High accuracy on-lattice side chain models
of PDB protein structures

Martin Mann[1], **Rhodri Saunders**[2], **Cameron Smith**[1], **Rolf Backofen**[1], and **Charlotte Deane**[2]

[1] University of Freiburg, Germany · [2] University of Oxford, England

mmann@informatik.uni-freiburg.de

http://cpsp.informatik.uni-freiburg.de

University of Freiburg

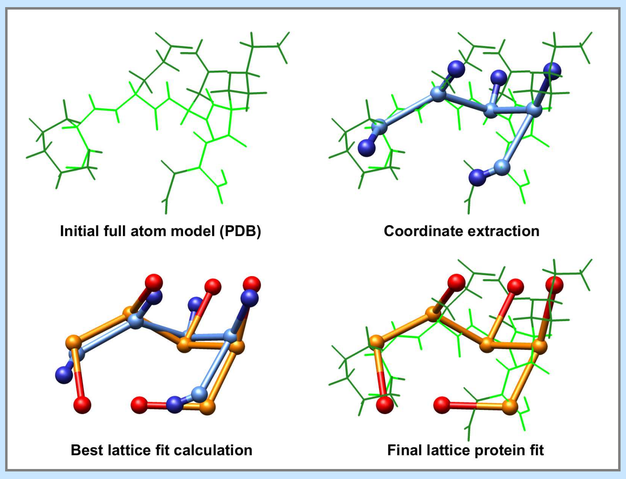
Bioinformatics

Introduction

The *protein chain lattice fitting (PCLF) problem* is to calculate a lattice protein model for a protein given in full atom representation, a problem shown to be NP-complete [1]. 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 [2].

The PCLF problem has been widely studied for backbone-only lattice protein models [2, 3, 4, 5]. The studies reveal that lattices with intermediate co-ordination numbers, such as the face-centred cubic (FCC) lattice, can produce high resolution backbone models [2]. However, the use of backbone models is limited since they do not account for the space required for side chain packing. Reva *et al.* have to our knowledge developed the only approach to solve the PCLF problem for lattice proteins including side chains [6].

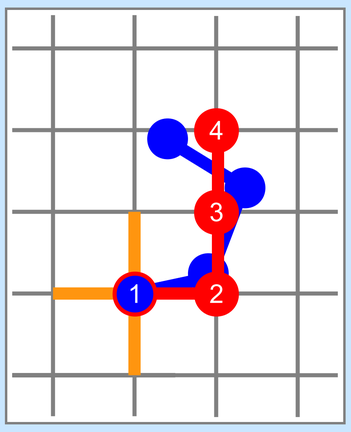
We present our tool `LatFit` that tackles the PCLF problem. It is available as both a stand-alone tool for high-throughput pipelines and a web interface for *ad hoc* usage. A new fitting procedure that optimises distance RMSD enables rotation independent lattice model creation of protein structures. The method is applicable to arbitrary lattices and handles both backbone and side chain representations with equivalent accuracy.



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$\mathring{A}$ RMSD.

Method

`LatFit` uses an RMSD-optimising chain-growth algorithm to build up the lattice protein model. To gain a reasonable modelling, all neighboring vectors $\vec{n} \in N$ of the used lattice L are scaled to a length of 3.8$\mathring{A}$, which is the mean distance between consecutive C_α atoms and close to the mean distance between a C_α atom and the side chain centroid ($\approx 3.6\mathring{A}$). Given a protein of length l in Protein Database (PDB) format [7], the positions for each amino acid i to be fitted, i.e. the C_α position of the backbone P_i^b, and the centroid P_i^s (geometric center) of all non-hydrogen atom co-ordinates of the side chain, are extracted from the PDB file.



The *distance (d)RMSD-optimisation* follows a greedy iterative chain-growth procedure related to [2] but optimising the rotation-independent dRMSD (Eq. 1). 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. We keep the best n_{keep} structures of length i for the next extension iteration according to dRMSD evaluation. (Depicted for backbone-only models and cRMSD optimisation on the left.)

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 [8]. The superpositioning with lowest co-ordinate RMSD (cRMSD, Eq. 1) is selected and finally returned.

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

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

The *coordinate (c)RMSD-optimisation* implements the method by Park and Levitt [2] and depends on the superpositioning of the protein and its model (see figure above). 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. But optimising lattice rotation slows down the method significantly.

Results

We use `LatFit` to derive protein models on the commonly used 3D cubic, FCC, and knights walk lattices [2]. Our test set was taken from the `PISCES` webserver [9] (40% sequence identity cut-off, chain length 50-300, R-factor ≤ 0.3 and resolution $\leq 1.5\mathring{A}$, no C_α-only chains). The resulting benchmark set contains 1198 proteins exhibiting a mean length of 160 ($\sigma = 64$).

In accordance with previous studies [2], cRMSD and dRMSD are used to assess model quality. cRMSD measures the similarity in co-ordinate position whereas dRMSD measures the similarity of interatomic distances. RMSD results are in $\mathring{A}$.

Each protein was fitted twice onto the lattice using either our dRMSD or cRMSD-optimising method. dRMSD-optimisation was parameterised with $n_{keep} = 1000$. For cRMSD-optimising runs, we used the parameters $r = 10$ and $r^{ref} = 5$ for backbone-only fits, and $r = 5$ and $r^{ref} = 3$ for side chain fits. A rotation range of $[0, \frac{\pi}{2}]$ and $n_{keep} = 5$ was used for initial rotations. Rotational refinement was applied onto the interval $\pm[0, \frac{\pi}{10}]$ around the best initial rotation to derive the final fit.

In the table below we compare the RMSD mean values (μ) and standard deviations (σ) from literature to the results from our `LatFit` cRMSD-optimisation methods for *backbone-only models* on three different lattices.

	Results taken from Park and Levitt [2]		Results taken from Ponty *et al.* [3]	`LatFit` cRMSD optimisation	
	dRMSD	cRMSD	cRMSD	dRMSD	cRMSD
	μ	μ	μ (rescaled to $\mathring{A}$)	μ / σ	μ / σ
cub	2.34	2.84	3.5 (0.923 · 3.8)	2.042 / 0.228	2.539 / 0.234
fcc	1.46	1.78	-	1.319 / 0.086	1.641 / 0.090
210	1.02	1.24	-	0.931 / 0.060	1.154 / 0.060

The following table gives the RMSD mean values (μ) and standard deviations (σ) of the results from our dRMSD- and the cRMSD-optimisation methods for *side chain models* on three different lattices.

	`LatFit` - dRMSD optimisation		`LatFit` - cRMSD optimisation	
	dRMSD	cRMSD	dRMSD	cRMSD
	μ / σ	μ / σ	μ / σ	μ / σ
cub	2.779 / 0.754	4.157 / 1.331	2.609 / 0.481	3.286 / 0.624
fcc	1.496 / 0.153	2.104 / 0.246	1.495 / 0.061	1.839 / 0.068
210	1.126 / 0.068	1.601 / 0.100	1.185 / 0.042	1.450 / 0.047

Conclusions

`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$\mathring{A}$ on the face-centred cubic lattice. To our knowledge, this is the first 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. The free web interface for *ad hoc* usage is accessible at

http://cpsp.informatik.uni-freiburg.de

References

[1] 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.

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

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

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

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

[6] 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.

[7] 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.

[8] W. Kabsch. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallographica*, A34:827–828, 1978.

[9] 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.