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

Martin Mann, Rhodri Saunders, Cameron Smith, Rolf Backofen, and Charlotte Deane

1 University of Freiburg, Germany - 2 University of Oxford, England

mmann@informatik.uni-freiburg.de

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

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 Å 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 $n \in N$ of the used lattice $L$ are scaled to a length of 3.6 Å, which is the mean distance between consecutive $C_{\alpha}$ atoms and close to the mean distance between a $C_{\alpha}$ atom and the side chain centroid ($\approx$ 3.6 Å). Given a protein of length $l$ in Protein Databank (PDB) format [7], the positions for each amino acid $i$ to be fitted, i.e. the $C_{\alpha}$ position of the backbone $M_{\alpha}$ and the centroid $P_{\alpha}$ (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 model’s backbone and side chain position ($M_{\alpha}$ and $M_{\beta}$) are placed arbitrarily but adjacent ($M_{\alpha'} - M_{\alpha} \in \mathbb{N}$). For each iteration $1 < i \leq l$, all valid placements of the next $M_{\alpha'}$ and $M_{\beta'}$, the lattice are calculated. We keep the best $n_{\text{keep}}$ 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 Kobasch [8].

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

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). 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 $\leq 0.3$ and resolution $\leq 1.5$ Å; no $C_{\alpha}$-only chains). The resulting benchmark set contains 1198 proteins exhibiting a mean length of 160 (σ = 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 Å.

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

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

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 Å 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


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


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

[9] G. Wang and Roland L. Dunbrack. PISCES: recent improvements to a PDB sequence usage is accessible at

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


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


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