A Multi-Objective Evolutionary Approach to Peptide Structure Redesign and Stabilization

Tim Hohm
Stiftung caesar
Research Group Functional Peptides
Ludwig-Erhard-Allee 2
53175 Bonn, Germany
Tim.Hohm@caesar.de

Daniel Hoffmann
Bingen University of Applied Sciences
Department II – Bioinformatics
Berlinstr. 109
55411 Bingen, Germany
Daniel.Hoffmann@fh-bingen.de

ABSTRACT

The prediction of the native structures of proteins, the so-called protein folding problem, is a NP hard multi-minima optimization problem for which to date no routine solutions exist. Using an evolutionary approach we have addressed a problem that is related to protein folding though much simpler: the computational improvement of small proteins or peptides with respect to stability and biological function. The solution of this problem is relevant for the life sciences, e.g. because it would help to optimize peptide drugs.

In a first experiment we used the proposed algorithm to stabilize a previously destabilized mutant of the otherwise stable folding Villin Headpiece. The algorithm generated amongst others a sequence that reverted the destabilizing mutation and introduced a second mutation. In terms of the used model this second mutation resulted in a more stable peptide than the original Villin Headpiece.

Categories and Subject Descriptors

J.3 [**Computer Applications**]: Life and Medical Sciences—*Biology and Genetics*; G.3 [**Mathematics of Computing**]: Probability and Statistics—*Probabilistic Algorithms*

General Terms

Algorithms

Keywords

Evolutionary algorithm, multi-objective optimization, peptide design

1. INTRODUCTION

Currently several peptides are used as drugs. Prominent examples are the well-known peptide insulin, the major drug against diabetes, or, more recently, the anti-HIV peptide T20 [16]. Nevertheless, in comparison to conventional small-molecule drugs, peptides are not ideal as drugs; for instance, they are relatively large and thus have difficulties to cross biological barriers [35], and they are often flexible or conformationally unstable which leads to high entropy losses on binding to their target molecules and thus to low affinities to these molecules [31].

Hence, methods that optimize the stability of the "active" peptide conformation while limiting peptide size are valuable. Today, such optimizations are usually achieved by labour-intensive experimental wet-lab methods, such as site-directed mutagenesis, phage display and others [33]. An *in silico* method that automatically optimizes peptide sequences would be an attractive supplement to these methods. Unfortunately, it is difficult to accurately predict peptide properties such as stability *ab initio* – such a prediction in a way implies the solution of the protein folding problem. However, if a reliable starting point is available, e.g. an experimentally determined peptide structure, point mutations could be applied in a stepwise fashion and their effects predicted more reliably. Since already small changes to the sequence have the potential of changing peptide properties significantly [13, 9, 21] even such a conservative *in silico* approach could be helpful. Similar approaches are in use for different applications in drug design [14, 15].

Here, we propose a multi-objective evolutionary algorithm that stabilizes a peptide in an active conformation, or more precisely, an algorithm that changes a peptide sequence such that the key parts of the peptide responsible for biological activity have a higher propensity of being in the active conformation.

Such an approach ranges in the field of *in silico* protein structure prediction, protein folding and drug design. Especially in the recent years there has been progress regarding the used methods [10, 34, 28, 24] and models [12, 30, 39, 22]. The achievements of these methods are documented in [20, 27, 29, 19].

We present first results obtained for the stabilization of the unstable mutant F18K of the Villin headpiece, a stably folding 36mer [25] for which an experimentally determined structure is available. The instability of mutant F18K – it carries a mutation at residue 18 from phenylalanine (F) to lysine (K) – has been measured experimentally [13]. During the redesign the algorithm was not only able to re-stabilize the mutant but also to predict stability enhancing mutations. Amongst some multi site mutants, which were mu-

GECCO'05, June 25–29, 2005, Washington, DC, USA.
Copyright 2005 ACM 1-59593-010-8/05/0006 ...$5.00.

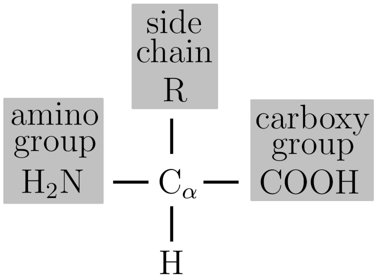

Figure 1: General chemical structure of an amino acid with central C_α atom, amino group, carboxy group and side chain.

tated at up to nine different sites, there was one mutant that reversed the destabilizing mutation and further improved the stability by mutating a second amino acid, namely a mutation of residue 34 from the native glycine to leucine, for short: G34L.

2. BIOLOGICAL BACKGROUND

We first give a few chemical basics on peptides to help the understanding of some features of the algorithm and of the methods used by the algorithm. Peptides are chains of amino acids. Amino acids consist of an amino group, a carboxylate group and a sidechain which are all linked to a central carbon atom (C_α) (Fig. 1). Peptides are formed by covalent "peptide" bonds between the carboxylate group of one amino acid and and the amino group of another amino acid. Fig. 2 shows a di-peptide, the smallest possible peptide. The main chain or "backbone" made up of successive C_α–C–N–C_α units of the peptide can be prolonged arbitrarily by adding more amino acids in this way at either end of the peptide. The sidechains are the variable regions of amino acids that can be positively or negatively charged, hydrophobic or polar, etc. Peptides and proteins are made up from a repertoire of the 20 most common amino acids, and the frequency and sequence of their respective sidechains determine the physico-chemical properties and biological function of the peptides [7].

Many proteins or larger peptides adopt a specific well-defined stable conformation or "native fold" which is solely determined by the respective sequence of amino acids [1]. According to equilibrium statistical mechanics this native fold corresponds to the minimum of the free energy of the peptide in its aqueous environment and thus to the most probable conformation. Minimizing the free energy of a conformation thus means maximizing the probability of the peptide of adopting this conformation. In our approach we equate this probability with the observed relative frequency of this conformation in simulations of the peptide dynamics.

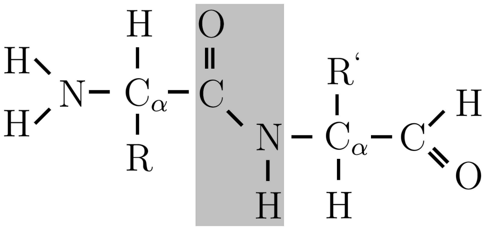

Figure 2: Two amino acids connected via a peptide bond – a di-peptide. By adding more amino acids in this way, peptides of arbitrary length can be generated.

3. APPROACH

We present an evolutionary algorithm that changes given peptide sequences towards sequences with increased propensity for a specific conformation.

Starting with a number of copies of a sequence with known structure (individuals) the algorithm carries out an iterative (generation based) optimization process that in each step slightly alters the individuals. Following the argument in the introduction we have restricted sequence changes in each generation to single point mutations. In this way we are able to predict rather reliably the effects of changes while we still have the potential of improving the sequence significantly. Larger changes such as crossover would in general cause global rearrangements that are much harder to predict.

Our algorithm preferentially changes the sequence in a way that promises to lead to improved conformational stability. The site and type of mutation are determined by a method that is inspired by *in vitro* alanine scan. However, instead of measuring the effect of a mutation experimentally we estimate the changes of free energy by approximations described in the Methods section.

Typically, functional peptides carry a few key residues that are essential for their biological action, often arranged in surface exposed loops. Therefore to preserve the bioactivity of the input peptide the algorithm allows to exclude a user defined loop of key residues from the mutation process.

The peptides were optimized with respect to at least two different attributes, namely the conformational stability of the peptide and the deviation of accessibility of the key residue loop. Quantifiers for an aggregated optimization approach are unknown. Hence, we have chosen a multi-objective EA (MOEA) with two fitness criteria: (1) structural stability of the peptide, and (2) root mean square deviation (RMSD) of the accessibility, whereas a third criterion, RMSD of key residues, is used to influence the mutation process. The fitness criteria are evaluated based on data from molecular dynamics (MD) trajectories of the peptides in aqueous solution (for details on MD see Methods section).

After having evaluated the offspring individuals a selection process takes place. The next generation is chosen from the set of offspring individuals and the former generation using tournament selection [2]. To identify the winning individual of a tournament the dominance-relation (see Def. 1) is used. The two fitness criteria stability and accessibility RMSD are regarded. If the individuals participating in a tournament are equal or incomparable in terms of the dominance-relation the tournament winner is randomly chosen. Apart from the individuals chosen via tournaments some individuals are chosen using elitism [11]. Therefore an archive of non-dominated (see Def. 2) individuals is kept which is updated after each generation. From this archive a fixed number of individuals is randomly chosen and inserted into the newly formed generation.

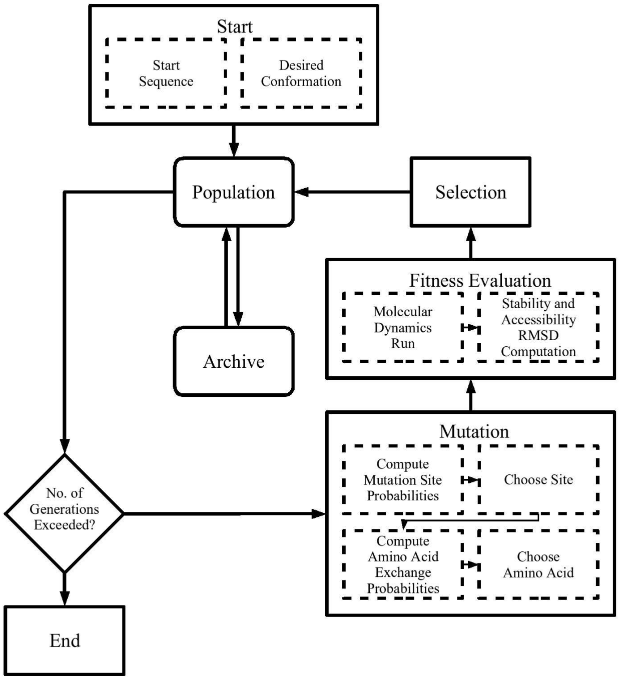
Start
Start Sequence
Desired Conformation
Population
Selection
Archive
Fitness Evaluation
Molecular Dynamics Run
Stability and Accessibility RMSD Computation
Mutation
Compute Mutation Site Probabilities
Choose Site
Compute Amino Acid Exchange Probabilities
Choose Amino Acid
No. of Generations Exceeded?
End

Figure 3: Scheme of the evolutionary algorithm used for peptide optimization. The contents of the dashed boxes are explained in detail in the Methods section.

Definition 1. Dominance-relation
Multi-objective optimization aims at simultaneously optimizing m objectives $F = (f_1, \dots, f_m) : \mathbb{R}^n \rightarrow \mathbb{R}^m$ depending on a vector of n parameters or decision variables $X = (x_1, \dots, x_n)$. These parameters may have to adhere to a set of k constraints $g_i(X) \geq 0 \quad \forall i \in \{1, \dots, k\}$. Without loss of generality it can be assumed that all objectives are to be minimized.
Given two individuals represented by the two vectors X_1 and X_2 it is said that X_1 strictly dominates X_2 (denoted $X_1 \succ X_2$) if following is true:

$$f_i(X_1) \leq f_i(X_2) \quad \forall i \in \{1, \dots, m\} \ ,$$

$$F(X_1) \neq F(X_2) \ .$$

Definition 2. Non-dominated individual
Given a set of individuals represented by vectors of decision variables, an individual $X_{non-dom}$ is *non-dominated* if no other member of the set dominates $X_{non-dom}$ in terms of the dominance-relation.

The optimization process stops after a previously determined number of generations. Figure 3 shows the scheme of the algorithm.

Stability

The basis of the stability estimation for a given sequence in a predefined conformation is a MD trajectory of the respective peptide molecule in aqueous solution over $10ns$ at room temperature. The trajectory file is divided into a set of frames, each displaying a conformation adopted during the MD run. For each of these frames the RMSD of all atoms of the conformation represented by the frame to all atoms of the start conformation is computed. Afterwards the proportion of the frames with a RMSD below a certain threshold is calculated. The size of this proportion is taken as stability measure. A threshold of 3.5Å is taken. This threshold is obtained empirically by analyzing the MD runs of two different stable folding peptides (cp. Fig. 4) recorded in the Brookhaven Protein Databank (PDB) [5], Villin headpiece (PDB ID 1VII) [25] and Ubiquitin (PDB ID 1UBQ) [37].

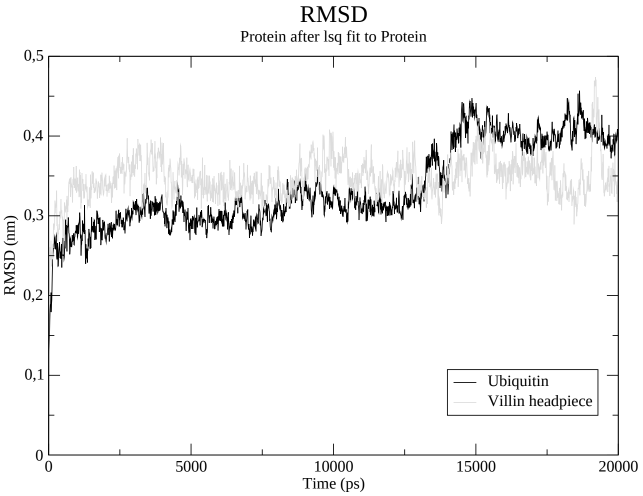

Figure 4: Traces of RMSD to native structures along $20ns$ MD trajectories for two peptides with stable folds, Villin headpiece and Ubiquitin. As start conformations the conformations recorded in the Brookhaven Protein Databank (PDB) were used.

Accessibility

Typically, binding sites and other functional regions are located on the surface of the molecule, accessible to the binding partner. Hence, we adopted as one optimization criterion the accessibility of the binding loop. Since it is difficult to compute the accessibility to the binding partner without knowledge of the geometry of this molecule, we used as a simple approximation the accessibility of the binding loop to water, which seems to be a minimum requirement. As the actual optimization objective we used the minimization of the residuewise RMSD of the solvent accessible surface of the loop in (a) the current and (b) the initial conformation, the latter being the reference conformation with respect to accessibility. The solvent accessible surfaces was computed with the program naccess [18].

RMSD of key residues

As mentioned above it is possible to exclude certain residues from the mutation process, e.g. key residues that presumably are essential for binding. We go a step further and use the RMSD of these key residues of the conformations adopted during a MD run from a given “active” conformation to influence the choice of the mutation site. This ought to result in conformations better reflecting the topography of the functional loop in its ”active” conformation and therefore resulting in good bio-activity.

4. METHODS

Molecular Dynamics (MD)

MD simulates trajectories of interacting atoms in space by integrating their equations of motion. As computing power is not sufficient for accurate simulations of large biomolecular systems at the quantum mechanical level, atoms are usually modeled by classical force fields using simplified bonded and non-bonded (Van der Waals, Coulomb, etc.) interaction potentials. This model is still sufficiently accurate for direct quantitative comparisons with experiment. For a detailed description of the MD program Gromacs and the potentials used in the present work see [4, 23, 36]. Despite the simplified potentials the run times for the generation of a MD trajectory are often large because many atom–atom interactions have to be considered. Typically, production of a 10ns trajectory of a peptide in aqueous solution took a runtime of the order of one day on a machine with a single XEON 3.04GHz processor and 1GB of RAM. This made MD simulation the most expensive component of our method. Nevertheless, we decided to use MD with a detailed model because this facilitates comparison with experiment. Mostly, several MD runs were performed in parallel on 32 processors of a PC cluster.

Alanine scan

In vitro alanine scans are a well established technique that is used to identify residues that have certain biological functions in proteins and peptides [6, 26]. Basically, it is a mutagenesis experiment in which each amino acid in turn is exchanged for an alanine. Alanine is the amino acid with the smallest sidechain, a single methyl group (Glycine is even smaller but has no sidechain at all and thus is very flexible; in this sense it is an untypical amino acid). In essence, replacement of a non-Glycine amino acid by alanine means reduction of the sidechain to a minimum. If an amino acid can be replaced by alanine without significant loss of "activity" (e.g. stability, or affinity to some other molecule) this means that the removed sidechain is probably not essential in this respect. On the other hand, a change of activity due to the loss of the sidechain implies some role of this sidechain.

We used a computational variant of alanine scanning to estimate the contribution of an amino acid to the conformational stability of a peptide in a desired conformation C_D, and to predict stabilizing mutations. Shortly, for a peptide of n amino acids of which k residues were considered to be key residues, we generated $n-k$ sequences by replacing residue i by alanine while leaving all other $n-k-1$ sequence positions unchanged. For each of these sequences we then computationally estimated the stability of C_D. The sequence position where mutation to alanine led to the greatest stabilization of C_D was considered a candidate position for a stabilizing mutation. At these candidate positions we then computationally tested the effects of all possible mutations.

In detail, the alanine scan was performed as follows. In order to quantify the stability of the key residue loop conformation C_D we had first to define two representative sets of conformations, one set with a loop conformation similar to C_D, and a second set with a larger conformational deviation. Stability of C_D then means that the first set has a lower free energy than the second one. Hence, first two sets of conformations were extracted from the MD trajectory of the peptide of the previous generation: a set N_{low} of conformations with low conformational RMSD (< 1.15Å) of the key residues to C_D, and a set N_{high} of conformations with a high RMSD to C_D (> 1.65Å). The thresholds of 1.15 Å and 1.65 Å have been determined empirically; the rationale for choosing thresholds between 1 and 2 Å is that structures with an RMSD of around 1 Å and less can satisfy the same local binding pattern of hydrogen bonds and hydrophobic contacts, whereas larger RMSDs are in general not compatible with the same binding pattern. If available, up to ten conformations below and above the thresholds, respectively, were chosen randomly from the MD trajectory for each of the two sets N_{low}, N_{high}. Two conformations were always put into the two sets: the conformation of lowest RMSD recorded in the trajectory was always a member of N_{low}, and the conformation of largest RMSD always part of N_{high}. Note that the use of RMSD thresholds implies optimization of peptides towards small RMSDs to C_D.

After having prepared N_{high} and N_{low} the stability of C_D for the parent sequence and for each of the $n-k$ sequences generated by alanine exchanges had to estimated. The essential physical quantity here is the free energy G, that we approximated by a widely used expression shown in Eq. (1), with the conformational part G_{ff} calculated with a classical force field, the non-polar part G_{np} of the peptide-water interaction assumed to be linearly dependent on the solvent accessible surface [8], and the polar part G_{es} of the peptide-water interaction computed with a classical continuum electrostatics approach [17]. Technically, the G_{ff} was computed with Gromacs [4, 23], G_{np} with a the solvent accessible surface obtained with naccess [18] and a surface tension constant of 40 J/Å^2, and G_{es} was computed with the program solvate [3] using PARSE van der Waals radii [32] for the peptide atoms.

$$G = \quad G_{ff} + G_{np} + G_{es} \ , \tag{1}$$
$$\Delta G_{j,i} = \quad G_{C_{U,j},i} - G_{C_D,i} \ . \tag{2}$$

Using Eq. 1 we could identify amongst the N_{low} conformations of the parent sequence the conformation with the lowest free energy. We treated this single conformation in the following as C_D. After this, all alanine mutants were prepared for C_D and all conformations in N_{high} using the modeling program WhatIf [38]. For all of these sequence–conformation pairs we computed the free energy difference $\Delta G_{j,i}$ (Eq. 2). Assuming a Boltzmann distribution of free energies, the ratio of probabilities of $C_{U,j}$ and C_D for the same sequence i is given by

$$\frac{p(C_{U,j},i)}{p(C_D,i)} = \quad \exp\ -\frac{\Delta G_{j,i}}{RT} \quad , \tag{3}$$

with the gas constant R and the absolute temperature T. It can be expected for a sequence position i that the higher the stability of C_D the smaller becomes the term P_i given in Eq. 4.

$$P_i = \quad \sum_{j=1}^{N_{high}} \frac{p(C_{U,j},i)}{p(C_D,i)} \ . \tag{4}$$

P_i was computed for all alanine mutation positions i and one of these positions selected at random with a probability inversely proportional to P_i.

After a position of a mutation had been chosen in this way by alanine scanning, we mutated this site into all possible amino acids. Using an analogous formalism and the same computational techniques as described above, we then selected one of these amino acids according to its stabilizing effect.

Our double-mutation strategy – first each residue testwise into alanine, then the actual mutation into some other, hopefully stabilizing, amino acid – samples only a small fraction of sequence changes that are possible in one step, namely $n - k + 19$ vs. $(n - k) \cdot 19$. Hence, it is likely that we missed mutations that increase stability. However, computing all the full energies of $(n - k) \cdot 19$ possible one-step mutants would be very costly, since already a single alanine scan followed by selection of a new amino acid took half a day on a 3.04GHz XEON dual processor. Thus our strategy is a compromise between the requirements of high accuracy and low cost.

5. RESULTS AND DISCUSSION

We tested our approach by carrying out computer experiments with Villin headpiece (VH). VH is a good test-case for such experiments for several reasons. Firstly, it is one of the smallest peptides known that folds autonomously into a stable and well-defined native conformation. Secondly, high resolution structural data for VH is available [25] which gives us validated conformational reference. Thirdly, the role of several residues for the stability of VH was investigated [13] experimentally; in these studies it was found that a set of hydrophobic amino acids are crucial for stability.

These experimental studies suggested a test for our algorithm: if we perturb the wild type sequence of VH by replacing one of the stabilizing hydrophobic amino acids by a strongly polar one, this should lead to a destabilized native conformation of this VH mutant; our algorithm should then be able to predict that reverting to the unperturbed wild type sequence will result in a stabilization of the native VH conformation.

To test our algorithm we created the VH mutant F18K, where the stabilizing hydrophobic phenylalanine at position 18 [13] was changed to a polar, and presumably destabilizing lysine. We first simulated the wild type and the F18K mutant using MD in aqueous solution and found indeed that the wild type sequence was stable in its experimentally determined native conformation whereas the mutant left this conformation quickly and remained highly mobile throughout the simulation (Fig. 5).

Then we subjected the F18K sequence to the evolutionary optimization method described above with the native conformation of wild type VH as desired conformation C_D. Run parameters are summarized in Tab. 1. The number of generations was limited by the available CPU-time to 15. A whole run then took about three weeks with a cluster of 14 CPUs. The outcome of the optimization was surprising. The method came up with variants that, according to our computational approximation, are more stable in the native VH conformation than the VH wild type sequence (Fig. 5).

In one of these variants the method reverted the mutation F18K (in generation three) by a back mutation K18F, and further on introduced a new mutation G34L (in generation one). The set of sequences covering all double mutants of a 36-residue peptide, such as VH, and using the full alphabet of 20 amino acids has about $(36 \cdot 20)^2 = 518400$ elements; the actual number is somewhat smaller because back-mutations can occur that do not contribute new sequences. It can be assumed that with respect to the wild type conformation most of these sequences are less stable than the wild type sequence and only a few sequences are more stable. Our algorithm generated $15 \cdot 8 = 120$ sequences. The fact that amongst these it found not only the K18F reversion but in addition a second mutant that, according to the calculation surpasses the stability of the wild type is encouraging in view of the accuracy of the physical model and the efficiency of the evolutionary algorithm.

On the other hand one may ask why nature has not found this more stable VH variant. Several reasons are imaginable: VH is part of the larger protein Villin, and nature may have optimized Villin as a whole and not only VH; the glycine at position 36 may have a particular biological function that cannot be fulfilled by leucine at that position; the G34L mutant may be not soluble in water and thus not functional; the computational prediction of the higher G34L stability

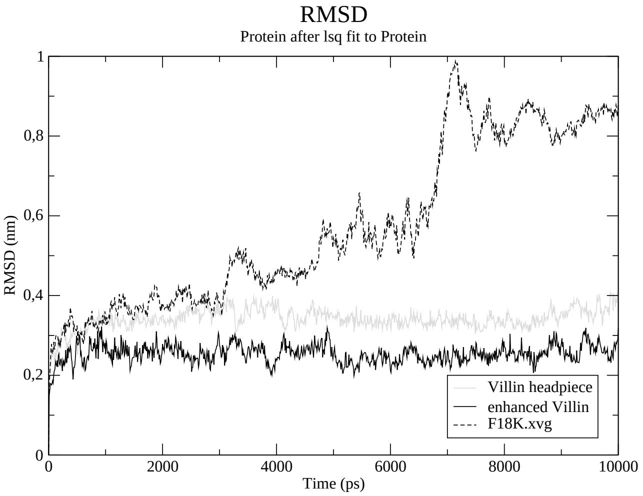

Figure 5: RMSD trace over $10ns$ MD simulations of three peptides to the native, experimentally determined structure of Villin headpiece (VH). Light grey: wild type VH sequence. Dashed: the unstable VH mutant F18K. Solid black: the G34L mutant of VH showing enhanced stability in the native VH conformation.

Table 1: Parameters of computer experiment with VH and mutants.

Parameter	Value
Algorithm Parameters	
Generations	15
Population size	8
Tournament size	2
Elitism	best individual
MD Parameters	
Duration of each MD run	$10ns$
Coulomb cut-off	$1.4nm$
van der Waals cut-off	$1.0nm$
Temperature	$300K$
Temperature coupling algorithm	Nose-Hoover

could be incorrect because we may not have accumulated enough data in the MD simulations that are the basis of stability estimation.

Apart from the formerly described double mutant the algorithm proposed an ensemble En_A of multi site mutated sequences. These four sequences carried five to nine mutations, which reversed the character of the present amino acids from hydrophobic to hydrophilic or vice versa. This change in peptide surface properties results in a change of the peptide backbone conformations. The RMSDs of these multi site mutants compared to the VH native conformation ranged from 5.5Å to 6.4Å whereas the double mutant showed a backbone RMSD of 2.2Å. We checked our results by repeating the run. Again the algorithm proposed an ensemble En_B of multi site mutants (four to seven mutations) with enhanced stability; the RMSD of En_B to the native VH conformation was 2.6Å to 2.8ÅInterestingly, the two ensembles (En_A, En_B) show a common, stability enhancing set of mutation sites. This set of sites consists of the sequence positions 2, 29, 32, 33, 34. The existence of such a common set supports the view that the results of the algorithm are reproducible.

Another interesting fact is that contrary to the studies of Frank *et al.* [13] there seem to exist VH mutants that are stable even without the F18, a phenylalanine at sequence position 18. There might be at least two explanations for this. Firstly Frank *et al.* used NMR and wet lab mutagenesis to investigate the impact of point mutations to the VH. Regarding the complexity of these methods their experiments were limited to single and double mutations. Given the fact that it took at least four different mutations to re-stabilize VH, Frank *et al.* could not observe a re-stabilized VH mutant without F18. Secondly we investigated the arrangement of the side chains in direct vicinity to the F18 or L18 respectively. Lysine has a relatively long side chain, only at its top carrying its hydrophilic group. Often this side chain is flexible and exposed to the solvent and in this way makes a stabilizing contribution to entropy. In this special case this flexible and coiled arrangement would severely disrupt the surrounding hydrophobic core, resulting in destabilization of the VH conformation while a totally stretched side chain would result in a penalty due to entropic losses. Now, the introduced mutations made it possible to arrange the lysine side chain in a conformation where its hydrophilic group is exposed to the solvent whereas the non-polar remainder of the sidechain is surrounded by hydrophobic sidechains and thus participates in the hydrophobic core.

Therefore it could be interesting to repeat the experiment using a F18D mutation instead of a F18K mutation. Aspartic acid and lysine both are hydrophilic residues and therefore should disrupt the hydrophobic core arranged around the replaced phenylalanine but their side chains differ in length. With a shorter side chain it is more difficult for the aspartic acid to expose its hydrophilic group. Therefore it needs drastic changes in sequence or conformation to expose this hydrophilic group which might increase the probability for the mutation of residue 18 back to a phenylalanine which could be achieved with only two mutations.

6. CONCLUSION AND OUTLOOK

We have presented a method for the evolutionary optimization of peptide sequences with respect to conformational stability. In this multi-objective approach three criteria are considered simultaneously that are crucial for peptide function: stability, accessibility of key residues, and, implicitly, the RMSD of key residues to a desired conformation.

Using this optimization *in silico* we have predicted variants of the Villin headpiece peptide that are more stable than the wild type. In collaboration with structural biologist, NMR experiments and calorimetric measurements are underway to test our predictions experimentally.

Should our method withstand this and other tests, there is still room for improvements in many directions, such as the implementation of a cheaper energy function, or the consideration of sequence insertions and deletions.

We hope that in this way our method will become a useful tool for the optimization of peptidic drugs such as T20.

7. ACKNOWLEDGMENT

This work was in part supported by a grant of the Bayer Hemophilia Awards Program.

8. REFERENCES

[1] C. B. Anfinsen. Principles that govern the folding of protein chains. *Science.*, 181:223–30, 1973.

[2] T. Bäck. *Evolutionary Algorithms in Theory and Practice*. Oxford University Press, New York, 1996.

[3] D. Bashford and K. Gerwert. Electrostatic Calculations of the pKa Values of Ionizable Groups in Bacteriorhodopsin. *J. Mol. Biol.*, 224:473–486, 1992.

[4] H. J. C. Berendsen, D. van der Spoel, and R. Drunen. GROMACS: A message-passing parallel molecular dynamics implementation. *Comp. Phys. Comm.*, 91:43–56, 1995. http://www.gromacs.org.

[5] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, and I. N. S. a nd P. E. Bourne. The Protein Data Bank. *Nucleic. Acids. Res.*, 28:235–42, 2000.

[6] H.-J. Böhm, G. Klebe, and H. Kubinyi. *Wirkstoffdesign*. Spektrum Akademischer Verlag, Heidelberg, Germany, 1996.

[7] C. Branden and J. Tooze. *Introduction to Protein Structure*. Garland Publishing, New York, 2 edition, 1999.

[8] T. E. Creighton. *Proteins – Structures and Molecular Properties*. W. H. Freeman and Company, New York, 1993.

[9] I. de la Mata, J. L. Garcia, C. Gonzalez, M. Menendez, J. Canada, J. Jimenez-Barbero, and J. L. A. sensio. The impact of R53C mutation on the three-dimensional structure, stability, and DNA-binding properties of the h uman Hesx-1 homeodomain. *Chembiochem.*, 3:726–40, 2002.

[10] G. M. de Mori, G. Colombo, and C. Micheletti. Study of the Villin headpiece folding dynamics by combining coarse-grained Monte Carlo evolution and all-atom molecular dynamics. *Proteins.*, 58:459–71, 2005.

[11] K. Deb. *Multi-Objective Optimization using Evolutionary Algorithms*. Wiley-Interscience Series in Systems and Optimization. John Wiley & Sons, Ltd., Baffins Lane, Chichester, West Sussex, England, 2001.

[12] A. R. Fersht. Nucleation mechanisms in protein folding. *Curr. Opin. Struct. Biol.*, 7:3–9, 1997.

[13] B. S. Frank, D. Vardar, D. A. Buckley, and C. J. McKnight. The role of aromatic residues in the

hydrophobic core of the villin headpiece subdomain. *Protein. Sci.*, 11:680–7, 2002.

[14] A. Globus, J. Lawton, and T. Wipke. Automatic molecular design using evolutionary techniques. *Nanotechnology*, 10:290–299, 1999.

[15] G. Goh and J. A. Foster. Evolving molecules for drug design using genetic algorithms. In D. Whitley, D. Goldberg, E. Cantu-Paz, L. Spector, I. Parmee, and H.-G. Beyer, editors, *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-2000)*, pages 27–33, Las Vegas, Nevada, USA, 2000. Morgan Kaufmann.

[16] H. Hardy and P. R. Skolnik. Enfuvirtide, a new fusion inhibitor for therapy of human immunodeficiency virus infection. *Pharmacotherapy.*, 24:198–211, 2004.

[17] B. Honig and A. Nicholls. Classical electrostatics in biology and chemistry. *Science.*, 268:1144–9, 1995.

[18] S. J. Hubbard and J. M. Thornton. *NACCESS, Computer Program, Department of Biochemistry and Molecular Biology.* University College London, 1993.

[19] I. A. Hubner, M. Oliveberg, and E. I. Shakhnovich. Simulation, experiment, and evolution: understanding nucleation in protein S6 folding. *Proc. Natl. Acad. Sci. U. S. A.*, 101:8354–9, 2004.

[20] B. Kuhlman, G. Dantas, G. C. Ireton, G. Varani, B. L. Stoddard, and D. Baker. Design of a novel globular protein fold with atomic-level accuracy. *Science.*, 302:1364–8, 2003.

[21] M. Lei, M. Yang, and S. Huo. Intrinsic versus mutation dependent instability/flexibility: a comparative analysis of the structure and dynam ics of wild-type transthyretin and it pathogenic variants. *J. Struct. Biol.*, 148:153–68, 2004.

[22] Y. Levy, S. S. Cho, T. Shen, J. N. Onuchic, and P. G. Wolynes. Symmetry and frustration in protein energy landscapes: a near degeneracy resolves the Rop dimer-folding myster y. *Proc. Natl. Acad. Sci. U. S. A.*, 102:2373–8, 2005.

[23] E. Lindahl, B. Hess, and D. van der Spoel. Gromacs 3.0: A package for molecular simulation and trajectory analysis. *J. Mol. Mod.*, 7:306–317, 2001.

[24] A. Liwo, M. Khalili, and H. A. Scheraga. Ab initio simulations of protein-folding pathways by molecular dynamics with the united-residue model of polyp eptide chains. *Proc. Natl. Acad. Sci. U. S. A.*, 102:2362–7, 2005.

[25] C. J. McKnight, D. S. Doering, P. T. Matsudaira, and P. S. Kim. A thermostable 35-residue subdomain within villin headpiece. *J. Mol. Biol.*, 260:126–34, 1996.

[26] K. L. Morrison and G. A. Weiss. Combinatorial alanine-scanning. *Curr. Opin. Chem. Biol.*, 5:302–7, 2001.

[27] J. Moult, K. Fidelis, A. Zemla, and T. Hubbard. Critical assessment of methods of protein structure prediction (CASP): round IV. *Proteins.*, Suppl 5:2–7, 2001.

[28] C. A. Rohl, C. E. Strauss, D. Chivian, and D. Baker. Modeling structurally variable regions in homologous proteins with rosetta. *Proteins.*, 55:656–77, 2004.

[29] J. Schonbrun, W. J. Wedemeyer, and D. Baker. Protein structure prediction in 2002. *Curr. Opin. Struct. Biol.*, 12:348–54, 2002.

[30] E. I. Shakhnovich. Modeling protein folding: the beauty and power of simplicity. *Fold. Des.*, 1:R50–4, 1996.

[31] S. K. Sia, P. A. Carr, A. G. Cochran, V. N. Malashkevich, and P. S. Kim. Short constrained peptides that inhibit HIV-1 entry. *Proc. Natl. Acad. Sci. U. S. A.*, 99:14664–9, 2002.

[32] D. Sitkoff, K. A. Sharp, and B. Honig. Correlating solvation free energies and surface tensions of hydrocarbon solutes. *Biophys. Chem.*, 51:397–403; discussion 404–9, 1994.

[33] J. F. Smothers, S. Henikoff, and P. Carter. Affinity selection from biological libraries. *Science*, 298:621–622, 2002.

[34] R. Srinivasan and G. D. Rose. Ab initio prediction of protein structure using LINUS. *Proteins.*, 47:489–95, 2002.

[35] H. Toyobuku, Y. Sai, T. Kagami, I. Tamai, and A. Tsuji. Delivery of peptide drugs to the brain by adenovirus-mediated heterologous expression of human oligopeptide tr ansporter at the blood-brain barrier. *J. Pharmacol. Exp. Ther.*, 305:40–7, 2003.

[36] D. van der Spoel, E. Lindahl, B. Hess, A. R. van Buuren, E. Apol, P. J. Meulenhoff, D. P. T. an, A. L. T. M. Sijbers, K. A. Feenstra, R. van Drunen, and H. J. C. Berendsen. Gromacs User Manual version 3.2., 2004. http://www.gromacs.org.

[37] S. Vijay-Kumar, C. E. Bugg, and W. J. Cook. Structure of ubiquitin refined at 1.8 A resolution. *J. Mol. Biol.*, 194:531–44, 1987.

[38] G. Vriend. WHAT IF: A molecular modeling and drug design program. *J Mol Graph.*, 8:52–56, 1990.

[39] Y. Xia and M. Levitt. Funnel-like organization in sequence space determines the distributions of protein stability and folding rate preferred by evolution. *Proteins.*, 55:107–14, 2004.