Run-time Estimates for Protein Folding Simulation in the H-P Model

A.A. Albrecht[1] and K. Steinhöfel[2]

[1] School of Computer Science
University of Hertfordshire
Hatfield, Herts AL10 9AB, UK
[2] Dept. of Computer Science
King's College London
Strand, London WC2R 2LS, UK

Abstract. The hydrophobic-hydrophilic (H-P) model for protein folding was introduced by Dill et al. [6]. A problem instance consists of a sequence of amino acids, each labeled as either hydrophobic (H) or hydrophilic (P). The sequence must be placed on a 2D or 3D grid without overlapping, so that adjacent amino acids in the sequence remain adjacent in the grid. The goal is to minimize the energy, which in the simplest variation corresponds to maximizing the number of adjacent hydrophobic pairs. Although the model is extremely simple, it captures the main features of the problem. The protein folding problem in the H-P model is NP-hard in both 2D and 3D. Recently, Fu and Wang [9] proved an $exp(O(n^{1-1/d}) \cdot \ln n)$ algorithm for d-dimensional protein folding simulation in the HP-model. Our preliminary results on stochastic search applied to protein folding utilize complete move sets proposed by Lesh et al. [15] and Blazewicz et al. [3]. We obtain that after $(n/\delta)^{c \cdot \Gamma}$ Markov chain transitions, the probability to be in a minimum energy conformation is at least $1 - \delta$, where n is the length of the instance, Γ is the maximum value of the minimum escape height from local minima of the underlying energy landscape, and c is a (small) constant. Γ depends on the move sets, and future research will focus on upper bounds of this value. To be competitive with the Fu/Wang bound, $\Gamma \leq n^{1-1/d}$ is required. From previous experiments on stochastic search applied to computationally hard problems [20] we expect Γ to be much smaller for real protein sequences.

1 Introduction

The protein folding problem is one of the most challenging problems in current biochemistry and is a very rich source of interesting problems in mathematical modelling and computational mathematics. Proteins are complex biological macromolecules that are composed of a sequence of amino acids, which is encoded by a gene in a genome [6, 11, 17]. Proteins mediate virtually all cellular functions. There are 20 different amino acids specified in the genetic code. Amino acids are joined end-to-end during protein synthesis by the formation of peptide bonds. The functional properties of proteins depend upon their three-dimensional structures. Unlike the structure of other biological macromolecules (e.g., DNA), proteins have complex structures that are difficult to predict.

The linear sequence of residues in a protein is called its primary structure. The smallest proteins, peptide-hormones, have about 25–100 residues, typical globular proteins about 100–500, and fibrous proteins may have more than 3000 residues [17].

One of the basic paradigms of structural proteomics is given by Anfinsen's thermodynamic hypothesis [1]: Proteins fold to a minimum energy state, and the information determining the three-dimensional structure (tertiary structure) of a protein resides in the chemistry of its primary structure (cf. [11, 21] for discussion of restrictions).

Proteins exhibit a variety of so-called secondary structure motifs that reflect common structural elements in a local region of the primary structure, such as α-helices and β-strands. Groups of secondary structures usually combine to form compact structures, which represent the tertiary structure of an entire protein. The key problem is to understand the entire folding pathway, i.e., the complete dynamics and chemical changes involved in going from an unfolded linear state into a compact folded state (tertiary structure).

The protein folding problem can be naturally posed as a numerical simulation, but there are several problems of scale, including the small energy differences between folded and unfolded states, and the extremely short interval (approximately 10–15 seconds) for which the dynamics equations remain valid, compared to the microseconds to milliseconds over which the folding takes place. The thermodynamic hypothesis [1] motivates the attempt to predict protein folding by solving certain optimization problems, but there are two main difficulties with this approach: The precise definition of the energy function that has to be minimised, and the extremely difficult optimization problems arising from the energy functions commonly used in folding simulations [11, 17].

A great variety of models has been developed for protein folding simulations, with different levels of detail (for a concise discussion, cf. [21]). In the present paper, we focus on *minimal models* [11, 21], and we distinguish roughly between lattice models [6] and off-lattice models [7, 17]. For a discussion of energy functions and justifications for the use of simplified (approximated) energy functions we refer the reader to [21]. One of the most popular models of protein folding is the hydrophobic-hydrophilic (H-P) model [6]. In the H-P model, proteins are modelled as chains whose vertices are marked either H (hydrophobic) or P (hydrophilic); the resulting chain is embedded in some lattice. H nodes are considered to attract each other while P nodes are neutral. An optimal embedding is one that maximizes the number of H-H contacts. The rationale for this objective is that hydrophobic interactions contribute a significant portion of the total energy function. Roughly, this objective favours conformations that have the hydrophobic amino acid residues clustered on the inside, covered by the hydrophilic ones. Unlike more sophisticated models of protein folding, the main goal of the H-P model is to explore broad qualitative questions about protein folding such as whether the dominant interactions are local or global with respect to the chain.

Lattice models of protein folding have provided valuable insights into the general complexity of protein structure prediction problems. For example, protein structure prediction has been shown to be NP-hard for a variety of lattice models [2, 11, 16]. The intractability results are complemented by performance guaranteed approximation algorithms that run in linear time [11, 13]. These results can be generalized to simple off-lattice protein models. However, these approximation algorithms have not proven helpful for finding minimum energy conformations. Since protein structure prediction is NP-hard, (local) search-based algorithms are a natural choice to tackle the problem, especially in lattice models; cf. literature in [11]. Lesh et al. [15] and Blazewicz et al. [3] proposed complete neighbourhood move sets for local search in 2D and 3D grids, respectively, and performed computational experiments on benchmark problems for protein folding in the H-P model. Recently, Fu and Wang [9] proved an $exp(O(n^{1-1/d}) \cdot \ln n)$ algorithm for d-dimensional protein folding simulation in the HP-model.

The present paper reports our preliminary results on stochastic search applied to protein folding in the H-P model. We utilize the complete move sets proposed in [15] and [3]. We obtain that after $(n/\delta)^{c \cdot \Gamma}$ Markov chain transitions, the probability to be in a

minimum energy conformation is at least $1 - \delta$, where n is the length of the instance, Γ is the maximum value of the minimum escape height from local minima of the underlying energy landscape, and c is a relatively small constant. Thus, to be competitive with the Fu/Wang run-time bound, we need to show $\Gamma \leq n^{1-1/d}$. However, based on our previous experiments on stochastic search applied to computationally hard problems [20] we expect Γ to be much smaller for real protein sequences. Future research will focus on proven upper bounds of Γ in the context of complete move sets for the H-P model, and on computational experiments on protein folding benchmark problems [3]. Furthermore, since local search methods have been proved to be useful in many applications, we plan to explore the applicability of local search methods developed in diverse areas of combinatorial optimization. For example, WalkSAT and related methods have been successfully used to solve SAT instances [10, 19]. Therefore, we intend to investigate if methods like WalkSAT are applicable to the protein folding problem in the lattice model; on the other hand, we also plan to study the performance of our logarithmic simulated annealing-based search on SAT instances.

2 Preliminaries

Our stochastic local search procedure for protein folding is based on simulated annealing [5, 14], where the underlying Markov chain is of inhomogeneous type [4, 12]. Simulated annealing algorithms are acting within a configuration space in accordance with a certain neighbourhood relation, usually of polynomial size. The particular transitions between adjacent elements of the configuration space are governed by an objective function. For simplicity of presentation, we focus on the 2D rectangular grid H-P model only.

According to Anfinsen's thermodynamic hypothesis [1], the problem of finding the native protein structure can be defined as an energy function minimization problem. In the 2D rectangular grid H-P model, one can define the minimization problem as follows:

$$\min_{\alpha} E(S, \alpha) \quad \text{for} \quad E(S, \alpha) := \xi \cdot HH_c(S, \alpha), \tag{1}$$

where where S is a sequence of amino acids containing n elements; $S_i = 1$, if amino acid on the i^{th} position in the sequence is hydrophobic; $S_i = 0$, if amino acid on the i^{th} position is polar; α is a vector of $(n-2)$ grid angles defined by consecutive triples of amino acids in the sequence; HH_c is a function that counts the number of neighbours between amino acids that are not neighbours in the sequence, but they are neighbours on the grid (they are topological neighbours); finally, $\xi < 0$ is a constant lower than zero that defines an influence ratio of hydrophobic contacts on the value of conformational free energy. The distances between neighbouring grid nodes is assumed to be equal to 1. We identify sequences α with conformations of the protein sequence S, and a valid conformation α of the chain S lies along a non-self-intersecting path of the rectangular grid such that adjacent vertices of the chain S occupy adjacent locations. Thus, we define the set of conformations (for each S specifically) by

$$\mathcal{F}_S := \{\, \alpha \ \text{ is a valid conformation for } \ S \,\}. \tag{2}$$

Since $\mathcal{F} := \mathcal{F}_S$ is defined for a specific S, we denote the objective function by

$$\mathcal{Z}(\alpha) \ := \ \xi \cdot HH_c(S, \alpha). \tag{3}$$

The neighbourhood relation of our stochastic local search procedure is determined by the set of *pull moves* introduced in [15] for 2D protein folding simulations in the H-P model (and, basically, extended to the 3D case in [3]). For details of the definition of the set of pull moves we refer the reader to [15].

Theorem 1 [15] *The set of pull moves is local, reversible, and complete within $\mathcal{F}$, i.e. any $\beta \in \mathcal{F}$ can be reached from any $\alpha \in \mathcal{F}$ by executing pull moves only.*

The set of neighbours of α that can be reached by a single pull move is denoted by $\mathcal{N}_\alpha$, where additionally α is included since the search process can remain in the same configuration. Furthermore, we set

$$N_\alpha := |\mathcal{N}_\alpha|; \tag{4}$$

$$\mathcal{F}_{\min} := \{\alpha : \alpha \in \mathcal{F} \text{ and } \mathcal{Z}(\alpha) = \min_{\alpha'} E(S, \alpha')\}. \tag{5}$$

In simulated annealing-based search, the transitions between neighbouring elements are depending on the objective function $\mathcal{Z}$. Given a pair of protein conformations $[\alpha, \alpha']$, we denote by $G[\alpha,\, \alpha']$ the probability of generating α' from α, and by $A[\alpha, \alpha']$ we denote the probability of accepting α' once it has been generated from α. As in most applications of simulated annealing, we take a uniform generation probability:

$$G[\alpha,\, \alpha'] := \begin{cases} \frac{1}{N_\alpha}, & \text{if } \alpha' \in \mathcal{N}_\alpha; \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

The acceptance probabilities $A[\alpha,\, \alpha']$ are derived from the underlying analogy to thermodynamic systems:

$$A[\alpha, \alpha'] := \begin{cases} 1, & \text{if } \mathcal{Z}(\alpha') - \mathcal{Z}(\alpha) \leq 0; \\ e^{-\frac{\mathcal{Z}(\alpha') - \mathcal{Z}(\alpha)}{t}}, & \text{otherwise,} \end{cases} \tag{7}$$

where t is a control parameter having the interpretation of a *temperature* in annealing processes. The probability of performing the transition between α and α' is defined by

$$\mathbf{Pr}\{\alpha \to \alpha'\} = \begin{cases} G[\alpha,\, \alpha'] \cdot A[\alpha,\, \alpha'], & \text{if } \alpha' \neq \alpha; \\ 1 - \sum\limits_{\alpha' \neq \alpha} G[\alpha,\, \alpha'] \cdot A[\alpha,\, \alpha'], & \text{otherwise.} \end{cases} \tag{8}$$

By definition, the probability $\mathbf{Pr}\{\alpha \to \alpha'\}$ depends on the control parameter t. Let $\mathbf{a}_\alpha(k)$ denote the probability of being in conformation α after k transition steps. The probability $\mathbf{a}_\alpha(k)$ is calculated in accordance with

$$\mathbf{a}_\alpha(k) := \sum_{\beta \in \mathcal{F}} \mathbf{a}_\beta(k-1) \cdot \mathbf{Pr}\{\beta \to \alpha\}. \tag{9}$$

The recursive application of (9) defines a Markov chain of probabilities $\mathbf{a}_\alpha(k)$, where $\alpha \in \mathcal{F}$ and $k = 1, 2, \ldots$. If the parameter $t = t(k)$ is a constant t, the chain is said to be a *homogeneous* Markov chain; otherwise, if $t(k)$ is lowered at any step, the sequence of probability vectors $\vec{\mathbf{a}}(k)$ is an *inhomogeneous* Markov chain.

In the present paper we are focusing on a special type of inhomogeneous Markov chains where the value $t(k)$ changes in accordance with

$$t(k) = \frac{\Gamma}{\ln(k+2)},\ k = 0, 1, \ldots. \tag{10}$$

The choice of $t(k)$ is motivated by Hajek's Theorem on logarithmic cooling schedules for inhomogeneous Markov chains [12]. To explain Hajek's result, we first need to introduce some parameters characterising local minima of the objective function:

Definition 1 *A conformation $\alpha' \in \mathcal{F}$ is said to be reachable at height h from $\alpha \in \mathcal{F}$, if $\exists \alpha_0, \alpha_1, ..., \alpha_r \in \mathcal{F}$ with $\alpha_0 = \alpha \wedge \alpha_r = \alpha'$ such that $G[\alpha_u, \alpha_{u+1}] > 0$, $u = 0, 1, \ldots, (r-1)$, and $\mathcal{Z}(\alpha_u) \leq h$ for all $u = 0, 1, \ldots, r$.*

We use the notation $H(\alpha \Rightarrow \alpha') \leq h$ for this property. The conformation α is a *local minimum*, if $\alpha \in \mathcal{F} \backslash \mathcal{F}_{\min}$ and $\mathcal{Z}(\alpha') \geq \mathcal{Z}(\alpha)$ for all $\alpha' \in \mathcal{N}_\alpha \backslash \{\alpha\}$.

Definition 2 *Let $\lambda_{\min}$ denote a local minimum, then $D(\lambda_{\min})$ denotes the smallest h such that there exists $\lambda' \in \mathcal{F}$ with $\mathcal{Z}(\lambda') < \mathcal{Z}(\lambda_{\min})$ that is reachable at height $\mathcal{Z}(\lambda_{\min}) + h$.*

The following convergence property has been proved by B. Hajek:

Theorem 2 [12] *For $t(k)$ from (10), the asymptotic convergence $\sum_{\alpha \in \mathcal{F}_{\min}} \mathbf{a}_\alpha(k) \underset{k \to \infty}{\longrightarrow} 1$ of the algorithm defined by (3), ..., (9) is guaranteed if and only if*

1. *$\forall \alpha, \alpha' \in \mathcal{F}\, \exists \alpha_0, \alpha_1, \ldots, \alpha_r \in \mathcal{F}$ such that $\alpha_0 = \alpha \wedge \alpha_r = \alpha'$ and $G[\alpha_u, \alpha_{u+1}] > 0$ for $u = 0, 1, \ldots, (r-1)$;*
2. *$\forall h : H(\alpha \Rightarrow \alpha') \leq h \iff H(\alpha' \Rightarrow \alpha) \leq h$;*
3. *$\Gamma \geq \max\limits_{\lambda_{\min}} D(\lambda_{\min})$.*

From Theorem 1 and the definition of $\mathcal{N}_\alpha$ we immediately conclude that the conditions (i) and (ii) are valid for $\mathcal{F}$. Thus, together with Theorem 2 we obtain:

Corollary 1 *If $\Gamma \geq \max_{\lambda_{\min}} D(\lambda_{\min})$, the algorithm defined by (3), ..., (10) and the pull move set from [15] tends to minimum energy conformations in the H-P model.*

3 Run-time Estimates of Simulations

For any $\alpha \in \mathcal{F}$ we introduce the following parameters:

$$s(\alpha) := |\{\alpha' : \alpha' \in \mathcal{N}_\alpha \wedge \mathcal{Z}(\alpha') > \mathcal{Z}(\alpha)\}|, \tag{11}$$

$$r(\alpha) := |\{\alpha' : \alpha' \in \mathcal{N}_\alpha \wedge \alpha' \neq \alpha \wedge \mathcal{Z}(\alpha') \leq \mathcal{Z}(\alpha)\}|. \tag{12}$$

Thus, from the definition of $\mathcal{N}_\alpha$ and (4) we have

$$s(\alpha) + r(\alpha) = N_\alpha - 1. \tag{13}$$

We observe that for $\mathcal{Z}(\alpha') > \mathcal{Z}(\alpha)$ the acceptance probability (7) can be rewritten as

$$e^{-(\mathcal{Z}(\alpha') - \mathcal{Z}(\alpha))/t(k)} = \frac{1}{(k+2)^{(\mathcal{Z}(\alpha') - \mathcal{Z}(\alpha))/\Gamma}},\ k \geq 0. \tag{14}$$

To simplify notations, we use $\gamma := \gamma(\alpha', \alpha) := (\mathcal{Z}(\alpha') - \mathcal{Z}(\alpha))/\Gamma$, in most cases not indicating the dependence on (α', α).

In (9), we separate the probabilities according to whether or not α' equals α, and the probability to remain in α is substituted by the defining equation from (8). Thus, we obtain:

$$\mathbf{a}_\alpha(k) = \sum_{\alpha' \in \mathcal{N}_\alpha} \mathbf{a}_{\alpha'}(k-1) \cdot \mathbf{Pr}\{\alpha' \to \alpha\}$$

$$= \mathbf{a}_\alpha(k-1)\cdot\mathbf{Pr}\{\alpha\to\alpha\} + \sum_{\alpha'\neq\alpha}\mathbf{a}_{\alpha'}(k-1)\cdot\mathbf{Pr}\{\alpha'\to\alpha\}$$

$$= \mathbf{a}_\alpha(k-1)\cdot\Big(1-\sum_{\alpha'\neq\alpha}\mathbf{Pr}\{\alpha\to\alpha'\}\Big) + \sum_{\alpha'\neq\alpha}\mathbf{a}_{\alpha'}(k-1)\cdot\mathbf{Pr}\{\alpha'\to\alpha\}.$$

The value of $\mathbf{a}_\alpha(k)$ is now expressed by using structural parameters as defined in (11) and (12):

Lemma 1 *The value of* $\mathbf{a}_\alpha(k)$ *can be calculated from probabilities of the previous step by*

$$\begin{aligned}\mathbf{a}_\alpha(k) &= \Big(\frac{s(\alpha)+1}{N_\alpha} - \frac{1}{N_\alpha}\cdot\sum_{i=1}^{s(\alpha)}\frac{1}{(k+1)^\gamma}\Big)\cdot\mathbf{a}_\alpha(k-1) + \sum_{i=1}^{s(\alpha)}\frac{\mathbf{a}_{\alpha_i}(k-1)}{N_{\alpha_i}} + \\ &\quad + \sum_{j=1}^{r(\alpha)}\frac{\mathbf{a}_{\alpha_j}(k-1)}{N_{\alpha_j}}\cdot\frac{1}{(k+1)^\gamma}. \qquad (15)\end{aligned}$$

The backwards expansion from Lemma 1 will be used as the main relation reducing $\mathbf{a}_\alpha(k)$ to probabilities from previous steps. The elements of the conformation space are distinguished by their minimum distance to $\mathcal{F}_{\min}$: Given $\alpha\in\mathcal{F}$, we consider a shortest path of length $\mathrm{dist}(\alpha)$ with respect to neighbourhood transitions from α to $\mathcal{F}_{\min}$. We introduce a partition of $\mathcal{F}$ in accordance with $\mathrm{dist}(\alpha)$:

$$\alpha\in M_i \Longleftrightarrow \mathrm{dist}(\alpha) = i \geq 0, \quad \text{and} \quad \mathcal{M}_{\mathrm{d\_m}} = \bigcup_{i=0}^{\mathrm{d\_m}} M_i, \qquad (16)$$

where $M_0 := \mathcal{F}_{\min}$ and d_m is the maximum distance. From the proof of Theorem 1 in [15] we conclude

$$d_m \leq n^{O(1)}. \qquad (17)$$

Since we want to analyze the convergence to elements from $M_0 = \mathcal{F}_{\min}$, we have to show that the value

$$\sum_{\alpha\notin M_0}\mathbf{a}_\alpha(k) \qquad (18)$$

becomes small as k increases. We assume $k\geq d\_m$ and we are going backwards from step k: At the same backwards transition from k to $(k-1)$, the neighbours of α are generating terms containing $\mathbf{a}_\alpha(k-1)$ as a factor in the same way as $\mathbf{a}_\alpha(k)$ generates terms with factors $\mathbf{a}_{\alpha_i}(k-1)$ and $\mathbf{a}_{\alpha_j}(k-1)$, see Lemma 1. If we now consider the entire sum $\sum_{\alpha\notin M_0}\mathbf{a}_\alpha(k)$, the terms corresponding to a particular $\mathbf{a}_\alpha(k-1)$ can be collected together to form a single expression. Firstly, we consider $\alpha\in M_i$, $i\geq 2$. In this case, α does not have neighbours from M_0, i.e., the expansion from Lemma 1 appears for all neighbours of α in the reduction of $\sum_{\alpha\notin M_0}\mathbf{a}_\alpha(k)$ to step $(k-1)$. Therefore, taking all terms together that contain $\mathbf{a}_\alpha(k-1)$, we obtain

$$\begin{aligned}&\mathbf{a}_\alpha(k-1)\cdot\Big\{\Big(\frac{N_\alpha - r(\alpha)}{N_\alpha} - \frac{1}{N_\alpha}\cdot\sum_{i=1}^{s(\alpha)}\frac{1}{(k+1)^{\gamma_i}}\Big) + \frac{1}{N_\alpha}\cdot\sum_{i=1}^{s(\alpha)}\frac{1}{(k+1)^{\gamma_i}} + \frac{r(\alpha)}{N_\alpha}\Big\} \\ &= \mathbf{a}_\alpha(k-1). \qquad (19)\end{aligned}$$

Secondly, if $\alpha\in M_1$, the neighbours from M_0 are missing in $\sum_{\alpha\notin M_0}\mathbf{a}_\alpha(k)$ at the step to $(k-1)$, i.e. they do not generate terms containing probabilities from higher levels. For

$\alpha' \in M_0$, the expansion from Lemma 1 contains the terms $\mathbf{a}_{\alpha_i}(k-1)/N_{\alpha_i}$ for $\alpha_i \in M_1$. Thus, the terms $\mathbf{a}_{\alpha_i}(k-1)/N_{\alpha_i}$ are not "available" for $\alpha = \alpha_i \in M_1$ in the reduction of $\sum_{\alpha \notin M_0} \mathbf{a}_\alpha(k)$ to step $(k-1)$. For each $\alpha \in M_1$, there are $r(\alpha)$ such terms related to neighbours from M_0. Therefore, in the expansion of $\sum_{\alpha \notin M_0} \mathbf{a}_\alpha(k)$, the following arithmetic term is generated when the particular α is from M_1:

$$\Big(1 - \frac{r(\alpha)}{N_\alpha}\Big) \cdot \mathbf{a}_\alpha(k-1). \tag{20}$$

We introduce the following abbreviations:

$$\varphi(\alpha, v) := \frac{1}{N_\alpha} \cdot \sum_{i=1}^{s(\alpha)} \frac{1}{(k+2-v)^{\gamma_i}} \quad \text{and} \quad D_\alpha(k-v) := \frac{s(\alpha)+1}{N_\alpha} - \varphi(\alpha, v). \tag{21}$$

Now, the relations expressed in (19) and (20) can be summarised to

Lemma 2 *A single step of the expansion of* $\sum_{\alpha \notin M_0} \mathbf{a}_\alpha(k)$ *results in*

$$\sum_{\alpha \notin M_0} \mathbf{a}_\alpha(k) = \sum_{\alpha \notin M_0} \mathbf{a}_\alpha(k-1) - \sum_{\alpha \in M_1} \frac{r(\alpha)}{N_\alpha} \cdot \mathbf{a}_\alpha(k-1) + \sum_{\alpha' \in M_0} \varphi(\alpha', 1) \cdot \mathbf{a}_{\alpha'}(k-1). \tag{22}$$

The diminishing factor $(1 - r(\alpha)/N_\alpha)$ is generated by definition for all elements of M_1. At subsequent reduction steps, the factor is "transmitted" successively to all probabilities from higher distance levels M_i because any element of M_i has at least one neighbour from M_{i-1}. The main task is now to analyse how this diminishing factor changes when it is transmitted to higher distance levels. We denote

$$\sum_{\alpha \notin M_0} \mathbf{a}_\alpha(k) = \sum_{\alpha \notin M_0} \mu(\alpha, v) \cdot \mathbf{a}_\alpha(k-v) + \sum_{\alpha' \in M_0} \mu(\alpha', v) \cdot \mathbf{a}_{\alpha'}(k-v), \tag{23}$$

i.e. the coefficients $\mu(\tilde{\alpha}, v)$ are the factors at probabilities after v steps of a backwards expansion of $\sum_{\alpha \notin M_0} \mathbf{a}_\alpha(k)$. Starting from step $(k-1)$, the probabilities $\mathbf{a}_{\alpha'}(k-v)$, $\alpha' \in M_0$, from (23) are expanded in the same way as the probabilities for all other $\alpha \notin M_0$.

We establish a recursive relation for the coefficients $\mu(\tilde{\alpha}, v)$ defined in (23). The recursive relation is derived by an inductive step from $(k-(v-1))$ to $(k-v)$, $v \geq 2$, where the probabilities $\mathbf{a}_{\tilde{\alpha}}(k-(v-1))$ are expanded in

$$\sum_{\alpha \notin M_0} \mathbf{a}_\alpha(k) = \sum_{\tilde{\alpha} \in \mathcal{F}} \mu(\tilde{\alpha}, v-1) \cdot \mathbf{a}_{\tilde{\alpha}}(k-(v-1)) \tag{24}$$

according to Lemma 1. We note that the particular summands in the expansion of $\mathbf{a}_{\tilde{\alpha}}(k-(v-1))$, i.e. the summands at the right hand side in Lemma 1 are multiplied by the corresponding $\mu(\tilde{\alpha}, v-1)$. Taking together all terms associated with a particular $\mathbf{a}_{\tilde{\alpha}}(k-v)$, we have

$$\begin{aligned} & \mathbf{a}_{\tilde{\alpha}}(k-v) \cdot \Big\{ \mu(\tilde{\alpha}, v-1) \cdot \Big(\frac{N_{\tilde{\alpha}} - r(\tilde{\alpha})}{N_{\tilde{\alpha}}} - \varphi(\tilde{\alpha}, v) \Big) + \\ & + \sum_{\alpha' > \tilde{\alpha}} \frac{\mu(\alpha', v-1)}{N_{\tilde{\alpha}}} \cdot \frac{1}{(k+2-v)^\gamma} + \sum_{\alpha'' \leq \tilde{\alpha}} \frac{\mu(\alpha'', v-1)}{N_{\tilde{\alpha}}} \Big\} \\ = \; & \mathbf{a}_{\tilde{\alpha}}(k-v) \cdot \mu(\tilde{\alpha}, v), \end{aligned}$$

where $\alpha' > \tilde{\alpha}$ is for $\mathcal{Z}(\alpha') > \mathcal{Z}(\tilde{\alpha})$ and $\alpha'' \leq \tilde{\alpha}$ for the reverse relation to simplify the notations. Thus, taking into account (21), we obtain the following parameterized representation:

Lemma 3 *The following recurrent relation is valid for the coefficients* $\mu(\tilde{\alpha}, v)$*:*

$$\mu(\tilde{\alpha},v)=\mu(\tilde{\alpha},v-1)\cdot D_{\tilde{\alpha}}(k-v)+\sum_{\alpha''<\tilde{\alpha}}\frac{\mu(\alpha'',v-1)}{N_{\tilde{\alpha}}}+\sum_{\alpha'>\tilde{\alpha}}\frac{\mu(\alpha',v-1)}{N_{\tilde{\alpha}}}\cdot\frac{1}{(k+2-v)^{\gamma}}. \tag{25}$$

It is important to note that the summands are divided by the same value $N_{\tilde{\alpha}}$.

We take advantage of the fact that for conformations α different from local and global minima the factor $D_{\alpha}(k-v)$, which is associated with the probability to remain in α, is smaller than $(1-1/(n+1))$, i.e. there is an upper bound independent of $(k-v)$; see (21). Therefore, for this type of conformations, it is possible to obtain an upper bound of $\mathbf{a}_{\alpha}(k)$ by straightforward calculations. Let MIN denote the set of all global and local minima. We set $\widehat{\mathcal{M}} := \{\alpha : r(\alpha) \geq 1\} = \mathcal{F}\backslash\text{MIN}$ and consider $\mathbf{a}_{\alpha}(k)$ defined by (8) and (9) when all probabilities on the right hand side are recursively substituted in the same way, where we break up the paths of the expansion that lead from some α to α' with $\mathcal{Z}(\alpha) > \mathcal{Z}(\alpha')$. Such transitions generate a factor $(k+2-u)^{-\gamma}$, which is then used as the crucial type of factors in the upper bound of $\mathbf{a}_{\alpha}(k)$. By analysing this type of expansions, we obtain:

Lemma 4 *If* $k > 2\cdot(n+1)^2\cdot\ln(k+2)^{\max\gamma}$*, then*

$$\sum_{\alpha\in\widehat{\mathcal{M}}}\mathbf{a}_{\alpha}(k) < \frac{3\cdot e\cdot(n+1)^3}{(k-2\cdot(n+1)^2\cdot\ln(k+2)^{\max\gamma})^{\min\gamma}}. \tag{26}$$

By $\mathcal{M}^{\text{lm}} \subset \text{MIN}$ we denote the set of all local minima. If $\alpha \in \mathcal{M}^{\text{lm}}$, we represent $\mu(\alpha, v)$ by $\mu(\alpha, v) = 1 - \nu(\alpha, v)$ and by straightforward calculations we obtain

$$\begin{aligned}\sum_{\alpha\notin M_0}\mathbf{a}_{\alpha}(k)-\sum_{\alpha\notin M_0}\mathbf{a}_{\alpha}(k') \;&<\; \frac{3\cdot e\cdot(n+1)^3}{(k-2\cdot(n+1)^2\cdot\ln(k+2)^{\max\gamma})^{\min\gamma}}+\\ &\quad+\sum_{\alpha\in\mathcal{M}^{\text{lm}}}\nu(\alpha,v')\cdot\mathbf{a}_{\alpha}(k).\end{aligned} \tag{27}$$

Thus, it remains to analyse $\nu(\alpha, v')$, $v' \geq d\_m + v$, for local minima:

Lemma 5 *If* $\alpha \in \mathcal{M}^{\text{lm}}$*, then*

$$\nu(\alpha,v') \;<\; \frac{4\cdot(n+1)}{(k+2-v')^{\min\gamma}}. \tag{28}$$

From (27) and Lemma 5 we obtain together with $\sum_{\tilde{\alpha}\notin M_0}\mathbf{a}_{\tilde{\alpha}}(k) = \sum_{\tilde{\alpha}\notin M_0}(\mathbf{a}_{\tilde{\alpha}}(k)-\mathbf{a}_{\tilde{\alpha}}(k'))+\sum_{\tilde{\alpha}\notin M_0}\mathbf{a}_{\tilde{\alpha}}(k')$ the main result:

Theorem 3 *If* $\Gamma \geq \max_{\lambda_{\min}} D(\lambda_{\min})$ *for the conformation space* $\mathcal{F}$ *from (2) and* $0 < \delta < 1$*, then*

$$k \;\geq\; 2\cdot\Big(\frac{8\cdot e\cdot(n+1)^3}{\delta}\Big)^{\Gamma/c}$$

implies for arbitrary initial probability distributions $\vec{\mathbf{a}}(0)$ *the relation*

$$\sum_{\tilde{\alpha}\notin\mathcal{F}_{\min}}\mathbf{a}_{\tilde{\alpha}}(k) < \delta \;\; \textit{and therefore,} \;\; \sum_{\alpha'\in\mathcal{F}_{\min}}\mathbf{a}_{\alpha'}(k) \;\geq\; 1-\delta,$$

where c *is determined by* $\min_{(\alpha',\alpha)}(\mathcal{Z}(\alpha')-\mathcal{Z}(\alpha))$.

References

[1] C.B. Anfinsen. Principles that govern the folding of protein chains. *Science* 181:223–230, 1973.

[2] B. Berger, T. Leighton. Protein folding in the hydrophobic-hydrophilic (HP) model is NP-complete. *J. Comput. Biol.* 5(1):27-40, 1998.

[3] J. Blazewicz, P. Lukasiak, M. Milostan. Application of tabu search strategy for finding low energy structure of protein *Artif. Intell. Med.* 2005; to appear.

[4] O. Catoni. Rough large deviation estimates for simulated annealing: Applications to exponential schedules. *Ann. Probab.* 20:1109–1146, 1992.

[5] V. Černy, A thermodynamical approach to the travelling salesman problem: An efficient simulation algorithm. *J. Optim. Theory Appl.* 45:41–51, 1985.

[6] K.A. Dill, S. Bromberg, K.Yue, K.M. Fiebig, D.P. Yee, P.D. Thomas, H.S. Chan. Principles of protein folding - A perspective from simple exact models. *Protein Sci.* 4:561–602, 1995.

[7] M.P. Eastwood, C. Hardin, Z. Luthey-Schulten, P.G. Wolynes. Evaluating protein structure-prediction schemes using energy landscape theory. *IBM J. Res. Dev.* 45(3/4):475–497, 2001.

[8] S.D. Flores, J. Smith. Study of fitness landscapes for the HP model of protein structure prediction. *Proc. IEEE Congress on Evolutionary Computation*, pp. 2338–2345, 2003.

[9] B. Fu, W. Wang. A $2^{O(n^{1-1/d}\cdot\log n)}$ time algorithm for d-dimensional protein folding in the HP-model. *Proc. ICALP 2004*, pp. 630–644, LNCS 3142, 2004.

[10] A. Gerevini, I. Serina. Planning as propositional CSP: From walksat to local search techniques for action graphs. *Constraints* 8(4):389–413, 2003.

[11] H.J. Greenberg, W.E. Hart, G. Lancia. Opportunities for combinatorial optimization in computational biology. *INFORMS J. Comput.* 16(3):211–231, 2004.

[12] B. Hajek, Cooling schedules for optimal annealing. *Mathem. Oper. Res.* 13:311–329, 1988.

[13] V. Heun. Approximate protein folding in the HP side chain model on extended cubic lattices. *Discrete Appl. Math.* 127(1):163–177, 2003.

[14] S. Kirkpatrick, C.D. Gelatt, Jr., and M.P. Vecchi, Optimization by simulated annealing. *Science* 220:671–680, 1983.

[15] N. Lesh, M. Mitzenmacher, S. Whitesides. A complete and effective move set for simplified protein folding. *Proc. RECOMB 2003*, pp. 188–195, 2003.

[16] A. Nayak, A. Sinclair, U. Zwick. Spatial codes and the hardness of string folding problems. *J. Comput. Biol.* 6(1):13–36, 1999.

[17] A. Neumaier. Molecular modeling of proteins and mathematical prediction of protein structure. *SIAM Rev.* 39(3):407–460, 1997.

[18] P.M. Pardalos, X. Liu, G. Xue. Protein conformation of a lattice model using tabu search. *J. Global Optim.* 11(1):55–68, 1997.

[19] S. Seitz, M. Alava, P. Orponen. Threshold behaviour of WalkSAT and Focused Metropolis search on random 3-satisfiability. *Proc. 8th Int. Conf. Theory and Applications of Satisfiability Testing (SAT 2005)*, pp. 475–481, LNCS 3569, 2005.

[20] K. Steinhöfel, A. Albrecht, C.K. Wong. An experimental analysis of local minima to improve neighbourhood search. *Comput. & Oper. Res.* 30:2157–2173, 2003.

[21] J.E. Straub. Protein folding and optimization algorithms. *The Encyclopedia of Computational Chemistry*, vol. 3, pp. 2184–2191, Wiley, 1998.