DOI: 10.1002/cphc.201601034

Genetic-Algorithm-Based Optimization of a Peptidic Scaffold for Sequestration and Hydration of CO_2

Elizabeth Brunk,[a] Marta A. S. Perez,[b] Prashanth Athri,[c] and Ursula Rothlisberger*[b]

Biomimicry is a strategy that makes practical use of evolution to find efficient and sustainable ways to produce chemical compounds or engineer products. Exploring the natural machinery of enzymes for the production of desired compounds is a highly profitable investment, but the design of efficient biomimetic systems remains a considerable challenge. An ideal biomimetic system self-assembles in solution, binds a desired range of substrates and catalyzes reactions with turnover rates similar to the native system. To this end, tailoring catalytic functionality in engineered peptides generally requires site-directed mutagenesis or the insertion of additional amino acids, which entails an intensive search across chemical and sequence space. Here we discuss a novel strategy for the computational design of biomimetic compounds and processes that consists of a) characterization of the wild-type and biomimetic systems; b) identification of key descriptors for optimization; c) an efficient search through sequence and chemical space to tailor the catalytic capabilities of the biomimetic system. Through this proof-of-principle study, we are able to decisively understand and identify whether a given scaffold is useful, appropriate and tailorable for a given, desired task.

The ability to determine the structure and oligomerization characteristics of peptidic scaffolds has laid the groundwork for the design of functional (i.e., catalytic) peptidic scaffolds. Numerous studies[1–5] have focused on de novo design of coiled coils that mimic metalloenzymes within a peptidic scaffold.[6] The coiled coil scaffold is less complex in structure and is mostly a first-coordination-sphere-only model, which makes it easier to predict structure–fold–reactivity relationships. Yet, the de novo design of peptide scaffolds to mimic reactions of native enzymes is a challenging task as it requires the precise binding and orienting of a specified substrate to efficiently catalyze the desired reaction. Furthermore, identifying the best sequence for a suitable biomimetic scaffolds remains a notable challenge and is the subject of recent papers and review articles.[7,8]

From a computational perspective, two limiting factors affect the optimization of these systems in silico: 1) searching effectively through chemical phase space to identify optimal sequence mutations and 2) performing computations that are capable of accurately predicting structural rearrangements and catalysis. Despite their simplified frameworks, the search for new sequences with unique structures or improved functions within these systems requires an exhaustive search of chemical and sequence space. Mutating even a small range of residues in a peptidic scaffold, (e.g., eight amino acids), requires a chemical search of the order of $20^8=2.56E^{10}$ possibilities. Subsequently, understanding how a system responds to such changes in the primary amino acid sequence entails the assessment of dynamic rearrangements of residues[9] and reordering of ligand and/or metal binding sites.[10] To this degree, genetic algorithms (GAs) provide a search heuristic inspired by natural evolution, which involve the phenomenon of genetic mutation, natural selection and inheritance. Recent advances in computational power together with a free-energy-based GA enable, to the best of our knowledge, the first application of a GA to optimize the structure–activity relationships in a peptidic scaffold in silico. As a proof-of-principle, we apply a GA to enhance the catalytic efficiency of a synthetic protein scaffold, a three-stranded coiled coil (3SCC), which has recently been reported to mimic human carbonic anhydrase (HCA, Figure 1).[3]

Understanding whether changes in the amino acid sequence enhance or diminish catalytic capabilities and structural stability of a system can be assessed using long-time-scale classical molecular dynamics (MD),[11] electronic structure calculations and quantum mechanical/molecular mechanical (QM/MM) molecular dynamics.[12] We first set out to determine whether the crystal structure of 3SCC is stable in solution by performing classical MD in an explicit solvent environment. The design of 3SCC consists of three alpha helices wrapped around each other, and incorporates a first coordination sphere around a pseudotetrahedral zinc ion (Zn^{II}) with three coordinated imidazoles and one water (H_2O/^-OH) molecule. Within 16 nanoseconds (ns) of a 100 ns trajectory, we find that the structure of 3SCC deviates from the crystallographic structure (average root mean square deviation of 3.5 Å). The total helical content, which is originally 83 %, drops to 51 % by the end of the trajectory. These findings are consistent with the high crystallographic B-factors for the 4–5 amino acids at the C-terminal end in two out of three helical peptides (Figure S1).

[a] E. Brunk
Department of Bioengineering,
University of California
San Diego, CA 92093 (USA)

[b] M. A. S. Perez, U. Rothlisberger
Laboratory of Computational Chemistry and Biochemistry,
Ecole Polytechnique Fédérale de Lausanne (EPFL)
1015 Lausanne (Switzerland)
Fax: (+41) 21 693 0320
E-mail: ursula.roethlisberger@epfl.ch

[c] P. Athri
Department of Computer Science,
Amrita School of Engineering
560035 Bangalore (India)

Supporting Information for this article can be found under:
http://dx.doi.org/10.1002/cphc.201601034.

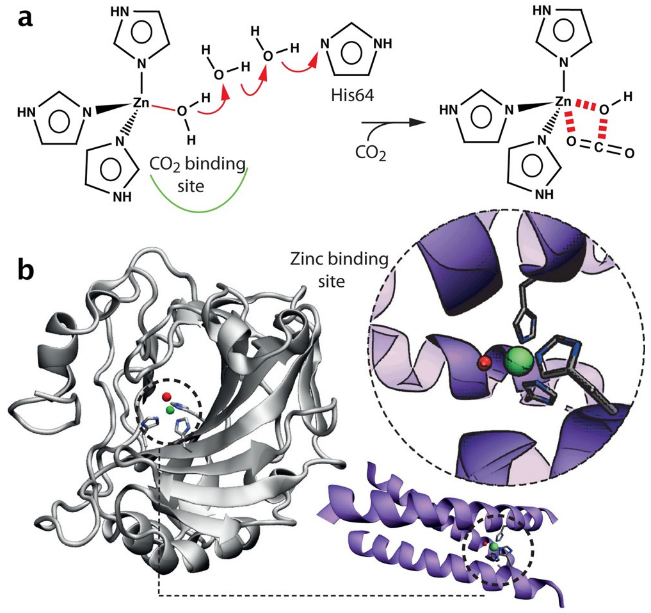

Figure 1. a) First step of the enzymatic reaction, the intramolecular proton transfer, of human carbonic anhydrase (HCA) for the hydration of CO_2. Arrows indicate proton movement. b) Comparison of the Zn3NO catalytic sites between HCA and a three-stranded coiled coil (3SCC), which is capable of catalyzing the same reaction at high pH (Hg not shown in Figure).

We were interested in understanding whether these structural changes affect the ability of the mimic to establish a hydrogen-bonded network of water molecules. Such networks are required to mediate the fast transfer of a proton from the zinc-bound water molecule to bulk solvent.[13] We probed the solvation characteristics of the 3SCC by evaluating the distribution of solvent molecules over 16 ns. We find that, for 80% of the time, one to two solvent molecules are found within the 3.6 Å vicinity of the metal binding site and are relatively short-lived, remaining in the active site for only 30 ps. One of the main factors limiting the formation of a stable hydrogen bonding network of water molecules is the lack of full second coordination sphere residues in 3SCC. Due to design of 3SCC and its minimal scaffold, the zinc-bound hydroxide is encapsulated in a pocket surrounded by leucine residues—not in a position to hydrogen-bond to any nearby polar amino acids.

Do the structural and solvation differences between wild-type and biomimetic systems lead to a diminished binding affinity to CO_2 in the scaffold? To address this question, we localized putative CO_2 binding sites and computed the free energy map for CO_2 binding using metadynamics[14] within the framework of classical MD simulations.[15,16] We detected two energy minima (M1 and M2) located 10 and 5 Å from the zinc binding site, respectively. M1, located near leucine residues at the 4rth position, is reminiscent of the second CO_2 binding site in CA,[17] a suggested CO_2 storage site (Figure S2). M2 is denoted as the main binding site in the mimic, similar to that of HCA[17,18] The binding of CO_2 at these respective sites is estimated to be of the order of -2.0 ± 0.8 kcal mol^{-1}, consistent with previously reported binding affinity in CA in both computational (-3.37 ± 1.1 kcal mol^{-1})[17] and experimental studies at a pH of 7.5 (-2.2 kcal mol^{-1} or 25 mmol).[17,18] Therefore, despite structural differences, we find that 3SCC binds CO_2 with similar affinity as wild-type.

Given the structural limitations of this particular scaffold, does 3SCC have the capacity to catalyze CO_2 hydration? Starting from the classically equilibrated structure, we studied the putative hydration reaction in 3SCC using QM/MM Car–Parrinello molecular dynamics (CPMD) simulations. The barrier for HCO_3^- formation mediated via the deprotonated form of the zinc bound water is found to be 7.2 ± 0.26 kcal mol^{-1} (Figure 2), whereas, in HCA, this process happens spontaneously (on a picosecond timescale). Throughout the simulation, the stability of the zinc-bound hydroxyl is, in part, dependent on the second water molecule, which occupies a similar position as the so-called "deep water molecule" in HCA (see Figure 3 a). Starting at a C· · ·Ow bond distance of 1.7 Å, the CO_2 molecule is no longer linear in structure and an oxygen atom from CO_2 donates electron density to the zinc ion, adopting bonding distances of 2.4 to 3 Å (see Figure 3 b). Release of the constraint at 1.5 Å spontaneously generates HCO_3^- and this moiety rotates such that it is bound to the zinc ion by one of the oxygen atoms of the initial CO_2 molecule. It is with this oxygen atom that the HCO_3^- molecule is finally bound to the metal center (see Figure 3 c). Upon formation of HCO_3^-, the Lipscomb product is the most stable intermediate configuration that forms during catalysis, similar to HCA.

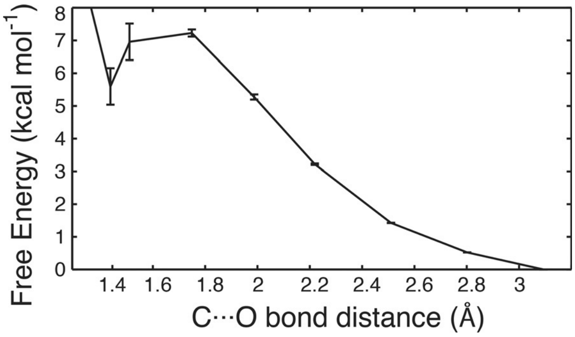

Figure 2. Free energy profile for the formation ofHCO_3^- in 3SCC. The reaction coordinate was chosen to be the distance between the carbon atom of CO_2 and the oxygen atom of the zinc-bound hydroxide moiety.

However, based on our findings, 3SCC lacks a stable hydrogen-bonding network of solvent molecules and is not expected to efficiently catalyze the initial proton-transfer step under ambient conditions (neutral pH and room temperature). This observation is consistent with the substantial decrease in the catalytic rate of the 3SCC at pH 7, which is reported to be 1.38 $\text{M}^{-1}\text{s}^{-1}$,[3] compared to the wild-type enzyme ($10^6\,\text{M}^{-1}\text{s}^{-1}$). It is widely recognized that, in HCA, the orchestration of the first *and* second coordination sphere around the zinc ion is one of the main strategies of the enzyme to increase catalytic efficiency. These interactions play a significant role in stabilizing the displacement of HCO_3^- from the metal site and the "hand off" to the next water molecule.[19] To evaluate the over-

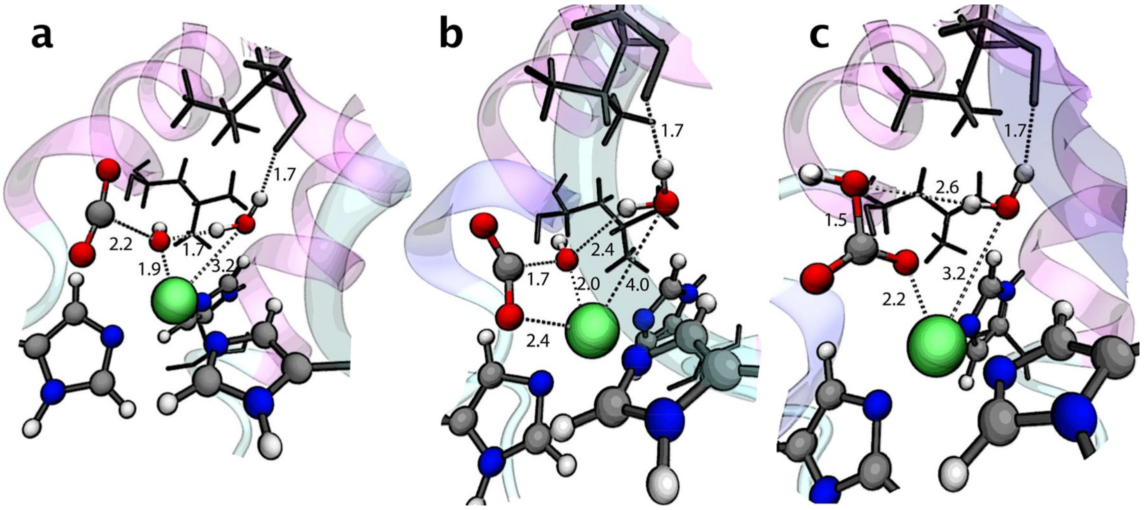

Figure 3. Nucleophilic attack of the zinc-bound hydroxide on CO_2. a) Nucleophilic attack of the zinc-bound hydroxide moiety on CO_2. b) Transition state for the conversion step and. c) Final (Lipscomb) state with HCO_3^- bound to the metal site.

all potential of this particular scaffold, we were interested in understanding whether this minimal scaffold could be re-engineered to optimize interactions in neighboring helices to further enhance its capacity to sequester and hydrolyze CO_2 on one hand, while on the other hand, tune the acidity of the zinc bound water molecule closer to the wild type.

To address these questions, we developed a genetic-algorithm-based sequence exploration tool with an appropriate fitness function based on efficient free-energy calculations (see the Supporting Information) to assess changes due to mutation of specific residues in the vicinity of the metal binding site (Figure 4a, Figure S3). The knowledge gained by the molecular simulations was used to redesign the three-stranded coiled coil mimic, by which two routes were taken for optimization: 1) increasing CO_2 binding affinity and 2) tuning the p*K*a of the Zn–OH_2 metal site. While the first route increases the ability of the mimic to bind CO_2 in the active site, the latter optimizes the rate-limiting step of catalysis, the initial deprotonation of the zinc-bound water molecule. The fittest individual for CO_2 binding, found after roughly 10 generations, lowers the $\Delta\Delta G_{bind}$ by approximately 0.9 kcal mol^{-1} (Figure 4b, Table 1, Supporting Information). In this variant, many of the nine amino acids from the original helical bundle are replaced by hydrophobic residues such as Val, Ile, and Leu, which mimics the nonpolar environment of HCA. For tuning the p*K*a, the fittest individual converges after 25–30 generations and shifts the p*K*a from 10.4 (original system) to 8.4 (variant mimic) (see Table 2). The experimental p*K*as of HCA and 3SCC are around 6.8[20] and 9.0[3] respectively (e.g., as estimated for esterase activity).

Table 1. CO_2 binding affinity optimization using an evolutionary algorithm. Nine amino acids at the C-terminal end of the protein were varied to influence the binding free energy of CO_2. The positions and amino acids of the original (3SCC) system are compared with the mutant that has an increased binding affinity to CO_2.

	$\alpha1$	$\alpha2$	$\alpha3$	ΔG [kcal mol^{-1}]
Position in helix	25 26	19 22 25 26 27	26 27	
3SCC	A L	L K A L E	L E	−2.0
Mutant	V I	H N I H L	G T	−2.9

Table 2. p*K*a tuning of the Zn-OH2 moiety using an evolutionary algorithm.

	$\alpha1$	$\alpha2$	$\alpha3$	p*K*a shift
Position in helix	25 26	19 22 25 26 27	26 27	
3SCC	A L	L K A L E	L E	10.4
Mutant	K K	K K K K K	K E	8.4

In the case of p*K*a tuning, the GA finds an amino acid sequence consisting of purely lysine residues, which clearly does not lead to a stable structure (we estimate a destabilization energy of 162.6 kcal mol^{-1} relative to the original system; Supporting Information). Similarly, the mutant with a slightly enhanced binding affinity to CO_2 is only realizable at the price of a destabilized structure (13.7 kcal mol^{-1} less stable than the original mimic structure). This presents an important conceptual conclusion: for this particular system, we inevitably reach a limit in terms of capacity for variation in sequence, which is not too surprising, considering the fact that it is mostly a first-coordination sphere model only. While the fitness functions appear adequate for optimizing and converging sequence space, given a biological objective, we find that the scaffold itself is too limited (in terms of size and interaction space) for sufficient optimization. Even allowing for all possible sequence variations does not generate the properties needed for improving catalysis to rival the natural system. Therefore, our findings suggest that sufficient optimization of biomimetic scaffolds require the system to be larger in size and allow at least for second sphere interactions.

Many of the GAs used for small molecule discovery apply fitness functions that are straightforward and easy to compute[21,22] and relatively few attempts exist in GA protein

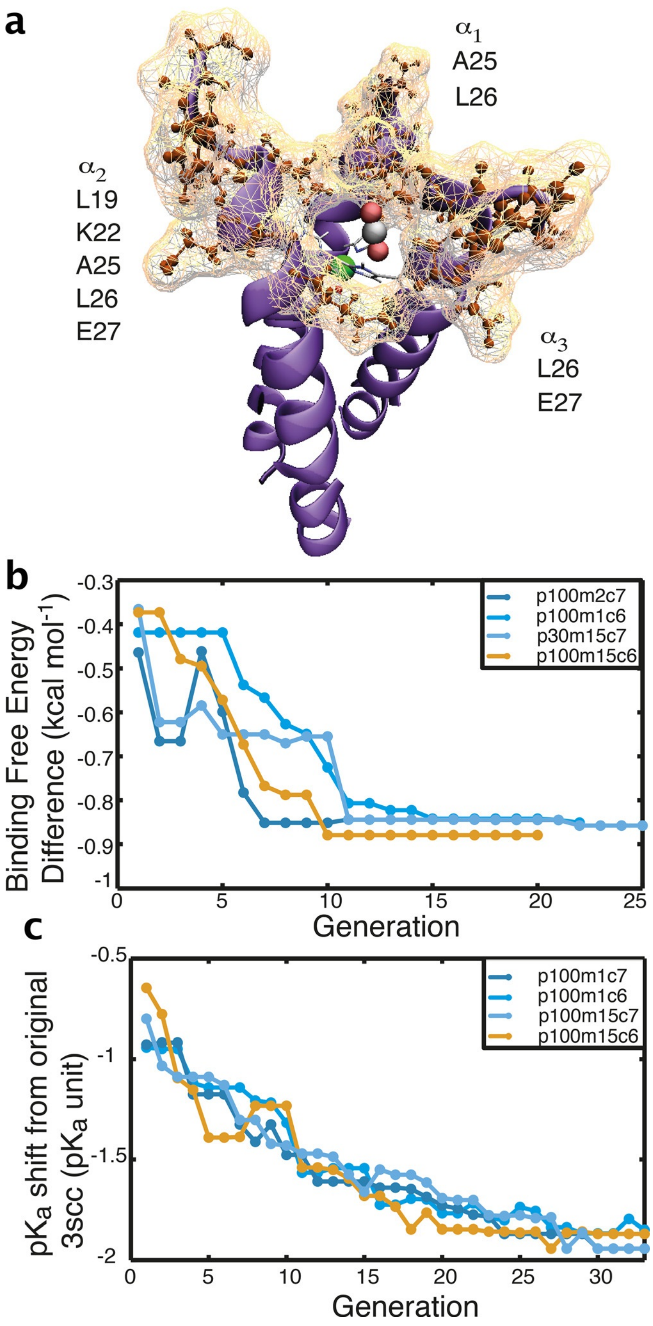

Figure 4. Heuristic-based optimization of CO_2 binding in 3SCC. Nine amino acids at the C-terminal end of the protein (a) were mutated to influence the binding free energy of CO_2 in (b). The evolution of the fitness of the best individual, found after the 10 generations, is shown by the orange curve. The parameters for optimization are given in the legend (see the Supporting Information for more details).

design based on simple scoring functions.[23] While optimizing amino acid sequence is a natural goal for protein engineering, (as protein sequence determines fold and function), only recently have computational resources been powerful enough to integrate molecular simulation tools (minimization and molecular mechanics) into the selection and optimization process. Molecular simulations and electronic structure calculations are computationally much more demanding and are typically thought to be less amenable to pairing with GAs, as balancing the speed of execution with the correlation to the goal of the design can be a non-trivial challenge. This capability now allows one to optimize directly for function while earlier attempts have all been based on selecting systems with maximal stability.

Using QM/MM simulations to compare wild-type and biomimetic systems has been a widely used approach in protein engineering.[24] Here we have used this approach to identify relevant descriptors that can quantify the differences in catalytic activity between the two systems. Finding the biological determinants that enable a highly optimized system to function is not trivial, even in the wild-type system. The findings from initial studies on Carbonic Anhydrase were used to define appropriate fitness functions for the subsequent GA optimization. Optimizing the sequence of the coiled coil in this way demonstrated that, even by allowing for all possible amino acid substitutions, changes in sequence only slightly enhance the catalytic activity of the biomimetic system. This is an important outcome as it shows that, for this particular template (3-helical bundle), optimization is intrinsically limited by size and is unable to achieve a higher efficiency to mimic HCA.

In summary, we present one of the first applications of a free-energy-based GA to the design and functional optimization of proteins and small peptides. Applying the GA-based protocol introduced here, we are able to decisively understand and identify whether a given biomimetic scaffold is useful, appropriate and tailorable for a given, desired task and, given a suitable scaffold, it can predict sequences that optimize catalytic properties. This proof-of-principle study confirms that our GA-based approach is capable of achieving fast optimizations even in large spaces (e.g. simultaneously optimizing nine amino acids corresponding to a search space of $19^9 = 3.23E^{11}$ possibilities). The GA converges within 10–20 generations and is likely to be of general interest for exploring possible sequence landscapes of different proteins and biomimetic templates. Most importantly, this approach can complement experimental analyses to predict which templates would be appropriate as starting points for the design of functional analogues.

Computational Details

The starting structure for the QM/MM molecular dynamics simulations considered two bound water molecules in the vicinity of the zinc binding site in an intact scaffold, to probe the maximum capacity of the mimic for catalysis. The reaction was examined by studying the formation of HCO_3^- with a single reaction coordinate, the C–*Ow* bond length, which signifies the nucleophilic attack of the zinc-bound hydroxide on CO_2. Starting from the equilibrium value of 3.2 ± 0.2 Å, the reaction coordinate was systematically decreased to 1.5 Å. The barrier for HCO_3^- formation is computed via thermodynamic integration (TI).[25]

For the genetic algorithm optimization, two primary objectives were considered: 1) increasing the binding affinity of the zinc-binding pocket for CO_2 and 2) shifting the p*Ka* for efficient deprotonation/reaquation at neutral pH. Optimizing considered mutation of nine amino acids at the C-terminal end of the protein (see Figure 4a), chosen on the basis of a clustering analysis of different

CO_2 binding positions. The fitness function for calculating the binding free energy of CO_2 consists of several steps: 1) the GA randomly chooses a new set of individuals; 2) selected amino acids were mutated to give a new set of individuals; 3) each mutant structure was minimized to remove high-energy interatomic interactions; 4) the binding free energy ($\Delta\Delta G_{bind}$) for each mutant was calculated. Several tests were performed to vary the population, mutation and crossover rate, to obtain the fittest possible individual (see Figure 4 b). For tuning the p*K*a, a similar procedure was taken. The p*K*a of the zinc-bound water moiety was calculated using the adaptive Poisson–Boltzmann solver (APBS) method.[26] Considering that each fitness evaluation takes three minutes (see Table S3) and that the fittest individual is found after ten generations, the total amount of time needed for one entire GA optimization is approximately two days: 10 generations × 100 individuals per generation × 3 minutes = 3000 minutes = 50 hours (machine characteristics: 8 cores, Xeon E5-2650v2, 64GBRAM).

Acknowledgements

The authors gratefully acknowledge support from the Swiss National Science Foundation (grant 200020-146645 and the NCCR MARVEL) as well as the IT domain (DIT) of EPFL, the Swiss National Computing Center (CSCS), and BlueGene for computing resources.

Keywords: biomimetic systems · catalytic properties · engineered peptides · enzymes · genetic algorithms

[1] G. R. Dieckmann, D. K. McRorie, J. D. Lear, K. A. Sharp, W. F. DeGrado, V. L. Pecoraro, *J. Mol. Biol.* **1998**, *280*, 897–912.

[2] J. Kaplan, W. F. DeGrado, *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 11566–11570.

[3] M. L. Zastrow, A. F. A. Peacock, J. A. Stuckey, V. L. Pecoraro, *Nat. Chem.* **2012**, *4*, 118–123.

[4] M. L. Zastrow, V. L. Pecoraro, *J. Am. Chem. Soc.* **2013**, *135*, 5895–5903.

[5] M. Tegoni, F. Yu, M. Bersellini, J. E. Penner-Hahn, V. L. Pecoraro, *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 21234–21239.

[6] F. Yu, V. M. Cangelosi, M. L. Zastrow, M. Tegoni, J. S. Plegaria, A. G. Tebo, C. S. Mocny, L. Ruckthong, H. Qayyum, V. L. Pecoraro, *Chem. Rev.* **2014**, *114*, 3495–3578.

[7] L. Wang, E. A. Althoff, J. Bolduc, L. Jiang, J. Moody, J. K. Lassila, L. Giger, D. Hilvert, B. Stoddard, D. Baker, *J. Mol. Biol.* **2012**, *415*, 615–625.

[8] D. Baker, *Protein Sci.* **2010**, *19*, 1817–1819.

[9] R. Elber, M. Karplus, *Science* **1987**, *235*, 318–321.

[10] J. Z. Ruscio, J. E. Kohn, K. A. Ball, T. Head-Gordon, *J. Am. Chem. Soc.* **2009**, *131*, 14111–14115.

[11] D. E. Shaw, R. O. Dror, J. K. Salmon, J. P. Grossman, K. M. Mackenzie, J. A. Bank, C. Young, M. M. Deneroff, B. Batson, K. J. Bowers, et al. in *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis*, ACM, **2009**, p. 65.

[12] E. Brunk, U. Roethlisberger, *Chem. Rev.* **2015**, 115, 6217–6263.

[13] K. K. Kannan, B. Notstrand, K. Fridborg, S. Lövgren, A. Ohlsson, M. Petef, *Proc. Natl. Acad. Sci. USA* **1975**, *72*, 51–55.

[14] A. Laio, M. Parrinello, *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 12562–12566.

[15] D. A. Case, T. A. Darden, T. E. Cheatham, C. L. Simmerling, J. Wang, R. E. Duke, R. Luo, M. Crowley, R. C. Walker, W. Zhang, et al., *J. Comput. Chem.*, **2008**, 26, 1668–1688. The Amber biomolecular simulation programs.

[16] M. Bonomi, D. Branduardi, G. Bussi, C. Camilloni, D. Provasi, P. Raiteri, D. Donadio, F. Marinelli, F. Pietrucci, R. A. Broglia, M. Parrinello, *Comput. Phys. Commun.* **2009**, *180*, 1961–1972.

[17] K. M. Merz Jr., *J. Am. Chem. Soc.* **1991**, *113*, 406–411.

[18] J. Y. Liang, W. N. Lipscomb, *Proc. Natl. Acad. Sci. USA* **1990**, *87*, 3675–3679.

[19] D. N. Silverman, R. McKenna, *Acc. Chem. Res.* **2007**, *40*, 669–675.

[20] J. A. Verpoorte, S. Mehta, J. T. Edsall, *J. Biol. Chem.* **1967**, *242*, 4221–4229.

[21] G. Schneider, U. Fechner, *Nat. Rev. Drug Discovery* **2005**, *4*, 649–663.

[22] S. Sengupta, S. Bandyopadhyay, *IEEE/ACM Trans. Comput. Biol. Bioinform.* n.d., 9, 1139–1154.

[23] S. Traoré, K. E. Roberts, D. Allouche, B. R. Donald, I. André, T. Schiex, S. Barbe, *J. Comput. Chem.* **2016**, *37*, 1048–1058.

[24] U. Rothlisberger, P. Carloni, *Int. J. Quantum Chem.* **1999**, *73*, 209–218.

[25] M. Sprik, G. Ciccotti, *J. Chem. Phys.* **1998**, *109*, 7737–7744.

[26] N. A. Baker, D. Sept, S. Joseph, M. J. Holst, J. A. McCammon, *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 10037–10041.

Manuscript received: September 22, 2016
Accepted Article published: October 5, 2016
Final Article published: October 21, 2016