Exploring the lower part of discrete polymer model energy landscapes

Michael T. Wolfinger[1], Sebastian Will[2], Ivo L. Hofacker[1], Rolf Backofen[2] and Peter F. Stadler[3,4]

[1] *Institute for Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Wien, Austria*
[2] *Bioinformatics Group, Department of Computer Science, University of Freiburg, Georges-Köhler-Allee, Geb. 106, D-79110 Freiburg, Germany*
[3] *Bioinformatics Group, Department of Computer Science, University of Leipzig, Härtelstrasse 16-18, D-04107 Leipzig, Germany*
[4] *The Santa Fe Institute, 1399 Hype Park Rd., Santa Fe NM 87501*

PACS. 87.15.-v – Biomolecules: structure and physical properties.
PACS. 87.15.Aa – Theory and modeling; computer simulation.
PACS. 87.15.Cc – Folding and sequence analysis.

Abstract. – We present a generic, problem independent algorithm for exploration of the low-energy portion of the energy landscape of discrete systems and apply it to the energy landscape of lattice proteins. Starting from a set of optimal and near-optimal conformations derived from a constraint-based search technique, we are able to selectively investigate the lower part of lattice protein energy landscapes in two and three dimensions. This novel approach allows, in contrast to exhaustive enumeration, for an efficient calculation of optimal and near-optimal structures below a given energy threshold and is only limited by the available amount of memory. A straightforward application of the algorithm is calculation of barrier trees (representing the energy landscape), which then allows dynamics studies based on landscape theory.

Introduction. – The concept of energy landscapes has proven to be of fundamental relevance in investigations of complex disordered systems, from simple spin glass models to biopolymer folding. In this picture, energy is viewed as an explicit function $E(S)$ of underlying conformational degrees of freedom S. The topological structure of the conformation space is determined in terms of the elementary moves that underly the dynamical behavior. Examples are single spin flips in spin glasses, the formation or breaking of a base pair in RNA folding models, or rotation around a bond in a protein folding model.

The geometric properties and topological details of the energy landscape, such as number of local optima, the saddle points separating them, as well as the size distributions of the basins of attraction, therefore directly influence the dynamics of the underlying system. A thorough understanding of these aspects of geometrical landscape structure is thus of wide interest. Various attempts to elucidate the topological structure of landscapes, and in particular of their low-energy regions, have been developed independently and proposed for different contexts, among them $\pm J$ spin models [1], potential energy surfaces (PES) for protein folding [2] and molecular clusters [3], as well as the kinetics of RNA secondary structure formation [4]. An extensive study elucidating the energy landscape and dynamics of short two-dimensional lattice heteropolymers based on exhaustive enumeration, that characterizes energy landscapes in similar terms as we do, was given in [5]. However, we focus on larger and more complex systems, where full enumeration is out of reach. This requires to develop methods for selectively enumerating the (kinetically most important) low energy part of the conformation space.

The decomposition of energy landscapes into basins and saddle points separating them is straightforward for non-degenerate landscapes. However, the situation becomes more complicated if the landscapes are degenerate (e.g. in the lattice protein case). Consider a flat landscape. It is not a trivial task to decide which points are local minima or saddle points in flat-land, however a rigorous formalism to answer questions like this was given in [6].

Energy landscapes are conveniently visualized by "barrier trees" (Figure 1) that give a reasonable impression on the overall shape and topology of the landscape. Formally, three things are needed to construct an energy landscape [7]: a) a set $\mathcal{X}$ of configurations, b) a notion $\mathcal{M}$ of neighborhood, nearness, distance or accessibility on $\mathcal{X}$ and c) an energy function $f : \mathcal{X} \rightarrow \mathbf{R}$. The *conformation space* $\mathcal{X}$ of a (biopolymer) sequence is the total set of configurations S compatible with this sequence. The move set $\mathcal{M}$ is an order relation on $\mathcal{X}$, defining adjacency between the elements of $\mathcal{X}$. It crucially determines the topology of the energy landscape.

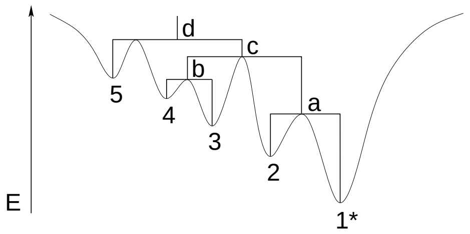


Figure 1 – Schematics representation of an energy landscape and its associated barrier tree. Local minima are labeled with numbers (1-5), saddle points with lowercase letters (a-d). The global minimum is marked with an asterisk.

Here, we consider lattice proteins. The aim of finding a structure x that minimizes the energy $E(\mathfrak{S}, x)$ for a certain sequence $\mathfrak{S}$ can be regarded as a combinatorial optimization problem, often termed *lattice protein folding problem.* In contrast to RNA, where efficient algorithms to determine the ground state exist [8], lattice protein folding was shown to be NP-complete, see e.g. [9]. However, there exists a fast and successful constraint-based approach to this problem, which we will use in this contribution [10]. This method, termed constraint-based protein structure prediction (CPSP), is the only available method that is capable of enumerating the ground state and near-optimal states of several three-dimensional lattice protein models completely. An efficient algorithm to generate the lower part of the density of states, like it was given for RNA [11], is not available for proteins. Nevertheless, several approximation algorithms have been proposed so far. A resource intensive genetic algorithm based on Monte Carlo techniques in the square lattice yields good results for fairly long chains up to a length of 60 monomers [12]. Further, the activation-relaxation technique (in combination with reduced off-lattice representations and a simple energy function) was successfully used to investigate the energy landscape of small peptides by starting from distinct low-energy conformations [13].

Description of the Method. – In this contribution, we present a generic, problem-independent approach for the exploration of the lower portion of energy landscapes. Generally, the energy function for a sequence with n residues $\mathfrak{S} = \mathfrak{s}_1\mathfrak{s}_2 \ldots \mathfrak{s}_n$ with $\mathfrak{s}_i \in \mathcal{A} = \{a_1, a_2, \ldots, a_b\}$, the alphabet of b residues, and an overall configuration $x = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ on a lattice $\mathcal{L}$ can be written as the sum of pair potentials. In the lattice models that we will consider in this contribution, this takes the form $E(\mathfrak{S}, x) = \sum_{i<j-1} \Psi[\mathfrak{s}_i, \mathfrak{s}_j]$ for $|\mathbf{x}_i - \mathbf{x}_j| = 1$. $\Psi[\mathfrak{s}, \mathfrak{s}']$ is called the *contact energy* for the monomers $\mathfrak{s}$ and $\mathfrak{s}'$. Within this contribution, we will use two different lattice protein models: First, one with a two-letter alphabet $\mathcal{A} = \{\mathbf{H}, \mathbf{P}\}$ where there is only one stabilizing interaction if, and only if hydrophobic residues ($\mathbf{H}$) are neighbors on the lattice but not along the chain. Polar residues ($\mathbf{P}$) do not explicitly contribute to the overall energy. Second, we will give an example for the four-letter **HPNX** model (see [14]) with alphabet $\mathcal{A} = \{\mathbf{H}, \mathbf{P}, \mathbf{N}, \mathbf{X}\}$ in three dimensions. The letters denote hydrophobic ($\mathbf{H}$), positive ($\mathbf{P}$), negative ($\mathbf{N}$) and neutral ($\mathbf{X}$) residues. This model extends the **HP** model by incorporating electrostatic interactions among polar residues.([1])

In addition, we assume a fixed move set giving rise to a symmetric neighborhood relation $\mathfrak{N} : \mathcal{X} \times \mathcal{X}$. A *walk* between two conformations x and y is a list of conformations $x = x_1 \dots x_{m+1} = y$ such that $\forall 1 \leq i \leq m : \mathfrak{N}(x_i, x_{i+1})$.

Given a threshold η, the lower part of the energy landscape (written as $\mathcal{X}^{\leq \eta}$) consists of *all* conformations x such that $E(\mathfrak{S}, x) \leq \eta$. For generating this lower part, a naive approach would exhaustively enumerate all conformations. However, this is only applicable to very short sequences because of the huge size the conformation space.

So we developed a method for investigating the lower part of the energy landscape selectively. This approach starts at low energy conformations and enumerates all "accessible" conformations. To exemplify the idea, for generating the lower part completely one starts with *all* local minima x with $E(\mathfrak{S}, x) \leq \eta$ (where x is a *local minimum* if for all y with $\mathfrak{N}(x, y)$ we have $E(\mathfrak{S}, y) \geq E(\mathfrak{S}, x)$). Iteratively, one visits all conformations that are neighbors of already seen conformations and stay below the energy threshold η.

According to [6], two conformations x and y are mutually accessible at the level η (written as $x \overset{\eta}{\longleftrightarrow} y$) if there is a walk from x to y such that all conformations z in the walk satisfy $E(\mathfrak{S}, z) \leq \eta$. The *saddle height* $\hat{f}(x, y)$ of x and y is defined by

$$\hat{f}(x, y) = \min\{\eta \mid x \overset{\eta}{\longleftrightarrow} y\}.$$

This gives rise to an ultrametric distance $d(x, y)$ between conformations x and y (see [15,16]). Given the set of all local minima $\mathcal{X}^{\leq \eta}_{\min}$ below threshold η, the lower energy part $\mathcal{X}^{\leq \eta}$ of the energy landscape can alternatively be written as

$$\mathcal{X}^{\leq \eta} = \{y \mid \exists x \in \mathcal{X}^{\leq \eta}_{\min} : \hat{f}(x, y) \leq \eta\}.$$

Of course, one does not have the complete set of local minima $\mathcal{X}^{\leq \eta}_{\min}$ as starting point of the construction in many practical applications. In this case, one can hope to enumerate a large part of the low energy conformations by starting from a restricted set of low energy conformations $\mathcal{X}_{\text{init}}$. In our application to the three-dimensional **HPNX** model, we use the method described in [10] for computing a set of excellent start conformations. This method, CPSP, is based on constraint optimization. Given a **HPNX**-sequence with n_t monomers of type $t = \mathbf{H}, \mathbf{P}, \mathbf{N}, \mathbf{X}$, it starts by enumerating all maximally compact hydrophobic cores of size $n_{\mathbf{H}}$. Then, for every hydrophobic core, all possible threadings of the sequence onto the selected hydrophobic core are generated. Since the maximally compact hydrophobic cores give only optimality with respect to the hydrophobic part of the energy function, we enumerate sub-optimal hydrophobic cores as well. Here, we can bound the degree of suboptimality by the maximal number of $\mathbf{N} - \mathbf{P}$ contacts for the given sequence, which is $\min(n_{\mathbf{P}}, n_{\mathbf{N}})$. Using this

([1])The contact energies $\Psi[\mathfrak{s}, \mathfrak{s}']$ for two neighboring $\mathbf{H}$'s in the **HP** model is -1. For the **HPNX** model, the contact energies for **H–H**, **N–P**, **P–P** and **N–N** are -4, -1, $+1$, and $+1$, respectively. All other contacts have energy 0.

approach, the method is able to enumerate the ground state and near-optimal states of three-dimensional lattice proteins in **HP**-type models completely. The method was successfully applied to complete enumeration of optimal conformations in the cubic lattice up to sequence length 48 and predicts optimal conformations up to length 300 in the face-centered cubic lattice. By using such states as start conformations we guarantee to cover the very lowest part of the energy spectrum. Since furthermore the method provably predicts all ground states, it can be used to identify sequences with unique ground state. This allows us to find good candidate sequences for further studies; we provide an example for one such sequence later.

Operationally, the lower part of the energy landscape can be generated using a fixpoint of a monotone operator that successively adds neighbors whose energy is below the threshold η. Given a set $\mathcal{X}$ of conformations whose energies are lower than η, then $\mathfrak{F}^{\leq\eta}(\mathcal{X})$ is defined as the following set of conformations:

$$\mathfrak{F}^{\leq\eta}(\mathcal{X}) = \{y \mid E(\mathfrak{S}, y) \leq \eta \wedge \exists x \in \mathcal{X} : \mathfrak{N}(x, y)\} \cup \mathcal{X}$$

It is easy to see that $\mathfrak{F}^{\leq\eta}$ is a monotone operator, and that the fixpoint

$$\bigcup_{n=1}^{\infty} (\mathfrak{F}^{\leq\eta})^n (\mathcal{X}_{\mathrm{min}}^{\leq\eta})$$

of applying $\mathfrak{F}^{\leq\eta}$ to $\mathcal{X}_{\mathrm{min}}^{\leq\eta}$ is the lower part $\mathcal{X}^{\leq\eta}$ of the energy landscape.

This operator can now be implemented efficiently. We denote the initial set $\mathcal{X}$ by $\mathcal{X}^0$, and define $\mathcal{X}^i$ for $i > 0$ to be

$$(\mathfrak{F}^{\leq\eta})^i(\mathcal{X}^0) = \mathfrak{F}^{\leq\eta}(\mathcal{X}^{i-1}).$$

The current set $\mathcal{X}^i$ of conformations is represented by a hash table. Note that by definition of the operator, the set $\mathcal{X}^i$ contains already all neighbors of $\mathcal{X}^{i-1}$. Hence, we need to consider in the step $\mathcal{X}^i \to \mathcal{X}^{i+1}$ only the conformations in the set $\mathcal{X}^i \backslash \mathcal{X}^{i-1}$, which is represented as a list of pointers to hash entries. For each single conformation x in the hash pointer list for $\mathcal{X}^i \backslash \mathcal{X}^{i-1}$, all neighbor conformations of x are generated. Given that its energy is below the threshold η, the hash table is used to determine for each neighbor of x if it has already been seen before. If this is true, the structure is skipped and the next structure is processed. Otherwise, it is inserted into the hash table and a pointer to the entry is put into a new hash pointer list. After all conformations from $\mathcal{X}^i \backslash \mathcal{X}^{i-1}$ are processed, the new hash pointer list replaces the previous one, and the next round is started. The end of the algorithm is reached as soon as a) a predefined amount of structures has been found(2) or b) all structures that are "reachable" from a distinct start-structure (constrained to an energy threshold) are found.

We have applied the algorithm to lattice proteins. However, the algorithm is readily applicable to any kind of discrete system, such as spin glass models or RNA secondary structures. The algorithm presented here does not - in contrast to previously mentioned algorithms - aim at finding *distinct* low-energy minima. It is rather designed to generate the whole low-energy portion of energy landscapes including the ground state(s) as well as suboptimal structures in order to enable dynamics studies based on landscape theory. An efficient approach for biopolymer folding dynamics calculations incorporating the energy landscape framework was given recently [17]. The main advantage of this approach, compared to e.g. exhaustive enumeration, is time efficiency as well as the possibility to explore certain regions of the energy landscape. It would thus be possible to selectively investigate a high-energy portion of the landscape. The algorithm is only limited by the available amount of memory.

---

(2)To given an impression of limitations by the size of RAM, currently approximately 85 million structures can be generated on machines with 4GB RAM.

Move Sets. – For the purpose of this contribution we will rely on a simple, yet efficient move whose ergodicity was proven for the simple (hyper)cubic lattice [18], called pivot move. This move set is N-conserving, i.e. the total number of beads along the chain is preserved. Pivot moves are non-local in a sense that the positions of a large fraction of beads along the chain are changed by one elementary step. Alternatively, a local move set consisting of crankshaft-, corner-, and end moves could has been implemented. Local move sets, altering only a few consecutive beads of the chain and leaving all other sites unchanged allow the chain to exhibit more fine-grained structural transitions. However, it was shown that every local, N-conserving move set is non-ergodic on simple (hyper)cubic lattices for sufficiently large N [19].

Results. – To illustrate the capabilities (and limitations) of this new approach we give two examples of lattice heteropolymer energy landscapes here. The first one is a 31-mer with the sequence HHHHHHHPPHPPHHPPHHPHPHPPHHPPHPH on the two-dimensional square lattice. Starting from a conformation with an energy of -16 (middle structure at the bottom of Figure 2) and an upper energy threshold of -10, we found a total of 22985151 conformations that are related to the start structure by means of pivot moves.

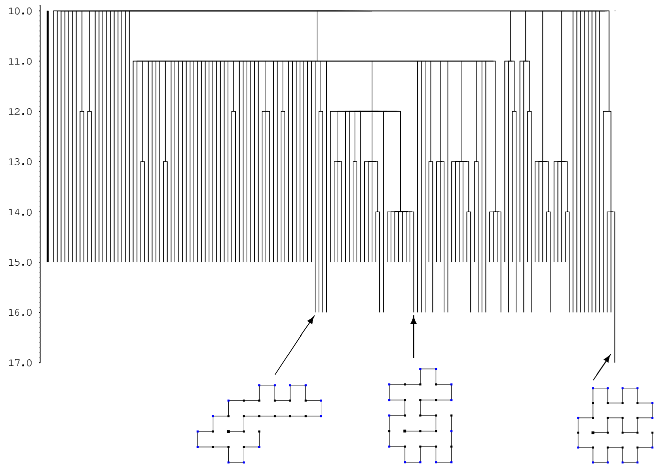
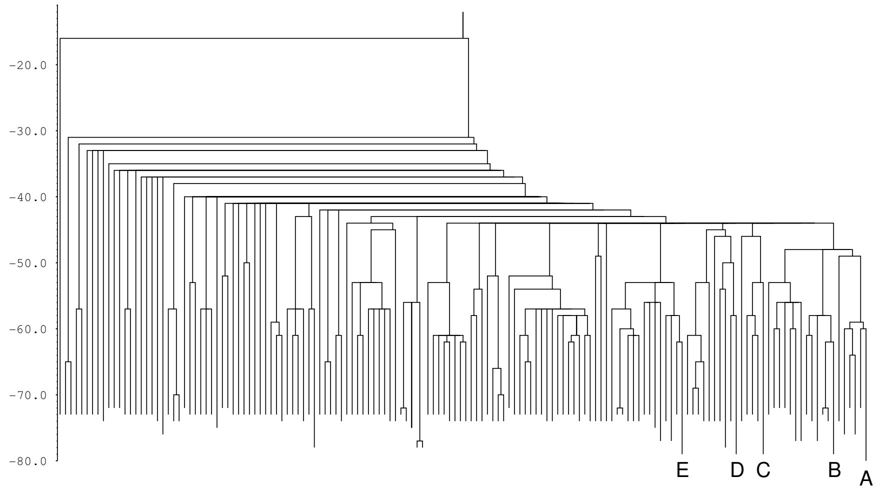


Figure 2 – Barrier tree of an 31-mer **HP**-kind lattice protein showing the 150 lowest lying minima of the energy landscape. Structures that can be inter-converted by symmetry operations (such as reflections) were not considered for calculation of the tree. A ground state (17 contacts) as well as two near-optimal conformations with 16 contacts are illustrated below. Note that there is one local minimum at the very left of the plot that is not attached to the rest of the tree, but there is a direct path connecting this minimum to the ground state with saddle height of $E = -6$.

The barrier tree in Figure 2 exhibits common features of lattice protein energy landscapes such as a high degree of degeneracy (i.e., there are many conformation having exactly the same energy). There are 35 minima with $E = -16$ and 114 minima with $E = -15$. Degeneracy can be seen as an artefact of the underlying model here, i.e. bond lengths/angles are fixed and the alphabet consists of only two letters. It is striking that many of the near optimal conformations are connected to the global optimum via a high energy barrier. This is due to the low connectivity (i.e., number of neighbors) of the two-dimensional lattice. At this point, it seems fair to ask whether it is correct to model a complex protein with such a coarse-grained model.

a)



b)

Conformation	Absolute move	Energy	Barrier
A	FFUUBBDLFRRRBLDRFLFUUBBRFF	-80	-
B	FFUUBDBBUFLLDRDLFRFUUBDLUF	-79	31
C	FUBBDDFLFRRRBLURFLUBBDDRUU	-79	35
D	FUBBDDFDFULLBRULFRUBBDDLUU	-79	35
E	FFUUBDBLURRRDLDRFLFUUBDRUF	-79	35

Figure 3 – Energy landscape for the sequence HHXHPHHHNPHHPHHHHNHPHNHHHNP; a) barrier tree generated by our method, using the ground state and the first excited states conformations as start set; b) conformations of the start set. Absolute moves are: F (forward), L (left), R (right), U (up), D (down).

In the second example, we used a **HPNX** sequence in the three-dimensional cubic lattice. Applying CPSP, we were able to prove that the 27-mer sequence HHXHPHHHNPHHPHHHHNHPHN-HHHNP has a unique ground state. In addition, we used CPSP to find all conformations on the first exited energy level. The resulting barrier tree for the lower part of the energy landscape is given in Figure 3. Although the tree still shows lattice protein artifacts like high degeneracy, the near optimal conformations are highly connected via low energy barriers. This more pronounced connectivity makes the energy landscape similar to those found in biologically relevant and well studied systems such as RNA [17]. This is in contrast to the two-dimensional case, and is due to the use of the three-dimensional grid *as well as* the extended energy model (**HPNX**).

Conclusion and Discussion. – We have designed a method for generating the lower portion of the energy landscape of discrete models of biopolymers, given a starting set of low energy conformations. Using this method, we are able to calculate the barrier tree representing topological details of the energy landscape such as local minima, basin sizes and barrier heights. This information can readily be used to study a coarse grained dynamics [17].

We have applied the method to lattice protein models in two and three dimensions. We combined it with the constraint-based protein structure prediction method (CPSP [10]), which allowed us to completely explore the low energy part in three dimensions for the first time. The experiments indicate that using three dimensions and a four letter alphabet yield results that are in good agreement with well-studied biological systems, such as RNA.

In addition, this is in good accordance with experimental as well as theoretical studies that have shown that the full sequence complexity of naturally occurring proteins is not necessarily required to design a functional, rapidly folding protein (see e.g. [20] and references therein). Proteins with a drastically reduced set of amino acids (compared to the 20 naturally occurring ones) have been successfully designed experimentally in the last years. Wang and Wang proposed an algorithm to systematically select reduced alphabets [21]. One of the optimally reduced sets they predicted was the five-letter IKEAG alphabet. A later study proposed the lower bound of amino acid types required for a protein to fold into a stable structure to be around ten [22]. The size of the alphabet influences both *foldability* [23] and *designability* (that is, the number of sequences that have the prescribed structure as their unique lowest-energy state) of structures [24]. Detailed computational investigations into both foldability and designability in lattice protein models are dependent upon an efficient method for generating and analyzing the low-energy states. The lattice flooder approach presented here sets the stage for such a research program.

* * *

Thanks to Christoph Flamm for stimulating and useful discussions. This work was supported in part by the EMBIO project in FP-6 (`http://www-embio.ch.cam.ac.uk/`) and the DFG Bioinformatics Initiative (BIZ-6/1-2).

References

[1] T. Klotz and S. Kobe. *J. Phys. A: Math. Gen*, 27:L95–L100, 1994.
[2] P. Garstecki, T. X. Hoang, and M. Cieplak. *Phys. Rev. E*, 60:3219–3226, 1999.
[3] J. P. Doye, M. A. Miller, and D. J. Welsh. *J. Chem. Phys.*, 111:8417–8429, 1999.
[4] C. Flamm, W. Fontana, I. Hofacker, and P. Schuster. *RNA*, 6:325–338, 2000.
[5] H. S. Chan and K. Dill. *J. Chem. Phys.*, 100:9238–9257, 1994.
[6] C. Flamm, I. L. Hofacker, P. F. Stadler, and M. T. Wolfinger. *Z.Phys.Chem*, 216:155–173, 2002.
[7] C. M. Reidys and P. F. Stadler. *SIAM Review*, 44:3–54, 2002.
[8] M. Zuker and P. Stiegler. *Nucl. Acids Res.*, 9:133–148, 1981.
[9] B. Berger and T. Leighton. *J. Comput. Biol.*, 5:27–40, 1998.
[10] R. Backofen and S. Will. *Journal of Constraints*, 11(1), 2006.
[11] S. Wuchty, W. Fontana, I. L. Hofacker, and P. Schuster. *Biopolymers*, 49:145–165, 1998.
[12] R. Unger and J. Moult. *J. Mol. Biol.*, 231:75–81, 1993.
[13] G. Wei, N. Mousseau, and P. Derreumaux. *J. Chem. Phys.*, 117:11379–11387, 2002.
[14] R. Backofen, S. Will, and E. Bornberg-Bauer. *Bioinformatics*, 15(3):234–242, 1999.
[15] A. M. Vertechi and M. A. Virasoro. *J. Phys. France*, 50:2325–2332, 1989.
[16] R. Rammal, G. Toulouse, and M. A. Virasoro. *Rev. Mod. Phys.*, 58:765–788, 1986.
[17] M. T. Wolfinger, W. A. Svrcek-Seiler, C. Flamm, I. L. Hofacker, and P. F. Stadler. *J. Phys. A: Math. Gen.*, 37:4731–4741, 2004.
[18] N. Madras and A. D. Sokal. *J. Stat. Phys.*, 50:109–189, 1988.
[19] N. Madras and A. D. Sokal. *J. Stat. Phys.*, 47:573–595, 1987.
[20] H. S. Chan. *Nature Struct. Biol.*, 6(11):994–996, 1999.
[21] J. Wang and W. Wang. *Nature Struct. Bio.*, 6:1033–1038, 1999.
[22] K. Fan and W. Wang. *J. Mol. Biol.*, 328:921–926, 2003.
[23] S. Govindarajan and R. A. Goldstein. *Proc. Natl. Acad. Sci. USA*, 93:3341–3345, 1996.
[24] R. Wroe, E. Bornberg-Bauer, and H. S. Chan. *Biophys J*, 88(1):118–31, 2005.