Constraint-based Local Move Definitions for Lattice Protein Models Including Side Chains

Martin Mann[1], Mohamed Abou Hamra[2], Kathleen Steinhöfel[3], and Rolf Backofen[1]

[1] University of Freiburg, Bioinformatics, 79110 Freiburg, Germany,
{mmann,backofen}@informatik.uni-freiburg.de,
[2] German University in Cairo, 5th Settlement New Cairo City, Cairo, Egypt,
mohamed.abdel-fattah@student.guc.edu.eg,
[3] King's College London, Department of Computer Science, London WC2R 2LS, UK,
kathleen.steinhofel@kcl.ac.uk,

Abstract. The simulation of a protein's folding process is often done via stochastic local search, which requires a procedure to apply structural changes onto a given conformation. Here, we introduce a constraint-based approach to enumerate lattice protein structures according to k-local moves in arbitrary lattices. Our declarative description is much more flexible for extensions than standard operational formulations. It enables a generic calculation of k-local neighbors in backbone-only and side chain models. We exemplify the procedure using a simple hierarchical folding scheme.

1 Introduction

The *in silico* determination of a protein's functional fold is a well established problem in bioinformatics. Since X-ray or NMR studies are still time consuming and expensive, computational methods for *ab initio* protein structure prediction are needed. Despite research over the last decades, a direct calculation of minimal energy structures in full atom resolution is currently not feasible. Thus, heuristics and a wide variation of protein models have been developed to identify fundamental principles guiding the process of structure formation. A common abstraction of proteins are lattice protein models [3, 13, 14, 20]. Their discretized structure space enables efficient folding simulations [29, 31] while maintaining good modelling accuracy [25].

Folding simulations are often based on stochastic local searches, e.g. Monte Carlo simulations [29]. Different procedures, so called *move sets*, have been developed to calculate the structural changes along the simulation, i.e. to enumerate the structural neighborhood of a certain structure. A method often applied in literature are *k-local moves* [28] that allow for structural changes within a successive interval of fixed length k. They are discussed in detail in Sec 3. Dotu and co-workers have used local moves for backbone-only HP models within a constraint-based large neighborhood search for optimal protein structures [9]. Lesh *et al.* introduced *pull moves* [15] that are widely used in recent studies [19,

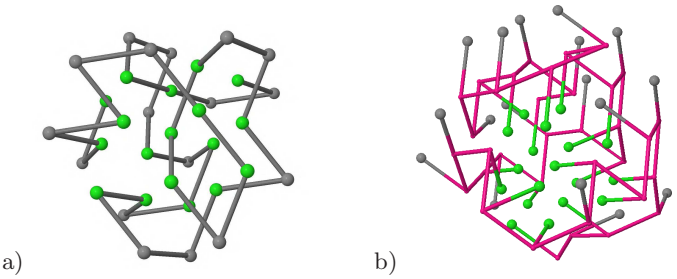
29]. *Pivot moves* allow for the rotation or reflection of subchains at an arbitrary Pivot position of the structure [17], while Zhang *et al.* suggested a sequential regrowth of structure fragments to enhance folding simulations [31].

All named move sets are currently restricted to backbone-only lattice protein models, i.e. only the C_α-trail of the protein is modeled. For more realistic protein models incorporating side chains, often a combination of different move sets is applied. Betancourt combined Pivot moves on the backbone with a new FEM move set [5], while Dima and Thirumalai have used a combination of 2-local moves on the backbone with a simple relocation of the side chain [8]. An exception is the advanced CABS model by Kolinski and co-workers [13], which represents the side chain in higher detail and requires more complex moves.

Here, we introduce a generic and flexible approach to enable folding simulations in backbone-only and side chain models using any k-local moves (i.e. any interval length k) in arbitrary lattices. The constraint programming (CP) based formulation focuses on a description of the targeted structural neighbors instead of an operational encoding of the moves possible. The introduced scheme is therefore easy to extend with new directives or can be used for other applications, e.g. fragment re-localization [31], as discussed later. Beneath applications in studies of the whole energy landscape [21], the approach is well placed to be applied within a local search following the framework of Pesant and Gendreau [26]. We apply our move set for side chain models within a simple folding simulation procedure in the style of [29] and evaluate the results with known protein structures.

2 Preliminaries

Given a lattice $L \subseteq \mathbb{Z}^3$ and an according neighborhood relation $\overset{L}{\sim}$ between coordinates of L. A *backbone-only* lattice protein of length n is described by (S, C) where $S \in \Sigma^n$ denotes the sequence over some alphabet Σ (e.g. the 20 proteinogen amino acids) and $C \in L^n$ the lattice nodes occupied. A valid lattice protein structure satisfies connectivity of successive monomers $\forall_{1 \leq i < n} : C_i \overset{L}{\sim} C_{i+1}$ and their self-avoidingness $\forall_{1 \leq i < j \leq n} : C_i \neq C_j$. A *side chain* lattice protein is defined by $(S, C^\mathrm{b}, C^\mathrm{s})$, i.e. a sequence $S \in \Sigma^n$, the backbone positions $C^\mathrm{b} \in L^n$ and the side chain positions $C^\mathrm{s} \in L^n$. The side chain position C^s represents the centroid of the amino acid's side chain atoms. A valid lattice protein structure including side chains satisfies connectivity of successive backbone monomers $\forall_{1 \leq i < n} : C^\mathrm{b}_i \overset{L}{\sim} C^\mathrm{b}_{i+1}$, the connection of backbone and side chain for each amino acid $\forall_{1 \leq i \leq n} : C^\mathrm{b}_i \overset{L}{\sim} C^\mathrm{s}_i$, and the selfavoidingness of all monomers $\forall_{1 \leq i < j \leq n} : C^\mathrm{b}_i \neq C^\mathrm{b}_j \wedge C^\mathrm{s}_i \neq C^\mathrm{s}_j \wedge C^\mathrm{b}_i \neq C^\mathrm{s}_j \wedge C^\mathrm{b}_i \neq C^\mathrm{s}_i$. We consider the contact based energy functions $E^\mathrm{b}(S, C) = \sum_{1 \leq i < j \leq n}^{(C_i \overset{L}{\sim} C_j)} e(S_i, S_j)$ for backbone-only and $E^\mathrm{s}(S, C^\mathrm{b}, C^\mathrm{s}) = \sum_{1 \leq i < j \leq n}^{(C^\mathrm{s}_i \overset{L}{\sim} C^\mathrm{s}_j)} e(S_i, S_j)$ for side chain lattice proteins for a given energy contribution function $e : \Sigma \times \Sigma \to \mathbb{R}$. Note, the energy function for side chain proteins considers (as in [7]) the contacts between side chain positions



a) b)

Fig. 1. HP-optimal structures of `HPPHHPPPPHPHHPHHPPPHPHPPPHHHPHHPPPHPHPH` 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. Calculation and visualization are done using the CPSP-package [22].

only! e^{20} denotes an empirical 20 amino acid contact potential as described in [4, 23]. e^{HP} represents the energy contribution function of the Hydrophobic-Polar (HP) model [14], i.e. it returns -1 if both amino acids are hydrophobic, 0 otherwise. Our hydrophobic/polar (H/P) assignment follows [29]. An *optimal structure* minimizes the energy function. In the following, we denote a structure *HP-optimal* if it minimizes the energy function based on e^{HP}. Figure 1 exemplifies HP-optimal structures for both lattice protein models. In the following, we assume a scaled lattice such that neighbored positions in the lattice have a distance of $3.8\mathring{A}$, the average C_α-atom distance in proteins.

3 Constraint-based Local Move Set Definition

To enable folding simulations we need a definition of structural changes that encodes the structural neighborhood of a given lattice protein structure. Here, we follow the idea of k-local moves, that confine the difference between the initial and the neighbored structure to a consecutive interval of maximal length of k. Therefore, we define the k-neighborhood $\mathcal{N}_k(C)$ of a given structure C as:

$$\mathcal{N}_k(C) = \left\{\text{valid structures } C' \mid \exists_{1\leq s\leq n} : \forall_{j\notin[s,\ldots,(s+k-1)]} : C_j = C'_j\right\} \quad (1)$$

In order to enumerate all valid structural neighbors $C' \in \mathcal{N}_k(C)$ of a given lattice protein C, we have to enumerate the neighbors for all possible interval lengths $1 \leq k' \leq k$ and interval starts $1 \leq s \leq (n - k' + 1)$. Since we want to calculate each neighboring structure only once, we have to enhance the k-local move definition to *strict k-local moves*. Here, we enforce in addition that both ends (C'_s and C'_{s+k-1}) of the successive interval of length k are changed, i.e. a strict k-local move does not cover a k'-local move with $k' < k$, as a normal k-local move in accordance with Eq. 1 does. This ensures a unique enumeration of structural neighbors for an increasing k'.

In the following, we will introduce the Constraint Satisfaction Problems (CSP) that describe all valid structural neighbors $C' \in \mathcal{N}_k(C)$ of a given lattice protein C according to strict k-local moves in a lattice L. A CSP is 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 $\mathcal{C}$. A solution of a CSP is an assignment $a_i \in \mathcal{D}(X_i)$ for each variable that satisfies all constraints in $\mathcal{C}$. To simplify the presentation, we utilize a binary neighboring constraint $\text{neigh}(X, Y)$ that ensures $\forall_{d_x \in \mathcal{D}(X)} : \exists_{d_y \in \mathcal{D}(Y)} : (d_x \overset{L}{\sim} d_y)$ and vice versa. Furthermore, we use the global all-different constraint by Régin [27] to enforce pairwise differences within a set of variables.

3.1 CSP for Backbone-only Models

Given a valid backbone-only lattice protein structure C of length n, a move interval length $k \leq n$, and the start of the interval $1 \leq s \leq (n - k + 1)$. We define k variables X_i, one for each position of the interval, with $\mathcal{D}(X_i) = L \setminus \{C_1, \ldots, C_{s-1}, C_{s+k}, \ldots, C_n\}$. These variables have to form a valid structure, therefore we post all-different$(X_1, \ldots, X_k)$ and $\forall_{1 \leq i < k} : \text{neigh}(X_i, X_{i+1})$. Since we describe a substructure, it has to be connected to the interval borders: if $s > 1 : \text{neigh}(X_1, C_{s-1})$ and if $(s + k - 1) < n : \text{neigh}(X_k, C_{s+k})$. Finally, we enforce that both ends of the interval are different from the old placement, i.e. $X_1 \neq C_i$ and $X_k \neq C_{i+k-1}$, to enumerate strict k-local move neighbors only.

The presented CSP is similar to the work of Dotu *et al.* [9], but in contrast ensures the uniqueness of each move. Thus, each neighbored structure is available only via a single interval. This is of high importance to enable a non-redundant enumeration of a structure's neighborhood in the fold space to access its energy landscape [21].

3.2 CSP for Models Including Side Chains

Given a valid side chain lattice protein structure $(C^{\text{b}}, C^{\text{s}})$ of length n, a move interval length $k \leq n$, and the start of the interval $1 \leq i \leq (n-k+1)$. We define k variables X_i^{b} and X_i^{s}, two for each position of the interval, with $\mathcal{D}(X_i^{\text{b}}) = \mathcal{D}(X_i^{\text{s}}) = L \setminus \{C_1^{\text{b}}, \ldots, C_{s-1}^{\text{b}}, C_{s+k}^{\text{b}}, \ldots, C_n^{\text{b}}, C_1^{\text{s}}, \ldots, C_{s-1}^{\text{s}}, C_{s+k}^{\text{s}}, \ldots, C_n^{\text{s}}\}$. To ensure a valid structure, we enforce all-different$(X_1^{\text{b}}, \ldots, X_k^{\text{b}}, X_1^{\text{s}}, \ldots, X_k^{\text{s}})$, $\forall_{1 \leq i < k} : \text{neigh}(X_i^{\text{b}}, X_{i+1}^{\text{b}})$, and $\forall_{1 \leq i \leq k} : \text{neigh}(X_i^{\text{b}}, X_i^{\text{s}})$. Since we describe a substructure, it has to be connected to the interval borders: if $s > 1 : \text{neigh}(X_1^{\text{b}}, C_{s-1}^{\text{b}})$ and if $(s + k - 1) < n : \text{neigh}(X_k^{\text{b}}, C_{s+k}^{\text{b}})$. Finally, we warrant the strictness of the k-local moves and enforce that both ends of the interval differ from the old backbone or side chain placement, i.e. $(X_1^{\text{b}} \neq C_i^{\text{b}} \vee X_1^{\text{s}} \neq C_i^{\text{s}})$ and $(X_k^{\text{b}} \neq C_{i+k-1}^{\text{b}} \vee X_k^{\text{s}} \neq C_{i+k-1}^{\text{s}})$.

4 Application

In the following, we applied the introduced move set to folding simulations of side chain lattice protein models in the 3D face-centered-cubic (FCC) lattice. In the FCC lattice, two lattice points l_1 and l_2 are neighbored, if and only if $(l_1 - l_2) \in \{\pm(1,1,0), \pm(1,0,1), \pm(0,1,1), \pm(1,-1,0), \pm(1,0,-1), \pm(0,1,-1)\}$. Thus, each point of the FCC lattice has 12 neighbored positions. k-local moves

are known to be non-ergodic for backbone-only models [16] depending on k, the used lattice, and the protein length. We expect the same for models including side chains, but using the FCC and an intermediate k should shift the problem to long chain lengths. Thus, we apply 3-local moves, i.e. with a maximal interval length $k = 3$ such that up to 6 monomers are moved (2 per amino acid). The implementation is based on Gecode [11]. To evaluate the structural difference between two structures $(C^{\mathrm{b}}, C^{\mathrm{s}})$ and $(\hat{C}^{\mathrm{b}}, \hat{C}^{\mathrm{s}})$ we calculate the distance and coordinate root mean square deviation (dRMSD and cRMSD) as given by Eq. 2 and 3, respectively. The needed superpositioning utilizes Kabsch's algorithm [12]. We apply the contact based energy function E^{s} that evaluates (only) side chain monomer contacts using the e^{20} contact energy potentials from Sec. 2 similar to the backbone-only studies in [4, 29]. In the following, we use C as an abbreviation for $(C^{\mathrm{b}}, C^{\mathrm{s}})$.

$$\mathrm{dRMSD} : \sqrt{\frac{\sum_{i<j}(|C_i^{\mathrm{b}} - C_j^{\mathrm{b}}| - |\hat{C}^{\mathrm{b}}{}_i - \hat{C}^{\mathrm{b}}{}_j|)^2 + (|C_i^{\mathrm{s}} - C_j^{\mathrm{s}}| - |\hat{C}^{\mathrm{s}}{}_i - \hat{C}^{\mathrm{s}}{}_j|)^2 + \sum_i (|C_i^{\mathrm{b}} - C_i^{\mathrm{s}}| - |\hat{C}^{\mathrm{b}}{}_i - \hat{C}^{\mathrm{s}}{}_i|)^2}{n^2}} \tag{2}$$

$$\mathrm{cRMSD} : \sqrt{\frac{\sum_i (|C_i^{\mathrm{b}} - \hat{C}^{\mathrm{b}}{}_i|)^2 + (|C_i^{\mathrm{s}} - \hat{C}^{\mathrm{s}}{}_i|)^2}{2 \cdot n}} \tag{3}$$

We derived a protein data from the Pisces web server [30] on June 23rd 2009. Only complete X-ray structures of $2.0\mathring{A}$ resolution or better with an R-value of 0.3 that contain side-chain data were considered. We used a 30% sequence identity cut-off. Since we are applying a simple contact-based energy function we filtered for short globular shaped proteins. Table 1 summarizes the used sequences and their corresponding Protein Data Bank (PDB) identifiers etc.

For each full atom PDB structure C_{PDB}, we derived a lattice protein structure C_{fit} that minimizes the dRMSD to C_{PDB}. This was done using LatFit from the LatPack-tools package v1.7.0[4] [19]. Table 1 summarizes the resulting dRMSD and cRMSD values.

Since the applied energy function is still a rough abstraction of the forces that guide the real folding process into C_{PDB}, no energy minimizing folding strategy will find the fitted lattice protein structure C_{fit}. Thus, we map C_{fit} to the according local minimum in the energy landscape. The mapping is done via a steepest decent or *gradient walk*. Starting from a given structure, at each step the neighbored structure with lowest energy is chosen for the next step until no such neighbor exists. Therefore, a gradient walk ends in a local minimum of the energy landscape, which we denote $g(C)$ for a given start structure C.

The $g(C_{fit})$ structures represent our "true" model to benchmark the following folding scheme. The energies of C_{fit} and $g(C_{fit})$ and their structural differences to each other and to C_{PDB} are given in Table 1.

The folding simulation procedure applied follows the idea of [29]. For each amino acid sequence S, we derive an according HP-sequence S_{HP} using the translation table used in [29]. The derived S_{HP} are given in Table 1. Following the observation of the hydrophobic collapse [1], we calculated HP-optimal

[4] Freely available at http://www.bioinf.uni-freiburg.de/Software/LatPack/

PDB ID - chain	Sequences (original and HP transform)
1BAZ-A	SKMPQVNLRWPREVLDLVRKVAEENGRSVNSEIYQRVMESFKKEGRIGA
	PPHPPHPHPHPPPHHPHHPPHPPPPPPPHPPPHHPPHHPPHPPPPPHPP
1J8E-A	GSHSCSSTQFKCNSGRCIPEHWTCDGDNDCGDYSDETHANCTNQ
	PPPPHPPPPHPHPPPPHHPPPHPHPPPPPHPPHPPPPPPPHPPP
1RH6-A	MYLTLQEWNARQRRPRSLETVRRWVRESRIFPPPVKDGREYLFHESAVKVDLNRP
	HHHPHPPHPPPPPPPPPHPPHPPHHPPPPHHPPPHPPPPPHHHPPPPHPHPHPPP
1Z0J-B	IEEELLLQQIDNIKAYIFDAKQCGRLDEVEVLTENLRELKHTLAKQKGGTD
	HPPPHHHPPHPPHPPHHHPPPPHPPHPPHPHHPPPHPPHPPPHPPPPPPPP
2DS5-A	GKLLYCSFCGKSQHEVRKLIAGPSVYICDECVDLCNDIIREEI
	PPHHHHPHHPPPPPPHPPHHPPPPHHHHPPHHPHHPPHPPPH
2EQ7-C	LAMPAAERLMQEKGVSPAEVQGTGLGGRILKEDVMRH
	HPHPPPPPHHPPPPHPPPPHPPPPHPPPHHPPPHHPP
2HBA-A	MKVIFLKDVKGMGKKGEIKNVADGYANNFLFKQGLAIEATPANLKALEAQKQ
	HPHHHHPPHPPHPPPPPHPPHPPPHPPPHHHPPPHPHPPPPPPHPPHPPPPP

PDB ID - chain	n	C_{fit} to C_{PDB} dRMSD	C_{fit} to C_{PDB} cRMSD	$E(C_{fit})$	$E(g(C_{fit}))$	$g(C_{fit})$ to C_{fit} dRMSD	$g(C_{fit})$ to C_{fit} cRMSD
1BAZ-A	49	0.886 Å	1.725 Å	-3.73	-31.51	4.050 Å	6.565 Å
1J8E-A	44	0.928 Å	1.939 Å	-3.54	-30.76	3.865 Å	6.857 Å
1RH6-A	55	0.921 Å	1.791 Å	1.33	-38.17	4.192 Å	8.243 Å
1Z0J-B	51	0.917 Å	2.095 Å	2.05	-35.95	3.185 Å	6.640 Å
2DS5-A	43	0.901 Å	1.750 Å	-4.35	-34.36	4.658 Å	7.755 Å
2EQ7-C	37	0.905 Å	1.813 Å	-3.07	-20.58	2.328 Å	4.751 Å
2HBA-A	52	0.890 Å	1.780 Å	-3.04	-30.62	3.224 Å	6.015 Å

Table 1. Used sequences, their HP transforms, length n, the quality of the fitted lattice protein model, and the according energies.

structure representatives utilizing the CPSP-approach [3, 22, 20] and its latest extension HPREP [18]. The resulting HP-optimal structures are named C_{HP}. For each C_{HP} we run gradient walks and evaluated the resulting local minima found. The corresponding energies are listed in Table 2. Furthermore, we performed a structural comparison of the resulting $g(C_{HP})$ structures to our "true" models $g(C_{fit})$ from the fitting. The RMSD values are given in Table 2.

In addition, we executed for each C_{HP} *random descending walks* in order to sample the local minima of the energy landscape accessible from the collapsed starting structures. Here, at each step a random neighbor with lower energy is selected following a uniform distribution until no such neighbor exists. The lowest reached local minimum of all random descending walks starting at C is denoted by $r(C)$. Energy and structural differences are given in Table 2.

5 Discussion

The gradient walks using 3-local moves starting from the fitted structures C_{fit} revealed that the currently applied contact based energy function using the energy potentials e^{20}, originally derived for backbone-only models [4], does not

PDB ID	average values	minimal values	
- chain	$\langle E(C_{HP})\rangle$	$\min E(g(C_{HP}))$	$\min E(r(C_{HP}))$
1BAZ-A	-10.67	-33.07	-34.60
1J8E-A	-12.45	-29.33	-32.35
1RH6-A	-13.09	-35.12	-37.59
1Z0J-B	-13.42	-34.71	-37.69
2DS5-A	-6.97	-31.00	-32.53
2EQ7-C	-6.55	-21.64	-25.10
2HBA-A	-11.07	-30.91	-35.56

PDB ID	$g(C_{HP})$ vs. $g(C_{fit})$		$r(C_{HP})$ vs. $g(C_{fit})$	
- chain	dRMSD	cRMSD	dRMSD	cRMSD
1BAZ-A	4.736 Å	8.797 Å	4.762 Å	9.360 Å
1J8E-A	3.384 Å	7.508 Å	3.196 Å	7.052 Å
1RH6-A	4.190 Å	9.645 Å	4.242 Å	10.156 Å
1Z0J-B	5.609 Å	10.166 Å	6.232 Å	11.438 Å
2DS5-A	3.588 Å	8.679 Å	3.425 Å	7.639 Å
2EQ7-C	3.427 Å	7.247 Å	4.177 Å	8.401 Å
2HBA-A	3.832 Å	8.848 Å	4.194 Å	9.075 Å

Table 2. Resulting energies and a structural comparison of the folding results.

reflect the real forces present for models including side chains. This can be observed when comparing the energies $E(C_{fit})$ to $E(g(C_{fit}))$ (see Table 1). An energy function that results in a smaller difference would be preferable, i.e. it would be a better model for the real forces guiding the folding process to C_{PDB}. In addition we could show, that the derived structures from our simple energy-optimizing folding simulation procedure are still quite dissimilar to the energy-optimized lattice fits of the real structures (see Table 2). We assume this mainly results from the simple energy function as well.

To improve the results, we plan to apply more advanced energy functions, e.g. following [13]. Most important: the energy function has to consider the backbone positioning as well, which is not done by the contact-based energy functions from Sec. 2. Additionally, we want to apply distance based energy potentials that allow for a more realistic energy evaluation. Another direction of ongoing research is to further constrain the allowed structures. Here, we will directly benefit from the CP-based formulation of k-local moves. Since we are formulating a CSP on valid structural neighbors, it is quite easy to post additional structural constraints. For instance, we can enforce a restriction on the allowed relative torsion angles along the protein chain (as e.g. done in [24]), that follows the observation of a limited degree of freedom in nature.

6 Conclusions and Summary

We introduced a CP-based approach to enumerate k-local neighbors of a lattice protein structure in backbone-only and side chain lattice protein models. The generic approach can be applied for any local move length k within arbitrary

lattices. Thus, it enables a fast prototyping of new folding simulation schemes or can be easily extended with additional constraints, e.g. restricted torsion angles. The CSP formulation enables the enumeration of the whole k-local move neighborhood $\mathcal{N}_k(C)$ of a given structure C or the calculation of a random neighboring structure $C_r \in \mathcal{N}_k(C)$ when applying a randomized search as possible in Gecode [11]. The application of symmetry breaking search [2] can be used to avoid the enumeration of symmetric structures, increasing the efficiency of folding simulations [10]. We plan the incorporation of the k-local move neighbor enumeration into our C++ energy landscape library (ELL) [21]. This will open an easy interface for folding simulations in arbitrary lattices utilizing any energy function of interest. Furthermore, this will enable full kinetics studies based on the energy landscape topology.

We will utilize the flexibility of the CP-based approach to incorporate additional structural constraints into the neighborhood generation. Following [24, 23], it is beneficial to restrict torsion angles along the backbone or to exploit secondary structure information.

Another advantage of the CP-based approach is its extensibility to constraint optimization problems (COP). Currently, we plan to incorporate the energy function as the objective into the CSP, as e.g. done in [6, 9]. Thus, by solving a COP while optimizing the energy function, we can directly calculate the lowest energy neighbor of a structure following the framework of Pesant and Gendreau [26], which is needed e.g. for a gradient walk in the energy landscape as done in Sec. 4. Furthermore, this would enable an extension of the work of Zhang *et al.* [31]. They showed (for backbone-only models) that the performance of Monte Carlo folding simulations can be significantly increased using a greedy sequential regrowth of subchains. Thus, we plan to directly apply the sketched COP to calculate the optimal fragments for lattice proteins including side chains. Finally, the presented CP-based move set formulation can be easily extended to any other local move definition of interest.

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. Excluding symmetries in constraint-based search. *Constraints*, 7(3):333–349, 2002.
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. Marco Berrera, Henriette Molinari, and Federico Fogolari. Amino acid empirical contact energy definitions for fold recognition in the space of contact maps. *BMC Bioinformatics*, 4:8, 2003.
5. M. R. Betancourt. Efficient monte carlo trial moves for polypeptide simulations. *J Chem Phys*, 123(17), 2005.
6. Raffele Cipriano, Alessandro Dal Palù, and Agostino Dovier. A hybrid approach mixing local search and constraint programming applied to the protein structure prediction problem. In *Proc of WCB'08*, page ?, 2008.

7. S. Bromberg K. A. Dill. Side-chain entropy and packing in proteins. *Protein Sci*, 3(7):997–1009, 1994.
8. R.I. Dima and D. Thirumalai. Exploring protein aggregation and self-propagation using lattice models: Phase diagram and kinetic. *Prot. Sci.*, 11(5):1036–1049, 2002.
9. Ivan Dotu, Manuel Cebrián, Pascal Van Hentenryck, and Peter Clote. Protein structure prediction with large neighborhood constraint programming search. In *Proc of CP'08*, volume 5202 of *LNCS*, pages 82–96. Springer, 2008.
10. Xiangchao Gan, Leonidas Kapsokalivas, Andreas A. Albrecht, and Kathleen Steinhöfel. A symmetry-free subspace for ab initio protein folding simulations. In *Proc. of BIRD'08*, volume 13 of *CCIS*, pages 128–139. Springer, 2008.
11. Gecode: Generic constraint development environment, 2007. Available as an open-source library from `www.gecode.org`.
12. W. Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica*, A32:922–923, 1976.
13. Sebastian Kmiecik and Andrzej Kolinski. Characterization of protein-folding pathways by reduced-space modeling. *PNAS*, 104(30):12330–12335, 2007.
14. Kit Lau and Ken Dill. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules*, 22(10):3986–3997, 1989.
15. Neal Lesh, Michael Mitzenmacher, and Sue Whitesides. A complete and effective move set for simplified protein folding. In *Proc of RECOMB'03*, pages 188–195. ACM, 2003.
16. Neal Madras and Alan D. Sokal. Nonergodicity of local, length-conserving Monte Carlo algorithms for the self-avoiding walk. *Journal of Statistical Physics*, 47(3-4):573–595, 1987.
17. Neal Madras and Alan D. Sokal. The pivot algorithm: A highly efficient Monte Carlo method for the self-avoiding walk. *J Stat Phys*, 50(1-2):109–186, 1988.
18. Martin Mann, Rolf Backofen, and Sebastian Will. Equivalence classes of optimal structures in HP protein models including side chains. In *Proc of WCB'09*, 2009.
19. 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.
20. 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.
21. Martin Mann, Sebastian Will, and Rolf Backofen. The energy landscape library - a platform for generic algorithms. In *Proc of BIRD'07*, volume 217, pages 83–86. OCG, 2007.
22. 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.
23. Alessandro Dal Palu, Agostino Dovier, and Federico Fogolari. Constraint Logic Programming approach to protein structure prediction. *BMC Bioinformatics*, 5(1):186, 2004.
24. Alessandro Dal Palù, Sebastian Will, Rolf Backofen, and Agostino Dovier. Constraint based protein structure prediction exploiting secondary structure information. In *Proc of CILC'04*, pages 16–17, 2004.
25. Britt H. Park and Michael Levitt. The complexity and accuracy of discrete state models of protein structure. *J Mol Biol*, 249:493–507, 1995.
26. Gilles Pesant and Michel Gendreau. A constraint programming framework for local search methods. *Journal of Heuristics*, 5(3):255–279, 1999.

27. Jean-Charles Régin. A filtering algorithm for constraints of difference in CSPs. In *Proc. of 12th National Conference on AI*, pages 362–367, 1994.
28. A. Sali, E. Shakhnovich, and M. Karplus. Kinetics of protein folding. A lattice model study of the requirements for folding to the native state. *J Mol Biol*, 235(5):1614–1636, 1994.
29. 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).
30. G. Wang and Roland L. Dunbrack. Pisces: a protein sequence culling server. *Bioinformatics*, 19(12):1589–91, 2003.
31. J. Zhang, S. C. Kou, and J. S. Liu. Biopolymer structure simulation and optimization via fragment regrowth Monte Carlo. *J Chem Phys*, 126(22):225101, 2007.