

# Investigations into the effect of multiobjectivization in protein structure prediction

Julia Handl, Simon C. Lovell, and Joshua Knowles

The University of Manchester, UK

{j.handl,simon.lovell,j.knowles}@manchester.ac.uk

**Abstract.** Physics-based potential energy functions used in protein structure prediction are composed of several energy terms combined in a weighted sum. ‘Multiobjectivization’ — splitting up the energy function into its components and optimizing the components as a vector using multiobjective methods — may have beneficial effects for tackling these difficult problems. In this paper we investigate the hypotheses that multiobjectivization can (i) reduce the number of local optima in the landscapes, as seen by hillclimbers, and (ii) equalize the influence of different energy components that range over vastly different energy scales and hence usually swamp each other’s search gradients. The investigations use models of two real molecules, the alanine dipeptide and Metenkephalin under the Amber99 energy function, and consider hillclimbers with a range of mutation step sizes. Our findings support the hypotheses and also indicate that multiobjectivization is competitive with alternative methods of escaping local optima.

## 1 Introduction

The accurate prediction of protein structure from sequence remains one of the biggest challenges in computational biology [1, 10, 19]. Recent work has suggested tackling the problem by decomposing the traditional physics-based energy function into two or more energy components, and optimizing the resulting multiobjective function using multiobjective EAs [3, 4, 18]. The principal argument offered for the attraction of this multiobjective approach is the observation of conflicts between some of the energy components in physics-based energy functions and the fact that an ensemble of candidate solutions rather than a single structure may be obtained [3]. In other words, these papers argue that the set of Pareto optimal solutions, taken as an ensemble, is likely to provide a better answer to the problem of protein structure prediction than would the single-objective optimum, usually a single structure.

In this paper, we are interested in a different aspect of multiobjective optimization, namely the way a decomposition of the energy function impacts on the difficulty of the fitness landscape ‘seen’ by an optimization method. This is closely related to previous work on ‘multiobjectivization’ [2, 9, 12], which argues that the introduction of additional objectives, or the decomposition of an objective into several, may influence the difficulty of a problem, making it easier [2, 9, 12, 14, 17] or harder [2]. The approach taken in this paper is an empirical one in which single- and multiobjective hillclimbers present themselves as useful tools to investigate changes in the difficulty of a fitness

landscape caused by a decomposition of the energy function. Such an empirical analysis is useful, as general results about the changes in the fitness landscape only directly apply to multiobjective algorithms without archives [7], whose use is rarely practicable in real problems.<sup>1</sup> Also, a straightforward visualization of the multiobjective landscape is not possible even for a two-dimensional problem, as the Pareto dominance relation provides us with a partial ranking of solutions only.

The remainder of the paper is structured as follows. Section 2 discusses the properties of physics-based potential energy functions and the motivations behind their decomposition, in terms of facilitating search. Section 3 discusses the main methods used in this paper, including the two molecular structures considered and the hillclimbers used to explore the resulting fitness landscapes. Experimental results are presented and discussed in Sections 4, 5 and 6. Section 7 considers the wider implications of these results and concludes.

## 2 Decomposition of physics-based potential energy functions

A prototypical physics-based potential energy function (here, Amber99 [5]) can be written as a linear combination of six terms:

$$E_s = E_{bs} + E_{ab} + E_{it} + E_{ta} + E_{vdw} + E_{cc},$$

where  $E_{bs}$ ,  $E_{ab}$ ,  $E_{it}$  and  $E_{ta}$  are the bonded terms constraining bond lengths, bond angles, improper torsion angles and torsion angles respectively.  $E_{vdw}$  and  $E_{cc}$  are the non-bonded forces, which arise from van der Waals attractive and repulsive forces and electrostatic interactions respectively.  $E_s$  is to be minimized. A decomposition into non-bonded and bonded components then considers a two-dimensional vector

$$E_v = (E_{vdw} + E_{cc}, E_{bs} + E_{ab} + E_{it} + E_{ta})^T,$$

rather than a single energy value. The set of solutions that are optimal with respect to  $E_s$  form a subset of those that are Pareto optimal with respect to  $E_v$ , so minimization of  $E_v$  as a Pareto multiobjective optimization problem ([6], page 24) is a valid means of finding a solution to  $E_s$ .

The fitness landscapes described by physics-based potential energy functions are highly rugged (multi-modal), which makes them very challenging to optimize. In addition, the scale of the variation within the different energy components differs strongly in these functions: the variation in the non-bonded energies (especially the van der Waals term) is several orders of magnitude larger than that of the bonded terms. Evidently, a large variation in a given term implies the existence of large local gradients in the same terms, which are bound to dominate the overall energy gradients in many areas of the search space. A distinct effect of a decomposition of the function into bonded and non-bonded components is, therefore, an increase in the influence of the bonded objective in those areas of the search space, as the differences in the scales are annihilated and

<sup>1</sup> In particular, [7] shows that multiobjectivization by decomposition causes the introduction of plateaus of incomparable solutions, which can only result in the removal but not in the introduction of local optima in the search space.

the influence of bonded and non-bonded terms is effectively equalized. Importantly, the same effect cannot easily be obtained through a scaling of the individual energy components, as this would not guarantee to preserve the actual energy minimum. The bonded term is smoother than the non-bonded term (as well as having a smaller energy range), so amplifying its influence may help the search process.

The above observation raises the question of whether an increased influence of the bonded components is something that is actually desirable during protein structure prediction. This question can partly be answered through consideration of relevant work in protein structure prediction. Several state-of-the-art prediction methods use mechanisms to suppress the dominating influence of non-bonded energies during the early stages of the search. These measures range from the reformulation or capping of van der Waals forces [19] to a division of forces into short- and long-range components, where long-range components are only periodically updated [8]. The very existence of such techniques suggests that increased guidance by means of bonded terms is seen as favorable at least by some authors.

### 3 Methods

**The alanine dipeptide** The alanine dipeptide is a well-known model system in the protein structure prediction literature [15], with only two degrees of freedom. Despite the simplicity of the peptide, its energy landscape already exhibits some fundamental features of the energy landscape of proteins, such as their multimodality and the dominant influence of non-bonded energies. Its small dimensionality, allows for extensive experimental testing and enabled us to visualize directly the (single-objective) energy landscape, algorithm trajectories and the location of local optima during the interpretation of experiments. Due to space limitations these visualizations are not included in the paper.

To create a model of the peptide suitable for optimization, the molecular modeling software TINKER [16] was used to enumerate all possible integer values (from -179 to 180) for the two dihedral angles, and to determine the potential energy of the resulting conformation using the Amber99 force field.

**Metenkephalin** The molecule Metenkephalin was used as an example of a more complex molecular structure. This protein consists of five amino-acids and has seventeen flexible dihedral angles, which correspond to the degrees of freedom or decision variables in our problem. A complete enumeration of the search space, as done for the alanine peptide, is no longer possible for this size of problem. Each evaluation therefore requires an explicit call to the TINKER molecular modeling software, making these experiments much more expensive, computationally. As a result, the global optimum for this molecule under the Amber99 energy function was not explicitly identified.

**Algorithms** Three different hillclimbers were used to explore the fitness landscapes under integer coding using a standard Gaussian creep mutation operator<sup>2</sup>. The single-objective hillclimber (SHC) always accepts the mutant solution if its objective is equal

<sup>2</sup> We take the floor of the value to make it an integer.

to or better than that of the parent solution. The multiobjective hillclimber (MHC) uses the basic mechanisms described in [13]. It maintains an archive of non-dominated solutions to avoid degradation of solutions (see [13]) and always accepts the mutant solution if it is indifferent or incomparable to the current solution and if it is not dominated by a solution in the archive. The third algorithm, a hybrid hillclimber (HHC), uses single-objective optimization but maintains an archive (of non-dominated solutions under the biobjective formulation) and switches to multiobjective optimization whenever it has failed to find a valid move for 20 consecutive iterations. It switches back to single-objective optimization as soon as an improvement upon the minimum energy value so far has been found.

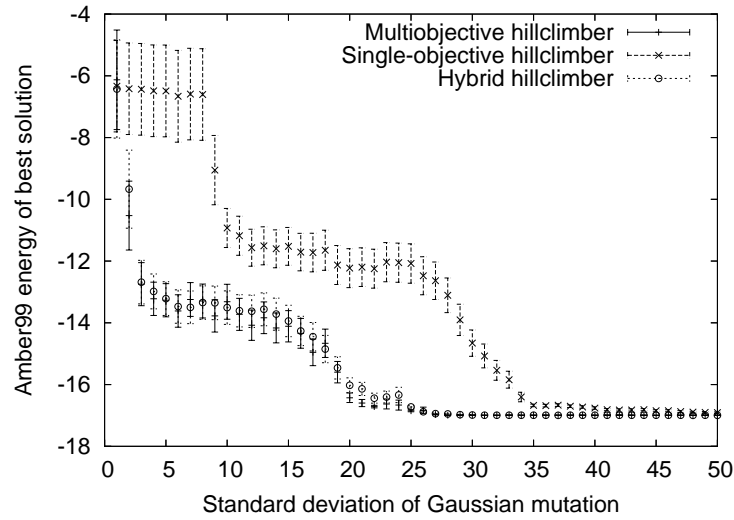
**Experimental details** In all our experiments (see Sections 4, 5, and 6) all three algorithms were run from identical starting positions with a standard mutation rate of  $\frac{1}{n}$ , where  $n$  is the number of decision variables, and for 10000 iterations. The multiobjective and hybrid hillclimbers used a large archive size of 1000 in order to simulate an unbounded archive and remove any influence of the archive’s performance on the search. All experiments were repeated from different starting positions 100 times for the alanine peptide and (due to the much larger computational costs) 15 times for the Metenkephalin molecule. Means of the minimum energy value found per run are reported, and standard errors and p-values (obtained using the Wilcoxon paired rank sum test) are included, where appropriate.

## 4 Comparative performance of the three hillclimbers

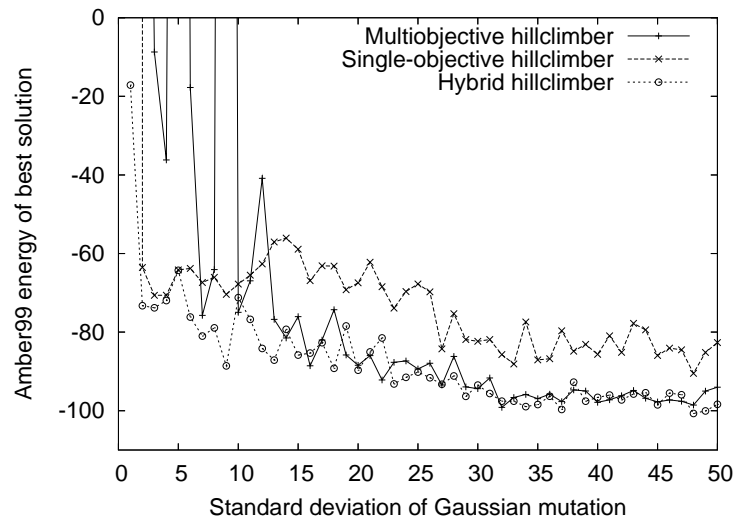
Figure 1 shows the performance of the three hillclimbers for the alanine dipeptide as a function of the standard deviation  $\sigma$  of the Gaussian mutation operator. For  $\sigma \geq 35$ , all three methods show reliable convergence to the global optimum of -16.98 indicating that escape from all local optima is possible using this size of mutation operator. The results also suggest that very large mutation sizes do not hinder convergence for any of the algorithms, but this is likely to be an artifact resulting from the small size of the search space for this particular problem.

Distinct differences between the algorithms can be observed in the regime for  $\sigma < 35$ . For this range of mutation sizes, the single-objective hillclimber converges to local optima with the highest frequency. There are two possible explanations for this result:

1. The hillclimbers utilizing multiobjective optimization can escape local optima more readily at a smaller mutation step size through the exploitation of plateaus of incomparable solutions. Analysis of the trajectories of the single-objective and hybrid hillclimbers provides some evidence for the validity of this latter explanation: the set of solutions accessed by the single-objective hillclimber is usually a subset of those accessed by the hybrid hillclimber. In other words, the trajectories of the two usually agree until a local optimum is met, where the single-objective hillclimber remains while the hybrid hillclimber switches to Pareto optimization and escapes (results not shown).



**Fig. 1.** Performance of the three hillclimbers as a function of the standard deviation  $\sigma$  of the Gaussian mutation operator on the alanine dipeptide. Shown are the averages and the standard error over 100 runs.



**Fig. 2.** Performance of the three hillclimbers as a function of the standard deviation  $\sigma$  of the Gaussian mutation operator on Metenkephalin. Shown are the averages over 15 runs. For  $\sigma < 15$ , SHC and HHC outperform MHC with p-values of 0.0001370 and 3.073e-08, respectively. SHC and HHC are not significantly different at the 0.01 level. For  $\sigma \geq 15$ , MHC and HHC outperform SHC with a p-value  $< 2.2e - 16$ . MHC and HHC are not significantly different at the 0.01 level.

2. The multiobjective formulation (in particular the larger emphasis on the bonded objective) provides better guidance in the search space and steers the multiobjective algorithm towards the global optimum and away from the local optima. The good performance of the hybrid hillclimber (which does not utilize multiobjective optimization during the early stages of the search) gives some indication that this is not happening. Nevertheless the validity of this explanation will be investigated further in the next subsections.

In cases where several hillclimbers find the same local or global optimum, the single-objective and hybrid hillclimber are usually more efficient at converging towards the optimal solution than their multiobjective counterpart (results not shown). This is a further side-effect arising from the introduction of plateaus of non-dominated solutions, which cause the multiobjective hillclimber to spend time exploring regions away from the global optimum. For Metenkephalin, this slower convergence actually results in a performance advantage of the single-objective and hybrid hillclimber for small mutation sizes, as shown in Figure 2. Overall, however, a performance advantage of the hillclimbers using multiobjective optimization also remains for this more complex molecule and can now mainly be observed for larger mutation step sizes.

## 5 Random decompositions

The above experiments indicate a performance advantage of hillclimbers utilizing multiobjective optimization. As mentioned above, one may speculate that, in addition to the presence of plateaus facilitating the escape from local optima, these methods may benefit from the stronger emphasis on the bonded objective, which may help to direct the search towards the global optimum (and away from local optima).

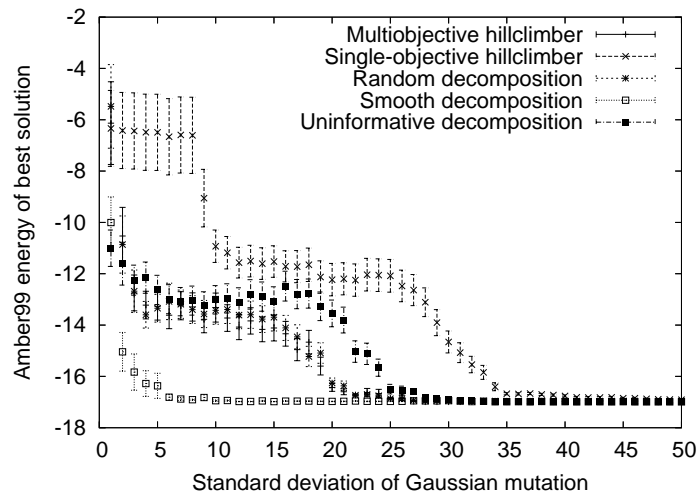
In order to test this hypothesis further, a set of control experiments were conducted on the alanine dipeptide that compared the performance of a number of alternative decompositions of the overall energy function. In particular, we considered biobjective formulations of the form:

$$F = (E_{bs} + E_{ab} + E_{it} + E_{ta} + E_{vdw} + E_{cc} - r, r)^T.$$

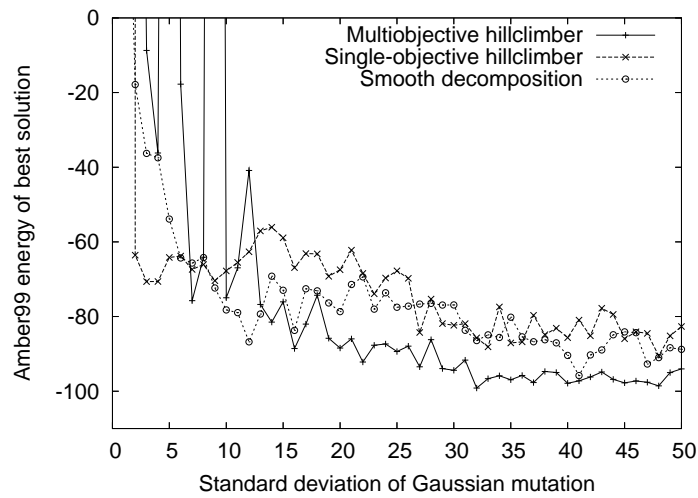
Evidently,  $E_v$  is a special case of such a decomposition, where  $r = E_{bs} + E_{ab} + E_{it} + E_{ta}$ . The alternative definitions of  $r$  considered were:

1.  $r$  has the same properties as the bonded objective, but is uninformative. This effect was obtained by choosing  $r$  to correspond to the bonded energy for a conformation with interchanged phi and psi angles.
2.  $r$  is extremely rugged. This effect was obtained by choosing  $r$  to correspond to the bonded energy with the phi and psi angles randomly permuted (a random mapping).
3.  $r$  presents a smooth gradient (in arbitrary direction). This effect was obtained by choosing  $r$  as the sum of all decision variables.

Figure 3 compares the performance of the multiobjective hillclimber on the alanine dipeptide using these alternative decompositions to that of the original multiobjective



**Fig. 3.** Performance of SHC and MHC, as well as MHC using three alternative decompositions, as a function of the standard deviation  $\sigma$  of the Gaussian mutation operator on the alanine dipeptide. Shown are the averages and the standard error over 100 runs.



**Fig. 4.** Performance of SHC and MHC, as well as MHC using the smooth decomposition, as a function of the standard deviation  $\sigma$  of the Gaussian mutation operator on the Metenkephalin molecule. Shown are the averages over 15 runs. For  $\sigma < 15$ , SHC outperforms MHC and the smooth decomposition with p-values of 0.0001370 and 1.234e-05, respectively. MHC and the smooth decomposition are not significantly different from each other at the 0.01 level. For  $\sigma \geq 15$ , MHC outperforms SHC and the smooth decomposition with a p-value  $< 2.2e - 16$ . HHC and the smooth decomposition are not significantly different at the 0.01 level.

and single-objective hillclimber. The results confirm that, on the alanine dipeptide, the second objective acts as an escape mechanism: the degree of information provided by  $r$  has very little impact and the performance of the hillclimbers is primarily influenced by the ruggedness of  $r$ , as a smooth gradient along  $r$  provides the facility to ‘drift’ out of a local optimum in the first objective. Consequently, the decomposition based on a smooth  $r$  turns out as the strongest performer in this comparison and was further evaluated for the Metenkephalin molecule (see Figure 4). For small mutation step sizes ( $\sigma \leq 15$ ), the experimental results on Metenkephalin appear to confirm those obtained for the alanine dipeptide and the decomposition based on the smooth  $r$  performs somewhat more robustly than the original decomposition (result not statistically significant). However, for  $\sigma \geq 15$  the original decomposition shows a significant performance advantage over the smooth decomposition, indicating that the method may, after all, benefit from the additional guidance provided by the bonded objective. Together with the good performance of the hybrid hillclimber, this result may indicate that increased emphasis on the bonded objectives mainly matters during the escape from local optima.

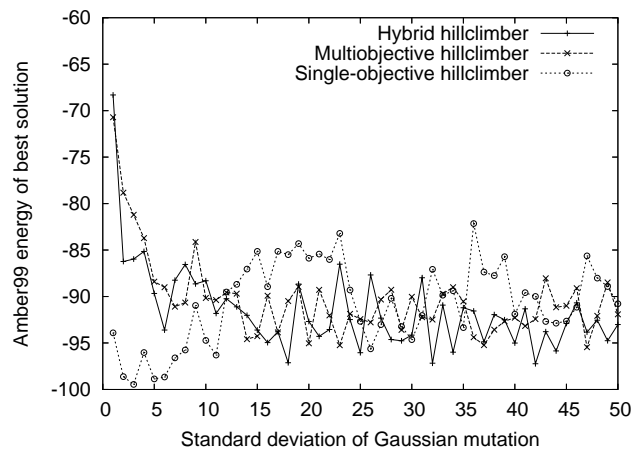
## 6 Alternative escape mechanisms

Some of the success of the multiobjective and hybrid hillclimber can be attributed to the introduction of plateaus that facilitate the escape from local optima. On the downside, the presence of these plateaus slows down convergence speed, which is at the root of the superior performance of the single-objective and hybrid hillclimbers on Metenkephalin for small mutation step sizes. There is thus a trade-off to be met regarding the introduction of plateaus, and it is likely that more effective escape mechanisms than multiobjectivization exist.

In a final experiment, we consider the relative performance of the three hillclimbers if they are furnished with additional escape mechanisms. In particular, the mechanism chosen here is the macromutation operator proposed in [11], which simulates uniform crossover between the current and a random solution. The offspring that inherits more than fifty per cent of its genes from the current solution is taken as the mutant solution. This macromutation is applied with a probability of 0.7 in every iteration.

Figure 5 shows the results obtained for Metenkephalin. The results obtained show a significant increase in the performance of the single-objective hillclimber for small mutation step sizes. In contrast, the performance of the multiobjective and hybrid hillclimbers suffers from the introduction of the macromutation, (which may appear surprising given their increased performance for large mutation step sizes — see Figure 2). This is probably because, for the multiobjective hillclimbers, the macromutation causes them to spend too much time in plateaus (of non-dominated solutions) in the search space, which affects the degree of convergence that can be achieved by them.

When comparing the results obtained using the best parameter settings for each of the three types of hillclimbers (best overall averages are obtained by the single-objective hillclimber with macromutation for  $\sigma = 4$ , the hybrid hillclimber without macromutation with  $\sigma = 48$  and the multiobjective hillclimber without macromutation with  $\sigma = 32$ ), no statistically significant difference can be observed.



**Fig. 5.** Performance of the three hillclimbers with macromutation as a function of the standard deviation  $\sigma$  of the Gaussian mutation operator on the Metenkephalin molecule. Shown are the averages and the standard error over 15 runs. For  $\sigma < 15$ , SHC outperforms MHC and HHC with p-values of  $4.534e-14$  and  $7.731e-11$ , respectively. MHC and HHC are not significantly different from each other at the 0.01 level. For  $\sigma \geq 15$ , MHC, SHC and HHC are not significantly different at the 0.01 level.

## 7 Conclusion

This study has explored the impact of multiobjectivization on the potential energy functions used in protein structure prediction. Compared with a simple hillclimber, a multiobjective hillclimber operating on a decomposed two-objective energy function finds lower overall energy solutions for the same number of evaluations - and does so over a range of mutation step sizes. Experiments to investigate this advantage indicate that multiobjectivization achieves a reduction in the number of local optima in the landscape whilst simultaneously maintaining some of the important “guidance” (or gradient) that the landscape possesses.

When comparing multiobjectivization to more advanced search methods, namely the inclusion of a well-respected macromutation operator to facilitate escape from local optima, we find no advantage in terms of the minimal energy achieved at the best mutation step-size settings. However, the multiobjective approach seems slightly more robust to different step-size choices. More importantly, multiobjectivization finds the low energy solutions at the same time as finding many other non-dominated trade-off solutions, and at no extra cost in function evaluations. Whether or not these additional trade-offs are valuable for identifying native structures is not considered here, but is the subject of our future work.

**Acknowledgments** JH gratefully acknowledges support by a Special Training Fellowship from the Medical Research Council (MRC), UK. JK is supported by a David Phillips Fellowship from the Biotechnology and Biological Sciences Research Council (BBSRC), UK

## References

1. R. Bonneau, J. Tsai, I. Ruczinski, D. Chivian, C. Rohl, C. Strauss, and D. Baker. Rosetta in CASP4: progress in ab initio protein structure prediction. *Proteins*, Suppl. 5:119–126, 2001.
2. D. Brockhoff, T. Friedrich, N. Hebbinghaus, C. Klein, F. Neumann, and E. Zitzler. Do additional objectives make a problem harder? In *Proceedings of the 9th Annual Conference on Genetic and Evolutionary Computation*, pages 765–772. ACM Press, New York, NY, 2007.
3. V. Cutello, G. Narzisi, and G. Nicosia. A multi-objective evolutionary approach to protein structure prediction. *J. R. Soc. Interface*, 3(6):139–51, 2006.
4. R. O. Day, J. B. Zydallis, G. B. Lamont, and R. Pachter. Solving the protein structure prediction problem through a multiobjective genetic algorithm. *Nanotechnology*, 2:32–35, 2002.
5. Y. Duan et al. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *Journal of Computational Chemistry*, 24(16):1999–2012, 2003.
6. M. Ehrgott. *Multicriteria Optimization*. Springer-Verlag, Berlin, Germany, 2005.
7. J. Handl, S. Lovell, and J. Knowles. Multiobjectivization by decomposition of scalar cost functions. In *Proceedings of the Eighth International Conference on Parallel Problem Solving from Nature*. Springer-Verlag, Berlin, Germany, 2008.
8. M. Jacobson, D. Pincus, C. Rapp, T. Day, B. Honig, D. Shaw, and R. Friesner. A hierarchical approach to all-atom protein loop prediction. *Proteins: Structure, Function, and Bioinformatics*, 55(2):351–367, 2004.
9. M. Jensen. Helper-objectives: Using multi-objective evolutionary algorithms for single-objective optimisation. *Journal of Mathematical Modelling and Algorithms*, 3(4):323–347, 2004.
10. D. Jones. Predicting novel protein folds by using FRAGFOLD. *Proteins*, Suppl. 5:127–132, 2001.
11. T. Jones. Crossover, macromutation, and population-based search. In *Proceedings of the Sixth International Conference on Genetic Algorithms*, pages 73–80. Morgan Kaufmann, San Francisco, CA, 1995.
12. J. Knowles, R. Watson, and D. Corne. Reducing local optima in single-objective problems by multi-objectivization. In *Proceedings of the Congress on Evolutionary Multiobjective Optimization*, pages 269–283. Springer-Verlag, Berlin, Germany, 2001.
13. J. D. Knowles and D. W. Corne. Approximating the nondominated front using the Pareto archived evolution strategy. *Evolutionary Computation*, 8(2):149–172, 2000.
14. F. Neumann and I. Wegener. Minimum spanning trees made easier via multi-objective optimization. *Natural Computing*, 5(3):305–319, 2006.
15. G. N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan. Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology*, 7:95–99, 1963.
16. P. Ren and J. W. Ponder. Polarizable atomic multipole water model for molecular mechanics simulation. *Journal of Physical Chemistry B*, 107:5933–5947, 2003.
17. J. Scharnow, K. Tinnefeld, and I. Wegener. The analysis of evolutionary algorithms on sorting and shortest paths problems. *Journal of Mathematical Modelling and Algorithms*, 3(4):346–366, 2004.
18. S. Schulze-Kremer. Application of evolutionary computation to protein folding with specialized operators. In G. B. Fogel and D. W. Corne, editors, *Evolutionary Computation in Bioinformatics*, pages 163–191. Morgan Kaufmann, San Francisco, CA, 2003.
19. K. Simons, C. Kooperberg, E. Huang, and D. Baker. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *Journal of Molecular Biology*, 268:209–225, 1997.