J Mol Evol (2005) 61:498–512
DOI: 10.1007/s00239-005-0027-1

JOURNAL OF MOLECULAR EVOLUTION



# Expanding Metabolic Networks: Scopes of Compounds, Robustness, and Evolution

**Thomas Handorf, Oliver Ebenhöh, Reinhart Heinrich**

Institute of Biology, Department of Theoretical Biophysics, Humboldt University Berlin, Invalidenstr. 42, 10115 Berlin, Germany

Received: 31 January 2005 / Accepted: 24 May 2005 [*Reviewing Editor:* Dr. David Pollock]

**Abstract.** A new method for the mathematical analysis of large metabolic networks is presented. Based on the fact that the occurrence of a metabolic reaction generally requires the existence of other reactions providing its substrates, series of metabolic networks are constructed. In each step of the corresponding expansion process those reactions are incorporated whose substrates are made available by the networks of the previous generations. The method is applied to the set of all metabolic reactions included in the KEGG database. Starting with one or more seed compounds, the expansion results in a final network whose compounds define the scope of the seed. Scopes of all metabolic compounds are calculated and it is shown that large parts of cellular metabolism can be considered as the combined scope of simple building blocks. Analyses of various expansion processes reveal crucial metabolites whose incorporation allows for the increase in network complexity. Among these metabolites are common cofactors such as $NAD^+$, ATP, and coenzyme A. We demonstrate that the outcome of network expansion is in general very robust against elimination of single or few reactions. There exist, however, crucial reactions whose elimination results in a dramatic reduction of scope sizes. It is hypothesized that the expansion process displays characteristics of the evolution of metabolism such as the temporal order of the emergence of metabolic pathways.

*Correspondence to:* Reinhart Heinrich; *email:* reinhart.heinrich@biologie.hu-berlin.de

**Key words:** Metabolic networks — Scope — Evolution of metabolic pathways — Stoichiometry — KEGG database — Robustness

## Introduction

Traditionally, the theoretical analysis of metabolic networks concerns the development of models for simulating their dynamical properties. Such a description is based on systems of ordinary differential equations and requires detailed knowledge of the stoichiometry, regulatory interactions, and kinetic characteristics of the enzymes (Heinrich and Schuster 1996). In general, these models describe systems of a relatively small size, such as single biochemical pathways or a small number of interacting pathways. With the emergence of biochemical databases such as KEGG (Kanehisa 1997; Kanehisa and Goto 2000) or Brenda (Schomburg et al. 2002), information about large–scale metabolic networks has become easily accessible. Such databases are particularly comprehensive regarding the type of reactions and the catalyzing enzymes. So far, the KEGG database includes data on over 170 organisms. Whereas for small networks, the wiring principles of the reactions and metabolites are easily comprehensible, for large networks a more thorough analysis is required. Such structural investigations are of particular interest since networks of biochemical reactions are the outcome of natural selection during evolution and can therefore be expected to show characteristic features

compared to chemical reaction networks of inanimate matter or random networks (Heinrich et al. 1991). This kind of analysis is feasible even without considering kinetic properties of the enzymes.

In recent years, several approaches for the structural analysis of metabolic networks have been introduced, such as flux balance analysis (Kauffman et al. 2003), the concept of elementary flux modes (Schuster et al. 2000) or extreme pathways (Papin et al. 2003), and graph theoretical analyses (Jeong et al. 2000, Wagner and Fell 2001).

In this work we present a new method for the analysis of structural properties of large metabolic networks. The presented method is based on the fact that there exists a hierarchical ordering of metabolic reactions (Ebenhöh et al. 2004). This ordering is expressed by the facts that only those reactions may take place which can utilize the available substrates and that these reactions produce new substances which may be used by further reactions. To account for such interdependencies of reactions we developed the method of network expansion and the concept of scopes. The expansion process starts from a small number of selected compounds and results in a series of networks growing in size. The final network defines the scope comprising those compounds which can in principle be synthesized from the initial compounds.

Using the KEGG database as the source of information, we perform the expansions from all compounds and calculate their scopes. Moreover, combined scopes of simple building blocks are determined from which large parts of cellular metabolism can be reconstructed. The analysis of these expansion processes reveals structural features such as crucial reactions and metabolites whose incorporation into the network allows for networks with a higher complexity. By analyzing the effects of deletions of one or more reactions, we draw conclusions on the robustness of scopes.

We propose that the expansion process reflects important aspects of the evolution of metabolism from simple chemical building blocks toward complex biochemical reaction networks. In particular, hypotheses can be formulated in which temporal order reactions were recruited for cellular metabolism and metabolic pathways have been established.

## Concept of Network Expansion

The expansion process starts with one or more initial compounds forming the seed of the resulting network. During the process new reactions are added to the network which utilize those compounds that are already present in the expanding network. The products of the new reactions become part of the network and may be used as substrates in subsequent steps of the expansion process. The reactions are taken from a base set of biochemically feasible reactions. In the present work we use as the base set all reactions from the reference pathways defined in the KEGG database. In this way the base set comprises all known metabolic reactions from a large number of organisms.

The expansion process is defined by the following algorithm.

1. Selection of one or more biochemical compounds acting as a seed of the expanding network. The seed represents the first generation in the series of expanding networks.
2. Identification of those reactions from the base set which use as substrates only compounds already present in the current network.
3. Incorporation of the identified reactions and their products into the network. This results in the next generation of expanding networks.
4. Repetition of steps 2 and 3 until no further reactions can be identified for incorporation.

After completing the process, the expanded network will contain all compounds which can be synthesized from the seed using the reactions from the base set. Since not all compounds can be synthesized from arbitrary seed compounds, the expansion process will in general not lead to a network containing all reactions from the base set.

The set of compounds which are contained in the expanded network resulting from a single seed compound $A$, we denote $\sum(A)$ and call it the scope of $A$. By the scope size $\sigma(A)$ we denote the number of compounds contained in the scope $\sum(A)$. Corresponding to the set of compounds, the final network also contains an associated set of reactions. Clearly, if a compound $B$ is included in the scope of $A$, then the scope of $B$ is a subset of the scope of $A$, formally:

$$B \in \sum(A) \text{ is equivalent to } \sum(B) \subseteq \sum(A) \quad (1)$$

Further, if two compounds $A$ and $B$ are interconvertible in the sense that $A$ can be produced from $B$ and $B$ can be produced from $A$ (without using other compounds), then $A$ is included in the scope of $B$ and $B$ is included in the scope of $A$. This implies that the scopes of $A$ and $B$ are identical, formally described by:

$$B \in \sum(A) \text{and } A \in \sum(B) \text{ are equivalent to } \sum(A) = \sum(B) \quad (2)$$

Scopes may also be defined for a seed consisting of more than one initial compound $A_1,\ldots,A_k$. This

results in the combined scope $\sum(A_1,...,A_k)$. It is evident that this combined scope cannot be smaller than the union of the single scopes $\sum(A_1),...,\sum(A_k)$ of the individual compounds.

For the results presented below we consider water always to be present. This means, in particular, that the calculated scopes represent in a strict sense the combined scope of the seed compounds plus water.

Metabolic networks contain reactions which are almost irreversible and therefore predominantly take place only in one direction. For most of our calculations we assume that all reactions can take place in both forward and backward direction. We refrained from distinguishing between reversible and irreversible reactions, since the actual rates for the forward and backward directions depend on the physiological conditions, in particular, on the concentrations of the initial substrates and end products. Incorporation of the direction of reactions would require additional information which is not provided by the structural properties of the network.

The base set of reactions which was retrieved from the KEGG database contains information on 5311 reactions from more than 170 different organisms. These reactions are connected to 4587 compounds. The KEGG database includes 6481 further compounds which do not play a role in our calculations since they are not connected to any biochemical reaction. The presented methods can in principle be applied to other base sets. Other meaningful choices of base sets include the reactions contained in the metabolic networks of specific organisms. Obviously, the specific results will depend on the selected base set.

## Results

### *Scopes of Single Compounds*

We have calculated the scope for each compound connected to the base set of reactions. It turns out that not all of these scopes are different. Specifically, the 4587 seed compounds result in 3345 distinct scopes. Figure 1 shows a histogram for the number of distinct scopes of a given size. It can be seen that the smallest scope size is 5 and the largest scope size is 2101 and that the distribution of scope sizes is very nonuniform. The histogram reveals that every scope size $5 \leq \sigma \leq 32$ exists, while larger sizes scopes occur only sporadically. For small sizes there typically exist several distinct scopes of the same size, whereas for large sizes different scopes generally also have different sizes. There exist nine different scopes of size larger than 1500. Three of them can be reached from several seed compounds. This holds true for the scopes of size 1557, 1623, and 2101.

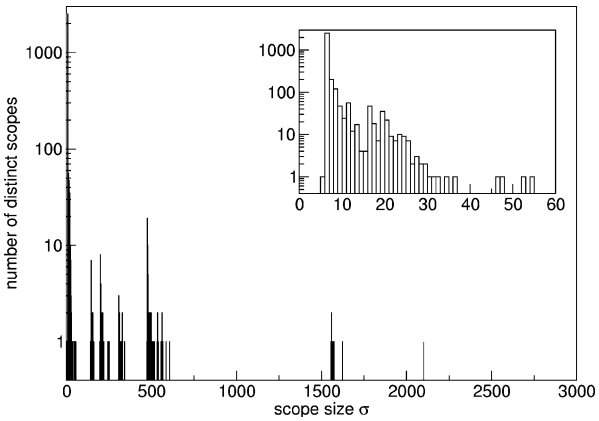


**Fig. 1.** Histogram of scope sizes showing how many distinct scopes have a certain size σ. The inset is a magnification for small scope sizes. All calculations include water as a seed compound. The smallest scope, of size $\sigma = 5$, corresponds to the scope of water ($H_2O$, $O_2$, $O_2^-$, $H^+$, $H_2O_2$).

The largest scope of size 2101 results from four different single seed compounds: adenosine 5'−phosphosulfate (APS), 3'−phosphoadenosine 5'−phosphosulfate (PAPS), dephospho−coenzyme A (CoA), and UDP−6−sulfoquinovose. APS and PAPS play an important role in the sulfur metabolism in many microorganisms. Dephospho−CoA is a direct precursor in the CoA biosynthesis pathway. UDP−6−sulfoquinovose plays a role in the nucleotide sugars metabolism.

The scope of size 1623 can be reached from six single seed compounds including phosphatidylserine, phosphatidylethanolamine, and CDP−diacylglycerol.

Of particular interest is the scope of size 1557 which can be reached by 103 different single seed compounds. Among them are central cofactors such as ATP and GTP as well as the corresponding mono− and diphosphates and the nicotinamide dinucleotides NADH and NADPH.

From Eq. (2) it follows that the seed compounds of each of the discussed scopes are interconvertible. Obviously, it is necessary for two compounds to be interconvertible that they are composed of the same chemical elements. However, not all compounds fulfilling this condition are interconvertible. This is the case if the reactions present in the base set do not have the capability to perform the interconversion. As an example, we consider the two compounds CoA and dephospho−CoA. Both substances consist of the same chemical elements. The only difference is that CoA contains three phosphate groups, whereas dephospho−CoA contains only two. Our calculations revealed that CoA is in the scope of dephospho−CoA, whereas the opposite does not hold true. Even though the base set includes the reaction dephospho−CoA + ATP ↔ CoA + ADP, it does not represent a direct interconversion between the two

compounds since it requires the presence of ATP or ADP. However, CoA can be produced from dephospho−CoA in a higher number of steps. This is possible since ATP is in the scope of dephospho−CoA (see above). In contrast, dephospho−CoA cannot be produced from coenzyme A since its scope does not contain ADP. In order to get an impression of how many of the compounds containing the same elements are really interconvertible, we have analyzed the interconvertiblitiy for all pairs of compounds containing only the elements C, H, and O. The database contains 1434 such compounds forming 1,027,461 different pairs. From these pairs of compounds, only 6713 pairs (0.65%) represent two compounds which can be interconverted. Analogously, we have analyzed all 183 compounds containing the elements C, H, O, N, P, and S. It turns out that 1.29% of all pairs of these compounds are interconvertible. Interestingly, for all 362 compounds containing the elements C, H, O, N, and P, over 8% of all pairs are interconvertible. This high percentage can be explained by the fact that many of those compounds are seed compounds of the scope of ATP.

Many scopes are subsets of larger scopes; see Eq. (1). For example, the scope of ATP is a subset of the scope of APS. From this it follows that ATP can be synthesized from APS. The opposite process is not possible, which simply follows from the fact that APS is composed of adenine, ribose, sulfate, and phosphate, whereas the adenosine phosphates AMP, ADP, and ATP contain the same building blocks except sulfate.

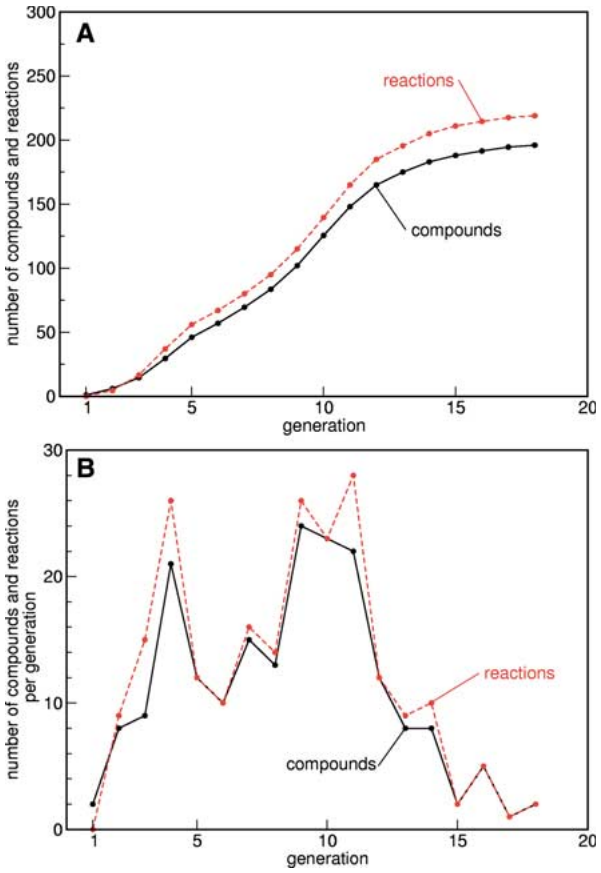


**Fig. 2.** The expansion process of glucose. **A** The total numbers of reactions and compounds present in the network of a certain generation. **B** Numbers of compounds and reactions incorporated in a given generation. The process starts in generation 1 with the two seed compounds glucose and water.

*The Expansion Process*

During an expansion process a network evolves from certain seed compounds by incorporation of new reactions and compounds. So far we have analyzed only the final result of this process, which defines the scope of the seed compounds. In this section we investigate the course of the expansion process, again starting with a single seed compound. In particular, we analyze two expansions, one starting from glucose, the other from ATP. The scope sizes are 197 and 1557, respectively.

Figure 2a shows the number of reactions and compounds over the network generation during the expansion process of glucose. The numbers of compounds and reactions increase slowly in the beginning, then faster in the middle phase, and saturate in the last phase. This behavior is expected: in the beginning the network is small and consequently there are only a few compounds which can serve as substrates for the reactions to be incorporated. Therefore only a few reactions can be added, even though there are many reactions still available from the base set. In the second phase the network has grown to such a size that it contains many compounds which can be used in the further development as substrates. Also, the set of available reactions is still considerably large. This results in a strong increase in the network size in this phase. Toward the end of the process those reactions which can use the compounds of the present network get exhausted. Therefore only a few reactions can be added per generation, which results in a final saturation phase. A closer inspection of the expansion curve for the scope of glucose reveals that in the middle phase the process temporarily slows down (generations 5–8) and again accelerates (generations 9–11). This becomes more apparent when looking at the number of compounds and reactions which are added per generation as shown in Figure 2b.

We investigated whether these characteristics of the expansion process are also typical for other seed compounds. Figures 3a and b depict the expansion process starting with ATP. It can be observed that the feature of slow-down and acceleration is even more pronounced.

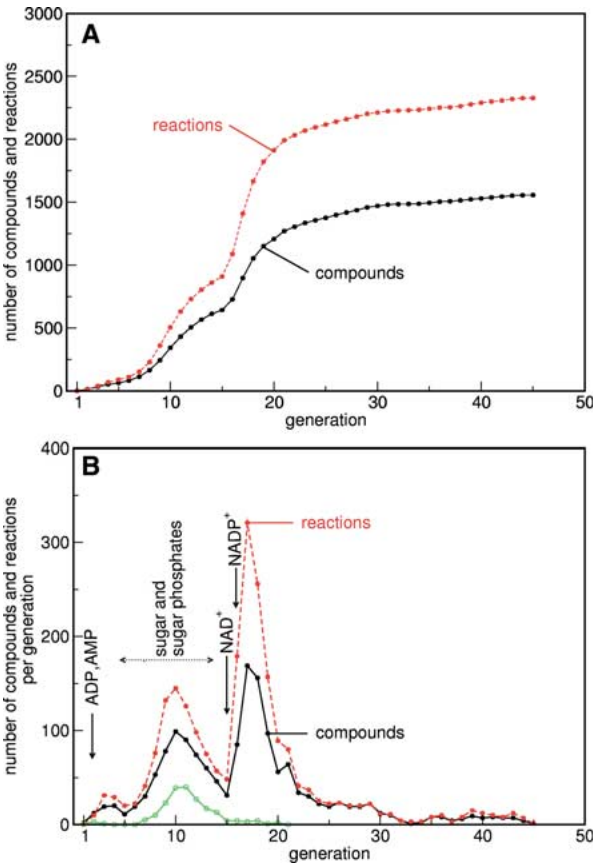


**Fig. 3.** The expansion process of ATP. **A** The total numbers of reactions and compounds present in the network of a certain generation. **B** Numbers of compounds and reactions incorporated in a given generation. Arrows indicate the incorporation of some important compounds leading to an acceleration of the expansion process. NADH and NADPH are synthesized in the generation following the acquisition of $NAD^+$ and $NADP^+$, respectively. Open circles indicate the numbers of compounds which are also found in the scope of glucose.

These expansion curves suggest that at critical stages in which the process becomes slow, certain key compounds are incorporated, which enables the network to expand again rapidly. In the following we analyze such critical stages in closer detail for the expansion starting with ATP.

The phenomenon of temporary slow-down of the expansion is most predominant at generations 5 and 15. This separates the expansion process into three stages. The first four generations are characterized by the inclusion of compounds which can be directly synthesized from ATP. Besides AMP, ADP, and adenine, these include ITP as well as R5P, PRPP, and ribose. The appearance of sugars and sugar phosphates leads to a strong acceleration of the expansion process after generation 5. In fact, between generation 5 and generation 15 many of the compounds incorporated are part of the scope of glucose (open circles in Fig. 3b).

Between generation 10 and generation 15 the number of available reactions which can utilize the sugars and sugar phosphates decreases, leading to a slow–down of the expansion process. The situation changes dramatically in generation 15, when $NAD^+$ is incorporated into the growing network. This compound is capable of acting as a cofactor in a large number of reactions. In fact, $NAD^+$ participates in 104 of the 179 new reactions added in generation 16. The expansion process is further accelerated in generation 16, mainly by incorporation of the compounds NADH and $NADP^+$. The exploitation of the capabilities of the nicotinamide dinucleotides to act as cofactors leads to a deceleration of the expansion process after generation 17. As a tendency, the deceleration continues until the end of the expansion process since the number of available reactions decreases.

During the whole process, the number of reactions added per generation is generally higher than the number of compounds added. This means that incorporation of a new reaction does not necessarily allow for the synthesis of additional compounds. The reason is the existence of alternative pathways for the production of certain compounds.

### *Distance Between Compounds*

The expansion process can be considered as a series of consecutive synthesis steps. In each step all those compounds are synthesized that can be produced by the set of available reactions using only those compounds as substrates which were provided by previous steps. Assuming that compound $B$ is in the scope of compound $A$, that is, $B \in \sum(A)$, we define the distance $d(A, B)$ from compound $A$ to compound $B$ by the number of required consecutive steps to produce $B$ exclusively from $A$. The distance $d(A, B)$ is not defined if $B$ is not in the scope of $A$. If $A$ is in the scope of $B$ and $B$ in the scope of $A$, both distances, $d(A, B)$ and $d(B, A)$, are defined but not necessarily the same. This asymmetry can be explained as follows. When producing $B$ from $A$, a number of additional end products are generally also synthesized. The direct inversion of this process would require these end products as substrates and therefore does not represent an expansion process starting from the only seed compound $B$. Therefore, the synthesis of $A$ from $B$ in general requires different synthesis steps. Obviously, the distance $d(A, B)$ is the integer which is smaller by 1 than the number of the generation in which compound $B$ is incorporated into the expansion process starting from the seed compound $A$.

As an example we consider the distance from pyruvate to citrate, and vice versa. As shown in Figure 4 the synthesis of citrate from pyruvate requires four synthesis steps, whereas pyruvate can be synthesized from citrate in only two steps.

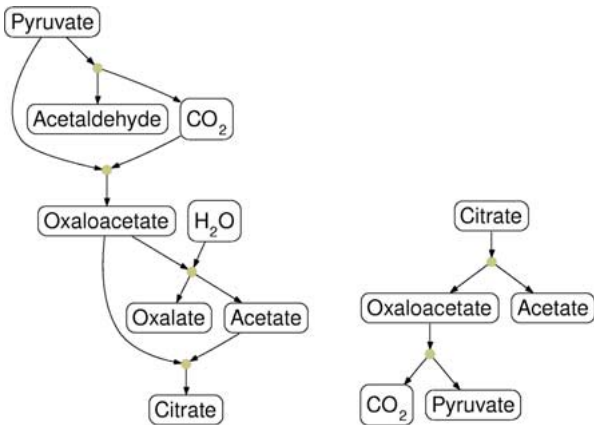


**Fig. 4.** Shortest paths from pyruvate to citrate (left) and citrate to pyruvate (right) defining the distances $d$(Pyr, Cit) = 4 and $d$(Cit, Pyr) = 2.

*Reconstruction of the Complete Network*

Using single compounds as seeds for the expansion process results in final networks which are subnetworks of the complete metabolism. The size of a resulting network depends on the synthesizing capacity of its seed compound. Consideration of further seed compounds will in general increase the size of the resulting network. It is possible to identify sets of compounds which expand to the full network. For finding a minimal set of such compounds we applied the following algorithm.

Initially, a seed compound $C_1$ is selected whose scope $\sum(C_1)$ is not smaller than the scope of any other compound. In the next step a second seed compound $C_2$ is determined which maximizes the size of the combined scope $\sum(C_1, C_2)$. This process is repeated in such a way that in step $i$ a seed compound $C_i$ is identified maximizing the size of the combined scope $\sum(C_1,..., C_{i-1}, C_i)$. For this, all sets which contain the compounds $C_1,..., C_{i-1}$ and one additional compound which is not yet included in the combined scope of step $i - 1$ have to be tested. Generally, the choice of the compound $C_i$ is not unique. In the case that there are several compounds leading to the same increase in scope size, one of these compounds is selected randomly.

A typical result of the algorithm is depicted in Figure 5. Shown is the development of the size of the combined scope versus the step number $i$. Initially, a strong increase in the scope size can be observed, while in a later phase of the process the scope size increases by only a small number of compounds per step. The process terminates after 680 seed compounds have been identified. The expansion process starting with this seed covers the complete network.

Interestingly, after only 10 steps the combined scope already covers 58% of the complete network (2651 of 4587 compounds). Table 1 contains the details of the first 10 steps of the algorithm. The seed compounds which were detected by this realization of the coverage process are listed in the second column; the corresponding increases in the size of the combined scope, in the third column. In the fourth column, we note the number of alternative seed compounds which yield the same increase in scope size. The alternative compounds in the first step are, besides APS, the other three compounds, PAPS, dephospho−CoA, and UDP−6−sulfoquinovose, which result in the largest scope, of size 2101. Interestingly, in the later phase of the covering process, the scope size increases mostly linearly with the step number. With a few exceptions, the size of the combined scope increases by two compounds per step over a period of more than 400 steps. In each such case, one of the two compounds is the new seed compound; the other, the new product which can be formed by the added reaction. Closer inspection reveals that these two compounds are generally very specialized in the sense that they exclusively participate in the added reaction. Therefore, neither of these two compounds serves as substrate later in the covering process.

The process of the reconstruction of the complete network is characterized by the stepwise inclusion of new seeds also enriching the diversity of the compounds in terms of their chemical elements. We therefore identified the chemical elements which are contained in the compounds within the combined scope in step $i$. In the last column in Table 1 those elements are listed which appear for the first time in a given step. There are six elements included in the first step. Their use in biochemical compounds is most productive since they can form a very large part of all compounds. In the further development of the reconstruction process, the number of elements increases slowly as the combined scope grows. After incorporation of Cl and Se in steps 3 and 7, respectively, the other elements are included later in the following order: Co(11), Fe(13), Mg(20), I(26), As(81), F(99), Hg(177), Br(352), Mn(670), where the numbers in parentheses denote the step number in which the element appears for the first time.

Since in each step of the algorithm the process can generally be continued with alternative seeds (see Table 1), we repeated the calculation many times using another choice of alternative seeds. It turns out that the length of the process remains almost constant (between 678 and 680 steps) and that the order of appearance of the elements remains essentially the same.

It should be noted that the 17 elements appearing in this reconstruction are only those which are constituents of biochemical compounds in metabolic reactions as documented in the KEGG database. Other elements such as $Na^+$, $K^+$, and $Ca^{2+}$ play a role in membrane transport, signal transduction, or other cellular processes and are not included.

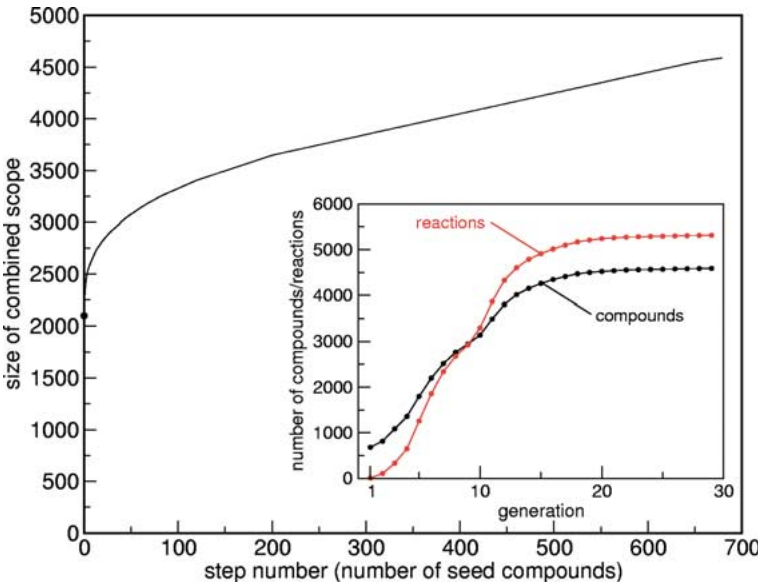


**Fig. 5.** Reconstruction of the network containing the complete reference set of reactions from the KEGG database using a minimal set of seed compounds. Shown is the size of the combined scope in a given step and the number of reactions of the corresponding network. The inset depicts the expansion process starting from the 660 compounds representing a minimal seed for the complete network.

**Table 1.** The first 10 steps in reconstruction of the network containing the complete reference set of reactions from the KEGG database

| Step no. | New seed | New compounds | Alternative seeds | New element(s) |
|---|---|---|---|---|
| 1 | APS | 2101 | 4 | C, H, N, O, P, S |
| 2 | α-Amino acid ester | 236 | 2 | |
| 3 | 1-Chloro-2,2-bis (4'-chlorophenyl) ethane | 99 | 11 | Cl |
| 4 | Tubulosine | 59 | 3 | |
| 5 | Lipoxin | 38 | 38 | |
| 6 | 6,7-Dimethyl-8-(1-D-ribityl)lumazine | 28 | 1 | |
| 7 | Methaneselenol | 26 | 4 | Se |
| 8 | α-Carotene | 22 | 22 | |
| 9 | Tetradecanoyl-[acp] | 21 | 16 | |
| 10 | α-Pinene | 21 | 12 | |

*Note*. Shown are the new seed compounds identified by the algorithm, the number of compounds by which the combined scope increases, the number of alternative seeds, and the chemical elements as they appear in the reconstruction process.

### *Combined Scopes of Small Building Blocks*

We have shown that APS as well as the other three compounds, PAPS, dephospho−CoA, and UDP−6−sulfoquinovose, possess the largest scope. These compounds are rather complex and produced by intracellular processes; for example, APS is produced by the enzyme sulfate adenylyltransferase, converting ATP and sulfate into APS and pyrophosphate. It is an intriguing question whether a network of size similar to the scope of APS can be obtained as a combined scope of a small number of simple compounds which are present in the environment. Guided by the elements contained in APS (see Table 1), we select the following set of seed compounds: $CO_2$, $NH_3$, $H_3PO_4$, and $H_2SO_4$. Interestingly, the set of compounds which can be synthesized from these simple compounds is exactly the same as the set of compounds which can be produced from APS. In other words, the combined scope of the former small compounds is identical to the scope of APS. However, in the case of APS, this scope is reached after 55 generations, whereas in the case of the small building blocks 61 generations are required.

Figure 6 depicts the expansion process starting with these four simple seed compounds. Its characteristic features are similar to those found in the expansion of the scopes of glucose and ATP (see Figs. 2 and 3). The expansion proceeds slowly at the beginning as well as near the end of the process. In between there are phases of deceleration and subsequent acceleration after the incorporation of key compounds such as ATP or $NAD^+$. Even more similarities are revealed when comparing the process with the expansion of APS (dashed line in Fig. 6). Both curves show three pronounced peaks. The first peak, appearing in both curves around generation 10, can be explained by the incorporation of the majority of carbohydrates. In both processes, the second strong increase in the number of incorporated compounds

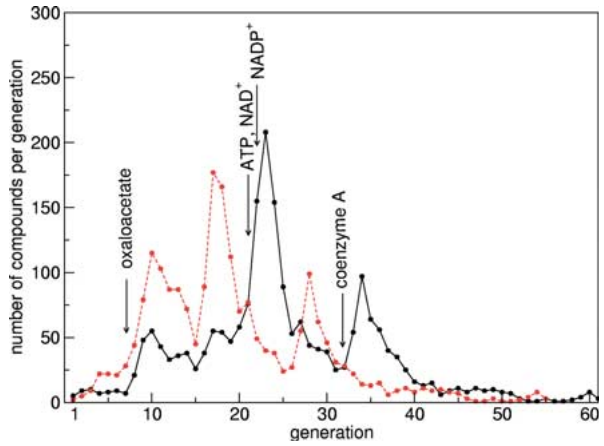


**Fig. 6.** The expansion process of the seed $CO_2$, $NH_3$, $H_3PO_4$, and $H_2SO_4$. Arrows indicate the incorporation of some key compounds leading to an acceleration of the expansion process. NADH and NADPH are synthesized in the generation following the acquisition of $NAD^+$ and $NADP^+$, respectively. For comparison, the expansion process of APS is shown (dashed line), which proceeds faster but results in the same final network.

results from the incorporation of $NAD^+$ into the network. Notably, in the expansion process starting with the four simple compounds this peak is shifted toward later generations. This can be understood by the fact that the adenine nucleotides, which are important precursors in the synthesis of $NAD^+$ can simply be extracted from APS while they have to be synthesized from the simple compounds in a larger number of reactions. After the second peak the two expansions proceed in an almost-identical way. From that we conclude that in this stage both expansion processes, one beginning with a single complex compound and the other starting with four simple molecules, have reached almost the same subnetwork. Consequently, in both processes the numbers of generations between the second and the third peaks are the same.

Starting the expansion process only with the building blocks $CO_2$, $NH_3$, and $H_3PO_4$, i.e., omitting the sulfur source, results in a combined scope which is identical to the scope of ATP. However, the expansion from the building blocks requires eight more generations. We have tested whether the expansions proceed similarly when replacing the carbon source $CO_2$ by $CH_4$. Interestingly, the expansion processes stop in very early stages, in generation 6 when sulfate is included and in generation 5 when it is excluded. They result in scopes of size 25 and 19, respectively, containing predominantly inorganic compounds. In both cases, the small scope sizes are due to the fact that all reactions utilizing methane require the presence of cofactors in a very early stage. The role of the cofactors is investigated further in the next section.

We have also calculated the combined scope of a set of compounds which are proposed as hypothetical inorganic precursors for the origin of life, namely, $H_2CO$ (carbonic acid), $CH_3SH$ (methanethiol), $NH_3$, and $P_2O_7^{4-}$ (pyrophosphate); see Martin and Russell (2003). Remarkably, the combined scope of these compounds is again identical to the scope of APS, the largest scope of the complete network. Extending the seed by $CO_2$, $CH_4$, and $CN^-$ (cyanide), compounds which are also discussed by Martin and Russell (2003), does not further increase the scope size.

### *Role of Cofactors*

We have shown above that expansion processes strongly accelerate after the appearance of the cofactors $NAD^+$/NADH and $NADP^+$/NADPH (see Figs. 3b and 6). The reason for this is the high number of reactions in which these compounds participate. For example, there are 583 reactions under participation of $NAD^+$ and 574 reactions under participation of NADH. Five hundred seventy-two reactions are redox reactions in which both compounds participate as cofactors. Thus, in the generations following the acquisition of $NAD^+$ and NADH, all these reactions become available for the expansion of the network provided that other substrates of them are already present.

It is an intriguing question in what way the expansion process is affected if these redox reactions are possible even without the presence of $NAD^+$ ($NADP^+$) or NADH (NADPH). Chemically, such reactions are in principle feasible by participation of other, for example, inorganic, electron acceptors, such as $Fe^{3+}$. Therefore, we studied modified expansion processes by allowing from the very beginning the incorporation of all redox reactions which would otherwise require the presence of the cofactors $NAD^+$/NADH or $NADP^+$/NADPH. Such a process, starting with the seed compounds $CO_2$, $NH_3$, $H_3PO_4$, and $H_2SO_4$ is depicted in Figure 7a (solid line). Since the original expansion (dashed line) includes the synthesis of these cofactors, the modified expansion results in the same final network but in a smaller number of generations. As expected, the strong increase in the speed of the expansion, which in the original process is evoked by the appearance of $NAD^+$, takes place much earlier. There is, however, an initial lag phase of about seven generations, which is required for the synthesis of those compounds participating in redox reactions.

We have also analyzed how other cofactors affect the expansion process. First, we assume that all phosphorylations which normally take place at the expense of ATP can also occur without the presence of ATP (by direct uptake of inorganic phosphate). This leads to the modified expansion process depicted in Figure 7b. Clearly, the modification results in an accelerated expansion but the effect is less pro-

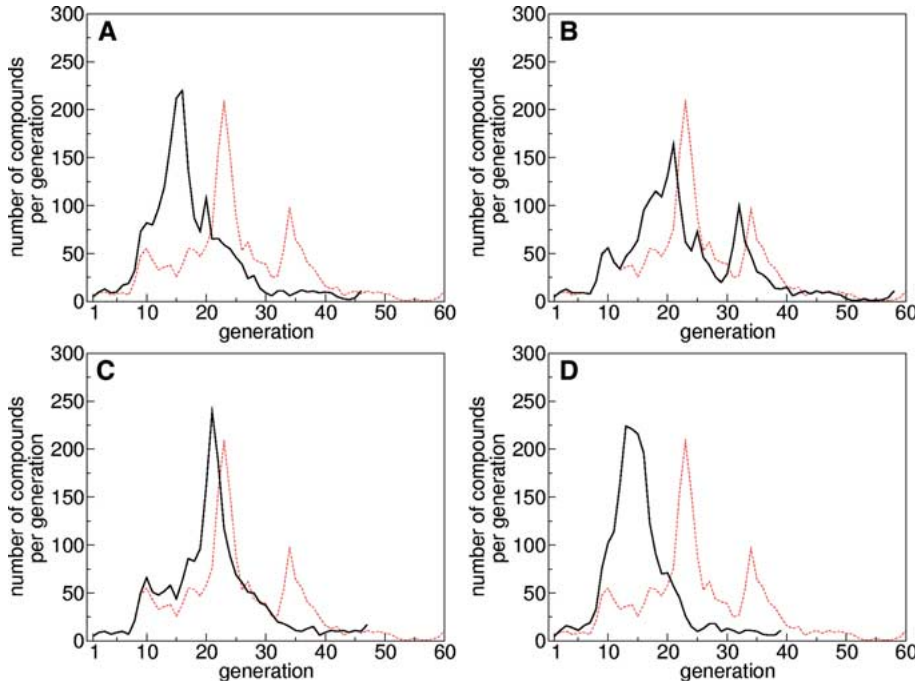


**Fig. 7.** Influence of the cofactors on the expansion process. In each part two expansions of the seed $CO_2$, $NH_3$, $H_3PO_4$, and $H_2SO_4$ are shown. The dashed lines represents the original process (cf. solid line in Fig. 6). The solid lines represent modified expansions where the functionalities of the cofactors are taken into account from the very beginning: (**A**) $NAD^+$ and $NADP^+$, (**B**) ATP, (**C**) CoA, and (**D**) $NAD^+$ and $NADP^+$, ATP, and CoA.

nounced than in the case of the replacement of the cofactor $NAD^+$. Second, we assume that acyl groups can be transferred even without the presence of coenzyme A. The resulting process is depicted in Figure 7c. Again, the process is accelerated, and as expected, the peak resulting in the original process from the incorporation of CoA does not appear.

Moreover, we considered the case where from the beginning the functionalities of the three cofactors are available. Not surprisingly, this combined replacement results in a process (Figure 7d) which is faster than any process obtained by replacement of a single cofactor. Whereas the original expansion process, and to a lesser extent also the modified processes with single replacements, show phases of temporary slow-down, this property is no longer observed in the case of combined replacement. Instead, the process accelerates continuously until generation 15 and subsequently decelerates.

Starting with other seed compounds, which in the original process do not produce the considered cofactors, the addition of the functionalities of the cofactors will generally lead to an increased size of the resulting network. The corresponding effects can be dramatic. For example, the original scope size of $CO_2$ is 17. Adding the functionalities of $NAD^+$, $NADP^+$, and CoA leads to a resulting network which contains 682 compounds. This again underlines the importance of cofactors in the establishment of large–scale biochemical reaction networks.

*Robustness of Scopes*

Scopes represent functional modules in the sense that they comprise all compounds which can be synthesized from the seed substrates. Which compounds are contained in the scope depends on the set of available reactions. In the calculations presented above we used all reactions which are currently included in the reference set of the KEGG database. As an extension of our previous work (Ebenhöh and Heinrich 2003), we analyze in the following the robustness of network functions against structural modifications. In particular, we investigate how the scopes are affected if the base set of reactions is changed.

Modifications can be caused by deletions of genes coding for metabolic enzymes, resulting in a loss of the corresponding biochemical conversions. Stronger modifications may result when considering specific metabolic networks of different organisms. Each organism has its own base set of reactions which is a subset of the reference set. It is of particular interest how robust the results described above are against such changes in the set of available reactions. Moreover, such analyses are of relevance since the KEGG database is not complete. If scopes turn out to be robust against modifications, one may expect that future updates of the KEGG database will not change the results dramatically.

We have calculated the effects of all single deletions from the base set of reactions on the scope sizes. In each case, those reactions were considered which are associated with the corresponding scope. In Figure 8, the resulting effects are depicted for the scopes of ATP, APS, and glucose and for the combined scope of $CO_2$, $NH_3$, $H_3PO_4$, and $H_2SO_4$. Each plot shows in how many cases a single deletion reduces the scope size by a given number of compounds. The diagram reveals that in all cases the majority of such deletions

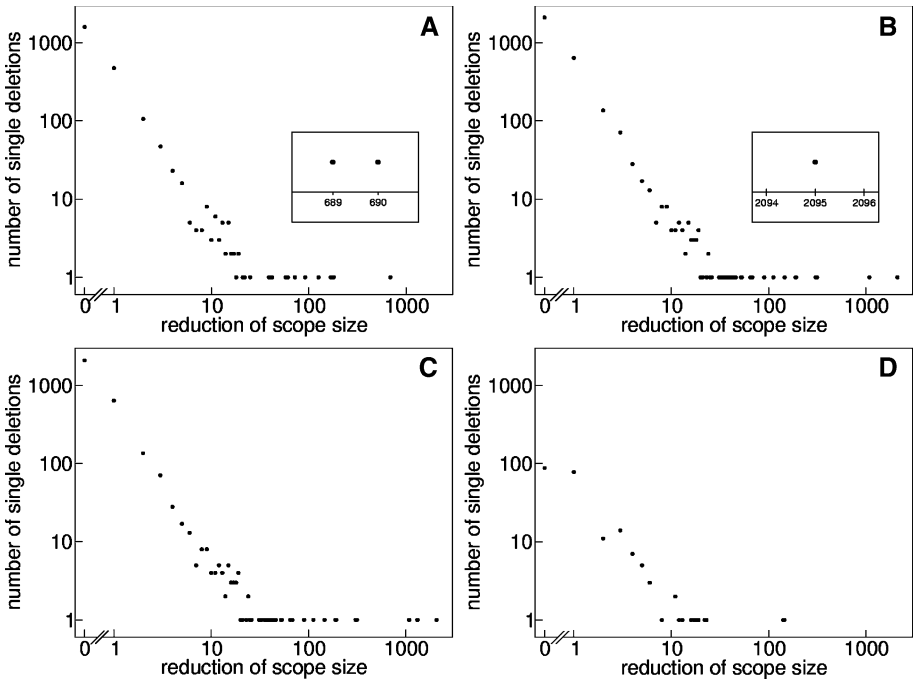


**Fig. 8.** Robustness of scopes. Plotted is the number of single deletions resulting in a given reduction of the scope sizes of (**A**) ATP, (**B**), APS, (**C**) the set $CO_2$, $NH_3$, $H_3PO_4$, $H_2SO_4$, and (**D**) glucose.

does not affect the scope size at all. Most of the other deletions have only a small effect on the scope size. However, there are a few reactions whose deletion from the base set significantly reduces the scope sizes.

Let us consider, as an example, the robustness of the scope of ATP (Fig. 8a). The network resulting from the expansion of ATP contains 1557 compounds and 2328 reactions. For 1589 reactions, a single deletion does not influence the scope size. For another 718 deletions, the reduction of the scope size is smaller than 20. Among the 21 reactions whose deletions have a larger effect, there exist two which are very critical for the scope size. The deletions of these reactions result in a reduction of the scope size by 689 and 690 compounds, respectively. A closer inspection reveals that these two reactions are

$$\text{L-Aspartate} + O_2 \leftrightarrow \text{Iminoaspartate} + H_2O_2 \quad (3)$$

catalyzed by the enzyme L-aspartate oxidase (EC 1.4.3.16) and

$$\text{Pyridine-2,3-dicarboxylate} + 2\ H_2O + \text{Phosphate} \leftrightarrow \text{Iminoaspartate} + \text{Glyceronephosphate} \quad (4)$$

catalyzed by a carbon lyase (EC 4.1.99.–). The fact that there exist a small number of reactions whose deletion affects the scope size dramatically is also visible in the other examples shown in Figure 8. In the case of APS, a deletion of the reaction

$$\text{APS} + H_2O \leftrightarrow \text{AMP} + \text{Sulfate} \quad (5)$$

which is catalyzed by the enzyme adenylylsulfate sulfohydrolase, results in a total collapse of the scope (Fig. 8b). This devastating effect can be explained by the fact that this reaction is the only one within the reference set which can metabolize the compound APS using no other compound except water. The elimination of this reaction stops the expansion process at the very beginning.

As mentioned above, the combined scope of $CO_2$, $NH_3$, $H_3PO_4$, and $H_2SO_4$ is exactly the same as the scope of APS. As expected, the analysis of the effect of single deletions yields similar results in both cases (cf. Figs. 8b and c). However, there exist differences for those reactions which have a strong impact on the scope size. Analysis of the robustness of the scope of glucose, which is significantly smaller than in the other examples (197 compounds connected to 220 reactions), shows a similar behavior (Fig. 8d).

Now we examine how the simultaneous removal of more than one reaction affects the scope sizes. Figure 9 shows how the size of the ATP scope decreases with increasing number of removed reactions. Specifically, a two−dimensional distribution is shown, with the shading indicating the probability that a reduction of a given number of reactions results in a certain scope size.

There exist distinct domains in which these probabilities are high. This results from the fact that there are a few reactions which alone have a strong impact on the scope size. For example, domain A contains cases in which one of the two reactions (3) and (4), reducing the scope size in single deletions by 689 or 690 compounds, has been removed. Similarly, domain B contains those cases in which at least one of a

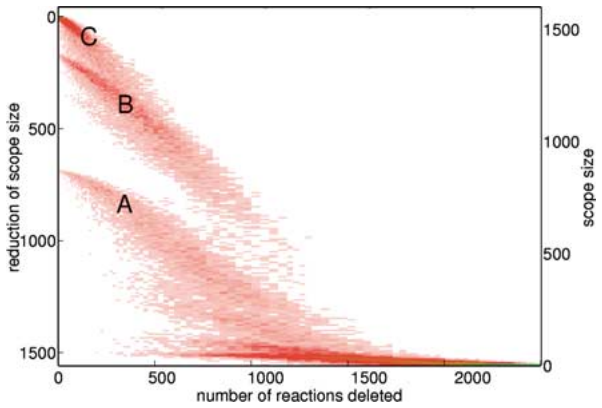


**Fig. 9.** The effect of multiple deletions on the scope size of ATP. The shading indicates the probability that the scope has a certain size ($y$-axis) when randomly deleting a certain number of reactions ($x$-axis). For any chosen number of deleted reactions, 400 random samples were calculated.

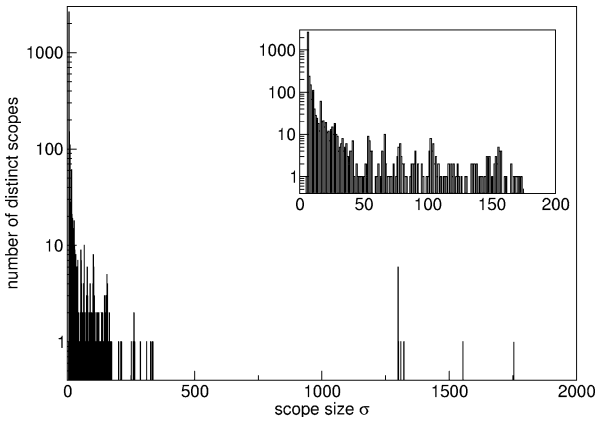


**Fig. 10.** Histogram of scope sizes taking into account directionality of reactions showing how many distinct scopes have a certain size σ. The inset is a magnification for small scope sizes.

group of reactions has been removed, which reduces the scope size by about 190. There exists a third domain (C) which contains those cases in which only reactions have been removed which do not have a strong effect on the scope size when considering single deletions. The existence of separate domains and the fact that within each domain the scope sizes decrease uniformly with increased number of reduced reactions indicates that the scope size is critically influenced by only a small number of reactions, whereas it is robust against elimination of most reactions.

### *Irreversible Reactions*

As argued in the beginning it makes sense to perform the expansion under the assumption that all reactions can in principle take place in forward as well as backward directions. However, network expansion can also be performed in the case that a part of the reactions takes place only in one direction. For that, the rules for the incorporation of reversible reactions remains unchanged (either all substrates or all products must be present in the network), whereas irreversible reactions can only be attached if all substrates are available.

Consideration of unidirectionality of some reactions will in general reduce the scope sizes, which can be seen as follows. First, all reactions are assumed to be reversible and all of them are replaced by two opposite unidirectional reactions. In each generation only those reactions are attached which can use the already existing compounds as substrates in its prescribed direction. Obviously, the scopes obtained in this way remain unchanged. Second, the existence of irreversible reactions is taken into account by eliminating those unidirectional reactions which are not allowed. Thus the consideration of unidirectionality corresponds to a deletion of reactions from the base set. As described in the previous section, this results in a decrease in the scope size.

We refrained from performing systematic calculations, since the necessary information about unidirectionality is still incomplete. However, a preliminary classification into reversible and irreversible reactions can be found in the KEGG database (2186 of 5311 reactions are considered to be irreversible). Therefore we performed some calculations by taking into account this classification. The distribution of scope sizes is shown in Figure 10. It is seen that the consideration of unidirectionality does not change the general characteristics of this distribution, cf. Figure 1. There exist a few oustandingly large scopes and a high number of small scopes. As expected, the scope sizes tend to be smaller when considering unidirectionality. For example, the scopes in Figure 1 with sizes around 500 do not exist anymore, whereas there are more scopes with smaller sizes in Figure 10. The property of interconvertibility of two compounds may be lost when considering unidirectionality meaning that the number of distinct scopes cannot decrease. In fact, the number of distinct scopes rises from 3345 to 3877. For example, the 103 seed compounds resulting in the scope of ATP in the case of reversible reactions, give rise to 47 different scopes when taking irreversible reactions into account.

### *Emergence of Pathways*

During expansion, an increasing number of reactions and metabolites are incorporated. It is an intriguing question in which stage of the process fundamental biochemical pathways have fully emerged. We expect that the answers will help us to understand the evolutionary history of metabolic networks.

For a corresponding analysis we again performed an expansion process with all available reactions

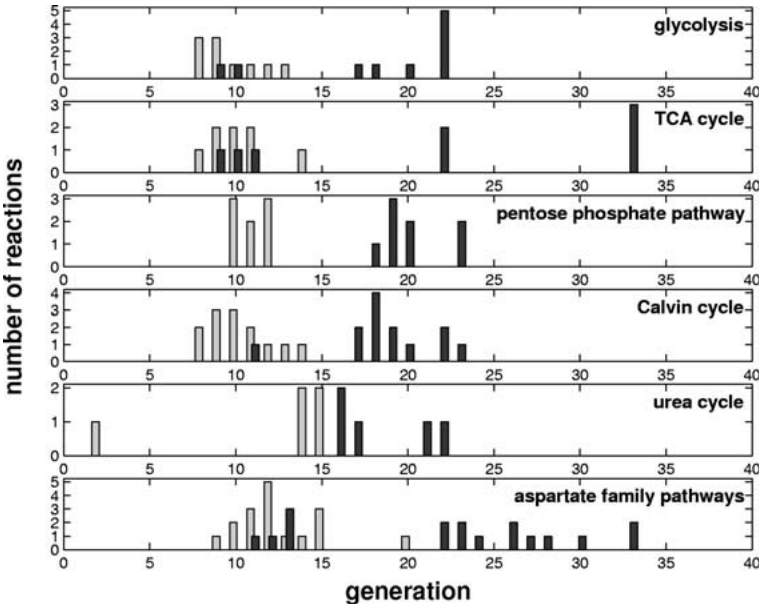


**Fig. 11.** Incorporation of pathways. The figure shows in which generation how many reactions of certain pathways appear in an expansion starting with the seed $CO_2$, $NH_3$, $H_3PO_4$, and $H_2SO_4$. The gray bars characterize the expansion with the functionalities of the cofactors taken into account from the very beginning. The black bars refer to the unmodified expansion.

using the small building blocks $CO_2$, $NH_3$, $H_3PO_4$, and $H_2SO_4$ as seed compounds. We focused on the appearance of the reactions constituting the following pathways: glycolysis, TCA cycle, pentose phosphate pathway, calvin cycle, urea cycle, and the pathways of the aspartate family.

Figure 11 shows in which temporal order the reactions of these six pathways are recruited. The bars indicate how many reactions of a given metabolic pathway are incorporated in a certain generation. Black bars refer to the unmodified expansion process in which cofactors have to be produced during the expansion, whereas gray bars refer to the modified process where the functionalities of the cofactors are present from the very beginning.

As expected, in all cases the pathways are discovered later if the cofactors have to be synthesized from the seed compounds. A comparison of Figures 7d and 11 reveals that the majority of the reactions of all these pathways are recruited in the middle phase where the expansion proceeds fastest (generations 10–35 for the unmodified process, generations 7–20 for the modified process).

A striking feature of these expansions is that the various metabolic pathways appear more or less in parallel. There is no case in which one pathway must be completely present before the reactions of another pathway can be incorporated. For example, all reactions of glycolysis have been recruited in generation 22. At this stage already five of the eight reactions of the TCA cycle have been included. Consequently, these reactions make use of substrates which are not provided by glycolysis but synthesized from the seed compounds by other reactions running in parallel.

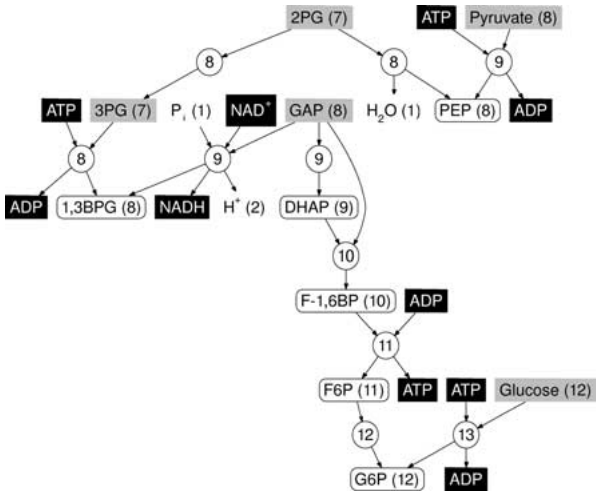


**Fig. 12.** Recruitment of the glycolytic reactions and metabolites in an expansion starting from the seed compounds $CO_2$, $NH_3$, $H_3PO_4$, and $H_2SO_4$. The numbers indicate the generation of appearance. Compounds in black boxes are cofactors whose functionalities are assumed to exist from the beginning. Gray boxes indicate those compounds which appear in the expansion for the first time as products of nonglycolytic reactions. White boxes indicate those compounds which appear for the first time as products of glycolytic reactions.

Figure 12 shows the course of the emergence of the glycolytic reactions for the expansion process where the functionalities of the cofactors are present from the beginning. The numbers at the reactions and compounds indicate in which generation they appear in the expansion. It is observed that the glycolytic reactions are not acquired in the same sequence as glycolysis proceeds. There are five compounds of glycolysis (glucose, GAP, 3PG, 2PG, and pyruvate) which in the expansion are provided by nonglycolytic reactions. Of the glycolytic reactions producing the

remaining compounds, those are incorporated first (in generations 8 and 9) which interconvert the three carbon sugar phosphates DHAP, 1, 3−BPG, and PEP. Subsequently (in generations 10–13), reactions involving the conversion of hexose phosphates are acquired.

In summary, our investigations suggest that metabolic reactions are not recruited in the same order as they are traditionally arranged in biochemical pathways starting from an initial substrate and ending in a final product. Our results also do not support the view that metabolic pathways evolve in a retrograde way. In this concept, proposed by Horowitz (1945), reactions producing a required end product appear first during evolution, and when the necessary substrates are no longer available, new reactions producing these compounds are incorporated into the metabolism. The fact that in the reconstruction of glycolysis, as depicted in Fig. 12, the lower part of the pathway appears prior to the upper part is not a consequence of a special demand for the production of pyruvate. It can rather be explained by the use of small building blocks as seed compounds, leading to an expansion in which the incorporation of small chemicals generally precedes the acquisition of large compounds.

## Discussion

In this work we present a new concept for the structural analysis of metabolic systems and its application to large−scale networks. This method is based on a hierarchy concerning possible successions of reactions, meaning that a particular reaction can only take place if its substrates are provided by certain other reactions. In particular, for given initial substrates, the interdependence of the reactions is not symmetrical. If reaction $Y$ is dependent on the occurrence of reaction $X$, this does not necessarily imply that reaction $X$ is dependent on reaction $Y$. The proposed network expansion is a structural method in the sense that only substrate–product relationships are included, reflecting the stoichiometry of the biochemical reactions, whereas kinetic properties are not taken into account.

Our approach differs from graph theoretical methods (Wagner and Fell 2001; Jeong et al. 2000; Milo et al. 2002) which represent metabolic networks as a set of vertices and edges, where vertices represent biochemical compounds and edges biochemical reactions. Such methods have the disadvantage that they are unable to reflect the precise substrate–product relations for reactions using two or more substrates or products. Therefore, relevant stoichiometric information is lost, which leads to the problem that a graph may be uniquely constructed from a metabolic network, whereas the original network cannot be reconstructed from the graph. This problem is not encountered in the expansion method presented in this work since it directly uses the metabolic pathway information without any simplification concerning its stoichiometric properties.

There exist in the literature other structural approaches which also use the full metabolic information. This concerns, for example, the method for calculating elementary flux modes (Schuster and Hilgetag 1994), representing minimal enzyme sets allowing for nonzero steady-state fluxes. Similarly to the elementary modes, the networks associated with the scopes represent functional subnetworks of the complete network. While the purpose of the concept of elementary modes is the calculation of possible flux distributions between initial substrates and end products, the aim of the expansion method is the reconstruction of biochemical reaction networks based on initial substrates and available reactions.

A disadvantage of the concept of the elementary modes is the combinatorial explosion of the number of modes when considering large networks. This generally holds true for each specific choice of external substrates and products (Dandekar et al. 2003). This implies that even for medium-sized networks a systematic analysis of all elementary flux modes becomes unfeasible. In contrast, for each specific choice of seed compounds there exists exactly one scope.

The expansion method and the concept of scopes can be used to answer the question which compounds can in principle be synthesized from a given set of substrates using a specified set of biochemical reactions. Moreover, we can identify reactions which are essential for the sustainment of the full scope as well as reactions which are redundant in the sense that their elimination does not affect the scope.

Furthermore, we can identify sets of compounds which can mutually be converted into one another. The compounds belonging to such sets have the property that they have the same scope. Moreover, the calculation of the scopes allows for certain conclusions about the chemical composition of the compounds. First, a compound B within the scope of another compound A must consist of a subset of the chemical elements of compound A. Second, two compounds having the same scope must consist of the same chemical elements. However, not all compounds consisting of the same set of chemical elements are interconvertible.

The method allows identification of small chemical building blocks from which a majority of all compounds can be synthesized. We could show, for example, that almost half of all compounds can be synthesized from the four building blocks carbon

dioxide, ammonia, sulfate, and phosphate together with water.

All calculations presented in this work have been performed using information contained in the KEGG database. Dealing with this database, we encountered the following problems. First, some compounds appear in the database more than once, with different names representing different levels of specification. For example, in some reactions the type of amino acid is not specified. This means that a reaction which uses, for example, alanine as a substrate, but is included in the KEGG database as a reaction using an unspecified amino acid, will not be incorporated during the expansion process even though alanine is already available. Second, the database includes compounds which represent chains of unspecified lengths such as fatty acids. This poses the problem that several distinct substances are represented by only one compound in the database. Another related problem arises from the fact that the oxidation levels of some inorganic chemicals such as iron are not included in the description. For example, $Fe^{2+}$ and $Fe^{3+}$ are considered as the same substance. Since these cases are rare, we expect that the overall expansion processes are not significantly influenced. These problems could be overcome by improving the database in such a way that all compounds are completely specified.

We expect that the expansion processes as introduced in this work show features which may also be characteristic for the evolution of metabolic systems. Both types of processes start from simple networks containing only a small number of reactions and evolve toward networks of very high complexity. Further, they share the property that certain reactions can only perform their function if other reactions exist providing the required substrates. One may argue that these dependencies imply a certain temporal order in which the different reactions may appear during evolution.

We have shown that expansion processes starting from simple inorganic building blocks proceed in such a manner that reactions as well as their products are incorporated in a well–defined order, where, as a tendency, compounds with a complex chemical structure appear later. While carbohydrates appear early in the expansion process, structurally complex cofactors such as ATP and $NAD^+$ appear later but still before coA. We have also shown that the capability to synthesize these compounds dramatically increases the capacity to incorporate new types of reactions.

We therefore hypothesize that the discoveries of such key metabolites have been critical events also during the evolutionary history of metabolic networks. This does not contradict the possibility that some functions of these metabolites, for example, the oxidizing capacity of $NAD^+$, could be carried out by other chemicals which were abundant in the environment. However, it is evident that cells having the ability to autonomously synthesize electron acceptors from simple building blocks will have a strong evolutionary advantage.

An interesting result of our investigation is that simple chemicals which are supposed to have been present under prebiotic conditions are possible seeds for very complex biochemical reaction networks covering a significant part of the present metabolism.

All calculations performed in this work are based on a reference set comprising all reactions present in the KEGG database irrespective of the organism in which they have been found. Analogous analyses can be performed on species–specific networks simply by changing the base set of reactions. It is challenging to extend our methods in such a way that they provide information not only about the temporal appearance of single reactions but also about the origination of whole metabolic networks of specific organisms from their precursors. We expect that such an extension will be helpful to reconstruct phylogenetic trees of organisms based on their metabolic information.

Our investigation gives reason to assume that the history of the evolution of metabolism can be, to some extent, deciphered from the existing metabolic networks.

## References


Dandekar T, Moldenhauer F, Bulik S, Bertram H, Schuster S (2003) A method for classifying metabolites in topological pathway analyses based on minimization of pathway number. BioSystems 70:255–270

Ebenhöh O, Heinrich R (2003) Stoichiometric design of metabolic networks: multi-functionality, clusters, optimization, weak and strong robustness. Bull Math Biol 65:323–357

Ebenhöh O, Handorf T, Heinrich R (2004) Structural analysis of expanding metabolic networks. Genome Informatics 15:35–45

Heinrich R, Schuster S (1996) The regulation of cellular systems. Chapman and Hall, New York

Heinrich R, Schuster S, Holzhütter HG (1991) Mathematical analysis of enzymic reaction systems using optimization principles. Eur J Biochem 201:1–21

Horowitz NH (1945) On the evolution of biochemical syntheses. Proc Natl Acad Sci 31:153–157

Jeong H, Tombor B, Albert R, Oltvai ZN, Barab'asi AL (2000) The large–scale organization of metabolic networks. Nature 407:651–654

Kanehisa M (1997) A database for post–genome analysis. Trends Genet. 13:375–376

Kanehisa M, Goto S (2000) KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acid Res 28:27–30

Kauffman KJ, Prakash P, Edwards JS (2003) Advances in flux balance analysis. Curr Opin Biotechnol 14:491–496

Martin W, Russell MJ (2003) On the origins of cells: a hypothesis for the evolutionary transitions from abiotic geochemistry to chemoautotrophic prokaryotes, and from prokaryotes to nucleated cells. Phil Trans R Soc Lond B 358:59–85

Milo R, Shen–Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U (2002) Network motifs: simple building blocks of complex metabolic networks. Science 298:824–827

Papin JA, Price ND, Wiback SJ, Fell DA, Palsson BO (2003) Metabolic pathways in the post–genome era. TIBS 28:250–258

Schomburg I, Chang A, Schomburg D (2002) BRENDA, enzyme data and metabolic information. Nucleic Acids Res 30:47–49

Schuster S, Hilgetag C (1994) On elementary flux modes in biochemical reaction systems at steady state. J Biol Syst 2:165–182

Schuster S, Fell DA, Dandekar T (2000) A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. Nature Biotechnol 18:326–332

Wagner A, Fell DA (2001) The small world inside large metabolic networks. Proc R Soc Lond B 268:1803–1810