Computational and Structural Biotechnology Journal 18 (2020) 162–176

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Mini Review

Computational reconstruction of atomistic protein structures from coarse-grained models

Aleksandra E. Badaczewska-Dawid, Andrzej Kolinski, Sebastian Kmiecik *

Faculty of Chemistry, Biological and Chemical Research Center, University of Warsaw, Pasteura 1, 02-093 Warsaw, Poland

ARTICLE INFO

Article history:
Received 10 December 2019
Accepted 10 December 2019
Available online 26 December 2019

Keywords:
Protein reconstruction
Structure prediction
Coarse-grained modeling
Structure refinement
Protein modeling

ABSTRACT

Three-dimensional protein structures, whether determined experimentally or theoretically, are often too low resolution. In this mini-review, we outline the computational methods for protein structure reconstruction from incomplete coarse-grained to all atomistic models. Typical reconstruction schemes can be divided into four major steps. Usually, the first step is reconstruction of the protein backbone chain starting from the C-alpha trace. This is followed by side-chains rebuilding based on protein backbone geometry. Subsequently, hydrogen atoms can be reconstructed. Finally, the resulting all-atom models may require structure optimization. Many methods are available to perform each of these tasks. We discuss the available tools and their potential applications in integrative modeling pipelines that can transfer coarse-grained information from computational predictions, or experiment, to all atomistic structures.

Contents

1. Introduction

Coarse-grained protein models (with some missing atomic details) are the outcome of many experimental or computational methods for the investigation of protein structures and their dynamics. For example, structures obtained via difficult comparative modeling and de novo simulation strategies often need further improvement. The complexity of the protein systems demands a multiscale approach, which requires easy and fast conversion between models of various resolutions and accurate reconstruction of atomic details. Coarse-grained modeling tools offer high efficiency and enable to overcome the limitations of all-atom tools on accessible system sizes and simulation time scales [1]. All-atom

* Corresponding author.
E-mail address: sekmi@chem.uw.edu.pl (S. Kmiecik).

https://doi.org/10.1016/j.csbj.2019.12.007

Table 1
Overview of protein reconstruction methods. The accuracy of some methods is evaluated using RMSD values between reconstructed and reference structures measured on: alpha carbons ($RMSD_{CA}$) or backbone ($RMSD_{BB}$) or side chain ($RMSD_{SC}$) heavy atoms. The accuracy of side chain reconstruction is also evaluated using chi angles, the first ($\chi 1$) and the second ($\chi 2$, if applicable).

Method, reference and year of the last publication	Software availability*	Reconstruction** task	Description***	Benchmark sets and comments***
Reconstruction from deeply coarse-grained representation or contact maps				
CONFOLD [31], 2015 **CONFOLD2** [32], 2018	server (confold) + standalone (confold2): http://protein.rnet.missouri.edu/confold/ https://github.com/multicom-toolbox/CONFOLD2/	CM → CA	The method translates contact maps into distance restraints and uses them as the input to distance geometry algorithm which builds tertiary structure models. CONFOLD2 predicts 200 models using various subsets of input contacts and selects five top models by clustering them.	CONFOLD2 is an improved version of CONFOLD method. Structure predictions for 150 proteins from the PSICOV dataset and for CASP12 targets showed that the for most protein sequences CONFOLD2 was able to capture the structural fold of the protein.
FT-COMAR [30], 2008	standalone http://bioinformatics.cs.unibo.it/FT-COMAR/	CM → CA	A heuristic procedure for building tertiary structure models **from a possibly erroneous and incomplete contact maps**.	Tested on 100 non-redundant single-domain protein chains (α, β, α+β, α/β; size from 55 to 786 residues) from SCOPE release 1.67. FT-COMAR is much more tolerant to under prediction than to over prediction of contacts. It can ignore up to 75% of the contact map and still compute a protein structure whose $RMSD_{CA}$ < 4 Å (assuming that the remaining 25% contains no errors).
GDFuzz3D [33], 2015	server + standalone: http://iimcb.genesilico.pl/gdserver/GDFuzz3D/	CM → AA + optimization	The method transforms contact maps into distance restrains and uses them as the input to MODELLER method [44], which generates protein models and REFINER method [138] for structure refinement.	Tested on 45 single-domain targets analyzed in the CASP10 experiment and 150 proteins of the PSICOV dataset. The tests showed that GDFuzz3D is slightly more accurate (based on TM-score and RMSD) than FT-COMAR and slightly inferior to PconsFold but more computationally efficient.
PconsFold [34], 2014	standalone: https://github.com/ElofssonLab/pcons-fold	CM → AA	Merges PconsC contact prediction tool [139] and the ROSETTA protein modeling tool [140]. The method has no intermediate stages of reconstruction.	Tested on 150 proteins (from 52 to 266 residues) of the PSICOV dataset. The input sequence can come from a PDB header (instead of an ATOM section) to avoid internal gaps of chain. This approach enables protein structure prediction of single-domain targets. PconsFold performance was also compared to that of GDFuzz3D [33].
SICHO [36], 2000	standalone: http://blue11.bch.msu.edu/mmtsb/rebuild.pl	SICHO → AA	Method for **reconstruction from the SICHO coarse-grained model** (see Section 2.2 and Fig. 2). Uses a library of fragments and a side chain center-based coordinate system to rebuild Cα positions and a complete backbone. Chooses side chain conformations from a rotamer library.	Tested on 13 high-resolution X-ray structures. Reconstruction quality $RMSD_{CA}$: < 0.6 Å on experimental structures.
SUReLib, 2019	standalone: http://biocomp.chem.uw.edu.pl/tools/surpass	SURPASS → CA	Method for **reconstruction from the SURPASS coarse-grained model** (see Section 2.2 and Fig. 2). Uses a knowledge-based library of 6-residue fragments and structural regularities observed in known protein structures.	Tested on PISCES_4600, BAKER_62 and other various proteins (α, β, α+β, α/β; size from 56 to 1016 residues). Reconstruction quality $RMSD_{CA}$: < 0.5 Å on experimental structures and 1–2 Å on distorted models.
Backbone reconstruction from CA-trace				
BBQ [38], 2007	standalone: http://biocomp.chem.uw.edu.pl/tools/bbq	CA → BB	Uses the library of 5148 backbone 4-residue fragments (quadrilaterals) and algorithm described by Milik et al. [141] with some modifications. All quadrilaterals are pre-computed (as Cα distances and a local coordinate system) and stored in a table. The algorithm is sequence-independent.	Tested on 81 non-redundant experimental protein structures and near-native decoys. Reconstruction quality $RMSD_{BB}$ < 0.7 Å on experimental structures. Available as part of the Bioshell package. The algorithm is implemented in java programming language. BBQ performance was also compared to that of PD2 and other tools [39] and can be improved by additional minimization [52].
BriX [42,142], 2010	standalone	CA → BB	Uses the library of high-resolution structural fragments between 4 and 14 residue long and local fit approximation algorithm. Newer version [142] uses additional Loop BriX **database of non-regular structure elements** (loops) and fragments from over 7000 non-homologous proteins from the Astral set. User provided structures can be covered on the fly with BriX fragments, **especially gaps or low-confidence regions** in these structures can be bridged.	Tested on all known human structures from the PDB (935, Park & Levitt protein set), with a global 0.48 Å RMSD [42] (improving existing results using smaller libraries [48,143]) and over 300 protein-peptide complexes from PepX database within 1 Å RMSD [144]. Irregular loop regions can be reconstructed from smaller (4–8 residues long) building blocks.

(*continued on next page*)

Table 1 (*continued*)

Method, reference and year of the last publication	Software availability*	Reconstruction** task	Description***	Benchmark sets and comments***
PD2 ca2main [39], 2013	server + standalone: http://www.sbg.bio.ic.ac.uk/~phyre2/PD2_ca2main/	CA → BB optimization	Uses the library of short 528 backbone fragments obtained using Gaussian mixture models (GMMs). The accuracy of reconstruction can be improved by additional (optional) energy gradient minimization.	Tested on 15 low-resolution and 28 high-resolution protein structures. Reconstruction quality $RMSD_{BB}$ < 0.4 Å on experimental structures. When combined with Rosetta, PD2 method **produced significantly lower energy all-atom models than other tested tools**. Except built-in minimization, another minimization scheme had been successfully tested [52]. The algorithm is implemented in C++ programming language.
SABBAC [40], 2006	server: http://bioserv.rpbs.jussieu.fr/SABBAC.html	CA → BB	Uses a 27-letter hidden Markov model-derived structural alphabet described by 155 backbone fragments from known protein structures and a greedy algorithm (based on the OPEP force field) to obtain an optimal combination of fragments. **Cα-trace coordinates remain unaffected** and only the missing backbone atoms are added. No further refinement is performed.	Tested on the Adcock subset of 14 proteins from 58 to 437 residues and a 7 PDB newcomers subset up to 666 residues. Reconstruction quality $RMSD_{BB}$ is near 0.4 Å for experimental structures. The **algorithm is robust to CA deviations** (for Cα-traces randomly perturbed by over 1 Å, SABBAC results were only marginally affected). SABBAC enables reconstructing single polypeptide chains.
Side chains reconstruction from backbone				
CIS-RR [67], 2011	server	BB → SC	Uses Dunbrack backbone-dependent rotamer library, SCWRL3-based scoring function and clash-reduction guided iterative search (CIS) with conjugate gradients optimization of rotamers (rotamer relaxation, RR). CIS-RR **detects the cysteine pair, which forms a disulfide bond**.	Tested on 180 proteins (SCWRL3 test set) and 65 high-resolution crystal structures of proteins. Compared to other tools (SCWRL4, IRECS and SCAP) reconstruction accuracy is similar but **removes atomic clashes much more effectively.** Also evaluated and compared with other tools in work [95].
IRECS [76], 2007	standalone: https://irecs.bioinf.mpi-inf.mpg.de/index.php	BB → SC	Uses a coarse-grained backbone-dependent rotamer library, heuristic greedy iteration scheme and effective score (based on knowledge-based scoring term ROTA 10 Å) for ranking all SC rotamers according to the probability of rotamer conformation.	Tested on 641 high resolution X-ray structures (194 with single conformation for all SCs and 447 with at least one SC of multiple conformations). Reconstruction accuracy similar to SCWRL3 and SCAP, $RMSD_{SC}$ ~1.5 Å. Allows the use of additional template of side-chain conformations.
NCN [92], 2004	standalone:available on request from the authors https://www.med.upenn.edu/wandlab/research.html	BB → SC	Uses optimized OPLS parameters for long-range and multi-body terms (van der Waals and electrostatic terms), hydrogen-bonding potential and frequency of rotameric states from PDB. The library contains 49,042 discrete rotamers.	Tested on 65 high resolution X-ray structures. Highly accurate tool for SC reconstruction ($RMSD_{SC}$: ~1 Å).
OPUS_Rota2 [81], 2019 **OPUS_Rota** [96], 2008	standalone: http://ma-lab.rice.edu/soft.php	BB → SC	Uses rotamer frequency and van der Waals potentials and two additional unique pairwise energy terms: short-range orientation-dependent (OPUS-PSP) for side chain packing interactions and explicit solvation effects. In newer OPUS_Rota2 version, OPUS-PSP had been replaced by OPUS-DASF term that describes relative positions of atoms on the side chains.	Tested on 65 high resolution X-ray structures and a 379-protein PISCES subset (sequence identity 30%, 1.8 Å) [77,81]. In the native tests sets, Opus_Rota2 was more accurate than other methods (OpusRota, SCWRL4, OSCAR-star variants) but slightly less accurate than OSCAR-o. In non-native test sets (with added random noise to the main-chain torsional angles) Opus_Rota2 was more accurate than any other tested method and also **several times faster** (except Upside).
OSCAR [97], 2011	standalone: https://sysimm.ifrec.osaka-u.ac.jp/OSCAR/	BB → SC	Uses a flexible (-o, slow modeling) or rigid (-star, fast modeling) rotamer model. The energy terms include distance and orientation-dependent potentials and side chain dihedral angle potential energy function. The library of sub-rotamers was derived by perturbation of dihedral angles of rotamers from Dunbrack and Cohen [145].	Tested on 218 proteins and a RAPPER decoy set. Oscar had similar accuracy in SC reconstruction for experimental structures as other available software and good accuracy in **selecting near-native conformations from loop decoys**. Also evaluated and compared with other tools in work [73].
PEARS [82], 2018	server: http://opig.stats.ox.ac.uk/webapps/newsabdab/sabpred/pears	BB → SC	Uses position-dependent **antibody-specific rotamer library** which is based on SC's χ_1 dependency on its immunogenetic positions. The method is **robust for uncertainties in the model backbone and detects disulphide bridges**. SC clashes are reduced during 200 rounds of Gaussian relaxation.	Tested on a set of 639 non-redundant and a blind set of 95 **antibody structures**. The approach is comparable to SIDEpro, RASP and SCWRL in reconstruction the side chains of crystal structures, while on computationally designed models PEARS achieves the highest average accuracy and the **smallest number of clashes**.

Table 1 (*continued*)

Method, reference and year of the last publication	Software availability*	Reconstruction** task	Description***	Benchmark sets and comments***
RASP [83], 2011	standalone	BB → SC rotamer optimization	Uses backbone-dependent rotamer library, an optimized energy terms and the clash elimination strategy to guide the optimization of side chain conformations. Combinatorial search includes dead-end elimination, graph theory-based, branch-and-terminate, backtrack and Monte Carlo algorithms.	Tested on 2412 high-resolution (≤1.8 Å) structures with complete side chains obtained from PISCES server. RASP had comparable prediction accuracy (%chi1, % chi1+2, RMSD) and returned **much fewer clashes** than SCWRL4, OPUS-Rota or IRECS. It was also much faster than these methods, but an order of magnitude slower than Upside. RASP performance was also evaluated and compared with other tools in works [73,95].
SCAP [98], 2001	standalone: http://honig.c2b2.columbia.edu/jackal	BB → SC	Heuristic approach using optimized CHARMM parameters for van der Waals torsion-angle terms in an iterative repacking protocol. The library contains 7562 discrete rotamers in terms of 1) Cartesian coordinates, 2) dihedral angles..	Tested on 33 high resolution protein structures (66–328 residues) not included in the creation of rotamer library. For multi-chain proteins, only the first chain was used. Reconstruction quality $RMSD_{SC}$ < 2 Å.
SCATD (ThreePack) [79], 2005	standalone: https://ttic.uchicago.edu/~jinbo/TreePack.htm	BB → SC rotamer optimization	Uses a backbone-dependent rotamer library (the same as SCWRL3), interaction scores by dead end elimination and energy minimization by tree decomposition. This tool does not attempt to regularize the backbone geometry or solve punched rings.	Tested on 180 experimental structures from the SCWRL3 benchmark set of proteins. This approach was several times faster than SCWRL3 especially on larger proteins or cases with heavy atomic clashes. SCATD is freely available and was only tested on a Debian Linux machine.
SCWRL4 [77], 2009	standalone: http://dunbrack.fccc.edu/scwrl4/	BB → AA rotamer optimization	Uses a backbone-dependent rotamer library based on kernel density estimates to provide rotamer frequencies and torsional angles, a tree decomposition algorithm to solve the side chain packing problem, specific potentials (anisotropic hydrogen-bonding, soft pairwise van der Waals), and fast collision detection. **Allows consideration of the crystal symmetry in the side-chain conformation prediction**. SCWRL4 is perhaps the most widely used SC reconstruction method, as shown by its high citation count.	Optimized on a set of 100 protein structures and tested on 379 X-ray structures with electron densities available from UEDS [146]. SCWRL4 performance was evaluated and compared with other tools in works [73,95]. SCWRL4 is also available as a **dynamic-linked library for incorporation into other software**. In comparison to its earlier version SCWRL3, SCWRL4 can be slower but converged in all cases tested, while SCWRL3 sometimes did not converge [77]. The software is freely available for academic research on request.
SIDEpro [99], 2012	server + standalone: http://sidepro.proteomics.ics.uci.edu/ http://scratch.proteomics.ics.uci.edu/	BB → SC rotamer optimization	Uses a machine learning approach based on 156 neural networks that are trained to compute an energy function based on pairwise contact distances and a backbone-dependent rotamer library (the same as OPUS-Rota [96]). The neural networks set the side-chains to the highest probability rotamers. The final optimizing procedure removes steric clashes.	Tested on the SCWRL4 benchmark set (379 proteins), 94 proteins from CASP9, 7 large protein complexes and a ribosome with and without RNA. SIDEpro can **use non-standard amino acids, post-translational modifications and external ligands**. It was several times faster and slightly better in accuracy than SCWRL4 and its $RMSD_{SC}$ remained ~1.0 Å also for complexes. SIDEpro performance was also evaluated and compared with other tools in work [95].
Upside [100], 2018	standalone: https://github.com/sosnicklab/upside-md	BB → SC rotamer optimization	Uses side chain free energy in a molecular dynamics simulations scheme. During the optimization of side chain packing, each rotamer state is represented by a single oriented CG bead (3 spatial and 2 orientation coordinates). Uses a combination of isotropic (excluded volume) and directional interactions (chemical character, e.g. polar, aromatic) for each pair of interacting side chains or backbones. The side chain model is trained by the maximum-likelihood scheme. The NDRD rotamer library [70] is used to define the atomic positions of side chains.	Tested on a large, non-redundant set of crystal structures of globular proteins from the PDB with 50–500 residues and resolution < 2.2 Å (6255 chains). The method gave similar accuracy of chi1 angle as SCWRL4 and RASP, but is **several (1–3) orders of magnitude faster**.
All-atom reconstruction from CA-trace				
ca_to_allatom (ROSETTA) [43], 2008	standalone: https://www.rosettacommons.org/	CA → AA AA optimization	The Rosetta protocol ca_to_allatom reconstructs AA structure and performs structure refinement. Uses the initial Cα-trace (with a user-defined parameter specifying how far Cα atoms are allowed to deviate from the initial model)) and rigid-body perturbation of secondary structure	Tested on 8 proteins (from 101 to 310 residues) from cryoEM maps at 5 and 10 Å resolution. **Original Cα positions are slightly changed** during the reconstruction process by harmonic oscillation [147]. **Successfully used in protein reconstruction from experimental data** [43]. Avail-

(*continued on next page*)

Table 1 (*continued*)

Method, reference and year of the last publication	Software availability*	Reconstruction** task	Description***	Benchmark sets and comments***
			fragments from known protein structures. The protocol includes **optional loop remodeling** (centroid mode) and all torsion angle minimization (all-atom).	able as an executable in the bin directory of ROSETTA package (bin/ca_to_allatom.*version*).
CG2AA [37], 2016	standalone	CA → AA SC optimization	Uses a strictly geometric approach based on Cα triplets and parameters from the Amber03 force field for rebuilding the protein backbone and Cβ. The side chain is rebuilt based on the definition of the united atom for the side group.	Tested on 5 experimental protein structures with reconstruction quality $RMSD_{BB}$: <0.8 Å, **stability of reconstructed models has been tested in MD simulations**. The algorithm is **implemented in Python**.
Modeller [44], 2016	standalone: https://salilab.org/modeller Modeller-based reconstruction script: https://bitbucket.org/lcbio/ca2all	CA → AA AA optimization	Uses protein template(s) in CG representation (it can be in Cα-trace) to create a set of distance restraints that guide the reconstruction. Stereochemical restraints (bond lengths and angles) are obtained from the CHARMM force field and statistical analysis of known structures. MODELLER employs various structure optimization techniques.	Available as part of the Modeller package. The algorithm is **implemented in Python**. Modeller-based script *ca2all* [148] is used by CABS-flex and CABS-dock multiscale modeling tools [149,150] for reconstruction of protein or protein-peptide models).
ModRefiner [53], 2011	server + standalone: https://zhanglab.ccmb.med.umich.edu/ModRefiner/	CA → AA AA optimization	Reconstructs and refines protein structures, first the BB only and, after adding SC, the entire structure. Both side-chain and backbone atoms are flexible during refinement simulations, while conformational search is driven by physics- and knowledge-based force-field. It can optionally use secondary structure assignment/prediction to drive the refinement. The method can start from the CA, BB or SC model.	Tested on 261 proteins up to 150 residues (148 hard targets for I-TASSER and 113 with good templates). Compared to other tools, ModRefiner was better in side chain packing and **improving hydrogen-bonding networks**. **Input CA coordinates can have unphysical distortions.** A standalone tool enables reconstruction of **dimeric proteins**, while server handles only single-chain proteins.
PULCHRA [54], 2008	standalone: http://cssb.biology.gatech.edu/skolnick/files/PULCHRA/index.html	CA → AA AA optimization	Uses backbone fragment library, rotamer library and backbone reconstruction algorithm described by Milik et al. [141] with some modifications. The initial Cα-trace and reconstructed backbone are minimized to **improve hydrogen-bonding networks**. Positions of SC united atoms (center of mass) can be used to improve the accuracy of full-atomic reconstruction.	Tested on 30 high-quality X-ray structures. (reconstruction quality $RMSD_{AA}$ 1.0–1.5 Å) and on a set of 500 low-resolution protein models. Initial Cα coordinates can be distorted. This approach enables **reconstruction of multi-chain models or a chain with breaks and solves punched rings**. The algorithm is implemented in C programming language.
RACOGS [55], 2007	server available on request from the authors http://www.kavrakilab.org/software.html	CA → AA AA optimization	Uses a geometric approach to place the backbone atoms at the average positions derived from known protein structures (based on the algorithm by Milik et al. [141] and Feig et al. [36]) with SC reconstruction using backbone dependent, coordinate rotamer libraries (algorithm described by Xiang and Honig [98]). The final stage of the procedure includes the addition of all hydrogen atoms and short all-atom minimization.	Tested on CG trajectories of SH3, S6 systems and a subset of 2945 non-redundant experimental structures from PDB. This approach enables reconstruction of all-atom details from large regions of **the protein folding landscape as folded, partially** folded or **random protein structures**.
REMO [41], 2009	server + standalone: https://zhanglab.ccmb.med.umich.edu/REMO/	CA → AA AA optimization	Uses backbone isomer libarary (528,798 fragments) and backbone-dependent rotamer library (SCWRL) for atomic details reconstruction. Backbone rebuilding stage includes **removing steric clashes and optimizing the hydrogen-bonding network** based on a consensus of PSIPRED preferred secondary structure distribution.	Tested on 230 non-redundant proteins up to 300 residues (experimental and CG decoys generated by I-TASSER in the CASP8). This approach can remove steric clashes, retain correct topology and improve the backbone hydrogen-bonding network.
Hydrogen atom reconstruction				
CHIMERA (AddH) [104], 2004	standalone: http://www.cgl.ucsf.edu/chimera/index.html	SC → AA	Adds missing hydrogen and OXT atoms. Uses the atom types and steric-only or H-bonds (default, slower) criterion to determine the number and positions of added hydrogens. Bond lengths are taken from Amber parm99 parameters.	**The positions of pre-existing atoms are not changed.** Protonation states of certain ionizable side chains **can be specified at specific pH** (default: physiological). This software is also a molecular visualization tool.
CNS [113], 1998	standalone: http://cns-online.org/v1.3/	SC → AA	The algorithm starts from random positions of hydrogen atoms and optimizes them using an iterative procedure of molecular dynamics simulations and Powell energy minimization steps. The energy function includes bonded terms and van der Waals.	The method is able to **compute also electrostatic interactions** if the required parameters are provided. It is flexible hierarchical software for macromolecular structure determination, especially crystallographic refinement or NMR structure calculations using NOEs, J-coupling or chemical shifts.

Table 1 (*continued*)

Method, reference and year of the last publication	Software availability*	Reconstruction** task	Description***	Benchmark sets and comments***
Computational Titration [106], 2009	server	SC → AA AA optimization	Uses a force field with the concept of **hydropathic interactions** (HINT) as its noncovalent force field and exhaustive enumeration for optimization. The method uses coordinate data for the protein, ligand and bridging water molecules (if available) and predicts the best combination of protonation states for each ionizable residue and/or ligand functional group as well as the Gibbs free energy of binding for the ionization-optimized protein-ligand complex.	Tested successfully in modeling binding affinities of protein-ligand complexes: β secretase (2va7), mutant HIV-1 reverse transcriptase (2opq) and human sialidase NEU2 complexed with an isobutyl ether mimetic inhibitor (2f11). The method **improves optimization of protonated amines and phosphines** and supports the use of **additional functional groups such as phosphates, sulfates, nucleotide backbone phosphates and sugars**.
GROMACS (pdb2gmx) [114], 2001	standalone: http://www.gromacs.org/Downloads	SC → AA AA optimization	Uses a geometry-based approach and performs molecular dynamics simulations. Uses different bond lengths and angles according to selected force field parameters. The energy function includes bonded terms, van der Waals and electrostatics.	The method enables **optimization of histidine protonation states** by attempting to satisfy neighboring hydrogen bonds. **Water hydrogen atoms are also predicted.** GROMACS is very fast at calculating non-bonded interactions.
HAAD [107], 2009	server + standalone: https://zhanglab.ccmb.med.umich.edu/HAAD/	SC → AA AA optimization	Combines local geometry restraints and conformational search that minimizes atomic overlap, encourage hydrogen bonding and optimize electrostatic interactions. Local geometries of the initial positions of H-atoms are taken from the CHARMM22 force field.	Tested on three sets of experimental data: high-resolution X-ray crystallography, structures from neutron diffraction, and NOE proton-proton distance restraints. Compared with other methods (CHARMM and REDUCE) HAAD was **faster** and had significantly **higher accuracy** and **better compatibility with NOE restraints.** The algorithm is implemented in FORTRAN90 programming language.
Hbuild (CHARMM) (X-PLOR) [115,117,151], 2005	standalone: http://charmm.chemistry.harvard.edu/ https://nmr.cit.nih.gov/xplor-nih/doc/current/xplor/	SC → AA	Searches hydrogen atom positions at intervals of 10° (ϕ = 10) or 3° (ϕ = 3) around the axis of a cone with a side equal to the bond length or places hydrogens using geometric criteria. Uses different bond lengths and angles according to the selected version of the CHARMM force field. The energy function includes torsion angle, van der Waals and electrostatics.	All hydrogen atoms (including non-polar) are described explicitly. **Water hydrogen atoms are also predicted.** This approach is quite fast and can be used before running molecular dynamics calculations or during large-scale homology modeling. The Hbuild algorithm is used in CHARMM and X-PLOR software packages.
MCCE2 [152], 2009 **MCCE** [108], 2002	standalone	SC → AA	Uses a geometry-based and molecular mechanics approach to place all non-hydroxyl hydrogen atoms. For hydroxyl and water hydrogens it uses systematic search of torsion angles. The energy function includes torsion angle (from CHARMM), van der Waals, solvation and continuum electrostatics.	**Cysteine residues cannot be treated as disulfide bridged.** This is a slower but more accurate approach that can be used for studies involving a specific protein, especially when the protonation states of ionizable residues and orientations of buried hydroxyls are relevant.
PyMOL, DeepView (SPV) [105], 1997	standalone: https://pymol.org/2/ https://spdbv.vital-it.ch/disclaim.html	SC → AA	Molecular visualization tools that use only geometric criteria, without minimization.	
Protonate3D [109], 2009	Standalone available on request from the authors http://www.chemcomp.com	SC → AA AA optimization	Predicts hydrogen geometry, ionization, and tautomer states for macromolecular structures based on 3D coordinates. The energy model includes van der Waals, electrostatics, solvation, rotamer, tautomer, and titration effects. Optimal states are chosen according to a chemical model derived from the MMFF94 force field.	Tested on ultra-high resolution X-ray structures. The method **considers side-chain flip, rotamer, tautomer, and ionization states** of all chemical groups, ligands, and solvent based templates are available in a parameter file. Close contacts and other poor geometry may cause structure distortions. The tool is not available for free.
Protoss [153,154,110], 2014	server: https://proteins.plus/	SC → AA AA optimization	Adds hydrogen atom positions based on optimal hydrogen bond networks in the protein-ligand interface. Networks are modeled as graphs. Uses an efficient dynamic programming approach with storing partial solutions and combining them to globally optimal solutions. The algorithm is split into two phases: initialization (performed only once) and optimization.	**Can be used to model the protein-ligand interface.** Predicted hydrogen positions were compared with those in high-resolution protein structures (the test set consisted of 34 hydrogen atoms from seven protein structures). This approach does not work well on strongly interconnected graphs (1ps3). **Samples 60 orientations for a water molecule.** The tool is faster than Protonate3D.

(*continued on next page*)

Table 1 (*continued*)

Method, reference and year of the last publication	Software availability*	Reconstruction** task	Description***	Benchmark sets and comments***
REDUCE (MolProbity) [111,155], 2010	standalone: http://kinemage.biochem.duke.edu/software/reduce.php a part of the MolProbity server: http://molprobity.biochem.duke.edu/	SC → AA AA optimization	Adds hydrogens based on expected atomic geometry lengths and angles. Places hydrogens to optimize local H-bonding networks, avoid steric overlaps and **detect the correct orientations of side chains for NQH residues, as well as imidazole ring, OH, SH, NH3+, Met methyls, HET groups.** The protonation state of histidine is adjusted based on the local environment.	**Both proteins and nucleic acids can be processed.** This approach is also efficient when a more intensive approach is desired. MolProbity evaluates X-ray and NMR structures (ensemble structures of up to 80 models, accepts an mmCIF file and automatically converts it to the PDB hybrid36 format) and rebuilds the model by removing outliers as part of the refinement cycle.
WHAT IF [112], 1990	server: https://swift.cmbi.umcn.nl/servers/html/index.html select option: Hydrogen, then Add Protons	SC → AA AA optimization	Adds all missing hydrogens to the structure. It contains several servers which additionally compute all possible hydrogen bonds, but in default they do not determine which bonds would be most favorable.	Uses the Optimal Hydrogen Bonds server for computing the best possible hydrogen bond network. The program works much slower when the system contains many water molecules. Dedicated for LINUX systems.
Reconstruction from coarse-grained protein complexes with other biomolecules				
BACKWARD [132], 2014	standalone: http://cgmartini.nl/index.php/back	Protein-lipid MARTINI → AA	Method for reconstruction from the MARTINI coarse-grained representation of **protein-lipid systems**. Uses a strictly geometric approach based on Cα triplets for rebuilding the protein backbone from coarse-grained beads. It is possible to map from MARTINI CG to united-aliphatic atom (GROMOS) or all-atom (CHARMM, AMBER) representation of **single and multimeric proteins**.	Tested on 6 systems including lipid bilayers, proteins in solution (YvoA), membrane proteins (ASIC) and peptides (WALP). Reconstruction quality $RMSD_{BB}$: <0.6 Å. The approach enables integral backmapping and reconstructing complete systems, including the solvent.
Stansfeld & Sansom [133], 2011	Standalone available on request from the authors MemProtMD database: http://memprotmd.bioch.ox.ac.uk/	Protein-lipid CG → AA optimization	Method for reconstruction from the MARTINI coarse-grained representation of **protein-lipid systems**. Uses fragment-based libraries for reconstructing CG complex protein-lipid bilayer systems. The protocol starts from the MARTINI CG model and uses all-atom force fields such as CHARMM36, GROMOS and OPLS for final energy minimization in MD simulations. Atomic details of protein structure are obtained by using MODELLER or PULCHRA. **Higher resolution of lipids is provided by a library of atomistic lipid fragments.**	Tested on 10 membrane protein-lipid bilayer systems of different size and complexity, generated by self-assembly CGMD simulations (leuT, aquaporin, ELIC, ASIC, Cyt Ox, KcsA, SERCA, β_2AdR/lysozyme, OmpC, OSC). This approach does not attempt to convert united water particles. The algorithm is implemented in perl programming language.
Shimizu & Takada [134], 2018	Standalone available on request from the authors	Protein-DNA CG → AA optimization	Method for reconstruction from coarse-grained representation of **protein-DNA complexes**. Uses a **DNA fragment library** to reconstruct all-atomic details of DNA and optimize side chain orientations of the protein-DNA interface. Other fragments of protein structure are modeled with PD2 and SCWRL4. The final stage of the procedure includes the addition of all hydrogen atoms by gmx (pdb2gmx) [156]. The method reconstructs atomic details from a CG protein-DNA complex (CafeMol representation), where an amino acid is replaced by a single bead at the Cα position and a deoxyribonucleotide by three beads for the sugar, phosphate and base.	A library of 22,347 DNA fragments is derived from high-resolution X-ray structures from PDB. Tested on 180 complex protein-DNA experimental structures with **single or multiple DNA chains** and CG models obtained from CGMD simulations. This approach provides the tilt of a base plane well and proper Watson-Crick base pairing of hydrogen bonds and **maintains the initial protein-DNA interface**. It should also be applicable to **other complexes as protein-ligand or multi-protein systems**.

* links to web servers or standalone methods have been provided only if working at the time of writing this publication.
** reconstruction tasks realized by outlined methods are summarized in the third column using the following shortcuts: contact map (CM), alpha carbon atoms (CA), backbone atoms (BB), backbone and side chain atoms (SC), all-atom representation that includes backbone, side chain and hydrogen atoms (AA), coarse-grained representation (CG).
*** some major or unique features are bolded for readers convenience.

resolution of protein structures is required for many practical structure-based studies, including drug design and protein design [2–4]. Therefore, the practical use of coarse-grained protein models and elastic network models requires integration with efficient tools for rebuilding atomic details [1,157]. Ideally, the reconstruction procedure should be effective not only for regularly packed folded protein structures, but also for models of disordered or partially unfolded proteins [157,158].

In this mini-review, we provide an overview of the available computational tools for reconstruction of all-atom protein structures from various levels of incomplete representation. The review is organized as follows. First, we present the typical reconstruction pipeline and visualize example coarse-grained protein models of various resolutions (Section 2.1). Then, we review the computational methods for consecutive reconstruction steps from low to high resolution levels: reconstruction from low-resolution and contact maps (Section 2.2), backbone reconstruction from the C-alpha trace (Section 2.3), side chain reconstruction from the backbone (Section 2.4), hydrogen atom reconstruction (Section 2.5) and final optimization/refinement of all-atom structures (Section 2.6). The reconstruction methods are described and reviewed in Table 1.

2. Protein structure reconstruction methods

2.1. Stages of protein reconstruction

Fig. 1 shows a typical reconstruction pipeline used in multiscale modeling methods that merge coarse-grained protein modeling tools with all-atom modeling. Coarse-grained protein models can present different levels of resolution [1]. In the case of low-resolution models (such as SICHO [5,6] or SURPASS [7,8]) the coarse-graining level can be so deep that it does not take into account even the explicit positions of alpha carbons (see Fig. 2).

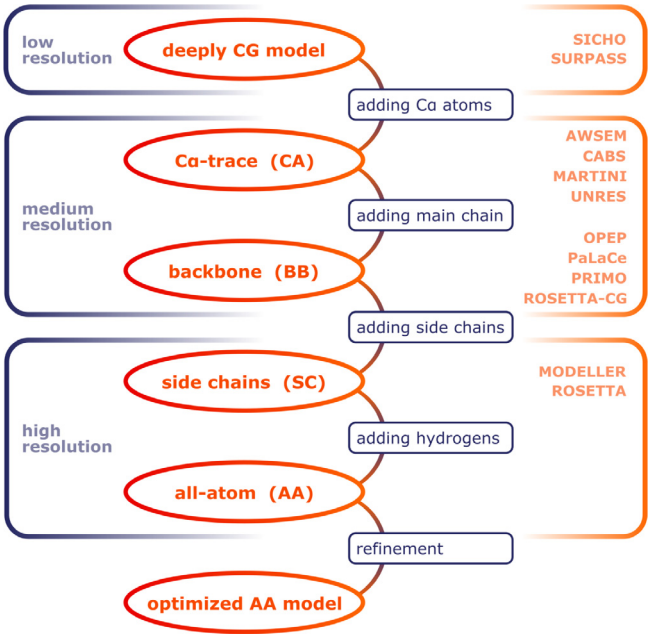

Fig. 1. Typical stages of protein structure reconstruction. The required range of reconstruction stages depends on the resolution of the initial models. For some deeply coarse-grained (CG) models, the first step is to reconstruct positions of C-alpha (CA) atoms. For most medium resolution CG models, recovering atomistic details starts with backbone (BB) reconstruction from the CA atoms that is followed by side-chain (SC) reconstruction and, subsequently, adding hydrogen atoms. The geometry of the final all-atom structure can be further improved using various refinement techniques.

In such cases, structure reconstruction requires an additional stage addressed to determine the C-alpha trace from the unified atoms that encode deeply averaged fragments of protein structure. This is not a trivial task due to the lack of unambiguous mathematical formula or simple geometric rules. However, as accurate as possible determination of the C-alpha trace plays crucial role for subsequent reconstruction of all-atom structure. C-alpha atoms are explicitly present in majority of medium resolution coarse-grained models (such as CABS [9], UNRES [10,11], AWSEM [12] or MARTINI [13], see Fig. 2) and C-alpha based elastic network models [157]. In these cases, the reconstruction procedure starts from the C-alpha trace level. Higher-resolution coarse-grained models, such as ROSETTA-centroid [14] (see Fig. 2), OPEP [15], PRIMO [16] or PaLaCe [17], require side chains reconstruction from protein backbone coordinates.

2.2. Reconstruction from low-resolution models and contact maps

Reconstruction from low-resolution coarse-grained protein models is a significant challenge and depends on the specificity of the model's simplification. For example, the SICHO [5,6] coarse-grained protein model (see Fig. 2) is based on an assumption that the protein spatial structure is determined and maintained by interactions between packed side chains. The single united atom per residue is located in the center of mass of the side group. Based on side chain center positions, the C-alpha trace and backbone heavy atoms can be reconstructed using a set of geometric criteria (for more details see the SICHO method in Table 1). Another low-resolution SURPASS model [7,8] assumes the averaging of short 4-residue long fragments of secondary structure to a single united atom lying in the center of their mass. As a result, the representation of regular secondary structure elements (α-helices and β-strands) in this model is almost linear. The procedure for recovering the C-alpha trace from SURPASS representation uses the SUReLib library (see Table 1), which consists of short fragments differentiated by the type of secondary structure. The positions of rebuilt C-alpha atoms maintain correct geometry and spatial orientation. Therefore, the reconstructed C-alpha trace can be used as a source of restraints (distances, angles or contacts) for higher resolution models or directly reconstructed to atomic resolution using the available tools.

Protein contact maps are another kind of low-resolution protein models generated by contact prediction methods [18]. The contact maps are usually defined as binary entries or distance maps between Cα or Cβ atoms [18]. Distance restraints can also be an outcome of low-resolution experimental data analysis (SAXS [19–21], NMR [22], cryo-EM [23], XL-MS [24], HDX-MS [25,26]). Prediction of contact maps (and their application in protein structure modeling) has become more accurate and effective by using evolutionary coupling analysis (DCA) of multiple sequence alignment (MSA) and deep neural networks to detect high-order correlation [27,28]. The reconstruction of three-dimensional protein structure based on a specific contact map is an NP-hard problem. Using the preferred contacts as restraints in de novo modeling can lead to more accurate structure predictions than template-based modeling, especially for proteins without close homologs [29]. The predicted contact maps often contain a fraction of false contacts. Some reconstruction from contact maps are robust to inaccurate or incomplete sets of preferred contacts (e.g. FT-COMAR [30], CONFOLD [31,32], GDFuzz3D [33], see Table 1). Contact maps are typically used as distance restraints between pairs of alpha carbons or as part of the force field in de novo structure modeling (e.g. CONFOLD, PconsFold [34]). Initial, partially random atomic positions are optimized in an iterative procedure to satisfy the specified distance restrictions.

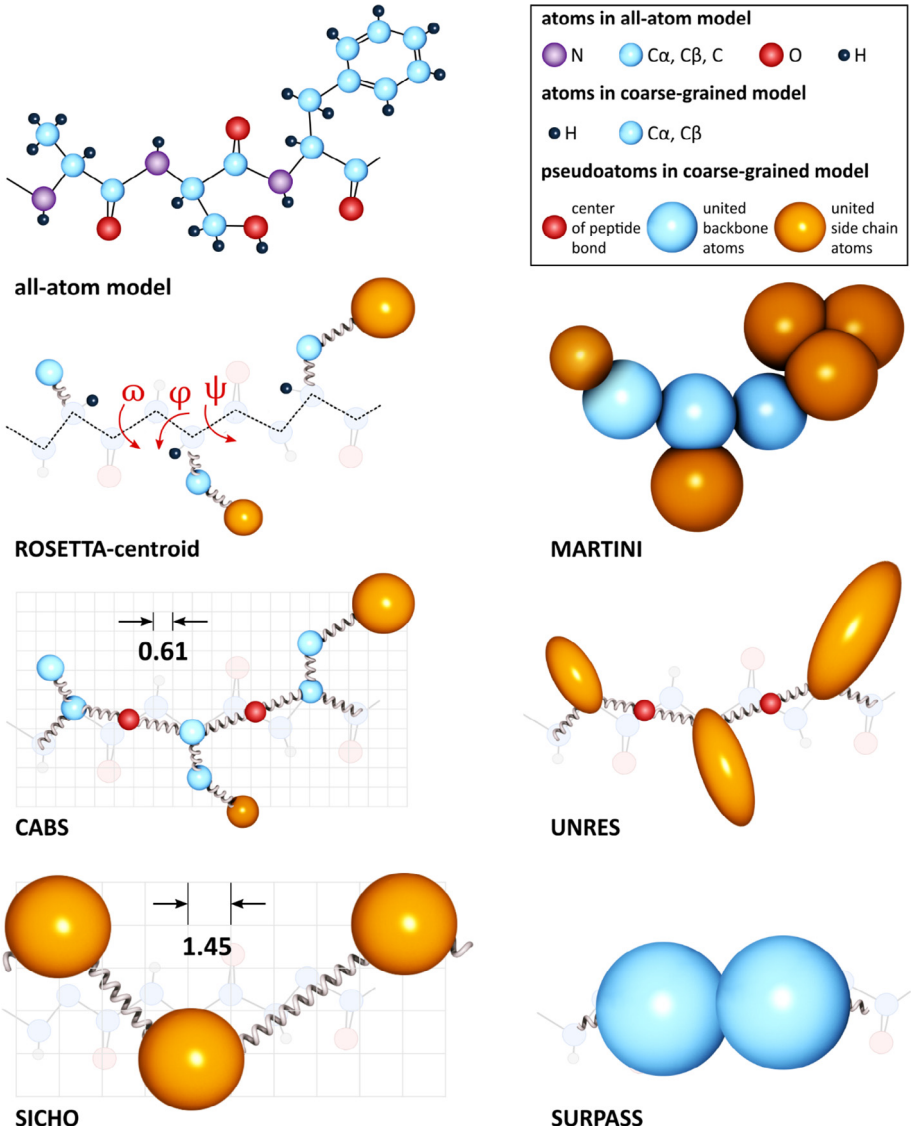

Fig. 2. Example tripeptide presented in all-atom and corresponding coarse-grained resolutions. Various coarse-grained modeling tools are shown: Rosetta-centroid, MARTINI, CABS, UNRES, SICHO and SURPASS. Note that most coarse-grained models use explicit positions of (pseudo) atoms while ROSETTA uses a set of torsional angles φ, ψ, ω to describe backbone geometry. The legend explaining the colors of atoms and pseudoatoms is presented in top right.

2.3. Backbone reconstruction from C-alpha positions

The arrangement of alpha carbons in the polypeptide chain is locally very regular with an average distance of 3.8 Å between neighboring Cα atoms. There are many methods dedicated to reconstruction of protein backbone coordinates, which provide models of protein backbone geometry (or complete all-atom structure) based on the C-alpha trace (see Table 1, section "Backbone reconstruction from CA-trace" and section "All-atom reconstruction from CA-trace"). Heavy atoms (N, C, O) in the main chain are usually added according to simple geometric criteria based on bond lengths and angles in the peptide plane (proline residues need separate treatment) [35–37]. The optimal roto-translation of the peptide plane is usually provided by the sequence-dependent statistical potential that assumes ideal bond lengths and phi-psi angles. Instead of inserting individual atoms, the other commonly employed approach is to use a library of peptide backbone fragments [38–41]. The fragments, typically from 4 to 15 residues long, are derived from non-redundant set of known protein structures and collected in the library. The size of libraries can be very wide and results from clustering strategy and adopted criteria. Some libraries are built from several hundred (e.g. 528 in PD2 method [39]) to even several thousand structural components (e.g. 5148 of 4-residue fragments in BBQ method [38]) with fixed or multiple overlapping fragment lengths [42]. The strategy of using protein fragments of various lengths is also successfully used by Rosetta [43], Modeller [44], and I-TASSER [136] packages for protein structure prediction. The large size and diversity of backbone libraries is likely to ensure high accuracy of reconstructed structures, but it increases the cost of calculations [45]. Therefore, much smaller size libraries (a dozen or several tens of fragments) are offered by methods based on structural alphabets such as Protein Blocks [46], SA-HMM [47] SABBAC [40] and other methods [48–50]. The structural alphabets are libraries consisting of short (from 3 to 7 residue long) usually fixed-length backbone fragments, that can be used as building blocks in protein reconstruction [42,46,51]. During the reconstruction procedure, overlapping fragments are selected from the library that best fit to the C-alpha trace. Selection of preferred fragments is based on energy scores, structural similarity, secondary structure assignment or geometric matching criteria [35].

Typically, the accuracy of the backbone reconstruction procedure is evaluated using measurement of the RMSD values (average; or of individual atoms: C, N, O; to a reference structure calculated on main chain heavy atoms) and Ramachandran dihedral angles (ψ, φ). For example, comparison of selected methods for protein backbone reconstruction from the C-alpha trace [39] showed that PD2 (especially with minimization step) and BBQ remain the most accurate due to RMSD and dihedral angle shifts criteria. Accuracy of those tools can be further improved by additional refinement of the protein backbone [52].

When considering the reconstruction of protein structures from coarse-grained modeling, a very important aspect that should be bear in mind is the ability of the method to handle unphysical distortions of the initial C-alpha trace. The various backbone reconstruction methods show different resistance to small unphysical local distortions in the Cα chain that are often present in coarse-grained models [1]. For some approaches, fragments of incorrect C-alpha trace geometry can result in missing parts of the rebuilt backbone or unphysical backbone distortions. These may have significant impact on the quality of subsequent side chain reconstruction, all-atom energy-minimization and scoring [39]. Some of the backbone reconstruction methods (like PD2 [39], SABBAC [40], ModRefiner [53], PULCHRA [54], RACOGS [55]) have been designed to be robust to small (~1 Å) distortions in the initial Cα chain of coarse-grained models. Methods like PULCHRA and ModRefiner offer additional optimization of the reconstructed main chain including Cα positions (see Table 1). Finally, it should be noted that methods based on fragment libraries, while usually effective in reconstruction of folded proteins, do not always cope with unstructured/disordered fragments of the protein chain.

2.4. Side-chain reconstruction from backbone

Side group interactions (hydrogen bonds, ionization, solvation, contacts) have a major role for the stabilization of three-dimensional protein structure [56–58] and binding interaction in protein complexes [59–62]. Therefore, the accurate side chains packing is important in structure prediction of proteins, their complexes and protein design [59,63–65]. Except for a few methods [66,67], most of the available side chain reconstruction methods are based on the position of backbone atoms and use rotamer/conformer libraries [68–72] with various strategies for the optimization of side chain packing [73–80]. Such backbone-dependent rotamer libraries define the probability of a given rotamer as a function of the main chain dihedral angles. Thus, backbone distortions (for example errors in backbone reconstruction) may have a significant influence on the accuracy of reconstructed side chains. However, minor backbone distortions are tolerated by some reconstruction methods [53–55,81,82].

The prediction of side chain conformations and packing usually involves three crucial modules:

- all-atom or coarse-grained rotamer library of discrete side chain conformations (conformer library) or the frequency distribution of rotational states (statistical rotamer library); rotamer models differ in flexibility (rigid or flexible), number of available rotameric states, packing conditions (e.g. force field, score function) and backbone dependencies
- set of energy functions to distinguish rotamer states (various combinations of van der Waals and electrostatic potentials, solvation effects, hydrogen bonds and orientation-dependent terms)
- search algorithm for efficient sampling of the conformational space of rotameric states: Monte Carlo Dynamics or Molecular Dynamics, simulated annealing scheme, neural networks, dead-end elimination, graph theory-based, self-consistent mean field, branch-and-terminate, backtrack and various combinations of these approaches [73,83].

The side chain reconstruction methods try to strike the balance between these modules by enhancing the sampling scheme [86,74,87,76], optimizing terms of energy function [78,88–91] or improving rotamers library [92–94]. Reconstruction of side chain geometry defining their proper spatial packing is a much more challenging task than reconstruction of the protein backbone. It is related to the high flexibility of side groups, especially for larger amino acids, defining a vast conformational space that needs to be considered [57]. The complexity of the side chain reconstruction problem can be simplified by using a finite number of variants of the spatial arrangement of side-chain rotamers. Rotational states are stored in the library, which can be efficiently searched even for large proteins or their complexes [68,85]. Rotamers are selected to avoid steric clashes and to provide favorable local interactions.

There are many software tools dedicated only to side-chain reconstruction that available mainly as standalone programs [76,79,77,92,96–99] (see Table 1, section "Side chains reconstruction from backbone"). The side-chain reconstruction methods are also available within integrated software for reconstruction of atomic details (including optimization of side chain packing) from the initial C-alpha trace [37,41,44,53–55] (see Table 1, section "All-atom reconstruction from CA-trace").

A comparison of the best performing methods in various residue environments (buried, surface, interaction interface, membrane-spanning) and protein types (membrane, mono- and multimeric) can be found in the comprehensive benchmark [73]. For all OSCAR (-o [78], -star [97]), OPUS (-Rota [96], -Rota2 [81]), Upside [100], SCWRL4 [77], RASP [83] methods the overall accuracy exceeded 85% of χ_1 angle, 75% of χ_1 + χ_2 angles and below 1.5 Å of average RMSD between all-atoms in the predicted and native side chain conformations. Interestingly, another evaluation of some best performing algorithms suggested that for buried residues in the protein, the algorithms are close to the best possible accuracy [95]. For exposed residues, there is large room for improvement and the scoring functions seem to be the main obstacle to correct side-chain packing [95]. Another room for improvement remains also in the design and specialization of rotamer libraries. This has been recently demonstrated in the work on the PEARS tool [82], a family specific side-chain predictor for antibodies, in which rotamers are binned according to their immunogenetics position rather than their local backbone geometry. The concept of PEARS is potentially generalizable to other protein families, provided that enough structural data is available.

The computational efficiency of these methods differs significantly. For example, the Upside method is extremely fast (Upside needs 0.006 s per 100 residues). RASP, OPUS-Rota2 and SCWRL4 methods are approximately 15, 150 and 300 times slower, respectively. The OPUS-Rota and the OSCAR-star are almost equally fast as the SCWRL4 and the OSCAR-o is 2 orders of magnitude slower [81,83,100].

Taking into account methods accuracy, efficiency and various features, different methods may be better in different applications (see Table 1). For example, Upside and OPUS-Rota2 methods have been tested in modeling of non-native conformers and can be very efficient as a component of multiscale modeling protocols for simulation of protein dynamics. SCWRL4 and OPUS-Rota methods are easy-to-use and well tested in the application to homology modeling. Also, SCWRL4 can improve the interactions of side chains within the crystal conformations, which can be useful in molecular replacement, structure refinement or prediction of protein-protein interfaces [77]. Both OPUS- and OSCAR- tools variants are sensitive to side chain orientations and used in selecting near-native conformations from decoys [101,102].

2.5. Hydrogen atom reconstruction

Hydrogen atoms account for nearly half of the atoms in protein structure. Omitting them in coarse-grained modeling enables significant simplification of the conformational space and acceleration of calculations by an order of magnitude. However, a more detailed analysis of system energy (e.g. ligand binding to a protein) requires an accurate physicochemical force field, in which hydrogens are treated in an explicit manner and their location significantly contributes to system energy (hydrogen bonds, ionization, solvation, contacts and structure stabilization). There are many tools for placing hydrogen atoms according to geometric criteria, and they also include specific effects, such as tautomeric or protonation states. The experimental structures or reconstructed models may have local stresses or clashes that require additional energy optimization. To minimize energy, some methods also refine the final structure using molecular dynamics simulations (see Table 1) or even quantum-mechanical calculations [103].

For most Protein Data Bank entries the experimental structures contain incomplete information about the proper location of hydrogen atoms. The main limiting factor for experimental techniques in the detection of hydrogen positions is their high mobility. However, the hydrogen occurring in various functional groups differs in rotational flexibility. Tautomeric states occur mainly in histidine and carboxyl groups. Torsional angle changes based on the rotation of the hydrogen position around the bond with the heavy atom involve mainly hydroxyl, thiol and amine groups. Protonation states differ in the number of hydrogens in the functional group due to losing (negative charge for carboxyl or thiol) or adding a proton (positive charge for amine or imidazole). Side chain flips occur in amide and imidazole groups and are particularly frequent for glutamine and asparagine residues.

Several tools that address the location of hydrogen atoms in protein structure have been developed (see Table 1, section "Hydrogen atom reconstruction"). Some of them add hydrogen according to simple geometric criteria (CHIMERA [104], PyMOL, DeepView [105]), while others take into account more subtle interactions and perform additional optimization (Computational Titration [106], HAAD [107], MCCE [108], Protonate3D [109], Protoss [110], REDUCE [111], WHAT IF [112]) or employ molecular dynamics (CNS [113], GROMACS [114], Hbuild [115]). Adding hydrogen atoms is a necessary step in crystallographic structure refinement, theoretical structure prediction, or calculation of associated binding energies [107,116]. A typical hydrogen reconstruction scheme involves initial placement of atoms according to geometric criteria which are then optimized by conformational search guided using empirical or physicochemical energy terms [113–117] or heuristic approaches [111,112]. Most methods are very effective in predicting the position of a hydrogen atom that is bonded to a tetrahedral geometry atom (both C and N), especially when the positions of the other three atoms are known. Quite good compatibility was also obtained for planar hydrogens and CH_2-type groups. It is slightly more difficult to predict the orientation of the CH_3 and NH_3 groups due to their high rotational flexibility and planar amine groups in asparagine, glutamine and arginine. In this case, geometry-based methods provide the highest accuracy (MCCE, WHAT IF) [116]. CHARMM software seems to be an efficient tool to predict hydroxyl and water hydrogens [116]. The HAAD [107] method is very effective in avoiding steric clashes in the densely packed hydrophobic protein core. REDUCE [111] and several recently developed tools such as Protoss [110] or Protonate3D [109] effectively take into account the effects of rotamers, tautomers and ionization states as well as side chain flips.

2.6. Optimization of all-atom structure

The accuracy of all-atom protein models, obtained using protein reconstruction methods and/or experimental techniques, can be further improved using physics-based energy-minimization and simulation techniques [84,6]. Most commonly, the optimization step is the last step of reconstruction procedures. However, energy minimization can be also combined with different reconstruction steps. This is the case of the ModRefiner method [53] which uses two-step atomic level minimization: the first one to refine the backbone only, and the second one to refine all-atom models.

Optimization of protein models can be short-timescale and aimed at local-scale improvement [118,119], i.e. side chain repacking, loop remodeling or optimization of hydrogen bonding in secondary structure elements. Much more challenging is deeper long-timescale optimization aimed at large conformational changes toward more accurate model [118,120–125]. The most common approach for optimization of protein models is all-atom Molecular Dynamics (MD) [120–123]. Long-timescale MD simulations require enormous computational resources but they can usually be significantly accelerated by proper sampling strategies [126–129], use of spatial restrains and knowledge-based information [120–123,159,160]. The recent evaluation of protein refinement techniques in the CASP12 experiment showed that the best performing approaches used restrained MD simulations alone, or in combination with other tools [122].

3. Summary

For successful reconstruction of all-atom protein models, computational methods most commonly use a set of geometric rules, libraries of protein fragments, various simulation techniques or their combinations. The most effective strategy for backbone reconstruction of folded proteins seems to be assembly from known protein fragments. This is because of the well-defined character of the protein backbone that is structurally conserved among homologous proteins and maintains major structural regularities in protein fragments of similar sequence. What's important to bear in mind, the accuracy of backbone reconstruction has significant impact on the accuracy of subsequent side-chain reconstruction and energy-based scoring of obtained models [39,81]. Reconstruction of side chain positions is a challenging problem and also in this case statistical regularities extracted from known protein structures can be useful [82]. The problem is NP-hard in nature and only suboptimal solutions are available. Nevertheless, for many reconstruction tasks such suboptimal solutions are satisfactory. Eventually, the performance of backbone and side chain reconstruction stages can be improved through combination with physics-based optimization techniques.

Methods of protein structure reconstruction from incomplete models are already commonly used and will be valuable components of modeling strategies that integrated data from various sources. Those sources include experiment (like SAXS, NMR, X-ray, cryo-EM [19–23,84] or measurements of the activity of mutant protein variants [130,131]) and theoretical predictions (like residue-residue contact predictions from evolutionary information [27,28] or simulation trajectories in coarse-grained resolution [1,157,158]). Since the all-atom MD is the most widely employed simulation method, the local quality and stability of reconstructed structures should be tested by using them as starting points for the all-atom MD. The growing number of experimental data or coarse-grained predictions on the structure of protein complexes also call for reconstruction methods designed for refining structural models of different biomolecules (the examples of methods for

reconstruction of protein-lipid [132,133] and protein-DNA [134] systems are presented in Table 1). This short review focuses on reconstruction tools which use various kinds of coarse-grained protein representations as the input. Note that there are also a number of tools, not discussed in this review, that enable filling the gaps of missing residues in protein structures [135–137].

Finally, we hope this short review can be a useful reference to existing protein reconstruction resources. They may be useful for design and development of new efficient molecular modeling tools, but also for a much larger community of bioscientists who may use reconstruction methods as supporting tools for deeper analysis and illustration of experimental data in structural biology, biomedicine and other branches of molecular biology. The tools available as web servers (see the availability column in Table 1) are probably the easiest to access and use.

4. Acknowledgments

AEB-D, AK, SK received funding from NCN Poland, Grant MAESTRO2014/14/A/ST6/00088.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Kmiecik S, Gront D, Kolinski M, Wieteska L, Dawid AE, Kolinski A. Coarse-grained protein models and their applications. Chem. Rev. 2016;116:7898–936. https://doi.org/10.1021/acs.chemrev.6b00163.
[2] Huang PS, Boyken SE, Baker D. The coming of age of de novo protein design. Nature 2016;537:320–7. https://doi.org/10.1038/nature19946.
[3] Śledź P, Caflisch A. Protein structure-based drug design: from docking to molecular dynamics. Curr. Opin. Struct. Biol. 2018;48:93–102. https://doi.org/10.1016/j.sbi.2017.10.010.
[4] Kuhlman B, Bradley P. Advances in protein structure prediction and design. Nat. Rev. Mol. Cell Biol. 2019. https://doi.org/10.1038/s41580-019-0163-x.
[5] Kolinski A, Jaroszewski L, Rotkiewicz P, Skolnick J. An efficient Monte carlo model of protein chains. Modeling the short-range correlations between side group centers of mass. J. Phys. Chem. 1998;102:4628–37.
[6] Stumpff-Kane AW, Maksimiak K, Lee MS, Feig M. Sampling of near-native protein conformations during protein structure refinement using a coarse-grained model, normal modes, and molecular dynamics simulations. Proteins Struct. Funct. Bioinf. 2008;70:1345–56. https://doi.org/10.1002/prot.21674.
[7] Dawid AE, Gront D, Kolinski A. SURPASS low-resolution coarse-grained protein modeling. J. Chem. Theory Comput. 2017;13:5766–79. https://doi.org/10.1021/acs.jctc.7b00642.
[8] Dawid AE, Gront D, Kolinski A. Coarse-grained modeling of the interplay between secondary structure propensities and protein fold assembly. J. Chem. Theory Comput. 2018;14:2277–87. https://doi.org/10.1021/acs.jctc.7b01242.
[9] Kolinski A. Protein modeling and structure prediction with a reduced representation. Acta Biochim. Pol. 2004;51:349–71. 035001349.
[10] Liwo A, Ołdziej S, Pincus MR, Wawak RJ, Rackovsky S, Scheraga HA. A united-residue force field for off-lattice protein-structure simulations. I. Functional forms and parameters of long-range side-chain interaction potentials from protein crystal data. J. Comput. Chem. 1997;18:849–73. https://doi.org/10.1002/(SICI)1096-987X(199705)18:7<849::AID-JCC1>3.0.CO;2-R.
[11] Czaplewski C, Karczyńska A, Sieradzan AK, Liwo A. UNRES server for physics-based coarse-grained simulations and prediction of protein structure, dynamics and thermodynamics. Nucleic Acids Res. 2018;46:W304–9. https://doi.org/10.1093/nar/gky328.
[12] Davtyan A, Schafer NP, Zheng W, Clementi C, Wolynes PG, Papoian GA. AWSEM-MD: Protein structure prediction using coarse-grained physical potentials and bioinformatically based local structure biasing. J. Phys. Chem. B 2012;116:8494–503. https://doi.org/10.1021/jp212541y.
[13] Monticelli L, Kandasamy SK, Periole X, Larson RG, Tieleman DP, Marrink SJ. The MARTINI coarse-grained force field: extension to proteins. J. Chem. Theory Comput. 2008;4:819–34. https://doi.org/10.1021/ct700324x.
[14] Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J. Mol. Biol. 1997;268:209–25. https://doi.org/10.1006/jmbi.1997.0959.
[15] Sterpone F, Melchionna S, Tuffery P, Pasquali S, Mousseau N, Cragnolini T, et al. The OPEP protein model: from single molecules, amyloid formation, crowding and hydrodynamics to DNA/RNA systems. Chem. Soc. Rev. 2014;43:4871–93. https://doi.org/10.1039/C4CS00048J.
[16] Gopal SM, Mukherjee S, Cheng YM, Feig M. PRIMO/PRIMONA: a coarse-grained model for proteins and nucleic acids that preserves near-atomistic accuracy. Proteins Struct. Funct. Bioinf. 2010;78:1266–81. https://doi.org/10.1002/prot.22645.
[17] Pasi M, Lavery R, Ceres N. PaLaCe: a coarse-grain protein model for studying mechanical properties. J. Chem. Theory Comput. 2013;9:785–93. https://doi.org/10.1021/ct3007925.
[18] Di Iena P, Nagata K, Baldi P. Deep architectures for protein contact map prediction. Bioinformatics 2012;28:2449–57. https://doi.org/10.1093/bioinformatics/bts475.
[19] Schindler CEM, de Vries SJ, Sasse A, Zacharias M. SAXS data alone can generate high-quality models of protein-protein complexes. Structure 2016;24:1387–97. https://doi.org/10.1016/j.str.2016.06.007.
[20] Lipfert J, Doniach S. Small-angle X-ray scattering from RNA, proteins, and protein complexes. Annu. Rev. Biophys. Biomol. Struct. 2007;36:307–27. https://doi.org/10.1146/annurev.biophys.36.040306.132655.
[21] Korasick DA, Tanner JJ. Determination of protein oligomeric structure from small-angle X-ray scattering. Protein Sci. 2018;27:814–24. https://doi.org/10.1002/pro.3376.
[22] Würz JM, Kazemi S, Schmidt E, Bagaria A, Güntert P. NMR-based automated protein structure determination. Arch. Biochem. Biophys. 2017;628:24–32. https://doi.org/10.1016/j.abb.2017.02.011.
[23] De Vries SJ, Chauvot De Beauchêne I, Schindler CEM, Zacharias M. Cryo-EM data are superior to contact and interface information in integrative modeling. Biophys. J. 2016;110:785–97. https://doi.org/10.1016/j.bpj.2015.12.038.
[24] Leitner A, Faini M, Stengel F, Aebersold R. Crosslinking and mass spectrometry: an integrated technology to understand the structure and function of molecular machines. Trends Biochem. Sci. 2016;41:20–32. https://doi.org/10.1016/j.tibs.2015.10.008.
[25] Konermann L, Pan J, Liu YH. Hydrogen exchange mass spectrometry for studying protein structure and dynamics. Chem. Soc. Rev. 2011;40:1224–34. https://doi.org/10.1039/c0cs00113a.
[26] Trabjerg E, Nazari ZE, Rand KD. Conformational analysis of complex protein states by hydrogen/deuterium exchange mass spectrometry (HDX-MS): challenges and emerging solutions. TrAC - Trends Anal. Chem. 2018;106:125–38. https://doi.org/10.1016/j.trac.2018.06.008.
[27] Kamisetty H, Ovchinnikov S, Baker D. Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. Proc. Natl. Acad. Sci. 2013;110:15674–9. https://doi.org/10.1073/pnas.1314045110.
[28] Ovchinnikov S, Park H, Varghese N, Huang P-S, Pavlopoulos GA, Kim DE, Kamisetty H, et al. Protein structure determination using metagenome sequence data. Science 2017;355:294–8. https://doi.org/10.1126/science.aah4043.
[29] Wang S, Sun S, Li Z, Zhang R, Xu J. Accurate de novo prediction of protein contact map by ultra-deep learning model. PLoS Comput. Biol. 2017;13:e1005324. https://doi.org/10.1371/journal.pcbi.1005324.
[30] Vassura M, Margara L, Dilena P, Medri F, Fariselli P, Casadio R. FT-COMAR: fault tolerant three-dimensional structure reconstruction from protein contact maps. Bioinformatics 2008;24:1313–5. https://doi.org/10.1093/bioinformatics/btn115.
[31] Adhikari B, Bhattacharya D, Cao R, Cheng J. CONFOLD: residue-residue contact-guided ab initio protein folding. Proteins Struct. Funct. Bioinforma. 2015;83:1436–49. https://doi.org/10.1002/prot.24829.
[32] Adhikari B, Cheng J. CONFOLD2: improved contact-driven ab initio protein structure modeling. BMC Bioinf. 2018;19:22–6. https://doi.org/10.1186/s12859-018-2032-6.
[33] Pietal MJ, Bujnicki JM, Kozlowski LP. GDFuzz3D: a method for protein 3D structure reconstruction from contact maps, based on a non-Euclidean distance function. Bioinformatics 2015;31:3499–505. https://doi.org/10.1093/bioinformatics/btv390.
[34] Michel M, Hayat S, Skwark MJ, Sander C, Marks DS, Elofsson A. PconsFold: improved contact predictions improve protein models. Bioinformatics 2014;30:i482–8. https://doi.org/10.1093/bioinformatics/btu458.
[35] Payne PW. Reconstruction of protein conformations from estimated positions of the Cα coordinates. Protein Sci. 1993;2:315–24. https://doi.org/10.1002/pro.5560020303.
[36] Feig M, Rotkiewicz P, Kolinski A, Skolnick J, Brooks CL. Accurate reconstruction of all-atom protein representations from side-chain-based low-resolution models. Proteins Struct. Funct. Genet. 2000;41:86–97. https://doi.org/10.1002/1097-0134(20001001)41:1<86::AID-PROT110>3.0.CO;2-Y.
[37] Lombardi LE, Martí MA, Capece L. CG2AA: backmapping protein coarse-grained structures. Bioinformatics 2016;32:1235–7. https://doi.org/10.1093/bioinformatics/btv740.
[38] Gront D, Kmiecik S, Kolinski A. Backbone building from quadrilaterals: a fast and accurate algorithm for protein backbone reconstruction from alpha carbon coordinates. J. Comput. Chem. 2007;28:1593–7. https://doi.org/10.1002/jcc.20624.
[39] Moore BL, Kelley LA, Barber J, Murray JW, MacDonald JT. High-quality protein backbone reconstruction from alpha carbons using gaussian mixture models. J. Comput. Chem. 2013;34:1881–9. https://doi.org/10.1002/jcc.23330.
[40] Maupetit J, Gautier R, Tufféry P. SABBAC: online structural alphabet-based protein backbone reconstruction from alpha-carbon trace. Nucleic Acids Res. 2006;34:W147–51. https://doi.org/10.1093/nar/gkl289.

[41] Li Y, Zhang Y. REMO: a new protocol to refine full atomic protein models from C-alpha traces by optimizing hydrogen-bonding networks. Proteins Struct. Funct. Bioinforma. 2009;76:665–76. https://doi.org/10.1002/prot.22380.

[42] Baeten L, Reumers J, Tur V, Stricher F, Lenaerts T, Serrano L, et al. Reconstruction of protein backbones from the BriX collection of canonical protein fragments. PLoS Comput. Biol. 2008;4:e1000083. https://doi.org/10.1371/journal.pcbi.1000083.

[43] Das R, Baker D. Macromolecular modeling with rosetta. Annu. Rev. Biochem. 2008;77:363–82. https://doi.org/10.1146/annurev.biochem.77.062906.171838.

[44] Webb B, Sali A. Comparative protein structure modeling using MODELLER. Curr. Protoc. Bioinforma. 2016;54:5.6.1–5.6.37. https://doi.org/10.1002/cpbi.3.

[45] Trevizani R, Custódio FL, Dos Santos KB, Dardenne LE. Critical features of fragment libraries for protein structure prediction. PLoS One 2017;12(1):e0170131. https://doi.org/10.1371/journal.pone.0170131.

[46] Etchebest C, Benros C, Hazout S, De Brevern AG. A structural alphabet for local protein structures: improved prediction methods. Proteins Struct. Funct. Genet. 2005;59:810–27. https://doi.org/10.1002/prot.20458.

[47] Camproux AC, Gautier R, Tufféry P. A hidden Markov model derived structural alphabet for proteins. J. Mol. Biol. 2004;339:591–605. https://doi.org/10.1016/j.jmb.2004.04.005.

[48] Camproux AC, Tufféry P. Hidden Markov Model-derived structural alphabet for proteins: the learning of protein local shapes captures sequence specificity. Biochim. Biophys. Acta - Gen. Subj. 2005;1724:394–403. https://doi.org/10.1016/j.bbagen.2005.05.019.

[49] Kolodny R, Levitt M. Protein decoy assembly using short fragments under geometric constraints. Biopolymers 2003;68:278–85. https://doi.org/10.1002/bip.10262.

[50] De Brevern AG, Etchebest C, Hazout S. Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. Proteins Struct. Funct. Genet. 2000;41:271–87. https://doi.org/10.1002/1097-0134(20001115)41:3<271::AID-PROT10>3.0.CO;2-Z.

[51] Pandini A, Fornili A, Kleinjung J. Structural alphabets derived from attractors in conformational space. BMC Bioinf 2010;11:97. https://doi.org/10.1186/1471-2105-11-97.

[52] Huang DY, Hor CY, Yang CB. Coordinate refinement on all atoms of the protein backbone with support vector regression. In: Perner P, editor. Advances in data mining. Applications and theoretical aspects. Lecture Notes in Computer Science, vol. 9728. Cham: Springer; 2016. pp. 212–223. ISBN 9783319415604.

[53] Xu D, Zhang Y. Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization. Biophys. J. 2011;101:2525–34. https://doi.org/10.1016/j.bpj.2011.10.024.

[54] Rotkiewicz P, Skolnick J. Fast procedure for reconstruction of full-atom protein models from reduced representations. J. Comput. Chem. 2008;29:1460–5. https://doi.org/10.1002/jcc.20906.

[55] Heath AP, Kavraki LE, Clementi C. From coarse-grain to all-atom: toward multiscale analysis of protein landscapes. Proteins Struct. Funct. Genet. 2007;68:646–61. https://doi.org/10.1002/prot.21371.

[56] Spassov VZ, Yan L, Flook PK. The dominant role of side-chain backbone interactions in structural realization of amino acid code. ChiRotor: a side-chain prediction algorithm based on side-chain backbone interactions. Protein Sci. 2007;16:494–506. https://doi.org/10.1110/ps.062447107.

[57] Nick Pace C, Martin Scholtz J, Grimsley GR. Forces stabilizing proteins. FEBS Lett. 2014;588:2177–84. https://doi.org/10.1016/j.febslet.2014.05.006.

[58] Marcos ML, Echave J. Too packed to change: side-chain packing and site-specific substitution rates in protein evolution. PeerJ 2015;3:e911. https://doi.org/10.7717/peerj.911.

[59] Gray JJ, Moughon S, Wang C, Schueler-Furman O, Kuhlman B, Rohl CA, et al. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. J. Mol. Biol. 2003;331:281–99. https://doi.org/10.1016/S0022-2836(03)00670-3.

[60] Camacho CJ. Modeling side-chains using molecular dynamics improve recognition of binding region in CAPRI targets. Proteins Struct. Funct. Genet. 2005;60:245–51. https://doi.org/10.1002/prot.20565.

[61] Li B, Kihara D. Protein docking prediction using predicted protein-protein interface. BMC Bioinf. 2012;13:7. https://doi.org/10.1186/1471-2105-13-7.

[62] Kirys T, Ruvinsky AM, Tuzikov AV, Vakser IA. Correlation analysis of the side-chains conformational distribution in bound and unbound proteins. BMC Bioinf. 2012;13:236. https://doi.org/10.1186/1471-2105-13-236.

[63] Rohl CA, Strauss CEM, Chivian D, Baker D. Modeling structurally variable regions in homologous proteins with rosetta. Proteins Struct. Funct. Genet. 2004;55:656–77. https://doi.org/10.1002/prot.10629.

[64] Canzar S, Toussaint NC, Klau GW. An exact algorithm for side-chain placement in protein design. Optim. Lett. 2011;5:393–406. https://doi.org/10.1007/s11590-011-0308-0.

[65] Burley KH, Gill SC, Lim NM, Mobley DL. Enhancing side chain rotamer sampling using nonequilibrium candidate Monte Carlo. J. Chem. Theory Comput. 2019;15:1848–62. https://doi.org/10.1021/acs.jctc.8b01018.

[66] Zhang W, Duan Y. Grow to fit molecular dynamics (G2FMD): an ab initio method for protein side-chain assignment and refinement. Protein Eng. Des. Sel. 2006;19:55–65. https://doi.org/10.1093/protein/gzj001.

[67] Cao Y, Song L, Miao Z, Hu Y, Tian L, Jiang T. Improved side-chain modeling by coupling clash-detection guided iterative search with rotamer relaxation. Bioinformatics 2011;27:785–90. https://doi.org/10.1093/bioinformatics/btr009.

[68] Dunbrack RL. Rotamer libraries in the 21st century. Curr. Opin. Struct. Biol. 2002;12:431–40. https://doi.org/10.1016/S0959-440X(02)00344-5.

[69] Shetty RP, De Bakker PIW, DePristo MA, Blundell TL. Advantages of fine-grained side chain conformer libraries. Protein Eng. 2003;16:963–9.

[70] Shapovalov MVV, Dunbrack Jr RL, Dunbrack RL. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. Structure 2011;19:844–58. https://doi.org/10.1016/j.str.2011.03.019.

[71] Larriva M, Rey A. Design of a rotamer library for coarse-grained models in protein-folding simulations. J. Chem. Inf. Model. 2014;54:302–13. https://doi.org/10.1021/ci4005833.

[72] Towse CL, Rysavy SJ, Vulovic IM, Daggett V. New dynamic rotamer libraries: data-driven analysis of side-chain conformational propensities. Structure 2016;24:187–99. https://doi.org/10.1016/j.str.2015.10.017.

[73] Peterson LX, Kang X, Kihara D. Assessment of protein side-chain conformation prediction methods in different residue environments. Proteins Struct. Funct. Bioinforma. 2014;82:1971–84. https://doi.org/10.1002/prot.24552.

[74] Canutescu AA, Shelenkov AA, Dunbrack RL. A graph-theory algorithm for rapid protein side-chain prediction. Protein Sci. 2003;12:2001–14. https://doi.org/10.1110/ps.03154503.

[75] Fromer M, Yanover C, Harel A, Shachar O, Weiss Y, Linial M. SPRINT: side-chain prediction inference toolbox for multistate protein design. Bioinformatics 2010;26:2466–7. https://doi.org/10.1093/bioinformatics/btq445.

[76] Hartmann C, Antes I, Lengauer T. IRECS: a new algorithm for the selection of most probable ensembles of side-chain conformations in protein models. Protein Sci. 2007;16:1294–307. https://doi.org/10.1110/ps.062658307.

[77] Krivov GG, Shapovalov MV, Dunbrack RL. Improved prediction of protein side-chain conformations with SCWRL4. Proteins Struct. Funct. Bioinforma. 2009;77:778–95. https://doi.org/10.1002/prot.22488.

[78] Liang S, Zhou Y, Grishin N, Standley DM. Protein side chain modeling with orientation-dependent atomic force fields derived by series expansions. J. Comput. Chem. 2011;32:1680–6. https://doi.org/10.1002/jcc.21747.

[79] Xu J. Rapid protein side-chain packing via tree decomposition. Res. Comput. Mol. Biol. 2005;3500:423–39.

[80] Xu G, Ma T, Zang T, Sun W, Wang Q, Ma J. OPUS-DOSP: a distance- and orientation-dependent all-atom potential derived from side-chain packing. J. Mol. Biol. 2017;429:3113–20. https://doi.org/10.1016/j.jmb.2017.08.013.

[81] Gang X, Tianqi M, Junqing D, Qinghua W, Jianpeng M. OPUS-Rota2: an improved fast and accurate side chain modeling method. J. Chem. Theory Comput. 2019. https://doi.org/10.1021/acs.jctc.9b00309.

[82] Leem J, Georges G, Shi J, Deane CM. Antibody side chain conformations are position-dependent. Proteins Struct. Funct. Bioinforma. 2018;86:383–92. https://doi.org/10.1002/prot.25453.

[83] Miao Z, Cao Y, Jiang T. RASP: rapid modeling of protein side chain conformations. Bioinformatics 2011;27:3117–22. https://doi.org/10.1093/bioinformatics/btr538.

[84] Joosten RP, Joosten K, Murshudov GN, Perrakis A. PDB-REDO: constructive validation, more than just looking for errors. Acta Crystallogr. Sect. D Biol. Crystallogr. 2012;68:484–96. https://doi.org/10.1107/S0907444911054515.

[85] Kirys T, Ruvinsky AM, Tuzikov AV, Vakser IA. Rotamer libraries and probabilities of transition between rotamers for the side chains in protein-protein binding. Proteins Struct. Funct. Bioinforma. 2012;80:2089–98. https://doi.org/10.1002/prot.24103.

[86] Pierce NA, Spriet JA, Desmet J, Mayo SL. Conformational splitting: a more powerful criterion for dead-end elimination. J. Comput. Chem. 2000;21:999–1009. https://doi.org/10.1002/1096-987X(200008)21:11<999::AID-JCC9>3.0.CO;2-A.

[87] Kingsford CL, Chazelle B, Singh M. Solving and analyzing side-chain positioning problems using linear and integer programming. Bioinformatics 2005;21:1028–36. https://doi.org/10.1093/bioinformatics/bti144.

[88] Jacobson MP, Friesner RA, Xiang Z, Honig B. On the role of the crystal environment in determining protein side-chain conformations. J. Mol. Biol. 2002;320:597–608. https://doi.org/10.1016/S0022-2836(02)00470-9.

[89] Liang S, Grishin NV. Side-chain modeling with an optimized scoring function. Protein Sci. 2002;11:322–31. https://doi.org/10.1110/ps.24902.

[90] Eyal E, Najmanovich R, Mcconkey BJ, Edelman M, Sobolev V. Importance of solvent accessibility and contact surfaces in modeling side-chain conformations in proteins. J. Comput. Chem. 2004;25:712–24. https://doi.org/10.1002/jcc.10420.

[91] Lopes A, Alexandrov A, Bathelt C, Archontis G, Simonson T. Computational sidechain placement and protein mutagenesis with implicit solvent models. Proteins Struct. Funct. Genet. 2007;67:853–67. https://doi.org/10.1002/prot.21379.

[92] Peterson RW, Dutton PL, Wand AJ. Improved side-chain prediction accuracy using an ab initio potential energy function and a very large rotamer library. Protein Sci. 2004;13(3):735–51. https://doi.org/10.1110/ps.03250104.

[93] Wang C, Schueler-Furman O, Baker D. Improved side-chain modeling for protein-protein docking. Protein Sci. 2005;14:1328–39. https://doi.org/10.1110/ps.041222905.

[94] Jain T. Configurational-bias sampling technique for predicting side-chain conformations in proteins. Protein Sci. 2006;15:2029–39. https://doi.org/10.1110/ps.062165906.
[95] Colbes J, Corona RI, Lezcano C, Rodríguez D, Brizuela CA. Protein side-chain packing problem: is there still room for improvement?. Brief. Bioinform. 2017;18:1033–43. https://doi.org/10.1093/bib/bbw079.
[96] Lu M, Dousis AD, Ma J. OPUS-Rota: a fast and accurate method for side-chain modeling. Protein Sci. 2008;17:1576–85. https://doi.org/10.1110/ps.035022.108.
[97] Liang S, Zheng D, Zhang C, Standley DM. Fast and accurate prediction of protein side-chain conformations. Bioinformatics 2011;27:2913–4. https://doi.org/10.1093/bioinformatics/btr482.
[98] Xiang Z, Honig B. Extending the accuracy limits of prediction for side-chain conformations. J. Mol. Biol. 2001;311:421–30. https://doi.org/10.1006/jmbi.2001.4865.
[99] Nagata K, Randall A, Baldi P. SIDEpro: a novel machine learning approach for the fast and accurate prediction of side-chain conformations. Proteins Struct. Funct. Bioinforma. 2012;80:142–53. https://doi.org/10.1002/prot.23170.
[100] Jumper JM, Faruk NF, Freed KF, Sosnick TR. Accurate calculation of side chain packing and free energy with applications to protein molecular dynamics. PLoS Comput. Biol. 2018;14:e1006342. https://doi.org/10.1371/journal.pcbi.1006342.
[101] Liang S, Zhang C, Standley DM. Protein loop selection using orientation-dependent force fields derived by parameter optimization. Proteins Struct. Funct. Bioinforma. 2011;79:2260–7. https://doi.org/10.1002/prot.23051.
[102] Xu G, Ma T, Zang T, Wang Q, Ma J. OPUS-CSF: a C-atom-based scoring function for ranking protein structural models. Protein Sci. 2018;27:286–92. https://doi.org/10.1002/pro.3327.
[103] Caldararu O, Manzoni F, Oksanen E, Logan DT, Ryde U. Refinement of protein structures using a combination of quantum-mechanical calculations with neutron and X-ray crystallographic data. Acta Crystallogr. Sect. D Struct. Biol. 2019;D75:368–80. https://doi.org/10.1107/S205979831900175X.
[104] Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera - a visualization system for exploratory research and analysis. J. Comput. Chem. 2004;25:1605–12. https://doi.org/10.1002/jcc.20084.
[105] Guex N, Peitsch MC. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 1997;18:2714–23. https://doi.org/10.1002/elps.1150181505.
[106] Bayden AS, Fornabaio M, Scarsdale JN, Kellogg GE. Web application for studying the free energy of binding and protonation states of protein-ligand complexes based on HINT. J. Comput. Aided. Mol. Des. 2009;23:621–32. https://doi.org/10.1007/s10822-009-9270-7.
[107] Li Y, Roy A, Zhang Y. HAAD: a quick algorithm for accurate prediction of hydrogen atoms in protein structures. PLoS One 2009;4:e6701. https://doi.org/10.1371/journal.pone.0006701.
[108] Georgescu RE, Alexov EG, Gunner MR. Combining conformational flexibility and continuum electrostatics for calculating pKas in proteins. Biophys. J. 2002;83:1731–48. https://doi.org/10.1016/S0006-3495(02)73940-4.
[109] Labute P. Protonate3D: assignment of ionization states and hydrogen coordinates to macromolecular structures. Proteins Struct. Funct. Bioinforma. 2009;75:187–205. https://doi.org/10.1002/prot.22234.
[110] Lippert T, Rarey M. Fast automated placement of polar hydrogen atoms in protein-ligand complexes. J. Cheminform. 2009;1:13. https://doi.org/10.1186/1758-2946-1-13.
[111] Word JM, Lovell SC, Richardson JS, Richardson DC. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. J. Mol. Biol. 1999;285:1735–47. https://doi.org/10.1006/jmbi.1998.2401.
[112] Vriend G. WHAT IF: a molecular modeling and drug design program. J. Mol. Graph. 1990;8:52–6. https://doi.org/10.1016/0263-7855(90)80070-V.
[113] Brünger AT, Adams PD, Clore GM, Delano WL, Gros P, Grossekunstleve RW, et al. Crystallography & NMR system: a new software suite for macromolecular structure determination. Acta Crystallogr. Sect. D Biol. Crystallogr. 1998;54:905–21. https://doi.org/10.1107/S0907444998003254.
[114] Lindahl E, Hess B, van der Spoel D. GROMACS 3.0: a package for molecular simulation and trajectory analysis. J. Mol. Model. 2001;7:306–17. https://doi.org/10.1007/S008940100045.
[115] Brünger AT, Karplus M. Polar hydrogen positions in proteins: empirical energy placement and neutron diffraction comparison. Proteins Struct. Funct. Bioinforma. 1988;4:148–56. https://doi.org/10.1002/prot.340040208.
[116] Forrest LR, Honig B. An assessment of the accuracy of methods for predicting hydrogen positions in protein structures. Proteins Struct. Funct. Genet. 2005;61:296–309. https://doi.org/10.1002/prot.20601.
[117] Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. J. Comput. Chem. 1983;4:187–217. https://doi.org/10.1002/jcc.540040211.
[118] Gront D, Kmiecik S, Blaszczyk M, Ekonomiuk D, Koliński A. Optimization of protein models. Wiley Interdiscip. Rev. Comput. Mol. Sci. 2012;2:479–93. https://doi.org/10.1002/wcms.1090.
[119] Kmiecik S, Gront D, Kolinski A. Towards the high-resolution protein structure prediction. Fast refinement of reduced models with all-atom force field. BMC Struct. Biol. 2007;7:43. https://doi.org/10.1186/1472-6807-7-43.
[120] Heo L, Feig M. What makes it difficult to refine protein models further via molecular dynamics simulations?. Proteins Struct. Funct. Bioinforma. 2018;86:177–88. https://doi.org/10.1002/prot.25393.
[121] Feig M, Mirjalili V. Protein structure refinement via molecular-dynamics simulations: what works and what does not?. Proteins 2016;84:282–92. https://doi.org/10.1002/prot.24871.
[122] Hovan L, Oleinikovas V, Yalinca H, Kryshtafovych A, Saladino G, Gervasio FL. Assessment of the model refinement category in CASP12. Proteins Struct. Funct. Bioinforma. 2018;86:152–67. https://doi.org/10.1002/prot.25409.
[123] Chopra G, Kalisman N, Levitt M. Consistent refinement of submitted models at CASP using a knowledge-based potential. Proteins Struct. Funct. Bioinforma. 2010;78:2668–78. https://doi.org/10.1002/prot.22781.
[124] Lin X, Schafer NP, Lu W, Jin S, Chen X, Chen M, et al. Forging tools for refining predicted protein structures. Proc. Natl. Acad. Sci. 2019;116:9400–9. https://doi.org/10.1073/pnas.1900778116.
[125] Bhattacharya D. refineD: improved protein structure refinement using machine learning based restrained relaxation. Bioinformatics 2019;35(18):3320–8. https://doi.org/10.1093/bioinformatics/btz101.
[126] Jung J, Mori T, Kobayashi C, Matsunaga Y, Yoda T, Feig M, et al. GENESIS: a hybrid-parallel and multi-scale molecular dynamics simulator with enhanced sampling algorithms for biomolecular and cellular simulations. Wiley Interdiscip. Rev. Comput. Mol. Sci. 2015;5:310–23. https://doi.org/10.1002/wcms.1220.
[127] Miao Y, McCammon JA. Unconstrained enhanced sampling for free energy calculations of biomolecules: a review. Mol. Simul. 2016;42:1046–55. https://doi.org/10.1080/08927022.2015.1121541.
[128] Maximova T, Moffatt R, Ma B, Nussinov R, Shehu A. Principles and overview of sampling methods for modeling macromolecular structure and dynamics. PLoS Comput. Biol. 2016;12:e1004619. https://doi.org/10.1371/journal.pcbi.1004619.
[129] Kolinski A. Toward more efficient simulations of slow processes in large biomolecular systems: comment on "ligand diffusion in proteins via enhanced sampling in molecular dynamics" by Jakub Rydzewski and Wieslaw Nowak. Phys. Life Rev. 2017;22–23:75–6. https://doi.org/10.1016/j.plrev.2017.07.003.
[130] Schmiedel JM, Lehner B. Determining protein structures using deep mutagenesis. Nat. Genet. 2019;51:1177–86. https://doi.org/10.1038/s41588-019-0431-x.
[131] Rollins NJ, Brock KP, Poelwijk FJ, Stiffler MA, Gauthier NP, Sander C, et al. Inferring protein 3D structure from deep mutation scans. Nat. Genet. 2019;51:1170–6. https://doi.org/10.1038/s41588-019-0432-9.
[132] Wassenaar TA, Pluhackova K, Böckmann RA, Marrink SJ, Tieleman DP. Going backward: a flexible geometric approach to reverse transformation from coarse grained to atomistic models. J. Chem. Theory Comput. 2014;10:676–90. https://doi.org/10.1021/ct400617g.
[133] Stansfeld PJ, Sansom MSP. From coarse grained to atomistic: a serial multiscale approach to membrane protein simulations. J. Chem. Theory Comput. 2011;7:1157–66. https://doi.org/10.1021/ct100569y.
[134] Shimizu M, Takada S. reconstruction of atomistic structures from coarse-grained models for protein-DNA complexes. J. Chem. Theory Comput. 2018;14:1682–94. https://doi.org/10.1021/acs.jctc.7b00954.
[135] Jarmolinska AI, Kadlof M, Dabrowski-Tumanski P, Sulkowska JI. GapRepairer: a server to model a structural gap and validate it using topological analysis. Bioinformatics 2018;34:3300–7. https://doi.org/10.1093/bioinformatics/bty334.
[136] Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. The I-TASSER suite: protein structure and function prediction. Nat. Methods 2015;12:7–8. https://doi.org/10.1038/nmeth.3213.
[137] Fleishman SJ, Leaver-Fay A, Corn JE, Strauch EM, Khare SD, Koga N, et al. Rosettascripts: a scripting language interface to the Rosetta Macromolecular modeling suite. PLoS One 2011;6:e20161. https://doi.org/10.1371/journal.pone.0020161.
[138] Boniecki M, Rotkiewicz P, Skolnick J, Kolinski A. Protein fragment reconstruction using various modeling techniques. J. Comput. Aided. Mol. Des. 2003;17:725–38. https://doi.org/10.1023/B:JCAM.0000017486.83645.a0.
[139] Skwark MJ, Abdel-Rehim A, Elofsson A. PconsC: combination of direct information methods and alignments improves contact prediction. Bioinformatics 2013;29:1815–6. https://doi.org/10.1093/bioinformatics/btt259.
[140] Rohl CA, Strauss CEM, Misura KMS, Baker D. Protein structure prediction using rosetta. Methods Enzymol. 2004;383:66–93. https://doi.org/10.1016/S0076-6879(04)83004-0.
[141] Milik M, Kolinski A, Skolnick J. Algorithm for rapid reconstruction of protein backbone from alpha carbon coordinates. J. Comput. Chem. 1997;18:80–5. https://doi.org/10.1002/(SICI)1096-987X(19970115)18:1<80::AID-JCC8>3.0.CO;2-W.
[142] Vanhee P, Verschueren E, Baeten L, Stricher F, Serrano L, Rousseau F, et al. BriX: a database of protein building blocks for structural analysis, modeling and design. Nucleic Acids Res. 2010;39:D435–42. https://doi.org/10.1093/nar/gkq972.
[143] Kolodny R, Koehl P, Guibas L, Levitt M. Small libraries of protein fragments model native protein structures accurately. J. Mol. Biol. 2002;323:297–307. https://doi.org/10.1016/S0022-2836(02)00942-7.
[144] Vanhee P, Stricher F, Baeten L, Verschueren E, Lenaerts T, Serrano L, et al. Protein-peptide interactions adopt the same structural motifs as monomeric protein folds. Structure 2009;17:1128–36. https://doi.org/10.1016/j.str.2009.06.013.

[145] Dunbrack RL, Cohen FE. Bayesian statistical analysis of protein side-chain rotamer preferences. Protein Sci. 1997;6:1661–81. https://doi.org/10.1002/pro.5560060807.
[146] Kleywegt GJ, Harris MR, Zou JY, Taylor TC, Wählby A, Jones TA. The uppsala electron-density server. Acta Crystallogr. Sect. D Biol. Crystallogr. 2004;60:2240–9. https://doi.org/10.1107/S0907444904013253.
[147] Di Maio F, Tyka MD, Baker ML, Chiu W, Baker D. Refinement of protein structures into low-resolution density maps using rosetta. J. Mol. Biol. 2009;392:181–90. https://doi.org/10.1016/j.jmb.2009.07.008.
[148] Badaczewska-Dawid AE, Khramushin A, Schueler-Furman O, Kolinski A, Kmiecik S. Protocols for all-atom reconstruction and high-resolution refinement of protein-peptide complex structures. Methods Mol. Biol. 2020. In press.
[149] Kurcinski M, Ciemny MP, Oleniecki T, Kuriata A, Badaczewska-Dawid AE, Kolinski A, et al. CABS-dock standalone: a toolbox for flexible protein-peptide docking. Bioinformatics 2019;35(20):4170–2. https://doi.org/10.1093/bioinformatics/btz185.
[150] Kurcinski M, Oleniecki T, Ciemny PM, Kuriata A, Kolinski A, Kmiecik S. CABS-flex standalone: a simulation environment for fast modeling of protein flexibility. Bioinformatics 2019;35(4):694–5. https://doi.org/10.1093/bioinformatics/bty685.
[151] Brünger AT. X-PLOR Version 3.1: a system for X-ray crystallography and NMR. London: Yale University Press; 1992. ISBN 9780300054026.
[152] Song Y, Mao J, Gunner MR. MCCE2: improving protein pKa calculations with extensive side chain rotamer sampling. J. Comput. Chem. 2009;30:2231–47. https://doi.org/10.1002/jcc.21222.
[153] Reulecke I, Lange G, Albrecht J, Klein R, Rarey M. Towards an integrated description of hydrogen bonding and dehydration: decreasing false positives in virtual screening with the HYDE scoring function. ChemMedChem 2008;3:885–97. https://doi.org/10.1002/cmdc.200700319.
[154] Bietz S, Urbaczek S, Schulz B, Rarey M. Protoss: a holistic approach to predict tautomers and protonation states in protein-ligand complexes. J. Cheminform. 2014;6:12. https://doi.org/10.1186/1758-2946-6-12.
[155] Chen VB, Arendall WB, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, et al. MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr. Sect. D Biol. Crystallogr. 2010;D66:12–21. https://doi.org/10.1107/S0907444909042073.
[156] Páll S, Abraham MJ, Kutzner C, Hess B, Lindahl E. Tackling exascale software challenges in molecular dynamics simulations with GROMACS. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); 2015; pp. 3–27.
[157] Kmiecik S, Kouza M, Badaczewska-Dawid AE, Kloczkowski A, Kolinski A. Modeling of protein structural flexibility and large-scale dynamics: coarse-grained simulations and elastic network models. Int. J. Mol. Sci. 2018;19(11):3496. https://doi.org/10.3390/ijms19113496.
[158] Ciemny MP, Badaczewska-Dawid AE, Pikuzinska M, Kolinski A, Kmiecik S. Modeling of disordered protein structures using monte carlo simulations and knowledge-based statistical force fields. Int. J. Mol. Sci. 2019;20(3):606. https://doi.org/10.3390/ijms20030606.
[159] Zhang J, Liang Y, Zhang Y. Atomic-level protein structure refinement using fragment-guided molecular dynamics conformation sampling. Structure 2011;19(12):1784–95. https://doi.org/10.1016/j.str.2011.09.022.
[160] Heo L, Park H, Seok C. GalaxyRefine: Protein structure refinement driven by side-chain repacking. Nucleic Acids Res. 2013;41:W384–8. https://doi.org/10.1093/nar/gkt458.