rijksuniversiteit groningen

Bachelor Thesis Life Science & Technology

The road towards *de novo* pathway engineering: from design to reality

Renske van Raaphorst
Supervised by Marnix Medema
Rainer Breitling, Eriko Takano

TABLE OF CONTENTS

ABSTRACT

Innovation in biotechnology has led to the possibility of engineering metabolic pathways in microorganisms in order to make useful products efficiently. Alteration of pathways and combination of different pathways enable the production of medicine, natural fuels and more. However, finding a pathway for a desired compound using these strategies is difficult, as the starting point is always an existing pathway – made for another product. The ultimate engineering strategy would therefore be *de novo* pathway engineering: designing a metabolic pathway from scratch in order to produce any desired compound. In order to realize this engineering strategy, several key steps have to be walked through. After rational design of the desired compound, possible metabolic pathways can be generated by a computational prediction system, which searches for theoretically efficient pathways and tests these pathways on few key criteria. Next, a suitable host organism has to be chosen. When that has been decided, a metabolic reconstruction model of that host organism has to be used to analyze the pathway *in silico*. This should lead to a step-by-step optimized production pathway which can be expressed in the host organism. While it is not possible yet to take these steps of *de novo* engineering, promising developments are made and challenges are taken up by researchers. The biggest challenge lies within the understanding of the metabolic network as a system and, thereby, the influence of a small part of the network on the whole system. Experiments on pathway engineering, developments of computational pathway prediction systems and refinements of metabolic reconstructions can finally lead to engineering and design of new compounds produced by novel pathways.

INTRODUCTION

The enormous number of different enzymes known and those yet to be discovered, catalyzing all kinds of reactions, trigger the imagination. Now it is becoming possible to manipulate the metabolism of an organism by adding an enzyme to its metabolic network (Dietrich et al., 2009); (Hanai, Atsumi, & Liao, 2007), overexpressing certain enzymes or knocking out genes of native enzymes (Fong & Palsson, 2004), the logical next step would be adding a completely new metabolic pathway to a microbial host. In this way, not only novel derivatives of native metabolites could be made, but it would even enable one to produce virtually any organic compound, provided that the enzymes are necessary to construct the pathway are available. Development of these *de novo* pathways would open the road towards a new way of developing and producing all kinds of compounds, from sustainable energy sources to novel antibiotics.

In the search for new drugs, rational design is becoming more and more important. For example, as few new antimicrobial drugs are discovered from natural resources and multi-resistant strains are evolving rapidly, known antibiotics are often being chemically modified to make more effective antibiotics (Fischbach & Walsh, 2009). Yet these modified compounds are not as effective as novel compounds could be, because modified compounds still attack microbes in a similar way as the defeated antibiotic did. *De novo* pathway engineering could be the way to make production of novel antibiotics with new antimicrobial strategies possible. Also in research on known pharmaceuticals such as the anti-malaria agent artemisinin, which is extracted from plant cells in very low concentrations, searching for novel – and in particular efficient – pathways has led to crucial

breakthroughs, with as a highlight the high yield production of the artemisinin precursor dehydroartemisenic acid in *Escherichia coli* in 2009 (Dietrich et al., 2009).

Sustainable energy sources may not seem to be the first compounds to think of in the context of biosynthetic pathways, but in fact, this field of biodiesel, bioethanol and other types of biofuel is the field where probably the most research on new pathways has been done (Atsumi, Hanai, & Liao, 2008). In order for biofuel to be a viable alternative to traditional fuel, the production by microorganisms seeds to be more efficient and use an abundant source which preferably is not used for other purposes as food production. By metabolic engineering, efforts are done to let host organisms metabolize organic compounds such as starch and cellulose in order to make biofuels efficiently.

The approach of *de novo* biosynthetic pathway engineering is essentially different from the classical microbiological approaches to manipulating the metabolism of a micro-organism. While in classical biological approaches, one looks at the working of the cell and tries to find ways to manipulate that. In *de novo* pathway engineering, however, one starts with defining the desired end product, and designs a route towards the synthesis of that product (Prather & Martin, 2008). The whole pathway is a new combination of enzymes, and the product can be either a new and unnatural, or a known natural or chemical product. This engineering approach gives a different perspective to biochemical innovation.

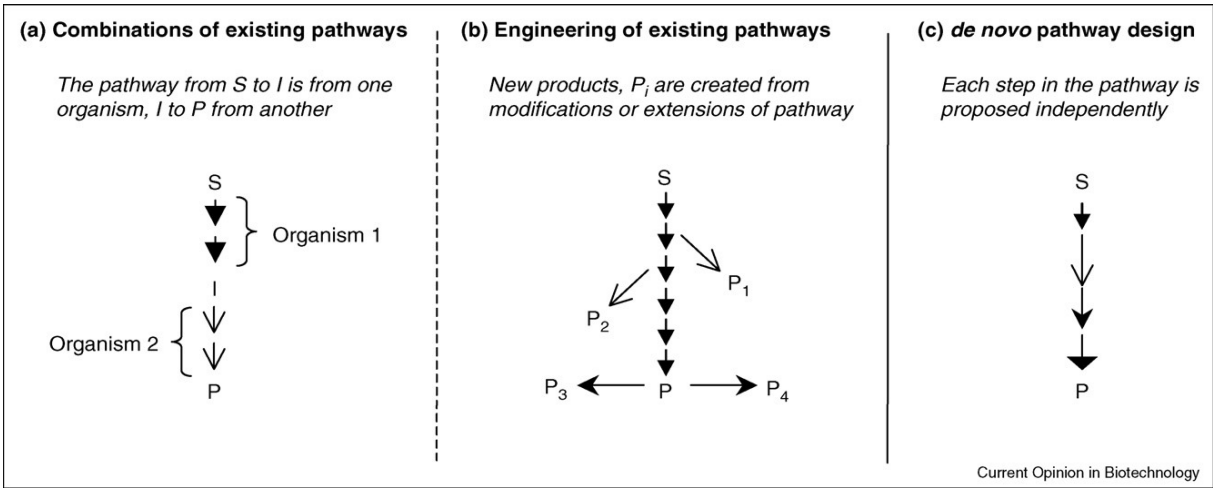

Figure 1. De novo *engineering compared to traditional pathway engineering and combinatorial pathways. Adapted from* Prather & Martin, 2008

Altogether, while much research is being done on altering existing pathways in order to obtain new compounds, or to make existing compounds more efficient, there are several reasons why creating pathways in a *de novo* fashion should be a more effective method to obtain these goals in the long term. The most important reason is the difference in perspective. The engineering approach allows, when possible, to consider every possibility in making a pathway for a certain target compound and finding the best pathway out of the possibilities. This should make the search for new pathways for new compounds more target-based and thereby efficient. However, there are several steps to be taken and tools needed to make *de novo* biosynthetic pathway engineering possible.

In this review I will give an overview of the past, current and future research on *de novo* biosynthetic pathway engineering. It will follow the various steps necessary for pathway construction, starting at

the definition of the target compound. On a global scale, I will focus on computational prediction of possible pathways within metabolic networks, and on a microscale on utilizing information on key enzyme properties: thermodynamics, kinetics and substrate specificity. Different considerations in choosing the right host organism for a pathway will be discussed, such as the prediction of effects of the new pathway to the host and ways to optimize the pathway. The last step is the insertion of the pathway into the host. The problems which come with this step will be discussed, as well as the possible solutions, in order to finally be able to give a comprehensive answer to this key question: how does one achieve *de novo* biosynthetic pathway engineering?

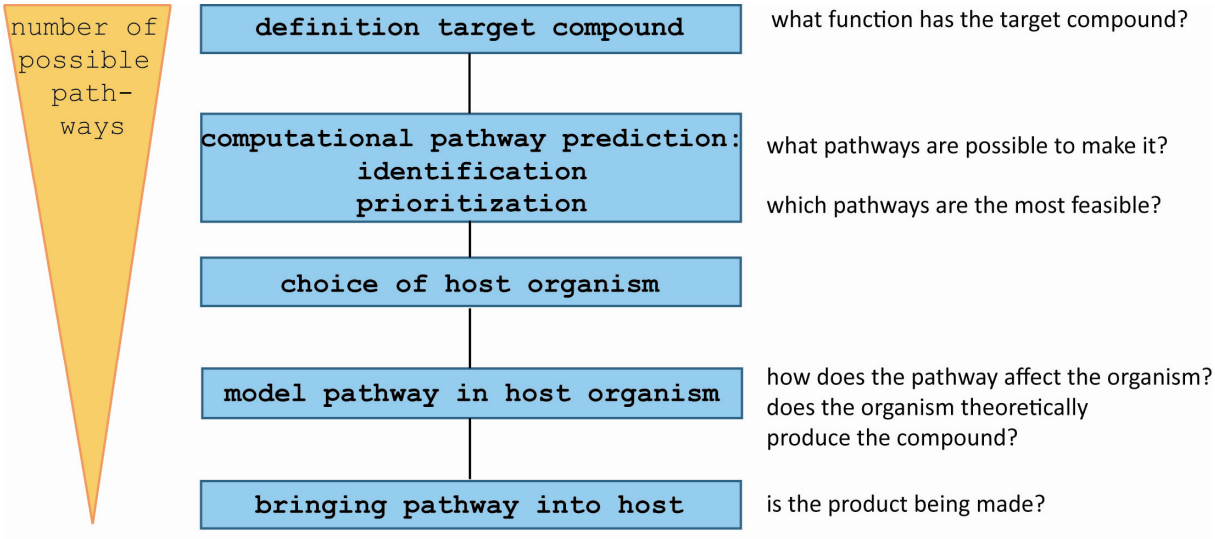

Figure 2. *Summary of steps to achieve* de novo *engineering of biosynthetic pathways. The different steps aim to fine tune the possible pathways until one optimal pathway is left.*

DEFINITION OF THE TARGET COMPOUND

The basis of engineering is that one starts with a clear definition of the product. While in for example the car industry it seems pretty logical that you have to know your design before starting to build, in biology this is less usual. However, for biosynthetic engineering the definition of the target compound is just as important as the design of every car, yet far less straight-forward.

In defining a suitable target compound there is often a conflict between the desired function of the compound and the ability to engineer the compound. The perfect compound will work optimal and is easy to produce, but often concessions have to be made. There are several properties of a compound one can think of cousing difficulties in production. The types of building blocks that a product consists of is one of those. One notoriously difficult group of natural products when it comes to heterologous expression is that of the polyketides; naturally produced by large enzyme complexes. The difficulties in producing this group, of which antibiotics are its famous members, will be discussed in a later part of this paper. Other challenging compound properties are for instance the stability of the compound and other chemical properties such as acidity and polarity are also to be taken into consideration since it has to be produced in a living organism. In the case of some toxic or not catalytically producable target compounds it can be favorable to produce a precursor rather than the target compound and carry out the last step outside the cells or even by using synthetic chemistry outside the culture (Dietrich et al., 2009).

These changes in target compound design to make the production process easier will mostly be made after more is known on the possible production pathways. The first design will therefore foremost be based upon the desired function. After that, computational prediction of possible pathways will tell if the design is probably feasible, or that one has to go back and reconsider the possibilities.

PREDICTION OF POSSIBLE PATHWAYS

Metabolic networks

All reactions catalyzed by enzymes in one single cell are connected in its metabolic network. In this network, many enzymes, co-factors and chaperones wok together and catalyze reactions to make many different compounds. This results in a highly complex, branched network, of which every part influences the system. Since the networks as a whole are so complex, it is more common to look at metabolic pathways: parts of the network which lead from a starting compound to a specific end product. In this sense, a metabolic pathway is nothing more than a module taken out of a metabolic network.

While every cell comprises its own metabolic network, a combination of all known metabolic networks gives insight into the potential of unusual combinations of enzymes. In the case of *de novo* engineering, having information on this "network of networks" is essential for predicting the possible new biosynthetic routes. This information is stored in many internet databases, of which the KEGG database is probably the most important. However, more elaborate information on for example enzyme localization and specificity has to be found in other databases. An overview of the databases and the type of information they contain is given in a review on metabolic modeling by Durot, Bourguignon, & Schachter, 2009, as shown in table 1 below.

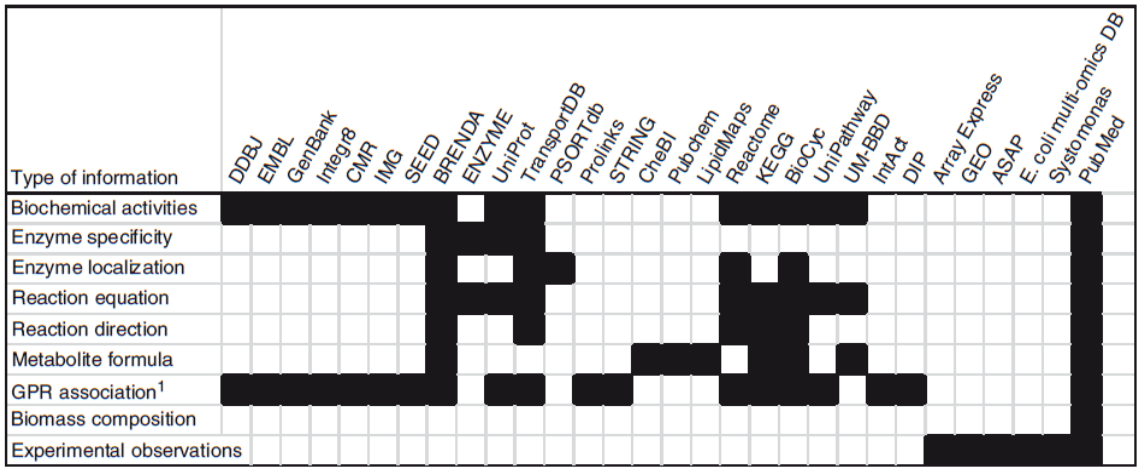

Table 1. *Overview of databases and types of information related to metabolic networks. Adapted from Durot et al., 2009*

Computational prediction systems

In the last decade, several pathway prediction systems have been made, as well as individual computational pathway predictions. There are two common problems with the most prediction systems, namely that they cannot be used for all types of target compounds or that they only comprise the metabolic network of one micro-organism. Two recent computational systems which

try to surpass these limitations to develop a broadly applicable method are BNICE, described by Hatzimanikatis et al., 2005, and the method described by Cho, Yun, Park, Lee, & Park, 2010. What is interesting about these methods is their essentially different approaches. While the work of Cho et al. primarily aims to advise which pathway to use to produce a certain compound, BNICE does something notably different: providing a way to discover potential novel compounds and find their production pathways.

BNICE is a computational method to predict novel pathways based on the reaction rules of the Enzyme Commission (EC) classification system. BNICE can predict novel pathways, given the starting compound and/or product, the requested length of the pathway and range of reactions searched in. With the last criterion is meant that one can choose to only search for a pathway using enzyme reactions out of one known pathway, for instance the tryptophan pathway, a combination of more pathways or the whole metabolic network. This makes it easy to search for shorter, more efficient pathways as well as investigate pathways for novel compounds.

BNICE can be a good first step in finding possible pathways, but a lot of analysis of the results is needed to get a useful outcome. In some searches, BNICE can predict more than 10.000 different pathways, given the few criteria the system relies on. In practical use, such a system is therefore not enough to start with. BNICE is not the only example of a pathway prediction system based on one type of information (in this case the EC classification system). In exception to this, Cho et al. tried to construct a system which predicts pathways in a more complete way. They developed a broad system which can predict pathways based on many criteria. Starting with a database of reactions, categorized by type of reaction and a database of reaction rules describing different reactions, the system first predicts many possible pathways. An algorithm was made which ranks every pathway based on binding site covalence, chemical similarity, thermodynamic favorability, pathway distance and organism specificity. In this way the system ends with one most favored pathway instead of thousands. Therefore, the work of Cho et al. is a step towards an integrated, widely usable method to not only predict possible pathways, but also predict which one will be the best to use.

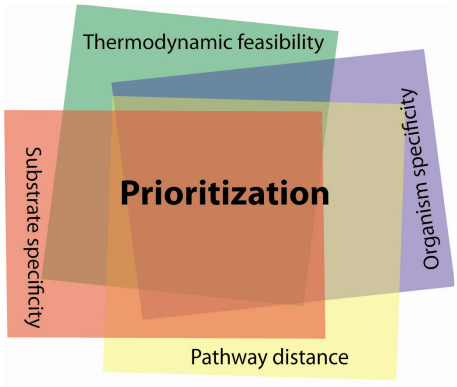

Figure 3. *The four different prioritization criteria globally used by different prediction systems: thermodynamic feasibility, organism specificity, pathway distance and substrate specificity. It can be discussed if all criteria should be used in this process.*

An important point of discussion in pathway prediction systems is the way to decide which pathway is theoretically the most favorable, sometimes over more than thousands of others. The first thing is that not everyone agrees with the criteria used. According to extreme pathway analysis by Papin et al., the distance of the pathway does not influence production rate (Papin, Price, & Palsson, 2002). That would mean pathway distance is not a suitable prioritization criterion. The biggest problem however is that few analyses can be done on the theoretical pathways without having a model of the host organism. One of the few factors which can be determined independently is the theoretical thermodynamic favorability. In most research, the thermodynamic favorability is measured by a group contribution method which measures the Gibbs free energy of formation of groups of atoms of the products and intermediates. These groups are added up to a total Gibbs energy of every reaction in the pathway. When the Gibbs energy of formation adds up below zero, the reaction is defined as thermodynamically favorable. Cofactors as such NADH and ATP have to be taken into account in this estimation(Bar-Even, Noor, Lewis, & Milo, 2010; Hatzimanikatis et al., 2005).

Cho et al. went beyond only using Gibbs free energy of formation to estimate thermodynamic feasibility (Cho et al., 2010). Also the fluctuation of Gibbs energy between the reactions in the pathway was taken into account: the less fluctuation, the more thermodynamically favorable the pathway, since every product of one reaction has to start as a reactant for the next.

Thermodynamic analysis narrows down the first search of pathways, but the way it is done can make a difference the outcome. The question is therefore: in which way can the thermodynamic feasibility best be estimated? While the method of group contribution is based on the most straightforward principle, it does not take into account the interaction of the whole pathway as a chain of reactions. More elaborate thermodynamic assays are therefore done in many occasions (Durot et al., 2009) in a later stage of pathway prediction, using a computational model of the chosen pathway in the metabolic network of the chosen host. One of these analysis methods, Thermodynamic Metabolic Flux Analysis (TMFA) will be discussed further in this paper.

Another factor which is theoretically independent of the host organisms metabolic network and topology is the substrate specificity. There is no standard method developed yet to predict this quickly and with many different reactions at the same time. Cho et al. try to do this by estimating the binding site covalence of the reactions they predict (Cho et al., 2010). Most other research on pathways where substrate specificity is investigated is not involved in the stage of pathway prediction, but in the stage where the chosen pathway has to be altered for better results. However, their information and methods can be used to get more insight in how substrate specificity influences the favorability of a pathway.

A practical use of analysis of the substrate specificity is shown in the research on artemisinin, mentioned earlier in the introduction (Dietrich et al., 2009). Here pathway engineering led to new pathways producing precursors of this drug. Dietrich et al. focused on making a pathway more efficient by altering one of the key enzymes of the novel pathway, cytochrome P450 from *Bacillus subtilis.* By using models and *in vitro* experiments, they found the parts of the enzyme which could be altered to make the non-natural substrate more favorable than the natural substrate of the enzyme.

In silico approaches to find enzymes with the right substrate specificity or to identify the right mutations to get those enzymes will be important in pathway prediction. Höhne et al. developed a method to find biocatalysts with a distinct enantiopreference (Höhne, Schätzle, Jochens, Robins, &

Bornscheuer, 2010). In the case of many regioselective drugs, chemical production mostly yields a mixture of (R) and (S) selective products, while only one of those is the effective drug. Höhne et al. deducted the way the desired enzyme probably is evolved. After searching for a possible candidate enzyme, which is mostly the known enzyme with the wrong enantiopreference, the desired mutations on this precursor enzyme were defined. The key difference in approach with other ones is their last step: the mutations are not carried out *in vitro* or *in vivo*. Instead, a database search is done in order to find enzymes already carrying the desired mutation. Successful searches take out the need for time-consuming techniques as metabolic engineering.

CHOICE OF HOST ORGANISM

The host organism has as much influence on the pathway as the latter has on its host organism. One cannot predict, make and investigate a new metabolic pathway without taking the host organism into account, because every new pathway has to take its place in a large native metabolic network. Competition with native metabolites, unpredicted side products and feedback loops are some of the possible effects of the host organism on the new pathway, while the new pathway might cause a defect in growth or even cell death.

There are different perspectives on what organism would be the best for a new pathway. Often, altered or copied pathways are brought into model organisms such as *E. coli* or *Saccharomyces cerevisiae*. The genomes, transcriptomes and metabolomes of these organisms are the most well studied, which has a great advantage: the predictability of the influence of the novel pathway on the host and vice versa. Because so much is known about the organisms, models can be made more easily which predict the behavior of the new pathway and its host. Another important advantage of using these organisms is that they have been used as industrial production hosts during the last decades and that accordingly the experience in manipulating them is great.

Different attempts to place existing pathways from one organism into model organisms show that for many classes of compounds, it is challenging to let their production pathways work in other organisms. One particularly interesting class is that of polyketides, in which are well-known for comprising various antibiotics. Many antibiotics are produced by *Streptomyces* bacteria, which are difficult to manipulate and culture on large scale. Most attempts to express the antibiotic producing enzymes, which are organized in large enzyme complexes, or are large multi-domain enzymes, into a manageable host as *Escherichia coli*, led to notorious problems (Boghigian & Pfeifer, 2008).

To begin with, natural product biosynthesis and especially that of polyketides, is often limited by an insufficient supply of the intracellular precursors needed to build the target compound. Therefore, precursor pathways have to be re-engineered or introduced as well. When that problem is overtaken and the pathways are expressed in the host, there is more. While during the last decade there are many efforts done to make model organisms produce polyketides by bringing in the heterologous pathways, all failed to gain production titers as high as in natural production or as theoretically possible. While many strategies, including the prediction of the influence of the new pathway by metabolic modeling, are heading in the right direction, production titers are still a problem.

One type of polyketide producing enzyme complexes has proved even harder to express heterologously. Expression of aromatic polyketide producing PKS type II's did only result in having the proteins as inclusion bodies and thereby incapable to produce the malonyl-CoA backbone. In

2008, Zhang et al. firstly managed to express a working minimal aromatic polyketide producing PKS in *E.coli* using an engineered fungal PKS as a template (Zhang, Zhao, Pan, Qiu, & Zhu, 2005). Still, there is a long road to climb before expression of polyketide pathways will be applicable for engineering purposes.

Another approach is to look for an organism which has the most enzymes in the designed pathway in its native metabolic network. In this way, as few as possible enzymes have to be introduced in the organism and thereby the metabolic network would be disturbed less. This idea is incorporated in the prediction system of Cho et al. which prefers pathways with many enzymes originating from the same organism (Cho et al., 2010). However, it has to be taken into account that this hypothesis is not been properly tested yet, since until now there are no *de novo* engineered pathways which are not based on any native pathway. Therefore one should be careful with stating that using a host with many usable native enzymes is the best option. It would mean that the new pathway will compete with native pathways for a fact, while trying to solve that by overexpressing native pathways probably also causes an extra flux in the native pathways. Partially knocking out the native pathway can be a solution to this problem if that pathway is not essential for the host.

Because the native pathway of a host organism can cause problems for the production of the desired compound, research is done on so-called reduced genome organisms. In these organisms, genes are deleted which are not necessary in an industrial environment. A reduced genome strain of *E. coli* for instance, lacks genes it needs in it natural environment, the mammalian intestines. To see if reducing of the genome is useful in metabolic engineering, Lee et al., 2009 expressed a L-threonine producing pathway into a reduced genome strain and a wild-type strain of *E. coli.* Comparison of the production efficiency showed that the reduced genome strain had a yield of L-threonine which was twice as high as the wild-type host. The researchers state that the reducing of metabolic burden and less competition with the engineered pathway might be responsible for this high production yields. Further research should however characterize the mechanism behind this significant effect.

What goes even one step further, is the development of the minimal and synthetic genomes. In spring this year, the J. Craig Venter Institute reported the creation of a bacterial cell controlled by a completely synthetic genome (Gibson et al., 2010). These exciting developments naturally trigger the imagination: it might just be the first step towards a synthetic, controlled template to which complete production pathways can be added. However, at the moment this is not more than a product of imagination. Therefore it the choice for the host organism depends on the type of pathway one wants to introduce. Shortly said, in the case of a *de novo* pathway consisting of large parts of existing pathways, it might be smart to introduce it into the organism that pathway is originating from when possible. A *de novo* pathway which consists of enzymes of different origins asks for a well-researched host as *E. coli* or *S. cerevisiae*, depending on the type of target compound. Reduced genome strains of these host organisms can probably be an improvement, reducing the competitive pathways.

PATHWAY ANALYSIS THROUGH METABOLIC MODELING

One of the most important, if not the most important developments enabling *de novo* pathway engineering in the future is the development of metabolic models. As the paragraphs above already make clear, the influence of the host organism on the production pathway and vice versa is substantial. Interference of the engineered pathway with the metabolic network of the host can result in several malfunctions as for instance growth deficiency. On the other hand, the target compound might be formed only at small rates due to competition with pathways of the host. In order to be prepared for problems like this and maybe even solve them before they actually occur *in vivo*, *in silico* testing based on several key properties as topology and metabolic flux is an essential step in *de novo* engineering.

As essential this step might be, prediction of the effects of the pathway in the host is not done often in metabolic engineering. Recently a few articles on *in silico* testing of novel pathways have been released. One particularly interesting article is that of Finley et al., which uses the previously discussed pathway prediction system BNICE to find novel pathways for biodegradation of the pollutant 1,2,4–trichlorobenzene by *Pseudomonas putida*, a known pollutant degrader (Finley, Broadbelt, & Hatzimanikatis, 2010). To analyze which pathway was most effective and determine the growth rate of the host with implemented pathway, they applied Thermodynamic Metabolic Flux Analysis (TMFA)(Edwards, Ramakrishna, & Schilling, 1999; Henry, Broadbelt, & Hatzimanikatis, 2007) to a computational model of the metabolic network of the host. The outcome of TMFA is based upon the influence a new pathway has on the metabolic flux of the network – to illustrate it, pathways and networks are often depicted with flux indicating arrows. To this method, thermodynamic constraints are added to indicate whether a pathway is likely to be blocked. This method is applicable to many organisms, because its simplicity compared to the use of complete metabolic models. However, that simplicity also implies directly that not every parameter can be taken into account. With that in mind, Metabolic Flux Analysis is a useful method to investigate whether a pathway is productive and how it affects the growth rate of an organism.

Finley et Al. used TMFA not only to predict the behavior of the pathway and its host, but also to choose one of the pathways BNICE created. Another system which aims to choose a pathway as well as monitor the effects of the new pathway on the host is OptStrain (Pharkya, Burgard, & Maranas, 2004) – a method which uses flux analysis in order to give advice on how the production could be optimized by altering the hosts gene expression. After constructing a hypothetical pathway, this system changes the pathway in such a way, that most enzymes from the pathway are native to the host organism. With the use of a purely stoichiometric model of the host's metabolic network, OptStrain predicts the effect of novel enzymes in the pathway and which genes could be up- or downregulated in order to increase the production yield.

While both prediction methods are of great interest for metabolic engineering, for the purpose of *de novo* engineering there are still steps to be made. The essence lies in the type of models these methods are applied to. While genome based models of metabolic networks are covering the essential properties needed for *in silico* testing of designed pathways, effects of key cellular processes are not taken into account. It is for instance useful to know which genes are transcribed in which condition and thereby which metabolic pathways are available at that point and at what rate certain enzymes occur. Most times, metabolic models are based on the exponential growth phase,

while many secondary metabolites are only produced in the steady state. For drugs as antibiotics, also novel ones, a different type of model is thereby needed. Also the amount of energy the cell uses for non-metabolic events, as compensation of membrane-leakage and protein polymerization costs, is a significant parameter (Feist et al., 2007). Therefore, in order to test different designs in a relatively quick and inexpensive but reliable fashion, as complete as possible models are needed.

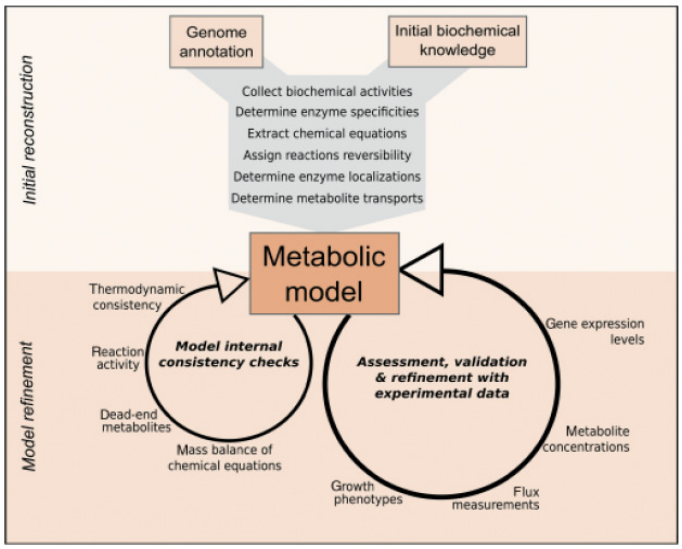

Figure 4. *Schematic view of reconstruction and refinement of a metabolic model. An initial model is reconstructed from genome annotations and from preexisting knowledge on the species' biochemistry and physiology. The resulting model is then corrected and refined, according to internal consistency criteria and by comparing its predictions to experimental data. Adapted from Durot et al. (2009)*

Seventeen metabolic reconstructions of different bacteria were developed until 2009 (Durot et al., 2009). These extensive reproductions are as close to reality as nowadays possible. One example is the *E. coli* metabolic reconstruction, which has been worked on since 1990 and is lastly improved in 2007 (Feist et al., 2007). The metabolic model developed with this reconstruction was adjusted to include effects of transcriptional regulatory effects and maintenance costs. Reaction kinetic effects were not included, since the researchers stated that there is not enough information yet available on *in vivo* kinetic parameters and concentrations.

When metabolic reconstructions are fully optimized, different types of analysis can be done. One potentially powerful tool to analyze metabolic networks is extreme pathway analysis. Extreme pathways are defined in an essentially different way from normal “traditional” pathways: they are mathematically derived vectors which can be used to characterize the phenotypic potential of a metabolic network (Papin et al., 2002). These vectors can be translated as pathways describing the production of a certain compounds catalyzed by enzymes, together with all co-factors, byproducts and precursors needed to make these products. Here lies the essential difference: while traditional pathways are depicted as a chain of enzyme reactions, extreme pathways also comprise everything else needed and produced in order to make the end product. An extreme pathway can therefore roughly be described as the minimal metabolic network needed to produce certain products. For *de novo* engineering, this means that novel pathways could be introduced and the extreme pathways could be identified. The influence of by-products, cofactors and many more can thereby be validated, and fluxes within these extreme pathways comprising the novel pathway. The major advantage of this analysis is that influences can be traced back in more detail. Problems with extreme pathway analysis are especially related to the large amounts of extreme pathways which are abundant in whole genome metabolic networks. Research is done to estimate the vastness of extreme pathways

and to find ways to analyze large numbers efficiently (Schilling, Schuster, Palsson, & Heinrich, 1999; Papin et al., 2002).

DISCUSSION

Promising perspectives

Several promising developments have passed in this paper. In general, the ability to computationally predict and model metabolic networks, made possible by the vast amount of data gained in the last decades, offers great perspectives for *de novo* engineering. These developments, which are quickly improving, will provide the key tools in understanding and using complex systems as metabolic networks.

The prediction of useful pathways is not yet fully applicable, but many different systems have been developed and are still worked on. At the moment, the reason for the choice of criteria on which these systems decide which pathway is the most feasible, is not always fully argued. More insight on how the efficiency of metabolic pathways is influenced by how they are made up, which could be gained by pathway analysis tools as Flux Balance Analysis and Extreme Pathway Analysis, could provide a better argument for criteria as pathway length and complexity. The described strategies to find enzymes with the desired substrate specificity are a promising step towards getting a pathway which is theoretically the best one possible.

A back-and-forward process

While the different steps in the process to achieve *de novo* engineering are presented as steps that follow each other up, in practice this process will be one of feedback and adjustments. This is already clear when starting with the first step, the design of the starting compound. Since it is not always predictable which compound will be easy to produce, the most important aspect in design is the functionality. Even while there are other aspects of the compound, such as toxicity, which could influence the production process, without the production pathway, it is difficult to predict the production efficiency. Recently, Papin et al. investigated the relation between compound complexity and pathway efficiency, but did not get a convincing correlation (Papin et al., 2002). With such uncertainties, feedback and adjustment is very important. In the case of compound design it would mean that when the pathway prediction process indicates that the product design is not optimal, it can be adjusted.

As another example more candidate pathways can be chosen to integrate into the model of the chosen organism instead of choosing one theoretically best pathway on relatively few criteria. In this way, a previously selected group of pathways can be compared to each other by looking at the product yield, vitality of the organism and the production time. Hereby a well argued choice can be made. Also in this stage feedback can be important, since even in the best pathways found, adjustments could be needed. When it is possible to find the exact point where a pathway does have competition problems, a lack of precursor compounds or other problems, the pathway can actively be adjusted until the optimal pathway is found.

Obstacles on the road

While the perspectives are promising, there are still major obstacles to overcome in the road towards *de novo* engineering. Even with the improvements of metabolic reconstructions and analyses which gain more insight into these complex systems, there is still a big gap of knowledge. One of the problems is the huge amount of data gained from metabolic network analysis. An example already noted in this paper is the amount of extreme pathways created when a full genome metabolic model is analyzed. At the moment this is a factor which blocks the use of extreme pathway analysis for full genome purposes. However, research is done to improve the ability to process the large amount of information extreme pathway analysis generates (Papin et al., 2002). Analysis on smaller systems, as the model system of the red blood cell, which comprises of a small amount of extreme pathways, showed promising results (Wiback & Palsson, 2002).

A problem related to the latter which cannot be underestimated is the unpredictability of how a host organism reacts on a novel pathway. Even while metabolic models might take away some unexpected behaviors, there are still many things we do not understand. What is clear, is that a host organism cannot be seen as a "template" on which one can place the functions one desires. A large list of attempts to place heterologous pathways into different host organisms prove that this point might be the bottleneck in *de novo* engineering. In many cases, improvements are made, but a full understanding of the host as a system is far away. For instance, heterologous expression of polyketide producing pathways in *E. coli* is effected positively by the presence of signaling molecules (Boghigian & Pfeifer, 2008). This useful effect is however not explained yet, and therefore difficult to use in rational design and modeling.

Concluding from this, one could say that the biggest problems in *de novo* pathway engineering are universal for the fields of bioengineering and synthetic biology, namely, the fact that biologists do not always know what the building blocks nor the templates look like. It is like building a house in the dark – the question therefore is, how long it will take for the sun to rise.

A matter of time?

For *de novo* engineering, a different way of thinking is needed than for traditional metabolic research. This change of mindset started about fifteen years ago when the term "synthetic biology" was introduced. While the synthetic biology research groups are springing from the ground around the world and the number of new techniques and information is increasing exponentially, it seems only a matter of time before *de novo* engineering is a widely used approach to create all different types of compounds.

How much time it will take to make *de novo* engineering possible is less obvious; it is unpredictable when natural systems as metabolic networks and cell metabolism will be understood as a system. This bottleneck of system understanding is a challenging research topic which has been taken on by many researchers around the world. The current techniques which produce massive amounts of data, used in the fields of genomics, transcriptomics and metabolomics will definitely give more background for reliable metabolic reconstruction of more organisms and industrial strains. The use of pathway prediction systems and metabolic models before expression *in vivo*, is already emerging and will be increasingly used as the systems get more reliable. Figure 5 shows some important developments of the last decades in the perspective of time.

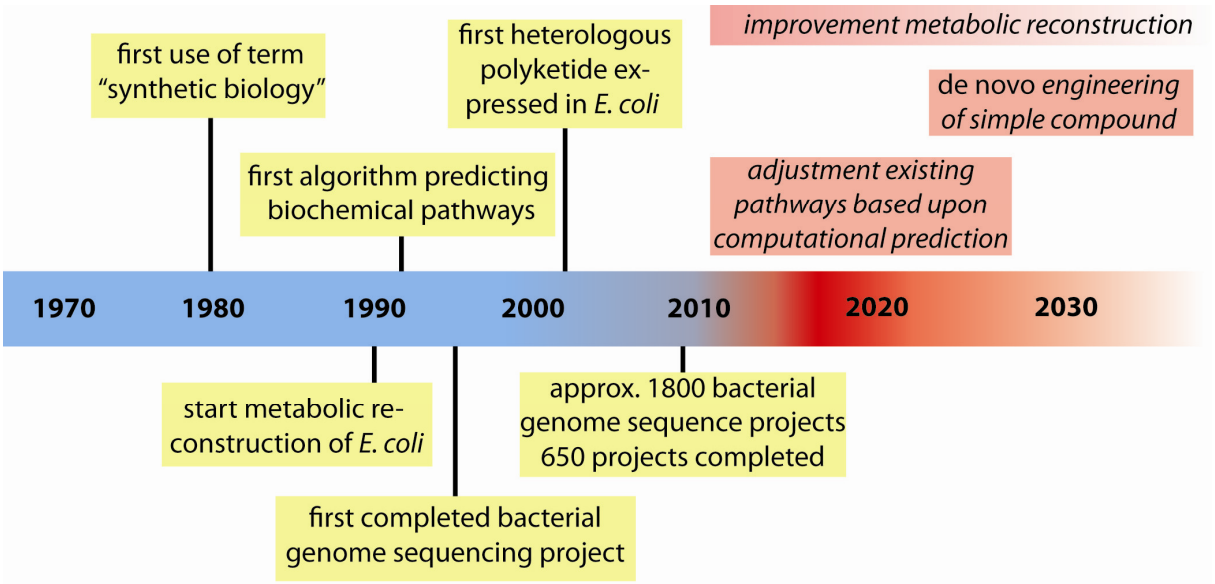

Figure 5. *Time scale from the last forty years to the future highlighting a few important points of developments concerning* de novo *pathway engineering. 1980: Barbara Hobom first uses the term "synthetic biology" to describe genetically engineered bacteria* (Benner & Sismour, 2005). *1990: The first report of the metabolic reconstruction of* E.coli. *Since then, it has been corrected and refined* (Feist et al., 2007). *1992: The first algorithm made which generated biochemical pathways from a database* (Mavrovouniotis, Stephanopoulos, & Stephanopoulos, 1992). *1995: The first genome sequenced of a bacterium,* Haemophilus influenza (Fleischmann et al., 1995). *2001: First succesful heterologous expression which of a PKS complex in* E. coli , *producing 6-deoxyerythronolide B (6dEB)* (Pfeifer, Admiraal, Gramajo, Cane, & Khosla, 2001). *2008: Checking the balance. Less than ten years after the first bacterial genome sequence, 1800 projects are running, 650 projects finished* (Durot et al., 2009). *In red, probable future events are stated. The use of computational prediction systems and metabolic model in traditional pathway engineering will pave the paths for the first* de novo *engineered pathways.*

After other forms of pathway engineering, like combinatorial and heterologous pathway engineering are optimized more and more, *de novo* engineering of pathways will come within reach. These more traditional engineering experiments will be necessary to overcome and gain insight to problems concerning expression of foreign pathways, which is still a very challenging topic. As seen in figure 5, it is has been less than ten years since the first heterologous expression of a polyketide pathway in *E. coli:* especially for these challenging natural product groups, much of the metabolic mechanism is not yet known. Moreover, the pathway engineering techniques used nowadays will provide experimental knowledge needed to make reliable metabolic models and test theoretically predicted pathways *in vivo.* With this knowledge it will be possible to truly predict which pathway and which organism, which prediction criteria and which analyses are necessary to make novel compounds with the use of novel pathways.

REFERENCES

Atsumi, S., Hanai, T., & Liao, J. C. (2008). Non-fermentative pathways for synthesis of branched-chain higher alcohols as biofuels. *Nature, 451*(7174), 86-89.

Bar-Even, A., Noor, E., Lewis, N. E., & Milo, R. (2010). Design and analysis of synthetic carbon fixation pathways. *Proceedings of the National Academy of Sciences, 107*(19), 8889-8894. doi:10.1073/pnas.0907176107

Benner, S. A., & Sismour, A. M. (2005). Synthetic biology. *Nature Reviews. Genetics, 6*(7), 533-543.

Boghigian, B., & Pfeifer, B. (2008). Current status, strategies, and potential for the metabolic engineering of heterologous polyketides in escherichia coli. *Biotechnology Letters, 30*(8), 1323-1330. doi:10.1007/s10529-008-9689-2

Cho, A., Yun, H., Park, J., Lee, S., & Park, S. (2010). Prediction of novel synthetic pathways for the production of desired chemicals. *BMC Systems Biology, 4*(1), 35.

Dietrich, J. A., Yoshikuni, Y., Fisher, K. J., Woolard, F. X., Ockey, D., McPhee, D. J., Renninger, N. S., Chang, M. C. Y., Baker, D., & Keasling, J. D. (2009). A novel semi-biosynthetic route for artemisinin production using engineered substrate-promiscuous P450BM3. *ACS Chemical Biology, 4*(4), 261-267.

Durot, M., Bourguignon, P. -., & Schachter, V. (2009). Genome-scale models of bacterial metabolism: Reconstruction and applications. *FEMS Microbiology Reviews, 33*(1), 164-190. doi:10.1111/j.1574-6976.2008.00146.x

Edwards, J. S., Ramakrishna, R., & Schilling, C. H. (1999). Metabolic flux balance analysis. *Metab Eng,* , 13-57.

Feist, A. M., Henry, C. S., Reed, J. L., Krummenacker, M., Joyce, A. R., Karp, P. D., Broadbelt, L. J., Hatzimanikatis, V., & Palsson, B. O. (2007). A genome-scale metabolic reconstruction for escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol Syst Biol, 3*, 121.

Finley, S., Broadbelt, L., & Hatzimanikatis, V. (2010). In silico feasibility of novel biodegradation pathways for 1,2,4-trichlorobenzene. *BMC Systems Biology, 4*(1), 7.

Fischbach, M. A., & Walsh, C. T. (2009). Antibiotics for emerging pathogens. *Science, 325*(5944), 1089-1093. doi:10.1126/science.1176667

Fleischmann, R., Adams, M., White, O., Clayton, R., Kirkness, E., Kerlavage, A., Bult, C., Tomb, J., Dougherty, B., Merrick, J., & al., e. (1995). Whole-genome random sequencing and assembly of haemophilus influenzae rd. *Science, 269*(5223), 496-512. doi:10.1126/science.7542800

Fong, S. S., & Palsson, B. O. (2004). Metabolic gene-deletion strains of escherichia coli evolve to computationally predicted growth phenotypes. *Nat Genet, 36*, 1056-1058.

Gibson, D. G., Glass, J. I., Lartigue, C., Noskov, V. N., Chuang, R., Algire, M. A., Benders, G. A., Montague, M. G., Ma, L., Moodie, M. M., Merryman, C., Vashee, S., Krishnakumar, R., Assad-Garcia, N., Andrews-Pfannkoch, C., Denisova, E. A., Young, L., Qi, Z., Segall-Shapiro, T., Calvey, C. H., Parmar, P. P., Hutchison, C. A., Smith, H. O., & Venter, J. C. (2010). Creation of a bacterial cell controlled by a chemically synthesized genome. *Science, 329*(5987), 52-56.

Höhne, M., Schätzle, S., Jochens, H., Robins, K., & Bornscheuer, U. T. (2010). Rational assignment of key motifs for function guides in silico enzyme identification. *Nat Chem Biol, 6*(11), 807-813.

Hanai, T., Atsumi, S., & Liao, J. C. (2007). Engineered synthetic pathway for isopropanol production in escherichia coli. *Applied and Environmental Microbiology, 73*(24), 7814-7818. doi:10.1128/AEM.01140-07

Hatzimanikatis, V., Li, C., Ionita, J. A., Henry, C. S., Jankowski, M. D., & Broadbelt, L. J. (2005). Exploring the diversity of complex metabolic networks. *Bioinformatics, 21*, 1603-1609.

Henry, C. S., Broadbelt, L. J., & Hatzimanikatis, V. (2007). Thermodynamics-based metabolic flux analysis. *Biophys J, 92*, 1792-1805.

Lee, J., Sung, B., Kim, M. Blattner, F., Yoon, B. Kim, J. Kim, S. (2009). Metabolic engineering of a reduced-genome strain of Escherichia coli for L-threonine production. *Microbial Cell Factories, 8(1),* 2

Mavrovouniotis, M., Stephanopoulos, G., & Stephanopoulos, G. (1992). Synthesis of biochemical production routes. *Computers & Chemical Engineering, 16*(6), 605-619. doi:DOI: 10.1016/0098-1354(92)80071-G

Papin, J. A., Price, N. D., & Palsson, B. Ø. (2002). Extreme pathway lengths and reaction participation in genome-scale metabolic networks. *Genome Research, 12*(12), 1889-1900. doi:10.1101/gr.327702

Pfeifer, B. A., Admiraal, S. J., Gramajo, H., Cane, D. E., & Khosla, C. (2001). Biosynthesis of complex polyketides in a metabolically engineered strain of E. coli. *Science, 291*(5509), 1790-1792. doi:10.1126/science.1058092

Pharkya, P., Burgard, A. P., & Maranas, C. D. (2004). OptStrain: A computational framework for redesign of microbial production systems. *Genome Research, 14*(11), 2367-2376. doi:10.1101/gr.2872004

Prather, K. L. J., & Martin, C. H. (2008). De novo biosynthetic pathways: Rational design of microbial chemical factories. *Current Opinion in Biotechnology, 19*(5), 468-474. doi: 10.1016/j.copbio.2008.07.009

Schilling, C. H., Schuster, S., Palsson, B. O., & Heinrich, R. (1999). Metabolic pathway analysis: Basic concepts and scientific applications in the post-genomic era. *American Chemical Society.* doi: 10.1021/bp990048k

Wiback, S. J., & Palsson, B. O. (2002). Extreme pathway analysis of human red blood cell metabolism [Abstract]. *Biophysical Journal, 83*(2) 808-818.

Zhang, W., Li, Y., Tang, Y. (2008). Engineered biosynthesis of bacterial aromatic polyketides in *Escherichia coli. PNAS, 105(*52) 20683–20688