

High-Quality Protein Backbone Reconstruction from Alpha Carbons Using Gaussian Mixture Models

Benjamin L. Moore, Lawrence A. Kelley, James Barber, James W. Murray, and James T. MacDonald*

Coarse-grained protein structure models offer increased efficiency in structural modeling, but these must be coupled with fast and accurate methods to revert to a full-atom structure. Here, we present a novel algorithm to reconstruct main-chain models from C traces. This has been parameterized by fitting Gaussian mixture models (GMMs) to short backbone fragments centered on idealized peptide bonds. The method we have developed is statistically significantly more accurate than several competing methods, both in terms of RMSD values and dihedral angle differences. The method produced Ramachandran dihedral angle distributions that are closer to that observed in real proteins and better Phaser molecular

replacement log-likelihood gains. Amino acid residue side-chain reconstruction accuracy using SCWRL4 was found to be statistically significantly correlated to backbone reconstruction accuracy. Finally, the PD2 method was found to produce significantly lower energy full-atom models using Rosetta which has implications for multiscale protein modeling using coarse-grained models. A webserver and C++ source code is freely available for noncommercial use from: http://www.sbg.bio.ic.ac.uk/phyre2/PD2_ca2main/. © 2013 Wiley Periodicals, Inc.

DOI: 10.1002/jcc.23330

Introduction

It is often useful to work with simplified coarse-grained models of protein structures for reasons of computational efficiency. De Mori et al.,^[1] for example, implemented a reduced C $_{\alpha}$ trace to analyze the molecular dynamics of the Villin headpiece. This complexity reduction afforded the authors a broader sampling of conformational space, while also permitting more targeted simulations to be run on a fine-grained full-atom model. A similar technique has also been applied to modeling the transition state ensemble of the src-SH3 domain.^[2] C $_{\alpha}$ coarse-grained models have also been successfully used for protein structure prediction applications^[3,4] and have the potential to be used in computational protein design.^[5,6] In order for the multiscale modeling approach to be viable, a fast and accurate method of interconversion between coarse- and fine-grained models is required.

Recent evidence has shown that at low resolutions, low crystallographic *R*-values are still achievable if the model is accurate, such as if it is restrained to a similar structure refined at high resolution.^[7,8] The use of modeling techniques is increasing in crystallography^[9] to generate *de novo* models for molecular replacement. The techniques presented here may help in the computationally efficient generation of more complex models.

When model building into electron density maps, C $_{\alpha}$ positions are usually the most accurate, as they can be defined using well-known secondary structure motifs. They are also at branches in the electron density between mainchain and side chain, and successive C $_{\alpha}$ atoms are almost always 3.8 Å or more rarely 3.0 Å apart. These facts make the C $_{\alpha}$ atoms good candidates as the starting points for the reconstruction of the rest of the atoms in the structure. A few PDB (Protein Data Bank) files have incomplete coordinates deposited. The

method described here is an appropriate method to remediate them into full main chain coordinate models. We have tested this against one example from the PDB to see whether crystallographic *R*-values can be improved by this method.

Numerous methods have attempted to solve the problem of reconstructing backbones from C $_{\alpha}$ coordinates by using fragments taken from known protein structures.^[10–18] Methods using backbone fragments have generally been successful and widely used, however they are likely to be highly dependent on the size of the fragment database and may not model loops particularly well. Alternatively, it is possible to rebuild backbones *de novo*, for example, by optimizing a potential of mean force by rotating a peptide group around the axis defined by two adjacent C $_{\alpha}$ atoms.^[19]

Another approach to this problem involves the use of a structural alphabet—a small library of short structural motifs which together can describe most of protein conformational space.^[20,21] A structural alphabet thus enables a three-dimensional protein structure to be represented as a series of one-

B. L. Moore, L. A. Kelley, J. Barber, J. W. Murray, J. T. MacDonald
Division of Molecular Biosciences, Imperial College, South Kensington Campus, London, United Kingdom
E-mail: j.macdonald@imperial.ac.uk

Contract/grant sponsor: Medical Research Council studentship (to BLM); contract/grant number: G0900187-1

Contract/grant sponsor: Biotechnology and Biological Sciences Research Council (BBSRC) (to LAK);

contract/grant number: BB/J019240/1

Contract/grant sponsor: BBSRC David Phillips Fellowship (to JWM); contract/grant number: BB/F023308/1

Contract/grant sponsor: BBSRC through a Eurocores (to JTM); contract/grant number: BB/J010294/1

© 2013 Wiley Periodicals, Inc.

Table 1. Comparison of root mean squared deviation (RMSD) between published PDB structures and backbones reconstructed from C_α traces by the method described in this work, alongside competing methods from BBQ,^[26] MaxSprout,^[14] Milik et al.,^[25] PULCHRA,^[27] SABBAC,^[17] and REMO.^[18]

Structure	RMSD (Å) from this work			RMSD (Å) from previously published methods					
	PD2-min	PD2+min	BBQ	BBQ [†]	MaxSprout	Milik et al.	PULCHRA	SABBAC	REMO
1CRN	0.316	0.306	0.470	0.456	–	0.408	0.358	0.317	0.509
1CTF	0.264	0.239	0.398	0.388	0.750	0.461	0.594	0.327	0.574
1TIM	0.572	0.569	0.621	0.643	0.668	0.595	–	–	–
1UBQ	0.225	0.198	0.214	0.259	0.320	0.324	0.376	0.267	0.490
2ALP	0.398	0.382	0.430	0.462	0.439	0.453	0.691	0.513	0.525
2CTS	0.378	0.366	0.432	0.422	0.484	0.369	0.493	0.417	0.612
2FOX	0.396	0.382	0.352	0.356	0.458	–	0.574	0.456	0.562
2MHR	0.277	0.268	0.269	0.262	0.532	0.457	0.425	0.459	0.545
2OZ9	0.152	0.143	0.218	0.221	0.407	–	0.378	0.204	0.488
2PRK	0.349	0.322	0.345	0.384	0.485	0.358	0.522	0.532	0.549
3APP	0.421	0.404	0.338	0.369	–	0.416	0.524	0.499	0.551
5CPA	0.407	0.377	0.467	0.493	–	0.480	0.698	0.460	0.723
5NLL	0.423	0.406	0.390	0.398	0.518	–	0.569	0.428	0.567
6PTI	0.410	0.381	0.440	0.444	0.558	0.381	0.518	0.509	0.659
9WGA	0.444	0.428	0.471	0.489	0.537	0.450	–	–	–
Mean	0.362	0.345	0.390	0.403	0.513	0.429	0.517	0.414	0.566
Std. Dev.	0.10	0.10	0.11	0.11	0.11	0.07	0.11	0.10	0.07

The RMSDs were calculated using the N, C, and O atoms only. PD2-min refers to the method introduced in this article without additional energy minimization, while PD2+min includes this additional step. Results from Milik et al.^[25] are retrieved from their publication, while other methods were rerun using their default parameters except BBQ[†] which used a database derived from the training set used in this work. Missing results include those structures that have since been superseded since the original Milik et al.^[25] test set, or instances where methods failed to reconstruct a backbone. Structures present in the MaxSprout fragment database were excluded from their results. This test set was only included as it was commonly used in previous papers. *Some structures from this test set have been obsoleted and were substituted as follows: 4FXN replaced by 2FOX, 3FXN replaced by 5NLL, and 2WRP replaced by 2OZ9.

dimensional “letters.”^[22–24] In this work, we derived a new structural alphabet and applied it to the interconversion of coarse- and fine-grained protein models, wherein “letters” fit to a reduced C_α representation can insert missing backbone atoms and build a full backbone model.

Previous methods similar to this approach include that of Milik et al.,^[25] wherein the authors built a library of 4-mer C_α fragments, which were described by three C_α distances oriented around a peptide bond. These distances were discretized to 0.3 Å intervals and combined with a sign representing handedness of the 4-mer. These discretized distances were then used as look-up keys describing approximately 4800 states. For each of these keys, the average C, O, and N vectors of the central peptide bonds from the training set were calculated in a local coordinate system. Then, for any given C_α 4-mer conformation, the C, O, and N vectors can be rapidly looked up and the missing backbone atoms reconstructed. Variations of this basic method were later reimplemented in the software packages BBQ^[26] and PULCHRA.^[27] Both methods achieved fast running times and high accuracy, with BBQ producing an average (N, C, and O atoms only) RMSD to the crystallographic backbone of 0.42 Å over a 35 protein test set.

Not all the methods referenced above were available as downloadable software or as online web services. The available methods that are able to reconstruct mainchain atoms while keeping the C_α atom positions fixed were compared to the method described in this article. Some methods^[5,18,27] are additionally able to refine C_α positions but this aspect is not explored in this article.

Methodology

In this work, we describe a novel method of constructing a structural alphabet using Gaussian mixture models (GMMs). In order to build a structural alphabet, a library of short fixed length backbone fragments from a high resolution training set was built. Each fragment's central peptide bond was centered upon an idealized *cis* or *trans* peptide bond (consisting of the C, O and N backbone atoms) constructed using residue definitions from the CHARMM forcefield.^[28] The training set C_α atoms either side of the peptide bond then formed a cloud of possible C_α positions relative to a fixed peptide bond. This cloud of C_α atoms was used to construct a new structural alphabet by fitting GMMs. Aligning the C_α coordinates of the best fitting “letter” to successive segments of the target C_α trace model could then be used to determine the coordinates of the central peptide bond of that segment and eventually rebuild the entire protein backbone.

A large dataset of fragments was derived from ASTRAL SCOP 1.75A,^[29] filtered by 40% sequence identity and retaining only X-ray crystallography-derived structures with high reliability and precision, as defined by an AEROSPACI quality score >0.5 (this score is approximately the reciprocal of the resolution).^[29] Finally, PSI-BLAST^[30] was used to filter structures that were homologous to those in our first test set (Table 1) to avoid overfitting. This left approximately 2800 high-resolution structures in our training set.

The training set of PDB structures was then decomposed into fragments. Several fragment sets were built for

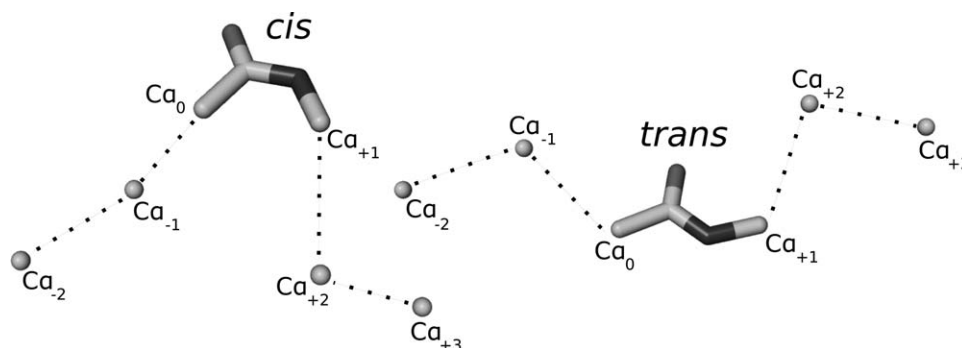


Figure 1. Example *cis* and *trans* components of a 6-mer alphabet. C_α are numbered with respect to $C_{\beta 0}$. β_0 was placed at (0,0,0), C_β at (1.5461, 0, 0), C_α was placed in the xy -plane at (-0.554, 1.436, 0). The O_0 , C_α , and N_{+1} coordinates were then calculated according to the CHARMM residue definitions for alanine for *trans* peptide bonds or proline for *cis* peptide bonds. Only the $C_{\alpha 0}$, O_0 , C_α , N_{+1} , and $C_{\alpha+1}$ were used for structural superposition in order to construct the fragment library training set.

comparison, ranging in length from 4 to 7 consecutive C_α atoms. Odd-length fragments were aligned to geometrically idealized central residue C_α , C_β , N, and carbonyl C_α atoms. Even-length fragments were instead centered on an idealized peptide bond built from CHARMM definitions.^[28] *Cis* and *trans* fragments were modeled with separate GMMs. Early comparisons suggested 6-mer fragments were an optimum value in this range (Supporting Information Fig. 1) and a library of 480,000 fragments was built from the training set.

A GMM was then fitted to C_α coordinates of length six fragments centered on an idealized peptide bond. This was done by converting the library of fragments to a dataset of 12-dimensional data points, whereby each point contained four sets of relative C_α coordinates in three-dimensional space. Using Figure 1 as a reference, a single point in the dataset would be:

$$\mathbf{x} = (C_{\alpha-2}^x, C_{\alpha-2}^y, C_{\alpha-2}^z, \dots, C_{\alpha+3}^x, C_{\alpha+3}^y, C_{\alpha+3}^z)$$

Note that $C_{\alpha 0}$ and $C_{\alpha+1}$ are adjacent to the fixed idealized peptide bond, hence their positions varied by very little. For this reason, they were not included in the data for GMM fitting.

The first step in fitting a GMM to this dataset was the hierarchical clustering of datapoints which provides starting values for the expectation-maximization (EM) algorithm. Due to the number of datapoints being considered, the initial clustering was performed on a random sample of 2000 points. These clusters were then used to initialize iterative rounds of EM, using the complete dataset, until convergence.

The Bayesian information criterion (BIC) was used to determine the optimum number of model components. This metric penalizes the inclusion of additional parameters weighted against the improved log-likelihood of the model to avoid overfitting. The BIC is more suitable in this instance relative to other measures, such as the Akaike information criterion, as it more strongly penalizes additional model components. We expect a greater number of components to result in a better performing alphabet, but the complexity of the model must be reasonably constrained in order to retrieve an alphabet small enough to maintain a fast reconstruction. A coarse-

grained BIC comparison combined with reconstruction accuracy tests lead to the selection of a 528 component unconstrained mixture model (Supporting Information Fig. 2).

GMMs are composed of a weighted sum of multivariate Gaussian distributions giving the log-likelihood function shown in eq. (1).

$$\ell_G = \sum_{i=1}^n \log \sum_{k=1}^G w_k \varphi(\mathbf{x}_i | \mu_k, \Sigma_k) \quad (1)$$

where w_k represents the weight of component k , while function φ is a multivariate Gaussian probability density function with vector of means, μ_k , and covariance matrix, Σ_k . Σ_k is eigenvalue decomposed [eq. (2)].

$$\Sigma_k = \lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T \quad (2)$$

where for each component k , the matrix of eigenvectors \mathbf{D} represents the cluster orientation, while the corresponding eigenvalue terms \mathbf{A} determines cluster shape. The scalar λ controls cluster volume. These parameters were not constrained. The R package MCLUST was used to fit the model to the data.^[31]

An uninformative inverse Wishart Bayesian prior^[32] was invoked upon the covariance matrix of the GMM to prevent convergence on singularities. In the case of unconstrained models with >100 components, it was common for a covariance term to converge on zero. The use of a conjugate prior in this way enabled the avoidance of such singularities evolving in unconstrained models, but otherwise has little effect on the outcome relative to a non-Bayesian EM and MLE (maximum likelihood estimation) approach. The 528 maximum *a posteriori* estimates of the component means (corresponding to C_α coordinates) together with fixed idealized central peptide bond atom positions then constituted the structural alphabet. For the purposes of this work, the covariance terms were discarded. Given a set of 6-mer C_α coordinates in the target structure, the member of the alphabet that minimized a weighted C_α RMSD superposition value using a fast quaternion-based method^[33,34] was determined. Weighting (W) was successively decreased by a factor of ten for C_α atoms at outer positions such that (relative to Fig. 1):

$$W(C\alpha_{-3}) = \frac{1}{100}, \quad W(C\alpha_{-2}) = \frac{1}{10}, \quad W(C\alpha_{-1}) = 1$$

$$W(C\alpha_0) = 1, \quad W(C\alpha_{+1}) = \frac{1}{10}, \quad W(C\alpha_{+2}) = \frac{1}{100}$$

This weighting helps to reduce the impact of averaging errors which may distort the geometry of the outermost C_α in some members of the alphabet. The factor of ten used here was not thoroughly optimized but gave adequate results. The rotation matrix from fitting the C_α atoms of the optimal "letter" was then used to determine the placement of the idealized peptide bond on the target structure. C_β and backbone amide hydrogen atom coordinates were then added using CHARMM residues definitions.^[28]

Our software can also provide an optional gradient energy minimization procedure which can further improve the rebuilt protein backbone at the expense of a longer runtime. The C_α atoms were kept fixed in position during minimization with all other backbone atoms free to move. This was performed using a previously described simple backbone potential energy function^[5] consisting of local structure molecular mechanics terms derived from the OPLS-UA force field (bond, torsion, improper torsion, 1–4 Lennard–Jones, 1–4 electrostatic, 1–5 Lennard–Jones, and 1–5 electrostatic), a soft steric repulsive term for atom pairs separated by more than four bonds and statistical backbone hydrogen bonding potential terms.

In fitting length six fragments to a C_α trace, there is inevitably a problem at the beginning and end of each polypeptide chain, where there are insufficient C_α atoms either side of the peptide bonds to anchor the fragment. We solved this problem in a manner similar to that of Milik et al.^[25] by adding pseudo- C_α to the N and C-termini with which we can then use to fully align fragments. Terminal backbone atoms are relatively unconstrained by C_α atoms so are generally poorly defined. Hence for fair comparisons, each backbone RMSD was calculated excluding the first NCONCON and last CONCONCO atoms for all methods.

In performance testing our program, two test sets were used: 15 structures previously used as a test set by Milik et al.^[25] and a second set of recently deposited PDB structures. To prevent overfitting, we removed homologues of the first test set from our training set that aligned with a PSI-BLAST E-value of ≥ 0.01 .

The second test set was made up of a filtered set of structures deposited to the RCSB PDB between 15/03/2012 and 08/06/2012, after the date ASTRAL-SCOP 1.75A was released. The structures in this set were required to be better than 2 Å resolution X-ray crystallography-derived protein structures with no more than 30% sequence similarity to each other. Additionally, this set was filtered for proteins with backbone chain breaks, unusual backbone bond lengths or multiple chains. This left 28 structures in the second test set.

Due to the disparate age and quality of the PDB structures in these two sets, they were analyzed separately. It is likely that the modern structures are much better refined due to methodological improvements.

The algorithm was implemented as part of a larger in-house protein structure modeling software package called PD2 and was written in the C++ programming language.

Results

The algorithm presented in this work was able to efficiently and accurately rebuild full-atom protein molecules from C_α traces. Table 1 shows a comparison of this method with previously described methods using the test set of PDB structures implemented in Milik et al.^[25] In nine of the 15 cases, our method outperforms the five other methods. Additionally, the mean backbone RMSD across the 15 structures is lowest for our method, both with and without energy minimization. When the energy minimization refinement step is included, the mean backbone RMSD improves by 0.02 Å on average across the test set, with no structures becoming worse.

The protein backbone with the lowest RMSD in this test set is that of 2OZ9, with a value of 0.143 Å for the average difference between N, C, and O backbone atoms. This structure is predominantly α -helical (76% α -helix, with six helices across 82 residues^[35]), whose regularity is more easily predicted relative to loop regions.^[21] In comparison, our least accurate reconstruction, that of 1TIM, is 45% helical and 17% β -sheet over 247 residues,^[36] in addition to being relatively low resolution (2.5 Å) with 9.6% of residues in Ramachandran outlier regions.^[37] Closer inspection of the individual residues contributing to the high RMSD, however, did not comprehensively correlate coil regions with high RMSD, with some high RMSD residues being located in α -helices and β -sheets. A combination of inaccuracies in the experimental PDB structure and potentially difficult to model loop regions may explain this poor result.

Overall, it should be noted that the test set used in Table 1 contains structures deposited in the 1970s and 1980s which are generally of lower quality than modern structures without modern advances in refinement techniques. An analysis of PD2 rebuilt structures in this test set found Ramachandran dihedral angles to be 93.9% favorable, 3.8% allowable, and 2.3% outliers, as defined by Lovell et al.^[38] These results are below the ideal 98%, 2%, 0% (favourable, allowable, and outlier respectively) expected of high-resolution structures but were an improvement on the torsion angles present in the original PDB files (91.1%, 5.2%, 3.7%). For these reasons, this test set may not be ideal for comparing the accuracy of the different methods. Nevertheless, it allowed comparison with existing methods and showed that the proposed method is capable of reconstructing protein backbones from potentially inaccurate C_α traces.

In a second comparison, we compared the results of the algorithms applied to a modern high resolution test set (Table 2). The additional energy gradient minimization step over this test set increased the accuracy of our reconstructions by an average of approximately 0.02 Å, but increased the algorithm's runtime. Over the test set shown in Table 2, runtime was found to be linearly dependent on chain length, building backbone residues at a rate of around 460 residues per second with an initialization time of 0.2 s on an Intel Core2 Duo T9300 2.50 GHz CPU. Adding the energy minimization step increased this average runtime to around 15 residues per seconds (Supporting Information Fig. 3). The extra runtime,

Table 2. Backbone atoms RMSD comparison using 28 recent, high-resolution (<2 Å) PDB structures with and without energy minimization refinement, alongside those methods also compared in Table 1.

Structure	RMSD (Å) from this work						RMSD (Å) from previously published methods					
	PD2-min	C	N	O	C	PD2+min	BBQ	BBQ [†]	MaxSprout	PULCHRA	SABBAC	REMO
4ANN	0.261	0.155	0.149	0.397	0.201	0.238	0.292	0.295	0.405	0.452	0.445	0.514
4E9L	0.289	0.158	0.138	0.455	0.205	0.267	0.315	0.313	0.403	0.546	0.544	0.540
4EG9	0.339	0.189	0.167	0.530	0.259	0.317	0.427	0.420	0.478	0.504	0.350	0.531
4EIU	0.366	0.198	0.180	0.573	0.224	0.334	0.339	0.368	0.468	0.587	0.555	0.556
4EO0	0.225	0.121	0.126	0.347	0.174	0.211	0.203	0.207	0.392	0.348	0.286	0.437
4EV1	0.253	0.144	0.144	0.388	0.200	0.239	0.304	0.319	0.396	0.411	0.385	0.504
4EXO	0.271	0.155	0.147	0.417	0.199	0.264	0.289	0.308	0.447	0.459	0.436	0.530
4EYO	0.287	0.160	0.155	0.443	0.220	0.271	0.368	0.364	0.384	0.569	0.344	0.558
4F78	0.385	0.212	0.186	0.604	0.250	0.364	0.424	0.386	0.552	0.558	0.477	0.616
4F7H	0.326	0.190	0.199	0.492	0.264	0.293	0.356	0.387	0.505	0.471	0.480	0.551
4F7V	0.458	0.241	0.202	0.727	0.242	0.444	0.410	0.421	0.436	0.617	0.450	0.643
4F8J	0.293	0.168	0.155	0.452	0.216	0.279	0.354	0.346	0.388	0.510	0.370	0.458
4F8X	0.363	0.200	0.178	0.569	0.246	0.330	0.380	0.364	0.538	0.562	0.488	0.542
4FAK	0.307	0.167	0.148	0.481	0.206	0.285	0.325	0.280	0.461	0.461	0.366	0.514
4FAT	0.261	0.152	0.153	0.396	0.201	0.227	0.413	0.398	0.359	0.540	0.431	0.632
4FB7	0.270	0.149	0.155	0.416	0.179	0.248	0.335	0.321	0.344	0.461	0.359	0.506
3VTF	0.325	0.178	0.172	0.505	0.209	0.310	0.342	0.350	0.468	0.551	0.401	0.538
4AVX	0.241	0.139	0.143	0.367	0.197	0.227	0.314	0.315	0.360	0.402	0.318	0.474
4AVZ	0.382	0.210	0.183	0.601	0.235	0.352	0.369	0.378	0.480	0.568	0.489	0.602
4FBR	0.438	0.231	0.186	0.698	0.253	0.404	0.399	0.408	0.528	0.586	0.552	0.622
4FCS	0.342	0.198	0.189	0.525	0.241	0.329	0.326	0.338	0.458	0.504	0.452	0.533
4FCU	0.315	0.169	0.142	0.498	0.204	0.298	0.342	0.346	0.455	0.458	0.501	0.502
4FD5	0.325	0.175	0.160	0.509	0.210	0.318	0.319	0.313	0.426	0.421	0.406	0.515
4FE3	0.287	0.161	0.154	0.445	0.216	0.275	0.318	0.312	0.489	0.449	0.450	0.507
4FE9	0.408	0.223	0.192	0.642	0.250	0.392	0.406	0.415	0.433	0.573	0.534	0.568
4FFK	0.362	0.189	0.195	0.564	0.242	0.355	0.447	0.451	0.390	0.604	0.441	0.607
4FHG	0.339	0.184	0.173	0.530	0.237	0.318	0.350	0.339	0.397	0.549	0.425	0.506
4FIK	0.393	0.214	0.181	0.619	0.231	0.374	0.356	0.370	0.288	0.568	0.575	0.563
Mean	0.325	0.180	0.166	0.507	0.222	0.306	0.350	0.351	0.433	0.510	0.440	0.542
Std. Dev.	0.06	0.03	0.02	0.10	0.02	0.06	0.05	0.05	0.06	0.07	0.08	0.05

The RMSDs were calculated using the N, C, and O atoms only. The method used in Milik et al.^[25] is not publicly available, so is not present in this set. BBQ[†] again represents the BBQ program run using a database derived from our training set, while BBQ is the same program run using its default database described in Gront et al.^[26] The PD2+min RMSD values were statistically significantly less than all the other methods at the level using the one-tailed Mann–Whitney U test (p -values: 0.002 (BBQ[†]), 0.002 (BBQ), 2×10^{-8} (MaxSprout), 4×10^{-8} (SABBAC), 1×10^{-10} (PULCHRA), 8×10^{-11} (REMO)). The PD2-min RMSD values were also significantly less than all methods (p -values: 0.03 (BBQ), 0.03 (BBQ[†]), 2×10^{-7} (MaxSprout), 7×10^{-7} (SABBAC), 5×10^{-10} (PULCHRA), 9×10^{-11} (REMO)).

however, resulted in a more accurate reconstruction in all the test set members (Table 2).

Overall, as with our previous comparison, our algorithm had the lowest mean RMSD across the test set, and this was further improved by the energy minimization step. The PD2 method with and without minimization was found have statistically significantly lower RMSD values across the second test set compared to all other methods (Table 2).

PD2 Ramachandran dihedral angles shifts compared to the original structure were statistically significantly less than the other methods (Fig. 2). 5.6% of Ramachandran angles displayed $>45^\circ$ shifts using the PD2 method, 4.2% for PD2+min, 6.9% for BBQ (both training sets), 9.7% for MaxSprout, 15.8% for PULCHRA, 11.5% for SABBAC, and 18.6% for REMO.

These shifts were examined for systematic or persistent errors (Supporting Information Figs. 3–5). It would be of concern, for example, if certain torsion angles were consistently being shifted from a favorable region to an anomalous or unfavorable conformation. Interestingly, the BBQ torsion shifts

do seem to show some systematic bias (Supporting Information Figs. 3(b) and 4(b)). Ramachandran angles found in the lower right allowed region (in the box defined by approximately (60, −180), (85, −90)) are all shifted out of this region. In comparison, our method retains most dihedrals in this region. BBQ also appears to be systematically shifting residues into the region roughly defined by the box (60, −90), (90, −20).

As a test of model structure quality, the entire training set was rebuilt using both the new algorithm (PD2+min) and BBQ, the highest performing previously described method. Ramachandran dihedral angles were divided into 2D histograms with $18^\circ \times 18^\circ$ bins. Each bin was given an additional pseudocount of 1 to prevent empty bins. Kullback–Leibler (KL) divergence^[39] values were calculated used a measure of difference between the training set and rebuilt Ramachandran dihedral angle distributions (Fig. 3). The PD2 method was found to be less divergent from the SCOP training set for all residue types except *cis*-prolines.

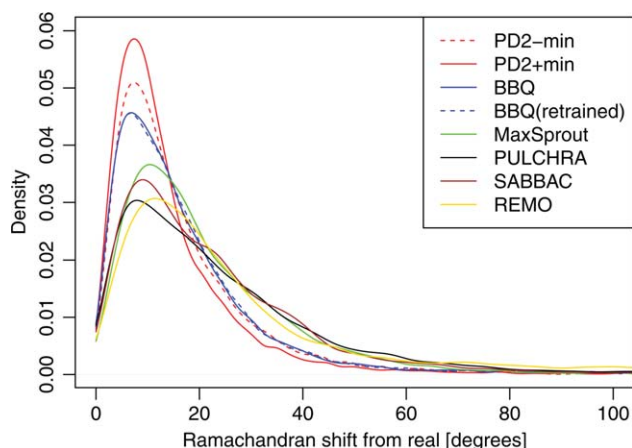


Figure 2. Ramachandran dihedral angle shifts for all *trans*-conformation residues in the second test set (Table 2) displayed by the different methods where the shift is defined as $\sqrt{\Delta\phi^2 + \Delta\psi^2}$. $\Delta\phi$ and $\Delta\psi$ were calculated taking into account periodicity. PD2+min shifts were found to be significantly less than that of other methods using the Mann–Whitney U-test (*p*-values: 0.002 (BBQ), 0.0002 (BBQ_r), 5×10^{-112} (MaxSprout), 6×10^{-125} (SABBAC), 3×10^{-160} (PULCHRA), 2×10^{-227} (REMO)). PD2+min values were also statistically significantly less than the other methods.

Biologically significant results

The rebuilt backbones from the second test set were used as models for molecular replacement using Phaser.^[40] The original crystal structures were stripped to main chain and C_β atoms, B factors set to 20, and occupancies to 1. The atoms of the reconstructed test set PDB files were also set to occupancy 1 and B factor 20. Where the method did not add C_β atoms, these were added using CHARMM residue definitions. The processed files were then used as molecular replacement models in Phaser, against the deposited structure factors. In all cases, the structure gave the correct molecular replacement solution, and the refined log-likelihood gain (LLG) score was noted. This is the LLG of the solution over the distribution of structure factors given by a Wilson distribution. The PD2+min method was found to have the highest LLGs for 22 out of 28 test structures and was statistically significantly higher than all other methods (Table 3).

In a molecular modeling context, the use of coarse-grained models is limited by the accuracy with which it is possible to reconstruct full atoms models. In order to test this, SCWRL4^[41] was used to predict sidechains using the reconstructed backbones (Fig. 4, Supporting Information Table 1 and Fig. 7). Despite the numerically small backbone RMSD differences and the fixed C_α positions, the accuracy of sidechain reconstruction was found to be statistically significantly correlated to backbone model accuracy (Fig. 4). It can be imagined that small errors in the backbone geometry could magnify to large errors in the sidechain placements due to the rough nature of all-atom energy landscape. Although, the PD2+min method gave the lowest mean RMSD, PD2, and BBQ methods were not statistically distinguishable from each other but they were both statistically significantly better than all other tested methods (Supporting Information Table 1). Interestingly, the crystal

structure backbones almost invariably gave the best SCWRL4 sidechain reconstruction accuracy.

A related question is whether, given a perfect C_α coarse-grained model (i.e. zero C_α RMSD), does the method of backbone reconstruction have a significant effect on further full-atom refinement and model selection? Using the Rosetta molecular modeling package,^[42] sidechains were added and repacked on the reconstructed backbones, and the resulting full-atom models were energy minimized. The PD2+min method gave a statistically significant lower energy and the PD2+min method was statistically significantly better than all methods except BBQ retrained with our training set (Supporting Information Table 2). The crystal structure backbones (with sidechains reconstructed by Rosetta) gave the lowest Rosetta energies in almost all cases.

Finally, we have tested whether the PD2 method can improve upon an incomplete main chain model of respiratory complex I at 3.9 Å resolution (PDB: 3M9C^[43]). The 3M9C coordinate set has a polyglycine backbone with no carbonyl oxygen atoms. *R*-values were calculated using 0 cycles of REFMAC^[44] of the deposited and reconstructed coordinates against the deposited structure factors. This gave an *R*-value of 52.9%. The PD2+min reconstructed PDB file gave an *R*-value of 47.4%. This shows that even without model building with reference to the crystallographic data, better models can be produced using a knowledge of protein structure alone.

Discussion

The method presented in this work compares favorably with previous methods, both when run with a secondary energy minimization step and without (Table 2). Of the five existing reconstruction methods used in comparisons, the BBQ program has proved to be the closest competitor, though it uses a much larger fragment library of 5148 relative to our 528.^[26] Additionally, the observed accuracy improvement was not merely a function of an expanded PDB database, as shown by comparisons with a retrained BBQ method using an identical training set to that of PD2 (BBQ[†]; Tables 1 and 2).

When energy minimization is applied, our program appears to give more accurate results for most structures (Table 2), so this longer process may prove beneficial when accuracy is more important than runtime. These system runtimes correspond to approximately 460 and 15 residues per second for unminimized and minimized reconstructions, respectively (on an Intel Core2 Duo T9300 2.50GHz), and around 0.2 s constant overhead time spent handling input/output of the alphabet and PDB files and initialization.

All the molecular replacement models gave good results that would have been sufficient to solve the structure (Table 3). However, in marginal cases the quality of model reconstruction could make the difference between solution and nonsolution of a structure. In this case, the PD2 approach would be the preferred option.

Backbone reconstruction accuracy was found to significantly impact both sidechain reconstruction accuracy and reconstructed full-atom model energy. In multiscale molecular

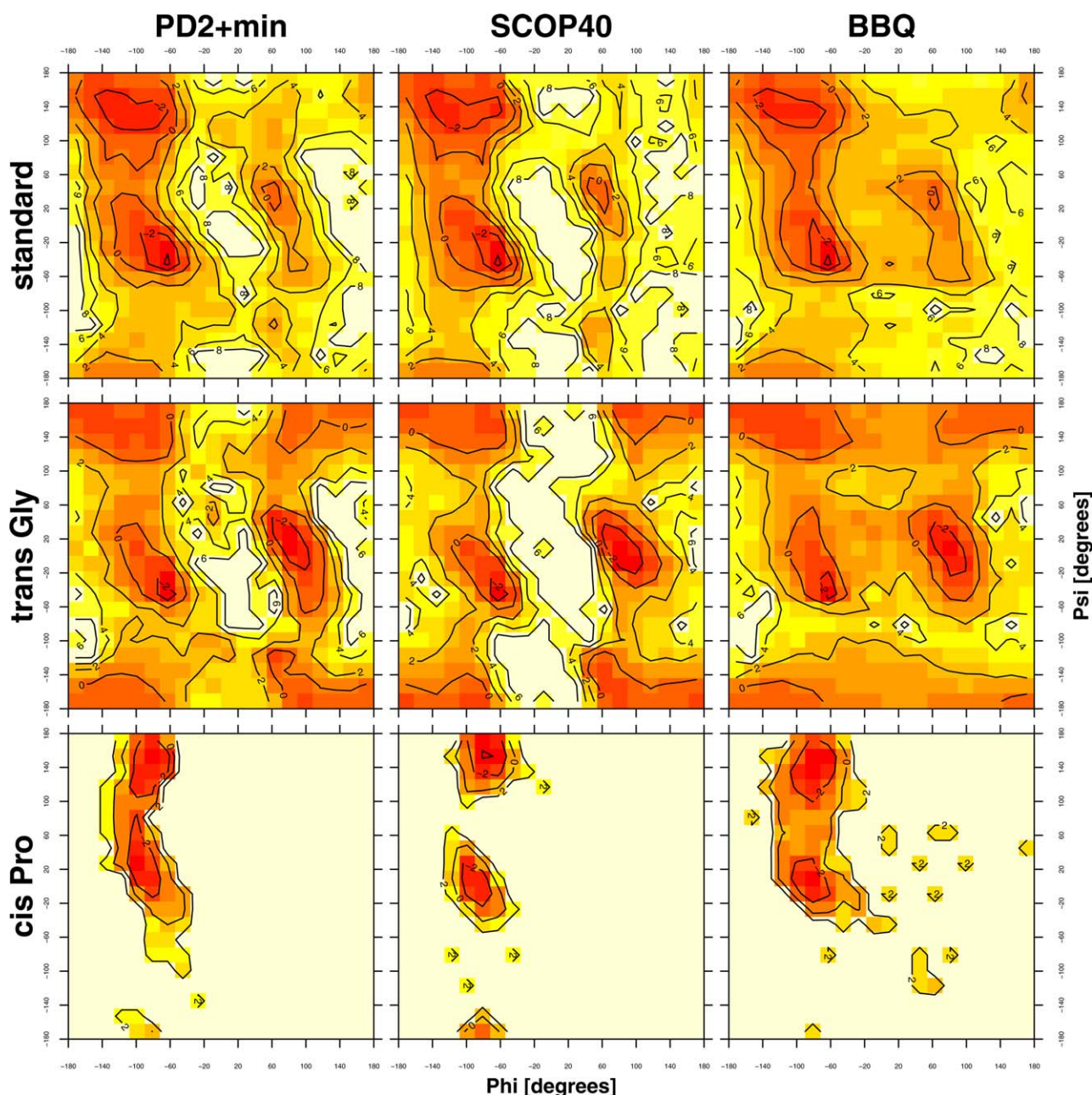


Figure 3. Log probability Ramachandran dihedral angle distributions of rebuilt SCOP training set backbones using the new method described in this article (PD2+min; left column), the best pre-existing method (BBQ_t trained using our training set; right column), and the distribution observed in the SCOP training set (SCOP40; central column). The distributions were split into $18^\circ \times 18^\circ$ 2D bins and the log probabilities are shown in the plot. The Ramachandran distributions were further subdivided into nonproline nonglycine *trans* conformation residues (standard), *trans* conformation glycines (*trans* Gly), and *cis* conformation prolines (*cis* Pro). For standard and *trans*-glycine residues, PD2+min was closer to the observed SCOP40 distribution (Kullback–Leibler divergence 0.05 and 0.19 bits, respectively) than BBQ (KL divergence 0.08 and 0.34 bits) but for *cis*-prolines it was more divergent to the observed SCOP40 distribution than BBQ (0.63 and 0.50 bits, respectively). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

modeling applications, this could result in low-energy states being missed and lead to problems in model selection. The observation that the crystal structure backbones lead to better SCWRL4 sidechain reconstruction accuracy shows that there is scope for further improvement (Fig. 4(b)).

In the future, as larger complexes are solved, at lower resolutions, tools such as PD2 can be used to help interpret maps using prior knowledge of protein structure.

Conclusions

We present a fast and accurate method of protein backbone reconstruction from C_α trace which has a range of potential applications in structural biology, such as protein structure prediction or *de novo* protein design. Our program features an optional energy minimization step which is able to further refine backbone atom placement.

Table 3. Phaser^[40] log-likelihood gains (LLG) using the same second high resolution test set and methods compared in Table 2.

Phaser LLG—This work					Phaser LLG—Previously published methods				
Structure	Phaser LLG using crystal structure	PD2—min	PD2+min	BBQ	BBQ [†]	MaxSprout	PULCHRA	SABBAC	REMO
4ANN	1252	1099	1123	1083	1082	936	953	951	916
4E9L	2160	1887	1924	1858	1849	1646	1619	1619	1616
4EG9	1146	1004	1030	981	969	845	867	965	895
4EIU	1150	953	986	979	995	853	842	850	856
4EO0	537	484	485	491	493	404	427	461	403
4EV1	1113	1014	1019	987	988	865	917	921	858
4EXO	843	717	725	727	731	598	640	643	639
4EYO	1915	1706	1728	1650	1642	1459	1467	1615	1460
4F78	1222	1047	1067	1028	1018	878	912	987	871
4F7H	785	685	703	688	683	563	612	604	589
4F7V	898	744	752	742	750	636	663	692	627
4F8J	2267	1912	1944	1920	1901	1597	1661	1831	1678
4F8X	2051	1638	1693	1617	1617	1319	1403	1514	1320
4FAK	938	816	828	827	831	671	778	788	724
4FAT	1315	1090	1141	1036	1042	895	896	906	773
4FB7	1948	1652	1685	1661	1654	1404	1484	1576	1367
3VTF	2316	2026	2041	1977	1981	1666	1807	1887	1738
4AVX	1167	1026	1033	1037	1046	844	958	993	896
4AVZ	3639	2940	3066	2963	2963	2530	2468	2679	2508
4FBR	1776	1363	1452	1360	1353	1110	1181	1183	1228
4FCS	2123	1741	1773	1768	1772	1495	1583	1607	1478
4FCU	1389	1241	1261	1221	1234	1063	1092	1063	1092
4FD5	1280	1099	1126	1114	1120	948	975	1015	910
4FE3	1781	1549	1577	1548	1548	1280	1384	1413	1322
4FE9	2166	1829	1865	1807	1816	1702	1616	1684	1648
4FFK	1145	926	935	904	898	837	818	859	779
4FHG	1635	1386	1417	1401	1404	1283	1234	1321	1229
4FIK	1625	1225	1269	1224	1225	1346	987	999	1133

BBQ[†] again represents the BBQ program run using a database derived from our training set, while BBQ is the same program run using its original database. The PD2+min LLG values were statistically significantly greater than all the other methods at the level using the one-tailed binomial test with (*p*-values: 1×10^{-5} (BBQ), 0.0005 (BBQ[†]), 1×10^{-7} (MaxSprout), 4×10^{-9} (SABBAC), 4×10^{-9} (PULCHRA), 8×10^{-11} (REMO)). The PD2—min LLG values were also significantly greater than all methods except BBQ (*p*-values: 0.3 (BBQ), 0.3 (BBQ[†]), 1×10^{-7} (MaxSprout), 4×10^{-9} (SABBAC), 4×10^{-9} (PULCHRA), 4×10^{-9} (REMO)).

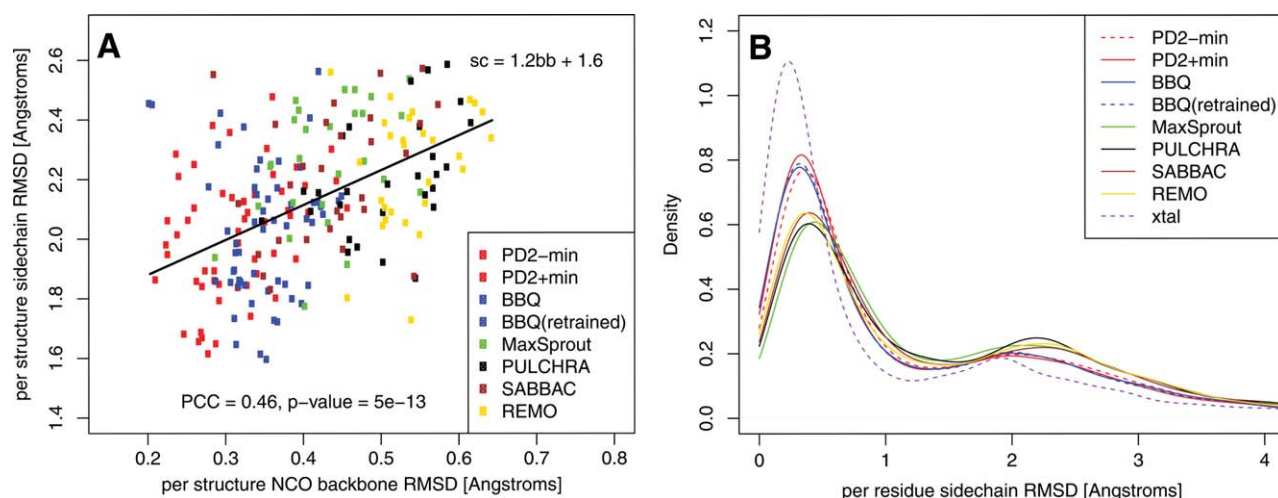



Figure 4. Sidechain reconstruction accuracy show (A) as a function of backbone RMSD and (B) by a histogram of reconstructed sidechain RMSD. Reconstructed backbones were taken from the second test set (Table 2) and using these backbones, the sidechains were reconstructed using SCWRL4.^[41] Backbone RMSDs were calculated using only the NCO backbone atoms having first superimposed the reconstructed backbones on the crystal structure using the C_{α} atoms only as some methods recenter the protein coordinates. Sidechain RMSDs were calculated using all sidechain heavy atoms excluding C_{β} . These two measures were found to have a statistically significant Pearson correlation coefficient of 0.46. Fitting a linear model gave a *y*-axis intercept of 1.6 Å which is close to the 1.7 Å mean SCWRL4 reconstructed sidechain RMSD using the crystal structure backbone (Supporting Information Table 1). The sidechain RMSDs appear to be bimodally distributed which is likely a result of the different size sidechains. Sidechains reconstructed using the crystal structure backbones appear to be substantially better than all methods.

Keywords: protein structure modeling · protein backbone · coarse-grained model · webserver · multiscale protein modeling

How to cite this article: B. L. More, L. A. Kelley, J. Barber, J. W. Murray, J. T. MacDonald *J. Comput. Chem.* **2013**, *34*, 1881–1889. DOI: 10.1002/jcc.23330

 Additional Supporting Information may be found in the online version of this article.

- [1] G. De Mori, G. Colombo, C. Micheletti, *Proteins* **2005**, *58*, 459.
- [2] F. Ding, W. Guo, N. V. Dokholyan, E. Shakhnovich, J. E. Shea, *J. Mol. Biol.* **2005**, *350*, 1035.
- [3] Y. Zhang, J. Skolnick, *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 7594.
- [4] Y. Zhang, *BMC Bioinform.* **2008**, *9*, 40.
- [5] J. T. MacDonald, K. Maksimiak, M. I. Sadowski, W. R. Taylor, *Proteins* **2010**, *78*, 1311.
- [6] I. Coluzza, J. T. MacDonald, M. I. Sadowski, W. R. Taylor, R. A. Goldstein, *PLoS One* **2012**, *7*, e34228.
- [7] G. F. Schröder, A. T. Brunger, M. Levitt, *Structure* **2007**, *15*, 1630.
- [8] G. F. Schröder, M. Levitt, A. T. Brunger, *Nature* **2010**, *464*, 1218.
- [9] F. DiMaio, T. C. Terwilliger, R. J. Read, A. Wlodawer, G. Oberdorfer, U. Wagner, E. Valkov, A. Alon, D. Fass, H. L. Axelrod, D. Das, S. M. Vorobiev, H. Iwai, P. R. Pokkuluri, D. Baker, *Nature* **2011**, *473*, 540.
- [10] T. A. Jones, S. Thirup, *EMBO J.* **1986**, *5*, 819.
- [11] M. Claessens, E. Van Cutsem, I. Lasters, S. Wodak, *Protein Eng.* **1989**, *2*, 335.
- [12] L. S. Reid, J. M. Thornton, *Proteins* **1989**, *5*, 170.
- [13] M. Levitt, *J. Mol. Biol.* **1992**, *226*, 507.
- [14] L. Holm, C. Sander, *J. Mol. Biol.* **1991**, *218*, 183.
- [15] M. Feig, P. Rotkiewicz, A. Kolinski, J. Skolnick, C. L. Brooks, *Proteins* **2000**, *41*, 86.
- [16] S. A. Adcock, *J. Comput. Chem.* **2004**, *25*, 16.
- [17] J. Maupetit, R. Gautier, P. Tufféry, *Nucleic Acids Res.* **2006**, *34*, W147.
- [18] Y. Li, Y. Zhang, *Proteins* **2009**, *76*, 665.
- [19] P. W. Payne, *Protein Sci.* **1993**, *2*, 315.
- [20] M. J. Rومان, J. Rodriguez, S. J. Wodak, *J. Mol. Biol.* **1990**, *213*, 327.
- [21] B. Offmann, M. Tyagi, A. G. de Brevern, *Curr. Bioinform.* **2007**, *2*, 165.
- [22] B. H. Park, M. Levitt, *J. Mol. Biol.* **1995**, *249*, 493.
- [23] C. Etchebest, C. Benros, S. Hazout, A. G. de Brevern, *Proteins* **2005**, *59*, 810.
- [24] A. Pandini, A. Fornili, J. Kleinjung, *BMC Bioinform.* **2010**, *11*, 97.
- [25] M. Milik, A. Kolinski, J. Skolnick, *J. Comput. Chem.* **1997**, *18*, 80.
- [26] D. Gront, S. Kmiecik, A. Kolinski, *J. Comput. Chem.* **2007**, *28*, 1593.
- [27] P. Rotkiewicz, J. Skolnick, *J. Comput. Chem.* **2008**, *29*, 1460.
- [28] B. R. Brooks, C. L. Brooks, III, A. D. Mackerell, Jr., L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caffisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kucera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, M. Karplus, *J. Comput. Chem.* **2009**, *30*, 1545.
- [29] J. M. Chandonia, G. Hon, N. S. Walker, L. L. Conte, P. Koehl, M. Levitt, S. E. Brenner, *Nucleic Acids Res.* **2004**, *32*, D189.
- [30] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman, *Nucleic Acids Res.* **1997**, *25*, 3389.
- [31] C. Fraley, A. E. Raftery, *J. Classif.* **1999**, *16*, 297.
- [32] C. Fraley, A. E. Raftery, *J. Classif.* **2007**, *24*, 155.
- [33] D. L. Theobald, *Acta Crystallogr. A* **2005**, *61*, 478.
- [34] P. Liu, D. Agrafiotis, D. Theobald, *J. Comput. Chem.* **2010**, *31*, 1561.
- [35] C. L. Lawson, R. Zhang, R. W. Schevitz, Z. Otwinowski, A. Joachimiak, P. B. Sigler, *Proteins* **1988**, *3*, 18.
- [36] D. Banner, A. Bloomer, G. Petsko, D. Phillips, I. Wilson, *Biochem. Biophys. Res. Co.* **1976**, *72*, 146.
- [37] V. B. Chen, W. B. Arendall, J. J. Headd, D. A. Keedy, R. M. Immormino, G. J. Kapral, L. W. Murray, J. S. Richardson, D. C. Richardson, *Acta Crystallogr. D* **2010**, *66*, 12.
- [38] S. C. Lovell, I. W. Davis, W. B. Arendall, III, P. I. W. de Bakker, J. M. Word, M. G. Prisant, J. S. Richardson, D. C. Richardson, *Proteins* **2003**, *50*, 437.
- [39] S. Kullback, R. A. Leibler, *Ann. Math. Stat.* **1951**, *22*, 79.
- [40] A. J. McCoy, R. W. Grosse-Kunstleve, P. D. Adams, M. D. Winn, L. C. Storoni, R. J. Read, *J. Appl. Crystallogr.* **2007**, *40*, 658.
- [41] G. G. Krivov, M. V. Shapovalov, R. L. Dunbrack, *Proteins* **2009**, *77*, 778.
- [42] A. Leaver-Fay, M. Tyka, S. M. Lewis, O. F. Lange, J. Thompson, R. Jacak, K. Kaufman, P. D. Renfrew, C. A. Smith, W. Sheffler, I. W. Davis, S. Cooper, A. Treuille, D. J. Mandell, F. Richter, Y. A. Ban, S. J. Fleishman, J. E. Corn, D. E. Kim, S. Lyskov, M. Berrondo, S. Mentzer, Z. Popovic, J. J. Havranek, J. Karanicolas, R. Das, J. Meiler, T. Kortemme, J. J. Gray, B. Kuhlman, D. Baker, P. Bradley, *Methods Enzymol.* **2011**, *487*, 545.
- [43] R. G. Efremov, R. Baradaran, L. A. Sazanov, *Nature* **2010**, *465*, 441.
- [44] G. N. Murshudov, P. Skubák, A. A. Lebedev, N. S. Pannu, R. A. Steiner, R. A. Nicholls, M. D. Winn, F. Long, A. A. Vagin, *Acta Crystallogr.* **2011**, *67*, 355.

Received: 19 January 2013
Revised: 01 April 2013
Accepted: 21 April 2013
Published online on 24 May 2013