path-finding concepts.

Automated Metabolic Reconstruction

Masanori Arita, the developer of the Automated Metabolic Reconstruction tool, introduced the use of a k-shortest path algorithm to search for routes with a minimum number of reaction steps in a metabolic network [Arita, 2000, 2003]. He also proposed the incorporation of atom mapping rules into the path search. For a chemical reaction, such a rule defines which educt atom is transferred or mapped to which product atom. Using these rules, paths found can be validated according to the structural moiety constraint. This constraint states that a biochemically feasible route transfers at least one atom of the source to the target metabolite.

A fundamental problem of the approach is that for approximately 30% to 40% of reactions stored in pathway databases, it is not possible to easily compute an atom mapping rule. These reactions contain metabolites without given structural information or general molecules like “an alcohol” or the reaction equation is unbalanced because of an incomplete or erroneous annotation. The reactions thus require time-consuming manual checking. Furthermore, the applied calculation of atom mapping rules is based on a maximum common subgraph approach that represents a heuristic solution to the problem and, therefore, fails to find the correct atom mapping rule in some cases. Additionally, the calculation requires some manual preprocessing of the reaction equations. A further problem with respect to the path search is that although the structural moiety constraint is fulfilled, the search for the shortest path with a minimum number of reaction steps still bears the risk of finding meaningless results with pool metabolites as intermediates. However, the approach represents a milestone in the field because it introduces a path validation concept based on atom mapping rules and the use of a k-shortest path algorithm.

PathMiner

The PathMiner approach [McShan *et al.*, 2003] is based on a chemically motivated heuristic to guide a search in a state space. The compounds are represented using chemical descriptors as points or states in a hyperspace based on the composition of their atoms and bond types (e.g. C, N, C-C, P=O and so on; 145 overall). Biochemical reactions are abstracted as transitions between the compound states and expressed as a state vector difference. Pathways are predicted by searching a route from an initial compound to the destination compound through a series of state transitions. The search is guided by best-first search using a heuristic evaluation function. The function is used to minimize the summed vector differences between the pairs of succeeding compound states in the final route.

Limitations of the PathMiner approach are that it computes only one metabolic route between a given source and target which is, of course, a drawback for studying alternative routes. Furthermore, it does not favor biochemical transformations that involve the transfer of larger functional groups between the metabolites like phosphate groups which appear in many metabolic processes that require the phosphorylation of compounds. The reason is that the heuristic is specialized on the transition of very similar compounds and therefore can only find pathways which are “chemically parsimonious”. Another problem is that there is no evaluation of the quality of the routes found with respect to experimentally determined pathways. The authors only compared the computational performance of the heuristic search to that of uninformed blind search approaches.

Pathway Hunter

The graph-representation of the Pathway Hunter tool [Rahman *et al.*, 2005] contains only compound nodes. Edges represent educt/product relationships between compounds in the same reaction. However, only structurally similar compounds are connected by edges, based on a mapping function. Therefore, the mapping function combines two measures. The first measure is the Tanimoto coefficient [Willet *et al.*, 1998] calculated from the chemical fingerprints of the compounds. The second measure is derived from the atomic mass contribution of an educt/product pair with respect to all compounds of a reaction. A breadth-first search algorithm calculates the shortest paths between a given source and a target compound.

A principle drawback of this approach is that the necessary structural information is not available for all compounds in the metabolic database used. Examples of such compounds are generic molecules like “an alcohol” or diverse macromolecules. Reactions which include these compounds are excluded from the standard approach. Given a source and a target compound, the number of computable transforming routes is limited to those that share the shortest length inferred using a breadth-first search algorithm.

Degree-weighted metabolic networks

In the degree-weighted metabolic networks approach [Croes *et al.*, 2006], the metabolic network of an organism is mapped on a bipartite graph, including all compounds and reactions as nodes. Directed edges connect the compound nodes (educts and products) with the reaction nodes. Both directions of a reaction are represented by two independent nodes per reaction. The key idea of a degree-weighted metabolic network is to assign each compound node a weight equal to its degree (e.g. the number of in- and outgoing edges) and each reaction node the weight 1 by default. The weight of a path in the graph is then defined as the sum of the weights of its nodes. This implies that the overall weight of a path is much larger if it contains highly connected compounds like typical pool metabolites or co-factors (e.g. NADP, ATP or water). Searching for paths of lowest weight significantly reduces the probability of finding unfeasible biotransformation routes that contain pool metabolites (network hubs) as intermediates between two successive reactions. Up to five paths of lowest weight (not a limitation of the algorithm) can be found by the use of a depth-first back-tracking algorithm.

An advantage is that the structural information of the compounds is not needed. However, a fundamental problem of the lightest-path search is its inability to handle important biotransformation routes involving the biosynthesis of pool metabolites (e.g. purine biosynthesis, in which AMP and ADP are intermediates). The method fails to reconstruct these routes because pool metabolites participate in many reactions of other transformation processes and, therefore, are assigned very large node weights. A further problem is that of routes passing pathways of the core metabolism like glycolysis or the TCA cycle, because highly connected metabolites like pyruvate or acetyl-CoA are involved. Fig. 3.5 shows more details of this issue. The transformation of adenylo-succinate to dADP is part of purine metabolism and is shown on the left side of Fig. 3.5. On the right side an alternative but biochemically irrelevant pathway is shown. For each reaction, main metabolites are drawn in black and side metabolites in red. Irrelevant intermediate steps, with respect to the adenylo-succinate/dADP conversion, are also drawn in red. Furthermore, the number of reactions (the weights) in which each intermediate participates as educt or product in a typical genome-scale metabolic network is presented within adjacent rectangles. Searching for the path with lowest weight will fail in this case because the irrelevant route obtains an overall weight of 33, which is significantly lower compared to that obtained for the textbook route (253).

Overall, this method represents a milestone in the field because it introduces the use of a weighting scheme as optimization criteria in order to detect

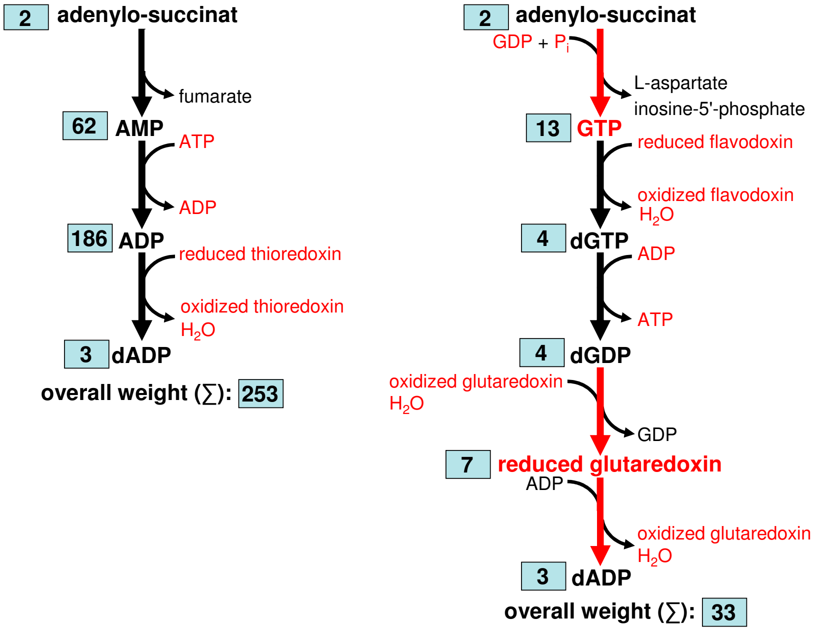


Figure 3.5: This figure depicts the problem of the degree-weighted metabolic networks approach to find relevant pathways that contain highly connected intermediates like AMP and ADP. Therefore, a relevant but heavy pathway is shown on the left and a light but irrelevant pathway on the right. Relevant transformation steps are drawn in black and irrelevant in red. In each step, main metabolites are drawn in black and side metabolites in red. Typical numbers of reactions (weights), in which each intermediate participates as educt or product are enclosed by rectangles.

meaningful pathways within the k-shortest or lightest paths. Furthermore, an evaluation approach is suggested and applied for validating the pathfinding performance against experimentally determined metabolic pathways extracted from EcoCyc. Such a systematic evaluation was not performed for the approaches described earlier in this section.

3.3 Metabolic pathway databases

An exhaustive list of databases focusing on metabolic pathways can be found at *http://www.pathguide.org/*. The two most popular databases are KEGG [Kanehisa, 1996] and BioCyc [Karp *et al.*, 2005] which will be briefly described in the following two sections.

3.3.1 KEGG

The Kyoto Encyclopedia of Genes and Genomes (KEGG) is a database that integrates manually curated genomic, chemical and systemic information in the form of metabolic and regulatory pathways. The KEGG project was initiated in 1995 and is maintained as part of the research projects of the Kanehisa Laboratories in the Bioinformatics Center of Kyoto University and the Human Genome Center of the University of Tokyo.

The aim of the KEGG project is to establish a computer representation of the biological system whereas biological objects (organisms, genes, enzymes, pathways, reactions, etc.) and their relationships are available as separate database entries and direct links. Each database entry or KEGG object is assigned a unique identifier which allows direct access to the corresponding database entry via the internet. Many other online biological databases are already linked to KEGG. Furthermore, KEGG is a valuable resource for bioinformatics and computational systems biology because a flat file version of the whole database is freely available and can be downloaded at *ftp://ftp.genome.jp/pub/kegg/*.

KEGG comprises 19 sub-databases which are completely described in Kanehisa [1996]. However, the six databases (KEGG GENOME, KEGG GENES, KEGG PATHWAY, KEGG COMPOUND, KEGG ENZYME and KEGG REACTION) can be considered the core databases and will be briefly described:

- **KEGG GENOME** contains genomic information on more than 800 organisms

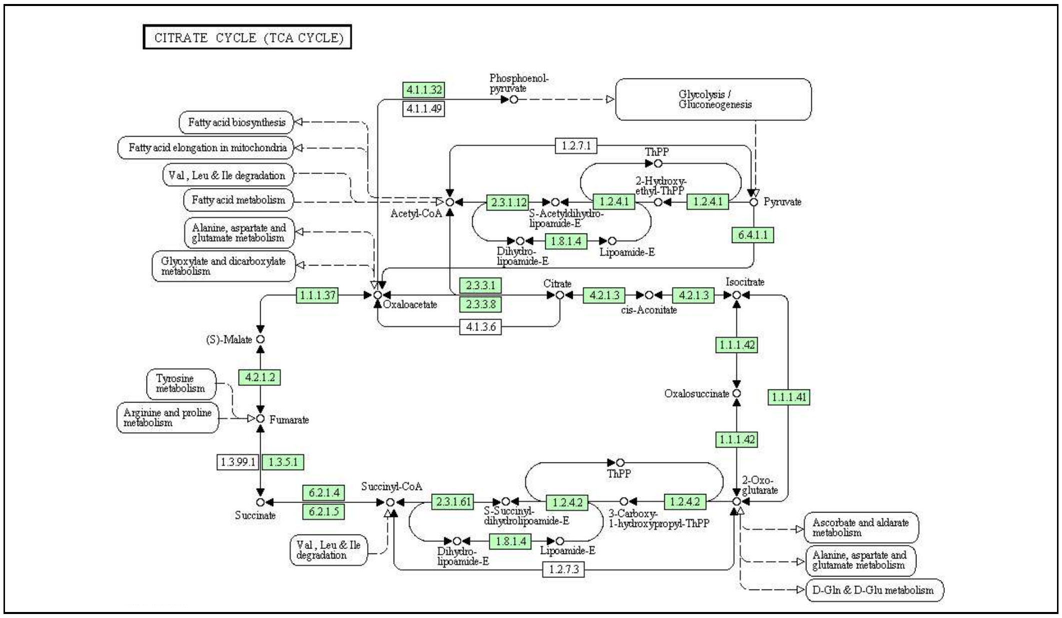


Figure 3.6: A screenshot of the KEGG wiring diagram that represents the TCA cycle reference map.

- **KEGG GENES** contains gene and protein sequence information from high-quality genomes
- **KEGG PATHWAY** contains reference maps in the form of wiring diagrams which combine pathway information from multiple organisms. The TCA cycle reference map is shown in Fig. 3.6. Each map can be colored in green to show which enzymatic reactions occur in a selected organism based on the set of enzymes identified from its genome.
- **KEGG COMPOUND** contains information about metabolites and other chemical compounds like trivial names, chemical formulas, links to other databases or the two-dimensional structure in MOL format.
- **KEGG ENZYME** contains all relevant information about enzymes.
- **KEGG REACTION** contains all details about chemical reactions like the assigned EC number(s) or the reaction equation.

3.3.2 BioCyc

The BioCyc project has been developed by the Bioinformatics Research Group at SRI International, directed by Dr. Peter Karp. BioCyc is a collection of more than 300 databases where each database describes the genome

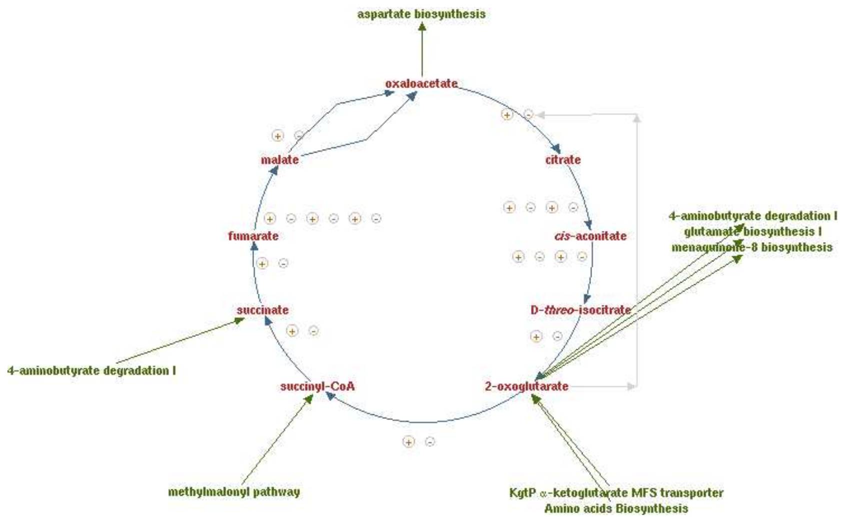


Figure 3.7: Screenshot of the EcoCyc TCA cycle pathway.

and metabolic pathways of a single organism. The EcoCyc database describes *E. coli* and is manually curated from the literature. The metabolic pathways stored in the remaining species-specific databases were computationally predicted using the Pathway Tools software [Karp *et al.*, 2002] based on the MetaCyc [Caspi *et al.*, 2006] pathway database. MetaCyc contains more than 1,100 experimentally verified metabolic pathways from more than 1,500 different organisms. Like EcoCyc, the MetaCyc database is curated from the literature.

The BioCyc collection offers electronic reference sources on the pathways and genomes of different organisms. The main difference between BioCyc and KEGG is the underlying ontology used to define pathways [Green and Karp, 2006]. The BioCyc ontology defines a metabolic pathway as a conserved atomic module within the metabolic network of a single organism. But there is a more or less strict distinction between biosynthesis and degradation modules as well as alternative pathways. KEGG pathways are on average 4.2 times larger than BioCyc pathways and represent the combined biosynthesis and degradation pathway information of multiple organisms; organism-specific aspects within such a reference pathway can be highlighted using green color. The corresponding TCA cycle example extracted from EcoCyc is shown in Fig. 3.7. Like KEGG, there is also a downloadable flat file version for each BioCyc database (*http://biocyc.org/download.shtml*).

Each database in the BioCyc collection is based on the same database scheme within an object database management system, the Ocelot database system. The system uses a complex object-oriented data model that is based on a taxonomic hierarchy of classes, instances of classes (called frames) and slots which represent attributes of the classes or relationships between them. Each class represents a biological entity like an organism, a reaction, an enzyme, a protein etc. The data model contains more than 1,000 class definitions, which demonstrates that the model is much more complex compared with the simpler KEGG model. The BioCyc scheme contains, for example, very detailed enzyme modulation types like ALLOSTERIC-INHIBITOR, PROSTHETIC-GROUP and so on. Furthermore, the class instances are annotated with numerous comments and extensive literature references.

Chapter 4

Approaches and results

The computational detection of all relevant pathways transforming a particular source into a product in genome-scale metabolic networks has numerous applications in systems biology. However, the combinatorial explosion of possible routes in large networks represents a challenging task.

In this chapter we introduce several novel approaches which can be used jointly in order to deal with the complexity of the underlying problem and to efficiently find the most relevant routes. Methodological details and results of the developed approaches, concerned with the calculation of atom mapping rules, the prediction of Gibbs reaction energies and the prediction of subcellular protein localization are described first. Relevant data obtained using these methods was integrated into a graph theory-based approach to metabolic pathways, presented subsequently. The performance of the whole approach was evaluated by the search for experimentally verified biotransformation routes in the genome-scale metabolic networks of *E. coli* and *A. thaliana*. In the last part of this chapter a brief overview of an implemented web interface for exploring genome-scale metabolic networks is given.

4.1 Calculation of Atom Mapping Rules

4.1.1 Introduction

Given a chemical reaction, an atom mapping rule describes which educt atom is transferred to which product atom. Fig. 4.1A shows an atom mapping rules using serine-pyruvate transaminase (EC 2.6.1.51) as an example. In this reaction, the whole carbon skeleton of serine is transferred to that of hydroxypyruvate and that of pyruvate to alanine. Furthermore, the amino-group of serine is mapped to alanine and the keto-group of pyruvate to hydroxypyru-

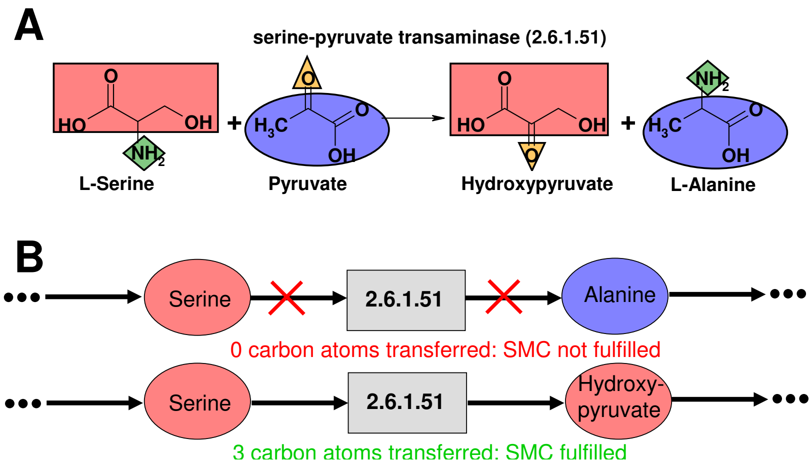

A
serine-pyruvate transaminase (2.6.1.51)
L-Serine
Pyruvate
Hydroxypyruvate
L-Alanine
B
Serine
2.6.1.51
Alanine
0 carbon atoms transferred: SMC not fulfilled
Serine
2.6.1.51
Hydroxy-pyruvate
3 carbon atoms transferred: SMC fulfilled

Figure 4.1: (A) An atom mapping rule using the reaction catalyzed by serine-pyruvate transaminase as an example is shown. The atom transfer between both sides of the reaction is represented using equal geometric shapes. (B) The concept of path validation based on the structural moiety constraint (SMC) is demonstrated. Three carbon atoms are transferred from serine to hydroxypyruvate and none to alanine.

vate. Within the scope of this thesis, we applied atom mapping rules for the validation of candidate pathways. Arita [2003] originally used atom mapping rules for this purpose and introduced the concept of the structural moiety constraint. According to this constraint, a pathway can only be biochemically feasible if at least one atom is transferred from the source to the product metabolite. For a given pathway this information is gained by tracing atoms through the pathway using atom mapping rules. For example, if the carbon skeleton of serine is reached only by a sequence of reaction steps during a path-finding algorithm, then it is clear that using EC 2.6.1.51 as subsequent reaction, the next pathway intermediate must be hydroxypyruvate and not alanine (shown in Fig. 4.1B). Using alanine as intermediate carbon carrier would violate the structural moiety constraint. The development of an improved approach for the fully automatic calculation of atom mapping rules is the topic of this section.

Representing the compounds of a chemical reaction as molecular graphs [1], atom mapping rules can be calculated using graph partition and graph isomorphism [Akutsu, 2004]. The underlying idea is that normally in chemical reactions only very few bonds are broken in order to transform the educts into

[1] where nodes of the graph represent atoms (ignoring hydrogen atoms) and edges stand for bonds in the original molecule

the products. Hence, we can find the mapping rules by removing a limited number of edges in the molecular graphs of the compounds and searching for graph isomorphisms between the remaining connected components. A valid atom mapping contains an isomorphic component of the product side for each connected component of the educt side and vice versa. However, the result of such a search, as presented in a previous work [Akutsu, 2004], is not necessarily unique and may contain biochemically meaningless mappings alongside the correct one. We were able to solve this problem by introducing the EC clustering approach. Using this approach, it is possible to detect the relevant mappings by clustering all mappings of those enzymatic reactions which have the first three digits of their EC number in common. The underlying idea is that only the first three digits describe the underlying reaction mechanism, and the last digit only enumerates the different chemical structures. This allows to select the atom mapping rule which best describes the reaction mechanism of the EC cluster or appears mostly in all reactions of the cluster.

The next section briefly describes the theoretical framework of the approach as introduced earlier [Akutsu, 2004] followed by the details of our practical algorithm for mapping calculation. The following section explains the EC clustering approach for filtering out irrelevant mappings. The results, when applying the approach to reactions extracted from KEGG as well as EcoCyc, are presented subsequently. After this, a brief discussion follows.

4.1.2 Problem definition and practical algorithm

Definition: A *chemical cut* [Akutsu, 2004] of size C is a partition of a graph G into connected components which are obtained by removing at most C edges whereas the nodes of each removed edge have to belong to different connected components after the removal.

In order to handle reactions modifying ring structures, we must extend the definition of a cut. A *pseudo cut* removes edges of a graph G which do not disconnect G. The total number of removed edges per compound may still not be larger than C. An example describing both types of cuts is shown in Fig. 4.2A.

Definition: Given the chemical reaction equation $E_1 + ... + E_e \leftrightarrow P_1 + ... + P_p$. $E_1, ..., E_e$ and $P_1, ..., P_p$ are molecular graphs representing educt and product compounds. The *mapping problem* is now to find a chemical cut of size C for each $E_1, ..., E_e$ and $P_1, ..., P_p$ such that the resulting multiset of connected components $\hat{E}_1 \cup ... \cup \hat{E}_e$ is equal to the multiset of connected components $\hat{P}_1 \cup ... \cup \hat{P}_p$. Elements of the multisets are equal if they are isomorphic.

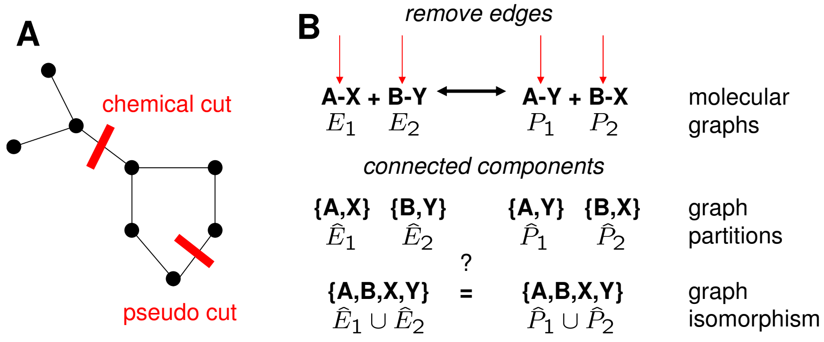


Figure 4.2: (A) Schematic illustration of chemical cuts and pseudo cuts. (B) The general mapping problem. The example shows a reaction with two educt (E_1, E_2) and two product compounds (P_1, P_2), and a cut-size C=1. Graph partitions ($\hat{E}_1$,$\hat{E}_2$,$\hat{P}_1$,$\hat{P}_2$) were created by removing at most one edge in the molecular graph for each compound. A mapping is found if the multisets $\hat{E}_1 \cup \hat{E}_2$ and $\hat{P}_1 \cup \hat{P}_2$ are equal.

Fig. 4.2B illustrates the mapping problem, for a simple example. For fixed values of p, q and C, the problem can be solved in polynomial time, since the number of combinations $(E_1, ..., E_e, P_1, ..., P_p)$ is $\mathcal{O}(n^{C(e+p)})$, where n is the maximum size of a compound in the reaction [Akutsu, 2004]. Practical algorithms solving the problem for the special case of $C = 1$ and $e = p = 2$ were presented earlier [Akutsu, 2004]. Here, we introduce a procedure for solving the general problem.

We distinguish two types of mapping rules. Given a chemical reaction, a fragment mapping rule defines which connected component (called fragment) of an educt molecular graph is isomorphic to which connected component of a product molecular graph. Such a rule consists of a list of isomorphic fragment pairs. An atom mapping rule defines which atom of an educt compound is transferred to which atom of a product compound. A rule of this type consists of a list of atom pairs. From the fragment mapping rules, we can deduce atom mapping rules using the canonical graph representations created by Morgan's algorithm [Morgan, 1965]. [2] We use unique SMILES [Weininger *et al.*, 1989] to detect isomorphic components. The advantage is that this permits the simple incorporation of stereochemical information and reduces the number of inferred irrelevant atom mapping rules. Furthermore, we define two functions which are necessary for the mapping calculation. The first function, $CSF(X)$, transforms the multiset X, which contains con-

[2] The algorithm assigns an unique integer label to each node in a molecular graph, based on the node degree and the degrees of its neighbors. Topologically equivalent nodes in isomorphic graphs get the same labels.

nected components as elements, to the multiset Y where the elements of X are replaced by their chemical formulas. Accordingly, the second function, $SMILES(X)$ replaces the elements of X by their unique SMILES.

Minimum cut algorithm: All valid atom mapping rules corresponding to a minimal cut size C can be computed as follows:

1. $C \leftarrow 0$
2. For the molecular graphs of the educts E_1,...,E_e and products P_1,...,P_p create all possible partitions $\hat{E}_{1_i}$,...,$\hat{E}_{e_j}$ and $\hat{P}_{1_k}$,...,$\hat{P}_{p_l}$ using cut size C.
3. Create all possible multisets of connected components $\tilde{E}_s = \hat{E}_{1_i} \cup ... \cup \hat{E}_{e_j}$ and $\tilde{P}_r = \hat{P}_{1_k} \cup ... \cup \hat{P}_{p_l}$.
4. Select all pairs ($\tilde{E}_s$,$\tilde{P}_r$) with $CSF(\tilde{E}_s) = CSF(\tilde{P}_r)$.
5. From all pairs calculated in Step 3 select all pairs ($\tilde{E}_s$,$\tilde{P}_r$) with $SMILES(\tilde{E}_s) = SMILES(\tilde{P}_r)$ and a minimum number of removed edges producing pseudocuts accumulated for all educts and products. Each pair represents a fragment mapping rule.
6. If no fragment mapping rule is found in Step 4: $C \leftarrow C + 1$, repeat from Step 2.
7. Extract the final atom mapping rules from the fragment mapping rules using the canonical graph representation calculated by Morgan's algorithm.

The third step was introduced to improve the calculation time significantly. It is not necessary to compute unique SMILES for all partitions. In the first iteration of the algorithm we simply compute the chemical formulas of the connected components and use them to collect a set of candidate partitions for the molecular graphs. Step 4 insures that the mappings found are based on a minimum number of removed edges. If we would search for all mappings allowing the maximum possible cut size C as well as the maximum number of edges producing pseudo cuts, the number of irrelevant mappings per reaction would be much higher. Note that a mapping found by the cut size $C = 0$ typically represents isomerization or oxidoreductive reactions.

4.1.3 EC clustering

For a significant number of reactions (approximately 40%, data not shown), there is more than one possible mapping rule. An example is shown in

Fig. 4.3A. Using cut size $C = 1$, there are three possible mapping rules for the reaction catalyzed by serine-pyruvate transaminase (EC 2.6.1.51). But only the first mapping rule describes the underlying reaction mechanism which exchanges the amino group of L-serine with a keto group of pyruvate. To filter out biochemically irrelevant mappings, we introduce the EC clustering approach. The idea is that the mechanism of many chemical reactions consists of shifting or exchanging small functional groups like amino, keto, methyl, phosphate or carboxyl groups. All reactions which have the first three digits of their EC number in common also share the reaction mechanism. The last digit only enumerates the different chemical structures operating as substrates. Typical examples are reactions transferring a phosphate (EC 2.7.1.-) or a methyl group (EC 2.1.1.-) from one molecule to another.

At first, we define an *EC cluster* (ECC) as a set of enzymatic reactions which have the first three digits of their EC number in common. Given an EC cluster, a reaction mechanism rule generally describes, for the reactions in the cluster, how the educts are transformed into the products. The aim is then to automatically infer the reaction mechanism rule by identifying the relevant functional groups or parts of the substrates. The next step is to select that fragment mapping rule and underlying atom mapping rule which correspond to the inferred reaction mechanism rule and to discard all the other fragment mapping rules.

Reaction mechanism rules are represented as strings and constructed from fragment mapping rules. The following syntax is used to describe them. The two sides of a reaction are separated by '='. The fragments of each compound are separated by ',' and enclosed by '<' and '>'. The first fragment representing a non-relevant structure, is designated with '$X1$', the second with '$X2$' and so on. Relevant fragments like the mentioned functional groups are represented using their SMILES (e.g. `N`, `O`, `C`, `OP(O)O`, `C(O)O`). [3] An empty fragment is represented by '$' and is used in graph partitions for compounds in which no edge is removed. The strings representing both the fragments and the whole reaction sides are alphabetically ordered to ensure uniqueness in the comparison with reaction mechanism rules from different reactions. Fig. 4.3B shows an reaction mechanism rule for each fragment mapping rule shown in Fig. 4.3A.

Note that there is no predefined list of relevant fragments. We generate all possible reaction mechanism rules from the fragment mapping rules of a given reaction by allowing each fragment to be relevant or not. Given an EC cluster and a reaction mechanism rule, the occurrence frequency of this

[3] Note that the SMILES shown lack double bonds since bond types (parallel edges) are ignored in our molecular graphs for simplicity.

rule accumulated over all reactions in the cluster is called the EC cluster score (ECCS). A reaction mechanism rule occurs in a reaction if it can be constructed from at least one fragment mapping rule of the reaction. From all generated reaction mechanism rules we select that to be relevant which has the highest score. The EC clustering procedure performs the following steps:

1. For each given fragment mapping rule containing n educt as well as product fragments, construct for all $\binom{n}{k}$ combinations with $k = 0, ..., n-1$, reaction mechanism rules in which k fragments are marked as non-relevant (represented as '$X1$', '$X2$', and so on).

2. For all reaction mechanism rules deduced from a fragment mapping rule of a reaction in an EC cluster, calculate the EC cluster scores.

3. Assign each fragment mapping rule of a reaction in an EC cluster the maximum ECCS of the reaction mechanism rules which were constructed from the fragment mapping rule.

4. For each reaction select the fragment mapping rule (and its corresponding atom mapping rule) with the highest score as the relevant mapping.

Considering the example shown in Fig. 4.3, it becomes possible to detect the first mapping rule as biochemically relevant, since the assigned score is significantly larger than the scores of the other two mapping rules. The score of 0.96 for the first reaction mechanism rule indicates that for 96% of the reactions in the EC cluster 2.6.1.- (overall 90 reactions using data from KEGG), the mechanism can be described as exchange of an amino group with a keto group. If there is more than one fragment mapping with the highest score or there is a reaction with no EC number, then we select the mapping as relevant with the minimum number of transferred atoms (the number of atoms of the relevant chemical groups).

4.1.4 Results

Atom mapping rules were inferred from chemical reactions extracted from the KEGG and the EcoCyc databases. The maximum cut-size was restricted to $C = 2$ and the maximum number of compounds permitted per reaction was set to 10. This ensured an efficient calculation. Reactions containing compounds for which the structural information was incomplete or non-existent, and reactions with an unbalanced reaction equation were not considered. This reduced the number of reactions from 6811 to 4621 for KEGG, and

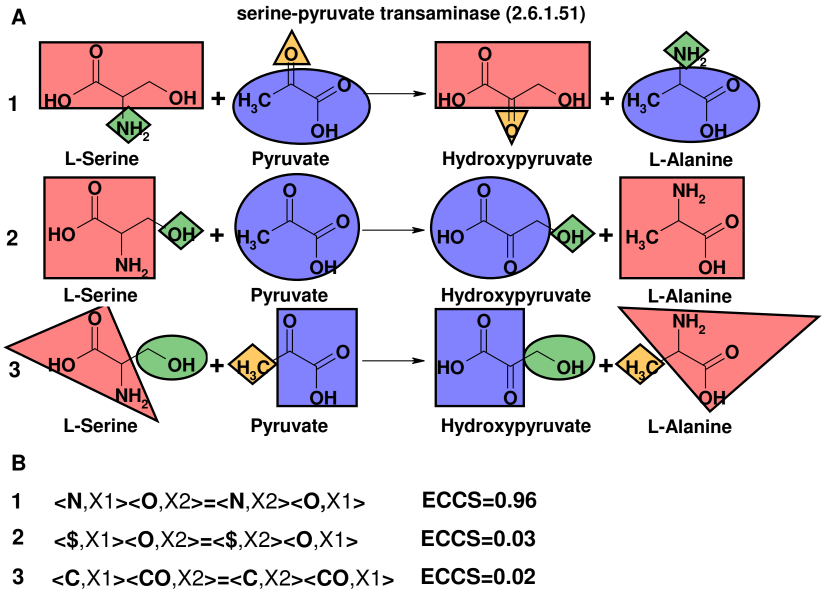

A
serine-pyruvate transaminase (2.6.1.51)
1
2
3
L-Serine
Pyruvate
Hydroxypyruvate
L-Alanine
HO
OH
O
NH_2
H_3C
+
B
1 <N,X1><O,X2>=<N,X2><O,X1> ECCS=0.96
2 <$,X1><O,X2>=<$,X2><O,X1> ECCS=0.03
3 <C,X1><CO,X2>=<C,X2><CO,X1> ECCS=0.02

Figure 4.3: (A) A reaction with multiple mapping rules. The atom transfer between both sides of the reaction is represented by equal geometric shapes. The different shapes within a compound also represent the connected components in the corresponding molecular graphs. Only the first rule is biochemically relevant. (B) Each mapping rule is assigned the maximum score (ECCS) of all reaction mechanism rules which were derived from the mapping. The mapping with the highest score is detected as the relevant mapping. For each mapping rule, the best reaction mechanism rule with corresponding score is shown.

Table 4.1: The results of the atom mapping calculation using the EcoCyc and KEGG data sets. 850 as well as 4621 reactions, with balanced equations and complete structural information of the compounds, were selected from EcoCyc and KEGG. For 98.0% as well as 97.7% of these reactions, at least one atom mapping rule could be calculated. More details are described in the text.

		EcoCyc	KEGG
reactions	overall	1348	6811
	selected	850 (63.1%)	4621 (67.9%)
	successful	833 (98.0%)	4516 (97.7%)
mappings	overall	1236	5913
	per reaction	1.51	1.31
cut size	$C = 0$	197 (24.0%)	807 (17.8%)
	$C = 1$	553 (67.4%)	3272 (72.5%)
	$C = 2$	71 (8.6%)	437 (9.7%)

from 1348 to 850 for EcoCyc. Tab. 4.1 summarizes the results of the calculation. For 833 (98%) of the reactions selected from EcoCyc and 4516 (97.7%) from KEGG, at least one atom mapping rule was found. The overall number of mappings per reaction was 1.51 (EcoCyc) as well as 1.31 (KEGG). The number of reactions with mapping rules using cut size $C = 0$ was 197 (23.6%) for EcoCyc and 807 (17.8%) for KEGG. These are typically stereoisomerization or oxidoreductive reactions in which the transfer of substructures between molecules was not necessary (e.g. EC 1.1.1.-). The majority of the reactions - 563 (67.6%) for EcoCyc and 3272 (72.5%) for KEGG - required atom mapping rules with cut size $C = 1$. Typical representatives are reactions transferring phosphate or methyl groups (e.g. EC 2.7.1.- or EC 2.1.1.-). Seventy-three (8.8%) of the EcoCyc and 437 (9.7%) of the KEGG reactions required atom mapping rules with the cut size $C = 2$. Examples are reactions belonging to EC 1.13.11.- in which two oxygen atoms, originating from molecular oxygen, are transferred. We manually inspected 17 reactions (2%) from EcoCyc and 105 (2.3%) from KEGG for which no atom mapping rule could be inferred. These reactions require mapping rules with a cut size greater than $C = 2$. The hydrolysis of allophanate resulting in two carbon dioxide molecules and two ammonia molecules (EC 3.5.1.54) is an example of a reaction requiring cut size $C = 3$. Another example is the uroporphyrinogen carboxy-lyase reaction (EC 4.1.1.37), in which four molecules of carbon dioxide are cleaved off from uroporphyrinogen ($C = 4$).

4.1.5 Discussion

A novel approach for inferring atom mapping rules from chemical reactions was developed. Fully automated and efficient calculation was the main target and was achieved by introducing pseudo cuts, the use of unique SMILES and EC clustering. The purpose of the EC clustering is to filter out biochemically irrelevant atom mapping rules but it also offers a way to extract the underlying mechanism of enzymatic reactions and, therefore, could also be used as a starting point for developing methods suited to large-scale classifications of reactions as well as automatic assignment of EC numbers. In addition to biochemical feasibility validation of candidate pathways inferred by path-finding approaches, calculated atom mapping rules can also be used for analyzing radioisotope tracer experiments, for consistency checking of pathway databases or visualizing conserved structural moieties along pathways.

We restricted the calculation of atom mapping rules to a maximum cut-size $C = 2$ to ensure an efficient calculation. Furthermore, reactions requiring a higher cut-size are very rare and it is not necessary to have calculated atom mapping rules for 100% of the reactions in EcoCyc and KEGG to support the main goal of this thesis, the inference of relevant biotransformaton routes (as described in Section 4.4.2). However, more work should be invested to ensure more efficient calculation even for very complex reactions. The main challenge is how to deal in general with larger compounds like NADH and acetyl-CoA or even larger ones like protoheme. Such compounds cause a high number of bond-breaking combinations that have to be considered. It is, however, obvious that there are bonds which have a higher breaking probability than others. The introduction of a suited chemical logic that helps identifying these bonds could significantly reduce the calculation effort.

4.2 Prediction of standard transformed Gibbs energies of biochemical reactions

4.2.1 Introduction

The Gibbs energy G represents the driving force for each biochemical reaction in a metabolic network. The standard Gibbs energy change $\Delta_r G^0$ of a reaction is related to the equilibrium constant K by

$$\Delta_r G^0 = -RT \ln K$$

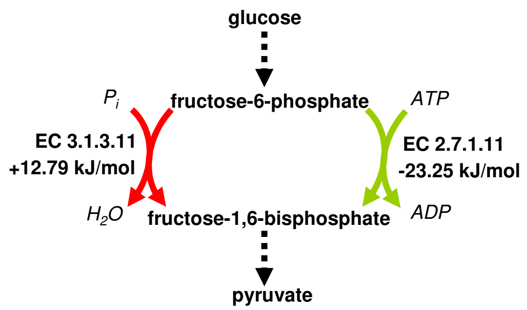


Figure 4.4: The chemical conversion of fructose-6-phosphate into fructose-1,6-bisphosphate using EC 3.1.3.11 and EC 2.7.1.11. $\Delta_r G'^0$ values (in kJ/mol) in the desired reaction direction, taken from Alberty [2005b], are also shown.

with gas constant R and absolute temperature T. Knowledge of $\Delta_r G^0$ as well as of K for each reaction step of a metabolic pathway supports thermodynamic pathway analysis. Whether a (novel or engineered) pathway is thermodynamically feasible and where to find bottlenecks and physiologically irreversible reactions are interesting questions. Their answers help us to understand cellular metabolism better. An important aim of this thesis was to consider and integrate thermodynamic information when searching for relevant pathways using a graph theory-based approach. The basic idea behind using Gibbs energy data for this purpose was the assumption and observation that pathways tend to use the reaction that is thermodynamically most favorable when several alternatives exist. For example, the chemical conversion of glucose to pyruvate in glycolysis requires the phosphorylation of fructose-6-phosphate to fructose-1,6-bisphosphate as an intermediate step. Ignoring regulatory aspects, the pathway could choose between two different reactions (see also Fig. 4.4). However, compared to the first reaction (EC 3.1.3.11), the second one (EC 2.7.1.11) is thermodynamically much more favorable in the desired direction under standard conditions. Therefore, it does not come as a surprise that the second reaction is known to be part of the textbook glycolysis pathway and the first reaction part of the opposed gluconeogenesis pathway, which transforms pyruvate back to glucose.

Without a given $\Delta_r G^0$ or K it is difficult to estimate even the favored direction of a reaction. Experimentally determined equilibrium constants are available only for a limited number of biochemical reactions. It is also possible to calculate $\Delta_r G^0$ of a reaction (if the standard Gibbs energies of formation $\Delta_f G^0$ of the educts and products are known) using the equation

$$\Delta_r G^0 = \sum_j \Delta_f G_j^0(p_j) - \sum_i \Delta_f G_i^0(e_i)$$

where e_i and p_j are the stoichiometric coefficients of the educts and products. Unfortunately, the availability of experimentally determined $\Delta_f G^0$ values of biochemical compounds is also limited. Based on data sources [Goldberg *et al.*, 2004; Alberty, 2005] that provide comprehensive thermodynamic information collected from the literature, we can annotate less than ten percent of all biochemical reactions stored in databases like KEGG or MetaCyc with $\Delta_r G^0$ or K values. With the rapidly increasing number of genome-scale metabolic networks stored in pathway genome databases, thermodynamic pathway analysis is becoming more important but is hindered by the lack of comprehensive information about equilibrium constants. Hence, there is a need for computational approaches for estimating or predicting Gibbs energy information given the educts and products of reactions with unknown equilibrium constants.

A group contribution method [Mavrovouniotis, 1990, 1991] has been developed for estimating $\Delta_r G^0$ of biochemical reactions in aqueous solution. To use this method, the chemical structures of the educts and products of a reaction have to be decomposed into functional groups of atoms. The basic idea is the assumption that the $\Delta_f G^0$ of a molecule is given by the linear combination of energy contributions from each constituent group multiplied by the number of occurrences of that group in the molecule. To this end, a predefined set of groups is provided by the authors and each group is assigned an energy contribution. Then $\Delta_r G^0$ of a reaction is equal to the difference between the sums of the group contributions of the products and educts. The contributions are estimated using multiple linear regression on data collected from the literature. The data set consists of a mixture of Gibbs energies of biochemical compounds and reactions in dilute aqueous solution at 298.15 K and pH 7. Reaction data are also used because $\Delta_r G^0$ of a reaction is given by the linear combination of net energy contributions from the educt and product groups. The typical error of an estimated $\Delta_f G^0$ is less than 2 kcal/mol (8.37 kJ/mol) but errors higher than 5 kcal/mol (20.92 kJ/mol) can occur.

Other works, not specialized for biochemical compounds in aqueous solution but related to the problem, are based on quantitative structure-property relationship (QSPR) techniques to predict the standard Gibbs energy of formation of organic compounds [Ivanciuc *et al.*, 2000, 2001; Toropov and Toropova, 2003; Yan, 2006]. The basic underlying idea is the fact that physiochemical properties like $\Delta_f G^0$ are determined by the chemical structure of the molecules. Molecular descriptors are used to encode the structures in a numerical form and linear or non-linear statistical methods (like multiple linear regression or neural networks) are used to model the complex relationship between $\Delta_f G^0$ and the selected descriptors.

A drawback of the group contribution method described is that it ignores effects on the thermodynamic equilibrium caused by the ionic strength I, the presence of metal ions like Mg^{2+} and the dissociation of biochemical compounds into several ionic species in aqueous solutions at pH 7. These effects can be significant and their consideration requires the adjustment and transformation of the Gibbs energy. Therefore, the IUPAC-IUBMB Joint Commission on Biochemical Nomenclature[Alberty, 1996] recommends the use of the apparent equilibrium constant K', which is written in terms of sums of species together with the standard transformed Gibbs energies $\Delta_f G'^0$ and $\Delta_r G'^0$ instead of K, $\Delta_f G^0$ and $\Delta_r G^0$ when analyzing biochemical reaction systems. These thermodynamic quantities should be based on biochemical standard conditions[4]. The importance of considering these parameters when analyzing metabolic pathways was also evaluated in an extensive study [Maskow and von Stockar, 2005]. Further drawbacks of using the group contribution method are that possible group interactions are neglected and that a couple of special correction rules have to be applied in order to get better estimations. Also, the decomposition of the chemical structures into non-overlapping groups of atoms is a non-trivial task. However, it should be noted that Forsythe *et al.* [1997] introduced an algorithm that complements the group contribution method by an automatic decomposition based on the application of SMILES [Weininger *et al.*, 1989]. Furthermore, special care has to be taken when dealing with pool compounds like ATP, ADP or NADH. These compounds have to be treated as single groups and are assigned special energy contributions when occurring in a reaction. Although the group contribution method represents a pioneering work in this field, the problems described and their drawbacks makes the estimation of equilibrium constants quite difficult.

Our contribution to the problem of estimating equilibrium constants was the development of a method that is easier to use, considers the recommendations described for analyzing biochemical reactions and provides acceptable predictions. To this end, we applied QSPR techniques, which allowed us to calculate a wide range of molecular descriptors fast and easy using a QSPR software package. Novel to our approach, but inspired by the group contribution method, is that biochemical reactions are represented as feature vectors created from the difference of the numerical molecular descriptor vectors between the products and educts of each reaction. Experimentally determined $\Delta_r G'^0$ values under approximate biochemical standard conditions for 484 reactions were collected using data extracted from relevant data sources [Goldberg *et al.*, 2004; Alberty, 2005]. Since we were primarily interested in

[4] T=298.15 K (or T=310.15 K), $P = 10^5$ Pa, pH 7, I=0.25 mol/l and pMg 3

the creation of a prediction model specialized for biochemical reactions involved in carbon, nitrogen, sulfur and phosphor metabolism, we disregarded reactions with compounds that contained atoms other than H, O, C, N, S or P. We used multiple linear regression and stepwise feature selection to calculate a model for the prediction of $\Delta_r G'^0$ for reactions given in the form of feature vectors. Based on an independent test procedure, the prediction error obtained for a typical reaction was 6.24 kJ/mol with a squared correlation coefficient of 0.9373 between the observed and predicted $\Delta_r G'^0$ values of the test reactions.

The following sections present the methods necessary to develop the approach including the data sources used, the creation of the training data set as well as feature vectors and the procedures applied for training and performance evaluation, followed by detailed prediction results with a concluding discussion.

4.2.2 Methods

TECRDB

The Thermodynamics of Enzyme-catalyzed Reactions Database (TECRDB) [Goldberg *et al.*, 2004] is a systematic collection of thermodynamic data on enzyme-catalyzed reactions. The data contains apparent equilibrium constants K' and molar enthalpies $\Delta_r H'^0$ of biochemical reactions measured in experimental studies. The database is available via a web interface and stores data for approximately 400 different enzyme-catalyzed reactions curated from approximately 1,000 published papers. The collected papers were also previously surveyed in six reviews [Goldberg *et al.*, 1993; Goldberg and Tewari, 1994,b, 1995,b, 1999].

For each entry (measured K' or $\Delta_r H'^0$) in the database the following information, if found in a paper, is given:

- literature reference
- the enzyme-catalyzed reaction written in terms of reactants (sum of species)
- Enzyme Commission (EC) number of the reaction
- the method of measurement
- the conditions of measurement (temperature, pH, ionic strength, buffer, cofactor(s) etc.)

- a subjective evaluation rating of the data

The subjective evaluation rating separates the data into four classes of quality (A for high, B for good, C for average and D for low quality). To carry out, the authors of the database considered the level of experimental details described in the corresponding study.

BasicBiochemData3

BasicBiochemData3 [Alberty, 2005] is a database written in Mathematica [Wolfram Research, Inc, 2005] that contains the standard Gibbs energy of formation ($\Delta_f G^0$) of species for 199 reactants of biochemical interest at 298.15 K and zero ionic strength. The standard enthalpies of formation ($\Delta_f H^0$) are also available for the species of 94 reactants. Furthermore, the database provides numerous programs for the calculation of the apparent equilibrium constant $K^{'}$ and other transformed thermodynamical properties of enzyme-catalyzed reactions.

Some of the collected species data stems directly from the NBS [Wagmann *et al.*, 1982] and CODATA [Cox *et al.*, 1989] thermodynamic tables. The thermodynamic properties are calculated from measurements of apparent equilibrium constants extracted from TECRDB, especially for larger biochemical compounds (e.g. acetyl-CoA).

Species data for 28 more reactants not included in the last version of BasicBiochemData3, but described in the literature [Alberty, 2006a,b, 2007], was also used in this work. These reactants represent the GTP, XTP, TTP, UTP, CTP and carbamoyl-phosphate series.

The species data and Mathematica programs provided can be used to calculate the standard transformed Gibbs energy of formation ($\Delta_f G^{'0}$) of the reactants at 298.15 K in the pH range from five to nine and ionic strength from zero to 0.35 mol/l.

Mining standard transformed Gibbs energies from experimental data

The standard transformed Gibbs energy of formation of a reactant at a specified temperature, pH and ionic strength can be calculated from the standard Gibbs energies of formation of the species involved in that reactant using Legendre transforms (described in Section 2.2.2). However, information about standard Gibbs energies of formation is available only for a limited number of biochemical species. Another way of obtaining $\Delta_f G^{'0}$ values for biochemical reactants without knowing $\Delta_f G^0$ values of its species was described by Alberty [1998]. This method can be applied if the standard transformed Gibbs energies of formation are given for all but one reactant in a biochemical reaction with experimentally determined $K^{'}$ (close to T=298.15 K and pH 7) simply by using the following equation:

$$\Delta_r G^{'0} = -RT \ln K^{'} = \sum_j (\Delta_f G_j^{'0} p_j) - \sum_i (\Delta_f G_i^{'0} e_i)$$

where e_i and p_j are the stoichiometric coefficients of the educts and products in the biochemical reaction. If there are two reactants A and B (one for each reaction side) with unknown standard transformed Gibbs energy, one can be assigned $\Delta_f G^0 = 0$ or $\Delta_f G^{'0} = 0$ by convention. The advantage is that this allows the calculation of $K^{'}$ for reactions where both reactants participate. However, it is not possible to calculate $K^{'}$ for reactions forming reactant A or B. Some of the thermodynamic properties present in the BasicBiochemData3 database are calculated according to this method from experimental data extracted from the database provided by Goldberg *et al.* [2004].

In order to increase the number of $\Delta_f G^{'0}$ values available, we performed an automated version of Alberty's method by combining BasicBiochemData3 with TECRDB. First, compounds and reactions from TECRDB were mapped to their corresponding entities in the KEGG and MetaCyc databases by comparing compound names and EC numbers. This had to be done because there is no structural information about the compounds in TECRDB. The information was required later for creating a QSPR training data set. The mapping candidates were detected computationally (by matching compound names) but selected manually to avoid false positives. We mapped TECRDB compounds only if the corresponding KEGG/MetaCyc compounds were annotated with complete structural information and only contained carbon, oxygen, hydrogen, nitrogen, sulfur or phosphorous atoms. All reactants in BasicBiochemData3 were also mapped to KEGG and MetaCyc and were, therefore, available in TECRDB. Then all valid TECRDB entries with measured $K^{'}$ were extracted. An entry was defined as valid if temperature T and pH values of the experiment were given. The information about ionic strength and cofactors was also extracted. Furthermore, all compounds occurring in the reaction equation had to be mapped to KEGG/MetaCyc. The EC number, literature reference and evaluation rating (ER) were always available. Based on these prepared TECRDB entries, the following iterative procedure was performed to automatically estimate $\Delta_f G^{'0}$ and $\Delta_r G^{'0}$ values under near biochemical standard conditions:

1. Calculation of the $\Delta_f G^{'0}$ for each reactant in BasicBiochemData3 from the $\Delta_f G^0$ of its species using Legendre transforms and adaptation to

ionic strength I=0.25 mol/l. These reactants and their $\Delta_f G^{'0}$ values form an initial list $L_{\Delta_f G^{'0}}$.

2. Setting the evaluation rate constraint variable: $\mathrm{ER}_{used} = \mathrm{A}$

3. Setting the experimental constraint variables:
   $T_{\mathrm{min}} = 298.15$ K; $T_{\mathrm{max}} = 298.15$ K; $\mathrm{pH}_{\mathrm{min}} = 7.0$; $\mathrm{pH}_{\mathrm{max}} = 7.0$;

4. Selection of all entries ($K^{'}$, T, pH, ER) which fulfill the current constraints:
   $T_{\mathrm{min}} \leq T \leq T_{\mathrm{max}}$; $\mathrm{pH}_{\mathrm{min}} \leq \mathrm{pH} \leq \mathrm{pH}_{\mathrm{max}}$; $\mathrm{ER}_{\mathrm{used}} = \mathrm{ER}$

5. For each reaction in the selected entries, $\Delta_r G^{'0}$ is calculated using equation $\Delta_r G^{'0} = -RT \ln K^{'}$. If more than one TECRDB entry is selected for a reaction, its $\Delta_r G^{'0}$ values are averaged. The reactions and their $\Delta_r G^{'0}$ values are added to the list $L_{\Delta_r G^{'0}}$. Once a reaction is added to this list, its $\Delta_r G^{'0}$ value cannot be overwritten or changed in a succeeding iteration.

6. For each reaction in $L_{\Delta_r G^{'0}}$ with exactly one reactant that is not found in $L_{\Delta_f G^{'0}}$, $\Delta_f G^{'0}$ of that reactant is calculated and added to $L_{\Delta_f G^{'0}}$ using the equation $\Delta_r G^{'0} = \sum_i v_i \Delta_{f_i} G^{'0}$ where v_i are the stoichiometric coefficients (negative for educts and positive for products). If $\Delta_f G^{'0}$ of a reactant can be calculated from multiple reactions, the average value is used. Once a reactant is added to the list, its $\Delta_f G^{'0}$ cannot be overwritten or changed in a succeeding iteration.

7. Repeat step 6 until no new reactant can be added to $L_{\Delta_f G^{'0}}$.

8. Increment/decrement experimental constraint variables:
   $T_{\mathrm{min}} = T_{\mathrm{min}} - 1$; $T_{\mathrm{max}} = T_{\mathrm{max}} + 1$
   $\mathrm{pH}_{\mathrm{min}} = \mathrm{pH}_{\mathrm{min}} - 0.2$; $\mathrm{pH}_{\mathrm{max}} = \mathrm{pH}_{\mathrm{max}} + 0.2$

9. If $T_{\mathrm{min}} \geq 293.15$ and $T_{\mathrm{max}} \leq 303.15$ and $\mathrm{pH}_{\mathrm{min}} \geq 6.0$ and $\mathrm{pH}_{\mathrm{max}} \leq 8.0$ continue with step 4.

10. Reduction of the evaluation rate constraint variable:
    If $\mathrm{ER}_{\mathrm{used}}$ is set to A then set $\mathrm{ER}_{\mathrm{used}} = \mathrm{B}$ and continue with step 3
    If $\mathrm{ER}_{\mathrm{used}}$ is set to B then set $\mathrm{ER}_{\mathrm{used}} = \mathrm{C}$ and continue with step 3.

After this procedure, the lists $L_{\Delta_r G^{'0}}$ and $L_{\Delta_f G^{'0}}$ contained additional thermodynamic data which could be used to create a comprehensive QSPR training data set.

DRAGON molecular descriptors

The DRAGON software package [Talete srl, 2007] was used to calculate molecular descriptors. It was possible to calculate up to 3,324 descriptors including functional group and fragment counts as well as topological, geometrical and molecular properties. We applied the package to chemical structures extracted from the KEGG and MetaCyc databases. The calculation was performed with the inclusion of hydrogen atoms and for 2D descriptors (2,425 overall) only.

Training data set and feature vector representation

The QSPR training data set consists of biochemical reactions with known $\Delta_r G^{'0}$. To this end, we extracted all distinct reactions from KEGG and MetaCyc that contained only compounds with known $\Delta_f G^{'0}$ and balanced educt/product atoms. We found 411 reactions. For each reaction, we calculated $\Delta_r G^{'0}$ using equation $\Delta_r G^{'0} = \sum_i v_i \Delta_{f_i} G^{'0}$. Furthermore, we included all reactions extracted from TECRDB with known $\Delta_r G^{'0}$ and with more than one compound for which the $\Delta_f G^{'0}$ value was missing. We found 73 reactions, which increased the training data set to 484 reactions overall.

The reaction equations were transformed into a feature vector representation. To this end, we computed DRAGON features (molecular descriptors) for all compounds participating in the reactions. Then for each reaction the difference between its educt and product feature vectors was computed using the equation

$$F_r = \sum_i v_i F_i$$

where F_r is the feature vector representation of a reaction r, F_i the DRAGON feature vectors of the educts and products and v_i the stoichiometric coefficients (negative for educts and positive for products).

The feature vector representation of reactions is inspired from the group contribution method that estimates $\Delta_r G^{'0}$ of a reaction by summing net energy contributions from the educt and product groups. The decision to develop a prediction method for $\Delta_r G^{'0}$ and not for $\Delta_f G^{'0}$ was based on several reasons. First, we have more comprehensive training data if we use $\Delta_r G^{'0}$ values. This allowed us to use an additional 73 reactions, each containing more than one compound with unknown $\Delta_f G^{'0}$, mined from TECRDB. Not all $\Delta_f G^{'0}$ values extracted from BasicBiochemData3 and other sources can be considered for a training data set. The reason is that the $\Delta_f G^{'0}$ of several reactants or the $\Delta_f G^0$ of one of its species is set to zero by convention. For example, the $\Delta_f G^{'0}$ values of the GTP series (GTP, GDP, GMP, guanosine, guanine) as calculated by Alberty [2006a] are based on the convention that $\Delta_f G^0$ of guanosine is zero. Otherwise it would not be possible to get $\Delta_f G'^0$ values for this series of reactants. In order to get correct absolute values for the GTP series, a certain amount of Gibbs energy remains to be specified and added to these relative $\Delta_f G'^0$ values. An advantage of relative values is that they can be used to calculate $\Delta_r G'^0$ of reactions like EC 2.7.1.30[5] or EC 3.2.2.1[6] that contain exactly one reactant of the series on the educt side and one on the product side. Furthermore, we believe that it is easier to predict $\Delta_r G'^0$ of reactions with balanced educt/product atom masses that normally undergo only slight molecular changes instead of using $\Delta_f G'^0$ of biochemical reactants which differ widely in their structural properties. For example, $\Delta_f G'^0$ of PRPP is -2978.51 kJ/mol and that of oxidized glutathione is 1219.74 kJ/mol. It is also more convenient to estimate the error of predicted $\Delta_r G'^0$ for a reaction if its $\Delta_r G'^0$ value can be directly inferred from a prediction system instead of using the predicted $\Delta_f G'^0$ of its educts and products.

Training and performance evaluation

The QSPR training was performed using multiple linear regression and stepwise feature selection starting with no pre-selected features. Minimizing the sum of the squared errors between observed and predicted values was the criterion for selecting the features using five-fold cross-validation.

Although other (non-linear) statistical learning approaches could be used, multiple linear regression was chosen because it allowed us to easily model two important properties of $\Delta_r G'^0$ that should be considered in a prediction method expressed as:

$$\Delta_r G'^0 = p(F_r)$$

where F_r is the feature vector of the query reaction and p the prediction function as a result of the model training. The two properties that should be supported by p can be defined as follows:

1. $p(F_r) = -p(-F_r)$ reverse reactions

2. $p(F_{r_1}) + p(F_{r_2}) = p(F_{r_1} + F_{r_2})$ reaction coupling

The first property describes the energetic behavior of reverse reactions, i.e. reversing the direction of a reaction with $\Delta_r G'^0$ will invert the Gibbs energy balance to $-\Delta_r G'^0$. Given a reaction represented by feature vector F_r, its

[5] GTP + glycerol $\rightleftharpoons$ GDP + glycerol 3-phosphate

[6] guanosine + H_2O $\rightleftharpoons$ guanine + ribose

reverse reaction is represented by the inverse feature vector $-F_r$ calculated from the difference between the educt/product molecular descriptor vectors. Hence, we can simply model the first property if coefficient b_0 of the multiple linear regression equation

$$y = b_0 + \sum_{i=1}^{m} b_i x_i + \epsilon$$

is constrained to zero ($b_0 = 0$). The second property represents the fact that changes in Gibbs energy of reactions are additive if the reactions are coupled (explained in Section 2.2.1). Given two coupled reactions represented by F_{r_1} and F_{r_2} with $\Delta_{r_1}G^{'0}$ and $\Delta_{r_2}G^{'0}$, the resulting overall reaction is represented by $\Delta_{r_1}G^{'0} + \Delta_{r_2}G^{'0}$ and $F_{r_1} + F_{r_2}$, which corresponds to the net molecular descriptors of the educts and products for both reactions.

An independent test procedure was applied to evaluate prediction performance. The test was performed by randomly selecting 50 reactions not to be used in the training phase. For each of the 50 test reactions, the standard transformed Gibbs energy was predicted from the regression model obtained using the remaining 434 reactions for training. The whole independent test procedure was performed ten times. However, there were two restrictions for selecting test reactions. Since we expected more data noise for the 73 reactions inferred from TECRDB using our data mining approach, we did not allow these reactions to be selected for testing. Furthermore, we did not select reactions for testing that contained some small and rarely occurring compounds with less than two carbon atoms except the frequently occurring compounds water, carbon dioxide, oxygen, hydrogen peroxide, ammonia, phosphate and pyrophosphate. All these reactions were used for training but not for testing, because we expected that these reactions could distort prediction performance. The results of the ten runs were merged into a unique independent set of test reactions to get more data for statistical evaluation. To this end, each reaction of the ten test runs was included only once. If a reaction appeared in more than one independent test run, we used the average value of the predicted Gibbs energies for this reaction. The resulting unique test set contained 285 reactions with predicted $\Delta_r G^{'0}$.

We used two standard statistical measures for estimating prediction performance. The first measure is the coefficient of determination R^2 which compares the variation between observed and predicted values to the variation within the observed values. In other words, R^2 is a measure of the quality of fit of a model and provides information about how well the predicted values approximate the observed real data values. The definition of

R^2 is given by

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \overline{y})^2}$$

where y_i are the observed values, $\hat{y}_i$ are the predicted values, $\overline{y}$ is the mean value of the observed values and N is the number of observed/predicted data value pairs. The standard error of estimate SEE is the second quality measure and it provides information about the expected accuracy of the model predictions and is calculated from the sum of the squared errors for each data value pair. The standard error of estimate is defined by

$$SSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}}.$$

Both quality measures were also used to evaluate the performance of the applied data mining approach, which automatically estimates $\Delta_f G'^0$ and $\Delta_r G'^0$ values from experimental data in TECRDB.

4.2.3 Results

Mining standard transformed Gibbs energies from experimental data

The result of estimating $\Delta_f G'^0$ for biochemical compounds from experimental data using the data mining approach, described in the methods section, is shown in Tab. 4.2. For each estimated $\Delta_f G'^0$, the table shows the relevant data for the automatically selected TECRDB entries, i.e., the EC number, temperature T, pH, ionic strength I, if given, measured equilibrium constant K' and the subjective evaluation rate EV. The approach produced $\Delta_f G'^0$ values for 31 new reactants which were not contained in BasicBiochemData3. To evaluate the accuracy and reliability of the whole approach, we performed the following experiment. For each reactant with given $\Delta_f G'^0$ inferred from its species data in BasicBiochemData3, we tried to estimate its $\Delta_f G'^0$ from experimental data using our data mining approach. We collected all successfully estimated $\Delta_f G'^0$ values (118 overall) and compared them with the corresponding (observed) values inferred from BasicBiochemData3. The result of this experiment is plotted in Fig. 4.5.

Table 4.2: Result of the mined $\Delta_f G'^0$ values for 31 reactants. For each $\Delta_f G'^0$, all relevant parameters of the automatically selected TECRDB entries are presented

reactant	$\Delta_f G'^0$	EC	T	pH	I/M	K'	$\Delta_r G'^0$	Ev
6-phospho-D-gluconate	-1568.69	1.1.1.44	298.15	6.9	0.18	0.079	6.29	A
phosphocreatine	-750.97	2.7.3.2	298.15	7.0	0.25	0.0058	12.78	A
(S)-2-methylmalate	-602.49	4.1.3.22	298.15	7.4	0.26	0.209	3.88	A
4-hydroxyphenylpyruvate	-183.38	2.6.1.5	298.15	7.5	0.32	0.88	0.32	A
phenylpyruvate	-20.21	2.6.1.5	298.15	7.5	0.33	1.0465	-0.11	A
O-acetyl-L-serine	-283.32	2.3.1.30	298.15	6.0	(0.25)	15.00	-6.71	A
UDP-glucose	-1723.83	2.4.1.13	303.15	7.0	(0.25)	6.7	-4.71	B
	-1725.20	2.7.7.9	303.15	7.0	(0.25)	0.286	3.10	B
D-arabino-3-hexulose 6-phosphate	-1302.76	5.3.1.-	303.15	7.0	(0.25)	188.00	-12.98	B
sucrose 6-phosphate	-1559.35	2.4.1.14	298.15	7.0	(0.25)	15.65	-6.11	B
formaldehyde	-45.69	4.1.2.-	303.15	7.0	(0.25)	4×10^{-5}	25.10	B
UDP-galactose	-1721.43	5.1.3.2	300.15	7.1	(0.25)	0.289	3.08	B
erythrulose 1-phosphate	-1160.60	4.1.2.2	301.15	7.4	(0.25)	4.3×10^{-4}	19.21	B
(S)-methylmalonyl-CoA	-340.23	5.1.99.1	303.15	7.4	(0.25)	1.0	0.00	B
N-acetyl-L-methionine	-143.48	3.5.1.14	298.15	7.5	(0.25)	3.6	-3.18	B
D-xylulose 5-phosphate	-1232.98	5.1.3.1	298.15	7.5	(0.25)	1.5	-1.01	B
D-erythrose 4-phosphate	-1164.67	2.2.1.1	298.15	7.6	(0.25)	0.084	6.14	B
sedoheptulose 7-phosphate	-1377.63	2.2.1.1	298.15	7.6	(0.25)	0.9	0.26	B
2-hydroxy-3-oxopropanoate	-486.32	1.1.1.60	296.15	7.6	(0.25)	1.9×10^{-6}	32.65	B
(-)-ureidoglycolate	-473.24	4.3.2.3	303.15	7.5	(0.25)	0.14	4.87	B
allantoate	-361.18	3.5.3.4	303.15	7.5	(0.25)	0.21	3.87	B
5-oxo-D-proline	-224.54	4.2.1.48	297.85	7.9	(0.25)	25.65	-8.04	B
2,2'-iminodipropanoate	-254.03	1.5.1.17	298.15	7.0	(0.25)	1.0×10^{-6}	34.25	C
adenylosuccinate	-1089.48	4.3.2.2	298.15	7.0	(0.25)	0.012	10.96	C
D-arabitol	-284.98	1.1.1.14	298.15	7.0	(0.25)	8×10^{-4}	17.68	C
6-phospho-2-dehydro-3-deoxy-D-gluconate	-1454.78	4.1.2.14	298.15	6.8	(0.25)	0.0016	15.96	C
2-dehydro-3-deoxy-D-galactonate 6-phosphate	-1452.70	4.1.2.21	298.15	6.8	(0.25)	0.0037	13.88	C
enol-phenylpyruvate	-14.50	5.3.2.1	298.15	7.8	(0.25)	0.1	5.71	C
GDP-glucose	-1469.60	2.7.7.34	303.15	7.8	(0.25)	0.25	3.44	C
5-dehydro-D-fructose	-468.28	1.1.1.124	303.15	7.0	(0.25)	4.78×10^{-4}	19.13	C
1-(indol-3-yl)glycerol 3-phosphate	-598.07	4.2.1.20	298.15	7.8	(0.25)	2300.0	-19.19	C
adenosine 5'-tetraphosphate	-3162.93	2.7.4.3	303.15	8.0	(0.25)	0.1	5.71	C

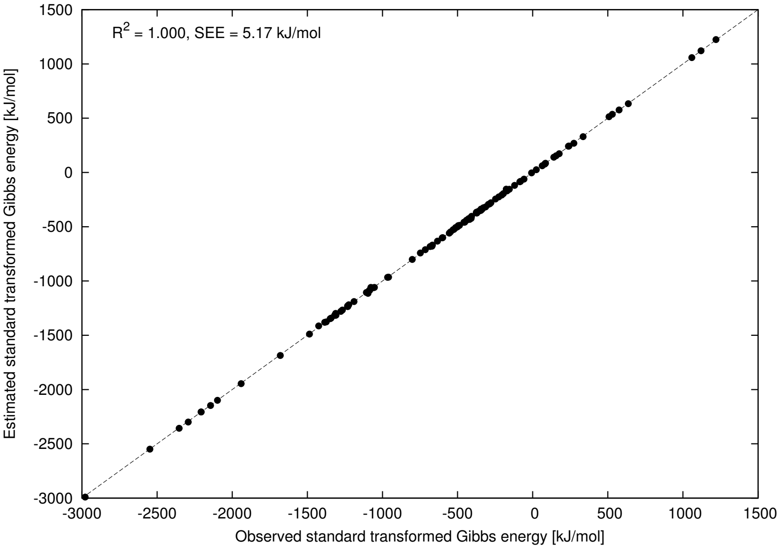


Figure 4.5: Performance evaluation of the estimation of $\Delta_f G'^0$ for biochemical reactants from experimental data. This plot shows the observed $\Delta_f G'^0$ against the estimated $\Delta_f G'^0$.

The plot shows the observed $\Delta_f G'^0$ against the estimated $\Delta_f G'^0$. The range of values of the observed $\Delta_f G'^0$ was between -2978.51 kJ/mol (PRPP) and 1219.74 kJ/mol (oxidized glutathione). The standard error of estimate of the experiment was 5.17 kJ/mol with an optimal R^2 (1.000).

We also applied the approach to estimate $\Delta_r G'^0$ values for 73 reactions where each reaction contained more than one compound with unknown $\Delta_f G'^0$. These reactions are listed in Appendix A together with all relevant parameters of the TECRDB entries used. To evaluate the quality of these estimated $\Delta_r G'^0$ values, we used the following experiment. First, we calculated the $\Delta_r G'^0$ of reactions which contained only compounds with known $\Delta_f G'^0$ (inferred from BasicBiochemData3). These (observed) $\Delta_f G'^0$ values were compared with the corresponding $\Delta_r G'^0$ values (105 overall) estimated using the data mining approach. The respective data plot can be seen in Fig. 4.6. This time, the range of values of the plotted data points was between -30 kJ/mol and 50 kJ/mol. The standard error of estimate was 5.36 kJ/mol and R^2 was 0.905.

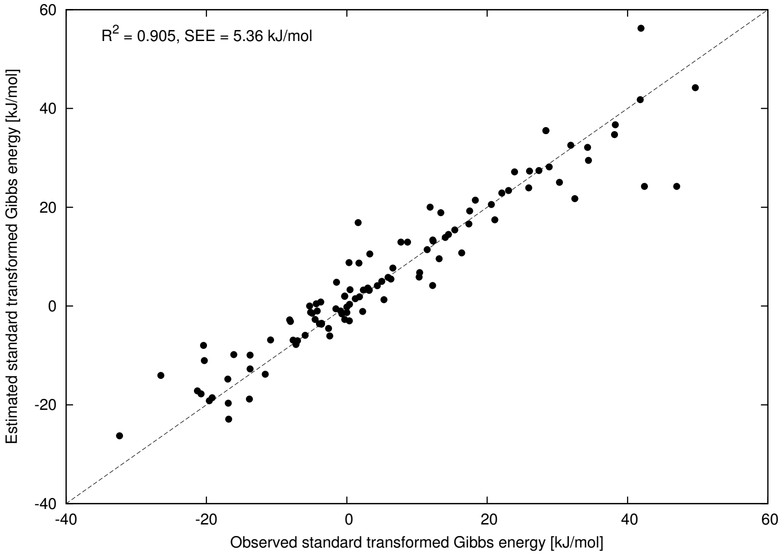


Figure 4.6: Performance evaluation of the estimation of $\Delta_r G'^0$ for biochemical reactions from experimental data. This plot shows the observed $\Delta_r G'^0$ against the estimated $\Delta_r G'^0$.

Prediction of standard transformed Gibbs energies of biochemical reactions using multiple linear regression

The evaluation of the prediction performance based on the independent test procedure, described in the methods section, can be seen in Fig. 4.7. The observed $\Delta_r G'^0$ of 285 reactions are plotted against the predicted $\Delta_r G'^0$ values. The quality of the prediction model was expressed by an R^2 of 0.9892. The corresponding standard error of estimate was 6.12 kJ/mol. The range of values of the plotted (observed) $\Delta_r G'^0$ was between -500 kJ/mol and 80 kJ/mol. Since the $\Delta_r G'^0$ values of 264 reactions (92%) were between -80 kJ/mol and 80 kJ/mol, we further analyzed these reactions in a separate plot which is shown in Fig. 4.8. This plot represents a cut-out of the plot shown in Fig. 4.7. Based on this sample subset, the model obtained an R^2 of 0.9373 and a standard error of estimate of 6.24 kJ/mol.

4.2.4 Discussion

We have developed a novel approach based on QSPR techniques for the prediction of $\Delta_r G'^0$ for biochemical reactions in dilute aqueous solution. The model was trained using stepwise multiple linear regression to select the best features for fitting the training data, which consisted of 484 reactions. The typical error rate for an estimated $\Delta_r G'^0$ was between six and seven kJ/mol.

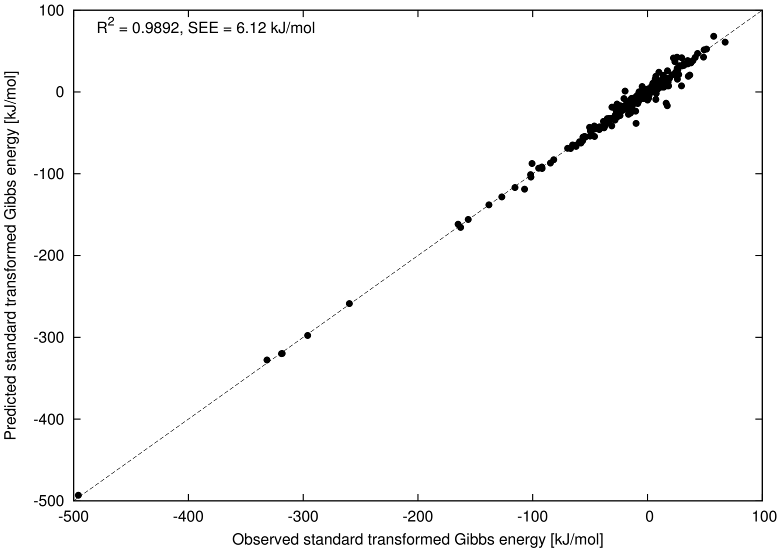


Figure 4.7: Performance evaluation of the predicted $\Delta_r G'^0$ based on the independent test with 285 biochemical reactions. This plot shows the observed $\Delta_r G'^0$ against the predicted $\Delta_r G'^0$ in the range of values between -500 kJ/mol and 80 kJ/mol.

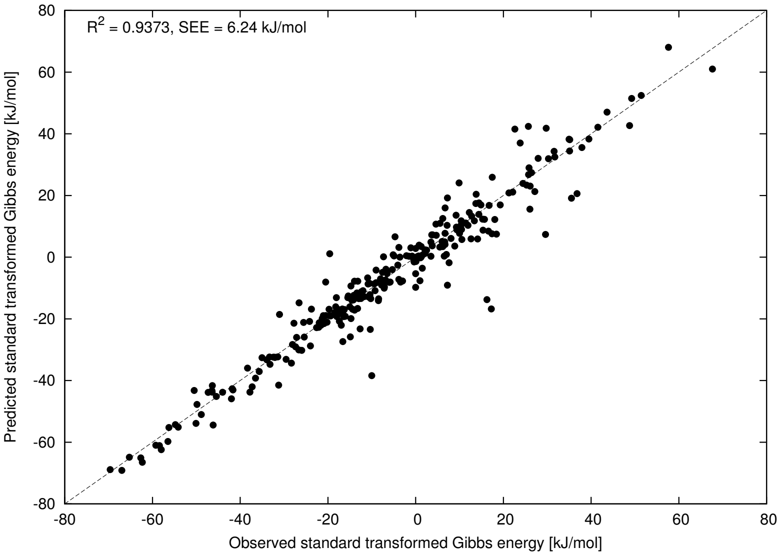


Figure 4.8: Performance evaluation of the predicted $\Delta_r G'^0$ based on the independent test with 264 biochemical reactions. This plot shows the observed $\Delta_r G'^0$ against the predicted $\Delta_r G'^0$ in the range of values between -80 kJ/mol and 80 kJ/mol.

Since the final regression model deliverers acceptable error rates and a significant correlation between observed and predicted values, we believe that our approach is very robust and is able to estimate $\Delta_r G'^0$ of reactions with unknown equilibrium constant. Furthermore, the approach is well suited for *in silico* thermodynamic analysis because it considers important effects on the equilibrium caused by the ionic strength, the presence of metal ions and the dissociation of biochemical compounds in several ionic species at pH 7. Although even better error rates are desirable for more accurate studies, we believe that our approach can give valuable clues about thermodynamic bottlenecks or physiologically irrelevant reactions in metabolic pathways and can help to select plausible and thermodynamically feasible biotransformation routes in large-scale reaction networks.

Using our model, we were able to predict standard energy changes for approx. 63% (or 4600) of the KEGG reactions. The remaining reactions had either unbalanced equations, missing structural information of the compounds or compounds that contained atoms other than H, O, C, N, P and S. Among the limitation with respect to these atom types, our method is restricted to enzymatic reactions, i.e. reactions causing relatively small chemical changes of the involved molecules. It is less suited to estimate the energy balance of overall pathway equations which typically contain structurally very different educts and products. To reach also this goal, the integration of $\Delta_f G'^0$ values for biochemical compounds into the training data set could be an option. Furthermore, the general addition of more training samples, retrieved from the literature or further sources, as well as the application of advanced (non-linear) statistical learning approaches like support vector machines could increase the prediction performance and thus the number of applications.

4.3 Prediction of subcellular protein localization

4.3.1 Introduction

A eukaryotic cell is organized into different membrane-surrounded compartments which are specialized for different cellular functions. However, most cellular proteins are synthesized in the cytoplasm and need to be transported to their final location to fulfill their biological function. The whole protein sorting process is not yet understood in all details but, in principle, it depends on signals in the amino acid sequence or signal patches on the protein surface.

There are diverse applications for the knowledge of the localization of the complete proteome, the localizome, in the fields of proteomics, drug target discovery and systems biology. Since subcellular localization is highly correlated with biological function, it is possible to draw conclusions from the knowledge of a protein's localization regarding its cellular role. Eisenhaber and Bork [1998] described subcellular localization as a key functional characteristic of proteins. Proteins destined for the cell surface are especially of pharmaceutical interest as they are easily accessible drug targets. The integration of large-scale localization data with diverse omics data, produced by high-throughput techniques, will help in understanding cellular function. Localization data can be used to validate or analyze protein-protein interactions inferred from two-hybrid experiments or biochemical pathways inferred from microarray expression data.

In Chapter 2, we described the spatial organization of enzymes as basic principle that supports regulation and fine-tuning of metabolism. The knowledge about the subcellular localizations of all enzymes in a metabolic network under study is therefore helpful to understand metabolism better. Localization information was considered in the path-finding process in order to support the main goal of this thesis, the development of an advanced graph theory-based approach for metabolic pathway analysis (described in Section 4.4.2). The biological meaning of this step derives from the assumption that the enzymes of a metabolic pathway are not distributed over various compartments by chance (see also Fig. 2.3) and tend to catalyze their reactions mainly within one or two localizations. To this end, we extended the weighting scheme of the network graph in order to penalize subsequent reactions that are catalyzed by differently localized enzymes. This allows a better analysis and ranking of alternative pathways that differ in their number of involved compartments.

In recent years large-scale sequencing projects have caused a rapid growth of sequence information and increased the number of proteins but without any further annotation in public databases. These databases also include relevant data sources specialized on metabolic pathways like KEGG and MetaCyc. Determining the localization of proteins using experimental methods alone is expensive and time-consuming.

Fast and accurate computational prediction methods provide an attractive complement to experimental methods. In the last decade numerous computational methods, which can be roughly divided into sequence-based and annotation-based methods [Emanuelsson *et al.*, 2007; Nair and Rost, 2005], have been developed. Sequence-based predictors only use the amino acid sequence of the query protein as input. They are based either on the detection of sequence-coded sorting signals like N-terminal targeting peptides

[Emanuelsson *et al.*, 1999; Bendtsen *et al.*, 2004; Emanuelsson *et al.*, 2000; Bannai *et al.*, 2002; Petsalaki *et al.*, 2006; Fujiwara *et al.*, 2001; Boden *et al.*, 2005; Small *et al.*, 2004; Cokol *et al.*, 2000] and nuclear localization signals (NLS) [Cokol *et al.*, 2000] or use the fact that the amino acid composition of a protein is correlated with its localization [Andrade *et al.*, 1998]. The latter methods [Cedano *et al.*, 1997; Reinhardt and Hubbard, 1998; Hua and Sun, 2001; Park and Kanehisa, 2003; Xie *et al.*, 2005; Guo and Lin, 2006; Nair and Rost, 2005; Pierleoni *et al.*, 2006; Cui *et al.*, 2004; Chou and Cai, 2003] use different kinds of composition information like the overall, paired, gapped-paired, surface or pseudo amino acid composition from the protein sequence or sequence profiles. More recent and advanced methods combine composition information with the detection of sorting signals [Horton *et al.*, 2007; Höglund *et al.*, 2006]. Annotation-based predictors search the sequence for functional domains and motifs [Chou and Cai, 2002; Scott *et al.*, 2004] or use textual information like Swiss-Prot keywords [Nair and Rost, 2002; Lu *et al.*, 2004], Gene Ontology (GO) terms [Lei and Dai, 2006; Huanq *et al.*, 2008] or PubMed abstracts [Brady and Shatkay, 2008; Fyshe *et al.*, 2008]. If such information is not available for the query protein most of these methods transfer annotation from close homologs. Nair and Rost [2002] showed that homology-driven subcellular localization assignment works because the localization is clearly conserved in the protein sequence. Annotation-based predictors often report higher performance than sequence-based predictors which, however, are more general and robust and can also be used for novel proteins for which no additional information is present and no annotated close homologs can be found. In addition to the predictors of the two categories, there are also hybrid approaches which combine sequence-based and annotation-based information [Shatkay *et al.*, 2007; Guda and Subramaniam, 2005; Bhasin and Raghava, 2004; Shen *et al.*, 2007; Chou and Cai, 2004] and can therefore profit from the advantages of both worlds.

Although there exist already numerous computational prediction methods, there is still room for improvement. This is due to the fact that the protein sorting process is very complex and not yet well understood. Only a small portion of proteins have clearly identifiable sorting signals in their primary sequence. As a consequence, available prediction methods are often either specialized for the prediction of very few localizations with higher accuracy or for the prediction of a wide range of localizations with reduced accuracy.

Our previously published support vector machine (SVM) based predictor MultiLoc [Höglund *et al.*, 2006] utilizes overall amino acid composition and the presence of known sorting signals. The aim of this work was to show that sequence-based predictors like MultiLoc can be improved by incorporating phylogenetic profiles and GO terms inferred from the primary sequence leading to a high-accuracy prediction system that covers all main eukaryotic subcellular localizations. Phylogenetic profiles encode evolutionary information in the form of patterns of protein inheritance among the species. Marcotte *et al.* [2000] successfully applied this approach to distinguish mitochondrial and non-mitochondrial proteins. GO terms were previously combined with sequence-based information in the form of pseudo amino acid composition [Shen *et al.*, 2007]. The GO terms are used as primary prediction criteria and pseudo amino acid composition is used if no GO term can be found. Our novel MultiLoc2 prediction system integrates composition and sorting signal information with phylogenetic profiles and GO terms towards a common localization prediction. The extended MultiLoc system was trained on two different datasets resulting in two versions with different resolutions. MultiLoc2-LowRes is a low resolution predictor that is specialized for globular proteins and predicts up to five localizations for animals, fungi and plants. MultiLoc2-HighRes is a high resolution predictor that covers all 11 main eukaryotic subcellular localizations.

MultiLoc2 was compared with current state-of-the-art tools (BaCelLo [Pierleoni *et al.*, 2006], LOCtree [Nair and Rost, 2005], Protein Prowler [Boden *et al.*, 2005], TargetP [Emanuelsson *et al.*, 2000] and WoLF PSORT [Horton *et al.*, 2007]) using independent datasets sharing very low sequence identity with the training datasets of all compared tools. We found MultiLoc2 to perform considerably better than related tools for animals and plants and comparably well for fungal proteins in a benchmark study with five localizations. Since GO terms are not always available, we evaluated MultiLoc2 as purely sequence-based and found the performance only slightly reduced but still better or comparable with other tools showing the robustness of our method. Furthermore, MultiLoc2-HighRes performed significantly better compared with WoLF PSORT using a second independent dataset that extends the benchmark study to all main eukaryotic subcellular localizations. In the following sections the MultiLoc2 system is described in detail together with the training and test datasets used, followed by the performance evaluation and the results of the benchmark studies. Both novel tools are available online at http://www-bs.informatik.uni-tuebingen.de/Services/MultiLoc2.

4.3.2 Methods

MultiLoc2 architecture

The MultiLoc prediction system described earlier [Höglund *et al.*, 2006] is based on the integration of the output of four sequence-based subclassifiers

(SVMTarget, SVMSA, SVMaac and MotifSearch) into a protein profile vector (PPV). The subclassifiers utilize the overall amino acid composition or search for specific sorting signals. MultiLoc2 extends the original architecture with two new classifiers based on phylogenetic profiles (PhyloLoc) and GO terms (GOLoc). As stated in the introduction, there are two versions of MultiLoc2 which differ in the number of predictable localizations. MultiLoc2-HighRes can deal with nuclear (nu), cytoplasmic (cy), mitochondrial (mi), chloroplast (ch), extracellular (ex), plasma membrane (pm), peroxisomal (pe), endoplasmic reticulum (er), Golgi apparatus (go), lysosomal (ly) and vacuolar (va) proteins. MultiLoc2-LowRes is specialized for globular proteins and predicts secretory pathway (SP) proteins (separated into the six classes ex, pm, er, go, ly, va in MultiLoc2-HighRes) as well as nu, cy, mi and ch. Similar to its previous version, MultiLoc2 is available for plant, animal and fungal protein localization prediction. An example of the overall architecture of MultiLoc2 is shown in Fig. 4.9. A query sequence is processed by a first layer of six subprediction methods. The results from these methods are collected in the PPV, which is used as input for the final layer of SVMs, which in turn yields the final localization prediction. In both layers one-vs.-one SVMs are used for classification. The corresponding figure of MultiLoc2-LowRes is available in Appendix B. The original four sequence-based classifiers are briefly described in the next section, followed by details of PhyloLoc and GOLoc.

Subprediction methods

SVMTarget: SVMTarget is based on the detection of N-terminal targeting peptides to predict ch, mi, SP and other (OT) localizations for plant proteins and only mi, SP and OT for animal and fungal proteins. A sliding window approach scans the N-terminal part of a given query sequence. The partial amino acid composition in the window is used as input for the the SVMs. The output of SVMTarget is a probability for each localization.

SVMSA: SVMSA scans the sequence for a signal anchor (SA) which can be present in membrane proteins of the secretory pathway instead of a signal peptide. Therefore, SVMSA complements SVMTarget. SAs are also detected using a sliding window approach based on partial amino acid composition. SVMSA is specialized for membrane proteins and is therefore not included in MultiLoc2-LowRes.

SVMaac: SVMaac is based on the overall amino acid composition of the query sequence and outputs a probability for each localization. In contrast to the original MultiLoc, the binary one-versus-all classification is replaced by a one-versus-one procedure.

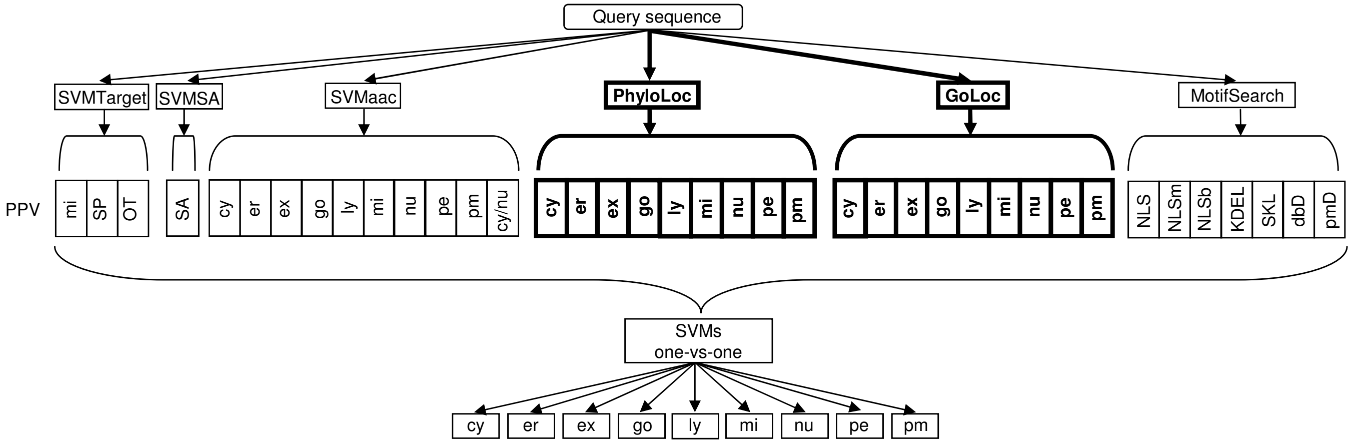


Figure 4.9: The architecture of MultiLoc2-HighRes (animal version). A query sequence is processed by a first layer of six subprediction methods (SVMTarget, SVMSA, SVMaac, PhyloLoc, GOLoc and MotifSearch). The individual output of the methods of the first layer are collected in the protein profile vector (PPV), which enters a second layer of SVMs producing probability estimates for each localization.

MotifSearch: MotifSearch outputs five binary features that encode the presence or absence of sequence motifs relevant to protein sorting like nuclear localization signals (NLSs). Two additional binary features represent the presence or absence of a DNA-binding domain or a plasma membrane receptor domain.

PhyloLoc: Proteins within the same subcellular localization tend to share a similar phylogenetic distribution of their homologs in organisms with known genome [Marcotte *et al.*, 2000]. This kind of information can be represented as a phylogenetic profile [Pellegrini *et al.*, 1999] which encodes the pattern of presence or absence of a given protein in known genomes. Marcotte *et al.* [2000] applied phylogenetic profiles for the distinction of mitochondrial and non-mitochondrial proteins using 31 genomes and a linear discrimination function. PhyloLoc is based on phylogenetic profiles derived from 78 fully sequenced genomes and SVMs to predict all of the localizations of the MultiLoc2 predictors. The genomes were downloaded from the National Center for Biotechnology Information (NCBI) web site. We used all available eukaryotic (20) and archaean (33) genomes and a non-redundant set of 25 bacterial genomes. (More details are available in Appendix B.) The input of PhyloLoc (as shown in Fig. 4.10) is a vector of similarities between the query sequence and the best sequence match in each genome using BLAST. The BLAST homology searches were performed using default settings. The bit score B_{qi} of the best sequence match of the query sequence q in genome i and the self bit score B_{qq} of q aligned with itself were used to calculate the similarity S_{qi} which is defined as: $S_{qi} = B_{qi}/B_{qq}$. Due to the fact that B_{qi} is always smaller than B_{qq}, the values of S_{qi} range from zero to one. Values close to one indicate presence of the query protein and values close to zero indicate absence. The calculation of phylogenetic profiles based on bit scores was also previously used for the functional annotation of bacterial genomes [Enault *et al.*, 2003]. An important point to note is that, although BLAST is used, creating phylogenetic profiles is not an annotation-based or homology-based method as sometimes described in the literature. The reason is that there is no annotation-transfer from the aligned sequences. Actually, it is irrelevant whether the proteins of the genomes are annotated or not. Proteins with similar phylogenetic profiles are co-inherited and do not have to be close homologs [Marcotte, 2000b].

GOLoc: The Gene Ontology (GO) is a controlled vocabulary for uniformly describing gene products in terms of biological processes, cellular components and molecular function across all organisms [Ashburner *et al.*, 2000]. It has been shown that GO terms can be used to improve the performance of subcellular protein localization prediction methods [Chou and Cai, 2003; Lu and Hunter, 2005]. In the literature to date, there are three possibilities

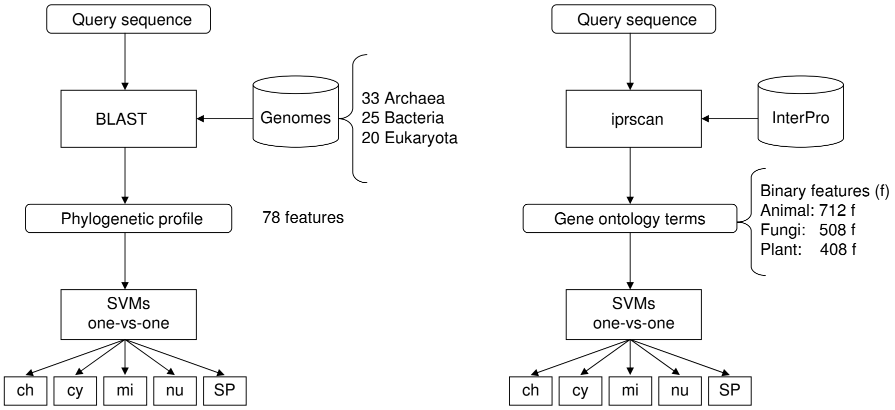


Figure 4.10: The architectures of PhyloLoc and GOLoc from MultiLoc2-LowRes. The input of PhyloLoc is a vector of similarities (phylogenetic profile) between the query sequence and the best sequence match in each genome inferred from BLAST. The input of GOLoc is a binary encoded vector representing the GO terms of the query sequence inferred from InterPro using InterProScan. PhyloLoc and GOLoc use one-versus-one SVMs to process their input and to calculate probability estimates for each localization.

for obtaining GO annotation terms for a query sequence. If the UniProt [Bairoch *et al.*, 2005] accession number is known, one can simply extract the GO annotation from the UniProt database [Shen *et al.*, 2007]. However, this procedure fails for novel proteins without accession number. Another possibility is to search for homologous proteins annotated with GO terms using BLAST [Lei and Dai, 2006; Huanq *et al.*, 2008]. This becomes difficult in cases where proteins have no close homolog or proteins have many homologs. In this case, no GO term can be obtained or GO terms might be ambiguous. Another method of inferring GO terms is InterProScan [Zdobnov and Apweiler, 2001] used, for example, by Chou and Cai [2004]. Given a protein sequence, the tool scans against various pattern and signature data sources collected by the InterPro project [Mulder *et al.*, 2007]. InterPro also provides a mapping of the detected protein domains and functional sites to GO terms.

Our subpredictor GOLoc is based on GO terms calculated using InterProScan. Since the GO terms are derived directly from the query sequence, we avoid the drawbacks of using accession numbers or BLAST. The input of GOLoc is a binary-coded vector which represents all GO terms of the training sequences (see Fig. 4.10). GO terms present in the query sequence are set to 1 in the vector and to 0 otherwise.

Datasets

BaCelLo: The datasets used for training and testing MultiLoc2-LowRes against comparable predictors were obtained from the BaCelLo website[7]. The homology-reduced training dataset was extracted from Swiss-Prot release 48 and contains 2597 animal, 1198 fungal and 491 plant proteins resulting in three kingdom-specific predictors. By ignoring proteins annotated as 'membrane' or 'transmembrane', only globular proteins were considered.

The animal and fungal proteins represent four localizations (nu, cy, mi, SP) and the plant proteins five localizations (with the addition of ch). The independent test dataset was extracted from Swiss-Prot release 54. Only proteins added to the database starting from release 49 were considered. Furthermore, proteins sharing a sequence identity >30% to at least one protein from release 48 were removed. This ensured that all test proteins were novel to the predictors in the benchmark study since all of them were trained using Swiss-Prot proteins up to release 48. In order to avoid a bias towards the prediction of over-represented protein classes, all sequences which share the same localization and align with an E-value lower than 10^{-3} using BLAST were clustered into 432 animal, 418 fungi and 132 plant groups.
MultiLoc: For training MultiLoc2-HighRes we used the original MultiLoc dataset [Höglund *et al.*, 2006], which contains 5959 eukaryotic proteins extracted from Swiss-Prot release 42. The data set covers 11 localizations (cy, ch, er, ex, go, ly, mi, nu, pe, pm, va). To also compare the prediction performance of MultiLoc2 with WoLF PSORT in regard to the localizations not present in the BaCelLo test dataset, we created a second independent dataset which covers seven localizations (er, ex, go, ly, pe, pm, va). To this end, animal, fungal and plant proteins of these localizations were extracted from Swiss-Prot release 55.3 in the same way as the BaCelLo independent dataset. However, in the case of the plant proteins, we increased the allowed sequence identity threshold to 40% in order to obtain enough data. We used BLASTClust to cluster the sequences using 30% pairwise sequence identity for the animal and fungal proteins and 40% for the plant proteins. The whole procedure delivered 158 animal, 106 fungi and 30 plant groups.

SVM training and performance evaluation

All building blocks (except MotifSearch) of MultiLoc2 were trained using SVMs [Vapnik, 1999] from the LIBSVM [Chang and Lin, 2003] software. We used the radial basis kernel function throughout, and optimized the c and g parameters by grid search. Furthermore, we defined weights for each class

[7] http://gpcr2.biocomp.unibo.it/bacello/index.htm

in order to reduce the over-prediction effect when using unbalanced training datasets. The probability estimates calculated by LIBSVM were used for ranking the final predicted localizations and choosing the most probable one.

We used five-fold cross-validation for training and evaluating the prediction performance. Additionally, independent datasets were used for testing MultiLoc2 and comparison with other prediction methods. Therefore, all test proteins share low sequence homology with proteins in the training datasets. Localization-specific performance results were expressed using sensitivity (SE) and the Matthews correlation coefficient (MCC). To evaluate the overall prediction performance, we used average sensitivity (AVG), which is also known as the average localization-specific accuracy, as primary measure. The average sensitivity is better suited than the overall accuracy (ACC), the percentage of correctly predicted proteins of all localizations. The reason is that all prediction methods were trained on unbalanced datasets with strongly varying numbers of proteins per localization. This often biases the prediction towards the localization with the most representations in the training dataset. Hence an unbalanced test dataset would also normally lead to a distorted performance evaluation when using the ACC only. To calculate the performance measures for the independent datasets, we used the average rates of true and false predicted proteins within each cluster.

4.3.3 Results

Cross-validation performance

The impact of the MultiLoc2 extensions on the overall prediction performance was evaluated using 5-fold cross-validation. The results are summarized in Tab. 4.3. The average sensitivity and overall accuracy of MultiLoc2-LowRes (trained on the BaCelLo dataset) and MultiLoc2-HighRes (trained on the MultiLoc dataset) were compared with those of the original MultiLoc architecture and MultiLoc extended by PhyloLoc as well as GOLoc only. Using the BaCelLo dataset, MultiLoc2-LowRes yielded a significantly higher AVG (85.0% for animals, 83.9% for fungi and 81.6% for plants) than the original MultiLoc (77.3%, 78.4% and 71.4% respectively). For the MultiLoc dataset the AVG was increased from 78.6% to 89.2% for animal, from 78.0% to 89.2% for fungal and from 78.0% to 89.4% for plant proteins by the MultiLoc2-HighRes system compared to the original MultiLoc. Note that the performance results for the original MultiLoc differed from those previously reported [Höglund *et al.*, 2006] since the SVMaac architecture has slightly changed.

Table 4.3: Cross-validation performance comparison of different MultiLoc architectures trained using the BaCelLo and the MultiLoc datasets. This table compares the average sensitivities (AVGs) and overall accuracies (ACCs) of MultiLoc2-LowRes and MultiLoc2-HighRes with those of the original MultiLoc and the extended architecture based on PhyloLoc as well as GOLoc only. The AVGs and ACCs are given in percent. The standard deviations (in parentheses) refer to the differences of the AVGs and ACCs of the different cross-validation models.

Dataset	**Method**	**Animals**			**Fungi**			**Plants**		
		No.	**avgACC**	**ovACC**	**No.**	**avgACC**	**ovACC**	**No.**	**avgACC**	**ovACC**
BaCelLo										
	MultiLoc	2597	77.3 (±2.9)	75.7 (±3.1)	1198	78.4 (±2.7)	71.0 (±2.6)	491	71.4 (±6.8)	67.8 (±3.8)
	+ PhyloLoc		80.1 (±2.4)	78.2 (±2.9)		80.0 (±2.5)	73.6 (±0.9)		78.6 (±3.6)	77.4 (±1.9)
	+ GOLoc		83.4 (±1.6)	82.4 (±1.8)		80.7 (±1.1)	75.1 (±1.7)		79.3 (±4.1)	74.9 (±3.8)
	MultiLoc2-LR		85.0 (±1.9)	83.9 (±2.6)		84.0 (±1.5)	78.9 (±1.8)		81.6 (±2.9)	79.6 (±4.3)
MultiLoc										
	MultiLoc	5447	78.6 (±1.2)	76.4 (±1.2)	5407	78.0 (±1.3)	76.6 (±1.2)	5856	78.0 (±1.8)	76.4 (±1.7)
	+ PhyloLoc		84.6 (±0.7)	84.0 (±0.6)		84.7 (±1.4)	84.4 (±0.9)		86.5 (±1.5)	84.3 (±0.7)
	+ GOLoc		87.5 (±1.7)	86.6 (±1.1)		87.6 (±0.6)	87.2 (±0.9)		87.1 (±1.4)	86.5 (±1.1)
	MultiLoc2-HR		89.2 (±1.5)	88.7 (±1.1)		89.2 (±0.7)	88.7 (±1.0)		89.4 (±0.8)	88.6 (±0.9)

Adding PhyloLoc or GOLoc individually to MultiLoc already increased the performance significantly whereas the performance gain caused by GOLoc was slightly higher compared to PhyloLoc. However, the best performance was achieved by the addition of both subpredictors in MultiLoc2. Similar trends could be detected regarding the overall accuracies. The standard deviations of the MultiLoc2-LowRes plant version were higher compared to the other versions which is due to the fact that the number of training sequences in the dataset is significantly lower.

Comparison with related tools

In a recently published study [Casadio *et al.*, 2008] five selected top-performing sequence-based prediction methods (BaCelLo, LOCtree, Protein Prowler, TargetP and WoLF PSORT) were compared using an independent dataset. Based on this benchmark study, we compared the performance of MultiLoc2 against these five methods using the same test setting. The benchmark study considered five subcellular localizations (nu, cy, mi, ch, SP). Furthermore, a virtual class nu/cy containing nu and cy proteins was created in order to ensure a fair comparison with TargetP and Protein Prowler which do not discriminate between these two localizations. To deal with WoLF PSORT and LOCtree, predicted sublocalizations of the secretory pathway were grouped into the SP class. A similar approach was followed for MultiLoc2-HighRes. Depending on the inclusion of the virtual nu/cy class, the number of tested classes was three or four for animals and four or five for fungi and plants. We also evaluated the performance of only sequenced-based predictions of MultiLoc2 by disregarding GO terms. The performance resembles the case of unavailability of GO terms. Tab. 4.4 shows the localization-specific performance results using sensitivity and MCC and Tab. 4.5 summarizes the overall performances using AVG and ACC. Note that the number of SP clusters for fungi (9) and plants (6) and the mi clusters for plants (6) is quite small compared to the remaining localizations. Therefore, some care should be taken when interpreting the prediction results. Small clusters have only a small influence on the ACC, however, a large influence on the AVG.

MultiLoc2-LowRes always yielded the highest ACCs and AVGs for animal and plant proteins and hence outperformed all other predictors. The reason for this outstanding result is that MultiLoc2-LowRes is, in general, better suited to discriminate between nu and cy and between mi and ch proteins (see Tab. 4.4), which is a known challenge in the prediction of subcellular protein localization.

Table 4.4: Comparison of the localization-specific prediction results using an independent dataset. The sensitivity (SE), given in percentages, and Matthews correlation coefficient (MCC) are listed for each localization (Loc). The number of clusters (No.) per localization is also shown. In Protein Prowler and TargetP, predictions for nu and cy are only available grouped as nu/cy.

Version	**Loc**	**No.**	**MultiLoc2-LR**		**MultiLoc2-HR**		**BaCelLo**		**LOCtree**		**Protein Prowler**		**TargetP**		**WoLF PSORT**	
			SE	MCC	SE	MCC	SE	MCC	SE	MCC	SE	MCC	SE	MCC	SE	MCC
Animals	SP	75	95	0.87	88	0.81	93	0.88	79	0.65	86	0.88	88	0.88	92	0.80
	mi	48	84	0.74	85	0.77	74	0.66	64	0.51	51	0.71	82	0.63	71	0.63
	nu	224	65	0.59	55	0.52	57	0.41	66	0.39					77	0.58
	cy	85	70	0.44	71	0.36	51	0.21	35	0.22					34	0.23
	nu/cy	308	91	0.83	91	0.79	93	0.83	84	0.64	98	0.79	89	0.75	89	0.76
Fungi	SP	9	78	0.59	78	0.61	100	0.74	78	0.35	93	0.20	89	0.56	89	0.73
	mi	77	66	0.61	56	0.56	79	0.58	42	0.38	33	0.51	50	0.44	53	0.44
	nu	152	59	0.35	46	0.29	72	0.38	63	0.22					93	0.35
	cy	180	57	0.27	58	0.21	32	0.19	35	0.15					11	0.19
	nu/cy	332	92	0.64	86	0.51	85	0.61	83	0.31	98	0.52	89	0.48	89	0.46
Plants	SP	6	83	0.57	83	0.51	100	0.66	83	0.60	100	0.61	100	0.61	33	0.24
	mi	6	67	0.46	67	0.42	17	0.40	58	0.30	67	0.40	50	0.26	42	0.52
	ch	72	71	0.67	54	0.52	71	0.54	77	0.66	7	0.40	55	0.49	61	0.43
	nu	36	94	0.76	86	0.75	88	0.60	72	0.61					72	0.52
	cy	17	35	0.33	37	0.20	27	0.38	33	0.39					24	0.28
	nu/cy	52	96	0.85	93	0.74	88	0.70	75	0.70	86	0.52	83	0.62	87	0.61

Table 4.5: Comparison of the overall performance results using an independent dataset. The average sensitivity and the overall accuracy (in parentheses) for the prediction of three and four classes for animals and fungi and four and five classes for plants are shown. Both measures are given in percentages. The top-scoring average sensitivity and average accuracy are highlighted in bold. Results for Protein Prowler and TargetP predictions are only available for a reduced number of classes since nu and cy are grouped as nu/cy.

Version	**Classes**	**Average accuracy (Overall accuracy)**						
		MultiLoc2-LR	**MultiLoc2-HR**	**BaCelLo**	**LOCtree**	**Protein Prowler**	**TargetP**	**WoLF PSORT**
Animals	3	**90** (**91**)	88 (90)	87 (**91**)	76 (81)	78 (**91**)	86 (88)	84 (88)
	4	**79** (**73**)	75 (67)	69 (64)	61 (62)			69 (71)
Fungi	3	79 (**87**)	73 (80)	**88** (84)	68 (75)	75 (86)	76 (82)	77 (82)
	4	65 (**60**)	60 (54)	**71** (57)	55 (47)			62 (51)
Plants	4	**79** (**81**)	74 (71)	69 (76)	73 (76)	65 (63)	72 (67)	56 (69)
	5	**70** (**73**)	65 (62)	61 (69)	65 (70)			46 (57)

For fungal proteins the ACCs were the highest and the AVGs were the second highest after the BaCelLo predictor. One reason for the reduced AVG performance is that on average only 34% of the fungal proteins were annotated with GO terms by InterProScan. The annotation-rate was higher for animals (43%) and plants (79%). Compared to MultiLoc2-LowRes, the performance of MultiLoc2-HighRes was, not surprisingly, reduced, since it is a more general predictor not specialized for globular proteins and covering a wider range of localizations. However, for animal and plant proteins the AVGs of MultiLoc2-HighRes were equal or higher compared to the other methods. Similar to MultiLoc2-LowRes, MultiLoc2-HighRes performed worse for fungal proteins. The AVGs were still better than LOCtree, however, worse compared with Protein Prowler, TargetP and WoLF PSORT.

Simulating the case in which no GO term was available for any test proteins, the overall performances of the MultiLoc2 predictors were slightly reduced but still better than the other methods for animal and plant and comparable for fungal proteins. Detailed results are available in Appendix B.

In a second benchmark study, MultiLoc2-HighRes and WoLF PSORT were compared using the MultiLoc independent dataset. In contrast to the other predictors, both methods allow the prediction of all main eukaryotic subcellular localizations. We further note that WoLF PSORT can also distinguishes the cytoskeleton within the cytoplasm. In this comparison we only considered those localizations (ex, pm, pe, er, go, ly, va) not tested in the previous study. Since it is known that discriminating between these classes is a big challenge, we also evaluated whether the tested proteins could be correctly predicted within the top three ranked localizations. The results of this study are summarized in Tab. 4.6. MultiLoc2-HighRes always achieved significantly higher AVGs. In particular, the AVG within the top three locations was about twice as high for MultiLoc2 than for WoLF PSORT. A similar result was observed regarding the ACCs. MutliLoc2-HighRes had a much lower bias towards overrepresented localizations and, thus, almost never showed zero sensitivity for a localization with few representatives. This again proves high robustness of MultiLoc2, even in cases of many localizations.

4.3.4 Discussion

Our new approach for predicting subcellular protein localization, MultiLoc2, integrates several subpredictors based on the overall amino acid composition, the detection of sorting signals, phylogenetic profiles and GO terms.

Table 4.6: The sensitivity (SE) and top three sensitivity (SE3) for each localization are shown. SE3 measures the fraction of correctly predicted proteins within the top three ranked localizations. The corresponding average sensitivity and overall accuracy are listed also, with the top-scoring highlighted in bold. All measures are given as percentages.

Loc	**animal**					**fungi**					**plant**				
	No.	**MultiLoc2-HR**		**WoLF PSORT**		**No.**	**MultiLoc2-HR**		**WoLF PSORT**		**No.**	**MultiLoc2-HR**		**WoLF PSORT**	
		SE	**SE3**	**SE**	**SE3**		**SE**	**SE3**	**SE**	**SE3**		**SE**	**SE3**	**SE**	**SE3**
ex	78	78	92	93	97	7	71	86	36	79	1	0	100	0	0
pm	34	51	78	41	59	29	7	28	59	79	6	17	50	83	83
pe	3	33	100	0	0	5	20	100	0	0	2	50	100	0	0
er	25	32	74	8	40	46	48	85	9	25	6	50	83	0	50
go	14	14	50	0	7	8	38	63	0	0	6	33	33	17	17
ly	4	25	75	0	25										
va						11	0	0	0	0	9	22	59	0	33
avgACC		**39**	**78**	24	38		**31**	**60**	17	31		**29**	**73**	17	31
ovACC		57	**82**	**58**	68		**31**	**59**	22	51		**30**	**63**	20	40

Compared to the original MultiLoc architecture, the robustness and prediction performance was significantly improved. The performances of the MultiLoc2 predictors were compared with current state-of-the-art sequence-based methods using independent datasets.

MultiLoc2-LowRes is specialized for globular proteins and offers kingdom-specific prediction of up to five localizations based on the BaCelLo dataset. On the other hand, MultiLoc2-HighRes is able to deal with membrane proteins and predicts all of the main eukaryotic localizations based on a dataset that consists of a mixture of animal, fungal and plant proteins. In comparison with five other methods, the MultiLoc2 predictors performed better for animal and plant proteins whereas MultiLoc2-LowRes outperformed MultiLoc2-HighRes in general. However, the performance of MultiLoc2-HighRes is remarkable since it is able to predict more localizations than the other tools except WoLF PSORT. We also simulated the scenario in which no GO term was available for any test proteins, which made the prediction sequence-based only. The resulting performance of the MultiLoc2 predictors was slightly reduced but still better for animals and plants and comparable for fungi. Therefore, we conclude that the MultiLoc2 approach is very robust and well suited for novel proteins without relevant sequence homology to annotated proteins but can also benefit from the presence of calculated GO annotation from the sequence using InterProScan.

In a second benchmark study we evaluated the prediction performance of MultiLoc2-HighRes compared to WoLF PSORT for proteins localized in the peroxisomes and in the sublocalizations of the secretory pathway. For all three eukaryotic kingdoms, MultiLoc2-HighRes performed significantly better. In particular, MultiLoc2-HighRes showed much better sensitivity throughout all localizations and yielded high robustness. However, the results indicate that the classification in all main eukaryotic localizations is still a challenging task and leaves room for improvement for future work.

We also demonstrated that our concept of a PPV is very useful since it can be easily extended and enables the integration of very heterogeneous but relevant information towards a common prediction of subcellular protein localization. Future improvements of our approach could be based on the integration of further relevant sequence-based or annotation-based information. An important issue that is not explicitly covered is the handling of proteins present in multiple localizations. Furthermore, with more annotated sequences available, the consideration of more locations like the mitochondrial or choroplast sub-compartments should be one of the next steps in order to increase the usefulness of the approach. At the moment, MultiLoc2 is only able to annotate eukaryotic proteins. However, an extended version suited for prokaryotes should also be taken into account.

4.4 Graph theory-based inference of feasible biotransformation routes

4.4.1 Introduction

Graph theory-based approaches (as introduced in Chapter 2) infer linear biotransformation routes using efficient (k-shortest) path-finding algorithms that can handle genome-scale networks. The big challenge for these approaches is to apply useful optimization criteria in order to find the most relevant routes within the k-shortest paths among a huge number of possible routes which are biologically irrelevant to a great extent.

Weighted graphs enable an easy integration of suitable information into the path search to overcome the above mentioned problem. For example, in the degree-weighted metabolic networks approach [Croes *et al.*, 2006], the metabolic network of an organism is mapped onto a bipartite graph, including all compounds and reactions as nodes. Directed edges connect the compound nodes (educts and products) with the reaction nodes. Both directions of a reaction are represented by two independent nodes per reaction. The key idea of a weighted metabolic network is to assign each compound node a weight equal to its degree (e.g. the number of in- and outgoing edges) and each reaction node the weight 1 by default. The weight of a path in the graph is then defined by the sum of the weights of its nodes. This implies that the overall weight of a path is much larger if it contains highly connected compounds like typical pool metabolites or co-factors (e.g. NADP, ATP, water). Searching for paths with lowest weight reduces the probability of finding unfeasible biotransformation routes which contain pool metabolites as intermediates between two subsequent reactions.

A fundamental problem of this lightest-path search is its inability to handle important biotransformation routes that involve the biosynthesis of pool metabolites (e.g. purine biosynthesis in which AMP and ADP are intermediates). The method fails to reconstruct these routes because pool metabolites participate in many reactions of other transformation processes and, therefore, are assigned very large node weights. Another problem are routes passing through pathways of the core metabolism like glycolysis or the TCA cycle because highly connected metabolites, for example, pyruvate or acetyl-CoA are involved.

The main goal of this thesis was to develop a novel graph theory-based approach which complements and improves existing methods. The approach is based on a novel graph representation of the metabolic network, called the metabolic transition graph, and on novel weighting schemes for a better detection of biologically meaningful routes. To this end, pre-calculated atom mapping rules were integrated into the graph representation and combined with the lightest-path search. To improve the graph weighting, novel weights were defined based on reaction context, thermodynamic and subcellular localization information. In the next sections, these novel concepts will be described followed by an extensive evaluation of the search performance of the approach.

4.4.2 Methods

Combining atom mapping rules with lowest weight paths

The mentioned problems of the degree-weighted metabolic networks approach can be reduced by combining the lightest-path search with atom mapping rules. The key idea is to use atom mapping rules to identify biochemically irrelevant paths of low weight. To this end, all relevant paths must fulfill the structural moiety constraint (SMC). The usefulness of the SMC when searching for relevant routes was briefly motivated in Section 4.1 where the calculation of atom mapping rules was described. Since the application of this constraint to the lightest-path search represents one of the most important aspects of this thesis, the topic is recaptured in this section. The structural moiety constraint can be defined as follows. A path and its corresponding biotransformation route can only be feasible if at least one atom of the source compound is transferred, via the intermediates, to the target compound. In many cases, this helps to filter out biochemically irrelevant lowest-weight paths. We also show that the combination of atom mapping rules with lowest-weight paths performs better than searching for the shortest path in the unweighted atom mapping graph. The example shown in Fig. 4.11 illustrates the concept of using the structural moiety constraint for path validation. 3-phosphoglycerate, also known as an intermediate in the degradation of glucose in glycolysis, is used as source metabolite and the amino acid L-alanine as target. The dashed arrows (1, 2, 3, 4A, 5, 6) describe a path which consists of six enzymatic steps for transforming 3-phosphoglycerate into L-alanine. Five intermediates are required. The rectangles mark the conserved substructures. In this example, 3-phosphoglycerate serves as carbon source for L-alanine. The sequential application of atom mapping rules, linking the educts and products in each reaction, enables a tracing of the conserved structure. Now, it is clear that this path fulfills the structural moiety constraint.

The path described by the solid arrows (1, 2, 3, 4B) requires only four steps in total. The enzymatic reaction with EC number 2.6.1.51 is used

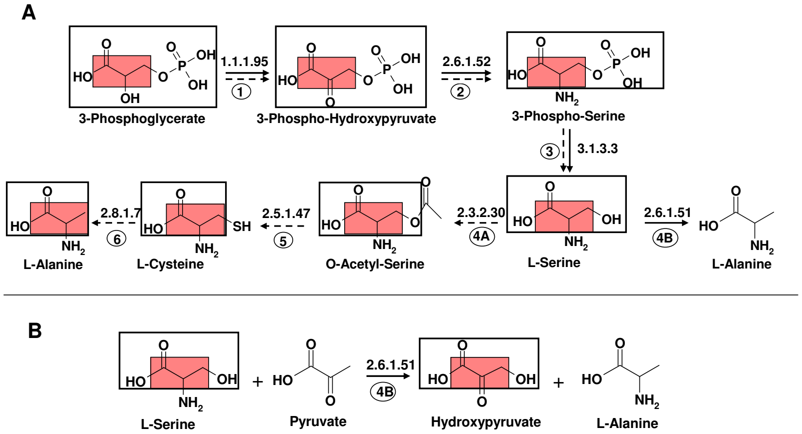


Figure 4.11: (A) Transformation of 3-phosphoglycerate to L-alanine. The dashed arrows represent a valid biotransformation route that conserves the structural moiety (shown using rectangles). No atom is transferred from 3-phosphoglycerate to L-alanine via L-serine in the route represented by solid arrows. (B) All atoms from 3-phosphoglycerate which are contained in L-serine are transferred to hydroxypyruvate according to the atom mapping rule of the reaction 2.6.1.51.

as final conversion step. However, the application of the atom mapping rule of that reaction implies that no atom could be transferred from 3-phosphoglycerate to L-alanine via L-serine. Hence, this path does not satisfy the structural moiety constraint.

Metabolic transition graph

Our graph representation of a metabolic network integrates pre-calculated atom mapping rules. Therefore, each reaction in the network is decomposed into a set of all possible educt/product pairs where at least one atom is transferred from the educt to the product according to the atom mapping rule of that reaction. Then each node in the graph represents a unique educt/product atom mapping pair (E_i, P_j). All reactions which have such a pair in common are associated with the corresponding node. This allows a more compact representation of reactions because frequent metabolic transitions like acetyl-CoA/CoA or glutamate/2-oxoglutarate, shared by several reactions, are summarized by a single node. The reverse reactions are represented by product/educt pair nodes. Each edge in the graph connects two nodes (E_i, P_j) and (E_k, P_l) if $P_j = E_k$ and if at least one atom is transferred from E_i to P_l according to the sequential application of the atom mapping rules of both nodes. Fig. 4.12 compares a transition graph with the common bipartite graph representation of an example network.

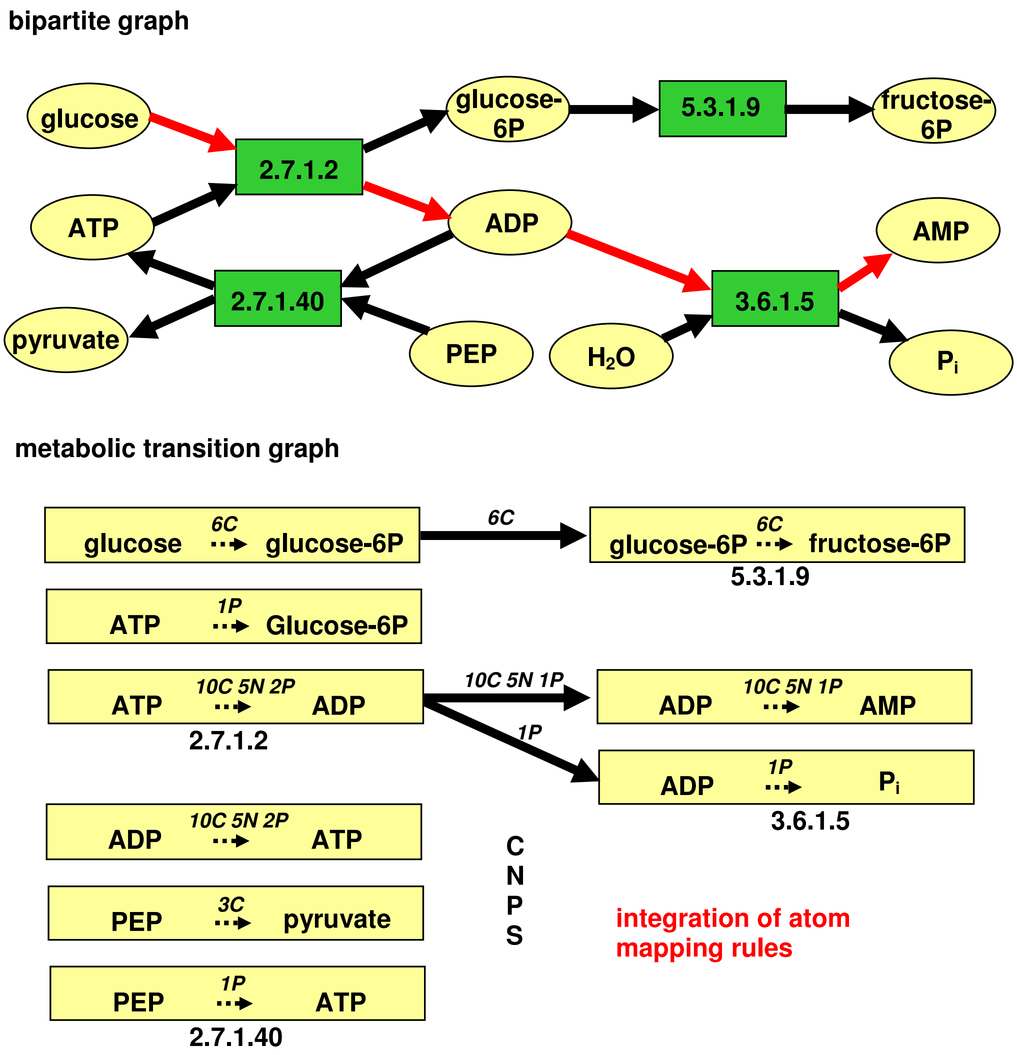


Figure 4.12: A bipartite graph (above) and a transition graph (bottom) representing a metabolic network that contains four reactions and corresponding metabolites. The transition graph integrates atom mapping rules and represents each biochemically feasible route that consists of two successive reaction steps by one edge. Irrelevant metabolic conversions like glucose → ADP → AMP, which are possible in the bipartite graph, are avoided.

Although this transition graph is more complex compared to other graph representations, it is more suited to our path-finding algorithm. Each biochemically feasible route consisting of two successive reaction steps is implicitly represented by one edge in the transition graph because each node codes one educt/product conversion. Therefore, the path-finding algorithm does not have to deal with a huge number of routes that contain irrelevant metabolic conversions like glucose → ADP → AMP because there will be no edge between the nodes representing the educt/product pairs glucose, ADP and ADP, AMP.

The computational complexity can be further reduced by the restriction to consider only one atom type when creating the nodes and edges of the transition graph. For example, if we are interested in questions concerning the carbon metabolism, we can simply ignore nodes and edges where no carbon atom is transferred. The same can be done for the nitrogen, sulfur or phosphorous metabolism.

Reaction context weight

As already described, the weighted metabolic networks approach [Croes *et al.*, 2006] fails to find routes involved in purine biosynthesis or routes passing the core metabolism (like glycolysis or TCA cycle) because frequently occurring compounds (like pyruvate or acetyl-coA) have to be traced. Therefore, we created an additional weight that also considers the context of the traced reactions as a counterpart to the weights derived from the network connectivities of the compounds (the compound weights). The context of a reaction contains all compounds of that reaction which are not used as intermediates in the path search. Each reaction $R_{k,(i,j)}$ associated with the transition node (E_i, P_j) gets a context weight for that node. The context consists of all educts $E_k, k \neq i$ and products $P_k, k \neq j$ of $R_{k,(i,j)}$. The weight of $R_{k,(i,j)}$ is the sum of the context weights of its context compounds. The higher/lower the number of reactions in the metabolic network a compound participates in, the lower/higher is its context weight. We used a function (based on piecewise linear interpolation) that maps the reaction count to the context weight and is defined as follows:

$$cw(i) = \begin{cases} 1 & \text{if } rc(i) > b_1 \\ 1 + \frac{(rc(i)-b_1)(b_3-1)}{b_2-b_1} & \text{if } b_2 \leq rc(i) \leq b_1 \\ b_3 + \frac{(rc(i)-b_2)(b_2-b_3)}{b_3-b_2} & \text{if } b_3 \leq rc(i) < b_2 \\ b_2 + \frac{(rc(i)-b_3)(b_1-b_2)}{1-b_3} & \text{if } 1 \leq rc(i) < b_3 \end{cases}$$

where:

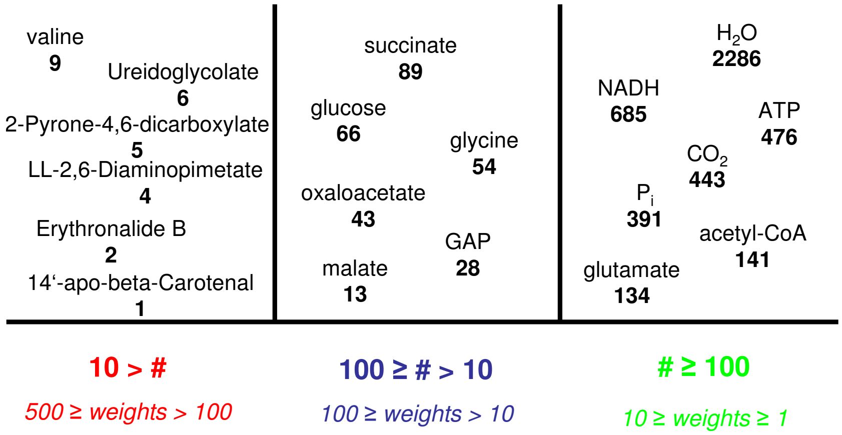


Figure 4.13: Motivation for using piecewise linear interpolation to transform compound frequencies (#) into context weights. The empirical bounds 10 and 100 separate the compounds in the KEGG network in frequent occurring compounds with context weights between 10 and one, in compounds of medium frequency with weights between 100 and 10 and in rare occurring compounds with weights between 500 and 100. The set of frequent occurring compounds contains typical pool metabolites like ATP participating in 476 KEGG reactions or acetyl-CoA participating in 141 reactions. Medium occurring compounds are typical intermediates of the core metabolism like succinate (with frequency 89) or GAP (frequency 28). Rare occurring compounds like LL-2,6-Diaminopimetate participate in less than ten reactions and are intermediates in very specialized or species-specific pathways.

- $cw(i)$: the context weight of compound i in the network

- $rc(i)$: the reaction count of compound i (number of reactions it is participating in as educt or product)

- b_1, b_2, b_3: empirical bounds which separate compounds in the network in frequent occurring compounds which will obtain low context weights, in compounds of medium frequency and in rare occurring compounds which will get high context weights (see also Fig. 4.13). For example, useful values are $b_1 = 500$, $b_2 = 100$, $b_3 = 10$ in case of the KEGG metabolic 'super network' with more than 5,000 distinct reactions or $b_1 = 100$, $b_2 = 20$, $b_3 = 2$ in case of the EcoCyc genome-scale metabolic network with more than 1,000 reactions. Since the EcoCyc network is approx. five times smaller, its empirical bounds are also reduced by factor five.

For example, the reactions EC 1.4.1.4 and EC 2.6.1.42 share the transformation of 2-oxoglutarate to glutamate where EC 1.4.1.4 uses NADPH, $NADP^+$, NH_3 and H_2O as co-substrates contrary to EC 2.6.1.42 using valine and 2-keto-isovalerate. The exclusive presence of pool metabolites in the context (all co-substrates) of EC 1.4.1.4 results in a significant lower weight compared to the weight for the context of EC 2.6.1.42. This weight makes sure that a biosynthesis route of glutamate via 2-oxoglutarate prefers to trace EC 1.4.1.4 instead of EC 2.6.1.42 which requires the production and consumption of further amino acids. The compound and reaction weights are incorporated into our transition graph where each edge represents the intermediate (metabolite) I_m of two subsequent educt/product transition nodes (E_i, I_m) and (I_m, P_j). The weight of each edge is assigned the number of reactions (in the network) in which I_m participates plus the minimum context weight of the reactions $R_{k,(m,j)}$ associated with the target transition node (I_m, P_j).

Note that a path in the graph can code multiple metabolic routes if more than one reaction is associated with at least one node in the path. The combined weighting is more suited to routes passing the core metabolism like glycolysis and the TCA cycle or routes of purine biosynthesis which involve 'hub compounds' like ADP as main intermediates.

Integration of thermodynamic information

The combined weighting scheme of the metabolic transition graph can be extended by the integration of thermodynamic information. Therefore, predicted standard transformed Gibbs energies $\Delta_r G'^0$ of the network reactions were transformed into positive energy weights. The biochemical meaning of this transformation derives from the assumption that cellular systems prefer to use the pathway, among a set of alternative ones, which is thermodynamically most efficient. Only positive energy weights are assigned to the edges in the graph. The use of negative weights for negative $\Delta_r G'^0$ and positive weights for positive $\Delta_r G'^0$ would make the path-finding problem very complicated. Furthermore, our preferred path-finding algorithm cannot handle negative edge weights. However, a simple shifting of the $\Delta_r G'^0$ range of values into a positive range would discriminate long pathways compared to very short but meaningless routes. Therefore, we applied a similar transformation like that used for creating the context weights. The transformation is defined by:

$$ew(k,i,j) = \begin{cases} 1 & \text{if } \Delta_r G'^0(k,i,j) < e_1 \\ 1 + \frac{(\Delta_r G'^0(k,i,j) - e_1)(b_3 - 1)}{e_2 - e_1} & \text{if } e_1 \leq \Delta_r G'^0(k,i,j) < e_2 \\ b_3 + \frac{(\Delta_r G'^0(k,i,j) - e_2)(b_2 - b_3)}{e_3 - e_2} & \text{if } e_2 \leq \Delta_r G'^0(k,i,j) < e_3 \\ b_2 + \frac{(\Delta_r G'^0(k,i,j) - e_3)(b_1 - b_2)}{e_4 - e_3} & \text{if } e_3 \leq \Delta_r G'^0(k,i,j) \leq e_4 \\ b_1 & \text{if } \Delta_r G'^0(k,i,j) > e_4 \end{cases}$$

where:

- $ew(k,i,j)$: the energy weight of reaction $R_{k,(i,j)}$
- $\Delta_r G'^0(k,i,j)$: the standard transformed Gibbs energy of reaction $R_{k,(i,j)}$
- e_1, e_2, e_3, e_4: empirical energy bounds which separate reactions in the network in reactions with negative or weak positive $\Delta_r G'^0$ which will obtain low energy weights, in reactions with medium positive $\Delta_r G'^0$ and in reactions with high positive $\Delta_r G'^0$ which will get high energy weights ($e_1 = -50$ kJ/mol, $e_2 = +10$ kJ/mol, $e_3 = +30$ kJ/mol, $e_4 = +50$ kJ/mol).
- b_1, b_2, b_3: empirical weight bounds which define the range of values of the energy weight mapping. Similar to the context weight definition, useful values are $b_1 = 250$, $b_2 = 100$, $b_3 = 10$ using the KEGG metabolic 'super network' or $b_1 = 50$, $b_2 = 20$, $b_3 = 2$ using the EcoCyc genome-scale metabolic network.

Reactions with very negative $\Delta_r G'^0$ receive very low energy weights compared to reactions with very high $\Delta_r G'^0$. All reactions for which no $\Delta_r G'^0$ can be calculated, are assigned the weight specified by b_3. These reactions contain compounds with incomplete structural information or with atoms other than C, O, N, P, S, H. The addition of the calculated energy weight of each reaction $R_{k,(i,j)}$ to its context weight finishes the integration of thermodynamic information into the weighting scheme.

Integration of subcellular localization information

The reactions of a metabolic network can be annotated with information about the subcellular localization of the protein monomers that build up the catalyzing enzymes. We can also use this information to extend the weighting schemes defined in the previous sections. The idea of integrating subcellular localization information derives from the assumption, that pairs of enzymes tend to be co-localized if they catalyze subsequent reaction steps within a metabolic pathway. We used experimentally verified localization information extracted from Swiss-Prot. If this information was not available, we used subcellular localizations predicted by MultiLoc2 (described in Section 4.3). Since the prediction of up to ten different eukaryotic subcellular localizations is a source of errors (see Section 4.3.3), virtual compartments were used. A virtual compartment consists of evolutionary related subcellular localizations that are difficult to discriminate by prediction methods like MultiLoc2. The virtual *other* compartment includes the nucleus, cytoplasm and peroxisomes. All subcellular localizations involved in the secretory pathway (ER, Golgi apparatus, lysosomes, vacuoles, plasma membrane and extracellular space) are summarized in the *SP* compartment. These two compartments are completed by mitochondrial and chloroplast organelles.

A further possibility to reduce the error rate of MultiLoc2 was to enable the prediction of multiple localizations. The idea is based on the fact that there are proteins or enzymes which operate in multiple localizations. MultiLoc2 does not directly predict multiple localizations. However, we assume that the probability estimates, calculated by MultiLoc2, are distributed more over multiple localizations compared to the probability estimates of proteins present in only a single localization. We assigned each localization, predicted with a MultiLoc2 score ≥ 0.25, to the query protein sequence. We used 0.25 as cut-off because this value would be assigned to the *other*, *SP*, mitochondrial and chloroplast compartment if the probability estimates are distributed equally. It should be noted, that the MultiLoc2 score for the virtual *other* and *SP* compartments are the sum of the MultiLoc2 scores calculated for their grouped single localizations.

For each pair of reactions $R_{k_1,(i,j)}$ and $R_{k_2,(j,l)}$ associated with the edge-connected transition nodes (E_i, P_j) and (E_j, P_l), we calculated a localization penalty weight which is defined by:

$$lpw(k_1,k_2) = \begin{cases} 0 & \text{if } |L_{k_1} \bigcap L_{k_2}| > 0 \text{ or } |L_{k_1}| = 0 \text{ or } |L_{k_2}| = 0 \\ p & \text{if } |L_{k_1} \bigcap L_{k_2}| = 0 \text{ and } |L_{k_1}| > 0 \text{ and } |L_{k_2}| > 0 \end{cases}$$

where:

- $lpw(k_1, k_2)$: the localization penalty weight for pairs of subsequent reactions $R_{k_1,(i,j)}$ and $R_{k_2,(j,l)}$
- L_{k_1}, L_{k_2}: the set of subcellular localizations assigned to reaction $R_{k_1,(i,j)}$ and $R_{k_2,(j,l)}$
- p: empirical weight which penalizes pairs of subsequent reactions in the network that do not share a common subcellular localization. Useful

values are $p = 100$ in case of the KEGG 'super network' or $p = 20$ for the EcoCyc network. Compared with the context and energy weights that are in the range between one and 500, single medium weights are used here as compromise.

Note that there are also reactions $R_{k,(i,j)}$ with no assigned subcellular localizations ($|L_k| = 0$). These reactions are either non-enzymatic and take place spontaneously or no enzymes as well as protein monomers are assigned to them so far. The final weight of each edge was assigned the minimum penalty weight of all reaction pairs associated with that edge.

Experimental setting

We performed several experiments for testing and comparing our approach. For this purpose, a bipartite graph and a metabolic transition graph were constructed from the EcoCyc and AraCyc databases. All reactions (1,348 and 1,284 respectively) from the small molecule metabolism were included representing the *E. coli* and *A. thaliana* metabolic networks at genome-scale. We investigated the search performances based on five different network types. The first four graph types were bipartite graphs used for comparison with the metabolic transition graph. The *blind search graph (bsg)* contains only the connectivity information extracted from the metabolic network. Using the *atom mapping graph (amg)*, pre-calculated atom mapping rules are available via the educt-reaction-product node relations and can be accessed in constant time. In the *weighted graph (wg)*, each edge representing a compound-reaction relation is assigned a weight equal to the connectivity of the compound in the whole metabolic network. This network type corresponds to the weighted metabolic networks approach [Croes *et al.*, 2006]. The only difference is that edges instead of nodes are assigned a weight. The *weighted atom mapping graph (wamg)* contains all of the available information as described for the other three network types. Finally, the *metabolic transition graph (mtg)* was used. The last two graph types represent our novel approach based on the integration of atom mapping rules and weighting schemes. The intension to test also the *blind search graph* and the *atom mapping graph* was to evaluate the impact of adding relevant information in the graph representation.

We always searched for feasible biotransformation routes between two given nodes (source and target). Using the blind search graph, feasible routes were found by searching for the shortest path. In the atom mapping graph, we searched for the shortest path that fulfills the structural moiety constraint. The lightest path was searched in the weighted graph. In the weighted atom mapping graph and in the metabolic transition graph, we searched for the lightest path fulfilling the structural moiety constraint. Paths between two given nodes were calculated using Eppstein's k-shortest path algorithm [Eppstein, 1998] which efficiently computes the first k-shortest or lightest paths in a directed graph. [8] Given two metabolites as source and target, we simply used the corresponding compound nodes for the search in the bipartite graphs. Using the metabolic transition graph, we had to create in each search a start node s and end node e representing the source metabolite E_s and product P_e where s was connected to all nodes (E_s, P_j) and e to (E_i, P_e). The weights of the edges connecting s and e with the graph were calculated as described in the section that defines the reaction context weight. Furthermore, the algorithm was adapted to consider the atom mapping rules. For this purpose, we used an analogous approach for path validation which was proposed by Arita [2003]. Each extracted path is validated by a sequential application of atom mapping rules. In the beginning, all atoms of the source metabolite are available for the mapping. After this, for each step, only those atoms of an educt are available for mapping to the next compound, which can be reached by a mapping in the step before. If no atom reaches the target metabolite, the path is rejected as not valid. Atom mapping rules were available for only 63% of the reactions (explained in Section 4.1). This fact is considered in the procedure. If a reaction without atom mapping rule is reached, the validation process is restarted with the next reaction that has an atom mapping rule. Hence, both the atom mapping graph and the metabolic transition graph can also find paths violating the structural moiety constraint. It should also be mentioned that oxygen and hydrogen atoms are ignored in the process. Hydrogen atoms are represented implicitly in the molecular graphs and not considered in the mapping calculation. Although oxygen atoms are considered in the mapping calculation, these atoms are ignored in the path validation process. The problem is that the water molecule is the most frequent pool metabolite and it is often impossible to detect a correct and unique mapping.

4.4.3 Results

The search performance of the presented network types and search strategies was evaluated by trying to find experimentally verified biotransformation routes in the metabolic networks of *E. coli* and *A. thaliana*. For this purpose, all annotated biotransformation routes of the small molecule metabolism with

[8] The algorithm creates an implicit representation of the k-lightest paths in a directed graph with n vertices and m edges in $\mathcal{O}(m + n \log n + kn)$, which can be traversed using breadth-first-search.

at least three reactions were extracted from EcoCyc and AraCyc (137 and 135 respectively).

It should be noted that, contrary to EcoCyc, the AraCyc database was computationally predicted using the pathway tools software [Karp *et al.*, 2002]. After this, AraCyc was manually curated and improved. However, AraCyc still contains a lot of pathways annotated with comments describing that the exact intermediate reaction steps are uncertain and remain to be validated experimentally. Since the EcoCyc database is of high quality, we used it as a gold standard. The AraCyc database was used only to evaluate the weighting scheme based on the integration of subcellular localization information. This was done because the established prediction method MultiLoc2 is based on eukaryotic training data only and outputs eukaryotic subcellular localizations.

Given the main source and target metabolites of the annotated routes as start and end nodes, we calculated the shortest (lightest) path constrained to use the first as well as the last reaction of the annotated route. If n annotated routes shared the same main source as well as target metabolites and start as well as end reaction, we computed the n shortest (lightest) paths. The quality of the routes found was measured by comparing the intermediate compounds and reactions with the annotated routes, and was expressed using sensitivity, specificity and relevance score, which are defined as follows:

$$sensitivity = \frac{tp}{tp+fn}$$
$$specificity = \frac{tp}{tp+fp}$$
$$relevance = \frac{sensitivity+specificity}{2} * smc$$

where:

- tp (true positives): The number of compounds and reactions of the route found which are also present in the annotated route. The first and last compounds and reactions are not considered.
- fp (false positives): The number of compounds and reactions of the route found which are not present in the annotated route.
- fn (false negatives): The number of compounds and reactions of the annotated route which are not present in the route found.
- smc (structural moiety constraint): This value is set to 1 if the route found fulfills the structural moiety constraint, and set to 0 otherwise.

If an extracted route was not identical to an annotated one and contained reactions without atom mapping rules, we manually checked the structural moiety constraint. Note that this evaluation procedure produces only relative performance measures useful for comparing different search strategies because novel routes could be very different compared to the annotated ones.

In case of the metabolic transition graph, we always used the carbon network (with carbon as the only traceable atom type) except for the sulfate reduction pathway (EcoCyc ID: SO4ASSIM-PWY) because sulfate and hydrogen sulfide were used as source and target metabolites. Here, we used the sulfur network (with sulfur as the only traceable atom type).

The search results are shown in Tab. 4.7. Searching for the shortest path in the blind search graph delivered poor search results. The average relevance score was only 0.31. Incorporating atom mapping rules for about two-thirds of the reactions in the graph doubled the search performance up to an relevance score of 0.61. A further improvement was achieved by searching for the lightest path in the weighted graph. This approach produced significantly more relevant routes. The relevance score was 0.77. But only 80% of the routes found fulfilled the structural moiety constraint which was clearly better in the atom mapping graph (+ 8%). The search for the lightest path in the weighted atom mapping graph further improved the search performance. The relevance score reached 0.86. Although atom mapping rules were available for only two-thirds of the reactions, 91% of the routes found fulfilled the structural moiety constraint, 11% more as for the weighted graph. The best search performance results with respect to all performance measures were produced by the metabolic transition graph. Compared to the weighted atom mapping graph, the performance was increased by approximately eight per cent. Now, nearly all of the extracted routes fulfilled the structural moiety constraint (99%). The routes especially of the core metabolism (glycolysis and TCA cycle) and the routes of the purine biosynthesis were better predicted.

In the next paragraph, we will demonstrate the search results of the three best approaches (*wg*, *wamg* and *mtg*) using glycolysis as an example.

Glycolysis

The biotransformation routes of glycolysis were searched given D-glucose-6-phosphate as source and pyruvate as target as well as EC 5.3.1.9 as start reaction and EC 2.7.1.40 as end reaction. Fig. 4.14A represents the two annotated routes extracted from EcoCyc. The first route (shown as red arrows) contains eight reactions. Three carbon atoms are transferred from the source to the target. An additional three atoms, resulting in a second molecule of pyruvate, are transferred via the second route (black arrows). This route contains dihydroxyacetone 3-phosphate (DHAP) as a further main interme-

Table 4.7: The search results for 137 experimentally verified biotransformation routes extracted from EcoCyc are shown here. The results of the verified routes present only in glycolysis, the TCA cycle and the purine biosynthesis are also shown. Each row represents one search approach: the blind search graph (bsg), the atom mapping graph (amg), the weighted graph (wg), the weighted atom mapping graph (wamg) and the metabolic transition graph (mtg). The columns show the average sensitivity (sens), specificity (spec), structural moiety constraint (smc) and relevance score (rel).

experiment	approach	sens	spec	smc	rel
all routes	**bsg**	0.34	0.41	0.47	0.31
	amg	0.61	0.66	0.88	0.61
	wg	0.82	0.87	0.80	0.77
	wag	0.86	0.87	0.91	0.86
	mtg	0.93	0.95	0.99	0.94
glycolysis	**wg**	0.37	0.79	0.00	0.00
	wag	0.73	0.80	1.00	0.77
	mtg	0.96	0.96	1.00	0.96
TCA cycle	**wg**	0.15	0.12	0.00	0.00
	wag	0.46	0.67	1.00	0.56
	mtg	1.00	1.00	1.00	1.00
purine syn.	**wg**	0.37	0.67	0.00	0.00
	wag	0.67	0.70	0.75	0.69
	mtg	0.93	0.94	1.00	0.94

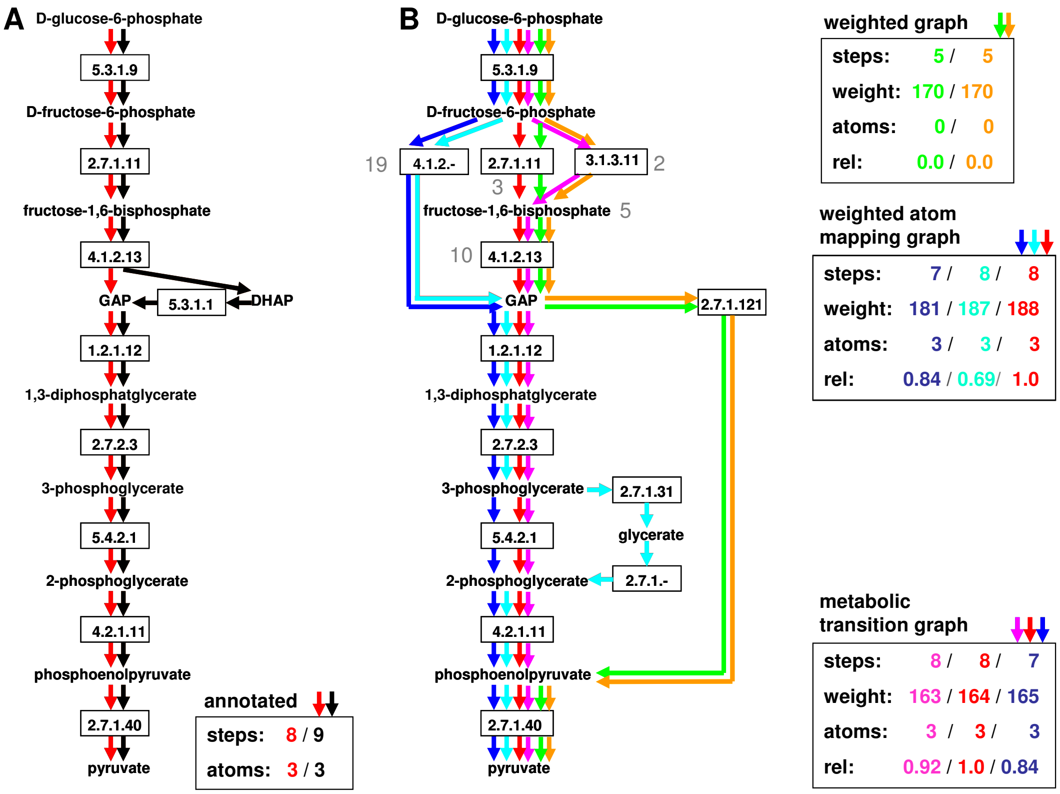


Figure 4.14: The annotated and the predicted routes for glycolysis, given D-glucose-6-phosphate as source metabolite, pyruvate as target and EC 5.3.1.9 as start reaction and EC 2.7.1.40 as end reaction. Different colors represent different routes. (A) The annotated routes extracted from EcoCyc. (B) The routes found using the weighted graph, the weighted atom mapping graph and the metabolic transition graph. For each graph type, the number of reaction steps, the overall weight, the number of transferred atoms from source to target, and the relevance (rel) of the routes found are shown.

diate which is transformed to D-glyceraldahyde 3-phosphate (GAP). All in all, the route contains nine reactions. The routes found by the lightest-path search in the graphs containing edge weights are shown in Fig. 4.14B. The first route (blue arrows), found by the weighted atom mapping graph, requires seven reactions, one and two less than the annotated routes, respectively. Once again three atoms are transferred from D-glucose-6-phosphate to pyruvate. The difference is that the route found needs only one reaction for transforming D-fructose-6-phosphate into GAP. The reaction with the EcoCyc ID RXN0-313 (EC 4.1.2.-) is very interesting since it is not assigned to a pathway in EcoCyc. The enzyme catalyzing this reaction is fructose-6-phosphate aldolase (gene name `fsa`) and was reported as a novel enzyme activity catalyzing an aldol cleavage of D-fructose-6-phosphate [Schürmann *et al.*, 2001]. The similarity to the standard glycolysis routes is reflected in a relevance score of 0.84. The second route (lightblue arrows), found by the *wamg* approach, also bypasses the annotated transformation of D-fructose-6-phosphate into GAP via fructose-6-phosphate aldolase. The difference is the alternative transformation of 3-phosphoglycerate into 2-phosphoglycerate via glycerate as an additional main intermediate. Three atoms are transferred again, but nine reactions are required. The relevance of this route is 0.69. However, reaction EC 2.7.1.31 is annotated in EcoCyc as physiologically favored in the opposite direction (glycerate as educt). The first annotated route (red arrows) was found on position three in the weighted atom mapping graph.

The first two routes found by the lightest-path search in the weighted graph are shown using green and orange arrows. Both routes contain only five reactions. However, no atom is transferred to pyruvate. The reaction with EcoCyc ID 2.7.1.121-RXN is responsible for the failed glycolysis reconstruction. The reaction transfers a phosphate group from DHAP to phosphoenolpyruvate. In the final reaction (EC 2.7.1.40), the phosphate group is cleaved so that no atom from DHAP can be transferred to pyruvate. The presence of the target pyruvate as an educt in reaction 2.7.1.121-RXN is a further reason for the irrelevance of this route. The failed reconstruction is reflected by an relevance score of 0.0.

The pink, the red and the blue routes were found by the metabolic transition graph. The top-ranked pink route is very similar to the glycolysis routes (0.92 relevance). The only difference is that EC 3.1.3.11 (fructose-1,6-bisphosphatase) is used instead of EC 2.7.1.11 (phosphofructokinase) to phosphorylate fructose-6-phosphate. However, EC 3.1.3.11 is known to be a key enzyme in the gluconeogenesis pathway for the conversion of fructose-1,6-bisphosphate to fructose-6-phosphate. The blue route, top-ranked using the *wamg* approach, was found one position after the red annotated route now. This is interesting because the blue route is shorter than the red route and contains only intermediary compounds that are also present in the red annotated route. The reason is the weighting used in the metabolic transition graph which combines the compound connectivity weight with the reaction context weight. Fig. 4.14B shows the weights of the relevant compounds and reactions drawn in grey. The total weight of EC 2.7.1.11, EC 4.1.2.13 and the intermediate fructose-1,6-bisphosphate is 18. This weight is slightly below the reaction context weight (19) of EC 4.1.2.- which is used in the alternative blue route. The context of EC 2.7.1.11[9] consists of the pool metabolites ATP and ADP. The reaction receives therefore a low context weight. The same is true for EC 3.1.3.11[10] with H_2O and phosphate in the reaction context. The context of EC 4.1.2.13[11] is given by DHAP and that of EC 4.1.2.-[12] by dihydroxyacetone. DHAP participates in 12 and dihydroxyacetone in only three reactions in the network. Compared to DHAP, dihydroxy-acetone is a rarely occurring compound in the pathways of the *E. coli* metabolism. Hence, EC 4.1.2.- receives a higher context weight compared to EC 4.1.2.13. The biological meaning, which is reflected in the context weight, is that using EC 4.1.2.13 instead of EC 4.1.2.- results in a higher probability that the occurred byproduct can be efficiently converted by other reactions.

Integration of thermodynamic information

To evaluate the impact on the search performance of the metabolic transition graph when integrating thermodynamic information into the weighting scheme, we ran three experiments. The setting of the first experiment was identical to that used in the section before. In experiment two and three we removed the constraint to only search for routes that pass through the first and last reaction of the annotated routes. This was done because using this constraint implies to a certain extent thermodynamic information and, therefore, makes the problem easier. The only difference of the third experiment compared to experiment two was that the most relevant route, among the first five routes extracted, was selected for performance evaluation. The results of the three experiments are summarized in Tab. 4.8. The result of experiment one was that the performance measures were only slightly improved. The already quite high relevance score of 0.94 was increased to 0.95. This was different in the other experiments where the search setting was more difficult. The final relevance scores are lower. However, the performance improvement

[9] D-fructose-6-phosphate + ATP $\rightleftharpoons$ D-fructose-1,6-bisphosphatase + ADP

[10] D-fructose-1,6-bisphosphatase + H_2O $\rightleftharpoons$ D-fructose-6-phosphate + phosphate

[11] D-fructose-1,6-bisphosphatase $\rightleftharpoons$ DHAP + GAP

[12] D-fructose-6-phosphate $\rightleftharpoons$ dihydroxyacetone + GAP

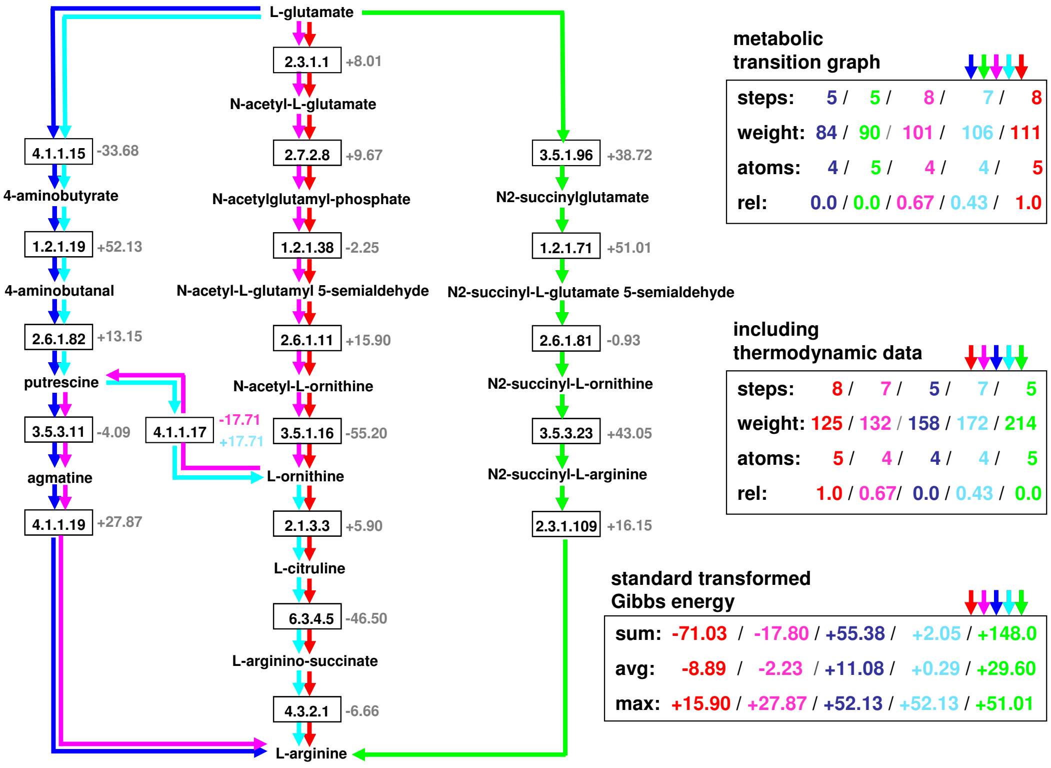
Table 4.8: The influence on the search results of the metabolic transition graph caused by integrating thermodynamic information into the weighting scheme is shown. Each row represents one search approach: the metabolic transition graph (mtg) and the metabolic transition graph with integrated thermodynamic information (mtg/T). The columns show the average sensitivity (sens), specificity (spec), structural moiety constraint (smc) and relevance score (rel).

experiment	approach	sens	spec	smc	rel
I	**mtg**	0.93	0.95	0.99	0.94
	mtg/T	0.95	0.96	0.99	0.95
II	**mtg**	0.65	0.72	0.98	0.68
	mtg/T	0.70	0.81	0.98	0.75
III	**mtg**	0.79	0.87	0.99	0.83
	mtg/T	0.85	0.93	0.99	0.89

was much more obvious when integrating thermodynamic information into the weighting scheme. The performance was significantly increased from 0.68 to 0.75 in experiment two and from 0.83 to 0.89 in experiment three. In the next paragraph, we describe the improvements based on thermodynamic information using arginine biosynthesis as an example.

Arginine biosynthesis

Arginine biosynthesis in *E. coli* (given L-glutamate as carbon source) is an interesting example because the reactions and intermediates of this pathway are completely different compared to the annotated arginine degradation pathway. Additionally, further alternative metabolic routes for converting L-glutamate into L-arginine are possible. Fig. 4.15 shows the first five routes found by the metabolic transition graph. The annotated route (shown in red) requires eight reaction steps for transferring five carbon atoms from L-glutamate to L-arginine. Using the standard metabolic transition graph without thermodynamic information, the annotated route was found on position five. The green route represents the reverse annotated arginine degradation pathway and was found on position two. The green route also transfers five carbon atoms but requires only five reaction steps. Also five reactions steps are used by the blue route which was ranked on position one. Furthermore, this route contains a completely different set of reactions and intermediates



metabolic
transition graph

steps:	5 /	5 /	8 /	7 /	8
weight:	84 /	90 /	101 /	106 /	111
atoms:	4 /	5 /	4 /	4 /	5
rel:	0.0 /	0.0 /	0.67 /	0.43 /	1.0

including
thermodynamic data

steps:	8 /	7 /	5 /	7 /	5
weight:	125 /	132 /	158 /	172 /	214
atoms:	5 /	4 /	4 /	4 /	5
rel:	1.0 /	0.67/	0.0 /	0.43 /	0.0

standard transformed
Gibbs energy

sum:	-71.03 /	-17.80 /	+55.38 /	+2.05 /	+148.0
avg:	-8.89 /	-2.23 /	+11.08 /	+0.29 /	+29.60
max:	+15.90 /	+27.87 /	+52.13 /	+52.13 /	+51.01

Figure 4.15: The predicted routes for arginine biosynthesis, given L-glutamate as source metabolite and arginine as target. Different colors represent different routes. The predicted $\Delta_r G'^0$ (in kJ/mol) of the reactions are drawn in grey. For each graph type, the number of reaction steps, the overall weight, the number of transferred atoms from source to target, and the relevance (rel) of the routes found are shown. Furthermore, for each route three scores which are calculated from the $\Delta_r G'^0$ values of the reactions involved are shown.

compared to the red and green routes. The pink and lightblue routes (ranked on position three and four) use parts of the annotated biosynthesis pathway and the blue route. Similar to the green route, the metabolic conversions of parts of the blue, the lightblue and the pink routes are physiologically more relevant in the reverse direction. For example, the reactions EC 1.2.1.19 and EC 2.6.1.82 are annotated in EcoCyc to be involved in the putrescine degradation. Further examples are the reactions EC 3.5.3.11 and EC 4.1.1.19 which are annotated to be also involved in the arginine degradation. Only the initial reaction of the blue top-ranked route takes place in the direction as indicated by the EcoCyc annotation as to be involved in glutamate degradation.

The ranking of these five routes was completely different when we integrated thermodynamic information in the form of predicted standard transformed Gibbs energy changes of the reactions. In Fig. 4.15 each reaction is annotated with its predicted standard transformed Gibbs energy $\Delta_r G'^0$ in kJ/mol. Now, the path-finding algorithm detects the annotated red route at first. The green and inverse arginine degradation route receives a significantly higher overall weight compared to the annotated route and drops from rank two to rank five. A further consequence is that the pink route which obtains a relevance score of 0.67 was found on position two. The blue and light blue routes were found on position three and four now. We also provide three interesting scores calculated for each route from the $\Delta_r G'^0$ values of its reactions in Fig. 4.15. The first score is simply the sum and the second the average of the energy values. The last score represents the maximum reaction energy within each route. An interesting result is that the relevance scores of the routes correlate quite well with the three energy scores. For example, the total energy of the red annotated route is -71.03 kJ/mol and is significantly lower compared to the values of the remaining routes. The reverse annotated arginine degradation pathway (green route) especially obtains a high positive energy sum of +148.0 kJ/mol. We also show the cumulative Gibbs energy landscape for the three routes that contain disjoint reaction sets (red, blue and green) in Fig. 4.16. This kind of a plot was also applied by [Chunhui *et al.*, 2004] for thermodynamic comparison of computationally created novel pathways. The energy landscape of the green reverse arginine degradation route clearly emphasizes that this pathway is thermodynamically not feasible. Although the blue biotransformation route starts with a thermodynamic very favorable reaction, it ends up clearly in the range of positive energy values. The first half of the red annotated route takes place in weak positive range. In the second half, however, the cumulative energy drops significantly into the negative range.

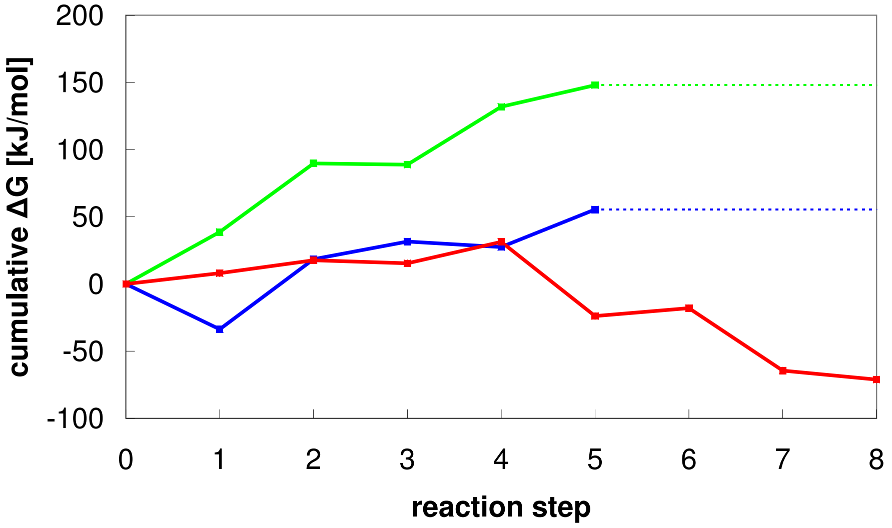


Figure 4.16: The landscape of the cumulative standard transformed Gibbs energy change for the three biotransformation routes (from glutamate to arginine) that contain disjoint reaction sets.

Integration of subcellular localization information

We used the same experimental setting as described in the previous section to evaluate the use of integrating subcellular localization information into the weighting scheme. As described earlier, we used *Arabidopsis thaliana* as a model organism instead of *Escherichia coli* because the subcellular localization information used is eukaryotic.

In each of the three experiments we compared the performance measures of the metabolic transition graph with respect to four different weighting schemes:

- mtg: the ordinary metabolic transition graph with integrated atom mapping rules, compound connectivity and context weight information.
- mtg/T: the metabolic transition graph with integrated thermodynamic information
- mtg/L: the metabolic transition graph with integrated subcellular localization information
- mtg/L/T: the metabolic transition graph with integrated subcellular localization and thermodynamic information

The results of the experiments are shown in Tab. 4.9. Although the number of reactions in the metabolic network and the number of annotated biotrans-

Table 4.9: Results for 135 experimentally verified biotransformation routes extracted from AraCyc. The table summarizes the influence on the search results of the metabolic transition graph caused by the integration of thermodynamic and subcellular localization information into the weighting scheme. Each row represents one search approach: the metabolic transition graph (mtg), the metabolic transition graph with integrated subcellular localization (mtg/L) as well as integrated thermodynamic information (mtg/T) and with the integration of both (mtg/L/T). The columns show the average sensitivity (sens), specificity (spec), structural moiety constraint (smc) and relevance score (rel).

experiment	approach	sens	spec	smc	rel
I	**mtg**	0.83	0.84	0.99	0.83
	mtg/L	0.83	0.82	0.99	0.83
	mtg/T	0.84	0.85	0.99	0.85
	mtg/L/T	0.85	0.85	0.99	0.85
II	**mtg**	0.52	0.61	0.97	0.56
	mtg/L	0.53	0.62	0.97	0.58
	mtg/T	0.56	0.67	0.98	0.61
	mtg/L/T	0.57	0.66	0.98	0.61
III	**mtg**	0.67	0.77	0.97	0.72
	mtg/L	0.68	0.78	0.97	0.73
	mtg/T	0.70	0.81	0.98	0.76
	mtg/L/T	0.70	0.81	0.98	0.76

formation routes extracted from AraCyc are comparable to those extracted from EcoCyc, the quality measures for the AraCyc data set were significantly lower compared to EcoCyc. This can be explained by the fact that AraCyc contains more noisy data, because the database was computationally predicted using the pathway tools software. After this, AraCyc was manually curated and improved. However, AraCyc still contains a lot of pathways annotated with comments that describe that the exact intermediate reaction steps are uncertain and remain to be validated experimentally. In each experiment, we tested different localization penalty weights (5, 10, 20, 40, 60). However, it was not possible to improve the relevance score of the mtg approach in experiment one by the addition of subcellular localization information (mtg/L). Using a penalty weight of 20 delivered a slightly increase of relevance from 0.56 to 0.58 in experiment two and from 0.72 to 0.73 in experiment three. The performance measures obtained by the integration of thermodynamic information (mtg/T) were significantly higher. Here, the relevance was increased from 0.83 to 0.85, from 0.56 to 0.61 and from 0.72 to 0.76 in experiment one, two and three. These improvements are comparable to those obtained using the EcoCyc data set. The simultaneous integration of subcellular localization and thermodynamic information led to the same relevance scores as those obtained by the integration of thermodynamic information alone in all experiments.

In the next paragraph, we will use the biosynthesis of DMAPP in *A. thaliana* as an interesting example to show that the consideration of subcellular localization information can still be useful when analyzing metabolic pathways.

DMAPP biosynthesis

The basic chemical units in isoprenoid biosynthesis are dimethylallyl diphosphate (DMAPP) and its isomer isopentenyl diphosphate (IPP). The biosynthesis of IPP and DMAPP is an interesting example because there are two different alternative routes in plants and bacteria [Bochar *et al.*, 1999; Rohmer, 1999]. These pathways operate in separated subcellular localizations [Lichtenthaler *et al.*, 1997; Kuzuyama *et al.*, 2003]. In *A. thaliana*, the methylerythritol phosphate pathway (MEP pathway) which is also called the mevalonate-independent or nonmevalonate pathway is localized completely in the chloroplasts. This differs from the mevalonate pathway (MVA pathway) which contains mevalonate as an intermediate and is referred to as cytoplasmic. 3-hydroxy-3-methylglutaryl-CoA (HMG-CoA) reductase (EC 1.1.1.34) represents an exception and is located within the endoplasmatic reticulum (ER) but has also been found within the cytoplasm [Leivar *et al.*, 2005]. However, two recent studies [Sapir-Mir *et al.*, 2008; Reumann *et al.*, 2007] reported experimental evidence regarding the peroxisomal localization for at least two enzymes (EC 2.3.1.9 and EC 5.3.3.2) of the MVA pathway. The additional presence of EC 5.3.3.2 in the chloroplasts was confirmed. This enzyme is the only one that is shared by the MVA and MEP pathways. The findings of both studies are based on more suited experimental techniques and are closer to the spatial organization observed in mammalian systems in which most of the MVA pathway enzymes are peroxisomal. Note that this literature discrepancy did not affect the weighting scheme of the search approach since cytoplasm and peroxisomes were grouped within the virtual “other” (OTH) compartment (described on page 93).

The annotated biotransformation routes of both pathways including the subcellular localizations involved are shown in Fig. 4.17. Given pyruvate as

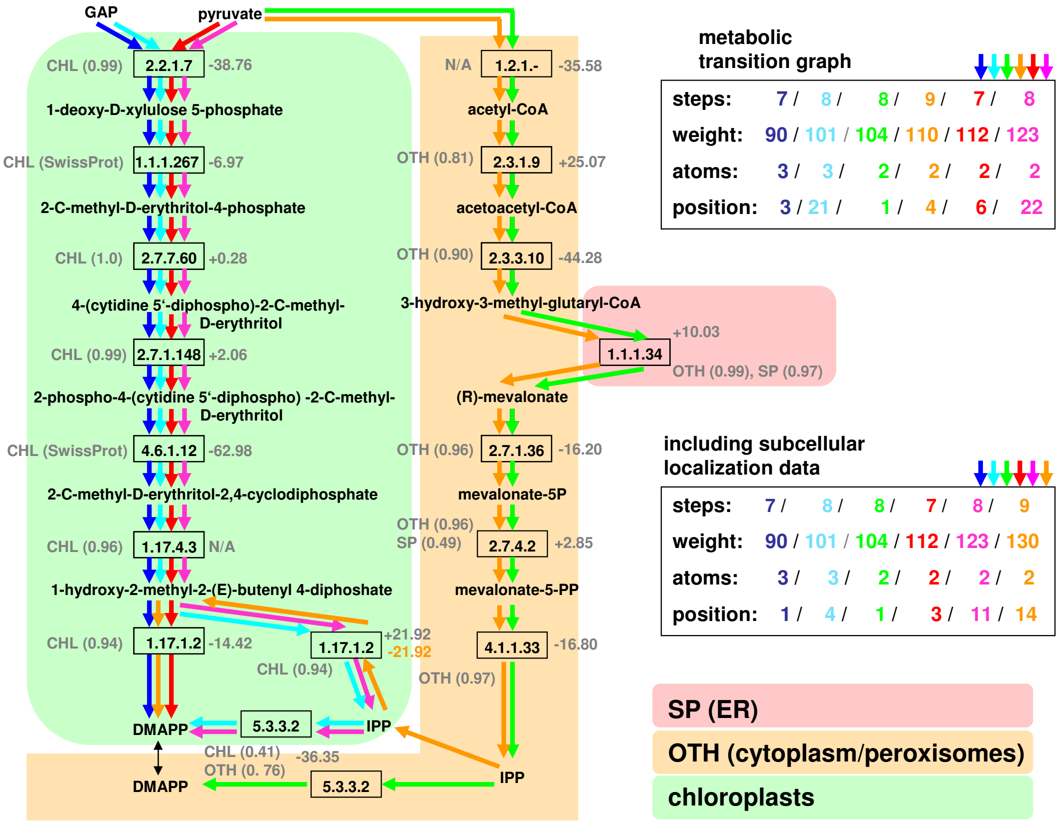

GAP
pyruvate
CHL (0.99) 2.2.1.7 -38.76
1-deoxy-D-xylulose 5-phosphate
CHL (SwissProt) 1.1.1.267 -6.97
2-C-methyl-D-erythritol-4-phosphate
CHL (1.0) 2.7.7.60 +0.28
4-(cytidine 5'-diphospho)-2-C-methyl-D-erythritol
CHL (0.99) 2.7.1.148 +2.06
2-phospho-4-(cytidine 5'-diphospho) -2-C-methyl-D-erythritol
CHL (SwissProt) 4.6.1.12 -62.98
2-C-methyl-D-erythritol-2,4-cyclodiphosphate
CHL (0.96) 1.17.4.3 N/A
1-hydroxy-2-methyl-2-(E)-butenyl 4-diphoshate
CHL (0.94) 1.17.1.2 -14.42
1.17.1.2 +21.92 -21.92
CHL (0.94)
DMAPP
5.3.3.2
IPP
CHL (0.41)
OTH (0. 76)
-36.35
DMAPP
5.3.3.2
IPP
N/A 1.2.1.- -35.58
acetyl-CoA
OTH (0.81) 2.3.1.9 +25.07
acetoacetyl-CoA
OTH (0.90) 2.3.3.10 -44.28
3-hydroxy-3-methyl-glutaryl-CoA
+10.03
1.1.1.34
OTH (0.99), SP (0.97)
(R)-mevalonate
OTH (0.96) 2.7.1.36 -16.20
mevalonate-5P
OTH (0.96)
SP (0.49)
2.7.4.2 +2.85
mevalonate-5-PP
4.1.1.33 -16.80
OTH (0.97)
metabolic transition graph
steps: 7 / 8 / 8 / 9 / 7 / 8
weight: 90 / 101 / 104 / 110 / 112 / 123
atoms: 3 / 3 / 2 / 2 / 2 / 2
position: 3 / 21 / 1 / 4 / 6 / 22
including subcellular localization data
steps: 7 / 8 / 8 / 7 / 8 / 9
weight: 90 / 101 / 104 / 112 / 123 / 130
atoms: 3 / 3 / 2 / 2 / 2 / 2
position: 1 / 4 / 1 / 3 / 11 / 14
SP (ER)
OTH (cytoplasm/peroxisomes)
chloroplasts

Figure 4.17: The annotated routes for DMAPP biosynthesis, given pyruvate as well GAP as source metabolite and DMAPP as target. Different colors represent different routes. Furthermore, the involved compartments are highlighted in green (chloroplasts), yellow (*other*) and red (secretory pathway). The predicted $\Delta_r G'^0$ (in kJ/mol) of the reactions and the predicted or SwissProt extracted subcellular localizations of the catalyzing enzymes are drawn in grey. For each graph type, the number of reaction steps, the overall weight, the number of transferred atoms from source to target, and the position of the routes found are shown.

carbon source, the MVA pathway is represented by the green route within the virtual other and SP compartments, shown on the right side of the diagram. The annotated MEP pathway is represented by the four routes within the chloroplast organelle on the left side. In addition to pyruvate, GAP is required as carbon source. The blue and lightblue routes transfer three carbon atoms each from GAP to DMAPP. Two carbon atoms are transferred from pyruvate via the red and pink routes. Contrary to the blue and red routes, the lightblue and pink routes contain IPP as intermediate. The last reaction (EC 5.3.3.2) of both routes converts IPP to DMAPP. This is also the final reaction step in the MVA pathway (green route). The orange route does not represent an annotated biotransformation route. It consists of the first seven reactions steps of the MVA pathway. The final IPP/DMAPP transformation of the MVA pathway is bypassed using the metabolic transformations catalyzed by MEP pathway enzyme EC 1.17.1.2. All reactions in the diagram are annotated with their predicted standard transformed Gibbs energy changes and the predicted or Swiss-Prot extracted subcellular localizations of the assigned protein monomers (according to AraCyc). Only two reactions in the MEP pathway are annotated with Swiss-Prot subcellular localization information. However, the predicted localizations with high probability estimates are consistent with the experimental localization observation of the MEP pathway. The predictions of the MVA pathway are also consistent because the pathway is known to be located outside the chloroplasts in the *other* as well as SP compartments. Using the ordinary mtg approach and given GAP as source compound, the annotated blue and lightblue routes were found on position three and 21. Including subcellular localization information, these routes were found on position one and four, which represents a significant improvement. The green route (MVA pathway) was found on position one using the mtg and the mtg/L search approach. The red and pink routes, with pyruvate as source, were found on position six and 22 using the mtg approach and found one position three and 11 including subcellular localization information. The non-annotated orange route drops from position four to 14 when using the mtg/L approach. The reason is that the orange route obtained a penalty weight because reaction EC 4.1.1.33 was predicted to be located in the *other* compartment (with score 0.97) and the subsequent enzyme EC 1.17.1.2 to be present in the chloroplasts (with score 0.94). Contrary to the annotated MVA pathway, the orange route requires the involvement of the chloroplasts as an additional compartment to take place.

4.4.4 Discussion

Based on atom mapping rules and weighting schemes, we introduced a novel graph theory-based approach for finding feasible biotransformation routes in metabolic networks. Constraining the lightest-path search to those paths where atoms are transferred between source and target nodes yielded improved predictions that are more consistent with experimentally verified biotransformation routes. Simply by checking sequentially the atom mapping rules of the transforming reactions, problematic routes like those present in glycolysis or purine synthesis were found better. The approach is generally more robust for biotransformation routes of the core metabolism or routes containing typical pool metabolites as intermediates compared to the ordinary lightest-path search in a degree-weighted graph. Further improvements were achieved by the integration of relevant information based on reaction context, thermodynamic and subcellular localization into the edge weights of the graph.

The combined use of the reaction context and thermodynamic weights makes sure that the path-finding algorithm selects the most plausible reaction among a set of candidate reactions that all convert the same pair of metabolites. To this end, reactions resulting in a negative change in standard transformed Gibbs energy and with no rare occurring side educts or products are preferred. Furthermore, the introduced metabolic transition graph with its integrated atom mapping rules and its specialized carbon, nitrogen, sulfur and phosphorus networks avoids the calculation of biochemically invalid routes to a very high extent.

The impact of using subcellular localization on the search performance was low compared to the remaining information. There are several reasons which may contribute to this somewhat disappointing result. As already mentioned, AraCyc contains noisy data. The exact assignment of reactions as well as genes and proteins to the individual pathways is uncertain in many cases. Only approximately 70% of the AraCyc reactions are assigned to genes. Further uncertainties come from failed MultiLoc2 predictions. Furthermore, the metabolic transition graph with its integrated atom mapping rules, compound and reaction weights as well as thermodynamic information already represents a lot of relevant biological meaning. Therefore, further improvements of the search performance are difficult. Providing integrated localization data in the context of metabolic networks, however, gives potential users of our approach interesting information about the organellar distribution of the catalyzing enzymes present in the pathways under study.

An open challenge for this approach is how to select optimally the parameters used to calculate the context weight, the thermodynamic weight and the localization penalty weight or how to integrate optimally these weights in the combined weight. A solution could be based on a cross-validation procedure that systematically analyzes pre-defined value ranges for all weighting parameters.

We believe that our approach could be used to complement existing approaches. It can bridge, for example, the gap between the raw genome-scale content stored in pathway databases and well curated local metabolic (sub-)networks necessary for applications in metabolic engineering. Fast and intelligent navigation through the network at genome-scale enables a goal-oriented refinement of the search by an iterative addition of constraints. Such constraints contain the identification of side metabolites and the sets of allowed, required or forbidden main intermediates as well as reactions. The search results can help scientists for designing experiments or biotechnologists for defining the constraints necessary for an efficient calculation of stoichiometrically balanced pathways using approaches based on computationally hard convex-analysis [Schuster *et al.*, 1999; Schilling *et al.*, 2000]. Too many constraints at the beginning of the analysis run the risk of missing relevant pathways, which is avoided by our approach.

4.5 Network navigation and visualization

4.5.1 Introduction

In this section, we present MetaRoute, a user-friendly tool for exploring metabolic networks. MetaRoute offers web access on efficient (and thus interactive) graph theory-based search and navigation through genome-scale metabolic networks combined with an easy-to-use visualization of the search results. Compared with other graph theory-based tools [Rahman *et al.*, 2005; Croes *et al.*, 2006; Klukas *et al.*, 2007] MetaRoute offers interactive speed, cross-species comparison, and the dynamic retrieval of local networks. It is possible to search in user-defined networks, in the network of a particular species or in that of multiple organisms based on data from the KEGG database [Kanehisa, 1996]. Another related web tool is MetaPath Online [Handorf *et al.*, 2007], an implementation of the network expansion approach that delivers information on which products can be synthesized in principle from a given set of seed compounds by the calculation of an expanded network. Additionally, the shortest route to a particular product can be extracted from such a network. This is different from MetaRoute where up to 500 different routes (between a source and a product) of increasing weight and size can be calculated and combined into a local network. This allows a systematic enumeration and a very compact representation of alternative routes. By enforcing biochemical constraints, MetaRoute strongly favors 'textbook pathways' over less relevant pathways. MetaRoute is available online at *http://www-bs.informatik.uni-tuebingen.de/Services/MetaRoute*.

4.5.2 Methods

The search engine of MetaRoute is an implementation of the graph theory-based approach which was described in Section 4.4. The basic algorithm uses atom mapping rules and the combined compound and reaction context weighting schemes to search for relevant paths in a directed graph representing the metabolic network.

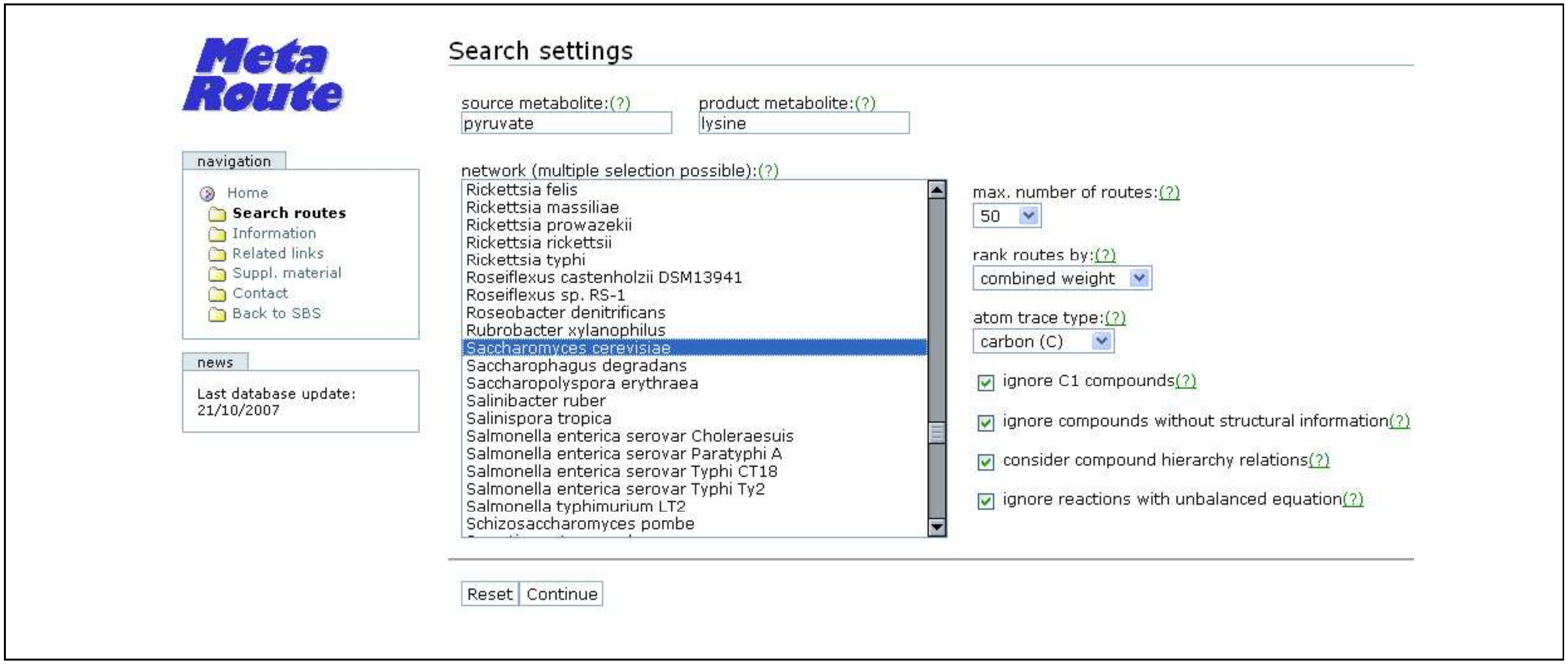
MetaRoute is built upon BNDB [Küntzer *et al.*, 2007] and BN++ [Sirava *et al.*, 2002; Küntzer *et al.*, 2006] and uses the metabolic data imported from KEGG [Kanehisa, 1996]. All reactions (approx. 7,000) and the compounds involved therein were used to create a metabolic 'super network'. Organism-specific networks are constructed by removing reactions catalyzed by enzymes absent in that organism.

We use the graph visualization software Graphviz (http://www.graphviz.org) for drawing the search results. Additional information within the search results is always available via small popup boxes. To this end, we use overLIB (http://www.bosrup.com/web/overlib/), a JavaScript library developed by Erik Bosrup.

4.5.3 Applications

The main application of MetaRoute is the exploration of genome-scale metabolic networks by the search for relevant routes transforming a source metabolite into a product.

Before starting a search, several search settings (shown in Fig. 4.18) can be defined by the user. However, only three (*source metabolite*, *product metabolite* and *network*) are necessary. The user can enter the name of the metabolite which will be used as source. Alternatively, the corresponding KEGG identifier (e.g. C00022) can be entered instead. However, it is also possible to leave the field empty and to search for metabolic routes with an unspecified source but that produce a given product metabolite. The product metabolite field is accordingly used.



Last database update: 21/10/2007

Search settings

source metabolite:(?) pyruvate

product metabolite:(?) lysine

network (multiple selection possible):(?)

Rickettsia felis
Rickettsia massiliae
Rickettsia prowazekii
Rickettsia rickettsii
Rickettsia typhi
Roseiflexus castenholzii DSM13941
Roseiflexus sp. RS-1
Roseobacter denitrificans
Rubrobacter xylanophilus
Saccharomyces cerevisiae
Saccharophagus degradans
Saccharopolyspora erythraea
Salinibacter ruber
Salinispora tropica
Salmonella enterica serovar Choleraesuis
Salmonella enterica serovar Paratyphi A
Salmonella enterica serovar Typhi CT18
Salmonella enterica serovar Typhi Ty2
Salmonella typhimurium LT2
Schizosaccharomyces pombe

max. number of routes:(?) 50

rank routes by:(?) combined weight

atom trace type:(?) carbon (C)

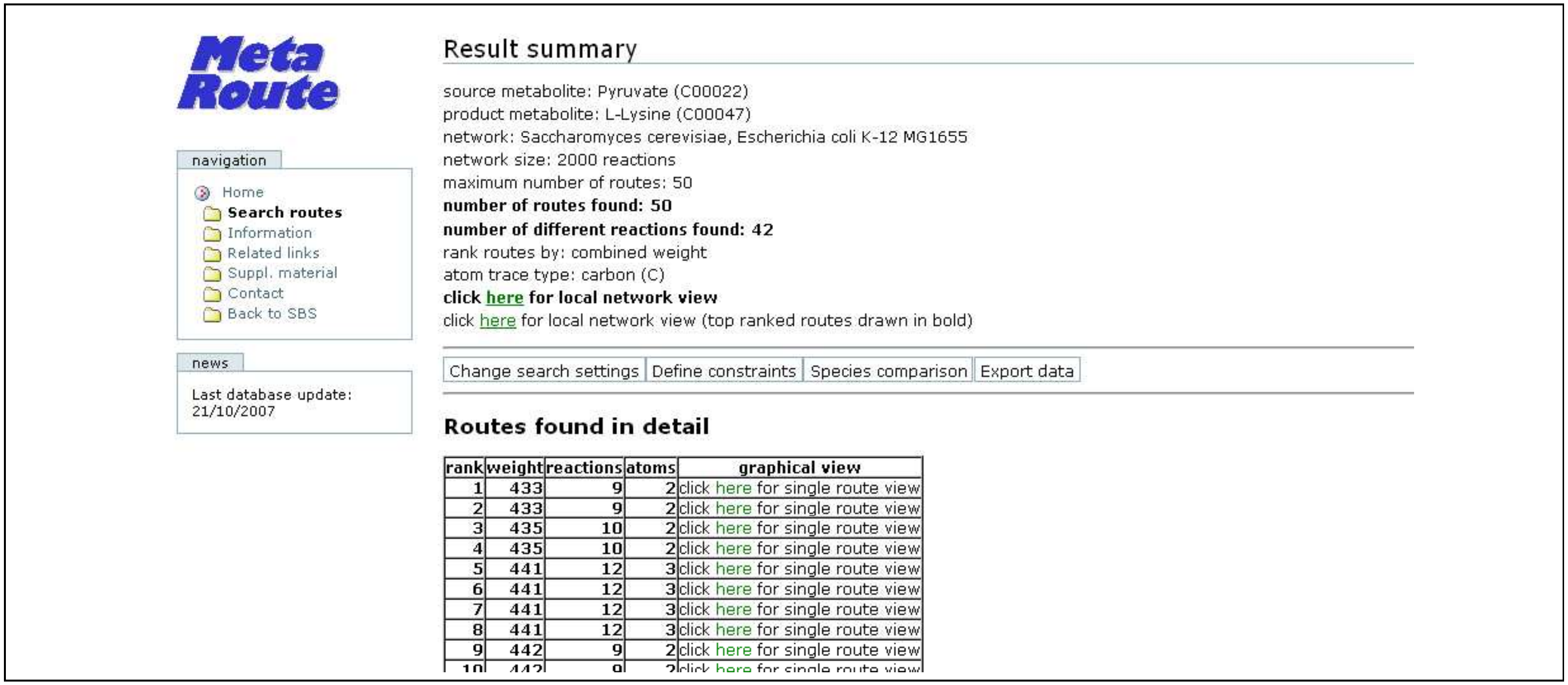
☑ ignore C1 compounds(?)

☑ ignore compounds without structural information(?)

☑ consider compound hierarchy relations(?)

☑ ignore reactions with unbalanced equation(?)

Figure 4.18: A screenshot showing the MetaRoute search settings window. The user has to define several search settings before MetaRoute can perform a search job.



Last database update:
21/10/2007

Result summary

source metabolite: Pyruvate (C00022)
product metabolite: L-Lysine (C00047)
network: Saccharomyces cerevisiae, Escherichia coli K-12 MG1655
network size: 2000 reactions
maximum number of routes: 50
number of routes found: 50
number of different reactions found: 42
rank routes by: combined weight
atom trace type: carbon (C)
click here for local network view
click here for local network view (top ranked routes drawn in bold)

Routes found in detail

rank	weight	reactions	atoms	graphical view
1	433	9	2	click here for single route view
2	433	9	2	click here for single route view
3	435	10	2	click here for single route view
4	435	10	2	click here for single route view
5	441	12	3	click here for single route view
6	441	12	3	click here for single route view
7	441	12	3	click here for single route view
8	441	12	3	click here for single route view
9	442	9	2	click here for single route view
10	442	9	2	click here for single route view

Figure 4.19: After the calculation of a search request, the results are presented in a summary page.

It is possible to search for relevant routes in the metabolic network of one organism of interest or in the combined network of multiple organisms, which is interesting for cross-species comparison. Furthermore, the user can search in the 'super network' containing all KEGG reactions as well as in an external or user-defined network. In the latter case, a file has to be submitted that contains a valid KEGG reaction ID (e.g. R00200) per line and at least 10 reactions overall. The remaining settings are pre-defined using default values which can be changed by experienced users.

As an example, we can search for lysine-producing routes with pyruvate as source in the combined network of yeast and *E. coli*. This is interesting because lysine biosynthesis is very different in each organism.

After the calculation of the submitted job is finished, the results are presented in a summary page (see Fig. 4.19). The summary page lists the user-defined search settings and shows useful information about the selected network size, the number of routes found and the number of distinct KEGG reactions within the routes.

After the first run, the user can refine the search by the addition of arbitrary constraints. Compounds and reactions can be marked as forbidden or required in the routes found. Furthermore, reactions can be defined by experienced users as irreversible (left-to-right or right-to-left). All reactions are set reversible by default. The reason is that detailed knowledge of thermodynamic data and physiological conditions is required in order to decide whether a reaction is irreversible.

All routes found can be visualized simultaneously in a local network showing only the main metabolites which were traced in the path-finding algorithm. If multiple organisms are selected (as in our example), the user can assign them to two groups which are differently highlighted for species comparisons. The resulting network for our search example, including the species highlighting, is shown in Fig. 4.20A. Another option is to visualize each route found in a separate window (see Fig. 4.20B). This time, for each reaction, the side metabolites are also drawn. Image maps offer direct links to the KEGG database, but the user also immediately gets more information in a small popup box simply by moving the mouse over the compounds and reactions (see Fig. 4.21).

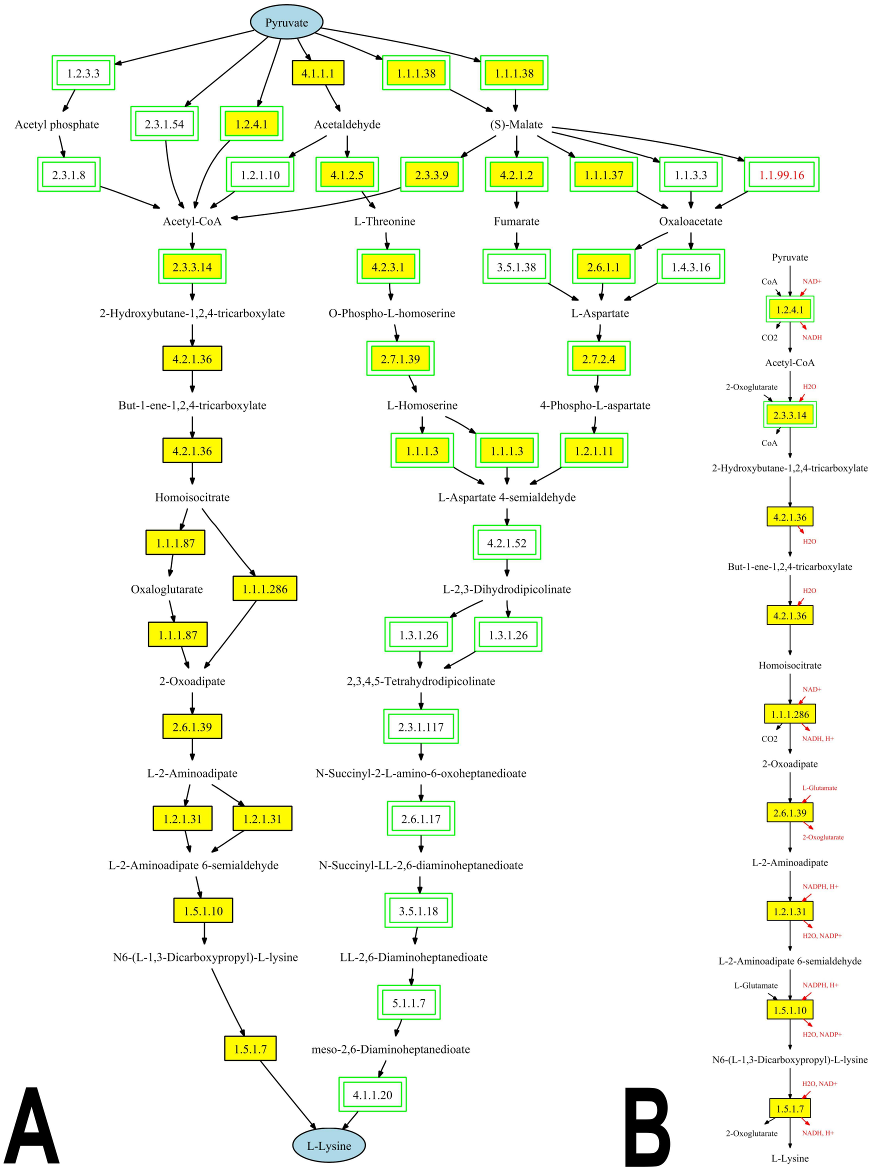
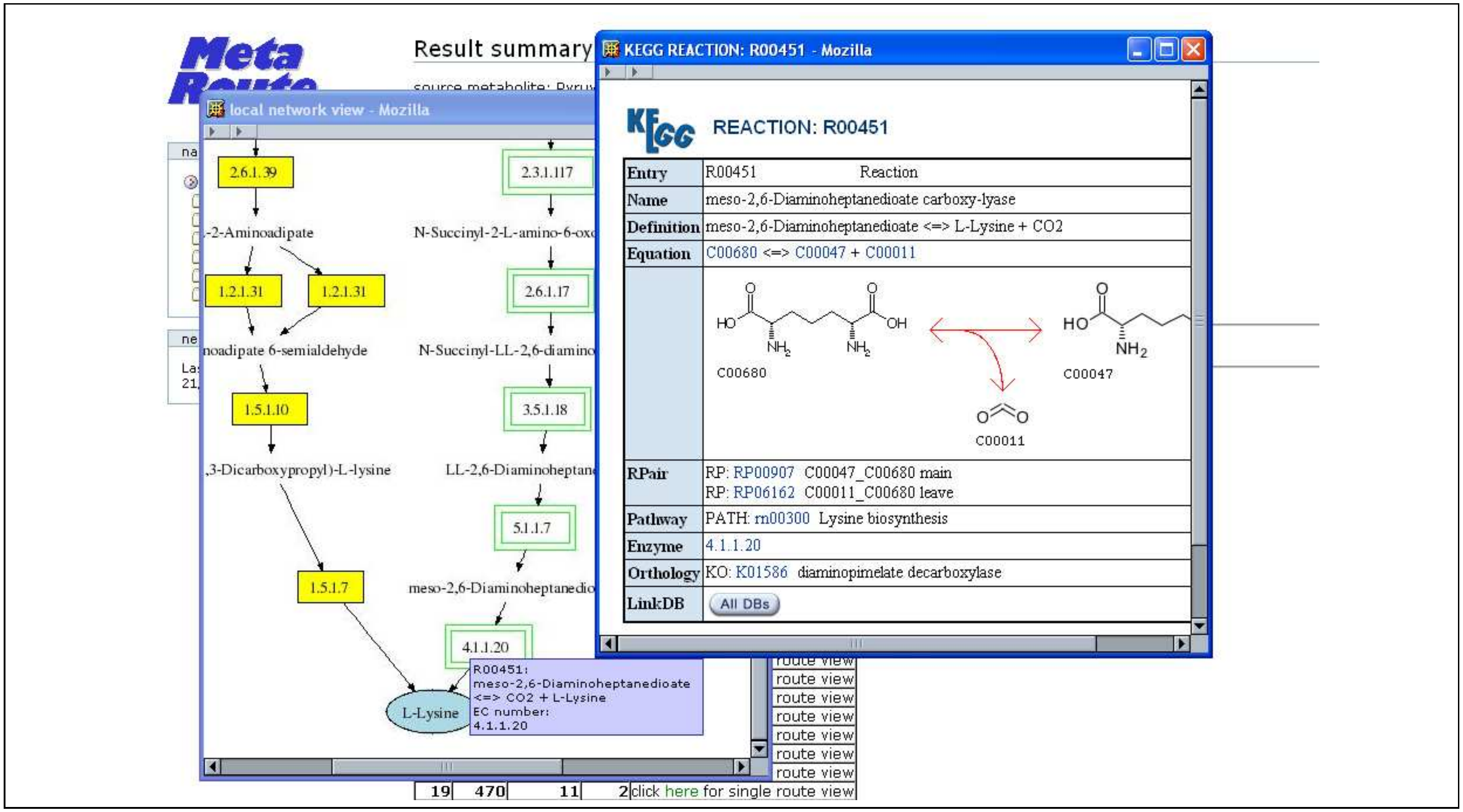


Figure 4.20: An example of the local network view is shown (A). Enzymatic reactions present in yeast are drawn in yellow boxes and those in *E. coli* in green double borders. Furthermore, an example of the single route view is shown (B). Here, the side metabolites are also drawn.



R00451:
meso-2,6-Diaminoheptanedioate
<=> CO2 + L-Lysine
EC number:
4.1.1.20

REACTION: R00451

Entry	R00451 Reaction
Name	meso-2,6-Diaminoheptanedioate carboxy-lyase
Definition	meso-2,6-Diaminoheptanedioate <=> L-Lysine + CO2
Equation	C00680 <=> C00047 + C00011
RPair	RP: RP00907 C00047_C00680 main RP: RP06162 C00011_C00680 leave
Pathway	PATH: rn00300 Lysine biosynthesis
Enzyme	4.1.1.20
Orthology	KO: K01586 diaminopimelate decarboxylase
LinkDB	All DBs

Figure 4.21: Pupup boxes and direct links to the KEGG database provide additional information to the user.

Besides its network exploration capabilities, MetaRoute can be used to complement related approaches. For example, Metatool [Kamp *et al.*, 2006], which is based on the elementary flux mode approach, allows sophisticated pathway analysis. However, the approach represents a computationally hard problem and cannot be used for genome-scale networks. Furthermore, a pre-defined distinction between internal and external (pool) metabolites is necessary. Using MetaRoute, the user can select a local network of moderate size and export it in the Metatool format including an automated definition of external compounds.

4.5.4 Conclusions

MetaRoute is a user-friendly tool for exploring genome-scale metabolic networks. It integrates efficient search for relevant routes with interactive network navigation and visualization. Due to the weighting scheme and the application of atom mapping rules in the path-finding approach, the resulting routes are close to textbook-like routes. Search results can be exported in various formats, e.g. Systems Biology Markup Language (SBML) or Metatool format [Kamp *et al.*, 2006] for further processing or analysis.

Potential applications of MetaRoute contain the systematic analysis of the KEGG database which includes the detection of novel or alternative pathways. In general, it can be used to support the design of knock-out or tracer experiments. Further applications exist in biotechnology (metabolic engineering) or biomedicine (drug target identification). MetaRoute can also complement other methods like elementary flux mode analysis where it can be used to identify (local) networks of moderate size as well as internal and external compounds.

Chapter 5

Conclusions

In this thesis, several computational approaches that support the analysis of metabolic pathways were developed. Each of these approaches provides useful applications in systems biology and their integration represents a way to deal with the complexity of metabolic networks. The main approach is based on graph theory and enables an efficient search for relevant biotransformation routes in genome-scale metabolic networks and it is available to the community via a web interface called MetaRoute. Compared to related work for analyzing metabolic pathways, a major step forward was the idea of combining pre-calculated atom mapping rules with the shortest-path search in a weighted graph representing the metabolic network. This was possible by introducing a novel graph representation as well as weighting scheme and a method for the automatic calculation of atom mapping rules that solves the problems of a previous approach. The weighting scheme was extended by the integration of further relevant information associated with biochemical reactions and their catalyzing enzymes. To this end, we developed two further novel approaches in the course of this thesis. The first predicts the standard transformed Gibbs energy changes of reactions from the chemical structures of educts and products and the second predicts the subcellular localizations given enzymatic amino acid sequences. The second method, called MultiLoc2, can also be accessed online as a web server.

Applying the approaches of this thesis to the genome-scale metabolic networks of *Escherichia coli* and *Arabidopsis thaliana*, in order to detect biologically meaningful pathways, showed encouraging results. However, there is still room for improvement, which is due to the fact that the interplay of different biochemical entities is very complex and also depends on environmental influences or the changing needs of the organism.

In future work, it should be possible to integrate further relevant information into the graph representation of metabolic networks and the path-finding algorithm for a more comprehensive analysis. Such information could consider pathway regulation or the evolutionary distance of enzymes, for example, in the form of phylogenetic profiles. As already described in Chapter 2, feedback inhibition and feed-forward activation are typical biochemical concepts that regulate or control the stability of metabolic pathways. Therefore, the presence of one or several of these concepts could indicate meaningful pathways. Phylogenetic profiles were already used in this thesis to improve the prediction of subcellular protein localization. This kind of information could also be applied to select, from a list of alternative pathways, the one with the highest degree of co-inherited enzymes during evolution. A suitable definition of such a degree remains to be specified but could be based on a pairwise comparison of phylogenetic profiles. Also, the incorporation of experimentally gained metabolomics or expression data extracted, for example, from public databases like the Golm Metabolome Database (GMD) [Kopka *et al.*, 2005] or Gene Expression Omnibus (GEO) [Barrett *et al.*, 2007], could lead to better results.

Although MetaRoute provides an easy-to-use visualization of the search results, the analysis of metabolic pathways based on complex networks and heterogeneous data requires more sophisticated visualization capabilities. The visual inspection of meaningful routes and local networks inferred by scientists can be significantly improved by, for example, visualizing the flow of atoms based on atom mapping rules, by highlighting the subcellular localizations of the involved enzymes or by highlighting potential thermodynamic bottlenecks with very positive $\Delta_r G'^0$ as well as more favorable reactions with very negative $\Delta_r G'^0$. Furthermore, suitable graph-layout algorithms that, for example, center top-ranked routes within a local network could offer attractive visual complements.

One nice feature of MetaRoute is that inferred local networks can be exported in a format that enables direct elementary flux mode analysis of these networks using external tools. The integration of an automatic elementary flux mode analysis of inferred local networks using efficient graph theory-based concepts into a joint method is desirable since this would enable the exploitation of the advantages of both worlds.

Having available meaningful pathways or flux modes and (predicted) standard transformed Gibbs energy changes of reactions, makes possible an embedded calculation of the equilibrium composition of the intermediates based on specified initial concentrations of the source compounds, using, for example, an iterative approach such that suggested by Alberty [2006cb]. This could support biotechnological applications which aim to increase the yield of commercially interesting compounds.

Appendix A

Mining standard transformed Gibbs energies for biochemical reactions

The result of estimating $\Delta_r G'^0$ for biochemical reactions from experimental data using the data mining approach, described on page 63, is shown in Tab. A.6. For each estimated $\Delta_r G'^0$, the table shows the relevant data of the automatically selected TECRDB entries, i.e. the EC number, temperature T, pH, ionic strength I if given, measured equilibrium constant K' and the subjective evaluation rate EV. The approach produces $\Delta_r G'^0$ values for 73 biochemical reactions where each reaction contains more than one compound with unknown $\Delta_f G'^0$. Additional information concerning detailed reaction equations is available in Tab. A.7.

Table A.1: Result of the mined $\Delta_r G'^0$ values for 73 biochemical reactions. For each $\Delta_r G'^0$, all relevant parameters of the automatically selected TECRDB entries are presented

reaction (EC)	T	pH	I/M	K'	$\Delta_r G'^0$	Ev
1.1.1.1b	298.15	7.5	(0.25)	17.17	C	
1.1.1.1	298.15	8.0	(0.25)	0.2	3.99	B
1.1.1.3	298.15	7.9	(0.25)	0.00063	18.27	C
1.1.1.3b	298.15	7.9	(0.25)	0.0088	11.73	C
1.1.1.4	300.15	7.4	(0.25)	0.00723	12.22	B
1.1.1.9	298.15	7.0	(0.25)	6.91×10^{-5}	23.75	B
1.1.1.14	298.15	7.0	(0.25)	0.013	10.76	C
1.1.1.25	303.15	7.0	(0.25)	0.0361	8.23	B
1.1.1.30	298.15	7.0	(0.25)	0.0146	10.48	A
1.1.1.31	298.15	8.0	(0.25)	0.0031	14.32	B
1.1.1.35	298.15	7.0	0.25	0.00025	20.56	C
1.1.1.37	298.15	7.5	(0.25)	5.3×10^{-5}	24.40	C
1.1.1.50	298.15	7.0	(0.25)	0.058	7.06	B
	298.15	7.0	(0.25)	0.092	5.91	B
1.1.1.61	298.15	7.1	(0.25)	3.9	-3.37	C
1.1.1.62	298.15	7.0	(0.25)	0.18	4.25	C
1.1.1.69	303.15	7.5	(0.25)	0.00011	22.59	C
1.1.1.97	303.15	7.6	(0.25)	0.18	4.25	C
1.1.1.108	303.15	7.0	(0.25)	0.00013	22.18	B
1.1.1.108b	295.15	8.0	(0.25)	0.00022	20.88	B
1.1.1.129	298.15	7.0	(0.25)	0.000342	19.78	C
1.1.1.150	303.15	6.9	(0.25)	7.8×10^{-8}	40.57	B
1.1.1.153	298.15	7.6	(0.25)	0.13	5.06	B
1.1.1.194	303.15	7.8	(0.25)	0.18	4.25	C
1.1.99.3	293.15	7.03	(0.25)	0.000723	17.93	B
1.3.99.11	293.15	7.2	(0.25)	0.00619	12.60	B
1.4.1.11	299.15	7.0	(0.25)	0.004	13.69	A
1.5.1.1	298.15	7.9	(0.25)	0.0036	13.95	B
1.5.1.3	298.15	7.5	(0.25)	6.1×10^{-5}	24.06	B
1.5.1.3b	295.15	7.0	(0.25)	19.4	-7.35	C
1.5.1.5	298.15	6.9	(0.25)	0.14	4.87	B
1.5.1.11	298.15	7.0	(0.25)	3×10^{-6}	31.52	C
1.8.1.4	298.15	6.87	0.25	0.138	4.91	A
	298.15	6.89	0.25	0.13	5.06	A
	298.15	7.08	0.25	0.267	3.27	A
	298.15	7.09	0.25	0.27	3.25	A
1.8.1.4b	295.15	6.9	(0.25)	0.18	4.25	B
	295.15	7.2	(0.25)	0.28	3.16	B

Table A.6: Continued

reaction (EC)	T	pH	I/M	K'	$\Delta_r G'^0$	Ev
2.1.2.1	303.15	7.3	(0.25)	0.125	5.15	A
2.1.2.4	298.15	7.0	(0.25)	3.1	-2.80	B
2.1.2.5	298.15	6.7	(0.25)	1.3	-0.65	B
2.2.1.1	298.15	7.6	(0.25)	0.015	10.41	B
2.3.1.2	299.15	7.0	(0.25)	0.011	11.17	B
	299.15	7.2	(0.25)	0.0099	11.44	B
2.3.1.6	298.15	7.03	0.25	11.7	-6.10	A
2.3.1.7	298.15	7.0	0.25	1.6	-1.17	A
2.3.1.8	303.15	6.85	(0.25)	0.14	4.87	B
2.4.1.67	298.15	6.5	(0.25)	4.0	-3.44	C
2.4.1.120	303.15	6.0	(0.25)	0.21	3.87	C
2.4.2.1	298.15	7.4	(0.25)	0.00036	19.66	C
2.4.2.10	301.15	8.0	(0.25)	1.4	-0.83	B
2.7.3.4	303.15	7.1	(0.25)	0.53	1.57	C
2.7.4.2	303.15	1.7	(0.25)	1.7	-1.32	B
2.7.4.14	303.15	7.5	(0.25)	1.49	-0.99	B
2.7.7.24	298.15	8.0	(0.25)	0.67	0.99	B
2.7.3.4	303.15	7.1	(0.25)	0.53	1.57	C
2.7.4.2	303.15	8.0	(0.25)	1.7	-1.32	B
2.7.4.14	303.15	7.5	(0.25)	1.49	-0.99	B
2.7.7.24	298.15	8.0	(0.25)	0.67	0.99	B
3.5.1.11	298.15	6.0	(0.25)	0.02	9.70	B
3.5.2.3	303.15	6.1	(0.25)	1.9	-1.59	B
3.5.4.9	298.15	7.0	(0.25)	11.0	-5.94	B
4.1.2.18	301.15	7.4	(0.25)	0.00037	19.59	B
4.1.2.18b	301.15	7.5	(0.25)	0.00012	22.38	B
4.1.3.3	298.15	7.5	(0.25)	0.034	8.38	A
	298.15	7.5	(0.25)	0.0348	8.32	A
4.1.3.32	298.15	8.0	(0.25)	0.5	1.72	B
4.2.1.10	302.15	7.4	(0.25)	15.0	-6.71	C
4.2.1.17	298.15	7.5	(0.25)	0.29)	3.07	C
4.2.1.49	298.15	7.5	(0.25)	69.8	-10.52	A
4.2.1.85	298.15	7.0	0.1	0.089	6.00	B
4.3.1.2	298.15	7.9	(0.25)	0.238	3.56	B
4.3.1.3	298.15	8.0	(0.25)	3.0	-2.73	C
4.4.1.5	303.15	7.0	(0.25)	9×10^{-5}	23.09	C
5.1.1.5	303.15	8.0	(0.25)	1.0	0.0	B
5.1.3.5	303.15	8.0	(0.25)	1.25	-0.55	B
5.1.3.6	298.15	7.5	(0.25)	2.6	-2.37	B
5.1.3.8	298.15	7.5	(0.25)	0.201	3.98	A
5.4.2.2	303.15	7.11	(0.25)	0.28	3.16	C
5.4.3.2	303.15	7.7	(0.25)	5.3	-4.13	C
5.4.99.6	298.15	7.5	(0.25)	0.66	1.03	A
5.5.1.1	303.15	7.5	(0.25)	0.041	7.92	B
	303.15	6.5	(0.25)	0.011	11.18	B
5.5.1.3	296.15	7.5	(0.25)	620.0	-15.94	C
5.5.1.6	298.15	7.6	(0.25)	7.6	-5.03	C

Table A.7: This table shows EC number and reaction equation for each mined biochemical reaction.

EC	reaction equation
1.1.1.1	cyclohexanol + NAD = cyclohexanone + NADH
1.1.1.1b	benzyl alcohol + NAD = benzaldehyde + NADH
1.1.1.3	L-homoserine + NADP = L-aspartate 4-semialdehyde + NADPH
1.1.1.3b	L-homoserine + NAD = L-aspartate 4-semialdehyde + NADH
1.1.1.4	(R,R)-2,3-butanediol + NAD = (R)-acetoin + NADH
1.1.1.9	L-threitol + NAD = L-erythrulose + NADH
1.1.1.14	galactitol + NAD = D-tagatose + NADH
1.1.1.25	shikimate + NADP = 5-dehydroshikimate + NADPH
1.1.1.30	(R)-3-hydroxybutanoate + NAD = 3-oxobutanoate + NADH
1.1.1.31	3-hydroxy-2-methylpropanoate + NAD = 2-methyl-3-oxopropanoate + NADH
1.1.1.35	(S)-3-hydroxyhexanoyl-CoA + NAD = 3-oxohexanoyl-CoA + NADH
1.1.1.37	meso-tartrate + NAD = (E)-dihydroxyfumarate + NADH
1.1.1.50	5-alpha-androstane-3alpha-ol-17-one + NAD = 5-alpha-androstane-3,17-dione + NADH
1.1.1.61	4-hydroxybutanoate + NAD = 4-oxobutanoate + NADH
1.1.1.62	estradiol-17-beta + NAD = estrone + NADH
1.1.1.69	D-gluconate + NADP = 5-oxo-D-gluconate + NADPH
1.1.1.97	3-hydroxybenzyl alcohol + NADP = 3-hydroxybenzaldehyde + NADPH
1.1.1.108	L-carnitine + NAD = 3-dehydrocarnitine + NADH
1.1.1.108b	D-carnitine + NAD = 3-dehydrocarnitine + NADH
1.1.1.129	L-threonate + NAD = 3-oxo-L-threonate + NADH
1.1.1.150	4-pregnene-11beta,17alpha,21-triol-3,20-dione + NAD = 4-pregnene-11beta,17alpha-diol-3,20,21-trione+ NADH
1.1.1.153	7,8-dihydrobiopterin + NADP = sepiapterin + NADPH
1.1.1.194	coniferyl alcohol + NADP = coniferyl aldehyde + NADPH
1.1.99.3	D-gluconate + NADP = 2-oxo-D-gluconate + NADPH
1.3.99.11	(S)-dihydroorotate + NAD = orotate + NADH
1.4.1.11	L-erythro-3,5-diaminohexanoate + NAD + H_2O = (S)-5-amino-3-oxohexanoate + NADH + ammonia
1.5.1.1	(S)-proline + NADP = D-pyrroline-2-carboxylate + NADPH
1.5.1.3	5,6,7,8-tetrahydrofolate + NADP = 7,8-dihydrofolate + NADPH
1.5.1.3b	2 7,8-dihydrofolate = folate + 5,6,7,8-tetrahydrofolate
1.5.1.5	5,10-methylenetetrahydrofolate + NADP = 5,10-methenyltetrahydrofolate + NADPH
1.5.1.11	N_2-(D-1-carboxyethyl)-L-arginine + NAD + H_2O = L-arginine + pyruvate + NADH
1.8.1.4	dihydro-alpha-lipoate + NAD = alpha-lipoate + NADH
1.8.1.4b	dihydrolipoamide + NAD = lipoamide + NADH
2.1.2.1	5,10-methylenetetrahydrofolate + glycine + H_2O = tetrahydrofolate + L-serine
2.1.2.4	5-formiminotetrahydrofolate + glycine = N-formiminoglycine + tetrahydrofolate
2.1.2.5	5-formiminotetrahydrofolate + L-glutamate = N-formimino-L-glutamate + tetrahydrofolate
2.2.1.1	D-fructose 6-phosphate + glycolaldehyde = L-erythrulose + D-erythrose 4-phosphate

Table A.7: Continued

EC	reaction equation
2.3.1.2	acetyl phosphate + imidazole = N-acetylimidazole + orthophosphate
2.3.1.6	acetyl-CoA + choline = CoA + O-acetylcholine
2.3.1.7	acetyl-CoA + L-carnitine = CoA + L-acetylcarnitine
2.3.1.8	formyl-CoA + orthophosphate = CoA + formyl phosphate
2.4.1.67	1-alpha-D-galactosyl-myo-inositol + raffinose = myo-inositol + stachyose
2.4.1.120	UDP-glucose + sinapate = UDP + 1-sinapoyl-D-glucose
2.4.2.1	nicotinamide + alpha-D-ribose 1-phosphate = nicotinamide riboside + orthophosphate
2.4.2.10	orotidine 5'-phosphate + pyrophosphate = orotate + 5-phospho-alpha-D-ribose 1-diphosphate
2.7.3.4	ATP + taurocyamine = ADP + N-phosphotaurocyamine
2.7.4.2	ATP + (R)-5-phosphomevalonate = ADP + (R)-5-diphosphomevalonate
2.7.4.14	ATP + dCMP = ADP + dCDP
2.7.7.24	dTTP + alpha-D-glucose 1-phosphate = dTDPglucose + pyrophosphate
3.5.1.11	penicillin G + H_2O = 6-aminopenicillanic acid + phenylacetic acid
3.5.2.3	(S)-dihydroorotate + H_2O = N-carbamoyl-L-aspartate
3.5.4.9	5,10-methenyltetrahydrofolate + H_2O = 10-formyltetrahydrofolate
4.1.2.18	2-dehydro-3-deoxy-L-pentonate = pyruvate + glycolaldehyde
4.1.2.18b	2-dehydro-3-deoxy-D-fuconate = pyruvate + (S)-lactaldehyde
4.1.3.3	N-acetylneuraminate = N-acetyl-D-mannosamine + pyruvate
4.1.3.32	2,3-dimethylmalate = propanoate + pyruvate
4.2.1.10	3-dehydroquinate = 3-dehydroshikimate + H_2O
4.2.1.17	(3S)-3-hydroxybutanoyl-CoA = trans-but-2-enoyl-CoA + H_2O
4.2.1.49	urocanate + H_2O = 4,5-dihydro-4-oxo-5-imidazolepropanoate
4.2.1.85	(2R,3S)-2,3-dimethylmalate = dimethylmaleate + H_2O
4.3.1.2	L-threo-3-methylaspartate = 2-methylfumarate + ammonia
4.3.1.3	L-histidine = urocanate + ammonia
4.4.1.5	(R)-S-lactoylglutathione = glutathione (reduced) + methylglyoxal
5.1.1.5	L-lysine = D-lysine
5.1.3.6	UDP-D-glucuronate = UDP-D-galacturonate
5.1.3.8	N-acetyl-D-glucosamine = N-acetyl-D-mannosamine
5.4.2.2	D-glucosamine 6-phosphate = D-glucosamine 1-phosphate
5.4.3.2	L-lysine = (3S)-3,6-diaminohexanoate
5.4.99.6	chorismate = isochorismate
5.5.1.1	2,5-dihydro-5-oxofuran-2-acetate = cis-cis-hexadienedioate
5.5.1.3	tetrahydroxypteridine = xanthine-8-carboxylate
5.1.3.5	UDP-L-arabinose = UDP-D-xylose
5.5.1.6	2',4,4'-trihydroxychalcone = (2S)-4',7-dihydroxyflavanone

Appendix B

MultiLoc2

B.1 Data used for the development of MultiLoc2

B.1.1 Phylogenetic profiles

We downloaded 453 fully sequenced genomes from the National Center for Biotechnology Information (NCBI) ftp site (*ftp.ncbi.nih.gov/genomes*) consisting of 20 eukaryotes, 33 archaea and 400 bacteria. However, only 78 genomes were used for the calculation of phylogenetic profiles. We used all eukaryotic and archaea genomes and selected only the 25 genetically most distant bacteria genomes in order to get an approximately equal distribution of genomes from the three kingdoms. We used the same genome subselection procedure as described in Sun *et al.*, 2005. The method uses the NCBI taxonomy information to reconstruct an evolutionary tree and exploits hierarchical information in a top down approach to select a preferably non-redundant set of genomes. The complete set of the genomes used is listed in Tab. B.1, Tab. B.2 and Tab. B.3.

Table B.1: The selected fully sequenced eukaryotic genomes used for the calculation of phylogenetic profiles.

Taxonomy ID	Organism name
3702	Arabidopsis thaliana
4932	Saccharomyces cerevisiae
5693	Trypanosoma cruzi
6239	Caenorhabditis elegans
7227	Drosophila melanogaster
9606	Homo sapiens
10090	Mus musculus
33169	Eremothecium gossypii
35128	Thalassiosira pseudonana
36329	Plasmodium falciparum 3D7
39947	Oryza sativa Japonica Group
214684	Cryptococcus neoformans var. neoformans JEC21
280699	Cyanidioschyzon merolae strain 10D
284590	Kluyveromyces lactis NRRL Y-1140
284591	Yarrowia lipolytica CLIB122
284592	Debaryomyces hansenii CBS767
284593	Candida glabrata CBS 138
284812	Schizosaccharomyces pombe 972h-
284813	Encephalitozoon cuniculi GB-M1
294381	Entamoeba histolytica HM-1:IMSS

Table B.2: The selected fully sequenced archaea genomes used for the calculation of phylogenetic profiles.

Taxonomy ID	Organism name
64091	Halobacterium sp. NRC-1
69014	Thermococcus kodakarensis KOD1
70601	Pyrococcus horikoshii OT3
178306	Pyrobaculum aerophilum str. IM2
186497	Pyrococcus furiosus DSM 3638
187420	Methanothermobacter thermautotrophicus str. Delta H
188937	Methanosarcina acetivorans C2A
190192	Methanopyrus kandleri AV19
192952	Methanosarcina mazei Go1
224325	Archaeoglobus fulgidus DSM 4304
228908	Nanoarchaeum equitans Kin4-M
243232	Methanocaldococcus jannaschii DSM 2661
259564	Methanococcoides burtonii DSM 6242
263820	Picrophilus torridus DSM 9790
267377	Methanococcus maripaludis S2
269797	Methanosarcina barkeri str. Fusaro
272557	Aeropyrum pernix K1
272569	Haloarcula marismortui ATCC 43049
272844	Pyrococcus abyssi GE5
273057	Sulfolobus solfataricus P2
273063	Sulfolobus tokodaii str. 7
273075	Thermoplasma acidophilum DSM 1728
273116	Thermoplasma volcanium GSS1
323259	Methanospirillum hungatei JF-1
330779	Sulfolobus acidocaldarius DSM 639
339860	Methanosphaera stadtmanae DSM 3091
348780	Natronomonas pharaonis DSM 2160
349307	Methanosaeta thermophila PT
362976	Haloquadratum walsbyi DSM 16790
368408	Thermofilum pendens Hrk 5
384616	Pyrobaculum islandicum DSM 4184
410358	Methanocorpusculum labreanum Z
415426	Hyperthermus butylicus DSM 5456

Table B.3: The selected fully sequenced bacteria genomes used for the calculation of phylogenetic profiles.

Taxonomy ID	Organism name
1140	Synechococcus elongatus PCC 7942
1148	Synechocystis sp. PCC 6803
59920	Prochlorococcus marinus str. NATL2A
60480	Shewanella sp. MR-4
62928	Azoarcus sp. BH72
62977	Acinetobacter sp. ADP1
64471	Synechococcus sp. CC9311
103690	Nostoc sp. PCC 7120
156889	Magnetococcus sp. MC-1
197221	Thermosynechococcus elongatus BP-1
203124	Trichodesmium erythraeum IMS101
232721	Acidovorax sp. JS42
240292	Anabaena variabilis ATCC 29413
243164	Dehalococcoides ethenogenes 195
251221	Gloeobacter violaceus PCC 7421
255470	Dehalococcoides sp. CBDB1
266779	Mesorhizobium sp. BNC1
290400	Jannaschia sp. CCS1
292414	Silicibacter sp. TM1040
292459	Symbiobacterium thermophilum IAM 14863
296591	Polaromonas sp. JS666
326442	Pseudoalteromonas haloplanktis TAC125
374463	Baumannia cicadellinicola str. Hc (Homalodisca coagulata)
387662	Candidatus Carsonella ruddii PV
413404	Candidatus Ruthia magnifica str. Cm (Calyptogena magnifica)

B.2 MultiLoc2-LowRes architecture

The architecture of the animal version of MultiLoc2-LowRes is shown in Fig. B.1. Compared with MultiLoc2-HighRes the SVMSA subpredictor is not used because MultiLoc2-LowRes is specialized for globular proteins.

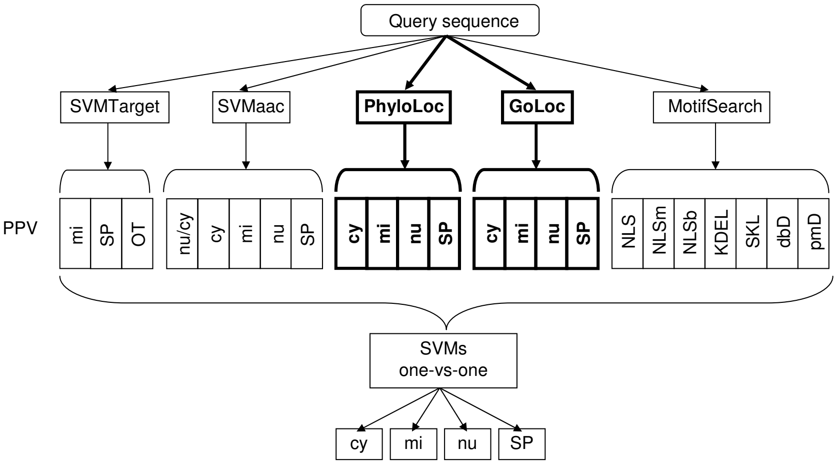


Figure B.1: The architecture of MultiLoc2-LowRes (animal version). A query sequence is processed by a first layer of five subprediction methods (SVMTarget, SVMaac, PhyloLoc, GOLoc and MotifSearch). The individual output of the layer one methods are collected in the PPV which enters a second layer of SVMs producing probability estimates for each localization.

B.3 Independent test without GO terms

The results of the simulation that no GO terms are available for all proteins of the independent data set are presented in this section. Tab. B.4 shows the localization-specific performance results using sensitivity and MCC and Tab. B.5 summarizes the overall performances using AVG and ACC.

B.3.1 MultiLoc2-LowRes

The animal prediction performance of MultiLoc2-LowRes is reduced by only one per cent regarding to AVG and ACC when predicting three classes and by two and four percent when predicting four classes. The reason is that more nuclear proteins are wrongly predicted if we discard the GO terms. The fungal prediction performance is almost unchanged which is mainly caused by the fact that on average only 34% of the fungal proteins are annotated with GO terms by InterProScan. The plant ACCs are decreased from 83% to 80% and from 76% to 71% for the prediction of three classes and four classes respectively. This is caused by the dropping sensitivity of the nuclear proteins (from 91% to 77%). The AVGs are reduced by nine per cent which seems to be a very significant performance lost at the first view. The reason is that the SP sensitivity is reduced from 83% to 50%. Only two SP proteins are additionally wrong predicted if we neglect the GO annotation. However, these two proteins have a large impact on the AVGs because the SP cluster contains only six proteins overall.

B.3.2 MultiLoc2-HighRes

Similar to the MultiLoc2-LowRes, the fungal prediction performance of MultiLoc2-HighRes is almost unchanged. This is the same for the plants in case of the prediction of four classes. The performance reduction by three per cent for the prediction of five plant classes is also moderate. However, very different to MultiLoc2-LowRes, the animal ACCs are reduced by nine percent and 11% respectively. We analyzed the additionally wrong predicted proteins and found out that this was caused by a failure in the clustering procedure performed by the curators of the data set [Casadio *et al.*, 2008]. The nuclear data set contains 56 proteins of the protamine-P1 family. Each protein represents one cluster which biases the prediction towards this over-represented protein class. The reason for the failed clustering are obviously the relatively short sequences of the proteins between 50 and 60 amino acids. Therefore, we reclustered the nuclear proteins using BLASTClust and 30% sequence identity. Now, the 56 proteins of the protamine-P1 family are clustered and the new number of clusters is 186 for the nuclear proteins and 277 for the nu/cy class. The comparison of the animal results based on the reclustered nuclear proteins delivers only a slightly performance reduction. We also applied BLASTClust on all other localizations and always received either the same number of clusters or a few more which indicates that the described clustering problem did not appear for the remaining classes.

Table B.4: Comparison of the localization-specific prediction results of the MultiLoc2 predictors using an independent dataset

Version	**Loc**	**Nr**	**MultiLoc2-LR**		**MultiLoc2-LR***		**MultiLoc2-HR**		**MultiLoc2-HR***	
			SE	MCC	SE	MCC	SE	MCC	SE	MCC
Animals	SP	75	97	0.89	97	0.88	87	0.79	88	0.60
	mi	48	89	0.81	86	0.78	83	0.75	83	0.74
	nu	224	62	0.57	56	0.52	58	0.54	36	0.34
	cy	85	72	0.43	72	0.38	71	0.39	72	0.37
	nu/cy	308	93	0.87	92	0.84	91	0.78	77	0.63
Animals$^+$	SP	75	97	0.89	97	0.88	87	*0.82*	88	*0.80*
	mi	48	89	*0.80*	86	0.78	83	0.75	83	*0.73*
	nu	*186*	*54*	*0.51*	*46*	*0.44*	*52*	*0.50*	*45*	*0.43*
	cy	85	72	*0.41*	72	*0.35*	71	*0.37*	72	*0.35*
	nu/cy	*277*	*92*	*0.85*	*91*	*0.83*	*91*	*0.79*	*89*	*0.77*
Fungi	SP	9	78	0.60	78	0.59	78	0.63	78	0.63
	mi	77	68	0.62	66	0.61	51	0.52	54	0.55
	nu	152	63	0.36	63	0.36	50	0.32	44	0.28
	cy	180	54	0.27	54	0.27	56	0.22	54	0.18
	nu/cy	332	92	0.63	93	0.66	84	0.48	83	0.47
Plants	SP	6	83	0.58	50	0.40	83	0.50	83	0.47
	mi	6	67	0.51	67	0.45	67	0.40	67	0.42
	ch	72	77	0.72	78	0.70	53	0.51	54	0.51
	nu	36	91	0.77	77	0.63	86	0.74	79	0.64
	cy	17	41	0.38	41	0.33	37	0.20	29	0.12
	nu/cy	52	94	0.84	88	0.76	93	0.74	91	0.70

The sensitivity (SE) and Matthews correlation coefficient (MCC) of MultiLoc2 (ML2) are listed for each localization (Loc). The number of clusters (Nr) per localization is also shown. The results for MultiLoc2-LowRes* and MultiLoc2-HighRes* are obtained by simulating that for all test proteins no GO term is available. The Animals$^+$ dataset was obtained by reclustering the nuclear proteins from the original animals dataset. Changes in performance are highlighted in italic.

Table B.5: Comparison of the overall performance results of the MultiLoc2 predictors using an independent dataset

Version	**Classes**	**Average sensitivity (Overall accuracy)**			
		MultiLoc2-LR	**MultiLoc2-LR***	**MultiLoc2-HR**	**MultiLoc2-HR***
Animals	3	93 (93)	92 (92)	87 (89)	83 (80)
	4	80 (73)	78 (69)	75 (68)	70 (57)
Animals$^+$	3	93 (93)	*91* (92)	87 (89)	*87 (88)*
	4	*78 (70)*	*75 (66)*	*73 (67)*	*72 (64)*
Fungi	3	79 (87)	79 (88)	71 (78)	72 (77)
	4	66 (60)	65 (60)	59 (52)	58 (51)
Plants	4	80 (83)	71 (80)	74 (70)	74 (70)
	5	72 (76)	63 (71)	65 (62)	62 (59)

The average sensitivity and the overall accuracy (in parenthesis) of MultiLoc2 (ML2) for the prediction of three and four classes for animals and fungi and four and five classes for plants are shown. The results for MultiLoc2-LowRes* and MultiLoc2-HighRes* are obtained by simulating that for all test proteins no GO term is available. The Animals$^+$ dataset was obtained by reclustering the nuclear proteins from the original animals dataset. Changes in performance are highlighted in italic.

Appendix C

Publications

1. Höglund,A., Dönnes,P., **Blum,T.**, Adolph,H.W., and Kohlbacher,O. (2005) Using N-terminal targeting sequences, amino acid composition, and sequence motifs for predicting protein subcellular localization. *Proceedings of the German Conference on Bioinformatics (GCB 2005)*, edited by Andrew Torda, Stefan Kurtz and Matthias Rarey, GI, 45-59.

2. Höglund,A., **Blum,T.**, Brady,S., Dönnes,P., Miguel,J.S., Rocheford,M., Kohlbacher,O., Shatkay,H. (2006) Significantly improved prediction of subcellular localization by integrating text and protein sequence data. *Pacific Symposium on Biocomputing (PSB 2006)*, 16-27.

3. Höglund,A., Dönnes,P., **Blum,T.**, Adolph,H.W., and Kohlbacher,O. (2006) MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition. *Bioinformatics*, **22**, 1158-65.

4. Küntzer,J., **Blum,T.**, Gerasch,A., Backes,C., Kaufmann,M., Kohlbacher,O., and Lenhof,H.P. (2006) BN++- A Biological Information System. *J. Integr. Bioinform.*, **3**, 34.

5. Shatkay,H., Höglund,A., Brady,S., **Blum,T.**, Dönnes,P., Kohlbacher,O. (2007) SherLoc: high-accuracy prediction of protein subcellular localization by integrating text and protein sequence data. *Bioinformatics*, **23**, 410-7.

6. Küntzer,J., Backes,C., **Blum,T.**, Gerasch,A., Kaufmann,M., Kohlbacher,O., and Lenhof,H.P. (2007) BNDB - the Biochemical Network Database. *BMC Bioinformatics*, **8**, 367.

7. **Blum,T.**, and Kohlbacher,O. (2007) Finding relevant biotransformation routes in weighted metabolic networks using atom mapping rules. *Proceedings of the German Conference on Bioinformatics (GCB 2007)*, edited by Claudia Walter, Alexander Schliep, Joachim Selbig, Martin Vingron and Dirk Walther, GI, 30-44

8. **Blum,T.**, and Kohlbacher,O. (2008) Using atom mapping rules for an improved detection of relevant routes in weighted metabolic networks. *J. Comput. Biol.*, **15**, 565-76.

9. **Blum,T.**, and Kohlbacher,O. (2008) MetaRoute: Fast search for relevant metabolic routes for interactive network navigation and visualization. *Bioinformatics*, **24**, 2108-2109.

10. Mitschke,J., Fuss,J., **Blum,T.**, Höglund,A., Reski,R., Kohlbacher,O., and Rensing,S. (2009) Prediction of dual targeting to plant organelles. *New Phytol.*, **183**, 224-235

11. **Blum,T.**, Briesemeister,S., and Kohlbacher,O. (2009) MultiLoc2: integrating phylogeny and Gene Ontology terms improves subcellular protein localization prediction, submitted.

12. Briesemeister,S., **Blum,T.**, Brady,S., Lam,Y., Kohlbacher,O., and Shatkay,H. (2009) SherLoc2: a high-accuracy hybrid-method for predicting subcellular protein localization, submitted.

13. Gerasch,A., Küntzer,J., Backes,C., **Blum,T.**, Keller,A., Kohlbacher,O., Lenhof,H.P., and Kaufmann,M. Interactive analysis of biological network data using BiNA., in preparation.

Bibliography

Alberty,R.A. (2007) Thermodynamic properties of enzyme-catalyzed reactions involving cytosine, uracil, thymine, and their nucleosides and nucleotides. *Biophys. Chem.*, **127**, 91-96.

Alberty,R.A. (2006) Thermodynamic properties of enzyme-catalyzed reactions involving guanine, xanthine, and their nucleosides and nucleotides. *Biophys. Chem.*, **121**, 157-162.

Alberty,R.A. (2006) Thermodynamics of the reactions of carbamoyl phosphate. *Arch. Biochem. Biophys.*, **451**, 17-22.

Alberty,R.A. (2006) Calculation of equilibrium compositions of systems of enzyme-catalyzed reactions. *J. Phys. Chem. B*, **110(48)**, 24775-24779.

Alberty,R.A. (2005) BasicBiochemData3, *http://library.wolfram.com/infocenter/MathSource/5704*.

Alberty,R.A. (2005) Thermodynamics of Biochemical Reactions. *Wiley-IEEE.*

Alberty,R.A. (1998) Calculation of Standard Transformed Gibbs Energies and Standard Transformed Enthalpies of Biochemical Reactants. *Arch. Biochem. Biophys.*, **353**, 116-130.

Alberty,R.A. (1996) IUPAC-IUBMB Joint Commission on Biochemical Nomenclature (JCBN). Recommendations for nomenclature and tables in biochemical thermodynamics. Recommendations 1994. *Eur. J. Biochem.*, **240(1)**, 1-14.

Alberty,R.A. (1992) Equilibrium calculations on systems of biochemical reactions at specified pH and pMg. *Biophys. Chem.*, **42(2)**, 117-131.

Alberty,R.A. (1992) Calculation of transformed thermodynamic properties of biochemical reactants at specified pH and pMg. *Biophys. Chem.*, **43(3)**, 239-254.

Aittokallio,T., Schwikowski,B. (2006) Graph-based methods for analysing networks in cell biology. *Briefings in Bioinformatics*, **7**, 243-255.

Akutsu,T. (2004) Efficient extraction of mapping rules of atoms from enzymatic reaction data. *J. Comput. Biol.*, **11**, 449-462.

Andrade,M.A., O'Donoghue,S.I., Rost,B. (1998) Adaptation of protein surfaces to subcellular location. *J. Mol. Biol.*, **276(2)**, 517-525.

Arita,M. (2000) Metabolic reconstruction using shortest paths. *Simulation Practice and Theory*, **8**, 109-125.

Arita,M. (2003) In silico atomic tracing by substrate-product relationships in Escherichia coli intermediary metabolism. *Genome Res.*, **13**, 2455-2466.

Ashburner *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25(1)**, 25-29.

Bairoch *et al.* (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **33**, D154-D159.

Bannai,H., Tamada,Y., Maruyama,O., Nakai,K., Miyano,S. (2002) Extensive feature detection of N-terminal protein sorting signals. *Bioinformatics*, **18(2)**, 298-305.

Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Rudnev,D., Evangelista,C., Kim,I.F., Soboleva,A., Tomashevsky,M., Edgar,R. (2007) NCBI GEO: mining tens of millions of expression profiles-database and tools update. *Nucleic Acids Res.*, **37**, D760-D765.

Bellman,R. (1958) On a Routing Problem. *in Quarterly of Applied Mathematics*, **16**, 87-90.

Bendtsen,J.D., Nielsen,H., von Heijne,G., Brunak,S. (2004) Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.*, **340**, 783-795.

Berg,J.,M., Tymoczko,J.,L., Stryer,L., (2002) Biochemistry. *Freeman*, 5th edition.

Bhasin,M., Raghava,G.P. (2004) ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. *Nucleic Acids Res.*, **32**, W414-W419.

Bochar,D.A., Freisen,J.A., Stauffacher,C.V., Rodwell,V.W. (1999) Biosynthesis of mevalonic acid from acetyl-CoA. *Comprehensive Natural Products Chemistry, Elsevier Science Ltd*, **2**, 15-44

Boden,M., Hawkins,J. (2005) Prediction of subcellular localization using sequence-biased recurrent networks. *Bioinformatics*, **21(10)**, 2279-2286.

Brady,S., Shatkay,H. (2008) EpiLoc: a (working) text-based system for predicting protein subcellular location. *Pac. Symp. Biocomput.*, 604-615.

Casadio,R., Martelli,P.,L., Pierleoni,A. (2008) The prediction of protein subcellular localization from sequence: a shortcut to functional genome annotation. *Brief Funct Genomic Proteomic.*, **7(1)**, 63-73.

Caspi,R., Foerster,H., Carol,A., Fulcher,A., Hopkinson,R., Ingraham,J., Kaipa,P., Krummenacker,M., Paley,S., Pick,J., Rhee,S.R., Tissier,C., Zhang,P., Karp,P.D. (2006) MetaCyc: A multiorganism database of metabolic pathways and enzymes. *Nucl. Acids Res.*, **34**, D511-D516

Cedano,J., Aloy,P., Prez-Pons,J.A., Querol.E. (1997) Relation between amino acid composition and cellular location of proteins. *J. Mol. Biol.*, **266(3)**, 594-600.

Chang,C.C., Lin,C.J. (2003) LIBSVM: a library for support vector machines, *http://www.csie.ntu.edu.tw/ cjlin/libsvm/.*

Chen,M., Hofestaedt,R. (2005) An algorithm for linear metabolic pathway alignment. *In Silico Biol.*, **5**, 111-128

Chou,K., Cai,Y. (2004) Prediction of protein subcellular locations by GO-FunD-PseAA predictor. *Biochem Biophys Res Commun.*, **320(4)**, 1236-1239.

Chou,K., Cai,Y. (2003) A new hybrid approach to predict subcellular localization of proteins by incorporating gene ontology. *Biochem Biophys Res Commun.*, **311(3)**, 743-747.

Chou,K., Cai,Y. (2003) Prediction and classification of protein subcellular location - Sequence-order effect and pseudo amino acid composition. *J. Cell Biochem.*, **90(6)**, 1250-1260.

Chou,K., Cai,Y. (2002) Using functional domain composition and support vector machines for prediction of protein subcellular location. *J. Biol. Chem.*, **277(48)**, 45765-45769.

Chunhui,L., Christopher,S., Henry,S., Jankowski,M.D., Ionita,J.A., Hatzimanikatis,V., Broadbelt,L.J. (2004) Computational discovery of biochemical routes to specialty chemicals. *Chem. Eng. Sci.*, **59**, 5051-5060

Cokol,M., Nair,R., Rost,B. (2000) Finding nuclear localization signals. *EMBO Rep.*, **1(5)**, 411-415.

Cordwell,S.J. (1999) Microbial genomes and "missing" enzymes: redefining biochemical pathways. *Arch. Microbiol*, **172**, 269-279.

Cox,J.D., Wagman,D.D., Medvedev,M.V. (1989) CODATA Key Values for Thermodynamics. *Hemisphere*, Washington, D. C..

Croes,D., Couche,F., Wodak,S.J., van Helden,J. (2006) Infering Meaningful Pathways in Weighted Metabolic Networks. *J. Mol. Biol.*, **356**, 222-236.

Cui,Q., Jiang,T., Liu,B., Ma,S. (2004) Esub8: a novel tool to predict protein subcellular localizations in eukaryotic organisms. *BMC Bioinformatics*, **5(66)**.

Diebold,R., Schuster,J., Dschner,K., Binder,S. (2002) The branched-chain amino acid transaminase gene family in Arabidopsis encodes plastid and mitochondrial proteins. *Plant. Physiol.*, **129**, 540-550.

Deville,Y., Gilbert,D., van Helden,J., Wodak,S.J. (2003) An Overview of Data Models for the Analysis of Biochemical Pathways. *Brief. Bioinform.*, **4(3)**, 246-259.

Dijkstra,E.W. (1959) A Note on Two Problems in Connexion with Graphs. *Numerische Mathematlk 1*, 269-271.

Enault,K., Suhre,C., Poirot,O., Clavarie,J.M.. (2003) Annotation of bacterial genomes using improved phylogenetic profiles. *Bioinformatics*, **19**, i105-i107.

Eisenhaber,F., Bork,P. (1998) Wanted: subcellular localization or proteins based on sequence. *Trends Cell Biol.*, **8**, 169-170.

Emanuelsson,O., Brunak,S., von Heijne,G., Nielson,H. (2007) Locating proteins in the cell using TargetP, SignalP and related tools. *Nat. Protocols*, **2(4)**, 953-971.

Emanuelsson,O., Nielson,H., Brunak,S., von Heijne,G. (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.*, **300**, 1005-1016.

Emanuelsson,O., Nielson,H., von Heijne,G. (1999) ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Protein Sci.*, **8(5)**, 978-984.

Eppstein,D. (1998) Finding the k Shortest Paths. *SIAM Journal on Computing*, **28**, 652-673.

Floyd,R.W. (1962) Algorithms 97: Shortest Path. *Comm. ACM*, **5**, 345.

Forsythe,R.G, Karp,P.D., Mavrovouniotis,M.L. (1997) Estimation of equilibrium constants using automated group contribution methods. *CABIOS*, **13(5)**, 537-543.

Fujiwara,Y., Asogawa,M. (2001) Prediction of subcellular localizations using amino acid composition and order. *Genome Inform*, **12**, 103-12.

Fyshe,A., Liu,Y., Szafron,D., Greiner,R., Lu,P. (2008) Improving subcellular localization prediction using text classification and the gene ontology. *Bioinformatics*, **24(21)**, 2512-2517.

Gille,C., Hoffmann,S., Holzhtter,H.G.. (2005) Combining Bioinformatics Resources for the Structural Modelling of Eukaryotic Metabolic Networks. *Genome Informatics*, **16(1)**, 223-232.

Goldberg,R.N., Tewari, Y.B., Bhat, T.N. (2004) Thermodynamics of Enzyme-Catalyzed Reactions - a Database for Quantitative Biochemistry. *Bioinformatics*, **20(16)**, 2874-2877.

Goldberg,R.N., Tewari,Y.B., Bell,D., Fazio,K., Anderson, E. (1993) Thermodynamics of enzyme-Catalyzed reactions: part 1. Oxidoreductases. *J. Phys. Chem. Ref. Data*, **24**, 1765-1801.

Goldberg,R.N., Tewari,Y.B. (1994) Thermodynamics of enzyme-Catalyzed reactions: part 2. Transferases. *J. Phys. Chem. Ref. Data*, **23**, 547-617.

Goldberg,R.N., Tewari,Y.B. (1994) Thermodynamics of enzyme-Catalyzed reactions: part 3. Hydrolases. *J. Phys. Chem. Ref. Data*, **23**, 1035-1103.

Goldberg,R.N., Tewari,Y.B. (1995) Thermodynamics of enzyme-Catalyzed reactions: part 4. Lyases. *J. Phys. Chem. Ref. Data*, **24**, 1669-1698.

Goldberg,R.N., Tewari,Y.B. (1995) Thermodynamics of enzyme-Catalyzed reactions: part 5. Isomerases and ligases. *J. Phys. Chem. Ref. Data*, **24**, 1765-1801.

Goldberg,R.N., Tewari,Y.B. (1999) Thermodynamics of enzyme-Catalyzed reactions: part 6 - 1999 update. *J. Phys. Chem. Ref. Data*, **28**, 931-965.

Green,M.L., Karp,P.D. (2006) The outcomes of pathway database computations depend on pathway ontology. *Nucleic Acids Res.*, **34(13)**, 3687-3697.

Guda,C., Subramaniam,S. (2005) pTARGET: A new method for predicting protein subcellular localization in eukaryotes. *Bioinformatics*, **21**, 3963-3969.

Guo,J., Lin,Y. (2006) TSSub: eukaryotic protein subcellular localization by extracting features from profiles. *Bioinformatics*, **22(14)**, 1784-1788.

Handorf,T., Ebenhöh,O. (2007) MetaPath Online: a web server implementation of the network expansion algorithm. *Nucl. Acids Res.*, **35**, W613-W618.

Hart,P.E., Nilsson,N.J., Raphael,B. (1968) A formal basis for the heuristic determination of minimum cost paths. *IEEE Trans. Syst. Sci. Cybern. SSC-4*, **2**, 100-107.

Hoffman,W., Pavley,R. (1959) A method for the selection of the Nth best path problem. *J. Assoc. Commun. Mach.*, **6**, 506-514.

Horton,P., Park,K.J., Obayashi,T., Fujita,N., Harada,H., Adams-Collier,C.J., Nakai,K. (2007) WoLF PSORT: protein localization predictor. *Nucleic Acids Res.*, **35**, W585-W588.

Höglund,A., Dönnes,P., Blum,T., Adolph,H.W., Kohlbacher,O. (2006) MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition. *Bioinformatics*, **22(10)**, 1158-1165.

Hrazdina,G., Jensen,R.A. (1992) Spatial organization of enzymes in plant metabolic pathways. *Annu. Rev. Plant. Physiol. Plant. Mol. Biol.*, **43**, 241-267.

Hua,S., Sun,Z. (2001) Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, **17(8)**, 721-8.

Huanq,W.L., Tunq,C.W., Ho,S.W., Hwang,S.F., Ho,S.Y. (2008) ProLoc-GO: utilizing informative Gene Ontology terms for sequence-based prediction of protein subcellular localization. *BMC Bioinformatics*, **9**:80.

Ivanciuc,O., Ivanciuc,T., Cabrol-Bass,D., Balaban,A.T. (2001) Evaluation in Quantitative Structure-Property Relationship Models of Structural Descriptors Derived from Information-Theory Operates. *J. Chem. Inf. Comput. Sci.*, **40**, 631-643.

Ivanciuc,O., Ivanciuc,T., Klein,D.J., Seitz,W.A., Balaban,A.T. (2001) Wiener Index Extention by Counting Even/Odd Graph Distances. *J. Chem. Inf. Comput. Sci.*, **41**, 536-549.

Johnson,D.B. (1977) Efficient algorithms for shortest paths in sparse networks. *Journal of the ACM*, **24**, 1-13.

Kamp,A., Schuster,S. (2006) Metatool 5.0: fast and flexible elementary modes analysis. *Bioinformatics*, **22(15)**, 1930-1931.

Kanehisa,M. (1996) Toward pathway engineering: a new database of genetic and molecular pathways. *Science & Technology Japan*, **59**, 34-38.

Karp,P.D., Ouzounis,C.A., Moore-Kochlacs,C., Goldovsky,L., Kaipa,P., Ahrn,D., Tsoka,S., Darzentas,N., Kunin,V., Lpez-Bigas,N. (2005) Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res.*, **33(19)**, 6083-6089.

Karp,P.D., Paley,S., Romero,P. (2002) The Pathway Tools Software. *Bioinformatics*, **18**, S225-32.

Keseler,I.M., Collado-Vides,J., Gama-Castro,S., Ingraham,J., Paley,S., Paulsen,I.T., Peralta-Gil, M., Karp,P.D. (2005) EcoCyc: A comprehensive database resource for Escherichia coli. *Nucl. Acids Res.*, **33**, D334-D337.

Klamt,S., Stelling,J. (2002) Combinatorial complexity of pathway analysis in metabolic networks. *Mol. Biol. Rep.*, **29**, 233-6.

Klamt,S., Stelling,J., Ginkel,M., Gilles,E.D. (2003) FluxAnalyzer: exploring structure, pathways, and flux distributions in metabolic networks on interactive flux maps. *Bioinformatics*, **19**, 261-269.

Klukas,C., Schreiber,F. (2007) Dynamic exploration and editing of KEGG pathway diagrams. *Bioinformatics*, **23**, 344-350.

Kopka,J., Schauer,N., Krueger,S., Birkemeyer,C., Usadel,B., Bergmller,E., Drmann,P., Weckwerth,W., Gibon,Y., Stitt,M., Willmitzer,L., Fernie,A.R., Steinhauser,D. (2005) GMD@CSB.DB: the Golm Metabolome Database. *Bioinformatics*, **21(8)**, 1635-1638.

Küntzer *et al.* (2006) BN++ - A Biological Information System. *Journal of Integrative Bioinformatics*, **3(2)**.

Küntzer *et al.* (2007) BNDB - The Biochemical Network Database. *BMC Bioinformatics*, **8:367**.

Küffner,R., Zimmer,R., Lengauer,T. (2000) Pathway analysis in metabolic databases via differential metabolic display (DMD). *Bioinformatics*, **16**, 825-836.

Kuzuyama,T., Seto,H. (2003) Diversity of the biosynthesis of the isoprene units. *Nat. Prod.Rep.*, **20**, 171-183.

Leivar,P., Gonzlez,V.M., Castel,S., Trelease,R.N., Lpez-Iglesias,C., Arr,M., Boronat,A., Campos,N., Ferrer,A., Fernndez-Busquets,X. (2005) Subcellular localization of Arabidopsis 3-hydroxy-3-methylglutaryl-coenzyme A reductase. *Plan. Physiol.*, **137**, 57-69.

Lei,Z. ,Dai,Y. (2006) Assessing protein similarity with Gene Ontology and its use in subnuclear localization prediction. *BMC Bioinformatics*, **7**:491.

Lichtenthaler,H.K., Rohmer,M., Schwender,J. (1997) Two independent biochemical pathways for isopentenyl diphosphate and isoprenoid biosynthesis in higher plants. *Plant. Physiol.*, **101**, 643-652.

Lu,Z., Hunter,L. (2005) Go molecular function terms are predictive of subcellular localization. *Pac Symp Biocomput.*, 151-161.

Lu,Z., Szafron,D., Greiner,R., Lu,P., Wishart,D.S., Poulin,B., Anvik,J., Macdonell,C., Eisner,R. (2004) Predicting subcellular localizations of proteins using machine-learned classifiers. *Bioinformatics*, **20(4)**, 547-556.

Marcotte,E.M., Xenarios,I., van der Bliek,A.M., Eisenberg,D. (2000) Localizing proteins in the cell from their phylogenetic profiles. *PNAS*, **97**, 12115-12120.

Marcotte,E.M. (2000) Computational genetics: finding protein function by nonhomology methods. *Curr Opin Struct Biol.*, **10(3)**, 359-365.

Maskow,T., von Stockar,U. (2005) How reliable are thermodynamic feasibility statements of biochemical pathways? *Biotechnol. Bioeng.*, **92(2)**, 223-230.

Mavrovouniotis,M.L. (1990) Group Contributions for Estimating Standard Gibbs Energies of Formation of Biochemical Compounds in Aqueous Solution. *Biotechnol. Bioeng.*, **36**, 1070-1082.

Mavrovouniotis,M.L. (1991) Estimation of Standard Gibbs Energy Changes of Biotransformations. *J. Biol. Chem.*, **266(22)**, 14440-14445.

McShan,D.C., Rao,S., Shah,I., (2003) PathMiner: predicting metabolic pathways by heuristic search. *Bioinformatics*, **13**, 1692-1698.

Morgan,H.L. (1965) The Generation of a Unique Machine Description for Chemical Structures - A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.*, **5**, 107-113.

Mulder,N.J. *et al.* (2007) New developments in the InterPro database. *Nucleic Acids Res.*, **35**, 224-228.

Nair,R., Rost,B. (2005) Mimicking Cellular Sorting Improves Prediction of Subcellular Localization. *J. Mol. Biol.*, **348**, 85-100.

Nair,R., Rost,B. (2002) Inferring sub-cellular localization through automated lexical analysis. *Bioinformatics*, **18**, S78-S86.

Nair,R., Rost,B. (2002) Sequence conserved for subcellular localization. *Protein Sci.*, **11(12)**, 2836-2847.

Paley,S., Karp,P.D. (2002) Evaluation of computational metabolic-pathway predictions for H. pylori. *Bioinformatics*, **18**, 715-724.

Papin,J.A., Price,N.D., Wiback,S.J., Fell,D.A., Palsson,B.O. (2003) Metabolic pathways in the post-genome era. *Trends Biochem Sci.*, **28(5)**, 250-258.

Park,K.J., Kanehisa,M. (2003) Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. *Bioinformatics*, **19(13)**, 1656-63.

Pellegrini,M., Marcotte,E.M., Thompson,M.J., Eisenberg,D., Yeates,T.O. (1999) Assigning protein functions by comparative genome analysis. *PNAS*, **96**, 4285-4288.

Petsalaki,E.I., Bagos,P.G., Litou,Z.I., Hamodrakas,S.J. (2006) PredSL: a tool for the N-terminal sequence-based prediction of protein subcellular localization. *Genomics Proteomics Bioinformatics*, **4(1)**, 48-55.

Pierleoni,A., Martelli,P.L., Fariselli,P.L., Casadio,R. (2006) BaCelLo: a balanced subcellular localization predictor. *Bioinformatics*, **22(14)**, 408-416.

Rahman,S.A., Advani,P., Schunk,R., Schrader,R., Schomburg,D. (2005) Metabolic pathway analysis web service (Pathway Hunter Tool at CUBIC). *Bioinformatics*, **21**, 1189-1193.

Reinhardt,A., Hubbard,T. (1998) Using neural networks for prediction of the subcellular location of proteins. *Nucleic Acids Res.*, **26(9)**, 2230-2236.

Reumann,S., Babujee,L., Ma,C., Wienkoop,S., Siemsen,T., Antonicelli,G.E., Rasche,N., Lüder,F., Weckwerth,W., Jahn,O. (2007) Proteome analysis of Arabidopsis leaf peroxisomes reveals novel targeting peptides, metabolic pathways, and defense mechanisms. *Plant. Cell.*, **19**, 3170-3193.

Rockafellar,R.T., (1970) Convex analysis,.*Princeton Landmarks in Mathematics*, *Princeton University Press.*

Rohmer,M. (1999) The discovery of a mevalonate-independent pathway for isoprenoid biosynthesis in bacteria, algae and higher plants. *Nat. Prod. Rep.*, **16**, 565-574.

Sapir-Mir,M., Mett,A., Belausov,E., Tal-Meshulam,S., Frydman,A., Gidoni,D., Eyal,Y. (2008) Peroxisomal localization of Arabidopsis isopentenyl diphosphate isomerases suggests that part of the plant isoprenoid mevalonic acid pathway is compartmentalized to peroxisomes. *Plant. Physiol.*, **148**, 1219-28.

Schilling,C.H., Letscher,D., Palsson,B. (2000) Theory for the Systematic Definition of Metabolic Pathways and their use in Interpreting Metabolic Function from a Pathway-Oriented Perspective. *J. Theor. Biol.*, **203**, 229-248.

Schomburg,I., Chang,A., Schomburg,D. (2002) BRENDA, enzyme data and metabolic information. *Nucl. Acids Res.*, **30**, 47-49.

Schuster,S., Dandekar,T., Fell,D.A. (1999) Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering., *Trends Biotechnol.*, **17**, 53-60.

Schuster,S., Fell,D., Dandekar,T. (2000) A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nat Biotechnol.*, **18**, 326-332.

Schürmann,M., Sprenger,G. (2001) Fructose-6-phosphate Aldolase is a Novel Class I Aldolase from Escherichia coli and is Related to a Novel Group of Bacterial Transaldolases., *J. Biol. Chem.*, **276**, 11055-11061.

Scott,M.S., Thomas,D.Y., Hallett,M.T. (2004) Predicting subcellular localization via protein motif co-occurrence. *Genome Res.*, **14(10A)**, 1957-66.

Shatkay,H., Höglund,A., Brady,S., Blum,T., Dönnes,P., Kohlbacher,O. (2007) SherLoc: high-accuracy prediction of protein subcellular localization by integrating text and protein sequence data. *Bioinformatics*, **23(11)**, 1410-7.

Shen,H.B., Yanq,J., Chou,K.C. (2007) Euk-PLoc: an ensemble classifier for large-scale eukaryotic protein subcellular location prediction. *Amino Acids*, **33(1)**, 57-67.

Sirava *et al.* (2002) BioMiner - modeling, analyzing, and visualizing biochemical pathways and networks. *Bioinformatics*, **18(2)**, S219-S230.

Small,I., Peeters,N., Legeai,F., Lurin,C. (2004) Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences. *Proteomics*, **4(6)**, 1581-1590.

Sun,J., Xu,J., Liu,Z., Liu,Q., Zhao,A., Shi,T., Li,Y. (2005) Refined phylogenetic profiles method for predicting protein-protein interactions. *Bioinformatics*, **16**, 3409-3415.

Talete srl (2007) DRAGON for Linux (Software for Molecular Descriptor Calculations). Version 1.4, *http://www.talete.mi.it/*.

Illustration of cellular compartmentalization and protein sorting, ©The Nobel Assembly at Karolinska Institutet (1999), *http://nobelprize.org/nobel_prizes/medicine/laureates/1999/med-cell-e.gif*.

Tolbert,N.E. (1981) METABOLIC PATHWAYS IN PEROXISOMES AND GLYOXYSOMES. *Ann. Rev. Biochem.*, **50**, 133-157.

Toropov,A.A., Toropova,A.P. (2003) QSPR Modeling of Alkanes Properties based on Graph of Atomic Orbitals. *THEOCHEM*, **637**, 1-10.

van Helden,J., Wernisch,L., Gilbert,D., Wodak,S.J. (2002) Graph-based analysis of metabolic networks. *Ernst Schering Res. Found. Workshop*, **38**, 245-274.

Vapnik,V.N. (1999) The Nature of Statistical Learning Theory. *Wiley, NY.*

Wagman,D.D., Evans,W.H., Parker,V.B., Schumm,R.H., Halow,I., Bailey,S.M., Churney,K.L., Nutall,R.L. (1982) The NBS Tables of Chemical Thermodynamic Properties. *J. Phys. Chem. Ref. Data*, **11**.

Weininger,D., Weininger,A., Weininger,J.L. (1989) SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.*, **29**, 97-101.

Willet,P., Barnard,J.M., Downs,G.M. (1998) Chemical similarity searching. *J. Chem. Inf. Comput. Sci.*, **38**, 938-996.

Wolfram Research, Inc (2005) Mathematica Edition: Version 5.2. *Wolfram Research, Inc.*

Xie,D., Li,A., Wang,M., Fan,Z., Feng,H. (2005) LOCSVMPSI: a web server for subcellular localization of eukaryotic proteins using SVM and profile of PSI-BLAST. *Nucleic Acids Res.*, **33**, W105-W110.

Yan,A. (2006) Modeling of Gibbs Energy of Formation of Organic Compounds by Linear and Nonlinear Methods. *J. Chem. Inf. Model.*, **46**, 2299-2304.

Yeung,M., Thiele,I., Palsson,B. (2007) Estimation of the extreme pathways for metabolic networks. *BMC Bioinformatics*, **8**, 363.

Zdobnov,E.M., Apweiler,R. (2001) InterProScan - an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, **17(9)**, 847-848.