Published online 19 April 2019 *Nucleic Acids Research, 2019, Vol. 47, Web Server issue* ***W451–W455***
doi: 10.1093/nar/gkz288

GalaxyRefine2: simultaneous refinement of inaccurate local regions and overall protein structure

Gyu Rie Lee†**, Jonghun Won**†**, Lim Heo and Chaok Seok***

Department of Chemistry, Seoul National University, Seoul 08826, Korea

Received January 31, 2019; Revised April 01, 2019; Editorial Decision April 08, 2019; Accepted April 11, 2019

ABSTRACT

The 3D structure of a protein can be predicted from its amino acid sequence with high accuracy for a large fraction of cases because of the availability of large quantities of experimental data and the advance of computational algorithms. Recently, deep learning methods exploiting the coevolution information obtained by comparing related protein sequences have been successfully used to generate highly accurate model structures even in the absence of template structure information. However, structures predicted based on either template structures or related sequences require further improvement in regions for which information is missing. Refining a predicted protein structure with insufficient information on certain regions is critical because these regions may be connected to functional specificity that is not conserved among related proteins. The GalaxyRefine2 web server, freely available via http://galaxy.seoklab.org/refine2, is an upgraded version of the GalaxyRefine protein structure refinement server and reflects recent developments successfully tested through CASP blind prediction experiments. This method adopts an iterative optimization approach involving various structure move sets to refine both local and global structures. The estimation of local error and hybridization of available homolog structures are also employed for effective conformation search.

INTRODUCTION

Template-based protein structure prediction methods can generate accurate protein models when sufficiently similar structural templates are available (1). Recently, template-free methods have also been able to produce highly accurate models (2,3) because of advances in coevolution analysis (4) and deep learning algorithms (5,6), which can extract protein structure information from protein sequences. However, the predicted protein structures may be inaccurate in regions for which there is not sufficient experimental data regarding template structures or related sequences available. Because model inaccuracy originates from a deficiency of information, researchers rely on physical principles to further refine structures. In recent CASP (Critical Assessment of techniques for protein Structure Prediction) blind prediction experiments, a model refinement category was introduced to evaluate existing model refinement methods and stimulate advances in the field (7).

The GalaxyRefine web server for protein structure refinement (8) was released on the GalaxyWEB server (9,10) in 2013. This server is based on a refinement method that performs short molecular dynamics (MD) relaxations after repeated side chain repacking perturbations. The GalaxyRefine server has been widely used in both experimental and computational studies. For instance, the server has been used by experimentalists in functional studies involving protein modelling to improve the quality of model structures obtained using other prediction methods (11–14). Additionally, developers of computational algorithms have combined the server algorithm with prediction methods employed in other research areas for improving prediction quality (15–17).

Here, we present GalaxyRefine2, an upgraded version of GalaxyRefine (8), which reflects a progress made during recent CASP experiments (18). In contrast to the previous version that focused on refinement in a local environment by a local move set, side chain repacking, GalaxyRefine2 introduces various local and global move sets and accumulates the conformational changes iteratively, enabling larger movements. The local and global move sets utilize estimated structure error to focus on refinement efforts in more inaccurate regions. Available structures of homologous proteins in the structure database are also used to enrich the possible structure pool (19). A benchmark test of CASP refinement targets showed that GalaxyRefine2 can improve model structures by 2-fold compared to GalaxyRefine in

*To whom correspondence should be addressed. Tel: +82 2 880 9197; Email: chaok@snu.ac.kr
Present addresses:
Gyu Rie Lee, Department of Biochemistry, University of Washington, Seattle, WA 98195, USA.
Lim Heo, Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI 48824, USA.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

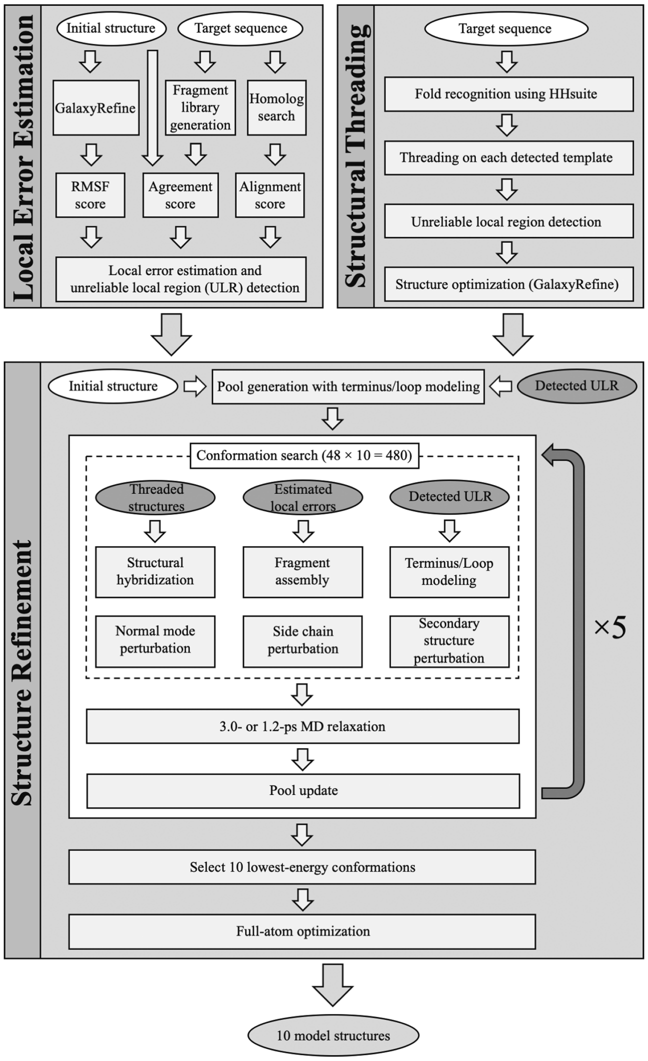


Figure 1. Flowchart of the GalaxyRefine2 protocol. The protocol consists of two pre-processing steps and the main refinement step.

terms of the accuracy measures GDT-HA (20) and LDDT (21).

THE GALAXYREFINE2 METHOD

Overall protocol

The computational protocol of GalaxyRefine2 is schematically shown in Figure 1. The protocol is a light version of the method presented in (19), with a smaller number of iterations. Details can be found in the reference, and the method is summarized below.

Pre-processing: local error estimation

Residue-wise error of the input structure is first estimated based on RMSF, FRAG and MSA scores. The RMSF score is residue-wise root-mean-square fluctuation in 24 runs of 14.4-ps MD relaxation involving side chain repacking every 1.2 ps (8). The FRAG score measures the agreement between backbone torsion angles of the input structure and those of fragments in the fragment library. The MSA score is the average of the position-specific scoring matrix components (22) from a multiple alignment of the sequences of homologs detected through HHsearch (23) against the input sequence. The alignment is generated by PROMALS3D (24). Residue-wise error is predicted using a linear model that combines these three scores. Finally, stretches of consecutive residues with high estimated errors are designated as unreliable local regions (ULRs).

Pre-processing: structure threading

Among the structures of detected homologs, only those with TM-score >0.5 (25) to the input structure are considered in this step. The input sequence is threaded onto each homolog structure based on the alignment of PROMALS3D (24). The threaded structure undergoes local error estimation and structure optimization by the GalaxyRefine algorithm (8). Local patches of the optimized structure, which are not part of the predicted ULRs are used by the 'structure hybridization' operator in the refinement step.

Structure refinement

An initial pool of 48 structures is generated from the input structure by re-building the termini and loops predicted as ULRs. At each iteration cycle, 480 trial structures are generated by applying structure operators 10 times to each pool structure. Structure operators include three operators that drive local refinement, 'fragment assembly', 'loop modelling' and 'side chain perturbation', and three that drive larger changes, 'normal mode perturbation', 'structural hybridization' and 'secondary structure perturbation'. 'Fragment assembly' re-builds regions with higher estimated errors through fragment assembly and triaxial loop closure (26,27). 'Loop modelling' either mixes the backbone torsion angles of a selected ULR with those of another pool structure or mutates them. 'Side chain perturbation', as in GalaxyRefine, repacks side chains (8). 'Normal mode perturbation' perturbs the structure toward one of low-frequency normal modes. 'Structure hybridization' hybridizes the structure with the structures threaded to homologs. 'Secondary structure perturbation' perturbs the orientations of the secondary structure chunks in a stochastic manner.

The 480 structures are then locally optimized by a 3.0-ps MD relaxation (1.2-ps after loop modelling). Each low-energy trial structure replaces a pool structure with higher energy that is structurally close enough. If a low-energy trial structure is not close to any pool structures, the highest energy pool structure is replaced. The criterion of closeness is gradually increased with iterations to facilitate broad sampling (19). After five iteration cycles, all 2400 generated structures are scored, and the 10 lowest-energy structures are selected. The 10 structures are subject to full-atom optimization to improve their stereochemical properties and are reported as the final refined structures.

Energy function

The energy function used for MD relaxation is a linear combination of physics-based energy, statistical potentials and

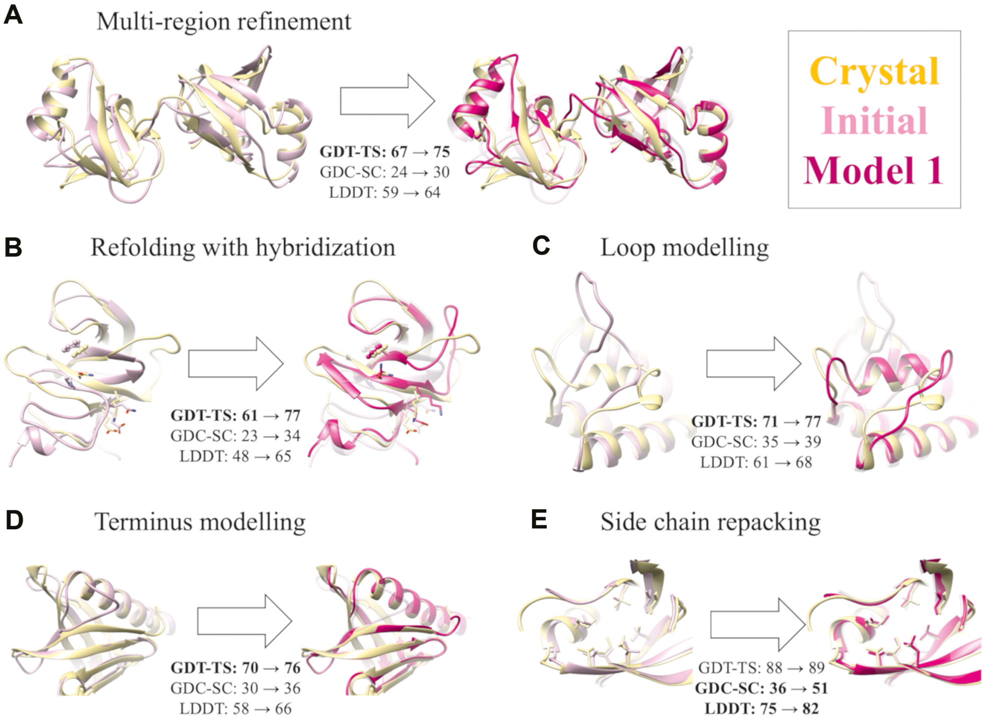


Figure 2. Successful refinement examples from CASP benchmark set. GalaxyRefine2 can improve structures at both the global level (**A**: TR462 and **B**: TR896) and local level (**C**: TR948, **D**: TR614 and **E**: TR488) by applying various structure operators simultaneously.

restraints, as described in (19). The restraints are derived from the input structure in terms of the Cartesian coordinates of alpha carbons and pair distances between alpha carbons or between backbone nitrogen and oxygen. To reduce the tendency of being restrained too strongly to the initial structure, higher 10% restraints are neglected during relaxation (28). The user can select the functional form of the restraints from either harmonic or Lorentzian. The default option of GalaxyRefine2 is Lorentzian, which allows for wider sampling than the harmonic form. The user may select the harmonic form by choosing the 'Conservative' option if the input structure is known to be reliable and only local refinement is desired. Energy without restraints is used in final scoring.

Method performance

The GalaxyRefine2 server was tested in the refinement category of CASP12 and CASP13 in a blinded manner, and the server named 'Seok-server' was ranked highly among all servers (18,29). The CASP13 results (http://www.predictioncenter.org/casp13/) are summarized in Table 1. GalaxyRefine2 was also compared to GalaxyRefine (8) on 114 refinement targets of CASP8-12 that are monomeric and have no missing residues in the middle. Homolog structures with a sequence identity of >40% were excluded during local error estimation and structure threading. As summarized in Table 2, the new version of GalaxyRefine, GalaxyRefine2, showed improved performance. The average magnitudes of improvement were 2-fold those of GalaxyRefine in terms of GDT-HA (20) and LDDT (21), when the server was run in the default mode. Several successful refinement examples improving models in various aspects are also illustrated in Figure 2.

GALAXYREFINE2 SERVER

Hardware and software

The server runs on a cluster of four Linux servers of 2.20-GHz Intel Xeon E5-2650 v4 12-core processors. The web application uses the Python programming language and the MySQL database. The GalaxyRefine2 pipeline is implemented using Python. The refinement method is implemented as part of the GALAXY program package (9,10) written in Fortran 90. JavaScript Protein Viewer (http://biasmv.github.io/pv/) is used to visualize the refined models.

Input and output

The required input is a protein monomer structure in PDB format. The number of residues in the input file is limited to 300 for computational efficiency. Structures with missing residues in the middle are not allowed. The user may choose to run the server in the conservative refinement mode when the backbone structure of the input is considered very reliable, such as in the cases of NMR structures or template-based models generated from templates with high sequence identities. The average run time is 6–10 h. Ten refined structures, ranked based on the energy, are visualized in the web browser and are downloadable in the PDB format. Information on structural changes between the input structure and refined structures is provided in terms of RMSD and MolProbity score (Figure 3).

Table 1. Performance comparison of server groups participated in CASP13 refinement category

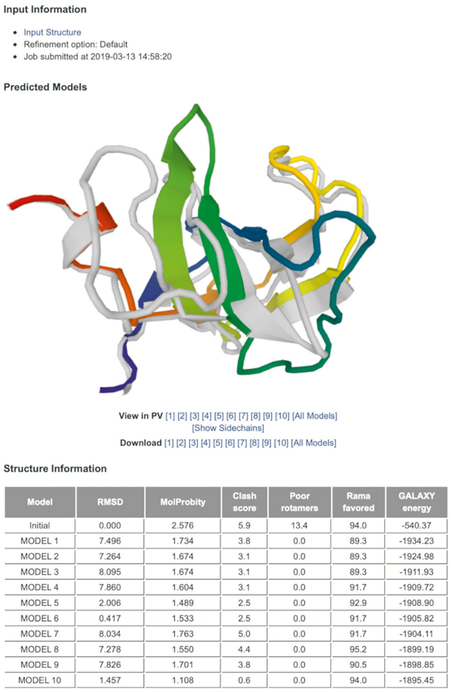
Group names	Mean improvement of Model 1 / Best among Model 1–5[a]			
	GDT-HA	GDC-SC	LDDT[b]	– MolProbity
Seok-server (GalaxyRefine2)	+1.46 / +2.68	+3.45 / +5.06	+2.55 / +3.21	+1.47 / +1.59
Bhattacharya-Server (29)	–0.37 / +1.75	+1.70 / +3.55	+0.86 / +1.79	+1.19 / +1.34
YASARA[c]	–1.21 / –1.21	+1.69 / +1.69	+0.57 / +0.57	+1.60 / +1.60
MUFold_server	–2.28 / –1.54	–0.69 / +1.17	–0.63 / –0.26	–0.40 / –0.17
3DCNN	–11.44 / –3.28	–6.52 / –1.05	–6.83 / –3.43	+0.65 / +0.87

[a] All evaluation values were obtained from CASP official homepage, http://predictioncenter.org/casp13/results.cgi.
[b] LDDT values were re-scaled to the range of [0, 100].
[c] YASARA group submitted only one model per target.

Table 2. GalaxyRefine2 benchmark test results for CASP8–12 refinement targets

Methods	Mean improvement of Model 1 / Best among Model 1–10		
	GDT-HA	GDC-SC	LDDT[a]
GalaxyRefine2 (default)	+0.92 / +2.72	+1.48 / +3.69	+1.58 / +2.43
GalaxyRefine2 (conservative)	+0.92 / +1.98	+0.98 / +2.11	+1.03 / +1.47
GalaxyRefine (8)	+0.58 / +1.41	+1.14 / +2.47	+0.73 / +1.16

[a] LDDT values were re-scaled to the range of [0, 100].



Input Information

- Input Structure
- Refinement option: Default
- Job submitted at 2019-03-13 14:58:20

Predicted Models

View in PV [1] [2] [3] [4] [5] [6] [7] [8] [9] [10] [All Models]
[Show Sidechains]
Download [1] [2] [3] [4] [5] [6] [7] [8] [9] [10] [All Models]

Structure Information

Model	RMSD	MolProbity	Clash score	Poor rotamers	Rama favored	GALAXY energy
Initial	0.000	2.576	5.9	13.4	94.0	-540.37
MODEL 1	7.496	1.734	3.8	0.0	89.3	-1934.23
MODEL 2	7.264	1.674	3.1	0.0	89.3	-1924.98
MODEL 3	8.095	1.674	3.1	0.0	89.3	-1911.93
MODEL 4	7.860	1.604	3.1	0.0	91.7	-1909.72
MODEL 5	2.006	1.489	2.5	0.0	92.9	-1908.90
MODEL 6	0.417	1.533	2.5	0.0	91.7	-1905.82
MODEL 7	8.034	1.763	5.0	0.0	91.7	-1904.11
MODEL 8	7.278	1.550	4.4	0.0	95.2	-1899.19
MODEL 9	7.826	1.701	3.8	0.0	90.5	-1898.85
MODEL 10	1.457	1.108	0.6	0.0	94.0	-1895.45

Figure 3. Example output page of GalaxyRefine2. Ten generated models are visualized using the JavaScript Protein Viewer. The models are downloadable in PDB format. Information such as structural changes from the input structure and MolProbity score is shown in the table.

CONCLUSIONS

GalaxyRefine2, an updated version of GalaxyRefine (8), refines local and global protein structures simultaneously by iterative conformational sampling, unlike GalaxyRefine, which was limited to local refinement. GalaxyRefine2 was successful in conducting blind prediction in CASP12 (18) and CASP13 refinement experiments. This server can, therefore, be used to improve predicted protein structures or low-resolution experimental structures for further interpretations or applications.

FUNDING

National Research Foundation of Korea (NRF), Korean government (MSIT) [2016R1A2A1A05005485, 2016M3C4A7952630]. Funding for open access charge: Seoul National University.

Conflict of interest statement. None declared.

REFERENCES

1. Marti-Renom,M.A., Stuart,A.C., Fiser,A., Sanchez,R., Melo,F. and Sali,A. (2000) Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.*, **29**, 291–325.
2. Ovchinnikov,S., Park,H., Varghese,N., Huang,P.S., Pavlopoulos,G.A., Kim,D.E., Kamisetty,H., Kyrpides,N.C. and Baker,D. (2017) Protein structure determination using metagenome sequence data. *Science*, **355**, 294–298.
3. Abriata,L.A., Tamo,G.E., Monastyrskyy,B., Kryshtafovych,A. and Dal Peraro,M. (2018) Assessment of hard target modeling in CASP12 reveals an emerging role of alignment-based contact prediction methods. *Proteins*, **86**(Suppl. 1), 97–112.
4. Kamisetty,H., Ovchinnikov,S. and Baker,D. (2013) Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 15674–15679.
5. Wang,S., Sun,S., Li,Z., Zhang,R. and Xu,J. (2017) Accurate de novo prediction of protein contact map by Ultra-Deep learning model. *PLoS Comput. Biol.*, **13**, e1005324.
6. Jones,D.T. and Kandathil,S.M. (2018) High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features. *Bioinformatics*, **34**, 3308–3315.
7. MacCallum,J.L., Hua,L., Schnieders,M.J., Pande,V.S., Jacobson,M.P. and Dill,K.A. (2009) Assessment of the protein-structure refinement category in CASP8. *Proteins*, **77**(Suppl. 9), 66–80.
8. Heo,L., Park,H. and Seok,C. (2013) GalaxyRefine: Protein structure refinement driven by side-chain repacking. *Nucleic Acids Res.*, **41**, W384–W388.

9. Ko,J., Park,H., Heo,L. and Seok,C. (2012) GalaxyWEB server for protein structure prediction and refinement. *Nucleic Acids Res.*, **40**, W294–W297.
10. Shin,W.-H., Lee,G.R., Heo,L., Lee,H. and Seok,C. (2014) Prediction of protein structure and interaction by GALAXY protein modeling programs. *Bio. Design*, **2**, 1–11.
11. Fenyk,S., Townsend,P.D., Dixon,C.H., Spies,G.B., de San Eustaquio Campillo,A., Slootweg,E.J., Westerhof,L.B., Gawehns,F.K., Knight,M.R., Sharples,G.J. *et al.* (2015) The potato Nucleotide-binding Leucine-rich Repeat (NLR) immune receptor Rx1 is a Pathogen-dependent DNA-deforming protein. *J. Biol. Chem.*, **290**, 24945–24960.
12. Ka,D., Lee,H., Jung,Y.D., Kim,K., Seok,C., Suh,N. and Bae,E. (2016) Crystal Structure of Streptococcus pyogenes Cas1 and Its Interaction with Csn2 in the Type II CRISPR-Cas System. *Structure*, **24**, 70–79.
13. Hajighahramani,N., Nezafat,N., Eslami,M., Negahdaripour,M., Rahmatabadi,S.S. and Ghasemi,Y. (2017) Immunoinformatics analysis and in silico designing of a novel multi-epitope peptide vaccine against Staphylococcus aureus. *Infect. Genet. Evol.*, **48**, 83–94.
14. Ali,M., Pandey,R.K., Khatoon,N., Narula,A., Mishra,A. and Prajapati,V.K. (2017) Exploring dengue genome to construct a multi-epitope based subunit vaccine by utilizing immunoinformatics approach to battle against dengue infection. *Sci. Rep.*, **7**, 9232.
15. Heo,L., Lee,H. and Seok,C. (2016) GalaxyRefineComplex: Refinement of protein-protein complex model structures driven by interface repacking. *Sci. Rep.*, **6**, 32153.
16. Baek,M., Park,T., Heo,L., Park,C. and Seok,C. (2017) GalaxyHomomer: a web server for protein homo-oligomer structure prediction from a monomer sequence or structure. *Nucleic Acids Res.*, **45**, W320–W324.
17. Keasar,C., McGuffin,L.J., Wallner,B., Chopra,G., Adhikari,B., Bhattacharya,D., Blake,L., Bortot,L.O., Cao,R., Dhanasekaran,B.K. *et al.* (2018) An analysis and evaluation of the WeFold collaborative for protein structure prediction and its pipelines in CASP11 and CASP12. *Sci. Rep.*, **8**, 9939.
18. Hovan,L., Oleinikovas,V., Yalinca,H., Kryshtafovych,A., Saladino,G. and Gervasio,F.L. (2018) Assessment of the model refinement category in CASP12. *Proteins*, **86**(Suppl. 1), 152–167.
19. Lee,G.R., Heo,L. and Seok,C. (2018) Simultaneous refinement of inaccurate local regions and overall structure in the CASP12 protein model refinement experiment. *Proteins*, **86**(Suppl. 1), 168–176.
20. Zemla,A., Venclovas,C., Moult,J. and Fidelis,K. (1999) Processing and analysis of CASP3 protein structure predictions. *Proteins*, **37**(Suppl. 3), 22–29.
21. Mariani,V., Biasini,M., Barbato,A. and Schwede,T. (2013) lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, **29**, 2722–2728.
22. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
23. Soding,J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics*, **21**, 951–960.
24. Pei,J., Kim,B.H. and Grishin,N.V. (2008) PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic Acids Res.*, **36**, 2295–2300.
25. Zhang,Y. and Skolnick,J. (2004) Scoring function for automated assessment of protein structure template quality. *Proteins*, **57**, 702–710.
26. Coutsias,E.A., Seok,C., Jacobson,M.P. and Dill,K.A. (2004) A kinematic view of loop closure. *J. Comput. Chem.*, **25**, 510–528.
27. Lee,J., Lee,D., Park,H., Coutsias,E.A. and Seok,C. (2010) Protein loop modeling by using fragment assembly and analytical loop closure. *Proteins*, **78**, 3428–3436.
28. MacCallum,J.L., Perez,A. and Dill,K.A. (2015) Determining protein structures by combining semireliable data with atomistic physical models by Bayesian inference. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, 6985–6990.
29. Bhattacharya,D. (2019) refineD: Improved protein structure refinement using machine learning based restrained relaxation. *Bioinformatics*, doi:10.1093/bioinformatics/btz101.