doi:10.2533/chimia.2014.642 Chimia 68 (2014) 642–647

Lessons from Nature: Computational Design of Biomimetic Compounds and Processes

Esra Bozkurt[a], Negar Ashari[a], Nicholas Browning[a], Elizabeth Brunk[ab], Pablo Campomanes[ac], Marta A. S. Perez[a], and Ursula Rothlisberger*[a]

Abstract: Through millions of years of evolution, Nature has accomplished the development of highly efficient and sustainable processes and the idea to understand and copy natural strategies is therefore very appealing. However, in spite of intense experimental and computational research, it has turned out to be a difficult task to design efficient biomimetic systems. Here we discuss a novel strategy for the computational design of biomimetic compounds and processes that consists of i) target selection; ii) atomistic and electronic characterization of the wild type system and the biomimetic compounds; iii) identification of key descriptors through feature selection iv) choice of biomimetic template and v) efficient search of chemical and sequence space for optimization of the biomimetic system. As a proof-of-principles study, this general approach is illustrated for the computational design of a 'green' catalyst mimicking the action of the zinc metalloenzyme Human Carbonic Anhydrase (HCA). HCA is a natural model for CO_2 fixation since the enzyme is able to convert CO_2 into bicarbonate. Very recently, a weakly active HCA mimic based on a trihelical peptide bundle was synthetized. We have used quantum mechanical/molecular mechanical (QM/MM) Car-Parrinello simulations to study the mechanisms of action of HCA and its peptidic mimic and employed the obtained information to guide the design of improved biomimetic analogues. Applying a genetic algorithm based optimization procedure, we were able to re-engineer and optimize the biomimetic system towards its natural counter part. In a second example, we discuss a similar strategy for the design of biomimetic sensitizers for use in dye-sensitized solar cells.

Keywords: Biomimetic compounds · Computational enzyme design · Density Functional Theory · Dye-sensitized solar cells · Green chemistry · Mixed quantum mechanical-molecular mechanical (QM/MM) simulations · Natural catalysts

1. Introduction

Nature can perform chemical reactions under mild and environmentally benign conditions for which laboratory experiments often have to resort to extreme pressures and temperatures. The idea to develop strategies that are inspired by living systems emerges therefore quite naturally. However, due to the high complexity of biological systems, it is far from trivial to pinpoint the most promising routes and to identify the best possible natural targets and choose suitable biomimetic scaffolds. However, once an appropriate system has been identified, the next step towards the development of biomimetic systems is to understand how Nature 'is doing it'. A comprehensive understanding of the inner workings of biological processes is often a very difficult task since these are notoriously complex events that can involve several orders in length and time scales and experimental techniques probing their mode of action often provide only indirect and partial information.

Computer simulations can nowadays give direct insights into biomolecular mechanisms but the combined challenge of extended size, large available configurational space and high accuracy required to describe small energy differences of the order of kT remains a challenge. In addition, the analysis of high-dimensional simulation data to identify the crucial factors responsible for catalysis that could guide the design of possible biomimetic systems is far from trivial. Furthermore, the size of the chemical and/or sequence space that has to be explored in the search for an optimal system with tailored properties is enormous and necessitates the use of special techniques.

Here, we discuss a general approach for the design of biomimetic systems and processes from the selection of the natural target to its characterization, the determination of its relevant features, the choice of biomimetic template and the inversion of the structure-function relationship, *i.e.* the sampling of chemical space in the search of compounds with desired properties.

This protocol is applied for the computational design of a biomimetic system for CO_2 fixation and to the optimization of biomimetic porphyrin dyes in dye-sensitized solar cells (DSSCs).

2. Computational Strategy for the Design of Biomimetic Systems

The computational strategy that is outlined in the following can be generally applied for the design of biomimetic systems with various functions. However, as an example, we will mainly focus on the development of biomimetic catalysts, *i.e.* compounds that are able to copy the chemistry of living systems. As a proof-of-principle, the approach is illustrated for the design of biomimetic catalysts for CO_2 fixation.

2.1 *Target Selection*

The first obvious step towards the development of a biomimetic system is the identification of a natural target that is able to accomplish the desired process.

*Correspondence: Prof. Dr. U. Rothlisberger[a]
[a]Laboratory of Computational Chemistry and Biochemistry
Ecole Polytechnique Fédérale de Lausanne EPFL
Avenue Forel
CH-1015 Lausanne
Tel.: +41 21 693 0325
E-Mail: ursula.roethlisberger@epfl.ch
[b]Current address: University of California Berkeley
Joint BioEnergy Institute (JBEI)
[c]Current address: German Research School for Simulation Sciences GmbH
D-52425 Jülich Germany

The search for a suitable biological system however can be nontrivial. In the case of biomimetic catalysts, the Enzyme Commission (EC) number,[1] which classifies enzymes according to the chemical reaction that they catalyze, can provide a useful starting point. This enzyme nomenclature scheme characterizes every enzyme-catalyzed reaction with four numbers a–d (EC a.b.c.d). a describes the enzyme family (oxidoreductases, transferases, hydrolases, lyases, isomerases and ligases); b specifies the functional group that the enzyme is acting on; c gives information about cofactors and d identifies the specific substrate. Helpful information can also be gained from the Comprehensive Enzyme Information System BRENDA (BRaunschweig Enzyme Database),[2] a highly inclusive enzyme repository system that lists the currently available data about enzymes and allows for easy searches *via e.g.* reactant or product specification. In addition, an analysis of experimentally identified metabolic pathways as compiled in the MetaCyc database[3] can greatly facilitate the search for suitable enzymatic systems. Furthermore, programs such as the Biochemical Network Integrated Computational Explorer (BNICE)[4] that allows the construction and evaluation of metabolic pathways using a database of generalized enzyme rules based on the EC classification also enables the discovery of novel pathways for the biosynthesis of chemical compounds. In this way, suitable enzymatic pathways for the production of a specific chemical can be identified and the involved enzymes can be re-engineered and optimized for non-natural substrate using atomistic simulations.[5]

The application of these tools for the discovery and design of a biomimetic catalyst for CO_2 fixation suggests different possible systems. In fact, CO_2 serves as a natural substrate in multiple biochemical processes such as for the enzyme RuBisCO (ribulose-1,5-bisphosphate carboxylase EC.4.1.1.39) that is involved in the first major step of carbon fixation during the Calvin cycle or for the ammonia-dependent carbamoyl-phosphate synthase (EC 6.3.4.16). Most of these enzymes however use additional (chemically complex) co-substrates and are comparably slow with catalytic rates of the order of seconds. The most basic system for CO_2 fixation that in fact only involves water as a co-substrate is achieved by carbonic anhydrases (CA) that reversibly catalyze the conversion of CO_2 and water to bicarbonate. In addition, CA are among the fastest enzymes known with rates between 10^4–10^6 s^{-1}. We have therefore chosen the highly proficient enzyme Human Carbonic Anhydrase II (HCAII) (EC 4.2.1.1) as a natural target for the design of a biomimetic catalyst.

2.2 *Target Characterization*

Once a suitable natural system has been identified, the next step consists in finding out how it works. State-of-the-art computer simulations allow for a mechanistic characterization of entire enzymatic cycles with atomistic/electronic detail. The methods of choice are usually mixed quantum mechanical/molecular mechanical (QM/MM) molecular dynamics simulations that can take environment and finite temperature effects into account. Here we use mixed QM/MM Car-Parrinello simulations based on the extended Lagrangian (Eqn. (1))[6]

$$L = \frac{1}{2}\mu\sum_i\int d\vec{r}\,\dot{\psi}_i^*(\vec{r})\dot{\psi}_i(\vec{r}) + \frac{1}{2}\sum_I M_I\dot{\vec{R}}_I^2 - E_{MM} - E_{QM/MM} - E_{QM} + \sum_{i,j}\Lambda_{i,j}\left(\int d\vec{r}\,\psi_i^*(\vec{r})\psi_j(\vec{r}) - \delta_{i,j}\right) \quad (1)$$

for a comprehensive characterization of the natural target. In these simulations, the active site residues and additional chemically active components (*e.g.* catalytic water molecules) are described at the first-principles (density functional theory) level while the remainder of the enzyme and solvent is treated with a classical force field. In combination with enhanced sampling methods (*e.g.* thermodynamic integration along reaction coordinates[7] and metadynamics)[8] all the enzymatic reaction steps can be characterized in detail in terms of reaction intermediates and free energy barriers that separate them.

In the case of HCAII, we applied these techniques for a mechanistic characterization of the deprotonation reaction of the zinc-bound water molecule as well as to the nucleophilic attack of the zinc-bound hydroxide ion with CO_2 and the conversion of the latter to bicarbonate. In addition, we studied substrate entry and binding as well as product release by using combinations of classical and QM/MM simulations. In agreement with previous experimental and computational studies, we find that the chemical steps of the cycle involve maximal barriers of the order of 10 kcal/mol consistent with the exceptionally high catalytic rate.

2.3 *Analysis of Simulation Data: Feature Selection*

QM/MM simulations provide a wealth of data about the structural, dynamic and electronic properties of the system during enzymatic catalysis. The analysis of this rich information in terms of catalytically important versus catalytically irrelevant features can be very difficult. In principle, computer experiments allow to probe the effect of every single residue on the reaction but considering the fact that QM/MM simulations at the first-principles level are rather costly, huge numbers of simulations needed to test the influence of different parts of the system are not feasible. It is therefore important to develop unbiased and systematic ways of analyzing highly complex simulation data to establish correlations between structural/electronic features and enzymatic action and identify causal relationships. Therefore, we apply a protocol based on the usage of feature selection algorithms[9] to systematically identify the most appropriate subset of features through a reduction of the dimensionality of an initial and as extensive as possible dataset, followed by causality inference techniques[10] to investigate the causal relationships between the features included in the previously generated reduced subset.

The design of bioinspired strategies and biomimetic compounds requires an identification of the parts of the system that play a crucial role. Feature selection methods are routinely employed to analyze high dimensional data in diverse areas such as text mining, bioinformatics, combinatorial chemistry, or multivariate imaging. In particular, these tools are commonly used in bioinformatics and biochemical applications in order to improve the efficacy of feature discovery techniques in sequence, microarray, and mass spectra analyses, to facilitate the discovery of more selective drugs from large chemical libraries in virtual screening studies, and to improve predictability of QSAR methodologies.[11]

Causality inference techniques have been applied to areas as diverse as neuroscience, economy, genetics, philosophy, ecology, or biomedical informatics. These techniques have been particularly useful in order to determine relative risks and benefits of medical treatments, including adverse drug effects, or to elucidate relationships between environmental factors and risks of diseases.[12] We have applied such analysis techniques to assess the role of the environment during enzymatic catalysis of the DNA repair enzyme MutY[13] and to identify the molecular determinants responsible for the spectral shifts during the photoactivation of the visual pigment rhodopsin.[14]

In the latter case, we employed a strategy that combines the usage of a feature selection algorithm, to reduce the dimensionality of an initial time series dataset generated using MD simulations that includes all the geometrical features required to describe the chromophore structure and its mutual

orientation with respect to the active site residues, followed by the application of causality inference techniques, to learn the causal structure of this reduced subset, in order to identify the factors that modulate the spectral shifts between the early intermediates along the rhodopsin photocycle. Between the various feature selection algorithms that can be found in the literature, we used the so-called Correlation Based Feature Selection (CBFS) to perform the attribute selection step in order to filter irrelevant, redundant and noisy geometrical features. CBFS is a filter algorithm able to identify the best subset of features from a given dataset such that variables highly correlated with the target, yet being uncorrelated to each other, are selected.[15] It ranks feature subsets according to a correlation based evaluation function:

$$CBFS_S = \frac{k\langle r_{ft}\rangle}{\sqrt{k + k(k-1)\langle r_{ff}\rangle}} \qquad (2)$$

where $CBFS_S$ is the heuristic merit of the subset S containing k features, $\langle r_{ft}\rangle$ is the mean feature–target correlation ($f \in S$), and $\langle r_{ff}\rangle$ is the average feature–feature correlation. Therefore, the numerator of this equation provides an indication of the predictive ability of a given subset of features while the denominator gives a measure of the redundancy among that group of features.

We employed the PC-LiNGAM algorithm[16] to infer causal relationships between variables and to estimate the underlying causal structure of our models as this algorithm has been reported to work well to estimate dependency structures in networks with arbitrary (both Gaussian and non-Gaussian) distributions, and therefore the possible absence of normality in our distributions is not an issue. This algorithm determines whether or not a particular variable influences another giving rise to a directed acyclic graph (DAG) showing the different relationships between variables. Applying this protocol, we were able to describe all spectral shifts among the early intermediates of the rhodopsin photocycle with a set of five structural variables.[14]

For the case of HCAII, two features that we identified as crucial for catalysis are the CO_2 binding affinity and the pKa of the zinc-bound water that is fine-tuned *via* an extensive hydrogen-bond network.

2.4 *Choice of Biomimetic Template*

There are a huge number of choices for possible biomimetic templates ranging from minimal synthetic models of the active site[17] to the *de novo* design of entire enzymes.[18] Here, we take an intermediate approach by choosing a library of small protein domains (<50 amino acids) (Fig. 1) as generic scaffolds to incorporate specific enzymatic functions. These mini proteins have the advantage that they tend to self-assemble and in spite of their compact size they still offer ample opportunities for structural modifications and chemical tuning.

A biomimetic system for HCAII based on such a template has recently been synthetized.[19] This HCAII mimic is based on a three-helical bundle containing metal binding sites and is able to convert CO_2 to bicarbonate at elevated pH.

Using classical molecular dynamics simulations in combination with metadynamics, we have been able to identify low affinity CO_2 binding sites in the vicinity of the active site zinc ion. However, both QM/MM simulations and estimates of the acidity of the zinc-bound water/hydroxyl based on Poisson-Boltzmann calculations showed that the pKa of the catalytic water molecule is too high preventing the balanced co-existence of water and hydroxyl forms at neutral pH that is required for efficient catalysis. This suggests that the catalytic efficacy of the biomimetic system could be improved both by enhancing its ability to capture CO_2 and by tuning the microenvironment of the zinc sites in such a way as to achieve optimal acid/base properties of the metal-bound water molecule.

2.5 *Searching Chemical and Sequence Space*

After the relevant catalytic features have been identified, they can be used for an optimization of the biomimetic system. However, the search for optimal biomimetic compounds requires extensive sampling of chemical or sequence space. It is easy to see that a full exploration of *e.g.* all possible amino acid sequences involving even a very small number of residues is quite impossible, *e.g.* even for a short sequence of eight amino acids, there are $8^{20} = 1.15\times10^{18}$ possibilities, which clearly prevents any systematic enumeration! Therefore, this search problem has to be approached in a different way and it is suggestive to use again a biomimetic strategy, *i.e.* to see how Nature has solved this optimization problem and try to copy the natural selection process of biological evolution.

In the field of drug design, genetic algorithms (GA) have been fervently employed in *de novo* design of small molecules and drugs,[20] since searching can be made more efficient *via* machine learning tactics. Many algorithms use fragment-based and atom-based manipulations of small molecules to generate new candidate structures *via* a graph-based representation of molecules that is encoded into the chromosome, either directly or indirectly. On the other hand, these types of algorithms have rarely been applied in the context of electronic structure calculations and molecular simulations.

2.5.1 *SMOGA: A Genetic Algorithm for the Design of Biomimetics*

Genetic algorithms are meta-heuristic optimization algorithms, first loosely proposed by Fraser *et al.,*[21] and later further developed by J. Holland.[22] These algorithms belong to the group of evolutionary algorithms[23] and in a broader sense are a subgroup of artificial intelligence. In short, evolutionary algorithms attempt to copy Nature's method of optimization through mimicking biological evolution; namely the concepts of natural selection, mutation, and reproduction/recombination.

In a genetic algorithm a population of candidate solutions, called organisms, are sequentially 'evolved' towards better solutions. Each organism is defined by a chromosome: a piece of data that fully defines all the organism's characteristics in view of the optimization problem to be solved. The evolutionary process begins from a population of randomly sampled organisms, where each successive population created through the evolutionary process is called a generation. In each generation the fitness of all solutions are evaluated, through a physically applicable objective

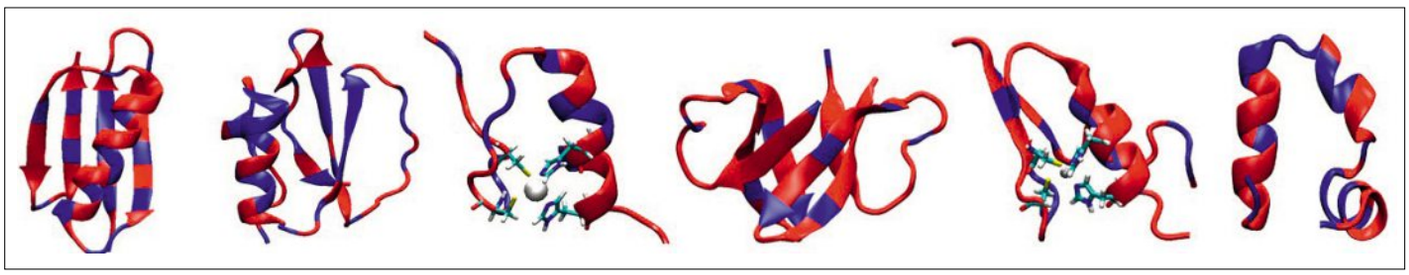

Fig. 1. Library of small protein domains for use as biomimetic scaffolds.

function. The more fit an organism is, the more likely this organism will be stochastically selected to combine with other organisms to create the children of the next generation. Each child resulting from this recombination process shares genetic information from both parents within its chromosome representation. Finally, this new generation of children is then allowed to stochastically mutate, before the children are redefined as parents to continue the evolutionary process.

Genetic algorithms are particularly good at finding near-optimal solutions in very large property spaces,[24] thus they pose an interesting application to the design of biomimetics. A single and multi-objective genetic algorithm software (SMOGA) has been developed in our group. This code provides a unique toolset for the rational design of proteins and small molecules based on defined objective functions. Fig. 2 provides a simplistic view of the algorithm procedure, with respect to protein sequence optimization.

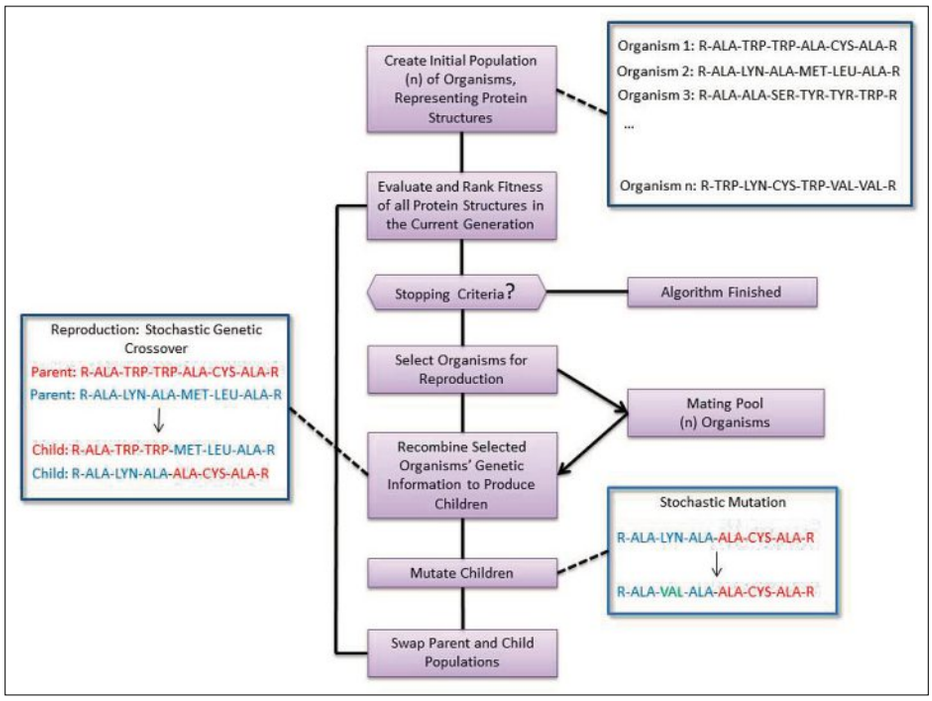

Fig. 2. Schematic of the procedure for sequence optimization *via* genetic algorithms.

2.5.2 Sequence Optimization of a Single α-Helix

As a first test application, we employed SMOGA for the sequence optimization of an ideal 20 amino acid long α-helix for which we optimized the central eight residues for a given property. In this case, the algorithm consisted of

(1) A well-defined chromosome constituted of 177 amino acid side chain rotamers based on the Richardson Rotamer Library.[25]

(2) A workable and modifiable objective function that evaluates the stability of each new GA mutation against the native structure (wild type) (Eqn. (3)).

$$f = \Delta\Delta E_{system}^{ref\rightarrow mut} = \Delta E_{system}^{mut} - \Delta E_{system}^{wildtype} = E_{system}^{mut} - E_{system}^{wildtype} + \sum_{i=1}^{n} E_{aa}^{wildtype} - \sum_{i=1}^{n} E_{aa}^{mut} \quad (3)$$

The objective function f in Eqn. (3) is defined as the difference in energy between a given sequence and the wild type structure plus the difference between the sum of the reference energies of the isolated amino acids for the wild type and those of the mutant sequence. This objective function is a simple measure for the intrinsic helical stability with respect to its amino acid constituents in various environments and is evaluated at the amberff10[26] level in combination with implicit solvent models of different dielectric constants.

We have used SMOGA to find near-optimal sequences for a single α-helix with the general sequence $A_6X_8A_6$, where X_8 indicates eight variable amino acids, flanked by six alanine residues on either side. As arbitrary reference structure, we use an ideal α-helix constituted of 20 ALA. The choice of α-helix was motivated by the fact that it is the most prevalent secondary structure element of proteins.

The GA search for the sequence with the best fitness is performed based on crossover (simulated binary crossover (SBX) technique with polynomial order 10 and single parameter genewise swap probability of 0.5),[27] mutation (polynomial) and selection (tournament selection with replacement), which are affected by crossover probability (c) mutation probability (m) and population size (pop). Numerous trials of SMOGA optimizations were run to determine the effects of these parameters on the search performance. Fig. 3 presents the best objective function within each generation, using m=0.04, c=0.7, pop=200 at a dielectric constant of 10. After 50 generations, more than 5000 individuals were analyzed, and the individual with the best fitness is $A_6W^{03}W^{03}M^{06}W^{02}W^{06}W^{03}W^{02}W^{01}A_6$ while after 100 generations the individual with the best value of the objective function is $A_6W^{03}W^{03}W^{03}W^{02}W^{03}W^{03}W^{04}W^{03}A_6$. Both high fitness sequences are entirely formed by different rotamers (indicated by the superscript) of TRP (a superimposition of the two helices is shown in Fig. 3). The GA is converging fast toward helices enriched in TRP due to the fact that TRP has a large hydrophobic side chain that can engage in strong vdW interactions, especially in hydrophobic environments mimicked by the low dielectric constant. In contrast, TRP sequences formed by a single rotamer only show less optimal values of the objective function. The best of these homo sequences is formed by the rotamer W04 with a fitness of –29.70 kcal/mol while the best SMOGA optimized sequence has a fitness of –37.60 kcal/mol showing that SMOGA is indeed proficient in finding nonobvious low energy sequences.

2.5.3 Optimization of the HCAII Biomimetic Helical Bundle

As a next application, we applied a GA to optimize the HCAII mimic for enhanced CO_2 binding and pKa tuning of the zinc-bound water molecule. The fitness for the former was evaluated from MMPBSA[28] calculations of the CO_2 binding optimizing residues around the previously identified binding pocket (Fig. 4). In the latter, the objective function was the pKa of the catalytic water molecule[29] estimated *via* an APBS protocol.[30]

Fig. 4 shows the putative CO_2 binding pocket of the HCAII biomimetic system and the residues that were optimized *via* the genetic algorithm. During the course of optimization, new sequences are discovered with slightly enhanced binding affinity for the substrate. However, the GA optimization of the acid/base properties of the zinc-bound water showed that even exploring a large sequence space, no configurations can be found with the desired properties indicating that this coiled-coil template might be too limited for the development of an optimal catalyst and that a slightly larger mini protein that also allows for variation of the second zinc coordina-

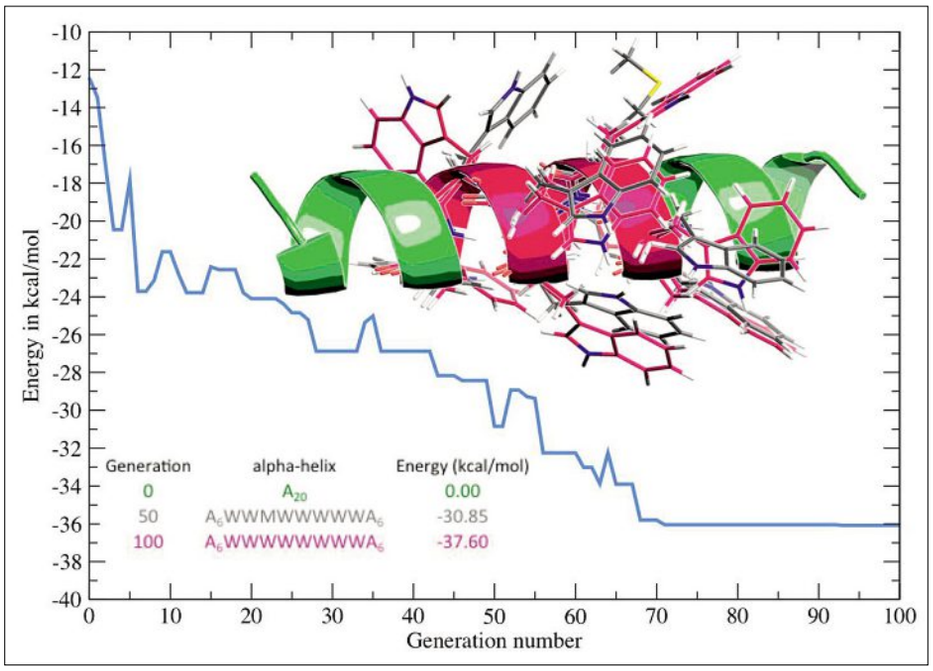

Fig. 3. Best values of the objective function within each generation, using mutation probability 0.04, crossover probability 0.7, population size 200 and at dielectric constant 10. The native structure is superimposed with the fittest structures after 50 and 100 generations, presented in green, grey and pink, respectively. For these structures, the rotamers sequence and respective objective values are also presented.

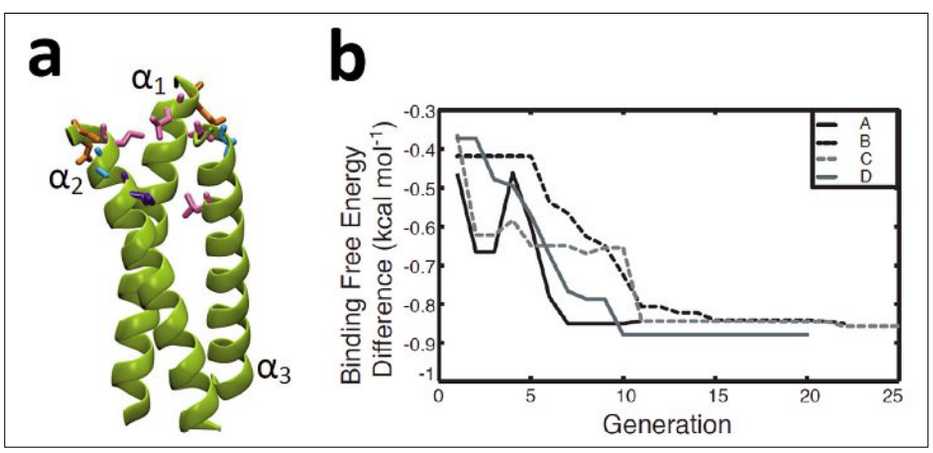

Fig. 4. (a) Residues around the CO_2 binding site that were subject to optimization. Leucine: pink, lysine: violet, alanine: blue, glutamate: red. (b) Different GA runs for the optimization of the binding affinity. A-D represent various GA runs which vary the total population number, mutation and crossover rate. For the GA run that converges the fastest (D), a population factor of 100, a mutation rate of 0.15, and a crossover rate of 6 were used.

tion sphere might be more successful for this task.

2.5.4 *Re-Engineering Protein Scaffolds for Catalysis: the B1 Domain of Streptococcal Protein G (GB1)*

For this reason, we have chosen the B1 domain of streptococcal protein G (GB1) which is composed of a four-stranded β-sheet and one α-helix[31] as a promising structural template for a biomimetic HCAII catalyst. The 56-residue B1 domain (Fig. 5) is of particular interest due to its relatively small size, unique structure and unusual thermostability.[32] Additionally, zinc[33] and iron[34] containing forms of the domain exist which renders this system a promising starting point for protein re-engineering towards specific functionalities. The tetrahedral zinc-GB1 variant might be a new alternative to mimic carbonic anhydrase activity and at the same time to perform mild (3 + 2) cycloaddition for unactivated nitriles due to the role of the zinc ion as Lewis acid. The latter is still an open problem, since a mild and general route for the cycloaddition of unactivated nitriles and azides is still missing.[35] With the aim of developing GB1-based metallo catalysts for various reactions, we are currently performing classical molecular dynamics (MD), quantum mechanics/molecular mechanics (QM/MM) simulations and model calculations using density functional methods. Indeed, our preliminary calculations indicate that activation energy barriers for zinc finger like models are lower than for the previously reported zinc salt catalyst, $ZnBr_2$[36].

2.5.5 *Design of Dye-Sensitized Solar Cells Using Biomimetic Porphyrin-Based Dyes*

Dye-sensitized solar cells (DSSCs)[37] have gained widespread attention in recent years because of their low production costs, ease of fabrication and tunable optical properties such as color and transparency. Many attempts have been made to optimize these devices towards their theoretical maximal performance. Among these, the modifications of the dye sensitizers play an important role in yielding higher efficiencies. Traditionally, Ru-based dyes have been used. However, despite many advantages, the difficulty of further improving the conversion efficiencies of these sensitizers are hampered by their low molar extinction coefficients (*e.g.* e <10000 $M^{-1}cm^{-1}$ for the metal to ligand charge transfer (MLCT) band of the black dye[38]) and the limited availability of precious ruthenium metal for practical applications.

Also in this case, it is of great appeal to use a biomimetic approach since essentially all natural systems have been optimized for the use of solar energy. The use of porphyrins as light harvesters in DSSCs is particularly attractive given their primary role in photosynthesis. In nature, porphyrin-based chromophores capture solar light and convert it into chemical energy. They are very good absorbers of electromagnetic radiation in the visible part of the spectrum. However, in nature there is no need for having high absorption efficiencies since plants only absorb as much light as they need for their daily consumption. For the purpose of solar cells, however, dyes have to harvest as much light as possible with a maximal overlap between their absorption spectrum and the solar spectrum.

In collaboration with the group of M. Graetzel at the EPFL, we are re-engineering the molecular structure of porphyrin-based dyes for optimal optical and redox properties. In this way, we were recently able to design new porphyrin sensitizers leading to DSSCs with a record efficiency of 13%.[39]

3. Summary and Conclusions

We have presented a general approach for the development of biomimetic systems from the identification of possible natural targets, their computational charac-

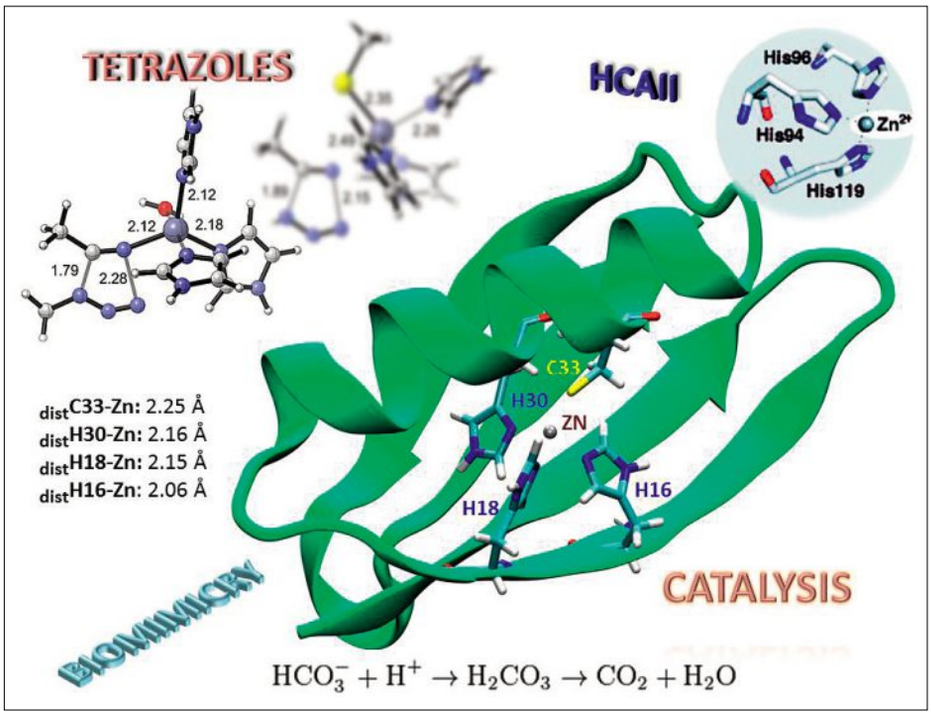

Fig. 5. A representative configuration from QM/MM simulation of the GB1 domain comprising a tetrahedral zinc binding site. Coordination distances for the cysteine and histidine residues and summary of the reactions relevant to this work are also given.

terization using QM/MM simulations, the systematic analysis of the simulation data using feature selection and causality inference algorithms from machine learning, to the possible choices of structural biomimetic templates and their optimization *via* bioinspired genetic algorithms. We have shown first applications of such a strategy for the design of a biomimetic catalyst for CO_2 fixation and for the development of biomimetic sensitizers in dye-sensitized solar cells.

Acknowledgements

Funding from the Swiss National Science Foundation *via* grant No. 200020-146645, and the interdisciplinary research programs NCCR MUST and MARVEL are gratefully acknowledged. We thank the IT domain (DIT) of EPFL, the Swiss National Computing Center (CSCS), and CADMOS project for computing resources.

Received: July 24, 2014

[1] E. C. Webb, 'Enzyme Nomenclature 1992', Academic Press, San Diego, California, **1992**.
[2] BRENDA. *http://www.brenda-enzymes.org/*.
[3] Metacyc. *http://www.metacyc.org/*.
[4] V. Hatzimanikatis, C. Li, J. A. Ionita, L. J. Broadbelt, *Curr. Opin. Struct. Biol.* **2004**, *14*, 300.
[5] K. C. Soh, L. Miskovic, V. Hatzimanikatis, *FEMS Yeast Res.* **2012**, *12*, 129.
[6] A. Laio, J. VandeVondele, U. Rothlisberger, *J. Chem. Phys*. **2002**, *116*, 6941.
[7] G. Ciccotti, R. Kapral, E. Vanden-Eijnden, *ChemPhysChem* **2005**, *6*, 1809.
[8] A. Laio, M. Parrinello, *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 12562.
[9] I. Guyon, A. Elisseeff, *J. Mach. Learn. Res*. **2003**, *3*, 1157.
[10] J. Pearl, 'Causality: Models, Reasoning, and Inference', Cambridge University Press, **2000**.
[11] a) M. A. Demel, A. G. K. Janecek, K. M. Thai, G. F. Ecker, W. N. Gansterer, *Curr. Comput-Aid. Drug* **2008**, *4*, 91; b) M. P. A. Sanders, R. McGuire, L. Roumen, I. J. P. de Esch, J. de Vlieg, J. P. G. Klomp, C. de Graaf, *Medchemcomm*. **2012**, *3*, 28; c) Y. Saeys, I. Inza, P. Larranaga, *Bioinformatics* **2007**, *23*, 2507.
[12] S. Kleinberg, G. Hripcsak, *J. Biomed. Inform.* **2011**, *44*, 1102.
[13] E. Brunk, J. S. Arey, U. Rothlisberger, *J. Am. Chem. Soc.* **2012**, *134*, 8608.
[14] P. Campomanes, M. Neri, B. A. Horta, U. F. Rohrig, S. Vanni, I. Tavernelli, U. Rothlisberger, *J. Am. Chem. Soc*. **2014**, *136*, 3842.
[15] M. A. Hall, Ph.D. Thesis University of Waikato, **1999**.
[16] A. H. P. Hoyer, R. Scheines, P. Spirtes, J. Ramsey, G. Lacerda, G. Shimizu, in 24th Annual Conference on Uncertainty in Artificial Intelligence, Helsinki, Finland, **2008**.
[17] L. Que, Jr., W. B. Tolman, *Nature* **2008**, *455*, 333.
[18] L. Jiang, E. A. Althoff, F. R. Clemente, L. Doyle, D. Rothlisberger, A. Zanghellini, J. L. Gallaher, J. L. Betker, F. Tanaka, C. F. Barbas, 3rd, D. Hilvert, K. N. Houk, B. L. Stoddard, D. Baker, *Science* **2008**, *319*, 1387.
[19] M. L. Zastrow, A. F. Peacock, J. A. Stuckey, V. L. Pecoraro, *Nat. Chem.* **2012**, *4*, 118.
[20] a) G. Schneider, U. Fechner, *Nat. Rev. Drug Discov.* **2005**, *4*, 649; b) S. Sengupta, S. Bandyopadhyay, *Ieee Acm. T. Comput. Bi.* **2012**, *9*, 1139.
[21] A. Fraser, D. Burnell, 'Computer Models in Genetics', McGraw-Hill, New York, **1970**.
[22] J. H. Holland, 'Adaptation in Natural and Artificial Systems', MIT Press, **1992**.
[23] D. Ashlock, 'Evolutionary Computation for Modelling and Optimization', Springer, **2006**.
[24] A. R. Simpson, G. C. Dandy, L. J. Murphy, *J. Water. Res. Pl-Asce.* **1994**, *120*, 423.
[25] J. M. Word, R. C. Bateman, Jr., B. K. Presley, S. C. Lovell, D. C. Richardson, *Protein Sci.* **2000**, *9*, 2251.
[26] Y. Duan, C. Wu, S. Chowdhury, M. C. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo, T. Lee, J. Caldwell, J. Wang, P. Kollman, *J. Comput. Chem.* **2003**, *24*, 1999.
[27] a) K. Deb, A. Kumar, *Complex Systems* **1995**, *9*, 431; b) K. Deb, R. B. Agarwal, *Complex Systems* **1995**, *9*, 115.
[28] a) M. Lepsik, Z. Kriz, Z. Havlas, *Proteins Struct. Funct. Bioinfom.* **2004**, 57, 279; b) W. Wang, P. A. Kollman, *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 14937.
[29] V. Zoete, O. Michielin, M. Karplus, *J. Comput. Aid. Mol. Des.* **2003**, *17*, 861.
[30] N. A. Baker, D. Sept, S. Joseph, M. J. Holst, J. A. McCammon, *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 10037.
[31] D. J. Sloan, H. W. Hellinga, *Protein Sci*. **1999**, *8*, 1643.
[32] a) S. M. Malakauskas, S. L. Mayo, *Nat. Struct. Biol.* **1998**, *5*, 470; b) M. Wunderlich, K. E. Max, Y. Roske, U. Mueller, U. Heinemann, F. X. Schmid, *J. Mol. Biol*. **2007**, *373*, 775.
[33] M. Klemba, K. H. Gardner, S. Marino, N. D. Clarke, L. Regan, *Nat. Struct. Biol.* **1995**, *2*, 368.
[34] E. Farinas, L. Regan, *Protein Sci.* **1998**, *7*, 1939.
[35] A. V. Gulevich, A. S. Dudnik, N. Chernyak, V. Gevorgyan, *Chem. Rev*. **2013**, *113*, 3084.
[36] F. Himo, Z. P. Demko, L. Noodleman, K. B. Sharpless, *J. Am. Chem. Soc*. **2003**, *125*, 9983.
[37] B. Oregan, M. Gratzel, *Nature* **1991**, *353*, 737.
[38] M. Kimura, H. Nomoto, N. Masaki, S. Mori, *Angew. Chem. Int. Ed.* **2012**, *51*, 4371.
[39] S. Mathew, A. Yella, P. Gao, R. Humphry-Baker, B. F. E. Curchod, N. Ashari-Astani, I. Tavernelli, U. Rothlisberger, M. K. Nazeeruddin, M. Gratzel, *Nat. Chem.* **2014**, *6*, 242.