11 Reconstruction of the stoichiometry of ATP and NADH-producing systems using evolutionary algorithms

O. Ebenhöh and R. Heinrich

Humboldt-Universität zu Berlin, Institut für Biologie, Theoretische Biophysik, Invalidenstraße 42, 10115 Berlin, Germany

It is theoretically analysed whether the structural properties of ATP and NADH-producing pathways, especially glycolysis and the citric acid cycle, can be explained by optimisation principles. It is assumed that real pathways have reached, as a result of their evolution, a high efficiency with respect to ATP production rates. On the basis of kinetic and thermodynamic principles conclusions are derived concerning the stoichiometry of such pathways (see [1]).

We consider reaction sequences converting an initial substrate (e.g., glucose) into a final product (e.g., lactate under anaerobic conditions or oxaloacetate under aerobic conditions) using the drop in free energy to produce a certain amount of ATP and NADH molecules. The pathways interact with a metabolic system producing ATP through the reducing power of NADH (in cells of aerobic organisms this corresponds to the functioning of oxidative phosphorylation). We also consider the alternative that NADH is being consumed directly without the production of ATP. Further, external ATP-consuming processes are incorporated.

It is examined in what respect real metabolic systems can be considered optimal by analysing a large number of alternative pathways and comparing them with respect to the resulting steady state ATP production rate. A formalism has been derived allowing for the description of chemically feasible alternative pathways. Alternative systems are generated by assembling *generic* reactions under consideration of a number of rules and boundary conditions, such as limiting the maximal number of phosphate groups bound to any intermediate to two. These rules ensure that all examined pathways are chemically feasible and represent theoretically possible reaction sequences. The set of generic reactions has to be defined according to the metabolic system on which the structural analysis is to be performed. As this work focuses on the analysis of ATP and NADH-producing systems, we restrict the set of generic reactions to ATP production (a), ATP consumption (A), NADH production (n), NADH consumption (N), phosphorylation (P)

and dephosphorylation (p) involving inorganic phosphate, reduction (H) and oxidation (h) and so-called *uncoupled* reactions (u). All reactions apart from the uncoupled ones act on the ligands but do not change any internal molecular structure of the substrate. The uncoupled reactions always leave the ligands unchanged but perform changes on the internal structure. Assembling these generic reactions in any order that fulfils the given boundary conditions yields an unbranched chain of chemical reactions which we denominate by strings of characters, with each character standing for exactly one generic reaction. All such pathways have in common that they produce a certain number of ATP and NADH molecules by consuming an energy-rich substrate.

For an unbranched chain C comprising r_C reactions, the steady state rate J_C can be expressed as (see [2])

$$J_C = \frac{X_0 \prod_{j=1}^{r_C} q_j - X_{r_C}}{\sum_{j=1}^{r_C} \frac{\tau_j(1+q_j)}{q_j} \prod_{k=j}^{r_C} q_k}. \tag{11.1}$$

Here, q_j and τ_j denote the equilibrium constants and relaxation times for the j-th reaction in chain C. For the reactions a, A and n, N these quantities depend on the concentrations of the corresponding cofactors. The concentrations of the initial substrate X_0 and the final product X_{r_C} are considered to be constant. The steady state rate for ATP consumption is assumed to be

$$J_{ATPase} = k_{ATPase} \cdot ATP, \tag{11.2}$$

where k_{ATPase} is a rate constant. The steady state rate of the production of ATP using the reducing power of NADH is expressed as

$$J_{Ox} = k_{Ox} \cdot ADP \cdot NADH, \tag{11.3}$$

and the rate of decoupled consumption of NADH as

$$J_d = k_d \cdot NADH, \tag{11.4}$$

where k_{Ox} and k_d denote the corresponding rate constants. The balance equations for ATP and NADH

$$d \cdot J_C - J_{ATPase} + \gamma J_{Ox} = 0, \tag{11.5}$$

$$n \cdot J_C - J_{Ox} - J_d = 0, \tag{11.6}$$

form two equations with two variables (ATP and $NADH$). Here, d and n denote the net number of produced molecules ATP and NADH, respectively, per consumption of one molecule glucose in reaction chain C and γ denotes the number of molecules ATP produced per consumption of one molecule NADH by flux J_{Ox}.

A realistic value is $\gamma = 3$ (see [3]). These equations can be solved numerically using the explicit expressions (11.1)–(11.4) for the fluxes.

By this means it is theoretically possible to examine all allowed arrangements of generic reactions and to calculate the resulting steady state ATP production rate they yield. However, the number of possible pathways is immense and therefore a different approach has been chosen.

Optimisation analysis has been performed by an evolutionary algorithm. This algorithm is initialised by generating a chosen number (*population*) of arbitrary pathways. This was made possible by the formalism described above. The size of the population can be varied, but during one special simulation it is kept constant. On these randomly generated sequences we repeatedly apply selection and mutation operations. The functionality of the whole algorithm strongly depends on reasonable definitions of these two operations.

Selection is carried out by defining a fitness function that depends monotonously on the steady state ATP production rate. Reaction sequences are doubled (they *reproduce*) with a probability proportional to their fitness. The size of the population is then reduced to its original size by *dilution*, i. e. the necessary number of sequences to be eliminated is picked randomly. Due to this selection process sequences yielding a higher steady state ATP production rate have a higher chance of reproduction and therefore of *survival* (survival of the fittest).

The mutation rules have been constructed on the basis of the formalism describing the reaction chains. A necessary condition for the evolutionary algorithm to work is that for any two arbitrary chosen but feasible sequences C_1 and C_2 there must exist a finite number of mutation operations which, when applied on C_1, result in C_2. The set of mutation classes we propose in the present work meets this condition. Further, the set is *minimal* in the sense that if any one mutation class is taken away from the set, the condition is not met. The importance of this condition is evident because any search algorithm must in principle be able to cover the whole space to be searched, here the sequence space consisting of all possible reaction chains. Moreover, the mutations have been defined in such a way that the changes resulting from the operations are small. The number of characters changed in one string when applying a mutation is always less than or equal to three.

Repeated application of selection and mutation operations result in a very efficient search for sequences yielding high steady state ATP production rates. The best sequences that were found during several simulations with the limitation that the number of ATP-consuming reactions occurring in the sequence does not exceed two, are given in Table 11.1.

All these sequences yield a high ATP production rate, the rate of the tenth sequence only being a 10^{10}-th fraction smaller than that of the first. On the other hand, randomly generated sequences generally do not yield an efficient production rate. The chance of randomly picking a sequence that yields a production rate larger than 90 % of the optimised rate is around 0.05 %, which demonstrates

Table 11.1 The best ten sequences fulfilling the boundary conditions given in the text.

1	hAhApNuhpuPHuhpHuHnnPHupHnnuHPnPaHaH
2	hAhApNpuuHnunHHununuHPnPaHaH
3	hAhApNpuHununHHununuHPnPaHaH
4	hAhApNupHhunuHPunpHHunnHunPPaHaH
5	hAhApNupnuuHHnunuHHnuPnPaHaH
6	hAhApNpuHnuunHuHnunuHPnPaHaH
7	hAhApNpuHnuunHHununuHPnPaHaH
8	hAhApNupnuHuuHnnuPHpHnuPnPaHaH
9	hAhApNuhpNHnunPuHpnNHunnHunHuPnPaHaH
10	hAhApNupHhhPupPHpHunPunpHHnunHunPPaHaH

the efficiency of the algorithm.

We found that all optimised sequences have certain common structural properties. Most obvious is the fact that all optimised sequences begin with the subsequence "hAhApN..." and end in the subsequence "...PaHaH". We also find this feature in glycolysis where two ATP-consuming (A) reactions catalysed by hexokinase and phosphofructokinase are situated at the beginning of the metabolic pathway and the ATP-producing (a) reactions catalysed by phosphoglycerate kinase and pyruvate kinase are located near the end.

Another important common feature is that the difference between the numbers of NADH-producing (n) and NADH-consuming (N) reactions amounts to four. This means that per consumed molecule of the initial substrate (e. g. glucose) four NADH molecules are effectively produced. This feature is found in the citric acid cycle where in each cycle exactly four reactions produce a molecule with a high reduction potential. These reactions are catalysed by isocitrate dehydrogenase (NADH), the α-ketoglutarate dehydrogenase complex (NADH), succinate dehydrogenase ($FADH_2$) and malate dehydrogenase (NADH).

The question arises if all efficient sequences necessarily have a similar internal structure. If so, these sequences could be interpreted as members of one quasi-species. Despite all common features, structural differences between efficient sequences can indeed be uncovered. For this purpose, two different analytical methods have been developed. In order to quantify the similarity / difference between sequences, a distance measure has been defined based on a Hamming distance. Using this tool, we examine the similarity between optimised sequences. We obtain the result that efficient sequences appear in "clusters". Within each cluster the sequences are very similar to each other, sequences from different clusters show large differences. This result is interpreted as the occurrence of different quasi-species all yielding very efficient ATP output rates. Interestingly, on a biological level this result means that there exists the possibility of several ATP and

NADH-producing pathways that show differences in their structural design but not in their efficiency regarding ATP production.

In order to understand the structural differences between quasi-species we defined a function describing the relative position of two types of generic reactions within one reaction chain. Thus we analyse which classes of reactions are preferable at the beginning of the chain and which are better located at the end. The general result for all efficient sequences is in agreement with features of the real pathways glycolysis and citric acid cycle: ATP-consuming reactions always tend to be located before ATP-producing reactions, as it is in glycolysis, NADH production also is always located near the end of a pathway, which resembles the fact that the citric acid cycle occurs "after" glycolysis in a sense that it uses the product of glycolysis (pyruvate). However, when comparing sequences belonging to different clusters, differences in the ordering of the reactions inside a chain are found. We conclude that the positioning of certain types of reactions is very important for the biological functioning of an ATP and NADH-producing metabolic pathway, whereas the positioning of some of the other reaction types does not seem to be as significant.

Summarising, we present a model that is capable of reproducing some important features of ATP and NADH-producing metabolic systems using very little information as input parameters. The calculations give an insight into underlying principles concerning the stoichiometric design of metabolic pathways. The results obtained by this model give rise to optimism that a similar approach will be successful when applied to the analysis of structures of other metabolic systems.

References

1. Ebenhöh, O. and Heinrich, R. (1999) Evolutionary Optimisation of Metabolic Pathways. Submitted to *Bull. Math. Biol.*

2. Heinrich, R. and Schuster, S. (1996) *The Regulation of Cellular Systems.* Chapman & Hall, New York.

3. Stryer, L. (1998) *Biochemistry* W. H. Freeman and Company, New York.