



HAL
open science

(Re-)Construction, characterization and modeling of genomes for synthetic biology

Maria Suarez Diez, Carolyn Lam, Audrey Leprince, Vitor Martins dos Santos

► **To cite this version:**

Maria Suarez Diez, Carolyn Lam, Audrey Leprince, Vitor Martins dos Santos. (Re-)Construction, characterization and modeling of genomes for synthetic biology. *Biotechnology Journal*, 2009, 4 (10), pp.1382. 10.1002/biot.200900173 . hal-00526194

HAL Id: hal-00526194

<https://hal.science/hal-00526194>

Submitted on 14 Oct 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



(Re-)Construction, characterization and modeling of genomes for synthetic biology

Journal:	<i>Biotechnology Journal</i>
Manuscript ID:	biot.200900173.R1
Wiley - Manuscript type:	Review
Date Submitted by the Author:	06-Sep-2009
Complete List of Authors:	Suarez Diez, Maria; Helmholtz Center for Infection Research, Systems and Synthetic Biology Group Lam, Carolyn; Helmholtz Center for Infection Research, Systems and Synthetic Biology Group Leprince, Audrey; Helmholtz Center for Infection Research, Systems and Synthetic Biology Group Martins dos Santos, Vitor; Helmholtz Center for Infection Research, Systems and Synthetic Biology Group
Keywords:	synthetic biology, genome-wide modeling, protocell, in-vitro systems, genome minimization



1
2 Biotech Highlight ((6088 words))
3

4 (Re-)Construction, characterization and modeling of systems for synthetic biology
5

6 &block Title change ok? &block

7 María Suárez Diez¹, Carolyn M.C. Lam¹, Audrey Leprince¹, Vítor A.P. Martins dos Santos¹
8

9
10 ¹ Systems and Synthetic Biology Group, Helmholtz Centre for Infection Research, Inhoffenstraße 7, D-
11 38124 Braunschweig, Germany.
12

13
14 **Keywords:** synthetic biology; protocell; genome minimization; genome-wide modeling, in-vitro
15 systems..
16

17
18
19 **Correspondence:**

20 Vítor A.P. Martins dos Santos
21

22
23 Systems and Synthetic Biology Group, Helmholtz Centre for Infection Research, Inhoffenstraße 7, D-
24 38124 Braunschweig, Germany.
25

26 Tel: ++49-531-6181-4008

27 Fax: ++49-531-6181-4199

28 e-mail: vds@helmholtz-hzi.de
29
30

31 Abbreviations: SB: synthetic biology.
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1 Introduction

Synthetic biology (SB) is a developing area of research that aims to design and build novel biological components and systems with functions that cannot generally be found in nature or biological systems with known properties using implementations or architectures different from historical evolution. SB proposes a new paradigm for the analysis of biological systems and the development of practical applications. It is an inherently cross-disciplinary approach to biological sciences, that integrates scientists, knowledge and techniques from fields like mathematics, physics, chemistry, engineering and computer science, within the large knowledge corpus developed in the different fields of biological sciences: molecular biology, systems biology, genetic engineering, comparative genomics etc. This emerging discipline is not yet uniform and comprises a plethora of approaches, projects, and goals. As a result, no single definition has been agreed on by the members of the community. Broadly, five categories or branches can be defined within SB: bioengineering, synthetic genomics, protocell synthetic biology, unnatural molecular biology and *in silico* synthetic biology [1].

The bioengineering branch, represented by the 'modularization, abstraction and standardization' motto, aims to design and construct novel biological systems by assembling standard, simplified and well characterized biological elements (parts) to form complex and functional modules (devices) that can be combined into new metabolic or regulatory pathways within the host (systems). This branch encourages an engineering approach to produce controllable and predictable constructs and aims at transforming biotechnology into a true engineering discipline [2]. Synthetic genomics aims at creating organisms with a (synthetic) minimal genome. Modern DNA chemical synthesis methods allow the synthesis of entire genomes that can serve to replace the original genome of the host and generate an artificial organism optimized for new targeted functionalities. The protocell branch aims at *in vitro* construction of fully artificial protocells endowed with the needed functions to be considered as living cells whereas the unnatural molecular biology branch pursues the synthesis of new life forms, based on modification of the rules of the molecular biology that evolution has shaped: new nucleic acids, new types of amino acids or existing ones with altered chiralities, different genetic code, orthogonal ribosomes. SB should be model-driven especially since it has to deal with an increasing degree of complexity and relies on tools such as computational models of complex systems and bioinformatics for exploring interactions among the basic building blocks and to predict new properties emerging from their interaction among themselves and with the environment. Thus, the goals of the *in silico* branch are to produce detailed models of the elements of the synthetic circuits and tools for their design, to establish computational methods to assemble the models and to simulate the final assembled models within their cellular context.

Standard biological chassis are intended to provide an optimal structure for the designed systems. They should be robust biological systems (cell, protocell or *in vitro* system) modified to allow inclusion of the new genetic elements and able to integrate signals from each subsystem in order to generate complex functions. We would like to emphasize that, in SB, construction and design are tightly bonded, so we will distinguish between 'scaffold' and 'chassis'. On one side we have the physical implementation of the host, the chassis, whereas by scaffold we would like to denote the chassis together with a quantitative and predictive model able to reproduce its biological behavior. Figure 1 depicts the distinction between these two concepts in SB.

The development and characterization of new biological scaffolds which are able to host designed circuits, optimized for new targeted functionalities, and endowed with a full quantitative and predictive model remains an open challenge for this new field. In the following section we will discuss about currently available chassis. Section three is devoted to the advances towards the construction of simplified hosts, whereas in section four we analyze chassis modeling and characterization. In section

1
2 five we provide an overview of computational tools and platforms for scaffold development, modeling
3 and design. Finally, we discuss some of the major challenges ahead. In each section we have included
4 references to recent reviews where interested readers can find more detailed information.
5
6

7 **2 Existing 'Chassis'**

9 The gram negative bacterium *Escherichia coli* has many desirable properties for a chassis: it is robust
10 under a variety of growth conditions, there exist well developed molecular tools for its genetic
11 manipulation as well as measuring techniques for its characterization. In addition, its physiology is very
12 well studied and there is a vast amount of knowledge on its genome, biochemical machinery and
13 regulatory mechanisms as well as a huge compendium of available high-throughput data that allows the
14 construction of genome scale metabolic and regulatory models. These characteristics have made this
15 strain the chassis of choice in the vast majority of the students' projects presented at the annual
16 international Genetically Engineered Machines competition, iGEM (<http://www.igem.org/>). As a
17 creative student training activity, the iGEM competition has helped to develop the community of SB
18 and is contributing to develop and enforce standards. SB projects using *E. coli* as a chassis include the
19 construction of detectors for natural and unnatural molecules, production of biofuels, construction of
20 transcription circuits featuring targeted output, exploring the mechanisms of cell-cell communication
21 and the construction of light-reacting bacteria among others. Nevertheless, *E. coli* presents some yet
22 unsolved drawbacks: production of human-affecting toxins, intra-cellular medium secretion, limited
23 metabolic and post-transcriptional modification capabilities...
24
25
26
27

28 Synthetic biologists have recently focused their efforts on the study of non-transcriptional regulatory
29 elements, like non-coding RNA. The model eukaryotic organism *Saccharomyces cerevisiae* has been
30 used as a chassis to study the effect of antiswitches, a type of trans-acting riboregulators. An antiswitch
31 contains an aptamer domain able to selectively bind specific ligands. Upon binding, the antiswitch
32 undergoes a conformational modification that enables the antisense domain to bind its target mRNA
33 and to modulate translation [3]. This host has also allowed the implementation of transcriptional
34 circuits with targeted behavior. Moreover, one of the most successful applications of SB, the microbial
35 production of the anti-malarial drug precursor arteminic acid, used this organism as a chassis [4].
36 Biofuel production by *S. cerevisiae* using starch, sugar, and cellulose as substrates is another active
37 field of research which presents serious challenges like the need to integrate exogenous genes into the
38 host cell or to improve its tolerance to the final product. Microbial production is one of the most
39 actively researched applications of SB[5].
40
41
42
43

44 Other energy related projects aim at the production of biofuels using the photosynthetic machinery of
45 the host cell (micro-algae or cyanobacteria) but the photosynthetic efficiency and lack of molecular
46 tools are yet to be improved. The development of more efficient organisms for biodegradation is guided
47 by the analysis of the routes that the pollutants follow within biodegradative microorganisms. SB in
48 mammalian cells will allow the development of new therapies: circuits with a time-oscillatory behavior
49 could regulate *in vivo* the administration of a given gene product. The use of viruses as chassis
50 promises to open new avenues for the development of biology based nano-technology.
51
52

53 The Registry of Standard Biological Parts (<http://partsregistry.org/>) contains nowadays (September
54 2009) more than 3200 parts contributed by the iGEM teams and suggests two cellular chassis for SB:
55 *E. coli* and the gram-positive bacterium *Bacillus subtilis*. *B. subtilis* is an endo-toxin free organism with
56 a high capacity of extracellular protein secretion due to its single membrane. It has been the focus of
57 extensive genetic, metabolic and pathway engineering in order to maximize the yield in bioproduction
58
59
60

1
2 process. Yet, it presents additional challenges like the insufficient number of available molecular tools
3 due to plasmid instability or problems arising when expressing foreign proteins (mis-assembled
4 proteins, aggregates formation...). Moreover, it expresses intracellular and extracellular peptidases that
5 may cleave and degrade the secreted proteins.
6

7 Each of the cited organisms has their own advantages and disadvantages, and there are many other
8 desirable properties for a chassis that are present in other organisms: higher ability for heterologous
9 protein expression, higher solvent or stress resistance, etc... that we would like to have in our chassis
10 for SB. In SB we can not define a globally optimal organism suitable for all the different applications.
11 For example, it is very likely that the optimal scaffold for photosynthetic biofuel production will not be
12 the most adequate for microbial production of complex biochemical compounds or for its use in bio-
13 remediation as a living factory able to degrade pollutants. The proposed solution is to generate a
14 diverse portfolio of scaffolds with different characteristics and each of them optimized for a specific
15 application.
16
17

18 19 **3 Chassis development**

20
21 The usefulness of a given chassis will be limited by several aspects. First of all, the new chassis may
22 require developing new standards, since existing standards, like Biobrick, are not guaranteed to work in
23 other hosts. Parts optimized for a chassis will most certainly need to be modified for a different chassis,
24 addressing issues like codon usage or restriction sites, as well as the different regulatory interplays and
25 network structures of the chassis. The potential users should also have access to a 'user manual'
26 containing information about optimal biomolecular tools (vectors, protocols etc.) and a guide on how to
27 manipulate the system such as optimal growing conditions, required media etc. With the increasing
28 orthogonalization of biological components and developments in the various disciplines regarding to
29 genome streamlining, genome synthesis and “from scratch” assembly of components, the vision in the
30 long run, would be that chassis *à la carte* can be designed for specific application and plug-in circuits.
31 We discuss some of these trends below.
32
33
34
35
36
37

38 **Orthogonality**

39 The addition of any module changes the nature of the scaffold and each expressed heterologous protein
40 can potentially interact with any other macromolecules already present in the medium, potentially
41 leading to an unpredictable behavior of the parts. Orthogonal regulators, working without interference
42 from other cellular process partially circumvent this problem. Computational protein design and
43 rational design will allow further development of orthogonal regulators. DNA interacting proteins have
44 been redesigned to obtain pairs of designed protein-mutated DNA able to interact among themselves
45 but not with their wild-type counterparts. Similarly, receptors with an increased specificity towards
46 synthetic molecules have also developed [6]. New synthetic expression systems have also been
47 developed by synthesizing mRNAs with ribosome binding sites unrecognizable for the native
48 ribosomes of the cell, as well as ribosomes that are able to translate only the synthetic mRNA.
49
50
51
52
53

54 **Minimal chassis**

55 Living organisms contain in their genomes, genes coding for functions that are not essential for
56 survival under the controlled and stable conditions present in a laboratory or a bioreactor. Removing
57 these unessential genes (albeit this essentiality has to be accounted for in its full network context)
58
59
60

1
2 would eliminate undesired cross-talking with the standard modules. Moreover, a less complex genome
3 will decrease the possibility of unknown genetic interactions and thus increase the predictive power of
4 the *in silico* models, in a process akin to the reduction of complexity in conventional engineering
5 processes. An ideal chassis would thus contain only the necessary functions to support the implemented
6 modules. Such a chassis may take the form of a simplified living cell or an artificial *in vitro*
7 environment, unable to grow, evolve or self-maintenance, and thus not alive.
8
9

10 11 **Protocells**

12
13 The genomics and the protocell branches follow opposite approaches that can also be denoted as top-
14 down or bottom-up respectively. These two categories not only pursue to construct a chassis for SB, but
15 their results can give fundamental insights into profound biological questions regarding the origins and
16 the contingency of life. As part of SB, the notion of minimal cells has been studied *in silico* and applied
17 in the laboratories. At the beginning of the last decade, spurred by the remarkable developments in
18 genomics, researchers oriented their studies on understanding what can be a minimal cell, its functions,
19 and its genome size.
20
21

22
23 The so-called bottom-up approach consists of the reconstitution of living cells (or protocells) from
24 scratch by designing and constructing new biological parts, devices, and systems in combination with
25 existing natural biological systems. The initial step is the analysis of *in vitro* systems to produce cell-
26 free chassis that are simpler than any living entity (lack of membranes and reduced number of
27 elements), while retaining the ability to express genes. The elements of the systems can be rationally
28 assembled, although the large number of involved species make this a formidable challenge. On the
29 other hand, the medium can be enriched with cell extract, although, whenever its molecular
30 composition is not fixed, we may introduce undesired interactions into the systems.
31
32

33 To construct a protocell, a membrane to separate it from its environment is needed and this is done by a
34 lipid bi-layer. Simple genetic constructs (a promoter, a coding sequence and cell-free extract) produce
35 the corresponding protein within this environment. More complex multistage reaction systems, like
36 protein production from cascade genes have also been accomplished within lipid vesicles, showing
37 their potential to sustain biologically functional modules. Lipid vesicles vary in size from a few
38 nanometers to tens of micrometers and can be used as self-assembled nanoreactors generating
39 environments for a wide range of reactions.
40
41

42
43 Research in protocells, unlike that on *in vitro* systems, aims at developing entities that can be
44 considered alive. Lipid addition to the membrane produces initial vesicle enlargement until an unstable
45 state is reached that triggers the eventual division into daughter protocells. The vesicles lack growth
46 and division-coordination mechanisms, causing that only some of the daughter vesicles will entrap all
47 the components allowing them to maintain their enzymatic capabilities. This aspect deprives the
48 protocells from the self-replication, self-maintenance and evolution properties that would allow us to
49 consider them as living entities.
50
51

52 Both the *in vitro* systems and the protocells are expected to become more apt for their use as chassis for
53 SB as they develop more complex functions, especially those related to self-maintenance: ATP
54 regeneration, protein regeneration (to avoid the effects of aging or of unfavorable conditions)...
55 Unfortunately, the addition of new molecular species enlarges the complexity of the systems and of the
56 mathematical models. Models describing protein synthesis in such systems are able to account for the
57 time varying concentrations of proteins and gene transcripts, however, modeling multi-enzymatic
58
59
60

1
2 reactions remains a challenge. Hold and Panke analyzed *in vitro* glycolysis in *E. coli* and concluded
3 that the assembly of a mathematical model from the kinetic equations of separate enzymes is not yet
4 fully possible due to the lack of a complete understanding of the kinetics of some of the involved
5 enzymes. Nevertheless, the simplicity (lack of membranes among others) of the *in vitro* systems
6 facilitates inference of the equations and their parameter values from dynamic experiments [7].
7
8
9

10 11 **Genome streamlining and minimization**

12
13 The top-down approach focuses on removing existing parts considered as non-essential for cellular
14 growth and division to reduce biological systems towards their functionally essential core. Free-living
15 organisms require more genes for survival in nature than parasitic organisms, so a reductionist rationale
16 proposes that differences in their genomes can establish a list of potentially non-essential genes for
17 survival[8].
18

19
20 In 1986, with initial knowledge of the yeast genomes at hand, the question of non-essential parts for
21 cell growth and division was raised. Ten years later a hypothetical 562kb minimal genome was inferred
22 from *B. Subtilis*. *Mycoplasma genitalium*, a parasitic bacterium and *Haemophilus influenza*, the first
23 sequenced free-living organism were analyzed in parallel to extract a so-called “minimal gene set”. By
24 comparative genomics of both strains, a core of 256 genes was established as a potential gene set
25 necessary and sufficient for sustaining the essential cellular functions. Three years later global
26 transposon mutagenesis was used for both *M. genitalium* and *M. pneumonia* to detect the position of
27 genes which could be considered as essential under specific growth conditions and 265 to 350 protein-
28 coding genes were highlighted.
29

30
31 Being essential may involve belonging to the minimal gene set. On the contrary, the minimal gene set
32 may not include only essential genes. So questions regarding the nature of the minimal gene set and its
33 relationship with the essential genes under specific growth conditions remain open. The analysis and
34 deletion of genomic islands recently acquired by *E. coli* led to a 15% reduction of its genome (under
35 laboratory conditions) concluding to the non-essentialness of the removed genes. The systematic
36 inactivation of *B. subtilis* genes, showed that 192 among the over 4000 genes of the organism were
37 indispensable.
38

39
40 These progresses lay a basis for future development on the streamlining of genomes. However,
41 although the knowledge on individual genes and their function (essential or not) remain important, a
42 determining factor is their role in the context of the cellular networks into which they participate.
43

44
45 Combining *in silico* comparative genomics and *in vivo* global knockout mutagenesis, most of the genes
46 of the minimal gene set described for *M. genitalium* have been proved to be either universal or
47 conserved in all bacteria. Current efforts tend to determine essential functions and adapt them for each
48 organism instead of typical essential genes. An active field of research is the streamline of genomes for
49 their use in SB [9], for example the EU funded project PROBACTYS, that aims at constructing a
50 bacterial cell devoid of most of its genome and endowed with highly coordinated, newly assembled
51 genetic circuits for the bio-degradation and bio-transformation of a range of chloroaromatics
52 (www.probactys.org).
53

54 55 **Synthetic genomes**

56
57 Modern DNA chemical synthesis methods have succeeded in chemically synthesizing and assembling
58 the complete 582,970–base pair *M. genitalium* genome, which was afterwards cloned into *S. cerevisiae*
59
60

1
2 cells and shown to reproduce the original phenotype. Similarly, *Mycoplasma mycoides* genome was
3 introduced in yeast and altered with yeast genetic systems, then it was transplanted in *Mycoplasma*
4 *capricolum* to produce a viable *M. mycoides* cell [10]. This significant breakthrough and the continuous
5 decrease of the prices of gene synthesis will soon allow the construction of fully artificial genomes. A
6 major hindrance to this, however, is that the function of many genes and of their interactions within the
7 cell remain poorly known. Thus, the technological developments on genome synthesis have to go hand-
8 in-hand to those on high-throughput functional genomics.
9
10

11
12 Biological systems show a high degree of modularity and the components of regulatory, signaling, and
13 metabolic networks are highly conserved in multiple organisms, enabling the use of comparative
14 genomics to reconstruct large functional modules in each of these networks. The same techniques have
15 also begun to shed light on the rules governing genome organization, regarding operons and their
16 distribution across the genome. The factorization of the bacteriophage T7 y replacing nearly a third of
17 its genome with an engineered DNA sequence produced a viable organism that maintained key features
18 from the wild-type but resulting in a simpler and easier-to-model organism than the original [11]. This
19 factorization shows the possibility to disentangle the intricate net of interrelations between the genes
20 and to enforce modularity in synthetic genomes. Projects aiming at producing protocells and minimal
21 cells will eventually produce the knowledge on the minimal set of genes and functions for an organism
22 to be viable and on their organization. This information will allow us to factor natural genomes and to
23 design synthetic genomes [12].
24
25
26
27
28
29

30 **4 Chassis modeling and characterization**

31
32 As we have stated in the introduction, in SB construction and design are tightly bonded, thus the
33 optimal biological frame for a given synthetic biology application, scaffold, includes predictive,
34 quantitative genome-scale models of the chassis, describing the interactions of all genes and
35 metabolites at the genome level. This way we will be able to analyze the behavior of the well-defined,
36 modular and standard parts when introduced into the already existing networks of the chassis. The need
37 for modeling arises from the idea that a living organism is more than the sum of its constituents and that
38 new emerging properties may arise from the interaction between its components. Systems biology
39 represents a holistic approach to understand living organisms and advances in this field have been
40 intimately linked to the development of sequencing technologies as well as the development of all
41 kinds of 'omics' data-sets. The huge amount of information generated by high-throughput data-
42 acquisition technologies and the developments in functional genomics are increasing our understanding
43 of life's processes. In the following subsections we will describe current efforts to describe, at genome
44 scale, two different type of networks present in living systems: transcriptional regulatory networks,
45 accounting (at least partially) for gene expression auto-regulation, and metabolic networks.
46
47
48
49
50

51 **Transcription Network Inference**

52
53 High-throughput data on gene expression levels offer the information to extract, at least partially, the
54 structure of transcription regulatory networks within the system, in such a way that heterologous gene
55 expression profiles can be predicted. However, the number of genes in an organism (even with a small
56 genome) is of the order of thousands while the number of experiments in a typical compendium is in
57 the best cases of hundreds, thus we are faced to a problem with more unknowns than data points. In
58
59
60

1
2 addition, there are other factors that regulate gene expression, such as metabolite concentrations,
3 enzymatic activity, diffusion across membranes, environmental factors inducing post-translational
4 modifications of proteins... These elements are not directly reflected in transcripts or protein levels and,
5 generally, we lack genome wide measurement of these effects. Methods to infer regulatory networks
6 deal with this problem by limiting themselves to consider only regulatory factors with measured levels,
7 which amounts to assume that regulation in the system is only caused by transcription factors.
8 Moreover, the dimensionality of the problem is further reduced by an initial clustering of co-expressed
9 genes, that are so assumed to be co-regulated. Three well-established approaches to network inference
10 can be distinguished [13].
11
12

13 Logic network inference methods produce digital models with genes that can only be in a discrete set
14 of states (usually only on/off states are considered) and with interactions described by binary operators.
15 The space of possible graphs is searched (by exhaustive or heuristic methods) to determine the
16 topology of the network that best fits the data.
17
18

19 The second type of approaches is based on constructing, for each pair of genes, a statistical estimator of
20 their mutual dependence through all measured conditions. The inference process determines the
21 background levels to extract significant values.
22
23

24 The third type of methods use differential equations to represent the time evolution of the concentration
25 of each protein or gene transcript, include contributions from a subset of the rest of the genes. The
26 inference work focuses on obtaining the set of parameters best fitting the data. The parameter adjusting
27 phase implies such a computational cost that only small groups of genes or experiments can be
28 considered at a time.
29
30

31
32 The quality of the predictions is further increased by the use of additional information: verified
33 transcriptional factor–gene interactions, operons, DNA sequence binding motifs, data from comparative
34 genomics etc. Perhaps the best approach should be to combine the broad topology prediction properties
35 of the first two methods with the detailed analysis performed by the third, to obtain not only the
36 topology of the network, but also the dynamical parameters needed to predict transcript levels and
37 model the chassis.
38
39

40 Realistic and complex *in silico* benchmarks have been developed to systematically assess the
41 performance of the inference methodologies. Cantone et al. have implemented a synthetic five gene
42 network in *S. cerevisiae* with an intricate topology (feed back loops and protein-protein regulatory
43 interactions). No other genes in the genome regulated any of the elements of the network and, albeit its
44 small size, this network provides a standard platform to validate reverse engineering approaches [14].
45
46

47 **Genome-scale metabolic models**

48
49 Systems biology has produced genome-scale metabolic reconstructions for many of the preferred
50 chassis in SB. These constraint-based models provide a framework to explore and understand global
51 microbial metabolism, allow the design of metabolic engineering strategies for desirable compounds
52 and can be used to predict the effect of mutations or gene deletions in the cellular metabolism as well
53 as to assess the metabolic cross talk between the devices and the chassis. These (quasi)steady state
54 models can be described using metabolic flux balance requiring only stoichiometric information. This
55 is a major advantage as compared to dynamic models, which, although able to predict the kinetics and
56 the time-dependent responses of the network to external stimuli require a detailed knowledge of the
57
58
59
60

1 enzymatic reactions and of their rates. In practice, this means that dynamic models usually can describe
2 only relatively small networks, whereas constraint-based metabolic models can account for the whole
3 metabolic and transport networks of the chassis under study.
4
5
6
7

8 9 **Static models**

10 Metabolic flux balance of a system with m species where r reactions occur, relies on the construction of
11 an $m \times r$ stoichiometric matrix, N , that describes the variations of the vector of concentrations of the
12 internal metabolites X as a function of the vector of reactions $V(X)$: $dX/dt = N \cdot V(X)$. The (quasi)
13 steady state hypothesis implies that metabolites are not consumed or produced and their overall
14 concentrations do not vary in time. As such, this is an under-determined system with infinitely many
15 different solutions, where algebraic techniques can be applied to explore the space of allowed solutions
16 representing meaningful routes through the networks. An elementary flux mode is a minimal set of
17 reactions that can lead to a steady state. This concept provides the tools to define the set of minimal
18 metabolic pathways feasible within a network and contains all the routes able to produce a metabolite
19 from a given substrate allowing to perform robustness and sensitivity analysis for the production of the
20 compound. A related approach is based on the study of extreme pathways, that differ from the
21 elementary flux modes in the treatment of internal reversible reactions (decomposing them into two
22 separate reactions) and can be used to perform similar analysis. Despite the mathematical similarity
23 between both approaches, they differ in practical aspects relevant to the analysis of real metabolic
24 systems [15].
25
26
27
28
29

30 A complementary approach, flux balance analysis, uses linear optimization techniques to look for a
31 single optimal solution that maximizes an objective function, provided that certain physiochemical
32 constraints on the fluxes are satisfied. Boolean relations between the genes can be added to represent
33 regulatory information and the inclusion of Gibbs free energy change rates ensures the thermodynamic
34 feasibility of the predicted pathways. To further narrow the search space empirical information on the
35 fluxes in the wild type strain can be added to predict the effects of mutations or gene deletions. The
36 cost function is usually related with essential cell functions, like ATP or biomass production. The
37 description of real cellular metabolism is nevertheless hindered by the fact that the cellular metabolism
38 may operate in a suboptimal mode regarding the chosen objective function. For biotechnological
39 applications the goal is to maximize the yield of a given bio-compound (while maintaining cell
40 viability). The simultaneous consideration of several competing objectives (e.g. growth and metabolite
41 formation) cannot define a single global maximum, but a set of optimal solutions, each of them
42 representing a different trade-off between the objectives. Thus complex systems, performing multiple
43 functions simultaneously, require the use of multi-objective optimization techniques to obtain the set of
44 solutions showing an optimal trade-off [16].
45
46
47
48

49 **Dynamic models**

50 Dynamic modeling of metabolic systems provides information of time-dependent variation of
51 metabolite concentration and reaction rates. Due to the complexity in elucidating and quantifying
52 enzyme kinetics, the size of those models is often restricted to small parts of the entire metabolic
53 network such as the tryptophan synthesis of *E. coli*, the glycolytic pathway of yeast, and the aspartate
54 metabolism of *Arabidopsis* etc. As metabolic reactions are closely associated with genetic regulation
55 and cellular signaling, the latter have also been modeled dynamically to shed some light on the control
56 mechanisms of cells and how these might affect metabolism. For example, the mammalian circadian
57 rhythm is found to be associated with cyclic metabolic behavior and the circadian transcriptional
58
59
60

1
2 regulation has been modeled to understand the dynamic relationship among the genetic network
3 components and to optimize the connectivity for targeted oscillations. The dynamic modeling of
4 signaling mechanisms can also be potentially connected to metabolic activities to give a more complete
5 picture of the dynamic interactions between metabolic reactions and genetic regulatory responses. The
6 major limitations in the predictive capability of quantitative dynamic models lie in the ambiguity of the
7 details of biological interaction of the modeled system and the multiple sets of possible parameter
8 solution resulting from a larger number of model parameters than the measured experimental data,
9 resulting in the need of an iterative approach to identify a biologically authentic model structure to suite
10 individual application. Currently, dynamic models are being used for designing synthetic gene
11 oscillators in bacterial and mammalian systems. With continuous development in dynamic modeling
12 towards systems of larger scale, as well as the ever increasing technological advances in high-
13 throughput analytical methods which would enable accurate measurements of model parameters, it may
14 eventually be possible to combine reaction kinetics with steady-state metabolic model at the genome-
15 scale to model the entire metabolic dynamics for cell optimization and re-programming.
16
17
18
19
20
21

22 **5 Computational tools and platforms**

23
24 Standard scaffolds should allow an iterative circuit design phase. In the first step the models of the
25 standard parts and devices will be computationally assembled. Afterwards, it is important to be able to
26 simulate their response in an environment where only the cellular resources they require are present.
27 Upon iteration, these steps would produce a set of parts and devices that, if introduced in a truly
28 orthogonal system, would show the required specifications. The following steps would be to simulate
29 the interaction with the chassis and to assess the mutual effect of the chassis and the biological modules
30 to either modify the initial design or to explore the parameter space of the different components to
31 assess the robustness of the design. Once the *in silico* analysis has analyzed the performance of the
32 designed system, then and only then would the construction phase start, see Figure 2. Software
33 initiatives like the E-Cell software, BioSPICE or the Virtual Cell aim at developing tools for the
34 integration of metabolic, signaling, and transcription networks with the goal of constructing a holistic
35 model allowing for precise whole cell simulations, coupling genome scale signaling networks and
36 global regulation. Other platforms are ToBiN and the JWS online database.
37
38
39

40 The integration of cumulative knowledge on parts and their characterization, devices, and chassis
41 structure and networks require the development and enforcement of standard exchange formats for the
42 models like the Systems Biology Markup Language (SBML), the cellular markup language, or
43 Minimum Information Requested in the Annotation of biochemical Models. In addition, visualization
44 tools and languages like the Systems Biology Graphical Notation and CELLDESIGNER allow
45 exporting the models to formats readily accepted by simulators [178]
46

47 Tools for the design of new parts have already been delivered by the *in silico* branch of SB as well as
48 tools for the design and optimization of small-sized circuits. New circuits are computationally designed
49 by modifying either their topology or their parameters. The resulting circuit is simulated and its
50 dynamics is compared to the target function through an appropriate fitness function. The space of
51 topologies and parameters is explored through an iterative optimization procedure [12].
52
53

54 Computational tools able to explore biochemical and pathways databases to locate pathways
55 consuming or producing a given molecule or metabolite have already been developed. Organic
56 chemistry developed long since the retro-synthesis approach, where the transformations on functional
57 groups are considered instead of the entire structure of the metabolite. A similar 'retro-biosynthetic'
58
59
60

1
2 approach specifies pathways for the production of compounds without a known natural counterpart
3 [18]. This approach requires protein evolution techniques like directed evolution or protein design [6]
4 to create highly proficient tailor-made enzymes with the required catalytic specificities.
5
6

7 8 **6 Conclusions**

9
10 Evolution has optimized the chassis currently used in SB for ecological fitness and, as such, they are
11 devoid of the main characteristics sought by synthetic biologists, like simplicity or standardization.
12 However, other characteristics, like modularity or robustness are a common feature of natural systems
13 that can be exploited to our advantage during the design and construction of new scaffolds.
14

15
16 As described above, standard scaffold developed for SB presents formidable challenges related to the
17 chassis construction, characterization, and modeling. In addition, these scaffolds should show a
18 standard and common interface to allow the integration of biological modules in such a way that their
19 behavior can be predicted and their cross-talk completely avoided, in the best cases, or at least kept to a
20 minimum.
21

22
23 The biotechnological applications of SB span a broad spectrum and it is unlikely that scaffolds optimal
24 for a particular microbial production will be the most adequate for other purposes like energy
25 production or development of new therapeutics targeting specific genes or tissues. Thus, a practical
26 solution might be to develop a set of standardized scaffolds, each optimized for a specific application
27 and endowed with a set of “compatible” parts and devices like orthogonal mechanisms for regulation or
28 devices encoding new functionalities that can be easily assembled within the chassis. The vision would
29 be thus the development of high-throughput design that would enable the generation of *à la carte*
30 chassis for cellular re-programming. These set of parts and devices should be well characterized and
31 contain common interfaces, and able to be readily integrated into the *in silico* model of the chassis
32 which is crucial in the design and construction of new biological systems.
33
34
35
36

37 **References**

- 38
39 1 Deplazes, A. Piecing together a puzzle. An exposition of synthetic biology. *EMBO reports* 10, 5,
40 428–432 (2009)
- 41
42 2 Andrianantoandro E, Basu S, Karig DK, & Weiss R. Synthetic biology: new engineering rules for
43 an emerging discipline. *Mol Syst Biol* 2, (2006).
- 44
45 3 Win, M.N. & Smolke, C.D. Higher-Order Cellular Information Processing with Synthetic RNA
46 Devices. *Science* 322, 456-460(2008).
- 47
48 4 Ro D, Paradise EM, Ouellet M, Fisher KJ et al. Production of the antimalarial drug precursor
49 artemisinic acid in engineered yeast. *Nature* 440, 940-943(2006).
- 50
51 5 French, C.E. Synthetic biology and biomass conversion: a match made in heaven? *J R Soc*
52 *Interface* 6:S547-S558(2009)
- 53
54 6 Suárez, M. & Jaramillo, A. Challenges in the computational design of proteins. *J.R.Soc Interface*
55 (2009) 6:S477-S491.
- 56
57 7 Hold, C. & Panke, S. Towards the engineering of in vitro systems. *J R Soc Interface* 6:S507-S521
58 (2009)
- 59
60 8 Fehér T, Papp B, Pal C & Pósfai G. Systematic genome reductions: theoretical and experimental
approaches. *Chem Rev.*107(8):3498-513(2007)

- 1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
- 9 Forster, AC. & Church, GM. Towards synthesis of a minimal cell. *Mol. Syst. Biol.* **2**, 45(2006).
- 10 Gibson, DG, Benders GA, Andrews-Pfannkoch C, Denisova EA, et al. Complete Chemical
11 Synthesis, Assembly, and Cloning of a Mycoplasma genitalium Genome. *Science* **319**, 1215-
12 1220(2008).
- 13 Lartigue C, Vashee S, Algire MA, Chuang RY, et al. Creating Bacterial Strains from Genomes That
14 Have Been Cloned and Engineered in Yeast. *Science*. 2009. DOI: 10.1126/science.1173759
- 15 Chan, LY, Kosuri, S. & Endy, D. Refactoring bacteriophage T7. *Mol Syst Biol.* **1**, 20050018(2005).
- 16 Carrera, J, Rodrigo, G & Jaramillo, A. Towards the automated engineering of a synthetic genome.
17 *Mol. BioSyst.* **5**:733-743(2009)
- 18 Bansal M, Belcastro V, Ambesi-Impiombato A, Bernardo DD. How to infer gene networks from
19 expression profiles. *Mol Syst Biol.* (2007);3:78.
- 20 Cantone I, Marucci L, Iorio F, Ricci MA et al. A Yeast Synthetic Network for In Vivo Assessment
21 of Reverse-Engineering and Modeling Approaches. *Cell* **137**, 172-181(2009).
- 22 Papin JA, Stelling J, Price ND, Klamt S, et al. Comparison of network-based pathway analysis
23 methods. *Trends Biotechnol* **22**, 400-405(2004).
- 24 Kim, H.U., Kim, T.Y. & Lee, S.Y. Metabolic flux analysis and metabolic engineering of
25 microorganisms. *Mol. BioSyst.* **4**, 113-120(2008).
- 26 Endler L, Rodriguez N, Juty N, Chelliah V et al. Designing and encoding models for synthetic
27 biology. *J R Soc Interface* **6**:S405-S417(2009)
- 28 Prather, K.L.J. & Martin, C.H. De novo biosynthetic pathways: rational design of microbial
29 chemical factories. *Current Opinion in Biotechnology* **19**, 468-474(2008).

Figure legends

Figure 1. A scaffold for SB is composed both by the biological chassis and a predictive genome scale model integrating knowledge on metabolism, transcriptional regulation, protein-protein interactions, signaling processes, cellular structure, cell-cell communication, and cell ecology... The chassis should also be endowed by molecular biology tools (plasmids, standard a parts) and a set of specifications and protocols for its manipulation.

Figure 2. The design cycle in SB. A detailed model of the biological chassis that will host the designed systems would allow a systematic simulation and assessment of the interaction between the introduced circuits and the host, which in turn would alter the design of the circuit itself.

For Peer Review

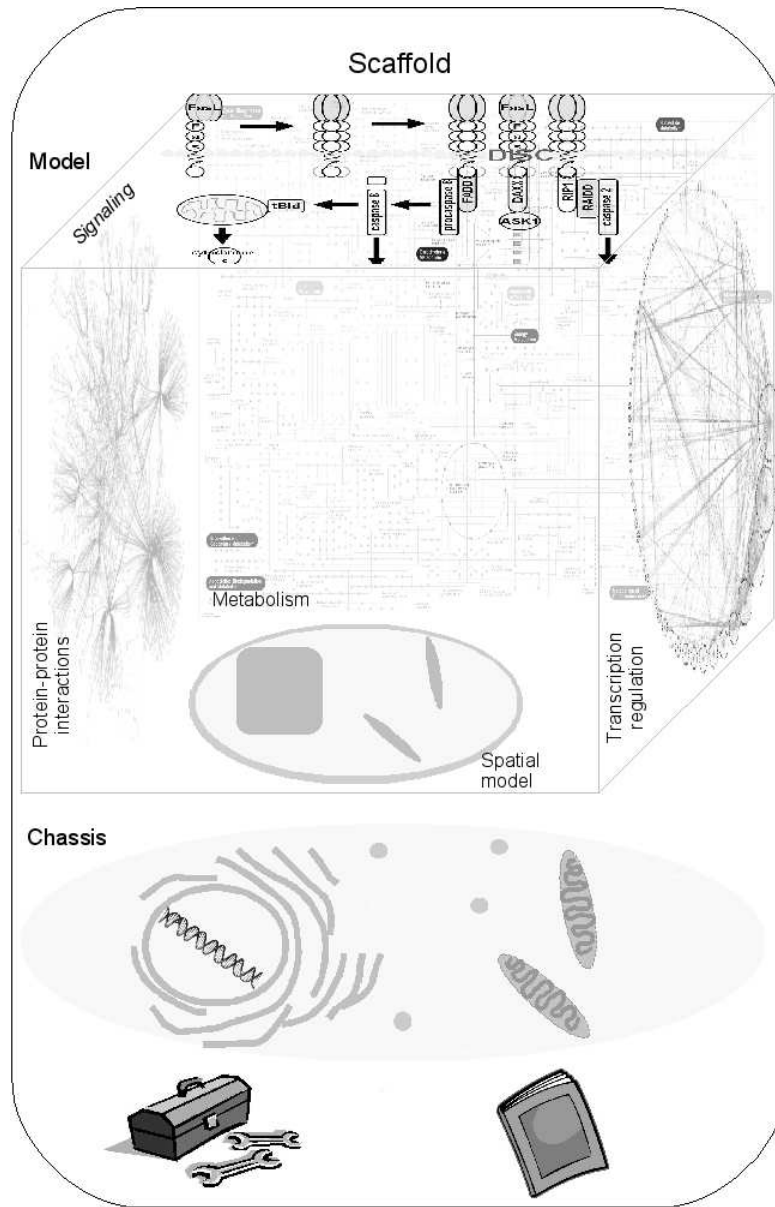


Figure 1. A scaffold for SB is composed both by the biological chassis and a predictive genome scale model integrating knowledge on metabolism, transcription regulation, protein-protein interactions, signaling processes, cellular structure, cell-cell communication, cell ecology, non-transcriptional regulation... The chassis should also be endowed by molecular biology tools (plasmids, standard parts) and a set of specifications and protocols for its manipulation.

229x353mm (72 x 72 DPI)

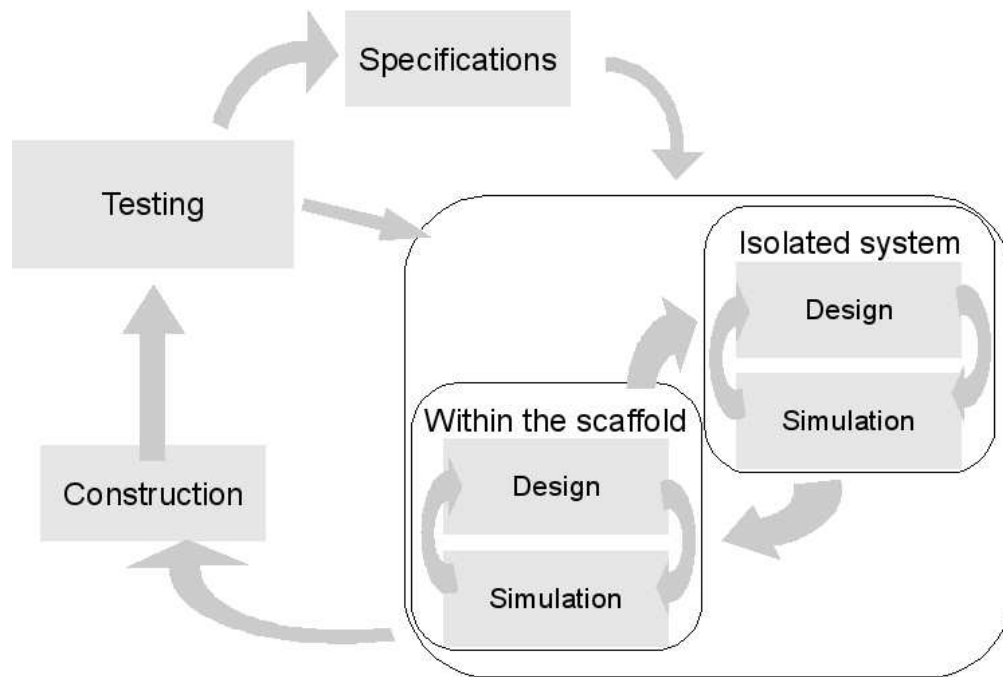


Figure 2. The design cycle in SB. A detailed model of the biological chassis that will host the designed systems would allow a systematic simulation and assessment of the interaction between the introduced circuits and the host, which in turn would alter the design of the circuit itself.

263x176mm (72 x 72 DPI)