Backbone Building from Quadrilaterals: A Fast and Accurate Algorithm for Protein Backbone Reconstruction from Alpha Carbon Coordinates

DOMINIK GRONT, SEBASTIAN KMIECIK, ANDRZEJ KOLINSKI
Faculty of Chemistry, Warsaw University, Pasteura 1 02-093, Warsaw

Received 16 August 2006; Revised 3 October 2006; Accepted 18 October 2006
DOI 10.1002/jcc.20624
Published online 6 March 2007 in Wiley InterScience (www.interscience.wiley.com).

Abstract: In this contribution, we present an algorithm for protein backbone reconstruction that comprises very high computational efficiency with high accuracy. Reconstruction of the main chain atomic coordinates from the α carbon trace is a common task in protein modeling, including *de novo* structure prediction, comparative modeling, and processing experimental data. The method employed in this work follows the main idea of some earlier approaches to the problem. The details and careful design of the present approach are new and lead to the algorithm that outperforms all commonly used earlier applications. BBQ (Backbone Building from Quadrilaterals) program has been extensively tested both on native structures as well as on near-native decoy models and compared with the different available existing methods. Obtained results provide a comprehensive benchmark of existing tools and evaluate their applicability to a large scale modeling using a reduced representation of protein conformational space. The BBQ package is available for downloading from our website at http://biocomp.chem.uw.edu.pl/services/BBQ/. This webpage also provides a user manual that describes BBQ functions in detail.

 J Comput Chem 28: 1593–1597, 2007

Key words: Protein backbone reconstruction; protein reduced models; protein modeling; protein structure predictions

Introduction

The successes of genomes sequencing projects and progress in protein structure prediction methods have led us to the next demanding task of structural genomics to obtain three-dimensional structures of all proteins. Although experimental structure determination methods are providing high resolution structure information, because of their costly and time-consuming procedures, they cannot be utilized on a large scale of entire genomes. For a considerable fraction of sequences whose structures will not be determined experimentally, computational methods provide valuable information.

Many theoretical prediction methods, especially purely *de novo* folding computations, various comparative modeling techniques, or hybrid methods utilizing different kinds of sparse experimental data employ simplified protein representation. This is necessary to be able to explore the vast conformational space of protein chains. Such approaches, based on reduced protein representation,[1] appeared to be very efficient and placed among the most successful during the last round of CASP (critical assessment of protein structure prediction) community-wide experiment. Employing the coarse-grained representation brings necessity of final models reconstruction to the all atom representation compatible with the classical all-atom modeling tools.

In the past few years, many groups have developed algorithms to construct all the atomic coordinates of a protein backbone and the side chains from known CA coordinates.[2–12] Unfortunately, only a few of them are available as a stand-alone application or as Internet services. Most of these programs implement quite complicated algorithms and cannot withstand large scale modeling experiments. The aim of this work is to bring a new high-throughput method that will be able to process as many as thousands of models in a reasonable time. Moreover, the desired algorithm should be as accurate as possible.

As in most of the previously proposed methods, we assume that the problem of reconstruction of an all-atom chain from a CA trace can be separated into two subsequent steps: (i) reconstruction of the all-atom backbone, and (ii) reconstruction of the side-chain geometry for a given backbone. In this contribution, we deal only with the first task, postponing the second step to the further work, which is now in progress.

Correspondence to: D. Gront; e-mail: dgront@chem.uw.edu.pl

Contract/grant sponsor: Polish Ministry of Scientific Research and Information Technology; contract/grant number: PZB-KBN-088/P04/2003

Many of the approaches that have been proposed so far utilize protein fragment libraries derived from known structures to locate possible fragments that do not violate a specified CA trace. The most favorable fragments to construct the entire backbone are selected using energy-based, homology-based, or geometric criteria. In the MaxSprout method,[6] a series of 50 best-matching segments are generated for each residue junction and then a dynamic programming algorithm is used to select the most compatible pairs of overlapping segments. Another, very elaborated algorithm by Levitt,[5] begins with enumeration of 40 database segments, each 3–4 residues long, which have a good CRMSD fit to CA trace. Such segment sets are built in the neighborhood of every residue. Each segment has an effective energy that is defined as a weighted average of the CRMSD distance error and the nonbonded interaction energy between the segment and its environment. Segments are combined by means of Monte Carlo sampling. Averaging coordinates over the resulting low-energy ensemble generates an initial guess for the protein backbone. These coordinates are refined in a subsequent energy minimization step.

In contrast to the homology-based methods, approaches with no reference to structural databases perform *de novo* construction of the backbone and try to minimize its energy. Kazmierkiewicz et al.[7] derived formulas describing energy for dipole–dipole interaction. Optimal alignment of peptide-group dipoles is constructed by means of Monte Carlo search. Payne[8] derived a statistical potential to score local conformations of the backbone. Global optimum is computed with a dynamic programming approach.

BBQ program, introduced in this work, is very robust and extremely efficient. Besides its computational efficiency, it provides reasonable accuracy. The general idea we follow is not new. It was originally invoked by Purisima and Scheraga[11] and then used for development of a method of reconstruction of protein backbone by Milik et al.[10] It has also been employed to all-atom reconstruction from approximate positions of the side groups centers of mass.[9] A part of this program (reconstruction of the main chain and side chains from CA coordinates) has been implemented by Rotkiewicz in his Pulchra program. In this approach, a four residue fragment is described by three internal coordinates—distances between CA atoms. These three distances form a three-dimensional grid in which average positions of C, O, and N atoms measured in a local coordinate system are acquired from the known PDB structures.

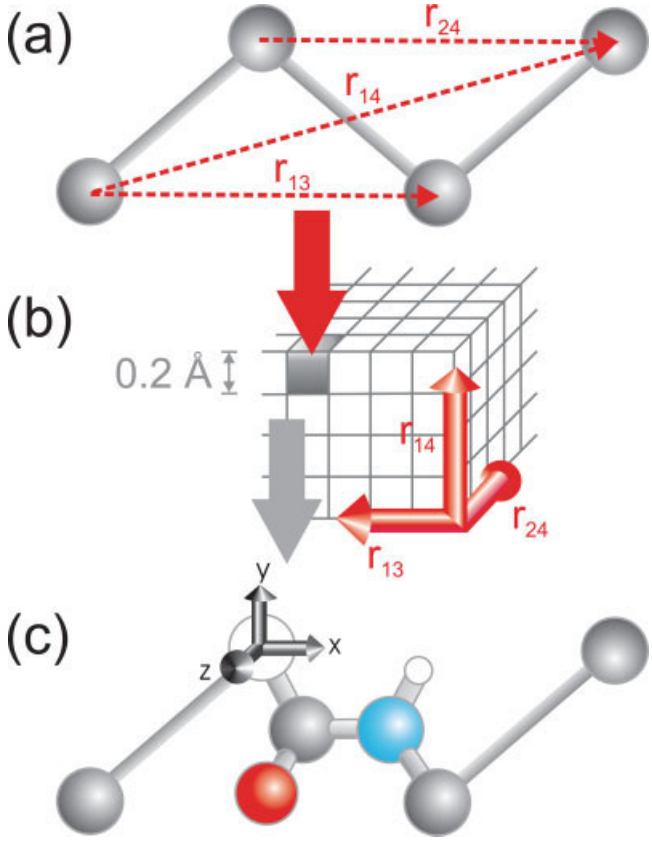

Figure 1. Protein backbone reconstruction flowchart: (a) Definition of the *R*-coordinates system. *R*-coordinates are computed as three distances marked by red-dashed line. They are used to describe a quadrilateral (black solid line and circles). (b) All quadrilaterals are stored in a look-up table. Each of the *R*-coordinates is dicretized with the mesh size of 0.2 Å. For each element of the grid average positions of N, C, and O, backbone atoms are computed. (c) In the reconstruction step, for a given quadrilateral composed of four CA atoms, local positions of the backbone atoms that belong to the central peptide plate are retrieved from a proper element of the grid, indexed by *R*-coordinates. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

Materials and Methods

To derive a statistics for positions of backbone atoms, we took a nonredundant protein database precomputed by the PISCES server.[13] The dataset contains 1259 protein chains with mutual pairwise sequence similarity not higher than 90%. Only high quality structures were included in the training set: X-ray resolution not worse than 1.6 Å and *R*-factor lower that 25.0. All these proteins were deposited to the database PDB[14] prior to April 2006. A different set of proteins has been used in evaluation of existing methods and the proposed here BBQ algorithm. We took all the protein structures deposited to PDB between April 1, 2006 and July 22, 2006 and discarded the redundant entries. Finally, the testing set contained 81 protein chains with mutual sequence similarity below 50%. The two sets did not contain any common or highly similar structures. Since the testing set contains only the newest structures, which are not included in the libraries of fragments employed by some methods, we eliminated any possible overfitting effects. All the calculations performed in this work, i.e., database extraction, structural calculations, and CRMSD evaluations were done using the recently developed BioShell[15] software package.

First of all, we define two coordination systems (Fig. 1). The first one (referred later as *R*-coordinates) is used to define a tetrapeptide, i.e., protein fragment of four amino acids. The second one (*L*-coordinates) defines a local Cartesian coordinate system centered on a given CA atom. We treat CA atoms as points and use them for all the geometric constructions. Each continuous fragment of four CAs is called a quadrilateral. Protein sequence is ignored in our algorithm.

In 3D space, four points have six internal degrees of freedom. Assuming that the distance between neighboring atoms (denoted as R_{12}) is constant we have only three free variables left. The

assumption is almost always true: the R_{12} distance is equal to 3.78 Å. The only exeption is cis conformation of proline, where $R_{12} = 3.2$ Å. For simplicity, we assumed for all the R_{12} distances a fixed value of 3.78 Å. Three other distances were chosen as local coordinates: R_{13} distance (between first and third CA atoms in a quadrilateral), R_{24}, and R_{14} (defined similarly to R_{13}). Unfortunately, the set of internal coordinates cannot distinguish between left-handed and right-handed conformations. Therefore, we define R_{24} coordinate as negative if the quadrilateral possesses the left-handed twist. Obviously, one can choose another set of three linearly independent variables to describe a quadrilateral, for instance, the two possible planar angles and a torsional angle. Our choice is motivated mostly by the computational efficiency requirements. For each quadrilateral, we also define a *L*-coordinates system as simple linear combinations:

$$\begin{aligned}\vec{v}_x &= (\vec{v}_{12} + \vec{v}_{23})/\,|\vec{v}_{12} + \vec{v}_{23}| \\ \vec{v}_y &= (\vec{v}_{12} - \vec{v}_{23})/\,|\vec{v}_{12} - \vec{v}_{23}| \\ \vec{v}_z &= \vec{v}_x \times \vec{v}_y \end{aligned} \quad (1)$$

where $\vec{v}_x$, $\vec{v}_y$, and $\vec{v}_z$ are the *L*-coordinates versors and $\vec{v}_{ij}$ denotes a versor pointing from *i*th to *j*th CA atom. The *L*-coordinates define the local positions of the backbone atoms.

In some preliminary tests, we had withdrawn those quadrilaterals for which *R*-factor was higher than the given threshold value. Interestingly, the final results, accuracy of backbone reconstruction, rather weakly depend on this parameter. Finally, we decided to keep only these quadrilaterals with *R*-factor below 50 (we used, for this purpose, the original values stored in PDB files) as a reasonable compromise between the statistics and the amount of noise introduced by the low-quality data. It is possible to create over 263,000 different quadrilaterals from our training data set.

For each of these quadrilaterals, we computed LCS coordinates of atoms that form central peptide plate (between the second and third CA atoms). We also computed *R*-coordinates. The *R*-coordinates were divided by 0.2 and rounded down to the nearest integer. This defined indices pointing to certain bins in a three-dimensional array. In this way, the continuous space described by *R*-coordinates has been discretized. We allow R_{24} and R_{13} distances vary from 4.0 Å to 7.6 Å, distance R_{14} in the range of 4.0–11 Å (or −11.0 Å to −4.0 Å for the left-handed conformations). The resulting three dimensional look-up table can hold 22,680 different quadrilaterals. In practice, the set of 263,000 quadrilaterals creates only 5148 discrete states. For every one of these states, we computed average positions for the N, C, and O atoms.

To reconstruct backbone atoms CA trace *R*-coordinates are calculated for subsequent quadrilaterals and a proper set of local coordinates for N, C, and O atoms are retrieved from the look-up table. In some rare cases, a specific combination of *R*-coordinates cannot be found in any protein observed from the training set. In such cases, program inspects the neighborhood of a given element of the grid, i.e., the 26 adjacent matrix elements. When all of them are empty, the program checks all quadrilaterals in the database and the entry, which minimizes the distance r^{QD} (see eq. (2)) between *R*-coordinates of the query (Q) and a element from training database (D):

$$r^{\mathrm{QD}} = \sqrt{\left(R^{\mathrm{Q}}_{13} - R^{\mathrm{D}}_{13}\right)^2 + \left(R^{\mathrm{Q}}_{24} - R^{\mathrm{D}}_{24}\right)^2 + \left(R^{\mathrm{Q}}_{14} - R^{\mathrm{D}}_{14}\right)^2} \quad (2)$$

Again, such situations, when a proper quadrilateral cannot be found are very rare. For example, in the test performed for 81 native proteins (18,651 total quadrilaterals), only in 0.35%, a grid neighbor was inserted. For 0.12% of all the cases, the all-grid neighbors were also empty and the best quadrilateral was found via the error minimization search.

Results

Although many methods have been proposed for backbone reconstruction, only a few are available as a stand-alone program or an Internet server. We compared our method with two currently available programs that are free for academic use (BB,[12] MaxSprout[6]), and Pulchra.[9] The version of Pulchra using CA traces as the input was kindly provided by Dr. Rotkiewicz. We also tested an algorithm by Claessens et al.[4] as implemented in SYBYL/Biopolymer commercial software from TRIPOS (St. Louis, MO). During the tests, we compared robustness, accuracy, and computational efficiency of these algorithms with those of BBQ.

Rebuilding Native Structures

In the first test, we attempted to rebuild backbone in carefully selected native structures. The results, summarized in the Table 1, show that the two methods employing protein fragments (Sybyl and MaxSprout) for some cases produced incomplete structures. They were unable to find a well-fitting fragment in their databases. Therefore, in columns 4, 5, and 6, we provide the results averaged

Table 1. Summary of the Results From the Reconstruction of Backbone in a Set of 81 Native Protein Structures.

		Results for 35 proteins rebuilt by all methods			
Method	Rebuilt structures (%)	Average CRMSD on backbone	Φ Correlation	Ψ Correlation	Average running time per protein (s)
MaxSprout	46.25	0.47	0.75	0.82	1.71
BB	100	0.64	0.52	0.65	56.98
Pulchra	100	0.59	0.65	0.78	1.06
Sybyl	91.25	0.39	0.77	0.86	172.6
BBQ	100	0.42	0.81	0.84	0.37

over 35 proteins for which all the four methods rebuilt successfully the entire backbones. For those methods that succeeded to rebuild all the proteins in our test set: BB, Pulchra and BBQ, an average CRMSD computed for 81 structures are very close to the corresponding values for 35 models (column 3): 0.65, 0.62, and 0.42 Å, respectively. According to Table 1, Sybyl and BBQ programs are the most accurate when the average CRMSD as well as the correlation coefficients between the experimental and predicted dihedral angles Φ and Ψ, are considered. However, Sybyl managed to construct only 91.25% of complete structures. Moreover, it is on average ~460 times slower than BBQ. Interestingly, BBQ algorithm is able to rebuild correctly backbone atoms in cis-proline. It is a result of a unique set of distances for proline in the quadrilateral data base.

Rebuilding Near-Native Decoys

The test described earlier provides a quantitative comparison of currently available methods for backbone reconstruction. Since native structures used in the test are known, various aspects of backbone modeling may be evaluated. However, from a practical point of view, this is not the most interesting situation. Only in very specific situations, the CA trace from a native structure needs the backbone reconstruction. The main goal for reconstructing algorithms is to provide a full-atom protein model starting from its reduced representation obtained in various modeling procedures. Therefore, we also assessed the quality of the reconstruction algorithms on a large set of near-native decoys.

Seven proteins from the testing set were selected for a decoy-building procedure: 2CJPA, 2CKLA, 2CL4X, 2GMKA, 2GR8A, 2GRRB, and 2GU3A. These structures are representative in respect to their chain length and secondary structure type. None of them has homologues structures in Sybyl or MaxSprout databases. We performed long Monte Carlo simulation with CABS model[16] to generate protein-like near-native decoys with CRMSD 0.35–3.0 Å from the native. For each protein ~800 decoys were randomly selected from a large set (60,000–150,000 structures, depending on a protein) using a uniform distribution of CRMSD from the native as a criterion of the selection.

Figure 2 illustrates the dependence between the CRMSD measured on the all backbone atoms after reconstruction and the CA CRMSD of the input decoys. The plot was prepared from over 5500 decoys reconstructed by the two best-performing methods: BBQ and Sybyl (Figs. 2a and 2b, respectively). The dependence is almost linear because during the reconstruction process the CA trace does not change significantly. Adding three additional atoms in the neighborhood of each CA has a very little effect on the CRMSD values measured on the whole structure. The most interesting part of the plot is in the range of low CRMSD values (see inserts in the diagrams). On average, CRMSDs for reconstructed backbones are somewhat worse than CRMSDs for the corresponding initial CA models. Nevertheless, the actual values depend on a method used for reconstruction. Sybyl implements a fragment-based approach. This creates an opportunity for the backbone structure improvement for a high accuracy models (better than 1.5 Å). This is because of the fact that the fragments of existing proteins are usually more accurate than the average geometry derived from the database. Obviously, only the local geometry of a backbone can be improved with fragment inserting procedures.

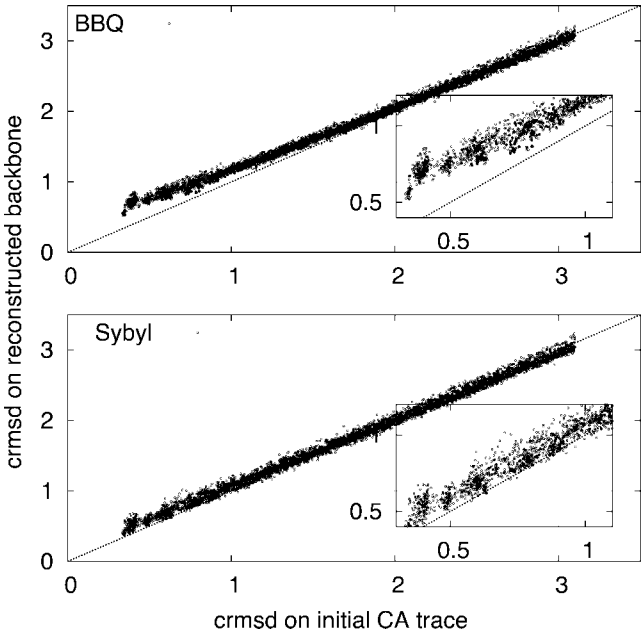

Figure 2. The dependence between CRMSD computed for CA atoms of initial model and CRMSD computed for all backbone atoms after reconstruction. Each circle represents a single decoy structure. The two plots in this figure: top and bottom correspond to the results from BBQ and Sybyl programs, respectively. In general, a linear dependence can be observed. The two methods slightly differ in their accuracy for high-resolution decoys (see the insets).

Detailed analysis of the results obtained in this work shows that the reconstruction accuracy depends on the secondary structure content. For all the methods, 2GRRB, all-α protein, was the easiest case among the seven proteins in our decoy set. Sybyl and Pulchra methods performed very well also on the 2GR8A set. This can be explained by the fact that almost 70% of 2GR8A residues buids a highly regular seven-stranded β-sheet. Most of the methods compared in this work, Sybyl, MaxSprout, BB, and Pulchra, alter somewhat positions of the CA atoms. The change of CRMSD measured on CA atoms of the decoys in respect to the native structure is order of 0.01 Å.

Summary

In this work, we present the results of comprehensive evaluation of performance of the BBQ, presented in this work (Tables 1 and 2), and four older methods for backbone reconstruction that are available as a stand-alone application. It can be clearly seen from our results that Claessens et al.[4] method employing longer fragments' insertions performs slightly better in the case of high accuracy models (resolution higher than 1.5 Å) than the methods that utilize short fragments (such as MaxSprout and BB) or averaged knowledge about backbone geometry (BBQ and Pulchra).

There is a wide range of applications of protein structure models, depending on their accuracy.[17] The accuracy of a comparative model is related to the percentage of sequence identity with the structural template (templates) on which it is based and it can be easily

Table 2. Summary of the Results From the Near-Native Decoys Reconstruction.

Method	Rebuilt decoys (%)	Average running time for 2CJPA decoy (320 residues) (s)
MaxSprout	70	2.6
BB	97	90
Pulchra	100	1.4
Sybyl	99	537
BBQ	100	0.48

estimated according to the prediction method. High accuracy models (better than 1.5 Å) can be expected from comparative modeling based on more than 30% sequence identity. Currently known protein structures allow to model about the half of all sequences deposited to SwissProt/TrEMBL database.[18] However, only 10% of the sequences are modeled on the basis of >30% sequence identity,[19] according to MODBASE.[20] This statistics shows that in a typical situation high-resolution models are not accessible because of the lack of suitable templates. In a daily practice, usually medium and low-resolution comparative models are created.

BBQ, when compared with Sybyl, is extremely fast. If we consider reconstruction for a set of 1000 protein models (2CJPA from the decoys testing set, which has 320 residues) on a standard PC workstation, it takes ∼8 min by BBQ and 149 h (almost a week) by Sybyl. It is also about three times faster than Pulchra. In summary, the proposed here BBQ algorithm is as accurate as the most accurate, but computationally very demanding methods. The exceptions are the very high resolution initial models rebuilt by Sybyl. This is however a very costly approach and of a limited practical applicability (as explained in the text earlier). The existing fast algorithms are significantly less accurate than the BBQ. Moreover, the BBQ performs rebuilding for entire structures, regardless the level of uniqueness of the encountered fragments. Therefore, BBQ seems be a method of choice for many typical procedures of protein modeling. Obviously, structures from BBQ (when it is needed) could be subject to a refinement procedure using more elaborate methods, what would lower significantly the computational cost of a high-resolution structure modeling.

Acknowledgments

The authors thank Dr. Piotr Rotkiewicz for providing Pulchra program. The computational part of this work was done using the computer cluster at the Computing Center of the Department of Chemistry, University of Warsaw.

References

1. Kolinski, A.; Bujnicki, J. M. Proteins 2005, 61, 84.
2. Jones, T. A.; Thirup, S. EMBO J 1986, 5, 819.
3. Blundell, T.; Carney, D.; Gardner, S.; Hayes, F.; Howlin, B.; Hubbard, T.; Overington, J.; Singh, D. A.; Sibanda, B. L.; Sutcliffe, M. Eur J Biochem 1988, 172, 513.
4. Claessens, M.; van Cutsem, E.; Lasters, I.; Wodak, S. Protein Eng 1989, 2, 335.
5. Levitt, M. J Mol Biol 1992, 226, 507.
6. Holm, L.; Sander, C. J Mol Biol 1991, 218, 183.
7. Kazmierkiewicz, R.; Liwo, A.; Scheraga, H. A. J Comput Chem 2002, 23, 715.
8. Payne, P. W. Protein Sci 1993, 2, 315.
9. Feig, M.; Rotkiewicz, P.; Kolinski, A.; Skolnick, J.; Brooks, C. L. Proteins: Struct Funct Genet 2000, 41, 86.
10. Milik, M.; Kolinski, A.; Skolnick, J. J Comput Chem 1996, 18, 80.
11. Purisima, E. O.; Scheraga, H. A. Biopolymers 1984, 23, 1207.
12. Adcock, S. A. J Comput Chem 2004, 25, 16.
13. Wang, G.; Dunbrack, R. L. Bioinformatics 2003, 19, 1589.
14. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E., Nucleic Acids Res 2000, 28, 235.
15. Gront, D.; Kolinski, A. Bioinformatics 2006, 22, 621.
16. Kolinski, A. Acta Biochim Polym 2004, 51, 349.
17. Baker, D.; Sali, A. Science 2001, 294, 93.
18. Boeckmann, B.; Bairoch, A.; Apweiler, R.; Blatter, M. C.; Estreicher, A.; Gasteiger, E.; Martin, M. J.; Michoud, K.; O'Donovan, C.; Phan, I.; Pilbout, S.; Schneider, M. Nucleic Acids Res 2003, 31, 365.
19. Chance, M. R.; Fiser, A.; Sali, A.; Pieper, U.; Eswar, N.; Xu, G.; Fajardo, J. E.; Radhakannan, T.; Marinkovic, N. Genome Res 2004, 14, 2145.
20. Pieper, U.; Eswar, N.; Braberg, H.; Madhusudhan, M. S.; Davis, F. P.; Stuart, A. C.; Mirkovic, N.; Rossi, A.; Marti-Renom, M. A.; Fiser, A.; Webb, B.; Greenblatt, D.; Huang, C. C.; Ferrin, T. E.; Sali, A. Nucleic Acids Res 2004, 32, D217.