Equivalence Classes of Optimal Structures in HP Protein Models Including Side Chains

Martin Mann, Rolf Backofen, and Sebastian Will

University of Freiburg, Bioinformatics, 79110 Freiburg, Germany,
{mmann,backofen,will}@informatik.uni-freiburg.de

Abstract. Lattice protein models, as the Hydrophobic-Polar (HP) model, are a common abstraction to enable exhaustive studies on structure, function, or evolution of proteins. A main issue is the high number of optimal structures, resulting from the hydrophobicity-based energy function applied. We introduce an equivalence relation on protein structures that correlates to the energy function. We discuss the efficient enumeration of optimal representatives of the corresponding equivalence classes and the application of the results.

1 Introduction

Proteins are the central players in the game of life. They are involved in almost all processes in cells and organisms, comprising replication, metabolism, and movement. To be able to perform their specific functions, proteins have to adopt a certain fold or structure, which is encoded by the protein's sequence. Thus, knowledge of a protein's structure elucidates the mechanisms it is involved.

Currently, it is not possible to calculate a protein's functional fold from its sequence nor to simulate the whole folding process in detail. Simplified protein models are used to reduce the computational complexity. A common abstraction are lattice proteins [5, 6]. Here, the structure space a protein can adopt is discretized and allows for efficient folding simulations [8, 13]. Nevertheless, it is difficult to determine minimal energy structures, which represent the functional folds in such models. Even in the most simple Hydrophobic-Polar (HP) model [7], the optimal structure prediction problem stays computationally hard (NP-complete) [4]. Despite this complexity, a fast calculation of non-symmetrical optimal structures in the HP model is possible using constraint programming techniques applied in the *Constraint-based Protein Structure Prediction (CPSP)* approach [3, 9, 10, 15].

Recently, we have introduced a significantly improved local search scheme for lattice protein folding simulations [14] using a full Miyazawa-Jernigan energy potential [11]. We take advantage of the efficient CPSP approach and initialize the folding simulations with optimal structures from the simpler HP model. This incorporates the phenomenon of hydrophobic collapse of protein structures, a driving force at the beginning of the folding process [1]. The already compact structures from the CPSP application form the starting point of the folding

driven by more complex interactions. This scheme outperforms folding simulations using a standard initialization with random structures and yields better results within shorter simulation time [14]. To increase efficiency of local search methods, usually many optimization runs from different starting points are done. Since the set of all HP-optimal structures is usually too large as a starting set, we are interested in a smaller subset that still covers the structural diversity of the whole set as good as possible. We achieve this by enumerating optimal structures that maintain a given minimal distance to each other.

Due to the hydrophobicity-focusing energy function, proteins in HP models show on average a huge 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.

Here, we introduce an equivalence relation to partition the set of (optimal) structures into according classes. Two structures are defined to be equivalent, iff they do not differ in the placements of their hydrophobic residues. We introduce an extension to the CPSP approach that enables an efficient calculation of the number of equivalence classes of optimal structures via enumerating one representative per class. The approach is presented for backbone-only and side chain incorporating HP models. We show that a sequence's number of representatives (later defined as core-degeneracy) is several magnitudes smaller than the overall number of all optimal structures (degeneracy).

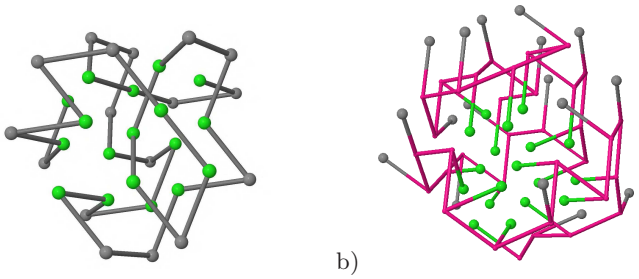
Thus, the set of optimal representatives is well placed to be used within the combined approach of CPSP and local search [14]. Furthermore, we propose another application of the equivalence classes: Since the equivalence relation is highly correlated to the HP energy function, the number of classes might be a better measure of structural stability than a sequences' degeneracy [12].

2 Preliminaries

A lattice protein in the HP model is specified by its sequence $S \in \{H, P\}^n$, where H and P denote hydrophobic and polar monomers, respectively. The structure positions are confined to nodes of a regular lattice $L \subseteq \mathbb{Z}^3$. A valid *backbone-only* structure $C \in L^n$ of length n is a self-avoiding walk (SAW) in the underlying lattice L, i.e. it holds connectivity $\forall_{1 \leq i < n} : (C_i - C_{i+1}) \in N_L$ and self-avoidance $\forall_{1 \leq i < j \leq n} : C_i \neq C_j$, where N_L denotes the set of distance vectors between neighbored points in L. An example is shown in Fig. 1a). The energy of a lattice protein structure is given by non-consecutive HH-contacts:

$$E(S, C) = \sum_{\substack{1 \leq i < j \leq n \\ (i+1) < j}} \begin{cases} -1 & : (C_i - C_j) \in N_L \wedge S_i = S_j = H \\ 0 & : \text{otherwise} \end{cases} \tag{1}$$

An *optimal structure* minimizes the energy function. The number of optimal structures is denoted as *degeneracy* of a sequence and is an important measure of structural stability [12].



a) b)

Fig. 1. Optimal structures of `HPPHHPPPHPHHPHHPPHPHPPHHHPHHPPHPHPH` in the face-centered-cubic lattice. (a) backbone-only model with energy -50, (b) side chain model with energy -55. Colors: green - H monomers, gray - P monomers, red - backbone in side chain models. Visualization by `HPview` from CPSP-package [10].

The *CPSP-approach* by Backofen and Will [3] enables the calculation of a sequence's degeneracy without full structure space enumeration [15]. It utilizes the observation that optimal structures show a (nearly) optimal packing of H monomers. Thus, the CPSP-approach can be sketched in two major steps:

1. *H-core construction:* Given the number n_H of H monomers from the target sequence S, all optimal packings of n_H monomers are calculated. These *optimal H-cores* show the maximal number of contacts possible. For a fixed sequence S and the corresponding n_H, we denote the set of optimal H-cores with $\mathcal{O}$. The calculation of $\mathcal{O}$ is computationally difficult on its own and was solved by us using constraint programming [2, 3].
2. *Structure threading:* Given S and $\mathcal{O}$ only structures are enumerated where the H monomers of S are confined to an optimal H-core $O \in \mathcal{O}$, i.e. they are "threaded" through the H-cores. Since all O show the maximally possible number of contacts between H monomers, each resulting structure is optimal according to Eq. 1 as well. The structure threading is done by solving a Constraint Satisfaction Problem (CSP) for each $O \in \mathcal{O}$ as given below.

Since step 1 depends only on the number of H monomers n_H and no further property of any sequence, we can precalculate the H-cores for different n_H and store them in a database. This significantly speeds up the approach and reduces the computation time to step 2, i.e. the structure threading.

It might happen, that we find no appropriate structure threading for a sequence S and the according set of optimal H-cores $\mathcal{O}$. Thus, we revert to the set of the best suboptimal H-cores $\mathcal{O}'$ that show at least one contact less than an optimal H-core $O \in \mathcal{O}$ and iterate the procedure. Still it holds: the first successive structure threading is an optimal structure, since no H monomer packing with more contacts was found before. Further details on the CPSP approach in [3].

The CSPs solved in step 2 are given by $(\mathcal{X}, \mathcal{D}, \mathcal{C})$, where we denote the set of variables $\mathcal{X}$, their domains $\mathcal{D}$, and a set of constraints $\mathcal{C}$. For each monomer $S_i \in S$ a variable $X_i \in \mathcal{X}$ is introduced. The SAW is modeled by a sequence of binary neighboring constraints $\text{neigh}(X_i, X_{i+1})$ and a global $\text{alldiff}(\mathcal{X})$ to enforce the self-avoidingness. The optimal H-core $O \in \mathcal{O}$ is used to define the domains $\mathcal{D}$:

$\forall_{i:S_i=H} : \mathcal{D}(X_i) = O$ and $\forall_{i:S_i=P} : \mathcal{D}(X_i) = L \setminus O$. Thus, if we find a solution of such a CSP, i.e. an assignment $a_i \in \mathcal{D}(X_i)$ for each variable that satisfies all constraints in $\mathcal{C}$, it will minimize the energy function in Eq. 1, i.e. an optimal structure.

3 Representative Optimal Structures

Revisiting the CSP we can see, that P monomers are constrained only by the SAW constraints. Imagine a sequence with a long tail of P monomers. Each valid placement of the subchain in front of the tail can be combined with a combinatorial number of possible SAWs of the tail. This leads to the immense degeneracy in the HP model.

Therefore, we set up an *equivalence relation* $\overset{H}{\sim}$ on structures (Eq. 2) that decomposes the set of all (optimal) structures into equivalence classes. In the following, the number of equivalence classes of optimal structures is denoted as *core-degeneracy*. As given by Eq. 2, structures from different equivalence classes differ in at least one H monomer placement.

$$C \overset{H}{\sim} \hat{C} \Leftrightarrow \forall_{i|S_i=H} : C_i = \hat{C}_i. \tag{2}$$

The representative enumeration (that corresponds to core-degeneracy calculation) can be done via an extension of the CPSP approach presented in Sec. 2. Instead of calculating all optimal structures, we want to calculate only one representative per equivalence class. This has to be ensured at two stages: (I) the solutions of each single CSP for a given H-core have to be different according to Eq. 2, and (II) the solutions from two CSPs for two different H-cores have to be different as well. The second condition (II) holds by definition, because $\overset{H}{\sim}$ is only defined on the H monomer placements that are constrained by different H-cores from $\mathcal{O}$ (differing in at least one position). In the following, we will discuss how to achieve the difference for solutions of a single CSP (I).

Note that the core-degeneracy, i.e. the number of different placements of H-monomers, or core-configurations, in optimal structures of a sequence, is *not equal* to the number of different H-cores, which are the sets of lattice points that are occupied by H-monomers. The latter number is easily obtained from the standard prediction algorithm, described in Sec. 2. It equals the number of cores, where the sequence is successfully threaded on.

Restricted Search for Enumeration of Representatives

The standard way to solve a CSP is a combination of domain filtering (i.e. constraint propagation) and depth first search. This results in a binary tree where each node represents a subproblem of the initial CSP (root) and edges represent the additional constraints added to derive the two subproblems from its predecessor node (CSP). The constraints c and $\neg c$ added to derive the leave nodes of a certain CSP are often of the form $c = (X_i \equiv d)$ by selecting a variable X_i from $\mathcal{X}$ and a value $d \in \mathcal{D}(\mathcal{X}_i)$ according to some heuristics. The constraint

solver traverse the binary tree until a solution was found or an inconsistency of a constraint from $\mathcal{C}$ was detected.

Therefore, a straightforward way to enumerate only one representative for each equivalence class can be sketched as follows: first, we restrict the search of the solving process onto the H associated variables. Then, we perform a single check for satisfiability, i.e. search for a single assignment of P monomer variables fulfilling all constraints in $\mathcal{C}$. Thus, we get only one P monomer placement for a given H monomer assignment if any exists.

The drawback of this approach is that we restrict the variable order of the search heuristics. But the performance of the CPSP approach mainly depends on the search heuristics applied to select a certain variable or value from its domain. It turned out that a mixed assignment of H and P associated variables yields the best runtimes. These heuristics can not be applied within the sketched procedure where we have to first assign H-associated variables, then P-associated ones. Thus, a lower CPSP performance is expected. But, we have to do less search which results in much faster runtimes than enumerating all optimal structures.

4 Representative Optimal Structures with Side Chains

Recently, we have introduced the extension of the CPSP approach [9] to *HP models including side chains* [6]. Here, each amino acid of a protein sequence is represented by two monomers: C_i^b representing the backbone atoms, and C_i^s representing the atoms of the side chain. Beneath the SAW condition on the backbone monomers C_i^b, we constrain each side chain to be neighbored to its backbone, i.e. $\forall_{1\leq i\leq n} : (C_i^s - C_i^b) \in N_L$. An example structure is given in Fig. 1b). The applied energy function E' exploits only HH-contacts of side chain monomers C_i^s:

$$E'(S, C^s) = \sum_{1\leq i<j\leq n} \begin{cases} -1 & : (C_i^s - C_j^s) \in N_L \wedge S_i = S_j = H \\ 0 & : \text{otherwise} \end{cases} \tag{3}$$

Therefore, the side chain models show an even higher degeneracy than the backbone-only models discussed so far, since all backbone monomers C_i^b are unconstrained by the energy function as well. Thus, an equivalence relation $\overset{H}{\approx}$ that focuses on the monomers constrained by the energy function is even more striking in HP models including side chains. The relation $\overset{H}{\approx}$ is given by

$$(C^b, C^s) \overset{H}{\approx} (\hat{C}^b, \hat{C}^s) \Leftrightarrow \forall_{i:S_i=H} : C_i^s = \hat{C}_i^s \tag{4}$$

Therefore, we will enforce that structures from one equivalence class show the same H monomer side chain positioning. The CPSP approach for HP models including side chains differs only in the CSP formulation from the original approach for backbone-only models [9]. This allows for the application of the same approach discussed in the previous section to enumerate non-equivalent optimal structure representatives. Thus, we restrict search to the H associated side chain variables first and only check for satisfiability on the remaining variables.

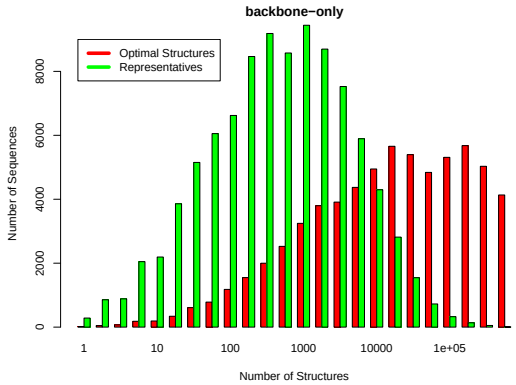


Fig. 2. Backbone-only models : Histogram of core-degeneracy (green) and degeneracy (red) with cut-off $\leq 10^6$. (Plots refer to 3D cubic lattice and sequence length 27.)

5 Results and Discussion

We exemplify the enumeration of representatives for backbone-only and side chain models. We focus on the comparison of the resulting core-degeneracy of a sequence and its overall number of optimal structures, i.e. degeneracy, because we are interested in a reduced set of optimal structures, e.g. for local search initialization (see introduction). All following results are given for HP-sequences of length 27 in 3D cubic lattice. Since the enumeration and check of all 2^{27} sequences ($> 10^8$) is computationally not feasible, we restrict each study to a large randomly chosen subset of 10^5 and 10^4 sequences, respectively.

The program HPREP implements the approach from section 3. It is integrated into the CPSP-tools package [10] version 2.4.0 and available online[1].

As discussed in Sec. 2, the structure threading step of the CPSP approach screens through a precomputed list of appropriate H-cores in decreasing number of contacts stored in a database. Therefore, it might occur that the available list from the database is exceeded without any solution, i.e. no optimal structure was computed. Still, the energy of the last H-core tried is a close lower bound on the energy this sequence can adopt. In the following, $\mathcal{B}$ denotes the subset of sequences where the current H-core database is not sufficient and thus the CPSP approach can give only a lower bound for now. The number of sequences in $\mathcal{B}$ is quite small. It is reasonable to assume that the degeneracy distribution among $\mathcal{B}$ is the same as for the remaining sequences or on average even higher.

Backbone-only models

We tested 10^5 random sequences in the backbone-only model in the 3D cubic lattice. Here, only 66% show a degeneracy below 10^6. $\mathcal{B}$ comprises about 4% of the sequences. The remaining 30% can adopt even more than 10^6 structures with minimal energy.

[1] `http://cpsp.informatik.uni-freiburg.de`

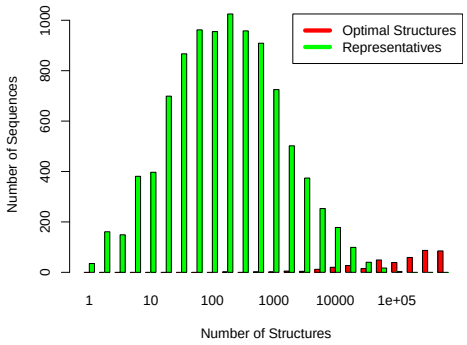


Fig. 3. Models with side chains : Histogram of the core-degeneracy *(green)* and degeneracy *(red)* with cut-off $\leq 10^6$. Note: only 408 sequences out of 10^4 showed a degeneracy below 10^6 as given in the text. (Plots refer to 3D cubic lattice and sequence length 27.)

Figure 2 summarizes the results: in *red* the degeneracy and in *green* the core-degeneracy distribution with cut-off 10^6 is presented. Thus, in *red* the degeneracy distribution comprises 66% of the sequences as given above. In contrast, *all* sequences show a number of optimal equivalence classes below 10^6 (in *green*)! The average degeneracy is reduced from 124800 (with cutoff 10^6) to a mean core-degeneracy of 4856. This reduction within two orders of magnitude results in reasonably small sets of representative structures e.g. to be utilized in local search initializations. Furthermore, the enumeration of representatives is on average six times faster than the enumeration of all optimal structures with a mean runtime of 2 seconds (Opteron 2356 - 2.3 GHz).

This increase of small sets of representatives compared to the complete sets of optimal structures shows the advantage of the approach: core-degeneracy does not show the huge combinatorial explosion of degeneracy. This gets even more striking in HP models including side chains, as shown in the next section.

Models including side chains

The degeneracy in HP models including side chains is much higher than for backbone-only models. This results from the simple energy function (Eq. 3) that does not constrain the backbone or P monomers. Therefore, an immense number of optimal structures is present. From the 10^4 HP-sequences tested only 408 show a degeneracy below 10^6. $\mathcal{B}$ comprises again about 3.1% of the sequences.

When investigating core-degeneracy the picture changes completely: *All* of the sequences tested have less than 10^6 representatives. Figure 3 summarizes the distribution. The average number of representatives is about 1550, which is again at least three orders of magnitude smaller than the average degeneracy. Since we have only a very rough lower bound of 10^6 on the average degeneracy (due to the cut-off), the real reduction ratio is expected to be even higher.

6 Conclusions

The introduced equivalence relations for HP models enables a energy function driven partitioning of structures. The presented CPSP approach extension enables an efficient calculation of representatives for all equivalence classes of optimal structures, i.e. calculation of a sequence's core-degeneracy. Using our implementation HPREP, we showed that sequences show several orders of magnitudes less optimal equivalence classes than optimal structures. This is most striking in models including side chains.

The sets of representatives are usually small. Furthermore, representatives show different hydrophobic core arrangements. Therefore, they are well placed to be used for the initialization of local search procedures that utilize more complex energy functions [14]. This emulates the hydrophobic collapse in the folding process.

Since a sequence's degeneracy is a measure of structural stability [12], we propose another application of our approach. The core-degeneracy might be used as a more reasonable *measure of stability* in the HP model compared to degeneracy. It ignores the HP model specific degeneracy blow-up due to unconstrained subchains of P monomers (see section 3). Thus, a structural stability analysis could be based on the presented equivalence classes instead of all possible structures.

References

1. Vishwas R. Agashe, M. C. R. Shastry, and Jayant B. Udgaonkar. Initial hydrophobic collapse in the folding of barstar. *Nature*, 377:754–757, 1995.
2. Rolf Backofen and Sebastian Will. Optimally compact finite sphere packings — hydrophobic cores in the FCC. In *Proc of CPM'01*, volume 2089 of *LNCS*, pages 257–272, 2001.
3. Rolf Backofen and Sebastian Will. A constraint-based approach to fast and exact structure prediction in three-dimensional protein models. *J Constraints*, 11(1):5–30, Jan 2006.
4. Bonnie Berger and Tom Leighton. Protein folding in the hydrophobic-hydrophilic (HP) model is NP-complete. *J Comp Biol*, 5:27–40, 1998.
5. K. A. Dill, S. Bromberg, K. Yue, K. M. Fiebig, D. P. Yee, P. D. Thomas, and H. S. Chan. Principles of protein folding – a perspective of simple exact models. *Protein Science*, 4:561–602, 1995.
6. S. Bromberg K. A. Dill. Side-chain entropy and packing in proteins. *Protein Sci*, 3(7):997–1009, 1994.
7. Kit F. Lau and Ken A. Dill. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromol.*, 22(10):3986–3997, 1989.
8. Martin Mann, Daniel Maticzka, Rhodri Saunders, and Rolf Backofen. Classifying protein-like sequences in arbitrary lattice protein models using LatPack. *HFSP Journal*, 2(6):396, 2008.
9. Martin Mann, Cameron Smith, Mohamad Rabbath, Marlien Edwards, Sebastian Will, and Rolf Backofen. CPSP-web-tool : a server for 3D lattice protein studies. *Bioinformatics*, 25(5):676–677, 2009.
10. Martin Mann, Sebastian Will, and Rolf Backofen. CPSP-tools - exact and complete algorithms for high-throughput 3D lattice protein studies. *BMC Bioinformatics*, 9:230, 2008.

11. S. Miyazawa and R. L. Jernigan. 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, 1996.
12. David Shortle, Hue S. Chan, and Ken A. Dill. Modeling the effects of mutations on the denatured states of proteins. *Prot Sci*, 1:201–215, 1992.
13. Kathleen Steinhöfel, Alexandros Skaliotis, and Andreas A. Albrecht. Stochastic protein folding simulation in the d-dimensional HP-model. In *Proc of BIRD'07*, pages 381–394. Springer, 2007.
14. Abu Dayem Ullah, Leonidas Kapsokalivas, Martin Mann, and Kathleen Steinhöfel. Protein folding simulation by two-stage optimization. In *Proc. of ISICA'09*, CCIS, Wuhan, China, Oct 2009. Springer. (accepted).
15. Sebastian Will and Martin Mann. Counting protein structures by DFS with dynamic decomposition. In *Proc of WCB'06*, pages 83–90, 2006.