Landscape Analysis for Protein-Folding Simulation in the H-P Model*

Kathleen Steinhöfel[1], Alexandros Skaliotis[1], and Andreas A. Albrecht[2]

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

Abstract. The hydrophobic-hydrophilic (H-P) model for protein folding was introduced by Dill et al. [7]. 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. The protein folding problem in the H-P model is NP-hard in both 2D and 3D. Recently, Fu and Wang [10] 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. [4]. We obtain that after $(m/\delta)^{O(\Gamma)}$ Markov chain transitions, the probability to be in a minimum energy conformation is at least $1 - \delta$, where m is the maximum neighbourhood size and Γ is the maximum value of the minimum escape height from local minima of the underlying energy landscape. We note that the time bound depends on the specific instance. Based on [10] we conjecture $\Gamma \leq n^{1-1/d}$. We analyse $\Gamma \leq \sqrt{n}$ experimentally on selected benchmark problems [15,21] for the 2D case.

1 Introduction

A great variety of models has been developed for protein folding simulations, with different levels of detail (for a concise discussion, cf. [20]). In the present paper, we focus on *minimal models* [11], and we distinguish roughly between lattice models [7] and off-lattice models [8,17]. For a discussion of energy functions and justifications for the use of simplified (approximated) energy functions we refer the reader to [20]. One of the most popular models of protein folding is the hydrophobic-hydrophilic (H-P) model [7]. 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 into 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

* Research partially supported by EPSRC Grant No. EP/D062012/1.

P. Bücher and B.M.E. Moret (Eds.): WABI 2006, LNBI 4175, pp. 252–261, 2006.

objective is that hydrophobic interactions contribute a significant portion of the total energy function. 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 [11].

Lattice models of protein folding have provided valuable insights into the general complexity of protein structure prediction problems: Protein structure prediction has been shown to be NP-hard for a variety of lattice models [3,11,16]. The intractability results are complemented by performance guaranteed approximation algorithms that run in linear time [11,13]. 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. [4] 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 [10] proved an $exp(O(n^{1-1/d}) \cdot \ln n)$ algorithm for d-dimensional protein folding simulation in the HP-model. It is interesting to note that this time bound almost exactly mirrors the folding time approximation $exp(\lambda \cdot n^{2/3} \pm \chi \cdot n^{1/2}/2)$ by Finkelstein and Badretdinov [9][1].

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 [4]. We obtain that after $(m/\delta)^{O(\Gamma)}$ Markov chain transitions, the probability to be in a minimum energy conformation is at least $1 - \delta$, where m is the maximum neighbourhood size of individual conformations, and Γ is the maximum value of the minimum escape height from local minima of the underlying energy landscape. Thus, the run-time estimation is *problem-specific*. To be competitive with the Fu/Wang run-time bound, we need to show $\Gamma \leq n^{1-1/d}$. Future research will focus on proven upper bounds of Γ in the context of complete move sets for the H-P model. In the present paper, we analyse the conjecture $\Gamma \leq \sqrt{n}$ experimentally on selected benchmark problems (taken from [15,21]) for the 2D case.

2 Preliminaries

Our stochastic local search procedure for protein folding is based on simulated annealing [6,14], where the underlying Markov chain is of inhomogeneous type [5,12]. For simplicity of presentation, we focus on the 2D rectangular grid H-P model only.

Anfinsen's thermodynamic hypothesis [2] 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]. In the

[1] The authors are grateful to one anonymous referee for drawing our attention to [9].

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 [4]). 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 $\mathbf{a}(k)$ 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 $\exists \alpha_0, \alpha_1, \ldots, \alpha_r \in \mathcal{F}$ 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

In this section, we outline a run-time estimation for finding optimum conformations with a certain confidence $\delta' = 1 - \delta > 0$. The run-time estimation is an extension of the convergence analysis from [1] to a more complicated objective function, and it relates the run-time to the landscape parameter Γ (cf. (10)), to the confidence parameter $\delta' = 1 - \delta$, and to the maximum size m of individual neighbourhood sets.

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

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

To simplify notation, 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:

$$\begin{aligned}\mathbf{a}_\alpha(k) &= \sum_{\alpha' \,\in\, \mathcal{N}_\alpha} \mathbf{a}_{\alpha'}(k-1) \cdot \mathbf{Pr}\{\alpha' \to \alpha\} \\ &= \mathbf{a}_\alpha(k-1) \cdot \Big(1 - \sum_{\alpha' \,\neq\, \alpha} \mathbf{Pr}\{\alpha \to \alpha'\}\Big) + \sum_{\alpha' \,\neq\, \alpha} \mathbf{a}_{\alpha'}(k-1) \cdot \mathbf{Pr}\{\alpha' \to \alpha\}.\end{aligned}$$

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*

$$\mathbf{a}_\alpha(k) = \left(\frac{s(\alpha)+1}{N_\alpha} - \frac{1}{N_\alpha}\cdot\sum_{i=1}^{s(\alpha)}\frac{1}{(k+1)^\gamma}\right)\cdot\mathbf{a}_\alpha(k-1) + \sum_{i=1}^{s(\alpha)}\frac{\mathbf{a}_{\alpha_i}(k-1)}{N_{\alpha_i}} + \\ + \sum_{j=1}^{r(\alpha)}\frac{\mathbf{a}_{\alpha_j}(k-1)}{N_{\alpha_j}}\cdot\frac{1}{(k+1)^\gamma}. \tag{15}$$

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 $\mathrm{dist}(\alpha)$:

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

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

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

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

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

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, 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:

$$\left(1 - \frac{r(\alpha)}{N_\alpha}\right)\cdot\mathbf{a}_\alpha(k-1). \tag{19}$$

We introduce the following abbreviations:

$$\varphi(\alpha,v) := \frac{1}{N_\alpha}\cdot\sum_{i=1}^{s(\alpha)}\frac{1}{(k+2-v)^{\gamma_i}} \quad \text{and} \quad D_\alpha(k-v) := \frac{s(\alpha)+1}{N_\alpha} - \varphi(\alpha,v). \tag{20}$$

Now, the backwards expansion can be summarised to

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

$$\sum_{\alpha \notin M_0} \mathbf{a}_\alpha(k) = \sum_{\alpha \notin M_0} \mathbf{a}_\alpha(k-1) - \sum_{\alpha \in M_1} \frac{r(\alpha)}{N_\alpha} \cdot \mathbf{a}_\alpha(k-1) + \sum_{\alpha' \in M_0} \varphi(\alpha', 1) \cdot \mathbf{a}_{\alpha'}(k-1). \quad (21)$$

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}. We denote

$$\sum_{\alpha \notin M_0} \mathbf{a}_\alpha(k) = \sum_{\alpha \notin M_0} \mu(\alpha, v) \cdot \mathbf{a}_\alpha(k-v) + \sum_{\alpha' \in M_0} \mu(\alpha', v) \cdot \mathbf{a}_{\alpha'}(k-v), \quad (22)$$

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 (22) are expanded in the same way as the probabilities for all other $\alpha \notin M_0$. Taking into account (20), we obtain the following parameterized representation for $\mu(\tilde{\alpha}, v)$:

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

$$\mu(\tilde{\alpha}, v) = \mu(\tilde{\alpha}, v-1) \cdot D_{\tilde{\alpha}}(k-v) + \sum_{\alpha'' < \tilde{\alpha}} \frac{\mu(\alpha'', v-1)}{N_{\tilde{\alpha}}} + \sum_{\alpha' > \tilde{\alpha}} \frac{\mu(\alpha', v-1)}{N_{\tilde{\alpha}}} \cdot \frac{1}{(k+2-v)^\gamma}. \quad (23)$$

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 $(1 - 1/(m+1))$ for $m := \max_\alpha N_\alpha$, i.e. there is an upper bound independent of $(k-v)$; see (20). Let MIN denote the set of all global and local minima. We set $\widehat{\mathcal{M}} := \{\alpha : r(\alpha) \geq 1\} = \mathcal{F} \backslash \mathrm{MIN}$ 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* $k > 2 \cdot (m+1)^2 \cdot \ln(k+2)^{\max \gamma}$ *for the maximum size* m *of neighbourhoods, then*

$$\sum_{\alpha \in \widehat{\mathcal{M}}} \mathbf{a}_\alpha(k) < O\Big(\frac{(m+1)^3}{(k - 2 \cdot (m+1)^2 \cdot \ln(k+2)^{\max \gamma})^{\min \gamma}}\Big). \quad (24)$$

By $\mathcal{M}^{\mathrm{lm}} \subset \mathrm{MIN}$ we denote the set of all local minima, and $\mathcal{A}$ stands for the RHS of (24). If $\alpha \in \mathcal{M}^{\mathrm{lm}}$, we represent $\mu(\alpha, v)$ by $\mu(\alpha, v) = 1 - \nu(\alpha, v)$ and by straightforward calculations we obtain

$$\sum_{\alpha \notin M_0} \mathbf{a}_\alpha(k) - \sum_{\alpha \notin M_0} \mathbf{a}_\alpha(k') < \mathcal{A} + \sum_{\alpha \in \mathcal{M}^{\mathrm{lm}}} \nu(\alpha, v') \cdot \mathbf{a}_\alpha(k).$$

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

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

$$\nu(\alpha, v') \;<\; O\Big(\frac{(m+1)}{(k+2-v')^{\min\gamma}}\Big). \tag{25}$$

From (25) and Lemma 5 we obtain the main result:

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

$$k \geq \Big(\frac{(m+1)^3}{\delta}\Big)^{O(\Gamma)} \;\textit{implies}\; \sum_{\alpha' \in \mathcal{F}_{\min}} \mathbf{a}_{\alpha'}(k) \;\geq\; 1 \,-\, \delta. \tag{26}$$

4 Landscape Analysis on Selected Benchmarks

As mentioned in Section 1 already, the run-time estimation (26) from Theorem 3 is problem-specific, i.e. depends on the parameter Γ of the landscape induced by an individual protein sequence. For a problem-independent upper bound we conjecture $\Gamma \leq n^{1-1/d}$, which complies with the result from [10]. However, for individual protein sequences one can proceed as follows: Given a sequence α, the parameter Γ is estimated in a pre-processing step (landscape analysis), where the maximum increase of the objective function is monitored in-between two successive improvements of the best value obtained so far. This approach usually overestimates Γ significantly. Therefore, we are searching for a suitable constant c such that $\Gamma' = G_{\mathrm{monit}}/c$ comes closer to Γ, where G_{monit} is the maximum of the monitored increases of the objective function in-between two successive total improvements of the objective function. This estimation Γ' is then taken (together with the length of α and a choice of δ for the confidence $1-\delta$) as the setting for the (slightly simplified) run-time estimation according to (26). In our computational experiments on 2D benchmark problems we indeed obtain optimum solutions for smaller values of Γ than $\sqrt{n}$.

The stochastic local search procedure as described in Section 2 was implemented and we analysed the following 2D benchmark problems (cf. [15,21]):

Table 1. Selected 2D benchmark problems from [15,21]

name/n	structure	$\mathcal{Z}_{\min}$
S36	3P2H2P2H5P7H2P2H4P2H2PH2P	-14
S60	2P3HP8H3P10HPH3P12H4P6HP2HPHP	-35
S64	12HPHPH2P2H2P2H2PH2P2H2P2H2PH2P2H2P2H2P2H 2PHPHP12H	-42
S85	4H4P12H6P12H3P12H3P12H3PH2P2H2P2H2PHPH	-53
S100	6PHP2H5P3HP5HP2H4P2H2P2HP5HP10HP2HP7H 11P7H2PHP3H6PHP2H	-48

Table 2. Results for selected 2D benchmarks; $1 - \delta = 0.51$

name/n	$\sqrt{n}$	G_{monit}	Γ'	$(n/\delta)^{\Gamma'}$	$T_{\max}$
S36	6.00	9.25	3.00	$\approx 4.0 \times 10^5$	29,341
S60	≈ 7.74	14.00	3.87	$\approx 1.2 \times 10^8$	30,319
S64	8.00	18.00	4.00	$\approx 2.9 \times 10^8$	259,223
S85	≈ 9.20	21.75	4.60	$\approx 2.0 \times 10^{10}$	13,740,964
S100	10.00	21.50	5.00	$\approx 3.5 \times 10^{11}$	57,195,268

Unfortunately, information about the exact number of ground states is not provided; the ground states are equally treated. In [15], three states are reported for S85, two states for S100.

Following the experimental part of [1], we use $(m/\delta)^{\Gamma'}$ as a simplified version of (26), where Γ' is $\approx \sqrt{n}/2$. We compare Γ' to G_{monit}/c, i.e. apart from trying to approximate the real Γ by Γ', we also try to relate Γ' to G_{monit}.

In Table 2 we report results where $\mathcal{Z}_{\min}$ was achieved for all five benchmark problems from Table 1. By $T_{\max}$ we denote the average number of transitions necessary to achieve $\mathcal{Z}_{\min}$ calculated from four successive runs for the same benchmark problem. The same applies to G_{monit}, which is the average from the four runs executed for each of the five benchmark problems. Although by definition Γ has to be an integer value, we allowed rational values for Γ'. The simplified version of (26) was calculated for $m = n$ and $\delta = 0.49$, i.e. for a confidence of 51%. As already mentioned, the value of Γ' was chosen $\approx \sqrt{n}/2$, which was used in (10) for the implementation.

As can be seen, the simplified version of (26) still overestimates the number of transitions sufficient to achieve $\mathcal{Z}_{\min}$ for the selected benchmark problems, which is at least partly due to the setting $m = n$. To incorporate improved upper bounds of m will be subject of future research. Based on the data from Table 2, the constant c in $\Gamma' = G_{\text{monit}}/c$ ranges from 3.08 to 4.73. Overall, the results encourage us to attempt a formal proof of the conjecture $\Gamma \leq \sqrt{n}$.

5 Concluding Remarks

We analyzed the run-time of protein folding simulations in the H-P model, if the underlying algorithm is based on the pull move set and logarithmic simulated annealing. We obtained that the probability to be in a minimum energy conformation is at least $1 - \delta$ after $(m/\delta)^{\kappa \cdot \Gamma}$ Markov chain transitions, where $m <$ sequence length n, κ is a small constant, and Γ is a crucial parameter of the landscape induced by the energy measure, the pull move set, and the individual sequence that has to be folded. Future research will be directed towards tight upper bounds of Γ in terms of the sequence length n, improved upper bounds of the maximum neighbourhood size m, on computational experiments on benchmark problems for the 3D case, and on landscape properties related to Levinthal's paradox [18], i.e. if there are "shallow" sub-landscapes with small Γ that imply fast folding.

References

1. Albrecht, A.A.: A stopping criterion for logarithmic simulated annealing. Computing (2006); in press.
2. Anfinsen, C.B.: Principles that govern the folding of protein chains. Science **181** (1973) 223–230.
3. Berger, B., Leighton, T.: Protein folding in the hydrophobic-hydrophilic (HP) model is NP-complete. J. Comput. Biol. **5** (1998) 27–40.
4. Blazewicz, J., Lukasiak, P., Milostan, M.: Application of tabu search strategy for finding low energy structure of protein. Artif. Intell. Med. **35** (2005) 135–145.
5. Catoni, O.: Rough large deviation estimates for simulated annealing: applications to exponential schedules. Ann. Probab. **20** (1992) 1109–1146.
6. Černy, V.: A thermodynamical approach to the travelling salesman problem: an efficient simulation algorithm. J. Optim. Theory Appl. **45** (1985) 41–51.
7. Dill, K.A., Bromberg, S., Yue, K., Fiebig, K.M., Yee, D.P., Thomas, P.D., Chan, H.S.: Principles of protein folding – A perspective from simple exact models. Protein Sci. **4** (1995) 561–602.
8. Eastwood, M.P., Hardin, C., Luthey-Schulten, Z., Wolynes, P.G.: Evaluating protein structure-prediction schemes using energy landscape theory. IBM J. Res. Dev. **45** (2001) 475–497.
9. Finkelstein, A.V., Badretdinov A.Y.: Rate of protein folding near the point of thermodynamic equilibrium between the coil and the most stable chain fold. Folding & Design **2** (1997) 115–121.
10. Fu, B., Wang, W.: A $2^{O(n^{1-1/d}\cdot \log n)}$ time algorithm for d-dimensional protein folding in the HP-model. Proc. ICALP'04, pp. 630–644, LNCS 3142, 2004.
11. Greenberg, H.J., Hart, W.E., Lancia, G.: Opportunities for combinatorial optimization in computational biology. INFORMS J. Comput. **16** (2004) 211–231.
12. Hajek, B.: Cooling schedules for optimal annealing. Mathem. Oper. Res. **13** (1988) 311–329.
13. Heun, V.: Approximate protein folding in the HP side chain model on extended cubic lattices. Discrete Appl. Math. **127** (2003) 163–177.
14. Kirkpatrick, S., Gelatt Jr., C.D., Vecchi, M.P.: Optimization by simulated annealing. Science **220** (1983) 671–680.
15. Lesh, N., Mitzenmacher, M., Whitesides, S.: A complete and effective move set for simplified protein folding. Proc. RECOMB'03, pp. 188–195, 2003.
16. Nayak, A., Sinclair, A., Zwick, U.: Spatial codes and the hardness of string folding problems. J. Comput. Biol. **6** (1999) 13–36.
17. Neumaier, A.: Molecular modeling of proteins and mathematical prediction of protein structure. SIAM Rev. **39** (1997) 407–460.
18. Ngo, J.M., Marks, J., Karplus, M.: Computational complexity, protein structure prediction, and the Levinthal paradox. In: K. Merz Jr., S. LeGrand (eds.), The Protein Folding Problem and Tertiary Structure Prediction, pp. 433–506, Birkhäuser, Boston, 1994.
19. Pardalos, P.M., Liu, X., Xue, G.: Protein conformation of a lattice model using tabu search. J. Global Optim˙**11** (1997) 55–68.
20. Straub, J.E.: Protein folding and optimization algorithms. The Encyclopedia of Computational Chemistry, vol. 3, pp. 2184–2191, Wiley & Sons, 1998.
21. Unger, R., Moult, J.: Genetic algorithms for protein folding simulations. J. Mol. Biol. **231** (1993) 75–81.