High Performance Computing in Synthetic Biology

JAVIER CARRERA* ** - GUILLERMO RODRIGO* - ALFONSO JARAMILLO* §**

* Instituto de Biologia Molecular y Celular de Plantas, CSIC, Universidad Politecnica de Valencia, Spain
** Instituto de Aplicaciones en Tecnologias de la Informacion y las Comunicaciones Avanzadas (ITACA), Universidad Politecnica de Valencia, Spain
*** Laboratoire de Biochimie, Ecole Polytechnique, CNRS, Palaiseau, France
*** Epigenomics Project, Université d'Evry Val d'Essonne, Genopole, CNRS, Evry, France

Abstract. – *Synthetic biology aims to the design or redesign of biological systems. We present the results of using a computational methodology to design biological systems such as proteins or networks of proteins. In particular, we will transform a thioredoxin protein into an esterase. The enlarged combinatorial space (due to the combination of docking with mutagenesis) challenges our current optimisation techniques. We use a detailed atomic model based only on physico-chemical principles and an automatic computational procedure to search through the sequence and conformational spaces. The new procedure aims to the design of functional proteins by coupling the automated protein design protocol to an automated active site scan. We simultaneously inspect all possible active site placements and completely optimise the surrounding amino acid sequence to allow for binding. We not only optimise the stability of the designed protein but also its function (ligand binding or catalytic activity). We treat the docking and protein design problems on the same footing and optimise them simultaneously. We have designed proteins with either a predefined ligand-binding function or a given catalytic activity using the thioredoxin protein fold as scaffold. We have experimentally tested our predictions in the biophysical lab of Prof. Sanchez-Ruiz (U. Granada). On the other hand, we have adapted recent transcriptional network inference methods to the redesign of global transcription by using HPC. By using a model based on ordinary differential equations, we are able to predict the regulatory network response under cellular changing environments, which opens the way to the whole genome design depending on the availability of computational resources.*

Introduction

The emerging discipline of Synthetic Biology [1] could be defined as the rational engineering of life or biological processes for practical use. It is a discipline at the intersection of protein and genetic engineering with systems biology. Modern computational tools based on algorithms with a high time-consuming activity have made possible to design artificial proteins, enlarging nature's repertoire, but in the future they will allow the design of custom-made organisms.

The present project proposes a HPC application in Synthetic Biology: a new systematic approach to the problem of introducing a binding site into inert thioredoxin protein scaffold (see Fig. 1) with the aim of redesigning new proteins with targeted function. We have used a detailed atomic model based only on physico-chemical principles and an automatic computational procedure to search through the sequence and conformational spaces.

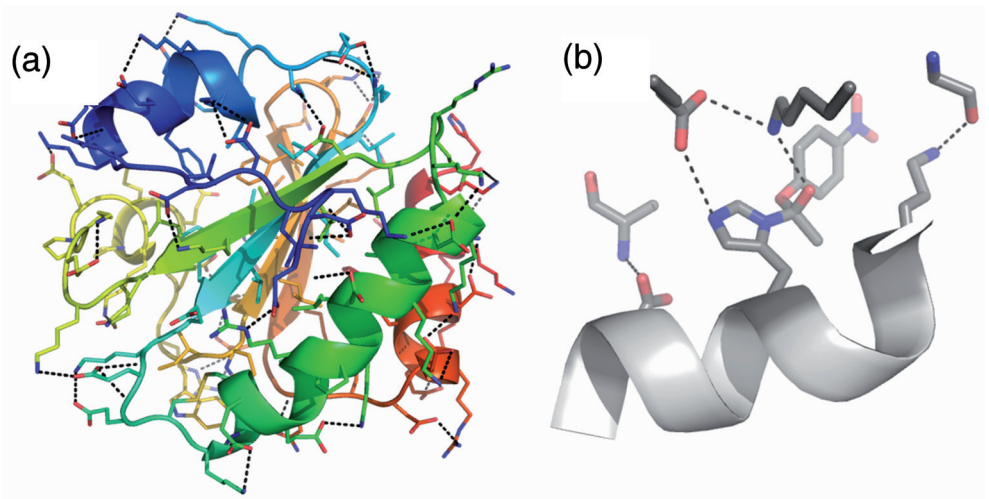

FIGURE 1. – *(a) Folding structure [2, 3] of the redesign of thioredoxin and, (b) detail of its active-site.*

It is possible to use site-directed mutagenesis to graft a catalytic triad into the scaffold of a protein. However, this method cannot be general due to the loss of stability produced by the mutations. Alternatively, directed evolution allows the exploration of many simultaneous mutations and at the selection for protein stability. But here the number of explored amino acid sequences (size of the combinatorial rotamer library) is much smaller than can be achieved by the latest computational optimization techniques and, on the other hand, the computational approach can simulate unnatural situations. In addition, a computational methodology [1] improves our basic understanding on protein structure and function.

Results

Our new procedure aims to the design of functional proteins by coupling the automated protein design protocol to an automated active site scan. We simultaneously inspected all possible active site placements and completely optimised the surrounding amino acid sequence to allow for binding. We did not only optimise the stability of the designed protein but also its function. For the catalytic activity, we

considered an atomic model of the high-energy tetrahedral transition state optimising the binding energy. We treated the docking and protein design problems on the same footing and optimised them simultaneously.

We redesigned a thioredoxin protein with predefined ligand-binding function and catalytic activity (by stabilising a transition state, as with catalytic antibody design). We computed the minimal number of mutations to transform a thioredoxin into an esterase. The enlarged combinatorial space (due to the combination of docking with mutagenesis) challenges our current optimisation techniques. The thioredoxin protein from *E. Coli* was used as a target system by previous enzyme design studies due to its experimental conveniences, to the availability of a high-resolution x-ray structure and to its small size. We also experimentally tested our predictions in the biophysical lab of Prof. Sanchez-Ruiz (U. Granada).

On the other hand, other application in Synthetic Biology requiring HPC has been done thanks to the interaction with the host institute. We developed a methodology [4] to construct a gene regulatory network model to predict the system dynamical behaviour in evolved environments. The identication of regulations is a high time-consuming activity, but the redesign of global transcription regulation is much more time-consuming and it will pose new challenges for future HPC projects. Our procedures were implemented in C++ to run on UNIX environments. We developed the InferGene software, which consists of different functional modules to compute firstly the network topology and then the corresponding kinetic parameters. The running-time scales with the number of genes and the square of the number of conditions. We developed a parallel code to apply our methodology with the full genome of *A. thaliana* comprising 22094 genes (1,187 transcription factors and 1,408,969 pairs of transcription factor interactions). Hence, distributed computing provided the necessary resources to apply our methodology to infer the regulations of genome-wide providing a quantitative global model of *A. thaliana* transcriptional regulation network.

We expect to continue our fruitful collaboration with them in the new HPC-Europa++ programme. Future visits will allow us to implement our algorithms and HPC resources to novel problems.

Acknowledgements. This work is supported by the HPC-Europa programme, funded under the European Commission's Research Infrastructures activity of the Structuring the European Research Area programme, contract number RII3-CT-2003-506079; the Spanish Ministry of Education and Science (ref. TIN 2006-12860), the Structural Funds of the European Regional Development Fund (ERDF), the EU grants BioModularH2 (FP6-NEST contract 043340) and EMERGENCE (FP6-NEST contract 043338), the ATIGE Genopole/UEVE and the MIT-France grants. GR acknowledges a graduate fellowship from the Conselleria d'Educacio de la Generalitat Valenciana (ref. BFPI 2007/160) and an EMBO Short-term fellowship (ref. ASTF-343.00-2007).

References

[1] D. ENDY, Foundations for engineering biology, Nature, **438**:449-453 (2005).

[2] A. JARAMILLO and S.J. WODAK, Computational Protein Design is a challenge for Implicit Solvation Models, Biophysical Journal, **88**:156-171 (2005).

[3] A. JARAMILLO, L. WERNISCH, et al., Folding free energy function selects native-like protein sequences in the core but not on the surface, Proceedings National Academy of Sciences, 99:13554-13559 (2002).

[4] J. CARRERA, G. RODRIGO and A. JARAMILLO, Control-based redesign of global transcription regulation, Nucleic Acids Research (2009, in press).