Lattice model refinement of protein structures

Martin Mann
Bioinformatics, University of Freiburg, Germany,
mmann@informatik.uni-freiburg.de

Alessandro Dal Palù
Dip. di Matematica, Università di Parma, Italy,
alessandro.dalpalu@unipr.it

Abstract

To find the best lattice model representation of a given full atom protein structure is a hard computational problem. Several greedy methods have been suggested where results are usually biased and leave room for improvement.

In this paper we formulate and implement a Constraint Programming method to refine such lattice structure models. We show that the approach is able to provide better quality solutions. The prototype is implemented in COLA and is based on limited discrepancy search. Finally, some promising extensions based on local search are discussed.

1 Introduction

Extensive structural protein studies are computationally not feasible using full atom protein representations. The challenge is to reduce complexity while maintaining detail [6, 11]. Lattice protein models are often used to achieve this but in general only the protein backbone or the amino acid center of mass is represented [1, 16, 18, 20, 26]. A huge variety of lattices and energy functions have previously been developed [5, 8, 28], while the lattices 2D-square, 3D-cubic and 3D face centered cubic (FCC) are most prominent.

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. In detail, given a full atom protein structure one has to find the best structure representation within the lattice model that minimizes the applied distance measure. Maňuch and Gaur have shown the NP-completeness of this problem for backbone-only models in the 3D-cubic lattice when minimizing coordinate root mean square deviation (cRMSD) and named it the *protein chain lattice fitting (PCLF) problem* [19].

The PCLF problem has been widely studied for backbone-only models. Suggested approaches utilize quite different methods, ranging from full enumeration [4], greedy chain growth strategies [17, 20, 23], dynamic programming [10], simulated annealing [25], or the optimization of specialized force fields [13, 27]. 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 neighborhood vector angles [23, 24]. Lattices with intermediate co-ordination numbers, such as the face-centered cubic (FCC) lattice, can produce high resolution backbone models [23] and have been used in many protein structure studies (e.g. [11, 12, 29]).

Most of the PCFL methods introduced are heuristics to derive good solutions in reasonable time. Greedy methods as chain growth algorithms [17, 20, 23] enable low runtimes but the fitting quality depends on the chain growth direction and parameterization. Thus, resulting lattice models are biased by the method applied and have potential for refinement.

This paper has the goal to provide some evidence that greedy methods can be effectively improved by subsequent refinement steps that increase the fitting quality. We present a formalization and a simple working prototype. Moreover we briefly discuss some potential methodologies that we expect could be effectively employed.

2 Definitions and Preliminaries

In order to define the Constraint Programming approach we first introduce some preliminary formalisms.

Given a protein in full atom representation of length n (e.g. in Protein Data Base (PDB) format [2]), we denote the sequence of 3D-coordinates of its C_α-atoms (its *backbone trace*) by $P = (P_1, \ldots, P_n)$.

A regular *lattice* L is defined by a set of neighboring vectors $\vec{v} \in N_L$ of equal length ($\forall_{\vec{v}_i,\vec{v}_j \in N_L} : |\vec{v}_i| = |\vec{v}_j|$), each with a reverse ($\forall_{\vec{v} \in N_L} : -\vec{v} \in N_L$, such that $L = \{\vec{x} \mid \vec{x} = \sum_{\vec{v}_i \in N_L} d_i \cdot \vec{v}_i \wedge d_i \in \mathbb{Z}_0^+\}$. $|N_L|$ gives the coordinate number of the lattice L, e.g. 6 for 3D-cubic or 12 for the FCC lattice. All neighboring vectors $\vec{v} \in N_L$ of the used lattice L are scaled to a length of $3.8\AA$, which is the mean distance between consecutive C_α-atoms in real protein structures.

A backbone-only *lattice protein structure* M of length n is defined by a sequence of lattice nodes $M = (M_1, \ldots, M_n) \in L^n$ representing the backbone (C_α) monomers of each amino acid. A valid structure ensures backbone connectivity ($\forall_{i<n} : M_i - M_{i+1} \in N_L$) as well as selfavoidance ($\forall_{i \neq j} : M_i \neq M_j$), i.e. it represents a selfavoiding walk (SAW) in the underlying lattice.

The *PCFL problem* is to find a lattice protein model M of a given protein's backbone P, such that a distance measure between M and P (dist(M,P)) is minimized [19].

In this contribution, we tackle the *PCFL refinement problem*. Here, a protein backbone P as well as a first lattice model M is given, e.g. derived by a greedy chain growth procedure [17, 20, 23]. The problem is to find a lattice model M', such that dist$(M',P) <$ dist(M,P), via a relaxation/refinement of the original model M.

In the following, we utilize distance RMSD (dRMSD, Eq. 1) as the distance measure dist(M,P). dRMSD is independent of the relative orientation of M and P since it captures the model's deviation from the pairwise distances of C_α-atoms in the original protein. Minimizing this measure optimizes the lattice model obtained.

$$\text{dRMSD}(M,P) \quad = \quad \sqrt{\frac{\sum_{i<j} \left(|M_j - M_i| - |P_j - P_i|\right)^2}{n(n-1)/2}} \tag{1}$$

3 Refinement of Lattice Models: a Constraint Model in COLA

In this section we formalize a Constraint Optimization Problem (COP) to solve the PCFL refinement problem (see Sec. 2), i.e. to refine a lattice model M of a protein P. The input is the original protein P and its lattice model M to be refined. The output is a lattice model M' derived from M via some relaxation that optimizes our distance measure dRMSD(M',P) (Eq. 1).

We first formalize the problem and show how to implement it in COLA, a COnstraint solver for LAttices [21]. This is followed by an altered formulation that utilizes limited discrepancy search [9].

3.1 The Constraint Optimization Problem

The COP can be formalized as follows:

$X_1 \dots X_n$ — variables representing $M' = (M'_1, \dots, M'_n)$

$D(X_i)$ — variable domains $= \{v \mid v \in L \wedge |v - M_i| \leq f_{\text{scale}} \cdot d_{\max}\}$,
i.e. an M_i surrounding sphere with radius $f_{\text{scale}} \cdot d_{\max}$

$SAW(X_1 \dots X_n)$ — self-avoiding walk constraint, e.g. split into a chain of binary `contiguous` and a global `alldifferent` constraint

O — objective function variable, implements dRMSD
$= \sum_{i<j}(|X_j - X_i| - |P_j - P_i|)^2$ to be minimized

Note that $d_{\max}$ refers to the number of lattice units used and thus it is scaled to the correct distance of $f_{\text{scale}} = 3.8$Å. Thus, the domains for $d_{\max} = 0$ only contain the original lattice point M_i (domain size 1), while $d_{\max} = 1$ results in M_i as well as all neighbored lattice points (domain size $1 + 12 = 13$ in FCC). The domain size guided by $d_{\max}$ defines the allowed relaxation of the original lattice model M to be refined. For more details about global constraints for protein structures on lattices, the reader can refer to [1, 22].

The COLA implementation takes advantage of the availability of 3D lattice point domains and distance constraints. The implementation changes the original framework only in the input data handling and objective function definition. A working copy of COLA and the COP implemented for this paper are available at `http://www2.unipr.it/∼dalpalu/COLA/`

3.2 Limited Discrepancy Search

A simple enumeration with $d_{\max} = 1$ and a protein of length 50, already shows that the search space of the COP from the previous section is not manageable. In this example, each point can be placed in 13 different positions in the FCC lattice, and even if the contiguous constraint among the amino acids is enforced, the number of different paths is still beyond the current computational limits.

We tried a simple branch and bound search an $X_1, \dots, X_n$, where the dRMSD bound is estimated by considering the possible placement of non labeled variables and the best dRMSD contribution provided by each amino acid. In detail, each amino acid s not yet labeled is compared to each other amino acid (s'). Each pair provides a range of different contributions to dRMSD measure, depending on the placement of s and the placement of the other amino acids (when not yet labeled). A closed formula computation (rather than a full enumeration of all combinations), based on bounding box of domain positions, is activated, in order to estimate the minimal contribution. Clearly, this estimation is not particularly suited, since we relax the estimation on $\mathbb{R}^3$, where the null (best) contribution can be easily found as soon as the bounds on $|X_s - X_{s'}|$ include the value $|P_s - P_{s'}|$. Unfortunately, the discrete version requires a more expensive evaluation that boils down to full pair checks. Therefore, the current bound is very loose and the pruning effects are modest.

A general impression is that the dRMSD measure presents a pathological distribution of local minima, depending on the placement of amino acids on the lattice. In general, due to the discrete nature of the lattice, the modification of a single amino acid's position can drastically vary its contributions to the measure.

Protein ID	8RXN	1CKA	2FCW
length	52	57	106

Table 1: Used proteins from the Protein Data Base (PDB) [2].

These considerations suggested us to focus on the identification of solutions that improve the dRMSD w.r.t. M rather than searching for the optimal one. In terms of approximated search we tried to capture the main characteristics of the COP and design efficient and effective heuristics.

A simple idea we tested is the *limited discrepancy* search [9]. This search compares the amino acid placements in the lattice models M and M'. Every time a corresponding amino acid is placed differently in the two conformations, we say that there is a *discrepancy*. We set a global constraint that limits the number of deviations to at most K. This allows to generate conformations that are rather similar to M, especially if $d_{\max}$ is greater than 1. The rational behind this heuristics is that we expect that potential conformations M' improve the dRMSD only when contained in a close neighborhood of the M structure.

The count of the number of discrepancies K is implemented directly in COLA at each labeling step.

3.3 Results

We summarize here the preliminary results coming from the COLA implementation of a K discrepancy search in 3D FCC lattice.

The initial lattice models to be refined were generated using the LatFit tool from the LatPack package [16, 17]. LatFit implements an efficient greedy dRMSD optimizing chain growth method and was parameterized to consider the best 100 structures from each elongation for further growth[1].

We test three proteins (Table 1) and for each of them we input the conformation M obtained from the greedy algorithm (LatFit). Table 2 reports the best dRMSD of our new model M' found depending on $d_{\max}$ and the number K of amino acids placed differently from the input conformation. Furthermore, time consumption for each parameterization is given.

Note that if either $K = 0$ or $d_{\max} = 0$ only the input structure resulting from the greedy LatFit run can be enumerated.

These results, yet preliminary, offer an interesting insight about the distribution of suboptimal solutions. It is interesting to note, e.g., that better solutions are found by allowing a rather large local neighborhood for a few amino acids ($d_{\max}$ parameter). On the other side, it seems that few modifications (K) are sufficient to alter the input sequence and obtain a better conformation.

In Figure 1 we exemplify the gain of model precision for the protein `8RNX`. Only the relaxation of $K = 4$ monomers enables the structural change that leads to a dRMSD drop from 1.2469 down to 1.0884, an improvement of about 13%. A movement of less monomers would not enable such a drastic change. This depicts the potential of a local search scheme that iteratively applies a series of such structural changes.

Investigating the time consumption (Table 2) one can see that the runtime increases drastically with K which governs the search tree size. The domain sizes implied by $d_{\max}$ do not show such an immense influence.

The behavior encountered is an indicator that a search based on exploring only the neighborhood should provide efficient and good suboptimal solutions. In the next section we briefly discuss some promising approaches that we plan to investigate.

[1] For details on the LatFit method see [17] and the freely available web interface at `http://cpsp.informatik.uni-freiburg.de`

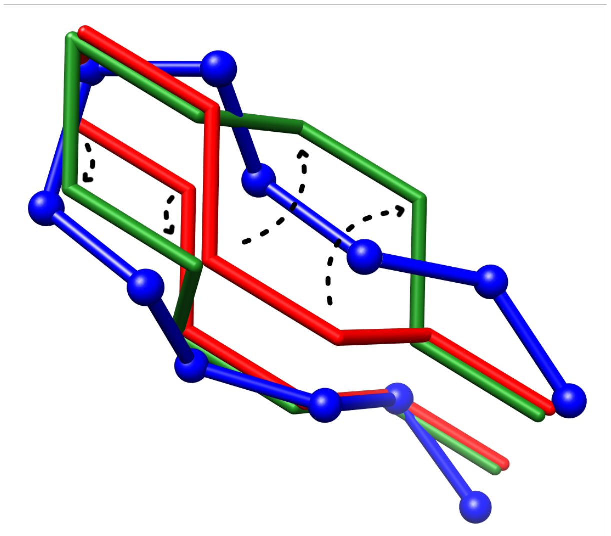

Figure 1: The initial lattice model M (red) of the protein chain P (blue, balls) and the final/refined lattice model M' (green) resulting from $d_{\max} = 2$ and $K = 4$ for protein 8RNX. Note, only the altered loop regions (residue 2-14) are shown, but the whole structure models M and M' were superpositioned to P independently.

dRMSD

		K			
8RXN		1	2	3	4
$d_{\max}$	0	1.2469	1.2469	1.2469	1.2469
	1	1.2319	1.2172	1.1639	1.1189
	2	1.2319	1.1674	1.1596	1.0884
	3	1.2319	1.1674	1.1596	1.0884
		K			
1CKA		1	2	3	4
$d_{\max}$	0	1.2370	1.2370	1.2370	1.2370
	1	1.2226	1.2226	1.2226	1.2226
	2	1.2026	1.1887	1.1887	1.1887
	3	1.2026	1.1887	1.1887	1.1887
		K			
2FCW		1	2	3	4
$d_{\max}$	0	1.1353	1.1353	1.1353	1.1353
	1	1.1353	1.1324	1.1317	1.1309
	2	1.1321	1.1300	1.1254	1.1200
	3	1.1321	1.1300	1.1254	1.1200

time in seconds

		K			
8RXN		1	2	3	4
$d_{\max}$	0	0.048	0.081	0.040	0.039
	1	0.112	0.790	2.365	20.70
	2	0.068	0.983	6.500	106.6
	3	0.106	0.499	7.399	124.0
		K			
1CKA		1	2	3	4
$d_{\max}$	0	0.031	0.030	0.027	0.037
	1	0.402	0.615	3.442	39.27
	2	0.225	0.456	7.595	120.6
	3	0.421	0.616	8.573	140.2
		K			
2FCW		1	2	3	4
$d_{\max}$	0	0.043	0.050	0.058	0.078
	1	0.118	1.997	49.99	1128
	2	0.294	7.192	341.8	14235
	3	0.332	8.129	394.5	16140

Table 2: $d_{\max}$ and K influence on discrepancy search measured in dRMSD and time.

3.4 Future work

In our opinion, a framework that integrates CP and Local Search is particularly suited to generate fast suboptimal solutions, potentially very close to the optimal one. We identify some possible directions that we believe are excellent candidates to model and solve approximately the PCLF problem:

- **local neighboring search [3, 7]**: this technique allows to integrate Gecode and Local Search frameworks. The framework handles constraint specifications and local moves within C++ programming language;
- ***k*-local moves [25]:** the idea here is to apply structural changes on *k* consecutive amino acids and repeat the process in a Monte-Carlo and/or simulated annealing style.
- **side chain model [15]**: our model can be extended to include side chains and we could exploit a similar set of local moves.
- **the framework presented in [30]**: COLA is here extended and combined directly to a Local Search approach based on *pull moves* [14].

4 Conclusion

In this paper we presented a Constraint Programming based model for the refinement of lattice fitting of protein conformations. A simple branching was shown to be ineffective and a limited discrepancy search was modeled and shown to be beneficial to the identification of suboptimal solutions. A prototypical implementation in the framework COLA and some preliminary results have shown the feasibility of the method. We believe that an extension of the framework to Local Search is particularly suited for the PCLF problem at hand.

Acknowledgments This work is partially supported by PRIN08 *Innovative multi-disciplinary approaches for constraint and preference reasoning* and GNCS-INdAM *Tecniche innovative per la programmazione con vincoli in applicazioni strategiche*.

References

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

[2] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissigand I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucl. Acids Res.*, 28(1):235–242, 2000.

[3] R. Cipriano, L. Gaspero, and A. Dovier. A hybrid solver for large neighborhood search: Mixing Gecode and EasyLocal++. In *Proc. of HM'09*, pages 141–155, Berlin, Heidelberg, 2009. Springer-Verlag.

[4] D. G. Covell and R. L. Jernigan. Conformations of folded proteins in restricted spaces. *Biochemistry*, 29(13):3287–3294, April 1990.

[5] K. A. Dill. Theory for the folding and stability of globular proteins. *Biochemistry*, 24(6):1501–9, 1985.

[6] K. A. Dill, S. B. Ozkan, M.S. Shell, and T. R. Weikl. The protein folding problem. *Annual Review of Biophysics*, 37(1):289–316, 2008.

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

[8] A. Godzik, A. Kolinski, and J. Skolnick. Lattice representations of globular proteins: How good are they? *J Comp. Chem.*, 14(10):1194–1202, 1993.

[9] W. D. Harvey and M. L. Ginsberg. Limited discrepancy search. In *Proceedings of IJCAI'95*, pages 607–613, San Francisco, CA, USA, 1995.

[10] 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.

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

[12] E. Jacob and R. Unger. A tale of two tails: Why are terminal residues of proteins exposed? *Bioinformatics*, 23(2):e225–30, 2007.

[13] P. Koehl and M. Delarue. Building protein lattice models using self-consistent mean field theory. *J. Chem. Phys.*, 108:9540–9549, June 1998.

[14] N. Lesh, M. Mitzenmacher, and S. Whitesides. A complete and effective move set for simplified protein folding. In *Proc. of RECOMB'03*, pages 188–195, 2003.

[15] M. Mann, M. A. Hamra, K. Steinhöfel, and R. Backofen. Constraint-based local move definitions for lattice protein models including side chains. In *Proc. of WCB'09*, 2009. arXiv:0910.3880.

[16] M. Mann, D. Maticzka, R. Saunders, and R. Backofen. Classifying protein-like sequences in arbitrary lattice protein models using LatPack. *HFSP J*, 2:396, 2008.

[17] M. Mann, R. Saunders, C. Smith, R. Backofen, and C. M. Deane. LatFit - producing high accuracy lattice models from protein atomic co-ordinates including side chains. (under review), 2010.

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

[19] J. Maňuch 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.

[20] 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.

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

[22] A. Dal Palù, A. Dovier, and E. Pontelli. 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, 2010.

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

[24] C. L. Pierri, A. Grassi, and A. Turi. Lattices for ab initio protein structure prediction. *Proteins*, 73(2):351–361, 2008.

[25] Y. Ponty, R. Istrate, E. Porcelli, and P. Clote. LocalMove: computing on-lattice fits for biopolymers. *Nucleic Acids Res*, 36(2):W216–W222, 2008.

[26] A. Renner and E. Bornberg-Bauer. Exploring the fitness landscapes of lattice proteins. In *Pac Symp Biocomput.*, pages 361–372, 1997.

[27] B. A. Reva, D. S. Rykunov, A. V. Finkelstein, and J. Skolnick. Optimization of protein structure on lattices using a self-consistent field approach. *Journal of Computational Biology*, 5(3):531–538, 1998.

[28] B. A. Reva, M. F. Sanner, A. J. Olson, and A. V. Finkelstein. Lattice modeling: Accuracy of energy calculations. *J Comp Chem*, 17(8):1025 – 1032, 1996.

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

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