Chapter 5

The Importance of Combinatorial Optimization in the Improvement of Models for Computational Protein Design

Optimization in computational protein design

Initial progress in the field of computational protein design (CPD) was accelerated by the development of mathematically rigorous optimization methods based on the dead-end elimination (DEE) theorem. The availability of these methods helped to instill confidence that provably optimal solutions could be found for astronomically combinatorial protein design problems based on the inverse-folding model. Although the utility of such methods was demonstrated by several successful designs, and many clever improvements were made to extend their applicability, their poor performance scaling soon began to limit the progress of CPD. Reliance on DEE-based optimization was especially problematic when applied in the context of more accurate sampling of side-chain conformational flexibility, the design of many positions simultaneously, or the modeling of substrates and enzymatic transition states.

In response to the limitations of DEE, stochastic optimization routines were developed based on Monte Carlo with simulated annealing (MC), FASTER, and genetic algorithms (GA). Although these methods do not guarantee the generation of optimal solutions, they can be run as long as desired to improve the quality of the solutions, and they always return a solution, regardless of the difficulty of the problem. In practice, we have found that, in contrast to the other stochastic methods, the improved FASTER procedure detailed in Chapter 2 is always able to find the DEE-derived solution when DEE can converge, and is able to converge to a single low-energy solution even for significantly more difficult problems.

Our experiences with the various types of exact and stochastic optimization techniques used in CPD strongly suggest that sampling of configuration space is not the limiting component in the application of single-state, inverse-folding models to real-world protein design problems. Even for the largest inverse-folding problems for which all possible pairwise energies between rotamers can be precomputed and stored in memory, the improved FASTER algorithm can converge to low-energy solutions that are believed (though not proven) to be optimal.

In contrast, the recent development of multi-state design (MSD) procedures has provided more fertile ground for the improvement and testing of optimization routines. MSD procedures must perform individual rotamer-optimization calculations to assess the fitness of each sequence analyzed, and therefore orders of magnitude fewer distinct sequences can be evaluated per unit time. Because scoring a sequence in MSD is so costly, efficient optimization algorithms for MSD must choose sequences to test much more carefully than would be required in single-state design (SSD) problems of equivalent combinatorial size. In Chapter 3, we saw that our implementation of MSD-FASTER significantly outperforms MSD-MC in all cases tested, often finding solutions better than the best ever found by MSD-MC. These results highlight the idea that, unlike SSD problems with precomputed pairwise energies, MSD problems can easily exceed the capabilities of existing sampling algorithms. Thus, more efficient optimization routines are expected to help generate more useful protein variants and accelerate the improvement of CPD models based on MSD.

Design protocols that compute energies on the fly have been investigated to a far lesser extent than those that rely on precomputed energy matrices. So far, the greater

computational expense of on-the-fly methods has precluded their use, despite the CPD model improvements their use enables. For example, on-the-fly methods are amenable to energy functions that cannot be expressed as sums of pairwise energies between positions, such as solvation functions that rely on exact descriptions of complete molecular surfaces. Furthermore, unlike precomputed energy methods, on-the-fly methods need not be limited to rigid main-chain structures. In on-the-fly design methodology, structure refinement and minimization moves can be applied concurrently with rotamer and amino acid changes, potentially facilitating the discovery during the design process of more appropriate scaffold conformations for evaluating the sequences of interest.

This strategy might be most useful in the context of MSD. A database of main-chain structures could be used to score individual sequences, and these structures could be refined during sequence optimization to better represent the sequences found over the course of the design. The database might include both target states and competing states for explicit negative design. Although such methods are expected to improve the predictive ability of CPD calculations, they will also be dramatically more time-consuming than the inverse-folding design calculations to which the field of CPD has become accustomed. These methods will only be rendered tractable by significant advances in computational hardware, as well as the development of conformational sampling algorithms that can handle the combinatorial explosion caused by the treatment of main-chain flexibility.

In Chapter 4, we found that CPD methods can help to predict combinatorial libraries of stable sequences, even when they cannot accurately correlate the experimental

and simulated stabilities of these sequences. Given this result, it seems worthwhile to question the utility of rigorous sampling in CPD calculations. Specifically, if the correlation between simulated and experimental fitness is low, then why bother spending additional time in an attempt to find solutions of better energy?

Characteristics of CPD as a tool for protein engineering

The high-throughput stability assessment of our designed libraries may provide insight into the level of simulation accuracy that might be required for CPD to be usefully applied in protein engineering. It is often postulated that, in order for CPD to display predictive power, it must adequately reproduce stability changes ($\Delta\Delta G$s) of mutation from experimental data sets. However, no correlation was observed between the simulation energies of the individual sequences we assayed and their experimental stabilities. Given this result, we were pleasantly surprised by the ability of our computational library design procedure to produce many well-folded and stabilized sequences based on each type of input structural data. Although it might be assumed that the sequence space of our designs contained an unusually large number of viable sequences, our own data and the reports of others soundly contradict this; we cannot reasonably conclude that the design problem we chose was serendipitously trivial.

So how can a protein design method successfully produce libraries of well-folded, stabilized variants without accurately predicting the relative stabilities of any given pair of mutants? This remarkable property of CPD may arise due to the same fundamental characteristics of proteins that make natural and directed evolution possible.

Although the ability of a protein sequence to fold to a stable and active structure is governed by a precarious balance of energetic contributions with large magnitudes and opposite signs, naturally occurring proteins are nevertheless sufficiently tolerant of substitution to enable the evolution of molecular function through mutagenesis and screening or selection. Starting with an existing functional protein, an area of sequence space enriched with active variants can be explored by iterative cycles of mutation or recombination. This process works because many substitutions can be accommodated by structural adjustments that maintain the general fold, and because the structural accuracy required for activity is not prohibitively high.

Now, we consider CPD methods in light of the biophysical properties of proteins that enable evolution. Inverse-folding design models (including those of the multi-state variety) ultimately score amino acid sequences in the context of one or more fixed scaffold conformations using molecular mechanics and heuristic energy functions. In order to rigorously assess the relative stabilities of any two sequences, a CPD procedure would need to find a representative ensemble of native and nonnative conformations for each sequence, and compute the free energy of each ensemble using a scoring function that accurately treats polar and nonpolar interactions and solvation effects. However, computational tractability requires that only a small subset of the possible conformational space be evaluated, and that approximate scoring functions which neglect explicit water and complex electrostatic effects be used. The finite set of representative structures used for a particular design will always be more appropriate for some sequences than for others. This leads to false positives, in which a sequence appears to stabilize the target ensemble but actually stabilizes alternative conformations more, and false negatives, in

which a sequence appears to destabilize the ensemble although slight adjustments to the ensemble would render it satisfactory. The unpredictability of these cases leads to the observed lack of correlation between simulation energies and experimental measures of fitness.

So, despite insufficient sampling and approximate energy functions, the forgiving nature of protein self-assembly enables CPD to find areas of sequence space likely to be compatible with a given structure and function. As described above, evolution can effectively explore sequence space because stable protein sequences are able to relax structurally and accommodate perturbing mutations. Likewise, CPD procedures are able to locate viable areas of sequence space because a sequence compatible with the simulated ensemble can also usually tolerate the minor relaxations that lead to the physically relevant conformational states that are not modeled. Since the exact nature of these relaxations, and the structures they lead to, cannot be predicted during the simulation, the energy of a sequence threaded on the ensemble does not correlate well with experimental reality. Explicit negative design provides an even greater challenge than positive design, since it demands sequences that *destabilize* an ensemble of competing conformations. Unmodelled structural relaxations are more problematic in competing states than in target states because they can transform an apparently destabilizing interaction into a stabilizing one, rendering a simulation-based fitness assessment *qualitatively* incorrect. Despite these issues, experimental validation of CPD calculations has shown that ensembles sufficiently representative of active states (and competing states, if available) can be used to identify regions of sequence space enriched with folded and functional members.

Although the structural adaptability of a protein native state renders untenable the accurate comparison of arbitrary sequences without prohibitive conformational sampling, it also enables the effective design of proteins under the same set of computational restrictions. Ultimately, we reach the surprising conclusion that accurate scoring of particular arbitrary sequences is neither necessary nor sufficient to find areas of sequence space enriched with functional variants.

In Chapter 4, we discussed how this view of current protein design methods leads to unorthodox proposals for the improvement of CPD. If the utility of CPD is derived primarily from its ability to choose variants that satisfy the native state, as it seems to, then two main avenues of inquiry arise. In the first, structural refinement, larger rotamer libraries, and better energy functions are used to improve the degree to which variants can be ranked based on their compatibility with the native state. However, the general difficulty of finding perfect structures for the evaluation of arbitrary sequences and the extreme sensitivity of molecular mechanics energy functions suggests that additional returns from this effort would diminish quickly; native state modeling is continually pushed to improve its predictive power. On the other hand, simulations of competing states have received scant attention in the context of protein design, and might represent lower-hanging fruit. Of course, the generation of appropriate structural templates for the simulation of competing states will be far from trivial.

The vastness of available conformational space will require redoubled efforts towards efficient sampling and optimization as the major simplifying approximations of CPD begin to be discarded. It seems clear that the development of more accurate design procedures must be driven by the availability of improved optimization methods and

move sets that enable protein sequences to be simulated more realistically. My intent with the projects described here was to push the boundaries of what can be attempted in CPD, to maximize the possibility of transformative breakthroughs derived from this technology. I consider it an honor to have had the opportunity to place my own small piece into this mighty puzzle.