

# Producing high-accuracy lattice models from protein atomic co-ordinates including side chains

Martin Mann<sup>1,2,\*</sup>, Rhodri Saunders<sup>3,\*</sup>, Cameron Smith<sup>1</sup>,  
Rolf Backofen<sup>1</sup> and Charlotte M. Deane<sup>3</sup>

<sup>1</sup> Bioinformatics, University of Freiburg, Georges-Köhler-Allee 106, 79110 Freiburg, Germany

<sup>2</sup> Theoretical Biochemistry, University of Vienna, Währingerstrasse 17, 1090 Vienna, Austria

<sup>3</sup> Department of Statistics, Oxford University, 1 South Parks Road, Oxford, OX1 3TG, UK

\* These authors contributed equally to this work.

**Keywords:** Protein structure, lattice protein, model fitting, RMSD

**Correspondence:** Charlotte Deane, Oxford University, Department of Statistics, 1 South Parks Road, Oxford, OX1 3TG, UK

Email: deane@stats.ox.ac.uk, Fax: +44 1865 272595, Phone: +44 1865 281301

## Abstract

Lattice models are a common abstraction used in the study of protein structure, folding, and refinement. They are advantageous because the discretisation of space can make extensive protein evaluations computationally feasible. Various approaches to the protein chain lattice fitting problem have been suggested but only a single backbone-only tool is available currently.

We introduce **LatFit**, a new tool to produce high-accuracy lattice protein models. It generates both backbone-only and backbone-side-chain models in any user defined lattice. **LatFit** implements a new distance RMSD-optimisation fitting procedure in addition to the known coordinate RMSD method. The program is freely available for academic download and as a web-server: <http://csp.informatik.uni-freiburg.de/LatFit/>

We tested **LatFit**'s accuracy and speed using a large non-redundant set of high resolution proteins (SCOP database) on three commonly used lattices: 3D cubic, face-centred cubic, and knight's walk. Fitting speed compared favourably to other methods, and both backbone-only and backbone-side-chain models show low deviation from the original data ( $\sim 1.5\text{\AA}$  RMSD in the FCC lattice). To our knowledge this represents the first comprehensive study of lattice quality for on-lattice protein models including side chains while **LatFit** is the only available tool for such models.

# 1 Introduction

It is not always computationally feasible to undertake protein structure studies using full atom representations. The challenge is to reduce complexity while maintaining detail [1–3]. Lattice protein models are often used to achieve this but in general only the protein backbone or the amino acid centre of mass is represented [4–12]. A huge variety of lattices and energy functions have previously been developed and applied [4, 13, 14].

In order to evaluate the applicability of different lattices and to enable the transformation of real protein structures into lattice models, a representative lattice protein structure has to be calculated. Mañuch and Gaur have shown the NP-completeness of this problem for backbone-only models in the 3D-cubic lattice and named it the *protein chain lattice fitting (PCLF) problem* [15].

The PCLF problem has been widely studied for backbone-only models [13, 16–24]. The most important aspects in producing lattice protein models with a low root mean squared deviation (RMSD) are the lattice co-ordination number and the neighbourhood vector angles [18, 23]. Lattices with intermediate co-ordination numbers, such as the face-centred cubic (FCC) lattice, can produce high resolution backbone models [18] and have been used in many protein structure studies (e.g. [3, 25, 26]). However, the use of backbone models is limited since they do not account for the space required for side chain packing.

To overcome this restriction lattice protein models that include side chains have been introduced [27–33]. Reva et al. [32] have, to our knowledge, developed the only previous approach to solve the PCLF problem including side chains. They apply dynamic programming to find an optimal solution according to their error function. Unfortunately, the approach is shown to often yield no solution in the 3D cubic lattice. The CABS-tools by Kolinski and co-workers utilize a hybrid on-lattice (backbone) and off-lattice (side chain) protein representation to study folding dynamics but do not attempt to answer the PCLF problem [31, 34].

In this manuscript we use the side chain model definition of Bromberg and Dill [28], where each amino acid is represented by two on-lattice monomers: one represents the side chain and one the  $C_\alpha$  atom. This explicit representation of side chains prevents unnatural collapse during structural studies [35] and enables the reconstruction of full atom protein data [36]. Full on-lattice protein models are constrained in their possible side chain placement but enable exhaustive studies of folding kinetics and structure space [11, 37, 38] not applicable within off-lattice side chain models like the CABS approach.

To the best of our knowledge, there is only one other publicly available implemented approach, namely `LocalMove`, to derive lattice protein models from real proteins despite a large number of published methods. `LocalMove` is a web interface introduced by Ponty et al. [22] for backbone-only models in 3D-cubic and FCC lattice and applies a Monte-Carlo search in order to find lattice protein models.

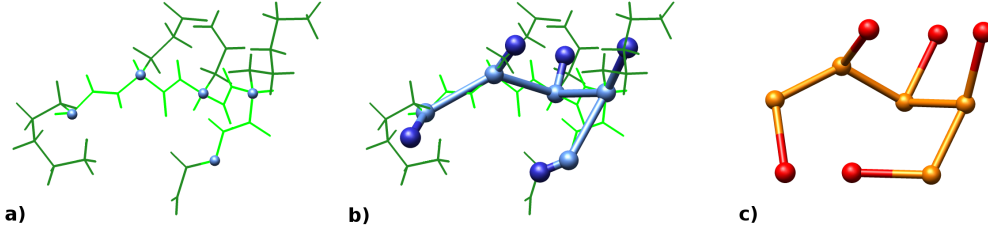


Figure 1: The diagram depicts the fitting process of **LatFit** for side chain models. a) Original full atom data is given. The five  $C_\alpha$  atoms of the segment are highlighted as balls while the backbone and side chain parts are given in light and dark green, respectively. b) The coordinates for each amino acid to fit are extracted, i.e. for side chain models the  $C_\alpha$  position (light blue) and the centroid of the side chain (dark blue). c) These positions are fitted to derive an according lattice protein model in the underlying lattice (here 3D knight’s walk lattice).

We present our tool **LatFit** to tackle this lack of available implementations. **LatFit** solves the PCLF problem, i.e. transforms a protein from full atom coordinate data to a lattice model, and is available as both a stand-alone tool for high-throughput pipelines and a web interface for *ad hoc* usage. A new fitting procedure that optimises distance RMSD enables rotation independent lattice model creation of protein structures. The method is applicable to arbitrary lattices and handles both backbone and side chain representations with equivalent accuracy. A depiction of the workflow is given in Fig. 1.

Utilising **LatFit** we present the first comprehensive study of lattice quality for protein models including side chains. In our test, **LatFit** fitted the majority of models on an FCC lattice within  $1.5\text{\AA}$  RMSD.

## 2 Material and Methods

In order to enable a precise formulation of the method we introduce some preliminary definitions. A lattice  $L$  is a set of 3D coordinates  $x$  defined by a set of neighboring vectors  $v \in N$ . The neighboring vectors are of equal length ( $\forall v, v' \in N : |v| = |v'|$ ), each with a reverse within the neighborhood ( $\forall v \in N : -v \in N$ ), such that each coordinate in  $L$  can be expressed by a linear combination of the neighboring vectors, i.e.  $L = \{ x \mid x = \sum_{v \in N} d \cdot v \wedge d \in \mathbb{Z}_0^+ \}$ .  $|N|$  gives the coordinate number of the lattice, e.g. 6 for 3D-cubic or 12 for the FCC lattice.

A lattice protein structure with side chains of length  $l$  is defined by a sequence of lattice nodes  $M^b = (M_1^b, \dots, M_l^b) \in L^l$  representing the backbone monomers of the protein (one for each amino acid) and the according sequence  $M^s = (M_1^s, \dots, M_l^s) \in$

$L^l$  for the side chain positions. A valid structure ensures backbone connectivity ( $\forall_{i < l} : M_i^b - M_{i+1}^b \in N$ ), side chain connectivity ( $\forall_i : M_i^b - M_i^s \in N$ ), as well as self-avoidance ( $\forall_{i \neq j} : M_i^b \neq M_j^b \wedge M_i^s \neq M_j^s$  and  $\forall_{i,j} : M_i^b \neq M_j^s$ ). The two sets together define the lattice protein structure  $M = (M^b, M^s)$ .

## Fitting Procedure

Given a protein structure of length  $l$  in Protein Database (PDB) format [39], `LatFit` builds up the lattice protein sequentially, one amino acid at a time, starting from the amino-terminus.

First, all neighboring vectors  $v \in N$  of the used lattice  $L$  are scaled to a length of  $3.8\text{\AA}$ , which is the mean distance between consecutive  $C_\alpha$  atoms and close to the mean distance between a  $C_\alpha$  atom and the associated side chain centroid. The latter distance was found to be on average  $\approx 3.6\text{\AA}$  within available PDB structures (data not shown). While this ignores the shorter CIS-PRO  $C_\alpha$  linkage and the non-existence of a side chain for Glycine, this scaling enables a reasonable mapping of proteins into the lattice, where each amino acid will be represented by two monomers and all covalent bonds are scaled to  $|v| = 3.8\text{\AA}$ . Therefore, all resulting measures will be directly interpretable in  $\text{\AA}$  units.

The positions for each amino acid  $i$  to be fitted, i.e. the  $C_\alpha$  position of the backbone  $P_i^b$ , and the centroid  $P_i^s$  (geometric center) of all non-hydrogen atom coordinates of the side chain, are extracted from the PDB file. They form the data to fit  $P = (P^b, P^s)$ .

The lattice model is derived by one of the following procedures optimising either a distance or coordinate RMSD. Both methods are introduced for lattice proteins including side chains but can be used to derive backbone-only lattice models as well. A sketch of the fitting workflow is given in Fig. 1.

## dRMSD Optimisation

The fitting follows a greedy iterative chain-growth procedure. The initial lattice model's backbone and side chain position ( $M_1^b$  and  $M_1^s$ ) are placed arbitrarily but adjacent ( $M_1^b - M_1^s \in N$ ). For each iteration  $1 < i \leq l$ , all valid placements of the next  $M_i^b$  and  $M_i^s$  on the lattice are calculated. A distance RMSD (dRMSD, Eqn. 1) evaluation is used to identify the best  $n_{keep}$  structures of length  $i$  for the next extension iteration. Since dRMSD is a rotation/reflection independent measure, symmetric structures must be filtered.

To calculate the final fit of the initial protein  $P$ , a superpositioning of the dRMSD-optimised structure  $M$  and a reflected version  $M'$  is done using the method by Kabsch [40]. The superpositioning translates and rotates  $M/M'$  in order to achieve the best mapping onto  $P$ . The superpositioning with lowest co-ordinate RMSD (cRMSD, Eqn. 2) is selected and finally returned.

$$\text{dRMSD} = \sqrt{\frac{\sum_{i < j} (|P_i - P_j| - |M_i - M_j|)^2}{l \cdot ((2 \cdot l) - 1)}} \quad (1)$$

with  $P = P^s \cup P^b$ , and  $M = M^s \cup M^b$ .

$$\text{cRMSD} = \sqrt{\frac{\sum_{i=1}^l (|P_i^b - M_i^b|)^2 + (|P_i^s - M_i^s|)^2}{2 \cdot l}} \quad (2)$$

### cRMSD Optimisation

A cRMSD evaluation according to Eq. 2 depends on the superpositioning of the protein and its model. Thus the best relative lattice orientation has to be identified in addition to the best model. Once the orientation is fixed, a cRMSD evaluation allows for a fast, additive RMSD update along the chain extension.

We implement a cRMSD optimising method following [6, 18] as an alternative fitting strategy. In general a user defined number of rotation intervals  $r$  are performed for each of the XYZ rotation axes. For each rotation, we transform  $P^b$  and  $P^s$  into  $\hat{P}^b$  and  $\hat{P}^s$ , respectively, to obtain the rotated current target structure.

The fitting procedure follows a chain-growth approach:  $P_1^b$  is placed onto an arbitrary lattice node  $M_1^b$ . The according side chain monomer  $M_1^b$  is placed to the adjacent node closest to the position  $P_1^s$  to be represented. Now, all valid placements of the next  $M_i^b$  and  $M_i^s$  on the lattice are calculated. Using the co-ordinate RMSD (cRMSD, Eqn. 2) we evaluate all derived models and keep the best  $n_{keep}$  for the next extension following [18] until all amino acids have been placed.

By applying the above cRMSD based fitting procedure we obtain the best fit for the current rotation. An iterative application of this procedure then results in the overall best fit for all screened rotations. Since our screen of XYZ rotations was discretised, the current rotation might be refineable. Therefore, another rotational refinement can be applied that investigates  $r^{ref}$  small rotation intervals around the best rotation from the first screen [6].

The run time of the cRMSD-method scales with respect to the lattice co-ordination number,  $n_{keep}$ , and most importantly the number of rotation intervals  $r$  and  $r^{ref}$  considered.

### Futher Features

Coordinate data in the PDB is often incomplete. For example flexible loop structures are hard to resolve by current methods [41]. This results in missing coordinate data for certain substructures within PDB files. **LatFit** enables a structural fitting of even such fragmented PDB structures and produces a lattice protein fragment for each fragment of the original protein.

Currently, **LatFit** supports the 2D-square, 3D-cubic (CUB, 100), 3D face centered cubic (FCC, 110) and 3D knights walk (210) lattice. The modular software

design of our open source program enables an easy and straight forward implementation of other lattices via a specification of the according neighboring vectors  $N$ .

The implementation is open source and freely available for academic use at

<http://www.bioinf.uni-freiburg.de/Software/LatPack/>

## Webserver

The web interface of **LatFit**, integrated into the **CPSP-web-tools** [42], enables *ad hoc* usage of the tool. Either a protein structure in PDB format can be uploaded or a valid identifier from the PDB database given. In the latter case, the full atom data is automatically retrieved from the database.

Our default parameters enable a direct application of **LatFit** resulting in a balanced tradeoff between runtime and fitting quality. The computations are done remotely on a computation cluster while the user can trace the processing status via the provided job identifier and according link. Results are available and stored for 30 days after production.

Supported output formats of **LatFit** are the PDB format, the Chemical Markup Language (CML) format, as well as a simple XYZ coordinate output. The output files are available for download. In addition, a highly compact string representation of the lattice protein is also given in absolute move strings that encode the series of neighboring vectors  $v \in N$  along the structure.

The generated absolute move string can be directly used to apply other lattice protein tools onto the resulting structures, e.g. from the **CPSP**-package for HP-type lattice protein models [10, 42] or from the **LatPack** tools for arbitrary lattice models [11, 38].

Results are visualised using **Jmol** [43] for an interactive presentation of the final protein structure. The final dRMSD and cRMSD values of the lattice protein compared to the original protein are given as well as the absolute move string encoding of the resulting structure. For an example of the **LatFit** web interface see Fig. 2.

Further details regarding the methods implemented, the output formats supported and the applicable parameterisation are located in the **LatFit** manual distributed with the source code. We provide an extensive help page and a frequently asked questions (FAQ) section within the web interface. Note, the web server is based on JavaServer Pages (JSP) technology and requires a connection via the JSP standard port 8080. A web interface for *ad hoc* usage is available at

<http://csp.informatik.uni-freiburg.de/LatFit/>  
<http://csp.informatik.uni-freiburg.de:8080/>

## 3 Results and Discussion

In the following, we evaluate the average fitting quality of our new **LatFit** tool to results known from literature [6, 13, 18]. Furthermore, we investigate the perfor-

## LatPack Tools - LatFit Result

**Menu**

[Home](#)

**HPstruct**  
structure pred.

**HPconvert**  
PDB, CML, ...

**HPview**  
3D visualization

**HPdeg**  
degeneracy

**HPnnet**  
neural network

**HPdesign**  
seq. design

**LatFit**  
PDB to lattice

**Results**  
direct access

[Help](#)

[FAQ](#)

Job ID: 6969628

Job 6969628 SUBMITTED @ 13:14:45 UTC+1 on 2010-03-18  
 Job 6969628 STARTED @ 13:15:02 UTC+1 on 2010-03-18  
 Job 6969628 COMPLETED @ 13:15:05 UTC+1 on 2010-03-18  
<http://cpsp.informatik.uni-freiburg.de:8080/trunk/LatFitResult.jsp?jobid=6969628> (30 day expiry) [Download Results](#)

**Input Parameters**

PDBFile	<a href="#">InputFile_6969628.pdb</a>
Atom	CoM
Chain Identifier	A
Model Number	1
Lattice Protein Type	Include Side-chains
Lattice Form	FCC
CA-CA bond length	3.8
Optimization Mode	dRMSD
Max. to keep per iteration	100
Output Format	PDB
Output points to fit	yes

**Output**

LatFit has produced a PDB file available for [download](#). Click [here](#) to view the results.

The following distance measures were obtained:

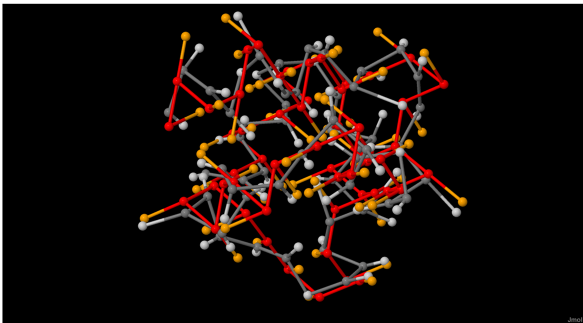
**cRMSD: 1.84493 Angstroms**  
**dRMSD: 1.41052 Angstroms**

The following absolute move strings have been produced and are available to view:

(BR)BD(BL)FL(LU)FR(BR)BD(BD)LD(BL)FL(RU)FR(BR)BD(BR)LD(FU)LU(BR)BU(FU)LU(LD)BD(RD)RU(RU)B

If the molecule does not display, click [here](#) or check the [FAQ](#).

Show LatFit Chain  
 Show Original Chain



■ LatFit Backbone   
 ■ LatFit Sidechain   
 ■ Original Backbone   
 ■ Original Sidechain

This result was obtained by using the CPSP-tools package with the following command and arguments:  
`latFit -pdbFile=7scratch/cpsp/bisge000/CPSP_results/result/2010-03-18_13~14~45_latFit_6969628.pdb*  
 -pdbAtom=CoM -pdbChain=A -pdbModel=1 -fitSideChain -lat=FCC -bondLength=3.8 -opt=D*  
 -nKeep=100 -outMode=PDB -outOrigData -v`

[Legal Disclosure and Contact](#)

Figure 2: A screenshot of the LatFit web interface result visualisation.



a) *Backbone-only models:*

	Park and Levitt [18]		Reva et al. [14][22]	Ponty et al. [22]	LatFit	
	cRMSD	dRMSD	cRMSD*	cRMSD*	cRMSD	dRMSD
CUB	2.84	2.34	2.84 (0.748 · 3.8)	3.46 (0.911 · 3.8)	2.97	<b>2.08</b>
FCC	1.78	1.46	-	-	1.89	<b>1.34</b>
210	1.24	1.02	-	-	1.29	<b>0.92</b>

b) *Side chain models:*

	LatFit	
	cRMSD	dRMSD
CUB	4.16	2.78
FCC	2.10	1.50
210	1.60	1.13

Table 1: Table (a) compares the RMSD mean values for *backbone-only* models for approaches from literature to the results from our LatFit dRMSD-optimisation method on three different lattices. Table (b) gives according results for side chain including models. \* Some reported values had to be rescaled to Å.

mance of the new dRMSD-based fitting procedure implemented in LatFit. To this end, we compare its results to the cRMSD-optimizing approach that follows [6, 18], both implemented within LatFit.

We use LatFit to derive protein models on the commonly used 3D cubic, FCC, and knights walk lattices [18] using the dRMSD-based approach, parameterised with  $n_{keep} = 1000$ . Our test set was taken from the PISCES webserver [44]. We enforced 40% sequence identity cut-off, chain length 50-300, R-factor  $\leq 0.3$  and resolution  $\leq 1.5\text{\AA}$  to derive a high-quality set of proteins to model. Given our requirement for side chains,  $C_\alpha$ -only chains were ignored. The resulting benchmark set contains 1198 proteins exhibiting a mean length of 160 ( $\sigma = 64$ ).

In accordance with previous studies [18], cRMSD and dRMSD are used to assess model quality. cRMSD measures the similarity in according co-ordinate position of two structures whereas dRMSD measures the similarity of intramolecular distances. Due to the scaling of our lattice, RMSD results are in Å rather than the scaled values provided by Ponty et al. [22].

Our backbone model RMSD values presented in Table 1 are competitive or superior to known fitting results known from the literature [6, 13, 18]. Both the new dRMSD- as well as the reimplemented cRMSD-optimisation method reproduce the high quality previously achieved by other methods using the FCC and 210 lattices. The slightly higher mean cRMSD values for the dRMSD method are due to the non-optimisation of that measure. Note, LatFit outperforms the results reported for LocalMove by Ponty et al. [22]. We found the LocalMove webserver currently not working for the proteins tested. Therefore, only results reported in [22] for the 3D cubic lattice and no FCC results are available.

**LatFit** is designed for side chain models and results here are strong (see Table 1b). In general, side chain models produce slightly larger RMSD values than the equivalent backbone-only model. This is due to the fact that the variation in distance between consecutive  $C_\alpha$  atoms (fitted in both models) is lower than that between  $C_\alpha$  atoms and their side chain centroid (fitted only in side chain models). In lattice models every distance is fixed at  $3.8\text{\AA}$  which results in a higher mean displacement of the side chain. Nevertheless, high accuracy fits are still attained. Results in our test set have mean dRMSDs of about  $1.2\text{\AA}$  and  $1.5\text{\AA}$  in the 210 and FCC lattice, resp., for both optimisation strategies. When comparing the dRMSD-optimisation with the cRMSD-optimising version, we observe very similar results. This is in accordance to our observations from the backbone-only models.

The strength of **LatFit** is its ability to produce both side chain and backbone-only lattice protein models. High accuracy models can be produced on the FCC lattice within seconds to minutes depending on the parameterisation. Fits on the 210 lattice take orders of magnitude longer for relatively little gain in model accuracy. For this reason we recommend using the FCC lattice for detailed high-throughput protein structure studies in both backbone-only and side chain representing lattice models.

## 4 Concluding remarks

**LatFit** enables the automated high resolution fitting of both backbone and side chain lattice protein models from full atomic data in PDB format. We demonstrate its high accuracy on three widely used lattices using a large, non-redundant protein data set of high resolution. Side chain fits show on average a higher deviation than backbone models, but both produce high quality fits with results generally less than  $1.5\text{\AA}$  on the face-centred cubic lattice. To our knowledge, this is the first study and publicly available implementation for side chain models in this field. Available via web interface and as a stand-alone tool, **LatFit** addresses the lack of available programs and is well placed to enable further, more detailed investigation of protein structure in a reduced complexity environment. Even now the **LatFit** webserver is in daily use worldwide (monitored via Google Analytics<sup>1</sup>), which shows the need for efficient implementations such as **LatFit**.

## Acknowledgement

The authors have declared no conflict of interest.

---

<sup>1</sup><http://www.google.com/analytics/>

## 5 References

- [1] Leonid Mirny and Eugene Shakhnovich. Protein folding theory: From lattice to all-atom models. *Annual Review of Biophysics and Biomolecular Structure*, 30(1):361–396, 2001.
- [2] Ken A. Dill, S. Banu Ozkan, M. Scott Shell, and Thomas R. Weikl. The protein folding problem. *Annual Review of Biophysics*, 37(1):289–316, 2008.
- [3] Sorin Istrail and Fumei Lam. Combinatorial algorithms for protein folding in lattice models: A survey of mathematical results. *Commun. Inf. Syst.*, 9(4):303–346, 2009.
- [4] K. A. Dill. Theory for the folding and stability of globular proteins. *Biochemistry*, 24(6):1501–9, 1985.
- [5] A. Renner and E. Bornberg-Bauer. Exploring the fitness landscapes of lattice proteins. In *Pac Symp Biocomput.*, pages 361–372, 1997.
- [6] J. Miao, J. Kleinseetharaman, and H. Meirovitch. The optimal fraction of hydrophobic residues required to ensure protein collapse. *J Mol. Bio.*, 344(3):797–811, 2004.
- [7] Rolf Backofen and Sebastian Will. A constraint-based approach to fast and exact structure prediction in three-dimensional protein models. *Constraints*, 11(1):5–30, 2006.
- [8] Fabien P. E. Huard, Charlotte M. Deane, and Graham R. Wood. Modelling sequential protein folding under kinetic control. *Bioinformatics*, 22(14):e203–210, 2006.
- [9] C. M. Deane, M. Dong, F. P. Huard, B. K. Lance, and G. R. Wood. Cotranslational protein folding—fact or fiction? *Bioinformatics*, 23(13):i142–148, 2007.
- [10] Martin Mann, Sebastian Will, and Rolf Backofen. CPSP-tools - exact and complete algorithms for high-throughput 3D lattice protein studies. *BMC Bioinf*, 9:230, 2008.
- [11] Martin Mann, Daniel Maticzka, Rhodri Saunders, and Rolf Backofen. Classifying protein-like sequences in arbitrary lattice protein models using LatPack. *HFSP J*, 2:396, 2008.
- [12] Rhodri Saunders, Martin Mann, and Charlotte Deane. Signatures of co-translational folding. *Biotechnology Journal, Special issue: Protein folding in vivo*, 6(6):742–751, 2011.
- [13] A. Godzik, A. Kolinski, and J. Skolnick. Lattice representations of globular proteins: How good are they? *J Comp. Chem.*, 14(10):1194–1202, 1993.
- [14] Boris A. Reva, Michel F. Sanner, Arthur J. Olson, and Alexei V. Finkelstein. Lattice modeling: Accuracy of energy calculations. *Journal of Computational Chemistry*, 17(8):1025 – 1032, 1996.
- [15] J. Mañuch and D. R. Gaur. Fitting protein chains to cubic lattice is NP-complete. *Journal of bioinformatics and computational biology*, 6(1):93–106, February 2008.
- [16] D. G. Covell and R. L. Jernigan. Conformations of folded proteins in restricted spaces. *Biochemistry*, 29(13):3287–3294, April 1990.
- [17] D. A. Hinds and M. Levitt. A lattice model for protein structure prediction at low resolution. *Proc Natl Acad Sci USA*, 89(7):2536–2540, 1992.
- [18] B.H. Park and M. Levitt. The complexity and accuracy of discrete state models of protein structure. *J Mol Biol*, 249:493–507, 1995.
- [19] D. S. Rykunov, B. A. Reva, and A. V. Finkelstein. Accurate general method for lattice approximation of three-dimensional structure of a chain molecule. *Proteins*, 22(2):100–109, 1995.
- [20] Boris A. Reva, Dmitrii S. Rykunov, Alexei V. Finkelstein, and Jeffrey Skolnick. Optimization of protein structure on lattices using a self-consistent field approach. *Journal of Computational Biology*, 5(3):531–538, 1998.
- [21] P. Koehl and M. Delarue. Building protein lattice models using self-consistent mean field theory. *J. Chem. Phys.*, 108:9540–9549, June 1998.

- [22] Y. Ponty, R. Istrate, E. Porcelli, and P. Clote. LocalMove: computing on-lattice fits for biopolymers. *Nucleic Acids Res*, 36(2):W216–W222, 2008.
- [23] C.L. Pierri, A. De Grassi, and A. Turi. Lattices for ab initio protein structure prediction. *Proteins*, 73(2): 351–361, 2008.
- [24] M. Mann and A. Dal Palu. Lattice model refinement of protein structures. In *Proc of WCB’10*, page 7, 2010. arXiv:1005.1853.
- [25] E. Jacob and R. Unger. A tale of two tails: Why are terminal residues of proteins exposed? *Bioinformatics*, 23(2):e225–30, 2007.
- [26] Abu Dayem Ullah, Leonidas Kapsokalivas, Martin Mann, and Kathleen Steinhöfel. Protein folding simulation by two-stage optimization. In *Proc. of ISICA’09*, volume 51 of *CCIS*, pages 138–145, 2009.
- [27] S. Sun. Reduced representation model of protein structure prediction: statistical potential and genetic algorithms. *Proteins*, 2(5):762–785, 1993.
- [28] S. Bromberg and K. A. Dill. Side-chain entropy and packing in proteins. *Protein Sci*, 3(7):997–1009, 1994.
- [29] W. E. Hart and S. Istrail. 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, 1997.
- [30] Volker Heun. Approximate protein folding in the HP side chain model on extended cubic lattices. *Discrete Appl. Math.*, 127(1):163–177, 2003. ISSN 0166-218X.
- [31] Andrzej Kolinski and Jeffrey Skolnick. Reduced models of proteins and their applications. *Polymer*, 45(2): 511 – 524, 2004.
- [32] B.A. Reva, D.S. Rykunov, A.J. Olson, and A.V. Finkelstein. Constructing lattice models of protein chains with side groups. *Journal of Computational Biology*, 2(4):527–535, 1995.
- [33] Y. Zhang, A. K. Arakaki, and J. Skolnick. TASSER: an automated method for the prediction of protein tertiary structures in CASP6. *Proteins*, 61 Suppl 7:91–8, 2005.
- [34] Andrzej Kolinski. Protein modeling and structure prediction with a reduced representation. *Acta biochimica Polonica*, 51(2):349–372, 2004.
- [35] Volker A. Eyrich, Daron M. Standley, and Richard A. Friesner. Prediction of protein tertiary structure to low resolution: performance for a large and structurally diverse test set. *J Mol Biol*, 288(4):725–742, May 1999.
- [36] M. Feig, P. Rotkiewicz, A. Kolinski, J. Skolnick, and C. L. Brooks. Accurate reconstruction of all-atom protein representations from side-chain-based low-resolution models. *Proteins*, 41(1):86–97, 2000.
- [37] M. Wolfinger, S. Will, I. Hofacker, R. Backofen, and P. Stadler. Exploring the lower part of discrete polymer model energy landscapes. *Europhysics Letters*, 74(4):725–732, 2006.
- [38] Martin Mann, Mohamed Abou Hamra, Kathleen Steinhöfel, and Rolf Backofen. Constraint-based local move definitions for lattice protein models including side chains. In *Proceedings of the Fifth Workshop on Constraint Based Methods for Bioinformatics (WCB09)*, 2009. arXiv:0910.3880.
- [39] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The Protein Data Bank. *Nucl. Acids Res.*, 28(1):235–242, 2000.
- [40] W. Kabsch. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallographica*, A34:827–828, 1978.
- [41] Yoonjoo Choi and Charlotte M. Deane. FREAD revisited: Accurate loop structure prediction using a database search algorithm. *Proteins*, 78(6):1431–1440, 2010.
- [42] Martin Mann, Cameron Smith, Mohamad Rabbath, Marlien Edwards, Sebastian Will, and Rolf Backofen. CPSP-web-tools: a server for 3D lattice protein studies. *Bioinformatics*, 25(5):676–7, 2009.
- [43] Angel Herráez. Biomolecules in the computer: Jmol to the rescue. *Biochem. Educ*, 34(4):255–261, 2006.
- [44] G. Wang and Roland L. Dunbrack. PISCES: recent improvements to a PDB sequence culling server. *Nucleic Acids Res*, 33(Web Server issue):W94–8, 2005.