Title
Exploring zipping and assembly as a protein folding principle

Journal
Proteins-Structure Function and Bioinformatics, 66(4)

ISSN
0887-3585

Authors
Voelz, Vince A
Dill, Ken A

Publication Date
2007-03-01

Peer reviewed

Exploring Zipping and Assembly as a Protein Folding Principle

Vincent A. Voelz[†] and Ken A. Dill[‡*]

[†] *Graduate Group in Biophysics and* [‡] *Department of Pharmaceutical Chemistry*

University of California at San Francisco, San Francisco, CA 94143

August 21, 2006

Abstract

It has been proposed that proteins fold by a process called "Zipping & Assembly" (Z&A). Zipping refers to the growth of local substructures within the chain, and assembly refers to the coming together of already-formed pieces. Our interest here is in whether Z&A is a general method that can fold most of sequence space, to global minima, efficiently. Using the HP model, we can address this question by enumerating full conformation and sequence spaces. We find that Z&A reaches the global energy minimum native states, even though it searches only a very small fraction of conformational space, for most sequences in the full sequence space. We find that Z&A, a mechanism-based search, is more efficient in our tests than the Replica Exchange search method. Folding efficiency is increased for chains having: (a) small loop-closure steps, consistent with observations by Plaxco et al. [1] that folding rates correlate with contact order, (b) neither too few nor too many nucleation sites per chain, and (c) assembly steps that do not occur too early in the folding process. We find that the efficiency increases with chain length, although our range of chain lengths is limited. We believe these insights may be useful for developing faster protein conformational search algorithms.

Introduction

How do proteins fold as quickly as they do? In test-tube refolding experiments, protein molecules begin in a disordered denatured state (a broad ensemble of microscopic conformations) and fold when native conditions are restored, sometimes averaging only microseconds to reach the ordered native conformation [2]. It raises the question of how the searching and sorting through the protein's large conformational space of disordered states happens so quickly. This puzzle has been called "Levinthal's Paradox" [3]. Even the simplest disorder-to-order transitions, like the crystallization of sodium chloride, takes days. It follows that protein folding cannot involve a random search over its very large number of degrees of freedom.

There have been two general approaches to modeling how proteins search their conformational spaces: stochastic methods, and "mechanism-based" models. On the one hand, folding is stochastic: different molecules follow different microscopic trajectories, subject to thermal fluctuations. Hence, computer simulations of physics-based models of protein folding commonly use Monte Carlo or molecular dynamics sampling methods. However, so far, such methods have been too slow, using atomically detailed physics-based models, to fold up proteins starting from extended conformations, in the computer.

On the other hand, much effort has also been invested in finding a folding *mechanism,* a sort of roadmap of the main features of the folding process with the fluctuations averaged out. In hierarchical models such as the *framework model,* secondary structures form early and hierarchically assemble into tertiary structures [4, 5]. The *hydrophobic collapse model* proposes that the initial collapse into a globular

*Correspondence to: Ken A. Dill, Department of Pharmaceutical Chemistry, University of California-San Francisco, San Francisco, CA 94143. E-mail: dill@zimm.compbio.ucsf.edu

state can drive the formation of secondary structure [6]. The *nucleation-condensation mechanism* [7] proposes that a diffuse transition state ensemble with some secondary structure nucleates tertiary contacts. Some proteins, such as helical bundles, appear to follow a hierarchical *diffusion-collision model* [8, 9]. Another proposed mechanism is the step-wise assembly of foldon units [10]. Features of both nucleation-condensation and hierarchical models have been observed in computer unfolding simulations [11]. In addition, there have been efforts to identify transition states of protein folding, in order to elucidate the folding mechanisms. Sampling the chain topomers has also been proposed as the main folding event [12, 13], although this class of models has recently been shown to be either inconsistent with experimental data or requiring unphysical search characteristics [14, 15].

However, such proposed folding mechanisms are mainly retrospective summaries of experimental data. They do not constitute a *general principle* that could be used to predict folding routes and rates for any arbitrary amino acid sequence. A viable folding principle should: (a) describe how the protein avoids so many of the possible wrong routes, without advance knowledge of the native state, (b) predict how the folding routes differ for different proteins, and (c) make explicit how such coarse-grained folding routes emerge from microscopic stochastic chain trajectories.

One possible folding principle is the Zipping & Assembly model. In Z&A, local structuring happens first in independent peptide fragment sites along the chain, then those structures either grow (zip) or coalescence (assemble) with other structures, along pathways involving topologically local contacts. This model: (1) is a microscopic description of folding, (2) treats the stochastic nature of the folding process explicitly, (3) describes a general process for reaching different native structures given different amino acid sequences, and (4) we find here, is highly efficient and searches only a small fraction of the conformational space, for most sequences.

We ask here whether the Z&A mechanism can efficiently find global minima for a significant fraction of sequence space. Toward that end, we use the HP lattice model, which is the only model currently available, as far as we know, for which the full conformational spaces can be enumerated exhaustively (in order to prove convergence to the true native state), for every sequence in the sequence space. Our aim here is not general testing and comparison of the many global optimization methods currently available.

Previous work provides evidence that Z&A is a viable model for protein folding. First, it has been shown in lattice models that zipping by itself can reach native states efficiently for a high percentage of foldable sequences [16]. Second, a kinetics model based on zipping and assembly predicts rates and rationalizes Φ values, given the native structures [17, 18, 19]. Third, zipping, which predicts that folding speed depends on the localness of contacts formed, is consistent with the observation of Plaxco et al that the fastest folders are proteins having predominantly local contacts in their native states [1]. In addition, Ozkan et al. have recently demonstrated that a conformational Z&A approach can efficiently reach native structures to near 2Å RMSD using atomically detailed physics-based molecular dynamics simulations [20], with much less computation needed than straightforward molecular dynamics efforts.

Methods

Testing the Z&A Method Using an Exact Model

In order to learn about the underlying complexity of the conformational search problem, we must rely on simple exact models, because no other models are currently able to elucidate the nature of the full conformational space or the conformations of sequences over the whole sequence space. We use the hydrophobic/polar (HP) model [21]. In the HP lattice model, a protein is represented as a self-avoiding chain of hydrophobic (P) and hydrophilic (H) residues living on a two-dimensional square lattice [21]. For a given conformation, each pair of nonbonded hydrophobic residues in contact contributes one unit of favorable energy, ε. The HP model is exactly enumerable, and therefore allows us to know for sure whether a search method reaches the true global minimum, or just a local minimum, and yet the model presents a search problem that has the same challenges a protein has – the folding process seeks a single lowest energy native state in a conformational space that grows exponentially with chain length, and its complexity arises from steric constraints due to chain connectivity and excluded volume, and energetic roughness. Because this model is sufficiently protein-like, we can study the algorithm's performance to gain insight as to how proteins can explore their conformational spaces as they fold. At the same time, simple exact models can address questions not addressible in more detailed models, where it becomes intractable to enumerate the full conformational and sequence spaces.

ZIPSEARCH: A generalized conformational searching algorithm

Here we describe a general algorithm for searching conformational space along zipping and assembly pathways. We model zipping and assembly as a graph-searching algorithm where each node on the graph is a *contact state* (elsewhere called a *contact map*), defined as collection of chain conformations sharing a unique set of contacts, and edges denoting a difference of a single contact. Starting from the top of this graph (in the "no-contact" state), the algorithm searches along edges that connect to topologically local contact states, performing conformational sampling along the way (Figure 1). Figure 2 illustrates the search process for the two-dimensional HP lattice model, which is described above.

To keep track of the current position of local conformational sampling, a list of nodes we call *search heads* is kept in memory. The number of search heads is an adjustable parameter, fixed for each run. Each iteration of the algorithm starts with proposing a number of possible contact states downstream from each search head, and performing local conformational sampling to determine if they are viable. Once this sampling has been performed, the viable states are connected to the graph, and added to each search head's *priority queue*, which is prioritized by lowest-free energy. Each search head is then allowed to move to a set of new nodes, chosen by the highest-priority node in its queue, and new contact state nodes are proposed on the next iteration. If a given search head's queue is empty, it may continue by taking the next highest-priority node from another search head, or else pause. This entire process is repeated until there are no nodes left in the queue, or until a specific target has been reached (in our case, the native state).

The shape of the contact state graph

The number of nodes, $N(c)$, on the graph having c contacts is a function with a peak at intermediate values of c. In an idealization in which there is no chain connectivity and in which chain monomers had no excluded volume, for L possible contacts, there would be 2^L possible contact states; each tier in the graph having c contacts would be populated by $\binom{L}{c}$ nodes (Figure 3a,b). However, a large fraction of such contact states are not viable due to excluded volume and chain connectivity (Figure 3c). For either the ideal or real graph, the number of nodes, $N(c)$, is maximal partway down the graph: there are few nodes representing the many open states (where c is small) and there are few nodes for the very small number of near-native states (where c is near-maximal); there are more contact states in between these limits.

Effective contact order

On the graph representing all the possible routes of folding (sequences of contacts), Z&A routes are those involving the formation of only topologically local contacts at each step. To characterize the extent to which an added contact is topologically local, we use a measure called the *effective contact order* (ECO). A related measure is the contact order (CO), which is defined as $|i - j|$ for two residues in contact at sequence positions i and j. The ECO, on the other hand, characterizes the effective size of a loop that is formed by a contact, given that other contacts already have formed [16]. The key features of the ECO measure are that it is much more closely related to the loop closure entropies along folding pathways, and that it is dependent on the order in which contacts form. Hence, the ECOs of contacts in a protein depend on the folding route, whereas the COs do not. The ECO can be easily calculated by shortest-path graph searching algorithm such as the Dijkstra algorithm [22].

Each edge in the contact state graph has an ECO value, and thus zipping and assembly can be regarded as a search over only low-ECO edges in the graph (Figure 3d). The low-ECO subgraph of protein contact states is much smaller than the full graph, yet usually retains a general binomial shape. This low-ECO subgraph is not guaranteed to have a path from the no-contact state to the native state, but does in many cases.

For the ZIPSEARCH algorithm to perform a low-ECO Z&A search on this subgraph, new nodes must be discovered either by low-ECO growth or assembly (Figure 4). Low-ECO growth is enforced by allowing the chain to grow only if a new low-ECO contact results. Similarly, low-ECO assembly is enforced by allowing substructures to assemble only if there is at least one new contact that is low-ECO.

Algorithm parameters. Our test set consists of the set of all HP sequences of a given chain length for which there is a single unique lowest-energy conformation. We call these *foldable* sequences. We call foldable sequences whose native states can be found by the Z&A algorithm *zippable*. We explored all

foldable HP sequences of chain lengths 12 through 18 monomers, taken from those published by Irback and Troein [23].

Calculating contact state free energies. We calculate free energy differences along contact state graph edges by approximating full chain enumerations from piecewise enumerations. This ignores end effects of the chain and excluded volume, but these effects are negligible for all practical purposes. For growth calculations, the known parent conformation(s) are held fixed and only the added chain links are enumerated. For assembly calculations, the two sets of parent conformations are held rigid while all possible assembly orientations are enumerated. In both cases, the free energy of a particular contact state ξ_j found by enumeration is calculated with respect to its upstream parent node ξ_i. (For sampling *via* assembly, the parent node with the search head is chosen as the upstream parent.)

$$F(\xi_j) = F(\xi_i) + \varepsilon\Delta c - k_B T \ln(N(\xi_j)/N(\xi_i))$$

where ε is the energy of a hydrophobic contact, Δc is the difference in the number of contacts, $N(\xi_i)$ is the number of microstates (i.e. individual conformations), counted for the parent contact state, $N(\xi_j)$ is the number counted for the child contact state, k_B is Boltzmann's constant, and T is the temperature. In practice, the value of ε is kept large enough ($\sim -10k_BT$) so that structures with the highest number of contacts are chosen with the highest priority, and entropic differences determine the priorities within a set of isoenergetic states. The free energy of the no-contact state, $F(\xi_0) = 0$, provides an arbitrary reference.

Replica-exchange Monte Carlo method. For our MC and REMC simulations, we use the MS2 move set, as described in [24]. For REMC, the temperatures T_i were chosen so that average energies $E(T_i)$ were uniformly spaced. $E(T)$ is derived directly from the density of states which we compute beforehand for each sequence. Eight replicas, starting from the extended state, were simulated with temperatures spanning the average melting temperature of each sequence, T_m, from $0.5 - 2.0k_BT_m$. A single nearest-neighbor exchange was attempted every 200 steps, with a typical acceptance ratio of 50-60%. The simulation terminates when the native state is found.

Results

We find that Z&A reaches native structures (global minima in free energy) in the HP model for a large fraction of sequences even when only a small fraction of conformational space is searched. These results indicate that Z&A is a plausible folding principle, which we expect to be applicable to more detailed models too.

Our Z&A algorithm is "greedy" insofar as it only searches locally optimal states, and thus is not guaranteed to find the globally optimal native state. However, the search for the native state can often be very efficient, because Z&A breaks a large global optimization problem down into much smaller local optimization problems. Such "divide-and-conquer" procedures don't work for arbitrary global optimization problems in general, but they do appear to work here for protein folding, at least for most sequences.

Z&A searching is efficient

Figure 5 shows an example of how zipping can find native states quickly, without searching much of conformational space. Along a particular pathway that leads to the native state, only 27 microstates need be searched out of the 15,037 in the entire ensemble. Even if every possible wrong turn is taken, searching the full low-ECO subgraph requires enumerating only 75 microstates.

If we let Ω_0 represent the total number of microstates available to the protein, and if Ω is the number of microstates searched by an algorithm, we can express the *search efficiency* as -$SE = \log_{10}(\Omega/\Omega_0)$, a sort of entropy-like quantity. For an algorithm that searches a full conformational space, $SE = 0$, meaning that the algorithm imparts no search efficiency. For an algorithm for which $SE = 4$, only 1 conformation is searched in every 10^4 total microstates that could have been searched. Thus a large value of the search efficiency SE indicates only a small fraction of the space is searched.

The distribution of search efficiencies across all zippable sequences is roughly Gaussian, with a spread of about 2 or 3 units in the value of SE (data not shown). That is, Z&A is highly efficient in finding native states for a few sequences; quite inefficient for a few others; and of intermediate efficiency for the large preponderance of sequences. The average search efficiency correlates with $N(c)$, indicating

that topological frustration is strong determinant of search efficiency. Zippable sequences having poor search efficiencies typically arise when there is high hydrophobic content, because there are many contact states and many possible low-ECO pathways, only a few of which lead to the native state. On the other hand, sequences that zip efficiently typically have less low-ECO pathways, and more of which lead to the native state. Indeed, the best-case scenario for the search efficiency of zipping would be a protein that forms nothing but low-ECO contacts on-pathway to the native state, with a total search time that scales linearly with the number of native contacts. If we examine how the number of microstates searched by the top 5% of the most efficiently zipped sequences scales with chain length, we find that such linear scaling cannot be maintained because there are invariably more places along the chain to locally zip.

Z&A trades off between search efficiency and sequence coverage

Greedy algorithms, such as Z&A, offer trade-offs between computational efficiency and the number of sequences for which the method can find the native structure. How greedy can a zipping-based search be and still be able to fold most sequences? We tested three greedy low-ECO strategies: one in which no step exceeds ECO = 3, one in which no step exceeds ECO = 5, and a lowest-ECO strategy in which, at each iteration, proposed states are limited to those with ECO $\leq \tau_i$, where τ_i is the lowest possible ECO found downstream of the i^{th} search head. We will refer to these strategies as *ECO-3*, *ECO-5*, and *lowest-ECO*, respectively. In each case, we calculate average search efficiency and the fraction of foldable sequences that are zippable using a particular strategy, as a function of chain length. Our results show an inherent trade-off between sequence coverage and search efficiency evident at all chain lengths (Figure 6).

For the chain lengths we examined, the average search efficiency for all strategies increases linearly with chain length (Figure 7). That is, while the number of microstates in the ensemble grows exponentially with chain length as a^N, where $a \approx 2.7$, the number of microstates explored in zipping and assembly increases roughly as b^N, where $b \approx 1.4$. Thus by breaking the search problem down into local subproblems, zipping and assembly effectively reduces the number degrees of freedom explored per unit chain length. This provides an intuitive explanation of how zipping and assembly can resolve Levinthal's paradox. Given the chain lengths examined in this study, we cannot be certain if this scaling law holds for longer chain lengths, although we have little reason to think otherwise.

We define a metric of "greediness" that we call the relative effective contact order (RECO) that applies independently of the chain length; it is the ECO cutoff divided by the chain length. For the lowest-possible ECO strategy, we calculate the RECO using the average ECO across all edges explored in the contact state graph. When the sequence coverage is plotted versus the RECO, a striking universal relationship is revealed (Figure 8). Regardless of whether the zipping and assembly strategy is ECO-3, ECO-5, or lowest-ECO, almost all foldable sequences in our model can be folded by zipping and assembly if the RECO is restricted to 30% of the chain length. For greedier searches, with even smaller local loop closures, less of the global foldable sequence space is zippable. The trade-off is that those sequences which are zippable can find the native state with high search efficiency. It is interesting to note that native structures of proteins in nature have average contact orders ranging from around 5-30% [25]. Many of the HP sequences we examined with non-local native topologies have few hydrophobic residues and unstable cores; these same sequences are searched with very poor efficiency by low-ECO zipping and assembly.

Z&A outperforms replica-exchange Monte Carlo (REMC) in reaching native states

Our interest here is not in general global optimization methods, but rather in exploring whether a particular proposed folding mechanism – Zipping & Assembly – could provide search efficiency for broad ranges of sequence space. Elsewhere, various search methods have been tested in lattice models [26, 27, 28, 29]. However, we were interested in comparing zipping and assembly with replica-exchange methods, which are generalized ensemble approaches to reducing barriers and increasing sampling efficiencies [30, 31]. We find that Z&A is consistently more efficient than REMC (replica exchange Monte Carlo) by several orders of magnitude (Figure 9). It suggests that compared to general optimization methods, mechanism-based searching may be efficient both for computers and for proteins.

Comparing Z&A to experimental protein folding kinetics

Most small proteins have smooth energy landscapes [32, 33], where the kinetic bottlenecks are determined mainly by topological frustration [34], rather than by deep kinetic traps. In contrast, the energy landscapes

of HP model proteins contain many traps, often because there are so many degenerate energy levels. This can be seen both from the time spent in off-pathway states during Monte Carlo simulations (not shown), and from the general shape of the contact state graph (Figure 10a). However, Z&A appears to avoid many of these traps, and instead shows a computational bottleneck near the middle of the contact state graph, the point of maximal topological frustration (Figure 10b), a key feature of protein folding. If we examine the number of microstates sampled as a function of the number of contacts, we see that the Z&A algorithm does most of its sampling at the point of greatest topological frustration. In the macroscopic view, along a reaction coordinate of the number of contacts, what we might define a "computational transition state" corresponding to this observed rate-limiting step. Zipping and assembly thus indicates how macroscopic bottlenecks can arise from the microscopic dynamics of the search.

Also, Z&A captures the experimental observation of Plaxco et al. that the logarithm of the folding rate depends on the topology of the native structure of a protein. This is measured by the relative contact order (RCO), defined as the average sequence separation of native state contacts divided by chain length [1]. We calculated the computational search rate as the inverse of the number of microstates sampled. For lowest-ECO zipping, the average search rate across all sequences sharing the same topology varies roughly linearly with RCO (Figure 11), although with much scatter, probably because of the simplicity of our model and shortness of the chains.

Optimal zipping and assembly

We calculated the distribution of search efficiencies for all zippable sequences with lengths from 12 to 18, varying two parameters. First, we adjusted the amount of parallelism in searching the graph of contact states by altering the number of search heads. For a single search head, the search is "depth-first" in free energy. As the number of search heads becomes very large (larger than the most populous tier on the contact state graph), the search becomes essentially "breadth-first". On average, it is more efficient to use fewer search heads, but we find that the average search efficiency is not very dependent on the number of search heads. This lack of sensitivity may be because we do not differentiate between parallel and sequential events. With many search heads (more than the number of initial nucleation sites), the efficiency suffers more dramatically.

Second, we varied the assembly threshold, i.e., the stage of folding at which substructures are allowed to assemble. For example, assembly can be attempted after the first chain contact has formed, or after the second, etc. On average, it is more efficient to delay assembly steps until after some growth has occurred. For example, a 2-contact assembly threshold gives a higher average search efficiency than a 1-contact threshold (data not shown).

To study the computational work required for early vs. late assembly, we examined the behavior of the zipping and assembly algorithm at various assembly thresholds, for all zippable sequences. Our results show that most sequences have a clearly-defined optimal assembly threshold, unique for each sequence. This optimal assembly point is readily interpretable by examining the profiles of growth versus assembly assessments performed by the zipping algorithm (Figure 12). Assembling too early can be expensive in terms of later sampling. Assembling too late can be expensive in terms of earlier sampling, at the growth steps. There is an optimal stage for overall search efficiency: some chain growth has already taken place to form substructures, but not too much.

We find that assembly occurs early for some sequences and late for others. For which sequences is early assembly most efficient, and for which sequences is late assembly most efficient? The main factor is simply the number of competing contact states, reflected in the size of the contact state graph. Sequences that have fewer competing contact states prefer earlier assembly, while sequences with many competing contact states prefer late assembly (Figure 13). Examining the shape of the contact state graph helps explain why this is the case: the point at which initial assembly steps are most efficient is usually before the widest tier of the contact state graph, where an assembly event could potentially bypass the computationally intensive steps of growing the chain in the most topologically frustrated region.

These results imply: (1) that assembly may occur at different stages in the physical folding process, and (2) that attempting assembly at strategic points in computational folding may increase the efficiencies of computational folding algorithms. Specifically, our results would suggest that for sequences with high secondary structural propensities (and hence fewer competing contact states) early assembly may be more efficient in leading to native states. Assembly processes have been explored in the diffusion-collision model [9, 35], which performs well for proteins with some degree of structural propensity, especially for helical bundle proteins. For structures with many more competing contact states (e.g. β-sheet structures with little secondary structural propensity), assembly is expected to occur at later stages of folding [36].

Excluded-volume "masking" increases the search efficiency.

In order to further increase the search efficiency of our ZIPSEARCH algorithm, we avoid searching over local steric violations by using a lookup table we call a *mask*. A k-contact mask is simply a function which, given that a particular set $\{c_1, c_2, \ldots c_k\}$ of contacts are present, returns the viability of a proposed additional contact, c_{k+1}:

$$M(c_{k+1}|\{c_1, \ldots c_k\}) = \begin{cases} 1 & \text{if viable} \\ 0 & \text{if not viable} \end{cases}$$

For the lattice enumerations, we compiled lookup tables for masks up to $k = 3$ contacts. Lookup tables for $k > 3$ are combinatorially impractical. A 1-contact mask reduced the conformational sampling to half, and a 2-contact mask reduced our conformational searching to one tenth of the searching in the absence of masks (Figure 14). 3-contact masks offer only slight improvements over the 2-contact masks. These results suggest a similar strategy may be effective in all-atom simulations by compiling conditional probabilities of contact topologies, from physical steric considerations or from the Protein Data Bank.

Conclusion: Zipping & Assembly is a viable folding principle.

We have focused here on principles-based conformational searching for protein folding. Our aim is to identify a conformational search process that has the following properties: (1) it can be applied to any protein, having any native topology or amino acid sequence, (2) it can explain how a protein can reach its native structure, a global minimum in the free energy, despite sampling only a very small fraction of conformational space, (3) it should explain the observation of Plaxco et al. that folding speeds are greatest for proteins having, on average, the most local contacts, and (4) the macroscopic folding routes it predicts should have a sound connection to the stochastic microscopic thermal sampling that the individual chain molecules undergo; hence it can be utilized within physical models.

We have studied Zipping and Assembly as a folding principle. It says that: (1) on the fastest time scales, small local chain fragments search conformations having low entropy loss, (2) such local searches happen at different sites independently along the chain, (3) those local structures that are metastable can then grow additional structure (zip) by recruiting neighboring chain segments, (4) on longer time scales, such local substructures can be combined (assembled), in such a manner that increasingly native-like structure emerges. Within the simple exact HP lattice model, we show that Z&A satisfies the above requirements of a folding principle. Its search efficiency comes from breaking a single large global optimization problem into multiple smaller local optimization problems. We find that it is considerably more efficient than Replica Exchange Monte Carlo in finding native structures, and that it gives a basis for understanding the observed correlation of Plaxco et al. We find that assembly steps are likely to be early in helical proteins and later in beta-sheet proteins.

Because the ZIPSEARCH procedure is a general graph-based method, it can be used with a broad range of protein models, not just with the HP lattice model. Ozkan et al, using a similarly directed zipping and assembly approach, has already demonstrated a proof of principle that Z&A can be effectively used with all-atom simulation models [20].

Acknowledgments

This work was supported by grant GM34993 from the National Institute of Health, and support from UC Discovery/Amgen. Vincent Voelz was supported in part by a Burroughs Wellcome Fund Interfaces in Science Fellowship. The authors would like to thank Michael Kim for helpful insight, Yihong Sui for supporting calculations, and John Chodera for a critical reading of the manuscript.

References

1. Plaxco KW, Simons KT, Baker D. Contact order, transition state placement and the refolding rates of single domain proteins. Journal of Molecular Biology 1998;277:985–994.
2. Kubelka J, Hofrichter J, Eaton WA. The protein folding 'speed limit'. Current Opinion in Structural Biology 2004;14:76–88.

3. Levinthal C. Are there pathways for protein folding? Journal de Chimie Physique et de Physico-Chimie Biologique 1968;65:44–.

4. Rose GD. Hierarchic organization of domains in globular proteins. J Mol Biol 1979;134:447–470.

5. Kim PS, Baldwin RL. Specific intermediates in the folding reactions of small proteins and the mechanism of protein folding. Ann Rev Biochem 1982;51:459–489.

6. Baldwin RL. How does protein folding get started? TRENDS in Biochemical Sciences 1989;14:291–294.

7. Fersht AR. Nucleation mechanisms in protein folding. Current Opinion in Structural Biology 1997; 7:3–9.

8. Karplus M, Weaver DL. Protein-folding dynamics. Nature 1976;260:404–6.

9. Myers JK, Oas TG. Preorganized secondary structure as an important determinant of fast protein folding. Nature Structural Biology 2001;8:552–558.

10. Maity H, Maity M, Krishna MMG, Mayne L, Englander SW. Protein folding: The stepwise assembly of foldon units. Proceedings of the National Academy of Science 2005;102:4741–4746.

11. White GWN, Gianni S, Grossmann JG, Jemth P, Fersht AR, Daggett V. Simulation and experiment conspire to reveal cryptic intermediates and a slide from the nucleation-condensation to the framework mechanism of folding. Journal of Molecular Biology 2005;350:757–775.

12. Debe DA, Carlson MJ, III WAG. The topomer-sampling model of protein folding. Proceedings of the National Academy of Science 1999;96:2596–2601.

13. Makarov DE, Plaxco KW. The topomer search model: A simple, quantitative theory of two-state protein folding kinetics. Protein Science 2003;12:17–26.

14. Wallin S, Chan HS. A critical assessment of the topomer search model of protein folding using a continuum explicit-chain model with extensive conformational sampling. Protein Science 2005; 14:1643–1660.

15. Wallin S, Chan HS. Conformational entropic barriers in topology-dependent protein folding: perspectives from a simple native-centric polymer model. Journal of Physics: Condensed Matter 2006; 18:S307–S328.

16. Fiebig KM, Dill KA. Protein core assembly processes. Journal of Chemical Physics 1993;98.

17. Weikl TR, Dill KA. Folding rates and low-entropy loss routes of two-state proteins. Journal of Molecular Biology 2003;329:585–598.

18. Merlo C, Dill KA, Weikl TR. Phi values in protein-folding kinetics have energetic and structural components. Proceedings of the National Academy of Science 2005;102:10171–10175.

19. Weikl T. Loop-closure events during protein folding: rationalizing the shape of phi-value distributions. Proteins 2005;60:701–11.

20. Ozkan SB, Wu GHA, Chodera JD, Dill KA. Protein folding by zipping and assembly, submitted.

21. Lau K, Dill K. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. Macromolecules 1989;22:3986–3997.

22. Cormen TH, Leiserson CE, Rivest RL, Stein C. Introduction to Algorithms. MIT Press and McGraw-Hill, second edition, 2001.

23. Irback A, Troein C. Enumerating designing sequences in the hp model. Journal of Biological Physics 2002;28:1–15.

24. Chan HS, Dill KA. Transition states and folding dynamics of proteins and heteropolymers. Journal of Chemical Physics 1994;100:9238–9257.

25. Bonneau R, Ruczinski I, Tsai J, Baker D. Contact order and ab initio protein structure prediction. Protein Science 2002;11:1937–1944.

26. Yue K, Dill K. Sequence-structure relationships in proteins and copolymers. Physical Review E 1993;48:2267–2278.

27. Beutler T, Dill KA. A fast conformational search strategy for finding low energy structures of model proteins. Protein Science 1996;5:2037–2043.

28. Bachmann M, Janke W. Multicanonical chain-growth algorithm. Physical Review Letters 2003;.

29. Shmygelska A, Hoos HH. An ant colony optimization algorithm for the 2d and 3d hydrophobic polar protein folding problem. BMC Bioinformatics 2005;6:30–41.

30. Mitsutake A, Sugita Y, Okamoto Y. Generalized-ensemble algorithms for molecular simulations of biopolymers. Biopolymers 2001;60:96–123.

31. Sugita Y, Kitao A, Okamoto Y. Multidimensional replica-exchange method for free-energy calculations. Journal of Chemical Physics 2000;113:6042–6051.

32. Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG. Funnels, pathways and the energy landscape of protein folding: A synthesis. Proteins: Structure, Function, and Bioinformatics 1995;215:167–95.

33. Dill KA, Chan HS. From levinthal to pathways to funnels. Nature Structrual Biology 1997;4:10–19.

34. Clementi C, Nymeyer H, Onuchic JN. Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? an investigation for small globular proteins. Journal of Molecular Biology 2000;298:937–953.

35. Karplus M, Weaver DL. Protein folding dynamics: The diffusion-collision model and experimental data. Protein Science 1994;3:650–658.

36. Searle MS, Ciani B. Design of beta-sheet systems for understanding the thermodynamics and kinetics of protein folding. Current Opinion in Structural Biology 2004;148:458–464.

Figures

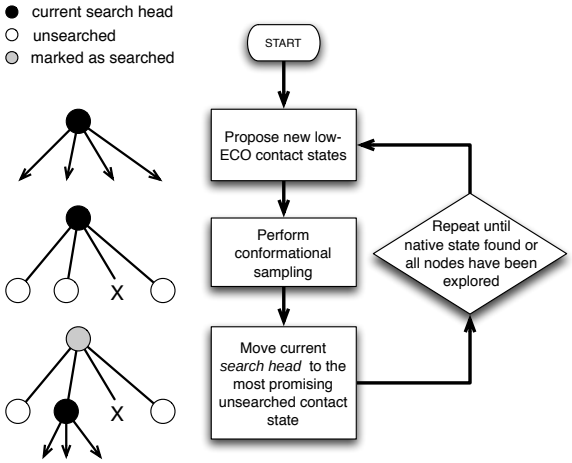

Figure 1: **A flowchart of the ZIPSEARCH algorithm.** ZIPSEARCH is a generalized algorithm for searching protein conformational space along zipping and assembly pathways, designed to navigate the graph of protein contact states. The black node represents a contact state currently being visited. The algorithm keeps track of its current position using a placeholder we call a 'search head'. A parallel search can be performed using multiple search heads. The white nodes are unsearched contact states, and the gray nodes are those which have been marked as searched. The 'X' denotes a contact state that was proposed but not encountered in the conformational sampling phase.

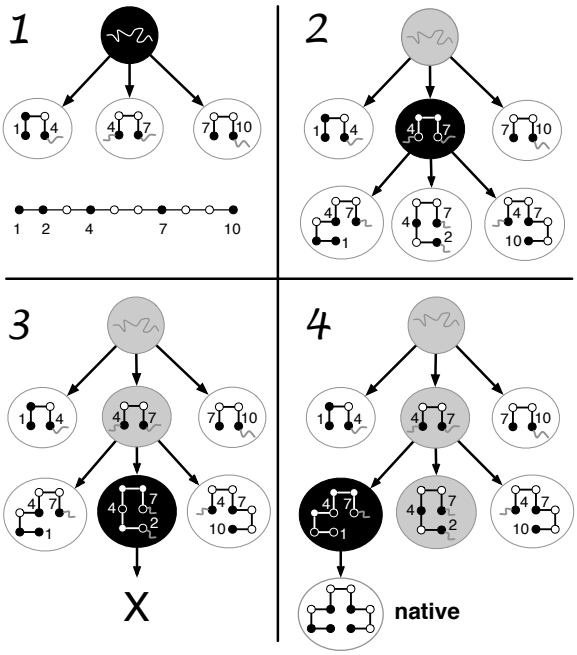

Figure 2: **Several iterations of the ZIPSEARCH algorithm showing how chain conformations are sampled for an example HP lattice model protein (see Methods).** The coloring conventions are as in the previous figure. On the first iteration, there are three possible local contacts that can be formed. The conformations of the chain fragments which contain each contact are enumerated and added to the graph. On the next iteration, the search head is moved to the most promising (lowest-free energy) unsearched node, and more conformational sampling of chain fragments is performed, keeping the parent structures fixed. If the algorithm reaches a dead end, the next-best unsearched contact state is explored (in this case, leading the native state).

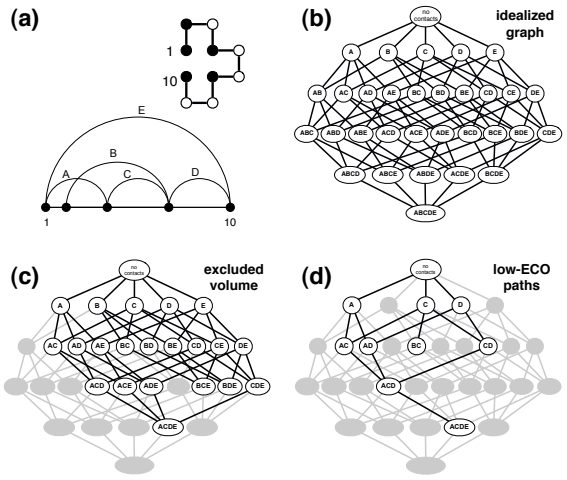

Figure 3: **Visualizing the graph of protein contact states.** Shown in the upper left is a model protein in its native conformation and a polymer graph showing all possible contacts available to the chain. In the absence of excluded volume, the full contact graph is an idealized binomial-shaped structure (upper right), but for real chains many contact states are not viable (lower left). The subgraph of contact states reachable solely by lowest-ECO loop closures has many fewer nodes, yet the native state is still reachable (lower right). Zipping and assembly is equivalent to searching over this subgraph, an efficient way to find the native state (ACDE). Despite the pruning, this low-ECO subgraph retains the basic binomial shape.

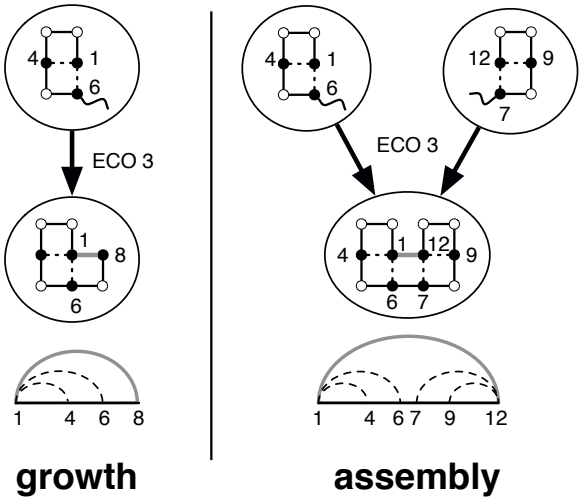

Figure 4: **New nodes on the graph can be proposed by either growth or assembly.** For a low-ECO Z&A search, proposed contact states are limited to only those which contain a new contact with low effective contact order (ECO). Lattice conformations illustrate the formation of new low-ECO contacts. For the 2D lattice, an ECO of 3 is the smallest loop closure possible.

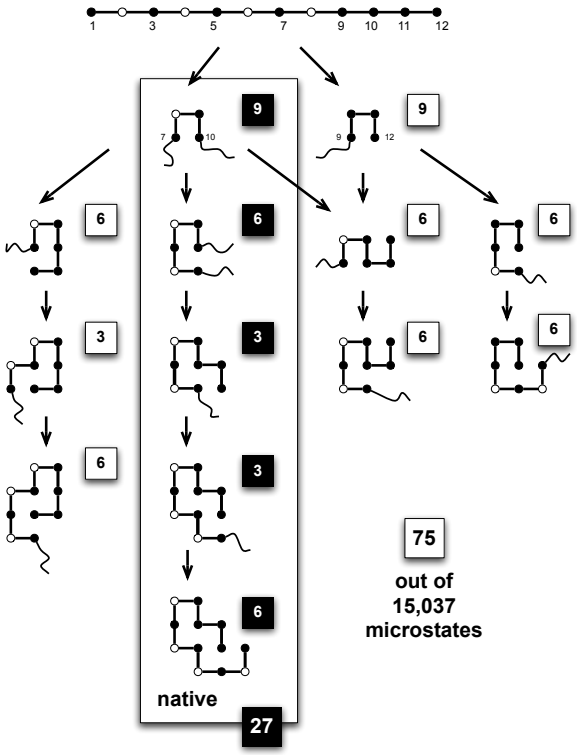

Figure 5: **A greedy strategy of low-ECO zipping can quickly lead to native structures.** For the hydrophobic/polar sequence above, the search is restricted to only contacts that can be made with ECO=3, the smallest loop possible in the lattice. First, consider the zipping sequence shown in the boxed area. After each iteration of the algorithm, the parent conformations are held fixed, and the chain is grown to find more contacts. The numbers in the black squares represent the number of new microstates (i.e. conformations) that are sampled at each step. For this particular pathway, the native state is found after 27 microstate enumerations, out of over 15,000 in the entire ensemble. Even in the worst-case scenario where the entire ECO 3 subgraph is searched, the maximum number of microstates that must be enumerated is 75, yielding a search efficiency of $SE = -\log_{10}(75/15037) \approx 2.3$, orders of magnitude faster than random search.

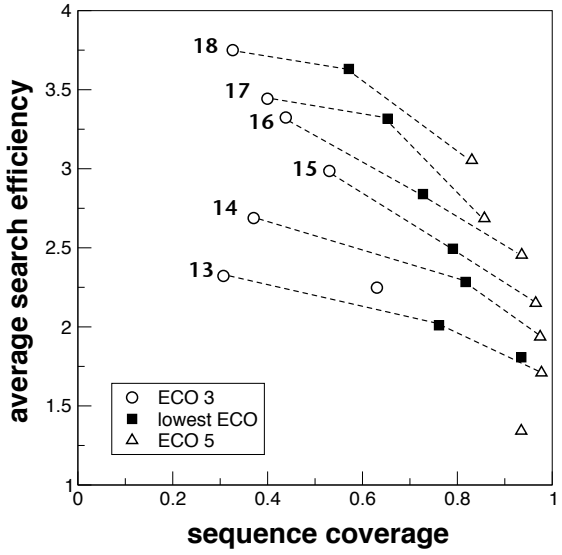

Figure 6: **The trade-off between search efficiency and sequence coverage, shown for various chain lengths (12-18) and low-ECO zipping strategies.** The global performance of the zipping and assembly algorithm across all foldable sequences displays an inverse relationship between the average search efficiency ($SE \equiv -\log_{10}(\Omega/\Omega_0)$) for zippable sequences and the fraction of foldable sequences which are zippable. This tradeoff is a hallmark of a greedy algorithm. The overall increase in search efficiency with larger chain length reflects that the average number of microstates that need to be explored to find the native state, Ω, grows at a slower rate than the conformational size of the full ensemble, Ω_0. The data points corresponding to the 12-mer are unlabeled, for clarity.

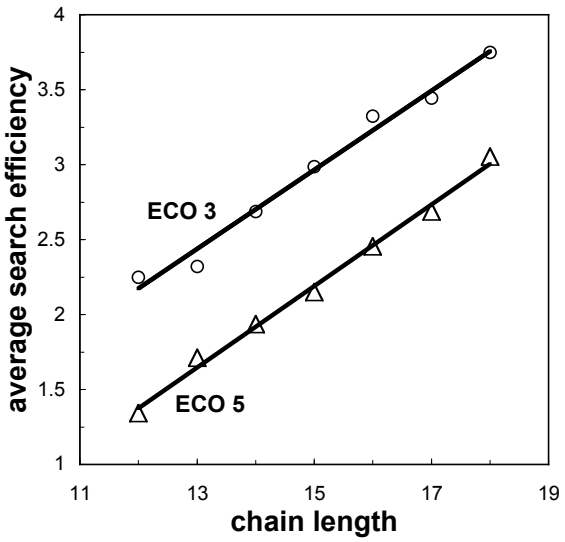

Figure 7: **Zipping and assembly effectively reduces the number degrees of freedom explored per unit chain length.** Shown here is the average search efficiency as a function of chain length, for *ECO-3* and *ECO-5* strategies, with lines showing the best least-squares-fit ($R^2 > 0.98$ in all cases). For the chain lengths we examined, the average search efficiency for all strategies increases linearly with chain length. That is, while the number of microstates in the ensemble grows exponentially with chain length as a^N, where $a \approx 2.7$, the number of microstates explored in zipping and assembly increases roughly as b^N, where $b \approx 1.4$. This reduced number for the effective number of degrees of freedom per unit chain length provides an intuitive explanation of how zipping and assembly can resolve Levinthal's paradox.

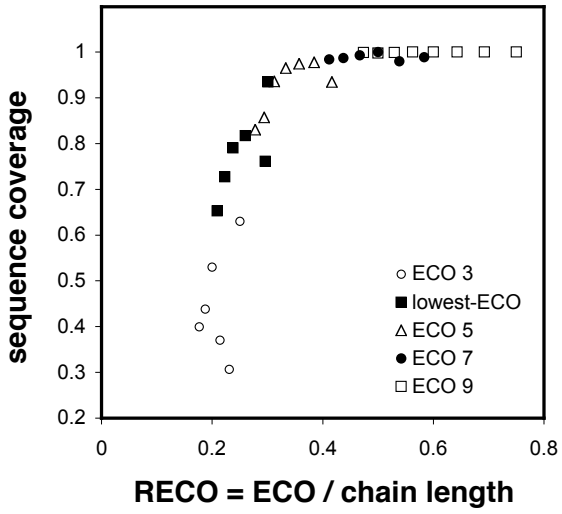

Figure 8: **A universal relationship between RECO cutoff and sequence coverage.** The sequence coverage data from the previous figure, along with data from ECO-7 and ECO-9 searches, when plotted as a function of the relative effective contact order (RECO), falls on a universal curve. Regardless of the zipping and assembly strategy, almost all foldable sequences in our model can be folded by zipping and assembly if the RECO is restricted to 30% of the chain length. For greedier searches, with even smaller local loop closures, less of the global foldable sequence space is zippable.

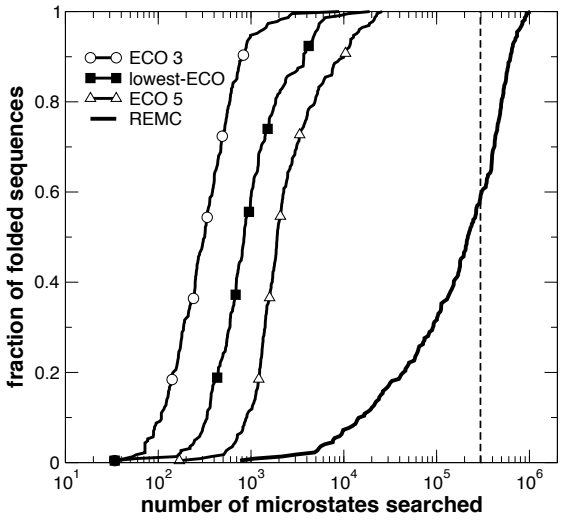

Figure 9: **Z&A is much more efficient than REMC in finding native states.** For the 228 ECO-3 zippable 15-mer HP sequences, the fraction of sequences that are folded by zipping and assembly in a given number of microstates is compared to the fraction of sequences folded in a corresponding replica exchange Monte Carlo (REMC) search. Zipping is consistently more efficient than REMC by about two orders of magnitude. The dotted line indicates the total number of 15-mer microstates (296806), as a reference.

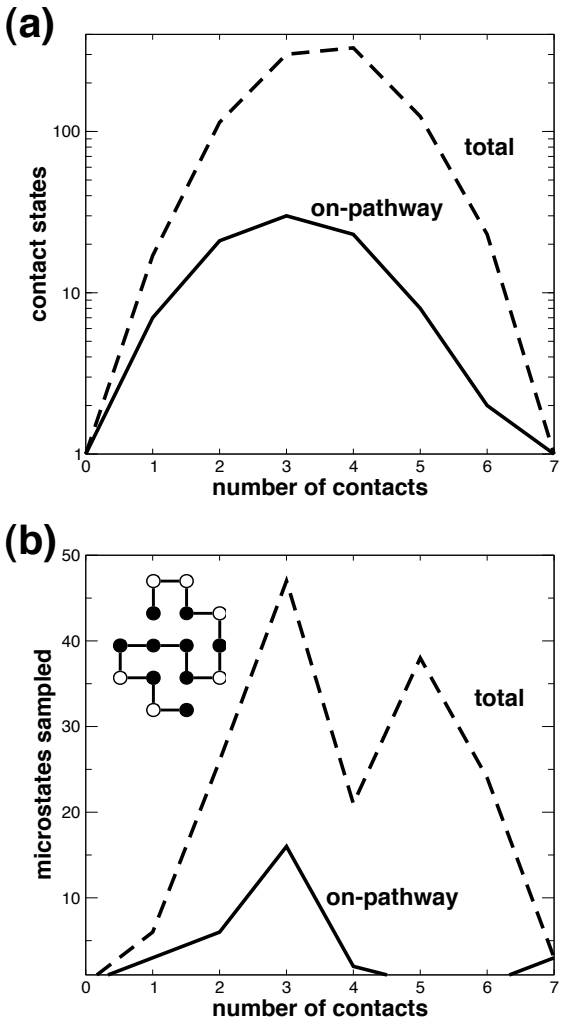

Figure 10: **A Z&A strategy has bottlenecks arising from topological frustration, a key feature of protein folding.** (a) The number of contact states as a function of the number of contacts is shown for a particular HP sequence. The total number of states is the dotted line, and the on-pathway states (contact states containing no non-native contacts) is the solid line. (b) The number of microstates sampled over the course of zipping and assembly as a function of the number of contacts shows that the algorithm spends most of its sampling in states with 3 contacts, closely coinciding with the maximum amount of topological frustration in the contact state graph.

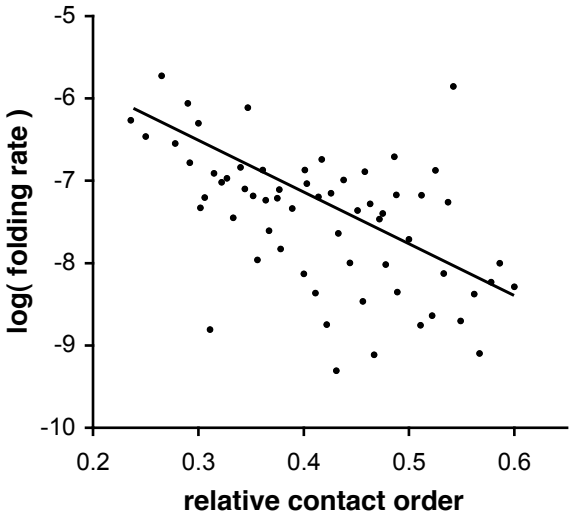

Figure 11: **Lowest-ECO Z&A shows topology-dependent folding rates, another key feature of protein folding.** Lowest-ECO searching for all zippable HP 18-mers shows that the logarithm of the folding rate (here defined as the inverse of the number of microstates searched) is proportional to the relative contact order (RCO) of the native structure. Each data point on the plot is the average log(folding rate) across all sequences sharing the same native topology. The line is to guide to eye.

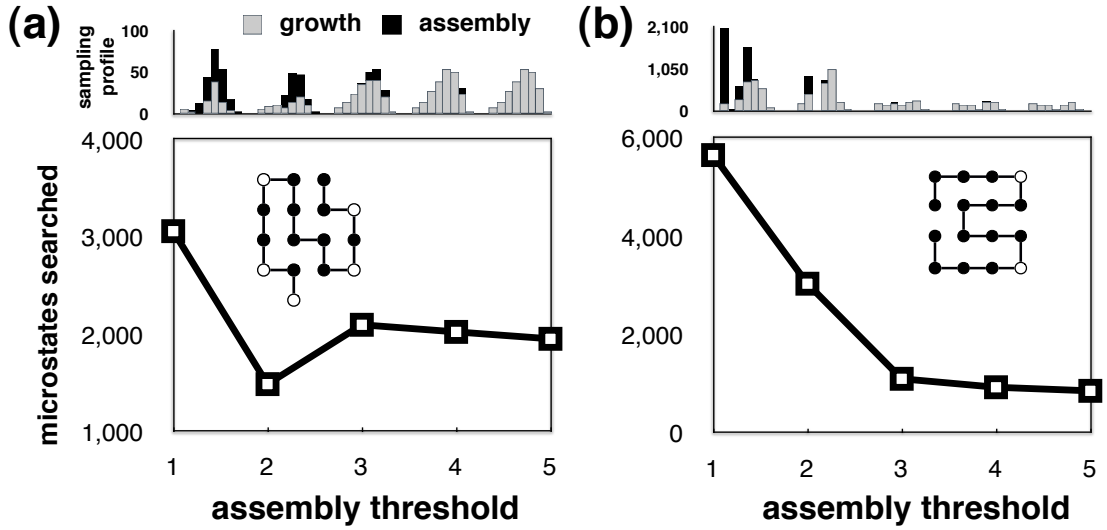

Figure 12: **Sequence-dependent optimal assembly thresholds require the right balance of growth and assembly.** ECO-3 zipping was performed on the 16-residue HP sequences shown, for various contact thresholds for assembly. The results for the sequence (a) show a "sweet spot" for efficient assembly: allowing assembly too early results in more searching than necessary, as does assembling too late. Sequence (b) prefers later assembly. The bar graphs at the top of each figure show a profile of the number of growth (gray) and assembly (black) assessments for each assembly threshold, along a reaction coordinate counting the number of contacts (from left to right).

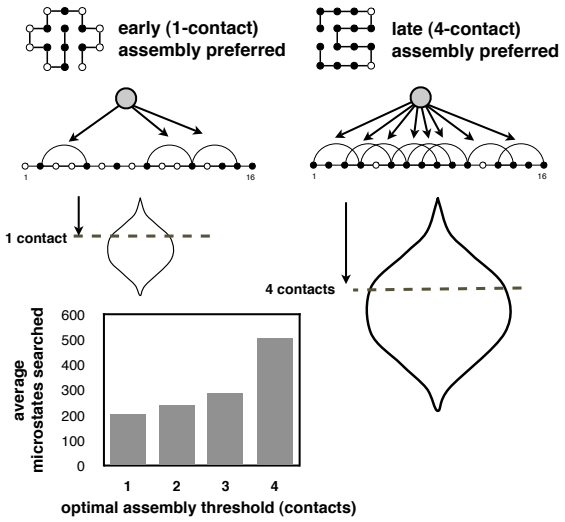

Figure 13: **Hierarchical assembly is a natural consequence of optimizing the efficiency of Z&A.** Sequences with few competing contact states (smaller contact state graphs) prefer early assembly, while sequences with many competing states (larger contact state graphs) prefer later assembly. Shown for example is a sequence that is folded most efficiently with an early (1-contact) assembly threshold (left) and sequence that is optimally assembled with a late (4-contact) assembly threshold (right). The polymer graphs shown below each sequence's native structure indicate the number of available lowest-ECO nucleation sites at the first step of zipping, an indication of the size of the contact state graph. A schematic diagram for each sequence shows that the optimal assembly threshold often comes before the contact state graph's widest point, where an assembly step may potentially bypass the most frustrated region. A survey of all zippable 16-mer HP sequences, binned by their optimal assembly threshold, shows early vs. late assembly depending on the amount of sampling needed.

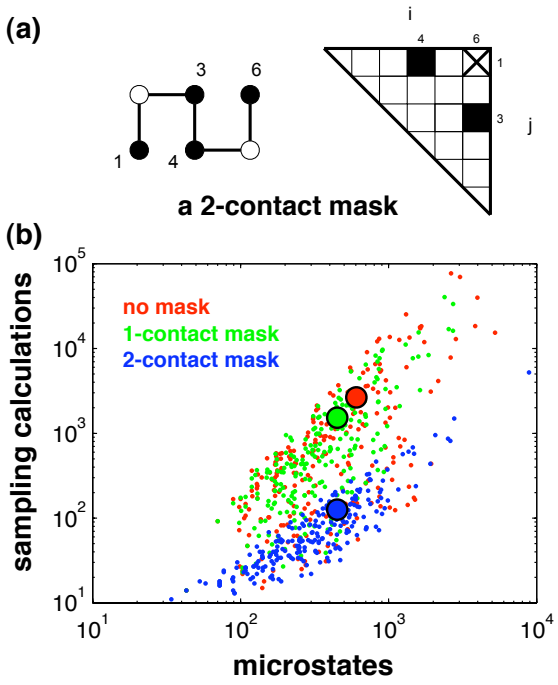

Figure 14: **Excluded volume "masks" help reduce dead-end calculations.** (a) An example of a 2-contact mask in the lattice model. If contacts (1,4) and (3,6) are already made, then (1,6) is not viable, as denoted by an 'X' on the contact map. (b) Masks reduce both the number of calculations that must be performed in the sampling stage of the algorithm, and the number of total microstates searched. The data shown is from an *ECO-3* search with HP 15-mers.