Exact methods for lattice protein models

Martin Mann[1] and Rolf Backofen[1–4]

[1]Bioinformatics Group, Department of Computer Science, University of Freiburg, Georges-Köhler-Allee 106, 79110 Freiburg, Germany
[2]Center for Biological Signaling Studies (BIOSS), University of Freiburg, Germany
[3]Center for Biological Systems Analysis (ZBSA), University of Freiburg, Germany
[4]Center for non-coding RNA in Technology and Health, University of Copenhagen, Grønnegårdsvej 3, DK-1870 Frederiksberg C, Denmark

Abstract

Lattice protein models are well studied abstractions of globular proteins. By discretizing the structure space and simplifying the energy model over regular proteins, they enable detailed studies of protein structure formation and evolution. But even in the simplest lattice protein models, the prediction of optimal structures is computationally hard. Therefore, often heuristic approaches are applied to find such conformations. Commonly, heuristic methods find only locally optimal solutions. Nevertheless, there exist methods that guarantee to predict globally optimal structures. Currently only one such exact approach is publicly available, namely the Constraint-based Protein Structure Prediction (CPSP) method and variants. Here, we review exact approaches and derived methods. We discuss fundamental concepts like hydrophobic core construction and their use in optimal structure prediction as well as possible applications like combinations of different energy models.

Key words: lattice protein, protein structure prediction, optimal structure, CPSP, HP-model

1 Introduction

Almost all cellular processes on earth are governed or guided by proteins [85]. Their functionality is always coupled to the formation of a specific three-dimensional structure named *functional, native, or biological fold/conformation*. It was shown via refolding experiments that a protein's functional fold is mainly encoded by its sequence of amino acids [1, 28]. Nevertheless, there is a large number of processes, like crowding effects [45], co-factor binding [93] or chaperons [94], that influence or support the structure formation process within living cells [24]. Still, the folding process is mainly governed by non-covalent intramolecular interactions [77] and misfolding can result in reduced or broken functionality [83]. Several diseases [50, 67, 73] and even cancer [80] are caused by misfolded proteins.

Two connected but independent problems arise: (I) To study the function of a protein, its native fold has to be identified. This is known as the *protein structure prediction* (PSP) problem. (II) To understand the mechanistic details of misfolding, the folding process itself has to be investigated e.g. using *folding simulations*. Here, we are focusing on the PSP problem, which resembles the result of the structure formation process.

In order to study the PSP problem various protein models have been developed. They range from full atom representations in three dimensions used in molecular dynamics simulations [46]; through off-lattice bead models [2]; and down to very coarse grained 2D lattice models [49]. Three more or less independent layers of abstraction can be found: structure space, sequence space, and energy function. The structure space dictates possible spatial properties of a protein, while sequence space and energy function describe the distinction of different amino acids and their interacting forces, respectively. Assuming the system to be in thermodynamic equilibrium [36], the native fold will be the structure with minimal free energy [1]. Thus, structure prediction can be approached with optimization methods for a given model.

Within this work, we will focus on PSP for lattice protein models. We first concentrate on the widely studied *Hydrophobic-Polar* (HP) model by Lau and Dill [49]. Even in this simplistic model, the identification of energy minimal structures is a hard computational problem [14, 27]. Since exhaustive structure enumeration approaches [43] are restricted to very short sequence lengths, usually heuristic methods are applied. They apply various techniques, e.g. simulated annealing [86, 87], quantum annealing [70], ant colonization optimization [64], evolutionary algorithms [84], energy landscape paving [53],

large neighborhood search [34, 35], constraint programming [29, 31, 32], or dedicated search heuristics [23, 54, 75, 81]. They enable performance-guaranteed approximations close to the optimum for both backbone-only [38, 65] and side chain models [39, 40]. For a detailed overview of approximating methods refer to reviews by Hart and Newman [37] or Istrail and Lam [44].

Here, we review exact methods for the PSP problem for lattice protein models, i.e. approaches that identify energy optimal structures only. Given the large amount of the mentioned recent local search PSP research, is is surprising that currently only two exact methods are known, both targeting the HP-model in three dimensions. The first approach was introduced by Yue and Dill and named *Constrained Hydrophobic Core Construction* (CHCC) [95]. It is based on the observation that globular proteins show a densely packed hydrophobic core both in reality as well as in the HP-model. Thus, energy optimal structures show an almost maximally dense packing of hydrophobic amino acids. This is utilized in the CHCC method by considering only (decreasingly) compact hydrophobic packings when searching for optimal structures. While the CHCC approach was the first exact PSP method, it was proven to be incomplete when enumerating all optimal structures [11]. In contrast, the *Constraint-based Protein Structure Prediction* (CPSP) approach [11, 60, 61] was shown to be exact and complete when calculating optimal structures. CPSP uses a similar strategy as CHCC and is so far the only exact and complete method for protein structure prediction in lattice protein models. It is often used as reference when testing local search methods and is discussed in detail within this review.

The CPSP approach facilitates methods for more sophisticated energy models. Among them are its extension and application to the HPNX model [12], which also takes polar interactions into account [17]. Furthermore, HP-optimal predictions can be used within a hierarchical prediction approach to search for energy optimal structures within full 20×20 pairwise potential models [74, 86, 87]. The latter requires the identification of HP-optimal conformations that are structurally diverse. This can be achieved by the introduction of an equivalence relation based on hydrophobic cores and the restricted enumeration of representatives for each equivalence class [56].

In the following, we will introduce the necessary formalisms to discuss lattice protein models and subsequently the CPSP approach. Afterwards, we will review the construction of (sub-)optimal hydrophobic core packings, a central prerequisite for both the CHCC and the CPSP method. Hydrophobic core information can be used to define equivalence classes of structures within the HP-model and we will present the enumeration of class representatives based on an extension of the CPSP method. Finally, the extension and application of the CPSP approach to enhanced lattice protein models is illustrated.

2 Lattice Protein Models

The strongest abstraction within lattice protein model is the allowed structure space, where conformations are discretized based on an underlying regular lattice. Such a *lattice* L is a set of 3D coordinates that form an additive group for any two points $\vec{u}, \vec{v} \in L$, i.e. it holds $(\vec{u} \pm \vec{v}) \in L$. The neighborhood $N_L \subset L$ is the minimal subset of vectors to encode the whole lattice by a linear combination of these vectors using positive integers only, i.e.

$$\forall_{\vec{u} \in L} : \vec{u} = \sum_{\vec{x} \in N_L} c_{\vec{x}} \cdot \vec{x} \quad \text{with } c_{\vec{x}} \in \mathbb{N} \tag{1}$$

Furthermore, we require N_L to contain also the reverse vectors, i.e. $\vec{x} \in N_L \rightarrow -\vec{x} \in N_L$, and only vectors of equal lengths, i.e. $\forall \vec{u}, \vec{v} \in N_L : |\vec{u}| = |\vec{v}|$. For instance, a 3D-cubic lattice is defined by $N_{\text{cubic}} = \{\pm(1,0,0), \pm(0,1,0), \pm(0,0,1)\}$. A wide variety of lattice has been studied in the context of lattice protein models. Common 3D lattices are the cubic (Fig. 1) and face-centered-cubic lattice (12 neighbors per point, Fig. 2) as well as the diamond lattice. For further lattices and details refer e.g. to [55, 69].

The coordinates of a given lattice define possible placements for protein monomers. Within backbone-only models, a protein is represented by its backbone C_α-atom positions only, while successive C_α-monomers have to be neighbored in the lattice according to N_L. Thus, a backbone-only lattice protein structure P of length n is defined by $P = (P_1, \ldots, P_n) \in L^n$ with $(P_{(i+1)} - P_i) \in N_L$. In addition, P has to be self-avoiding, i.e. $\forall i \neq j : P_i \neq P_j$. Side chain models extend the amino acid abstraction with a second side chain monomer, which has to be neighbored to the according C_α-monomer. For examples see Fig. 1. The modeling quality of real protein conformations by lattice protein structures depends strongly on the underlying lattice used [57, 59, 69]. Figure 1 depicts the exponential lattice-dependent growth of the structure space that is detailed in [55] for different lattices.

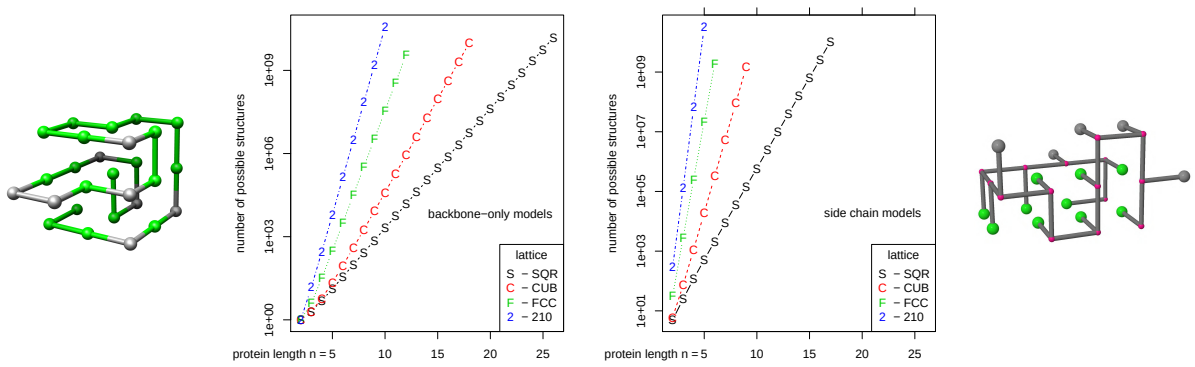

Figure 1: Examples for (left) backbone-only and (right) side chain lattice protein models in the 3D-cubic lattice. Colors encode an HP-model: hydrophobic (green), polar (gray) and backbone monomers (pink). (center) Sequence-length and lattice-dependent exponential growth of the symmetry free structure space for different lattices (SQR - 2D-square, CUB - 3D cubic, FCC - 3D face centered cubic, 210 - chess-knight). Figures are taken from [55].

The sequence space abstraction and energy function are the final determinants for a lattice protein model. Various reductions of the amino acid alphabet of proteins are known, which have been combined with according energy functions. Distance-based energy functions either incorporate sequence-based pairwise potentials scaled by the distance [88] or apply special distance dependent potentials [66]. For details refer to the review by Hart [37]. In the following, we will discuss contact based energy models, which can be used to define exact PSP methods. Therein, the energy E of a lattice protein structure P with sequence S is determined by the summation of all pairwise sequence-dependent contact potentials $e(S_i, S_j)$. Two non-successive monomers P_i and P_j are in contact, if they are neighbored within the lattice.

$$E(S,P) = \sum_{1 \le i+1 < j \le n} \Delta(P_i, P_j) \cdot e(S_i, S_j) \tag{2}$$

$$\Delta(P_i, P_j) = \begin{cases} 1 & \text{if } P_i - P_j \in N_L \\ 0 & \text{else.} \end{cases} \tag{3}$$

Most abstract is the HP-model [49], which models the central impact of hydrophobic forces within structure formation [71]. Here, hydrophobic amino acids are repelled from water due to their non-polar nature, resulting in a crowding of hydrophobic residues within the protein core [68] (see Fig. 1). This so called *hydrophobic core* is present in almost all globular water-solved protein structures. It is central to the CPSP approach and was also used in recent local search methods [75, 87]. Beneath the suggested energy potentials by Dill and coworkers ($e(S_i, S_j) = -1$ if S_i and S_j hydrophobic; 0 otherwise [33]), other potentials have been applied [17, 52].

Banavar and coworkers introduced the THP-model [13] to incorporate context specific hydrophobic contact contributions. Another extension, the HPNX-model introduced in [12, 17] with different potentials, distinguishes four different amino acid groups, namely Hydrophobic, Positively charged, Negatively charged, and X for all remaining neutral residues. Still, hydrophobic contacts are the strongest potentials. An improved version, the hHPNX-model [41], follows the amino acid grouping of the YhHX-model suggested by Crippen [28] where Alanine and Valine are treated as special group (h). Bornberg-Bauer introduced an integer conversion of the real valued YhHX-model that maintains approximately the same ratios of entries [17]. His potentials were corrected by Hoque and coworkers [41]. Full 20×20 pairwise potentials were pioneered by Miyazawa and Jernigan [62, 63] and derived from real protein structure information. Simplified potentials have been suggested in [15].

In first studies, Chan and Dill found contact-based potentials sufficient to enable realistic distributions of secondary structure elements in structure space [21] by the study of small molecules in the 3D cubic lattice. While it is clear to see that very simple lattice models, e.g. in 2D, are hard to map into the real protein structure space, Vendruscolo and Domany showed general limitations of contact-based potentials to mimic real protein structures [90]. Thus, to enable lattice proteins to predict real native structures,

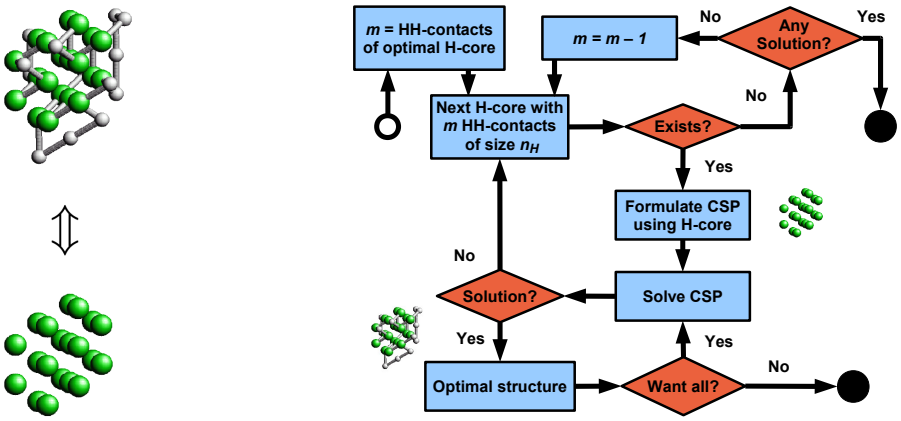

Figure 2: (left) Coherence between an optimal structure (top) and the optimally dense packing of H-monomers, i.e. its H-core (bottom) in the 3D-face-centered-cubic (FCC) HP-model. (right) CPSP Workflow. Start and end are marked by open and filled circles respectively. The number of hydrophobic amino acids within the sequence is denoted by n_H and m is initialized by the maximal number of hydrophobic contacts possible for n_H monomers in the underlying lattice. Figures from [55].

additional constraints are needed. For instance, the prediction and incorporation of secondary structure information yields very promising results [30, 31]. But even without such extensions, lattice protein models enable studies of general features of protein structure formation and related problems. They have been used to investigate the folding process [72], native structure properties [37], sequence evolution [18], cooperative/competitive folding [22], and co-translational folding [42, 79], to name but a few.

In the following, we will first focus on structure prediction the HP-model in three-dimensional lattices. Later on, exact PSP approaches for enhanced energy models will be discussed.

3 Constraint-based Protein Structure Prediction

As discussed in the introduction, the *Constraint-based Protein Structure Prediction* (CPSP) approach is currently the only exact and complete PSP method for lattice protein models. Its basic version is tailored for 3D backbone-only HP-models [8, 11, 61, 92] but was extended other models too. Among them the extension to the HPNX-model [7, 12] or to side chain structure models [60] both later discussed in detail. In contrast to most tools and approaches in the field, a CPSP implementation is available for local installation [25, 61] as well as ad hoc web usage [26, 60]. Thus, it has spawned the compilation of extensive benchmark data sets of protein-like sequences with various folding features [58]. Among them are the existence of a unique optimal fold, proven via CPSP, and the accessibility of this structure via unrestricted or co-translational folding using the LatPack package [48].

Within the HP-model, only hydrophobic contacts are contributing to the energy. It is therefore sufficient to consider and optimize H-monomer interactions only. The CPSP approach follows, as the CHCC method, the observation that HP-optimal structure feature an (almost) optimal packing of H-monomers and thus maximizing their contacts, see Fig. 2. This spawns the central idea of CPSP: if a structure shows an optimal packing of H-monomers it is has to be an HP-optimal structure! Following this idea, the CPSP screens (sub)optimal packings of H-monomers, so called H-cores, for their compatibility with the given sequence. If it is possible to identify a structure that confines the sequence's H-monomers to the current optimal H-core, an optimal structure was found. This structure threading is done via constraint programming techniques. The overall CPSP workflow is given in Fig. 2.

In the following, we will first sketch the structure threading step for a given sequence and H-core. In the next section the computation of H-cores is discussed. A detailed presentation of the whole approach if provided in [92].

3.1 Optimal Structure Identification

The CPSP approach uses constraint programming techniques to solve the PSP problem. Constraint programming enables the definition of constraint satisfaction problems (CSP) and offers a generic and

efficient framework for satisfiability checks [78]. A CSP is defined by a set of variables $\mathcal{X}$, their value domains $\mathcal{D}$ and the set of constraints $\mathcal{C}$ on $\mathcal{X}$, which define valid variable assignments, i.e. solutions. A sophisticated iterative reasoning and pruning of violating values from the variable domains combined with tuned search strategies enables an efficient identification and enumeration of constraint-conform solutions. For further details on general constraint programming techniques e.g. refer to [78].

The CSP formulated for the optimal structure identification step in the CPSP approach is based on both the protein's sequence S of length n and the current H-core of interest $\mathcal{H}$ in the underlying lattice L. An H-core $\mathcal{H} \subset L$ is a set of lattice points with maximally compact positioning (details on their generation are discussed later). Only S-compatible H-cores are considered, i.e. the size of the core equals the number of H-monomers in the sequence.

In the following, for simplicity we will first discuss the CSP for backbone-only models [8, 11, 61, 92]. For each backbone position, a lattice position variable $\mathcal{X}_i$ is defined. The domain $\mathcal{D}_i$ of $\mathcal{X}_i$ depends on the according amino acid S_i: if hydrophobic ($S_i = \mathrm{H}$), the variable domain is defined by the H-core positions ($\mathcal{D}_i = \mathcal{H}$). Otherwise, the monomer has to be placed outside of the core ($\mathcal{D}_i = L \setminus \mathcal{H}$). As discussed above, this ensures HP-optimality of any produced structure as long as the H-core shows an optimal packing. In order to encode valid structures only, additional constraints are enforced. First, self-avoidingness of the chain via global difference of assigned values [76] and second the chaining of successive monomers given the neighborhood of the lattice, i.e. $\exists \vec{x} \in \mathcal{D}_i : \exists \vec{y} \in \mathcal{D}_{i+1} : \vec{x} - \vec{y} \in N_L$. While formulated here for domains on lattice coordinates, a coordinate integer encoding can be used in order to apply standard finite domain constraint solvers [7, 55, 92]. Provided the H-core shows an maximal number of contacts, any solution satisfying these constraints will represent an optimal structure according to the HP-model.

If no solution was found, the next H-core with similar compactness is tried. If no optimally packed H-core yields a solution, it is proven that no optimal structure with lower or equal number of hydrophobic contacts exists. Thus, the CPSP approach relates to suboptimal H-cores, which show the highest non-optimal number of hydrophobic contacts and iterates the procedure. Therefore, solutions for such suboptimal H-cores are still globally optimal, since no solution for more compact H-cores was found. Usually no or only a few suboptimality iterations are necessary. Note, the constraint programming framework enables the enumeration of all valid solutions for a CSP. This was, the set of all optimal structures can be enumerated by the CPSP approach when screening all according H-cores. The full workflow is depicted in Fig. 2.

The complete enumeration of optimal structures has to exclude symmetric conformation. That is, identical structures resulting from rotation or reflection in the lattice have to be avoided. The CPSP approach employs efficient symmetry breaking techniques within the solution search [10] resulting in a fast and symmetry free solution enumeration. To this end, for each solution the CSP is enhanced by lattice specific constraints that exclude symmetric assignments within the remaining enumeration [3, 10]. The symmetry free number of optimal structures, known as a sequence's *degeneracy*, is often taken as measure of structural stability of the native fold [82].

When extending the approach to side chain lattice protein models [55, 60], only an adaption of the CSP is needed while the overall workflow is kept. The contact-based energy computation in side chain models is restricted to side chain monomers only [55]. In the HP-model, only hydrophobic side chain interactions are contributing to the energy. All other contacts are neutral. Therefore, the CSP is adapted as follows: for each amino acid two variables are defined, one for the backbone and one for the side chain monomer position. The backbone monomer is confined to positions outside of the current H-core, while a sequence specific domain assignment, analogous to the backbone model, is used for the side chain monomers. In addition to the connectivity constraints applied to the backbone monomers, also side chain-backbone connectivity for each amino acid has to be ensured. Self-avoidance is enforced for all variables, i.e. all monomer positions. This CPSP modeling extension enables for the first time the exact identification of optimal structures in the side chain HP-model in three-dimensional lattices [55, 60]. The complete enumeration of optimal side chain structures is enabled too, but the enormous structure space growth (see Fig. 1) and the high degeneracy of the HP-model yield often hundreds to millions optimal structures. This problem is faced when extending the CPSP approach for the enumeration of representatives for H-core equivalence classes, discussed in Sec. 4.

3.2 Hydrophobic Core Construction

An H-core can be generated from an optimal conformation by removing all bonds, and considering only H-monomers (see Fig. 3A). As introduced, the CPSP approach depends on the availability of (sub)optimal H-cores for the lattice of interest. This is based on the relation of the number of contacts in the H-core

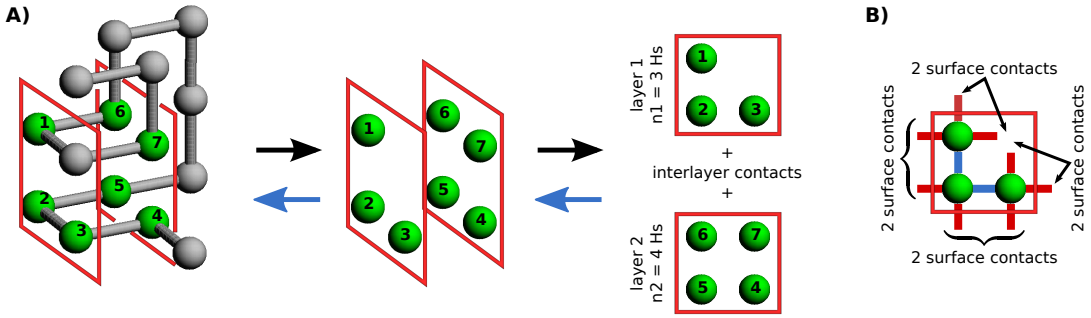

Figure 3: H-Core Construction. A) shows an optimal conformation for the sequence PPHPHHPPPPHHHHPPP in the 3D-cubic lattice together with the associated hydrophobic core and its decomposition into layers. B) Relation between HH-contacts (blue) and surface contacts (red) exemplified for layer 1.

and the number HH-contacts in a conformation featuring this H-core. Formally, the contact number of an H-core $\mathcal{H}$ is defined by

$$I(\mathcal{H}) = |\{\ \{\vec{x},\vec{y}\} \mid \vec{x},\vec{y} \in \mathcal{H}\ \wedge\ \vec{x}-\vec{y} \in N_L\ \}|. \tag{4}$$

Now let a conformation P for a sequence S have an associated H-core $\mathcal{H}$. Then the number of HH-contacts $E(S,P)$ relates to $I(\mathcal{H})$ by

$$E(S,P) = I(\mathcal{H}) - \text{number of HH-bonds}. \tag{5}$$

Here, HH-bonds denote the bonds between successive H-monomers in the sequence, which are not contributing to the energy. Thus, optimal conformations have also optimal or near optimal H-cores, where we define an H-core $\mathcal{H} \subset L$ to be optimal if its contact number $I(\mathcal{H})$ is maximal for this H-core size $|\mathcal{H}|$.

Figure 3A) shows the basic idea of using a hydrophobic core construction, as first introduced by Yue and Dill [95]. It also depicts the decomposition of H-cores in order to efficiently determine optimal ones. Here, an H-core is broken into individual layers (right part of Figure 3A), such that we can dissemble a cores contacts $I(\mathcal{H})$ into layer and interlayer contacts. Thus, the hydrophobic core construction reverses direction (indicated with blue arrows). Based on branch-and-bound techniques, (sub)optimal layers are generated, which are then composed into (sub)optimal hydrophobic cores. Finally, conformations are threaded onto the optimal cores as described above. Thus, the success of the CPSP approach relies on good bounds for the layer generation to effectively generate hydrophobic cores.

Therein, the basic question is the following: Given a specific distribution of H-monomers onto each layer (e.g., in Fig. 3A: $n_1 = 3$ and $n_2 = 4$), what is the maximal contact number of an associated H-core? As indicated before, one distinguishes between layer and interlayer contacts for this purpose, which we will illustrate with the example from Fig. 3. So let's concentrate first on the number of layer contacts of the three H-monomers in in layer 1 (Fig. 3B). This placement produces two HH-contacts. As shown by Yue and Dill [95], the number of layer contacts is related to the surface of the minimal rectangle around all H-monomers. For this, we need to identify the surface contacts of a specific layer, which is the number of unoccupied neighboring positions of H-monomers. Figure 3B highlights the 8 surface contacts for layer 1 in red. Since every H-monomer has four neighbors in a 3D cubic layer, we know that four times the number of H-monomers must be equal to twice the number of HH-contacts plus the number of surface contacts ($4 \times n = 2 \times \text{HH} + \text{surfaces}$). Thus, we get the equation $4 \times 3 = 2 \times 2 + 8$. This implies that minimizing the surface in a layer is maximizing the number of HH-contacts. Furthermore, the number of surface contacts is exactly the perimeter of the minimal rectangle around all H-monomers (both 8 in Fig. 3B). For that reason, a placement of n H-monomers that optimizes the number of HH-contacts is found by using the minimal rectangle (in the following called frame) with height $a = \lceil \sqrt{n} \rceil$ and width $b = \lceil \frac{n}{a} \rceil$, or vice versa.

This upper bound can be improved in several ways. If one considers four H-monomers to be placed in one layer, then the optimal frame has the dimension 2×2. However, four H-monomers cannot be placed into this frame if the protein's subsequence containing this four H-monomers is HHPHH. The reason is that the enclosed P, which is called P-singlet [95], cannot be placed outside the 2×2 frame, which contains all H-monomers, while being in contact to its flanking H-monomers. Thus, the optimal frame is not compatible with the sequence at hand. This can be solved by including these P-singlets in the number of H-monomers to be placed when calculating the optimal surrounding frame. Another improvement was introduced in [5] by using a special property of the cubic lattice. Given a point $\vec{p} = (x,y,z)$ in the cubic

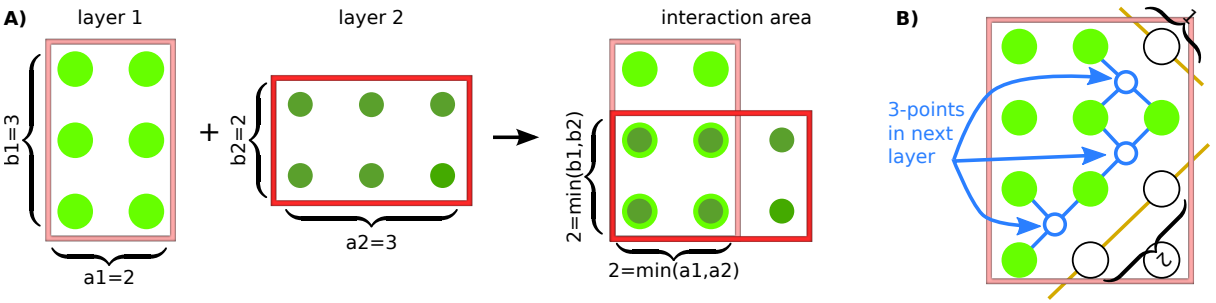

Figure 4: A) Two successive layers in the 3D cubic lattice filled with six H-monomers each. The associated frames have dimensions (a_1, b_1) and (a_2, b_2). The maximal number of interlayer contacts is $4 = \min(a_1, a_2) \times \min(b_1, b_2)$. B) Placement of H-monomers in a layer of the FCC lattice. The longest intersection of 45° diagonals with unoccupied positions are indicated with yellow diagonals. The sum of the associated quantities is three, which is the number of positions in the next FCC-layer which have exactly three H-neighbors in the current layer when filled with an H-monomer. These points are called 3-points and are indicated with blue circles, together with their possible contacts.

lattice, we define a point to be *even* if $x + y + z$ is even, and *odd* otherwise. Since the neighbors of any point in the cubic lattice differ by one in exactly one coordinate, an even point has only odd neighbors, and vice versa. In a sequence, we also have even and odd monomers, and this parity also translates to the parity of occupied positions. So instead of considering a distribution of the number of H-monomers per layer, one can consider distributions of even and odd H-monomers to layers [5]. If there are two even and two odd monomers, as in the case of Fig. 3A for layer 2, then the minimal frame has the dimension 2×2, which yields four HH-contacts. If, instead, three even and one odd (or vice versa) monomers are to be placed, then the optimal frame has dimension 2×3 with only three HH-contacts [5].

Concerning the number of interlayer contacts, the problem is relative easy for the cubic lattice since each H-monomer in a layer can have at most one HH-contact to exactly one H-monomer in the following layer. Thus, given two successive layer frames, one possible upper bound for the number of interlayer contacts is given by the maximal number of overlap positions between both frames. The dimensions of this overlap are defined by the minimum of the heights and widths of the associated frames (see Fig. 4).

The problem gets more complicated for the 3D FCC lattice. The layers, when using a corresponding splitting, are again equivalent to positions in the square lattice. Thus, most of the results for layers described above for the cubic lattice can be transferred to the layers in the FCC. However, the upper bound for the interlayer contacts is much more complex. To the best of our knowledge, only one upper bound for the interlayer contacts in the FCC lattice has been developed so far [4, 6]. The problem is that an H-monomer can have zero to four contacts to H-monomers in the next layer. For that reason, Backofen [6] developed bounds for the number of positions in the next layer having one to four contacts (called 1-, 2-, 3- and 4-points), given a frame with additional properties in the current layer. To give an example, it was shown that the number of 3-points for a given placement of H-monomers can be determined by considering the longest intersection of 45° diagonals with unoccupied positions in the frame [6] (see Fig. 4B). This implies that the $*$-point composition of the layers has to be considered for determining a bound on the interlayer contacts.

Once upper bounds for layer and interlayer contacts are found for a given H-core size, dynamic programming is used to identify optimal frame sequences for a contact bound [9, 11, 92]. These frame sequences are then instantiated to optimal H-cores via constraint programming. This frame sequence and H-core enumeration and its extension to suboptimal H-core generation is beyond the scope of this review and we refer to the according literature [9, 11, 91, 92].

4 Hydrophobic Core Equivalence

The hydrophobicity-focusing energy function in HP models results on average in a vast number of optimal structures. Since polar residues do not contribute to the energy, optimal structures usually show a much higher variation in the placement of polar than hydrophobic residues. While the latter form a compact core, polar residues are placed arbitrarily loose around it. This is even more severe in side chain models

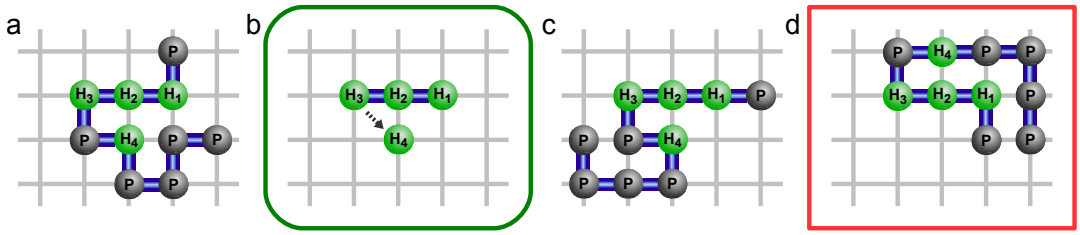

Figure 5: Different optimal structures (a,c) for the sequence PHHHPHPPPP in the 2D-square lattice that show the same relative H-monomer placement (b), i.e. are within one equivalence class. The structure on the right highlighted in red (d) is also part of this equivalence class when considering reflection. This exemplifies the symmetry problem for equivalence detection. Figures taken from [55].

where the majority of the monomers is unconstrained by the energy function, namely both backbone and polar side chains monomers.

When considering the relative positioning of a protein's hydrophobic monomers only, only few distinct patterns occur as depicted in Fig. 5. These H-core placements enable the definition of an *equivalence relation* $\overset{H}{\sim}$ in order to group optimal structures accordingly. Formally, two backbone-only structures $P = (P_1, \ldots, P_n)$ and $\hat{P} = (\hat{P}_1, \ldots, \hat{P}_n)$ for a sequence S are said to be equivalent if they show identical H-monomers placements according to some symmetrical shift. This is given by

$$P \overset{H}{\sim} \hat{P} \longleftrightarrow \exists_{r \in \mathcal{R}} \, \exists_{\vec{t} \in L} : \forall_{i|S_i = \mathtt{H}} : P_i = \hat{P}_i R_r + \vec{t}\,, \tag{6}$$

where $\mathcal{R}$ denotes the set of all symmetry functions (according to rotation and reflection) within the underlying lattice L, while R_r represents the rotation/reflection matrix for a symmetry $r \in \mathcal{R}$ (refer to [55, 92] for details). The translation vector $\vec{t} \in L$ shifts the symmetric structure into mapping. See Figure 5 for examples. This equivalence definition can be extended to side chain lattice protein models, see [55, 56]. This results in according equivalence classes within the structure space for a given sequence.

The CPSP approach can be used to enumerate all equivalence classes comprising optimal structures for a given protein [55, 56]. To this end, one representative structure per class is identified. It is based on the observation that only optimal structures confined to the same H-core can be equivalent. Thus, the overall CPSP workflow can be kept and only the CSP solution identification has to be adapted in order to avoid the enumeration of equivalent structures. This is achieved by a dedicated variable assignment strategy. Only for hydrophobic monomers, a complete solution enumeration is done excluding symmetries. For each H-monomer assignment, a satisfiability check for the placement of the remaining monomers is done, which results in a single representative optimal structure for the equivalence class if possible. Note, the symmetry problem is already solved within the CPSP workflow as discussed above.

Figure 6 compares the number of optimal structures (a sequence's *degeneracy*) with the number of equivalence classes for a large set of sequences. It reveals that the number of equivalence classes is several orders of magnitude smaller than a sequence's degeneracy. Furthermore, it reveals the extreme increase of optimal structures in side chain HP-models compared to backbone models due to energetically unconstrained polar and backbone monomers. Here, the difference to equivalence classes is even more extreme. The distributions of equivalence classes are similar when comparing backbone-only and side chain models, while backbone-only models show a slightly higher number on average. This results from the increased freedom in hydrophobic side chain monomer positioning in side chain models, since only neighborhood to the backbone monomer has to be ensured. In backbone-only models, each H-monomer has to be neighbored to the preceding and succeeding backbone monomer in the chain. Thus, optimal side chain model structures show on average a more compact H-core compared to optimal backbone-only structures for the same sequence. Since the number of H-cores decreases for increasing compactness, fewer H-cores and thus less equivalence classes are found for side chain structure models.

Since a sequence's degeneracy, which can be computed by the standard CPSP approach, is a measure of structural stability [82]. But as discussed, it is flawed in the HP-model due to the missing energetic constraints on the non-hydrophobic monomers. Thus, the number of equivalence structures was suggested as a new *measure of structural stability* in the HP model as an alternative to degeneracy [55, 56].

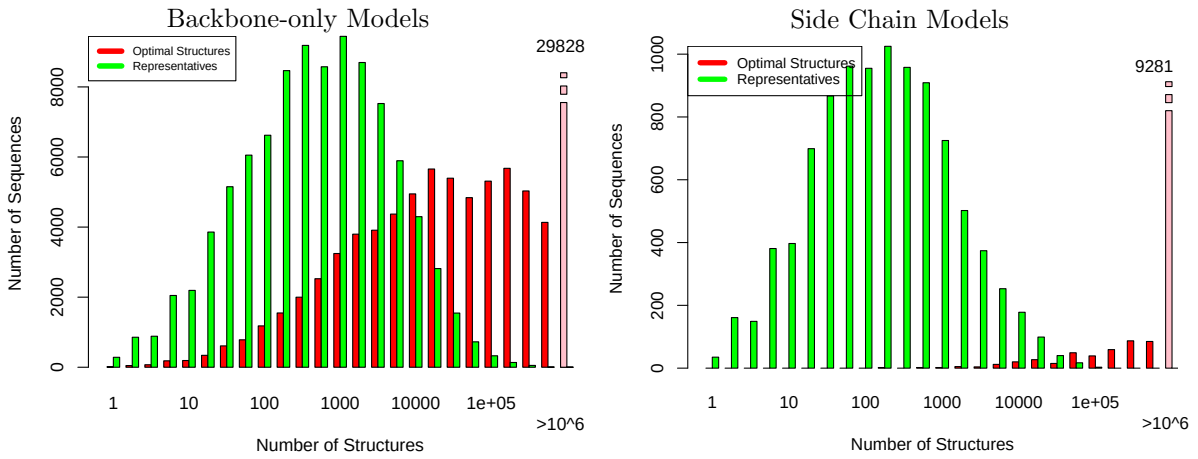

Figure 6: Distribution of the number of representatives (green) versus the overall number of optimal structures (degeneracy, red). They are exemplified for for HP models in the 3D-cubic lattice for sequence length 27. On the [left], results correspond to backbone-only predictions for 100,000 random sequences, the [right] figure depicts side chain model predictions for 10,000 random sequences. The distributions are only shown with exact values for degeneracy below 10^6, all sequences with a higher degeneracy are collected in the right most bar in pink representing "$> 10^6$". Figures taken from [55].

5 Enhanced Models

The HP-model implements a strong abstraction of the energetics underlying the protein folding process, since it only models hydrophobic forces. As discussed above, several more fine grained energy models have been introduced in literature. While the CPSP is currently the only exact method for protein structure prediction in lattice protein models, it is intrinsically tailored towards the HP-model. Nevertheless, it can be extended to HP-related energy models as HPNX-model [7, 12, 92]. Furthermore, it was shown that the HP-model can be used within hierarchical approaches to enhance the prediction of optimal structures in more sophisticated energy models providing 20×20-potentials [86, 87]. Both directions are discussed in the following.

5.1 Prediction in HP-type models

The HPNX model [17] is an extension of the HP-model, where polar amino acids are split into (P)ositively and (N)egatively charged amino acids and neutral (X) monomers. yields within N-P-contacts Still, it focuses on hydrophobic interactions, such that H-H-contacts have the strongest energy contribution of 4, while H-contacts to all other amino acids are neutral. N-P-contacts have an energy contribution of -1, while P-P- and N-N-contacts are penalized by +1. Contacts to X are neutral (contact potential is 0).

Backofen and Will have introduced an exact constraint programming-based PSP approach for the HPNX-model [3, 7, 12, 92]. The approach applies a branch-and-bound search on a constraint optimization problem, which encodes feasible lattice protein structures (similar to the CPSP-model) in concert with the described HPNX-energy function to be minimized. The performance is based on sophisticated lower energy bounds for partially defined solution structures, in order to prune energetically unfavorable parts of the search space. This way, the authors yield an efficient and exact PSP method that allows optimal structure prediction and enumeration [7, 12, 92].

5.2 Optimized prediction in full potential models

For more sophisticated energy models, as e.g. the 20×20 pairwise potentials by Miyazawa and Jernigan [62, 63], no efficient and exact PSP approach is known so far. Here, usually local search strategies are applied. Local search is a generic optimization scheme that minimizes a given objective function. Given a neighborhood relation within the search space to traverse, local search is able to follow the gradient to identify local or even global minima [47, 89]. Local search approaches work well in practice for more sophisticated energy functions but usually require a large number of steps to converge. Each search usually starts with a random start conformation to enable a good coverage of the structure space.

Ullah and co-workers introduced a protein folding simulation procedure that employs two stages of optimization in order to find structures of minimal energy [86]. The method mimics the hydrophobic collapse during protein folding, i.e. the hydrophobicity-driven, fast formation of an initial structure and the following folding into the functional fold. Thus, the protein sequence is first collapsed into a compact HP-optimal structure using the CPSP approach. Successively, the CPSP output is given as input to a Simulated Annealing-based local search procedure which employs the pairwise 20×20 energy potentials introduced by Berrera and co-workers [15]. Thus, the method combines the efficiency and performance of exact structure prediction in the simpler HP-model with established local search schemes in the enhanced energy model using pull moves [16, 51]. When comparing this two-step optimization pipeline with standard Simulated Annealing procedures based on random start structures, faster convergence to energetically lower structures is observed. This results from the energetically lower start conformations, since HP-optimal structures show on average about two orders of magnitude lower energies in the enhanced energy model compared to random structures [55, 86]. The impact of such an HP-optimal initialization scheme seems to be strongly connected to the subsequently applied optimization procedure, since Rashid and colleagues found their genetic algorithm to be more efficient with random initializations [74].

6 Summary

Lattice protein models enable detailed studies of protein folding processes in a discretized, finite but yet exponentially large structure space. The latter renders exhaustive structure enumeration useless for interesting protein lengths. In order to find native, i.e. energy optimal, structures, mainly local search schemes are applied, which neither ensure to find optimal structure nor enable an thorough investigation of the lowest energy spectrum. Only for hydrophobicity-focusing energy functions, namely the HP- and HPNX-model, exact methods are known. They are based on the *Constraint-based Protein Structure Prediction* (CPSP) approach [8, 11, 60, 61] applying efficient constraint programming techniques [3, 10].

The CPSP approach uses precomputed sets of maximally compact H-cores in order to identify optimal structures only. The computation of (sub)optimal H-cores is a hard computational problem on its own and was solved based on dynamic programming and constraint programming, too [3, 6, 9, 92]. The sequence-independence of H-cores enables a precomputation of an H-core database and its recomputation-free use within the CPSP workflow. This ensures a fast and reliable prediction of HP-optimal structures in 3D lattices. While tailored for the HP-model, several extensions of the CPSP approach have produced exact methods for other lattice protein models and applications. Beside the mentioned extension to the HPNX-model [7, 12, 61, 92], side chain models were targeted too [55, 60].

The CPSP approach is the only method enabling a complete and exclusive enumeration of optimal structures. This revealed an extremely high degeneracy, i.e. number of optimal structures, within the HP-model. Most of the optimal structures show an equivalent H-monomer placement, while they are only differing in the energetically unconstrained monomers. An extension of the CPSP approach enables the enumeration of according equivalence classes [55, 56], revealing a dramatically lower number of distinct H-monomer positionings among optimal structures. This phenomenon is most prominent in side chain lattice models. Given the properties of the hydrophobicity focusing energy function, the number of equivalence classes was suggested as an alternative measure of structural stability beside the commonly used degeneracy. Both measures can only be exactly identified using the CPSP method.

The existence of a unique native fold, i.e. lowest energy conformation, is a common criterion to name a model sequence protein-like. So far, the CPSP approach is the only method to enable the identification of such sequences. Thus, it was used to generate benchmark sets of protein-like sequences showing different folding properties [58]. Beside the unique native fold, its accessibility via co-translational and unrestricted folding is taken into account. Furthermore, the CPSP approach is often used as reference to determine the lowest energy accessible for a given sequence to evaluate local search schemes [23, 34].

The use of HP-optimal structure samples was shown to boost protein structure prediction in more sophisticated models [55, 74, 86, 87]. Here, no exact methods are available and thus local search schemes are applied. It was shown that the initialization of local search with HP-optimal structures increases convergence and enhances prediction results. Such a two-step optimization scheme resembles the hydrophobic collapse during globular protein folding.

Beside protein structure prediction, the CPSP approach was also used to tackle the inverse folding problem [55, 61], i.e. to find a sequence, which has a given structure as its native fold. Here, the CPSP method was needed for (a) to confirm that the targeted structure is among the optimal ones and (b) to check if a sequence's degeneracy is within a targeted upper bound (usually 1). The method can be easily extended to be constrained by equivalence classes if needed. With such a tool at hand, neutral evolution

studies are enabled [18–20]. Therein, the sequence space is screened for mutation-connected subsets that show identical native folds [61, 92].

References

[1] Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science*, 181(96):223 – 230.

[2] Angelani, L. and Ruocco, G. (2009). Saddles of the energy landscape and folding of model proteins. *Europhysics Letters*, 87(1):18002.

[3] Backofen, R. (2000a). *Optimization Techniques for the Protein Structure Prediction Problem*. Habilitation, Ludwig-Maximilians-Universität München. Available at `http://www.bioinf.uni-freiburg.de/Publications/`.

[4] Backofen, R. (2000b). An upper bound for number of contacts in the HP-model on the Face-Centered-Cubic Lattice (FCC). In Giancarlo, R. and Sankoff, D., editors, *Proceedings of the 11th Annual Symposium on Combinatorial Pattern Matching (CPM 2000)*, volume 1848 of *LNCS*, pages 277–292, Montréal, Canada. Springer-Verlag, Berlin.

[5] Backofen, R. (2001). The protein structure prediction problem: A constraint optimisation approach using a new lower bound. *Constraints*, 6:223–255.

[6] Backofen, R. (2004). A polynomial time upper bound for the number of contacts in the HP-model on the face-centered-cubic lattice (FCC). *J Discr Alg*, 2(2):161–206.

[7] Backofen, R. and Will, S. (1998). Structure prediction in an HP-type lattice with an extended alphabet. In *Proc of German Conference on Bioinformatics (GCB'98)*.

[8] Backofen, R. and Will, S. (2001a). Fast, constraint-based threading of HP-sequences to hydrophobic cores. In *Proc. of the 7th International Conference on Principle and Practice of Constraint Programming (CP'2001)*, volume 2239 of *LNCS*, pages 494–508.

[9] Backofen, R. and Will, S. (2001b). Optimally compact finite sphere packings — hydrophobic cores in the FCC. In *Proc of CPM'01*, volume 2089 of *LNCS*, pages 257–272. Springer.

[10] Backofen, R. and Will, S. (2002). Excluding symmetries in constraint-based search. *Constraints*, 7(3):333–349.

[11] Backofen, R. and Will, S. (2006). A constraint-based approach to fast and exact structure prediction in three-dimensional protein models. *Constraints*, 11(1):5–30.

[12] Backofen, R., Will, S., and Bornberg-Bauer, E. (1999). Application of constraint programming techniques for structure prediction of lattice proteins with extended alphabets. *Bioinformatics*, 15(3):234–242.

[13] Banavar, J. R., Cieplak, M., and Maritan, A. (2004). Lattice tube model of proteins. *Phys. Rev. Lett.*, 93(23):238101.

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

[15] Berrera, M., Molinari, H., and Fogolari, F. (2003). Amino acid empirical contact energy definitions for fold recognition in the space of contact maps. *BMC Bioinformatics*, 4(1):8.

[16] Böckenhauer, H.-J., Ullah, A. Z. D., Kapsokalivas, L., and Steinhöfel, K. (2008). A local move set for protein folding in triangular lattice models. In *Proc. of WABI '08*, LNBI, pages 369–381.

[17] Bornberg-Bauer, E. (1997). Chain growth algorithms for HP-type lattice proteins. In *Proceedings of RECOMB'97*, pages 47–55.

[18] Bornberg-Bauer, E. (2002). Randomness, structural uniqueness, modularity and neutral evolution in sequence space of model proteins. *Z. Phys. Chem.*, 216:139 – 154.

[19] Bornberg-Bauer, E., Beaussart, F., Kummerfeld, S., Teichmann, S., and Weiner, J. (2005). The evolution of domain arrangements in proteins and interaction networks. *Cell Mol Life Sci.*, 62(4):435–445.

[20] Bornberg-Bauer, E. and Chan, H. S. (1999). Modeling evolutionary landscapes: mutational stability, topology, and superfunnels in sequence space. *Proc. Natl. Acad. Sci.*, 96(19):10689–10694.

[21] Chan, H. S. and Dill, K. A. (1990). Origins of structure in globular proteins. *Proc Natl Acad Sci USA*, 87(16):6388–92.

[22] Citossi, M. and Guigliarelli, G. (2005). Lattice protein models: A computatinoal approach to folding and aggregation phenomena. In *Frontiers of Fundamental Physics*, volume IV, pages 355–358. Springer Netherlands.

[23] Citrolo, A. G. and Mauri, G. (2014). A local landscape mapping method for protein structure prediction in the HP model. *Natural Computing*, pages 1–11.

[24] Clark, P. L. (2004). Protein folding in the cell: reshaping the folding funnel. *Trends Biochem Sci*, 29(10):527–534.

[25] CPSP-home (2008). CPSP-tools : Constraint-based protein structure prediction. Available as an open-source package from `http://www.bioinf.uni-freiburg.de/sw/cpsp/`.

[26] CPSP-webtools (2009). CPSP-webtools : Constraint-based protein structure prediction webserver. Available at `http://cpsp.informatik.uni-freiburg.de`.

[27] Crescenzi, P., Goldman, D., Papadimitriou, C., Piccolboni, A., and Yannakakis, M. (1998). On the complexity of protein folding. *Journal of Computational Biology*, 5(3):423–65.

[28] Crippen, G. M. (1991). Prediction of protein folding from amino acid sequence over discrete conformation spaces. *Biochemistry*, 30(17):4232–4237.

[29] Dal Palù, A., Dovier, A., and Fogolari, F. (2004). Constraint Logic Programming approach to protein structure prediction. *BMC Bioinformatics*, 5:186.

[30] Dal Palù, A., Dovier, A., and Pontelli, E. (2005). A new constraint solver for 3D lattices and its application to the protein folding problem. In *Proc. of Logic for Programming, Artificial Intelligence, and Reasoning (LPAR'05)*, pages 48–63. Springer.

[31] Dal Palù, A., Dovier, A., and Pontelli, E. (2007). A constraint solver for discrete lattices, its parallelization, and application to protein structure prediction. *Softw., Pract. Exper.*, 37(13):1405–1449.

[32] Dal Palù, A., Dovier, A., and Pontelli, E. (2010). Computing approximate solutions of the protein structure determination problem using global constraints on discrete crystal lattices. *J of Data Mining and Bioinformatics*, 4(1):1 – 20.

[33] Dill, K. A., Bromberg, S., Yue, K., Fiebig, K. M., Yee, D. P., Thomas, P. D., and Chan, H. S. (1995). Principles of protein folding – a perspective of simple exact models. *Protein Science*, 4:561–602.

[34] Dotu, I., Cebrián, M., van Hentenryck, P., and Clote, P. (2008). Protein structure prediction with large neighborhood constraint programming search. In *Proc of CP'08*, volume 5202 of *LNCS*, pages 82–96. Springer.

[35] Dotu, I., Cebrián, M., van Hentenryck, P., and Clote, P. (2011). On lattice protein structure prediction revisited. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 8(6):1620–1632.

[36] Finkelstein, A. V. and Badretdinov, A. Y. (1997). Rate of protein folding near the point of thermodynamic equilibrium between the coil and the most stable chain fold. *Folding and Design*, 2(2):115 – 121.

[37] Hart, W. and Newman, A. (2006). *Handbook of Molecular Biology*, chapter Protein structure prediction with lattice models, pages 1–24. Chapman & Hall/CRC Computer and Information Science Series. CRC Press, New York.

[38] Hart, W. E. and Istrail, S. C. (1996). Fast protein folding in the hydrophobic-hydrophilic model within three-eighths of optimal. *Journal of Computational Biology*, 3(1):53–96.

[39] Hart, W. E. and Istrail, S. C. (1997). Lattice and off-lattice side chain models of protein folding: linear time structure prediction better than 86% of optimal. *Journal of Computational Biology*, 4(3):241–59.

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

[41] Hoque, T., Chetty, M., and Sattar, A. (2009). Extended HP model for protein structure prediction. *Journal of Computational Biology*, 16(1):85–103.

[42] Huard, F. P. E., Deane, C. M., and Wood, G. R. (2006). Modelling sequential protein folding under kinetic control. *Bioinformatics*, 22(14):e203–210.

[43] Irbäck, A. and Sandelin, E. (2000). On hydrophobicity correlations in protein chains. *Biophys J*, 79(5):2252–2258.

[44] Istrail, S. and Lam, F. (2009). Combinatorial algorithms for protein folding in lattice models: A survey of mathematical results. *Commun. Inf. Syst.*, 9(4):303–346.

[45] Jefferys, B., Kelley, L., and Sternberg, M. J. E. (2010). Protein folding requires crowd control in a simulated cell. *Journal of Molecular Biology*, 397(5):1329–1338.

[46] Karplus, M. and Kuriyan, J. (2005). Molecular dynamics and protein function. *PNAS*, 102(19):6679–6685.

[47] Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, 220(4598):671–680.

[48] LatPack-home (2008). LatPack : Lattice protein folding package. Available as an open-source package from `http://www.bioinf.uni-freiburg.de/Software/`.

[49] Lau, K. F. and Dill, K. A. (1989). A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromol.*, 22(10):3986–3997.

[50] Laurèn, J., Gimbel, D. A., Nygaard, H. B., Gilbert, J. W., and Strittmatter, S. M. (2009). Cellular prion protein mediates impairment of synaptic plasticity by amyloid-β oligomers. *Nature*, 457:1128–1132.

[51] Lesh, N., Mitzenmacher, M., and Whitesides, S. (2003). A complete and effective move set for simplified protein folding. In *Proceedings of the seventh annual international conference on Research in computational molecular biology (RECOMB'03)*, pages 188–195.

[52] Li, H., Helling, R., Tang, C., and Wingreen, N. (1996). Emergence of preferred structures in a simple model of protein folding. *Science*, 273(5275):666–669.

[53] Liu, J., Song, B., Liu, Z., Huang, W., Sun, Y., and Liu, W. (2013). Energy-landscape paving for prediction of face-centered-cubic hydrophobic-hydrophilic lattice model proteins. *Phys. Rev. E*, 88:052704.

[54] Maher, B., Albrecht, A. A., Loomes, M., Yang, X.-S., and Steinhfel, K. (2014). A firefly-inspired method for protein structure prediction in lattice models. *Biomolecules*, 4(1):56–75.

[55] Mann, M. (2011). *Computational Methods for Lattice Protein Models*. PhD thesis, Albert-Ludwigs-University Freiburg. Available at `http://www.freidok.uni-freiburg.de/volltexte/8156/`.

[56] Mann, M., Backofen, R., and Will, S. (2009a). Equivalence classes of optimal structures in HP protein models including side chains. In *Proceedings of the Fifth Workshop on Constraint Based Methods for Bioinformatics (WCB09)*.

[57] Mann, M. and Dal Palù, A. (2010). Lattice model refinement of protein structures. In *Proc of WCB'10*, page 7. arXiv:1005.1853.

[58] Mann, M., Maticzka, D., Saunders, R., and Backofen, R. (2008a). Classifying protein-like sequences in arbitrary lattice protein models using LatPack. *HFSP Journal*, 2(6):396–404. Special issue on protein folding: experimental and theoretical approaches.

[59] Mann, M., Saunders, R., Smith, C., Backofen, R., and Deane, C. M. (2012). Producing high-accuracy lattice models from protein atomic co-ordinates including side chains. *Advances in Bioinformatics*, 2012(Article ID 148045):6.

[60] Mann, M., Smith, C., Rabbath, M., Edwards, M., Will, S., and Backofen, R. (2009b). CPSP-web-tool : a server for 3D lattice protein studies. *Bioinformatics*, 25(5):676–677.

[61] Mann, M., Will, S., and Backofen, R. (2008b). CPSP-tools - exact and complete algorithms for high-throughput 3D lattice protein studies. *BMC Bioinformatics*, 9:230.

[62] Miyazawa, S. and Jernigan, R. L. (1985). Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules*, 18(3):534–552.

[63] Miyazawa, S. and Jernigan, R. L. (1996). Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J Mol Biol*, 256(3):623–44.

[64] Nardelli, M., Tedesco, L., and Bechini, A. (2013). Cross-lattice behavior of general aco folding for proteins in the hp model. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, SAC '13, pages 1320–1327. ACM.

[65] Newman, A. (2002). A new algorithm for protein folding in the HP model. In *Proceedings of the thirteenth annual ACM-SIAM symposium on Discrete algorithms*.

[66] Ngo, J. T. and Marks, J. (1992). Computational complexity of a problem in molecular structure prediction. *Protein Eng.*, 5(4):313–321.

[67] Nunnally, B. K. and Krull, I. S., editors (2003). *Prions and Mad Cow Disease*. CRC Press.

[68] Pace, C., Shirley, B., McNutt, M., and Gajiwala, K. (1996). Forces contributing to the conformational stability of proteins. *FASEB J.*, 10(1):75–83.

[69] Park, B. and Levitt, M. (1995). The complexity and accuracy of discrete state models of protein structure. *J Mol Biol*, 249:493–507.

[70] Perdomo-Ortiz, A., Dickson, N., Drew-Brook, M., Rose, G., and Aspuru-Guzik, A. (2012). Finding low-energy conformations of lattice protein models by quantum annealing. *Sci Rep*, 2:571.

[71] Perunov, N. and England, J. L. (2014). Quantitative theory of hydrophobic effect as a driving force of protein structure. *Protein Science*, 23(4):387–399.

[72] Potzsch, S., Scheuermann, G., Wolfinger, M., Flamm, C., and Stadler, P. (2006). Visualization of lattice-based protein folding simulations. In *In Proc. of V'06: Conference on Information Visualization*, pages 89–94.

[73] Prusiner, S. B. (1998). Prions. *Proceedings of the National Academy of Sciences of the United States of America*, 95(23):13363–13383.

[74] Rashid, M. A., Newton, M. H., Hoque, M. T., and Sattar, A. (2013a). Mixing energy models in genetic algorithms for on-lattice protein structure prediction. *BioMed Research International*, 2013:15.

[75] Rashid, M. A., Newton, M. H., Hoque, M. T., Shatabda, S., Pham, D., and Sattar, A. (2013b). Spiral search: a hydrophobic-core directed local search for simplified PSP on 3D FCC lattice. *BMC Bioinformatics*, 14(Suppl 2):S16.

[76] Régin, J.-C. (1994). A filtering algorithm for constraints of difference in CSPs. In *Proc. of 12th National Conference on AI*, pages 362–367.

[77] Rose, G. D., Fleming, P. J., Banavar, J. R., and Maritan, A. (2006). A backbone-based theory of protein folding. *PNAS*, 103(45):16623 – 16633.

[78] Rossi, F., van Beek, P., and Walsh, T. (2006). *Handbook of Constraint Programming (Foundations of Artificial Intelligence)*. Elsevier Science Inc., New York, NY, USA.

[79] Saunders, R., Mann, M., and Deane, C. (2011). Signatures of co-translational folding. *Biotechnology Journal, Special issue: Protein folding in vivo*, 6(6):742–751. RS and MM have contributed equally to this work.

[80] Scott, M. D. and Frydman, J. (2003). Aberrant protein folding as the molecular basis of cancer. In *Protein Misfolding and Disease*, volume 232 of *Methods in Molecular Biology*, pages 67–76. Humana Press.

[81] Shatabda, S., Newton, M. A. H., Pham, D. N., and Sattar, A. (2013). A hybrid local search for simplified protein structure prediction. In *International Conference on Bioinformatics Models, Methods and Algorithms*, page 6.

[82] Shortle, D., Chan, H. S., and Dill, K. A. (1992). Modeling the effects of mutations on the denatured states of proteins. *Prot Sci*, 1:201–215.

[83] Smith, A. (2003). Protein misfolding. *Nature*, 426(6968 (Insight)):883–909. Special Insight issue on protein misfolding edited by A. Smith.

[84] Tsay, J.-J. and Su, S.-C. (2013). An effective evolutionary algorithm for protein folding on 3d fcc hp model by lattice rotation and generalized move sets. *Proteome Science*, 11(Suppl 1):S19.

[85] Tyers, M. and Mann, M. (2003). From genomics to proteomics. *Nature*, 422:193–197.

[86] Ullah, A. D., Kapsokalivas, L., Mann, M., and Steinhöfel, K. (2009). Protein folding simulation by two-stage optimization. In *Proc. of ISICA'09*, volume 51 of *CCIS*, pages 138–145, Wuhan, China. Springer.

[87] Ullah, A. D. and Steinhöfel, K. (2010). A hybrid approach to protein folding problem integrating constraint programming with local search. *BMC Bioinformatics*, 11(Suppl 1):S39.

[88] Unger, R. and Moult, J. (1993). Finding the lowest free energy conformation of a protein is an NP-hard problem: proof and implications. *Bull Math Biol*, 55(6):1183–98.

[89] Černý, V. (1985). Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm. *J of Optimization Theory and Applications*, 45(1):41–51.

[90] Vendruscolo, M. and Domany, E. (1998). Pairwise contact potentials are unsuitable for protein folding. *The Journal of Chemical Physics*, 109(24):11101–11108.

[91] Will, S. (2002). Constraint-based hydrophobic core construction for protein structure prediction in the face-centered-cubic lattice. In *Proc. of the Pacific Symposium on Biocomputing*, pages 661–672.

[92] Will, S. (2005). *Exact, Constraint-Based Structure Prediction in Simple Protein Models*. PhD thesis, Friedrich-Schiller-Universität Jena. Available at `http://www.bioinf.uni-freiburg.de/Publications/`.

[93] Wittung-Stafshede, P. (2002). Role of cofactors in protein folding. *Acc Chem Res*, 35(4):201–208.

[94] Ying, B. W., Taguchi, H., and Ueda, T. (2006). Co-translational binding of groel to nascent polypeptides is followed by post-translational encapsulation by groes to mediate protein folding. *J Biol Chem*, 281(31):21813–21819.

[95] Yue, K. and Dill, K. A. (1995). Forces of tertiary structural organization in globular proteins. *Proc Natl Acad Sci.*, 92(1):146–150.