BRIDGING THE GAP BETWEEN DESIGN AND REALITY
A Dual Evolutionary Strategy for the Design of Synthetic Genetic Circuits

J. S. Hallinan, S. Park and A. Wipat
School of Computing Science, Newcastle University, NE7 4RU, Newcastle upon Tyne, U.K.

Keywords: Synthetic biology, Evolutionary computation, Directed evolution, Genome-scale design.

Abstract: Computational design is essential to the field of synthetic biology, particularly as its practitioners become more ambitious, and system designs become larger and more complex. However, computational models derived from abstract designs are unlikely to behave in the same way as organisms engineered from those same designs. We propose an automated, iterative strategy involving evolution both *in silico* and *in vivo,* with feedback between strands as necessary, combined with automated reasoning. This system can help bridge the gap between the behaviour of computational models and that of engineered organisms in as rapid and cost-effective a manner as possible.

1 INTRODUCTION

The nascent field of synthetic biology aims to produce engineered organisms with novel, desirable behaviour. To date, synthetic genetic circuits have primarily been designed manually, by a domain expert with an in-depth knowledge of the biological system of interest. This approach has been moderately successful; bacteria, and even plants, have been engineered to perform tasks as diverse as the detection of arsenic in well water, the identification of explosive residues in soil, and the performance of a range of computational tasks such as the operation of logic gates and mathematical functions (Khalil and Collins, 2010).

However, the ultimate aim of synthetic biology is the large-scale engineering of entire genomes. Important strides in this direction have been made (Cello et al., 2002); (Smith et al. 2003); (Tumpy et al., 2005). In 2010 Gibson and colleagues announced the synthesis of a completely synthetic genome, and its insertion into a living bacterium which had previously been denuded of its genome (Gibson et al., 2010). However, all of the work done in this area to date has focussed upon the re-creation, with slight modifications, of existing genomes. To date the design of entire genomes with appreciable novel functionality has not been achieved.

It is becoming increasingly apparent that the design of novel, genome-scale biological systems will require computer-aided design (CAD) and computational simulation prior to implementation (Cohen 2008). Several CAD systems (Chandran et al., 2009); (Czar et al., 2009); (Pedersen, 2009); (Beal et al., 2011), including a data and workflow management system (www.clothocad.org) have been designed specifically for synthetic biology. In addition, a synthetic biology-specific ontology, SBOL, (http://hdl.handle.net/1721.1/66172) is under active development.

However, manually-oriented CAD systems will almost certainly not scale to the genome level. In order to design large-scale synthetic biological systems the complex process of genetic circuit design, implementation, evaluation, modification and iterative refinement will have to be automated as fully as possible.

A design for a synthetic genetic circuit is usually initially in the form of a conceptual diagram, which can be converted into a simulateable model in a standard modelling language. However, converting such a model into a DNA sequence which can be inserted into a living organism is not so straightforward; there is a gap between design and successful implementation, which must be addressed.

Natural systems have arisen via the process of evolution, and there has been considerable interest in the application of evolutionary approaches to the design of novel genetic circuits. In this paper we briefly review the application of both computational

S. Hallinan J., Park S. and Wipat A..
BRIDGING THE GAP BETWEEN DESIGN AND REALITY - A Dual Evolutionary Strategy for the Design of Synthetic Genetic Circuits.
DOI: 10.5220/0003887002630268
In *Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms* (BIOINFORMATICS-2012), pages 263-268
ISBN: 978-989-8425-90-4

and directed evolution to the design of biological systems, and present our vision of a dual evolutionary strategy to bridge the gap between *in silico* design and *in vivo* reality.

2 BACKGROUND

2.1 Evolutionary Computation for Genetic Circuit Design

The manual design of circuits has the advantage of producing simple, well-understood circuit layouts. However, this approach relies heavily upon domain expertise; a biologist with extensive knowledge of the circuit to be engineered, and any extraneous components to be incorporated, is essential. An alternative approach is to incorporate techniques inspired by the only process yet known to have successfully produced life – evolution.

Evolutionary computation (EC) has been around almost since computers became a consumer item (Box 1957). Based on biological evolution, EC attempts to use random changes in a problem solution, together with a fitness function and fitness-proportional selection, to generate solutions to complex problems. EC is therefore ideally suited to problems in complex, poorly-understood domains, where a good, but not necessarily optimal, solution is essential, but the precise nature of the solution is not. There are many variants of EC (Hallinan and Wiles, 2002)but the basic principles are common to all.

EC has been applied to metabolic engineering, for tasks such as identifying the appropriate genes to knock out in order to maximize the production of biochemicals (Patil et al., 2005) and to optimize parameters for allosteric regulation of enzymes (Gilman and Ross, 1995). Some of the results have been interestingly counter-intuitive (Patil et al., 2005).

The applicability of EC to the design of genetic circuits is clear. Multiple runs of an algorithm will produce different, equally fit, solutions which can be compared for efficiency, cost and practicality of implementation, among other factors. Since the detailed workings of many genetic circuits are poorly-understood, EC is a promising approach to the generation of new circuit designs.

2.2 Directed Evolution *in vivo*

The relationship between a DNA sequence and the structure and function of the protein it encodes is indirect. Many factors affect the relationship, including post-transcriptional and –translational modifications to DNA, RNA and proteins; the presence or absence of protein chaperones; protein folding; and the cellular context. It is therefore non-trivial to design a protein with a required functionality, such as a transcription factor with a given binding strength. An extremely successful way to overcome this problem is to extend, or completely replace, the rational design approach with directed evolution (Romero and Arnold, 2008).

Directed evolution involves the application, to a population of cells, of iterative rounds of mutation and artificial selection. With each round of selection the desired behaviour is more closely approximated, and the process can be ended when the protein function is deemed to be close enough to the target behaviour. Directed evolution has been shown repeatedly to be both powerful and flexible in its outcomes (Aharoni et al., 2005).

There are two ways in which directed evolution is generally used. In the biotechnology industry the output of a particular biological pathway is often of primary interest; companies need to optimise the production of a specific compound (Lee et al., 2008). In this case directed evolution has the effect of optimising entire pathways. Alternatively, the evolutionary process can be aimed at manipulating individual proteins, developing, for example, specific enzymes (Brustad and Arnold, 2011).

Originally, much of the work in this area was performed in large-scale chemostats. The use of smaller volumes then made it possible to automate much of the directed evolution process using liquid-handling robots (Felton, 2003). Such robots, however, still work with relatively large numbers of cells at a time. At this scale the stochasticity inherent in biological systems is averaged out when measurements are made over whole populations of cells, prohibiting analysis of the behaviour of single cells.

Recently, however, there has been considerable interest in the use of microfluidic technologies in synthetic biology (Gulati et al., 2009); (Szita et al., 2010); (Ferry et al., 2011); (Vinuselvi et al., 2011). Operating at micrometre scales, microfluidic devices allow the manipulation and observation of single cells or small groups of cells. Biological stochasticity can thus be explored in detail. Importantly, microfluidics devices can be fully automated, with tasks such as the input of fresh media, removal of waste, selection of individual cells and control of cellular environment completely controlled by an attached computer. Microfluidics

provides an ideal environment for the directed evolution of cells for synthetic biology.

3 A DUAL EVOLUTIONARY STRATEGY FOR SYNTHETIC BIOLOGY

3.1 Design in Synthetic Biology

Synthetic genetic circuits tend to be designed in isolation, and often incorporate a number of simplifying assumptions. However, a designed circuit *in vivo* is operating in a complex genetic and environmental context, and an engineered microbe may not behave in the same way as the *in silico* model upon which it is based.

The creations of synthetic biologists must operate predictably in a complex, noisy environment in which they are subject to global selection pressures as yet poorly understood. However, a number of important issues have been identified. Sprinzak and Elowitz (Sprinzak and Elowitz, 2005) nominate “parameter sensitivity, the lack of effective rules to simplify complex circuits, and the difficulty of incorporating extrinsic noise". To this list we add: nonlinearity; crosstalk; scalability; evolvability; and genetic context.

Rather than attempting to eliminate this noise and complexity, we believe that it should be possible to harness the incredibly powerful forces that have shaped life on this planet for the past 3.8 billion years for the controlled design of organisms with novel, valuable behaviours.

3.2 A Dual Evolutionary Strategy

Our proposed dual evolutionary strategy involves *in silico* and *in vivo* experimentation and evolution carried out in serial, with results from each strand feeding back to the other strand as required.

The process starts, as does all formal engineering design, with requirements gathering, leading to a formal functional specification of the desired system. Data from a variety of sources are then integrated to inform automated reasoning, leading to an initial design for the system.

The field of data integration is increasingly being recognised as important to bioinformatics, whose practitioners routinely deal with the large datasets produced by high-throughput technologies such as microarrays and proteomics. Although much of this data is freely available in the over 1300 online databases currently available (Galperin and Cochrane, 2011), the sheer scale of data generation means that much of this data does not make it into the literature. It is not feasible to manually trawl databases for more than a small number of genes. Data, and thus information, can effectively be lost to the research community.

Tools such as the Ondex data integration platform (Kohler et al., 2006) rescue this data by bringing it together in a common format and integrating diverse datasets into a single resource, which can be viewed as a network, or accessed and manipulated computationally. Ondex incorporates an underlying ontology, so individual concepts, which can be of any type (gene, protein, publication, protein family, etc.) and their interactions are annotated in a structured manner. Ondex graphs are therefore well suited to the application of automated reasoning algorithms, which have already been applied to good effect in bioinformatics (King et al., 2004). An Ondex knowledgebase has recently been produced for the model Gram positive bacterium *Bacillus subtilis* (Misirli et al., 2011).

The initial reasoning / design process is iterative, as individual designs are scrutinised for genetic components and their desired interactions, which are then reasoned over to predict the systems behaviour and to suggest modifications to the design. Once the initial design is determined it is translated into a computational model in a standard modelling language such as SBML (Hucka et al., 2003) or CellML (Cooling et al.m 2008).

Simulation of the model and experiments on the engineered microbe are conducted in serial. Intially, the model is run, to determine whether it behaves as predicted. Simulation modelling can also establish factors such as sensitivity to variations in parameter values, and to determine which model elements are most important to the generation of the desired behaviour, observations which can be used to guide measurements made on the *in vivo* system. Models may be run multiple times using stochastic algorithms to investigate the range of behaviours possible from a single design (Hallinan et al., 2010).

If the behaviour generated by the model is not sufficiently close to the target, an evolutionary algorithm is used to modify the circuit until is behaves as desired. The modified model is analyzed in the same way as the original model.

Once the modelling results are satisfactory the design is converted into a synthesizable DNA sequence using a tool such as MoSeC (Misirli et al., 2011), an approach which preserves the automated nature of the process. Alternatively, if standard

cloning approaches are too be used, the design may be manually translated into a set of DNA components and the strategy with which to manipulate them.

The *in vivo* evolutionary cycle is then executed, with laboratory experimentation replacing model simulation, measurements made as indicated by the results of the modelling, and directed evolution replacing EC.

The end result of the *in silico* and *in vivo* experimentation is the amassing of large amounts of new data about the construct and its behaviour. This data is added to that in the original integrated dataset to form the basis for further computational reasoning.

Computational evolution of the model will produce multiple, variant designs for genetic circuits with the same functionality, many of which will never have existed in nature. Similarly, directed evolution of the engineered microbes will almost certainly produce a number of organisms which have behaviour closer to that desired. Sequencing of their genomes post-evolution will permit comparison with the original sequence, and thus facilitate the generation of testable hypotheses about the significance of any mutations observed.

All of the data generated by experiments on both the original design and the evolved variants then feed back into the reasoning / design loop, and the process can continue iteratively until an organism is achieved with behaviour close enough to the target.

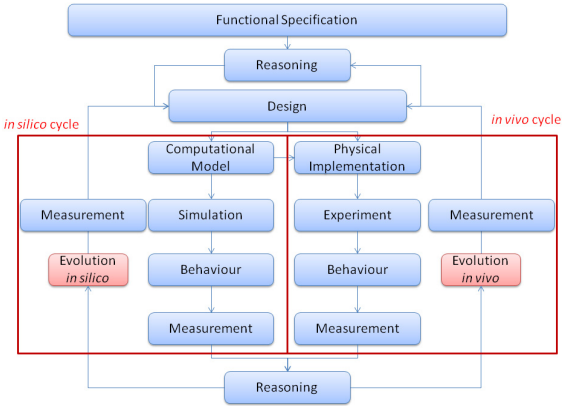

Figure 1: A dual evolutionary strategy for synthetic biology.

4 DISCUSSION

Currently most design in synthetic biology is done on a small scale, in close consultation with a domain expert. Although it may never be possible to move completely away from specialised expertise, the design of large scale genetic systems will clearly require a high level of automation, including automated reasoning over large amounts of data. Synthetic biology builds upon molecular and systems biology (Church 2005), but has a different aim from either of those disciplines: to engineer entirely novel biological systems, performing tasks which are not within the scope of existing organisms. In order to achieve these aims we contend that large scale systems must be engineered; such systems will be of of a size and complexity of which the human brain cannot maintain a complete overview.

Large scale synthetic biology therefore requires sophisticated computation and extensive automation. The algorithms and hardware required to achieve this task are rapidly becoming available. New technologies promise to extend the capabilities of laboratories in many different directions. DNA synthesis technology is increasing in speed, while decreasing in cost (details). Cloud and Grid computing make available enormous amounts of CPU time cheaply (Craddock et al., 2008), and, because these technologies are highly parallel, quickly. Microfluidics provides an exciting, albeit challenging, approach to the manipulation and measurement of cells, either wild type or engineered, in very small numbers.

The development of these technologies permits approaches such as directed evolution at the single-cell level in time scales which are not very different from those required to run multiple computational simulations. Computational and *in vivo* experiments provide different, but overlapping, windows onto the biology of synthetic genetic systems. We therefore propose a bipartite strategy for engineering synthetic genetic circuits, involving both *in silico* and *in vivo* experiments.

One important component of our approach is computational reasoning. The amount of data which can be collected from a single experiment is vast. At present, it is usually the task of the human experimenter to decide which parameters should be measured, and how those measurements should be used in the development of new experiments. Automated computational reasoning has been applied with success to the generation of new testable hypotheses, and appropriate experimental protocols, as in the case of the Robot Scientist (King et al., 2009). There is clearly considerable scope for the application of this approach to automated decisions about which aspects of an experiment to

measure, and which experiments to conduct, in the field of synthetic biology.

The other fundamental aspect of our approach is the use of evolution to refine the designs arrived at by humans or machines. The uncertainty inherent in biological systems—whether arising from inherent stochasticity or our lack of knowledge about the structure and function of many biomolecules—means that a completely rational design strategy in synthetic biology, as espoused by hard-core engineers, is simply not practical at this point in time. By harnessing evolution to refine our design, and then comparing the products of evolution with our original designs, we have the potential to learn not only how to better engineer the organisms in which we are interested, but also how these organisms work in the absence of engineering. Molecular and systems biology form the basis for synthetic biology; but synthetic biology also promises to provide unique insights into the fundamental workings of the cell.

A highly automated approach, incorporating computational intelligence wherever possible, and operating at the level of one or a few cells, appears to us to offer the best prospects for designing, implementing and testing large-scale novel genetic systems, thus bridging the gap between design and reality in synthetic biology. Although there are still many technical hurdles to be overcome in the construction of such a system, all of the individual technologies are currently in place, and the construction of such a synthetic biology factory is a realistic goal in the near future.

REFERENCES

Aharoni, A., L. Gaiducov, et al. (2005). "The 'evolvability' of promiscuous protein functions." *Nature Genetics* 37(1): 73 - 76.

Beal, J., T. Lu, et al. (2011). "Automatic compilation from high-level biologically-oriented programming language to genetic regulatory networks." *PLoS ONE* 6(8): e22490.

Box, G. E. P. (1957). "Evolutionary operation: A method for increasing industrial productivity." *Applied Statistics* 6(2): 81 - 101.

Brustad, E. M. and F. H. Arnold (2011). "Optimizing non-natural protein function with directed evolution." *Current Opinion in Chemical Biology* 15(2): 201 - 210.

Cello, J., A. V. Paul, et al. (2002). "Chemical synthesis of poliovirus cDNA: Generation of infectious virus in the absence of natural template." *Science* 297(5583): 1016 - 1018.

Chandran, D., F. T. Bergmann, et al. (2009) "TinkerCell: modular CAD tool for synthetic biology." *Journal of Biological Engineering* 3 DOI: doi:10.1186/1754-1611-3-19.

Church, G. M. (2005). "From systems biology to synthetic biology." *Molecular Systems Biology* 1: 0032.

Cohen, J. (2008). "The crucial role of CS in systems and synthetic biology." *Transactions of the ACM* 51(5): 15 - 18.

Cooling, M. T., P. Hunter, et al. (2008). "Modelling biological modularity with CellML." *Systems Biology, IET* 2(2): 73-79.

Craddock, T., C. R. Harwood, et al. (2008). "e-Science: Relieving bottlenecks in large-scale genomic analyses." *Nature Reviews Microbiology* 6: 948 - 954.

Czar, M., Y. Cai, et al. (2009). "Writing DNA with GenoCAD." *Nucleic Acids Research* 37(W40-7).

Felton, M. J. (2003). "Product review: Liquid handling: Dispensing reliability." *Analytical Chemistry* 75(17): 397A - 399A.

Ferry, M. S., I. A. Razinkov, et al. (2011). "Microfluidics for synthetic biology: From design to execution." *Methods in Enzymology* 497: 295 - 372.

Galperin, M. Y. and G. R. Cochrane (2011). "The 2011 Nucleic Acids Research database issue and the online molecular biology database collection." *Nucleic Acids Research* 39(Suppl 1).

Gibson, D. G., J. I. Glass, et al. (2010). "Creation of a bacterial cell controlled by a chemically synthesized genome." *Science* 329(5987): 52 - 56.

Gilman, A. and J. Ross (1995). "Genetic-algorithm selection of a regulatory structure that directs flux in a simple metabolic model." *Biophysical Journal* 69(4): 1321 - 1333.

Gulati, S., V. Rouilly, et al. (2009). "Opportunities for microfluidic technologies in synthetic biology." *Journal of the Royal Society Interface* 6(Suppl_4): S493-S506.

Hallinan, J., G. Misirli, et al. (2010). Evolutionary computation for the design of a stochastic switch for synthetic genetic circuits. *32nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2010)*, Buenos Aires, Argentina.

Hallinan, J. and J. Wiles (2002). Evolutionary algorithms. The Encyclopedia of Cognitive Sciences. L. Nadel. New York, *Palgrave Macmillan*.

Hucka, M., A. Finney, et al. (2003). "The systems biology markup language (SBML): A medium for representation and exchange of biochemical network models." *Bioinformatics* 19: 524-531.

Khalil, A. S. and J. J. Collins (2010). "Synthetic biology: Applications come of age." *Nature Reviews Genetics* 11(5): 367 - 379.

King, R. D., J. Rowland, et al. (2009). "The Automation of Science." *Science* 324(5923): 85-89.

King, R. D., K. E. Whelan, et al. (2004). "Functional genomic hypothesis generation and experimentation by a robot scientist." *Nature* 427(6971): 247-252.

Kohler, J., J. Baumbach, et al. (2006). "Graph-based analysis and visualization of experimental results with ONDEX." *Bioinformatics* 22(11): 1383-1390.

Lee, S. K., H. Chou, et al. (2008). "Metabolic engineering of microorganisms for biofuels production: From bugs to synthetic biology to fuels." *Current Opinion in Biotechnology* 19(6): 556-563.

Misirli, G., J. Hallinan, et al. (2011). CS-TR-1237. *Technical Reports*, Newcastle University.

Misirli, G., J. S. Hallinan, et al. (2011). "Model annotation for synthetic biology: automating model to nucleotide sequence conversion." *Bioinformatics* 27(7): 973-979.

Patil, K., I. Rocha, et al. (2005). "Evolutionary programming as a platform for in silico metabolic engineering." *BMC Bioinformatics* 6(1): 308.

Pedersen, M. P., A.; (2009). "Towards programming languages for genetic engineering of living cells." *Journal of the Royal Society Interface*.

Romero, P. A. and F. H. Arnold (2008). "Exploring protein fitness landscapes by directed evolution." *Nature Reviews Molecular Cell Biology* 10: 866 - 876.

Smith, H. O., C. A. Hutchison, et al. (2003). "Generating a synthetic genome by whole genome assembly: phiX174 bacteriophage from synthetic oligonucleotides." *Proceedings of the National Academy of Sciences 100*(26): 15440 - 15445.

Sprinzak, D. and M. Elowitz (2005). "Reconstruction of genetic circuits." *Nature* 438: 443 - 448.

Szita, N., K. Polizzi, et al. (2010). "Microfluidic approaches for systems and synthetic biology." *Current Opinion in Biotechnology* 21(4): 517-523.

Tumpy, T. M., C. F. Basler, et al. (2005). "Characterisation of the reconstructed 1918 Spanish Influenza pandemic virus." *Science* 310(5745): 77 - 80.

Vinuselvi, P., S. Park, et al. (2011). "Microfluidic technologies for synthetic biology." *International Journal of Molecular Sciences* 12(3576 - 3593).