BIOCHEMISTRY
including biophysical chemistry & molecular biology

Current Topic

pubs.acs.org/biochemistry

Protocols for Molecular Modeling with Rosetta3 and RosettaScripts

Brian J. Bender,[†,‡] Alberto Cisneros, III,[‡,§] Amanda M. Duran,[‡,∥] Jessica A. Finn,[‡,⊥] Darwin Fu,[‡,∥] Alyssa D. Lokits,[‡,#] Benjamin K. Mueller,[‡,∥] Amandeep K. Sangha,[‡,∥] Marion F. Sauer,[‡,§] Alexander M. Sevy,[‡,§] Gregory Sliwoski,[‡,∥] Jonathan H. Sheehan,[‡] Frank DiMaio,[@] Jens Meiler,[†,‡,§,∥,⊥,#] and Rocco Moretti*[,‡,∥]

[†]Department of Pharmacology, Vanderbilt University, Nashville, Tennessee 37232-6600, United States
[‡]Center for Structural Biology, Vanderbilt University, Nashville, Tennessee 37240-7917, United States
[§]Chemical and Physical Biology Program, Vanderbilt University, Nashville, Tennessee 37232-0301, United States
[∥]Department of Chemistry, Vanderbilt University, Nashville, Tennessee 37235, United States
[⊥]Department of Pathology, Microbiology and Immunology, Vanderbilt University, Nashville, Tennessee 37232-2561, United States
[#]Neuroscience Program, Vanderbilt University, Nashville, Tennessee 37235, United States
[@]Department of Biochemistry, University of Washington, Seattle, Washington 98195, United States

Supporting Information

ABSTRACT: Previously, we published an article providing an overview of the Rosetta suite of biomacromolecular modeling software and a series of step-by-step tutorials [Kaufmann, K. W., et al. (2010) *Biochemistry 49*, 2987−2998]. The overwhelming positive response to this publication we received motivates us to here share the next iteration of these tutorials that feature *de novo* folding, comparative modeling, loop construction, protein docking, small molecule docking, and protein design. This updated and expanded set of tutorials is needed, as since 2010 Rosetta has been fully redesigned into an object-oriented protein modeling program Rosetta3. Notable improvements include a substantially improved energy function, an XML-like language termed "RosettaScripts" for flexibly specifying modeling task, new analysis tools, the addition of the TopologyBroker to control conformational sampling, and support for multiple templates in comparative modeling. Rosetta's ability to model systems with symmetric proteins, membrane proteins, noncanonical amino acids, and RNA has also been greatly expanded and improved.

Obtaining atomic-detail accurate models for all proteins, natural and engineered, in all relevant functional states, alone and in complex with all relevant interaction partners by crystallography or nuclear magnetic resonance (NMR) is impaired by the vast number of possible protein sequences and interactions. In some cases, it is complicated by experimental obstacles and is often time and cost intensive. Additional difficulties arise when the dynamic properties of proteins and their interactions with other molecules are to be studied from crystallographic snapshots. Here, computational modeling of the structure and dynamics of proteins and interactions can complement experimental techniques. Such computational models add atomic detail not present in low-resolution or limited experimental data, model states that are not tractable for experimental structure determination, simulate conformational flexibility and plasticity of states, and prioritize states for crystallization or study with other experimental techniques.

At the same time, prediction and design of protein structure *in silico* is a formidable task: the need to model thousands of atoms instantiates the sampling challenge of testing a large number of possible arrangements or conformations. The need to complete these calculations in a finite time creates the scoring challenge of developing an energy function that is rapid but still accurately identifies biologically relevant, low-free energy states.

The Rosetta software suite represents a compilation of computational tools aimed at obtaining physically relevant structural models of proteins and their interactions with other proteins, small molecules, RNA, and DNA. Rosetta has contributed to the advancement of structural biology by tackling challenges in *de novo* protein design,[1−3] comparative modeling,[4,5] protein design,[6−11] protein−protein docking,[12−15] and protein−small molecule docking.[16−18] Additionally, Rosetta can be applied to RNA/DNA structure prediction,[19,20] the incorporation of noncanonical amino acids,[21,22] and other difficult structural challenges such as membrane protein structure prediction[23] and modeling of symmetric proteins.[24,25]

Rosetta developers follow the hypothesis that a single, unified energy function should be able to accomplish all of these complex tasks; furthermore, the continuous optimization of this

Received: May 5, 2016
Revised: July 29, 2016
Published: August 4, 2016

DOI: 10.1021/acs.biochem.6b00444
Biochemistry 2016, 55, 4748−4763

Table 1. Publically Accessible Web Servers Running Rosetta[a]

server	address	protocols offered
ROSIE	rosie.rosettacommons.org	many, including small molecule docking, protein design, RNA design, etc.[46]
Robetta	robetta.bakerlab.org	structure prediction[51]
Rosetta.design	rosettadesign.med.unc.edu	protein design[104]
FlexPepDock	flexpepdock.furmanlab.cs.huji.ac.il	flexible peptide docking[105]
RosettaBackrub	kortemmelab.ucsf.edu/backrub	backbone remodeling and design[106]
FunHunt	funhunt.furmanlab.cs.huji.ac.il	classification of protein−protein complex interactions[107]
CS-Rosetta	csrosetta.bmrb.wisc.edu	structure prediction based on chemical shift data
RosettaDiagrams	rosettadiagrams.org	setup protocols through visual diagrams

[a]All web servers listed are free for noncommercial use.

energy function to improve one structural problem will ultimately improve performance for other modeling tasks. Important components of the energy function are statistically derived, i.e., using protein models derived from high-resolution crystallographic data in the Protein Data Bank (PDB) as a knowledge base.[1,6,16,23,26−35] For speed, the energy function is pairwise decomposable and employs a distance cutoff. For many sampling tasks, Rosetta employs a Monte Carlo search steered by the Metropolis criterion (MCM).[27] Rosetta is continually developed and rigorously tested by a consortium of international academic laboratories known as the RosettaCommons (www.rosettacommons.org). Herein, we present a global review of generalized Rosetta protocols and applications, as well as descriptions of novel functionalities recently introduced.[26,36,37]

Detailed tutorials and examples are included as Supporting Information. The tutorials herein supersede our previous tutorials put forward in "Practically Useful: What the Rosetta Protein Modeling Suite Can Do for You".[38]

MAKING ROSETTA ACCESSIBLE

Rosetta is extremely powerful for many applications in structural biology, but for many years, it was limited by the fact that users needed an extensive background in C++ and the Unix environment to be able to construct new protocols. An ongoing effort by many groups has been taken to eliminate these boundaries, allowing greater flexibility and ease of use for the novice and intermediate user. These updates include customizable protocols using XML or Python. The updates using XML (RosettaScripts)[36] or Python (PyRosetta)[39] allow users to customize protocols without learning C++, by combining prewritten Rosetta objects and defining their behavior without having to write and recompile new C++ code. In addition, the Rosetta community now offers multiple web interfaces for application-specific tasks.

Other tools have been added, not to run Rosetta but to improve users' experience, such as graphical user interfaces (GUIs) to visualize Rosetta operations and to generate input files,[40] and PyMOL integration for real-time molecular visualization.[41] These tools offer users intuitive control over structural modeling without sacrificing flexibility and power.

RosettaScripts. RosettaScripts is an XML-like language for specifying modeling protocols through the Rosetta framework.[36] It allows users to define a set of Rosetta objects and execute them in a defined order to develop full protocols. Rosetta objects in RosettaScripts fall under four main categories: Movers, which are objects that modify a structure in some way; Filters, which evaluate properties of a structure; TaskOperations, which control the degrees of freedom of Rosetta's side-chain placement routines; and ScoreFunctions, which evaluate the energy of a structure. By combining these four elements, users are able to leverage many different sampling and scoring algorithms, with fine control over sampling degrees of freedom and protocol flow. All objects defined under these categories are customizable, which is a distinct advantage of RosettaScripts over conventional command line applications. For example, a user can define multiple score functions to be used in different sections of a protocol and then combine several protocols into a single XML protocol (i.e., protein−protein docking and design). This flexibility has made a number of scientific advances possible, such as *de novo* design of an influenza binder,[10] protein−protein docking based on hybrid structural methods,[42] and HIV vaccine design.[43]

PyRosetta. Because of the popularity of Python as a programming language in the computational biology community, a Python-based implementation of Rosetta was developed, termed PyRosetta.[39] PyRosetta consists of Python bindings for the major functions and objects of Rosetta, allowing all of these objects to be run from a Python environment. One advantage is the ability to combine Rosetta protocols with other popular structural biology software, such as PyMOL[44] and BioPython.[45] PyRosetta includes access to the same set of Rosetta objects for sampling and scoring that are described above for RosettaScripts, as well as many others. Unlike RosettaScripts, PyRosetta can be run in either script mode or interactive mode. Interactive mode allows the user to inspect their objects in real time while prototyping a new protocol.[39] Notably, PyRosetta is available for Windows in addition to Linux and Mac OSX, expanding the availability of Rosetta to researchers who use a Windows environment.

Web Interfaces. We are aware of eight Web servers that have been created to allow nonexperts to make use of Rosetta's functionality (Table 1). These Web servers allow Rosetta to be used with almost no learning curve, making the boundary to entry even lower than that of the scripting protocols mentioned above. In particular, ROSIE [the Rosetta Online Server that Includes Everyone (http://rosie.rosettacommons.org)][46] has been set up to easily provide a web interface for new Rosetta protocols.

Other Tools. Since the publication of RosettaScripts and PyRosetta, new tools have been developed to make running a Rosetta protocol even more intuitive. An interface to PyMOL was developed by Baugh et al., which allows users to visualize their molecules being manipulated by Rosetta as the protocol is being run.[41] While the viewer was originally developed for use with PyRosetta, it has since been extended for RosettaScripts. This visualization tool is especially useful for new users with experience in structural biology but new to computation.

DOI: 10.1021/acs.biochem.6b00444
Biochemistry 2016, 55, 4748−4763

Table 2. Standard Rosetta Score Function Terms

score term	definition
low-resolution scoring terms	
env	hydrophobicity term for each amino acid
vdw	steric repulsion between two residues
pair	probability of two residues interacting
rg	radius of gyration
cbeta	solvation term based on a number of surrounding residues
hs_pair, ss_pair, and sheet	secondary structure terms
high-resolution scoring terms (talaris2014)	
fa_atr, fa_rep, and fa_intra_rep	decomposed 6−12 Lennard-Jones potential
fa_sol	EEF1 solvation term
pro_close	proline ring closure energy
omega	omega backbone dihedral potential
dslf_fa13	updated disulfide geometry potential
rama	potential of ϕ and ψ angles for each amino acid
p_aa_pp	probability of an amino acid given a set of ϕ and ψ angles
fa_dun	rotamer likelihood
hbond_sr_bb, hbond_lr_bb, hbond_bb_sc, and hbond_sc	combined covalent−electrostatic hydrogen bond potentials for α-helices, β-sheets, side-chain backbone, and side-chain−side-chain interactions, respectively
yhh_planarity	tyrosine hydroxyl out-of-plane penalty
fa_elec	Coulombic electrostatic potential between two residues with a distance-dependent dielectric (deprecates fa_pair)

In addition, several GUIs for Rosetta have been developed to eliminate the need to run Rosetta exclusively through the Unix command line.[40] The PyRosetta Toolkit was developed to serve as a GUI for running PyRosetta, with menus to guide the user through the relevant Rosetta options that are needed for a protocol.[47] InteractiveRosetta is a GUI for running Rosetta protocols with an integrated molecular visualization window and user-friendly controls for implementing common Rosetta protocols.[40] Through these GUIs, users can generate input files for Rosetta protocols using a "point and click" interface while also running protocols seamlessly in the same window.

SAMPLING AND SCORING IN ROSETTA

Rosetta Sampling. While the approaches used by different protocols vary, in general Rosetta utilizes a Monte Carlo Metropolis sampling algorithm to quickly and efficiently determine the quality of structural trajectories. Rosetta further differentiates between sampling backbone and side-chain conformations within two separate refinement tasks. In addition, backbone sampling can be performed on a global or local scale. Large-scale backbone sampling utilizes 3-mer and 9-mer fragments derived from the Protein Data Bank (PDB), while local refinements of the backbone optimize ϕ and ψ angles without disturbing the global fold. Side-chain sampling also utilizes information derived from the PDB to create a "rotamer" library of observed conformations to reduce the conformational search space. For a more detailed discussion of Rosetta sampling, see ref 27.

Rosetta Scoring. The Rosetta score, or energy function, is a linear, weighted sum of terms combining knowledge- and physics-based potentials gathered from protein structural features within the PDB. The score function is used during Rosetta modeling to evaluate Monte Carlo sampling and for scoring the final output pose. With the implementation of Rosetta3, the score function is treated as a separate entity such that it can be repeatedly called and rapidly processed in a manner independent of the protocol at hand.[26] Additionally, score terms are grouped into a hierarchy based on potentials related to one entity (i.e., χ-angle probability), two interacting entities (i.e., hydrogen bonding potential), and terms that require the analysis of the entire model (i.e., radius of gyration).

Low- versus High-Resolution Scoring. In low-resolution scoring, or "centoid" mode, the side chain of each residue is removed and represented instead as a super atom ("centroid"), at a position that roughly approximates the center of mass of that side chain, averaged across likely side-chain states (or at the Cα atom for glycine). This greatly reduces the degrees of freedom that must be sampled during low-resolution backbone movement while preserving chemical and structural features of a given residue. Typical low-resolution sampling involves replacement of the backbone conformation with peptide fragments three and nine amino acids in length that are derived from the PDB. Peptide fragments are generated from the primary sequence of the protein. Centroid-mode scoring and sampling are used during the initial stages of protein modeling where exhaustive searches of conformational space are performed such as *de novo* protein folding, loop building, and rigid-body protein−protein docking.[1,12,27,28] Common score terms used in centroid mode are listed in Table 2.

High-resolution scoring, or "full-atom" mode, allows for full representation of all atoms of each side chain. In full-atom mode, conformational sampling relies on evaluating side-chain rotamers (derived from the PDB) during a Metropolis Monte Carlo simulated annealing protocol to find the global minimum.[29] Full-atom scoring was originally developed for protein design but has seen several improvements throughout Rosetta's history to the current *talaris2014* score function.[6,30−32,37] We have provided an additional example tutorial for the user on the basics of Rosetta scoring; see the scoring_and_prep folder in the Supporting Information.

Score Function Optimization. The score function is a linear weighted sum of energy terms; therefore, the weights can be parametrized to generate meaningful scores for predicted models. These are often fit against benchmark sets of modeling challenges to guide prediction of native structures. An algorithm "optE" was developed to streamline this weighting term optimization.[32] This algorithm excels at setting reference weights for amino acids. Using the approximation that a native,

DOI: 10.1021/acs.biochem.6b00444
Biochemistry 2016, 55, 4748−4763

evolved sequence is close to the optimal sequence for a structure,[30] optE attempts to find reference weights that minimize the divergence from native sequence profiles. Via optimization of the *talaris2014* score function for sequence recovery (~40%), performance in novel design tasks is also improved.

Like previous iterations of the full-atom score function, *talaris2014* sums separate physics- and knowledge-based potentials. It was found that combining physics- and knowledge-based information in a given score term led to improved Lennard-Jones and hydrogen bonding score terms.[31] The combined covalent−electrostatic hydrogen bonding terms were further updated with improved geometry and parametrization for sp^2-hybridized hydrogen bond acceptors.[37] Scoring potentials of knowledge-based score terms were smoothed with the use of bicubic-spline interpolation.[32] An updated rotamer library was included with an adaptive kernel formulation, which allows for smoother potentials of Ramachandran-based score terms.[33] Ideal atomic coordinates for amino acids, the geometry of disulfide bonds, and the hydroxyl sampling of serine and threonine residues were also expanded and improved. The free energies of solvation (LK_DGFREE) were updated to improve the EEF1 solvation energy potential of buried residues. Lastly, a new term that describes the Coulombic electrostatic potential between two residues with a distance-dependent dielectric (fa_elec) was introduced and replaces the previous statistics-based potential (fa_pair).[32] Further refinements were made to reduce the influence of the hydrogen bonding terms. This resulted in improved sequence recovery, rotamer recovery, and model discrimination.[37] As of writing, these updates culminated in the *talaris2014* score function, which is the default for current versions of Rosetta. All *talaris2014* score terms are listed in Table 2.

Continual optimization of the Rosetta score function means that the default score function varies with Rosetta version: *score12* for versions prior to Rosetta 3.5, *talaris2013* for weekly releases until 2016.10, and *talaris2014* for Rosetta 3.6 and weekly releases since 2016.11. Further score function refinement is ongoing, and it is likely that future Rosetta releases will have a different default score function. Additionally, while Rosetta strives to have a single all-atom score function to encompass all modeling tasks, several application-specific scoring potentials have been developed to include new score terms and optimized score term weights. These include, but are not limited to, modified score functions for small molecule docking,[16] protein−protein docking,[12,34] and membrane protein modeling,[23,35] as well as specialized score functions for low-resolution sampling stages.

Limitations and Caveats. Ongoing improvements made by the Rosetta community have led to increasingly accurate modeling protocols; however, there are still several hurdles that must be overcome for Rosetta to accurately produce nativelike models. First, Rosetta sampling is stochastic in nature. Therefore, not every modeling trajectory will sample a regional minimum on the score function. Second, the score function is heuristic and abbreviated for speed. It fails to fully recapitulate the fundamental forces. Therefore, minima of the energy function are not guaranteed to describe biologically relevant states. Third, even with its rapid score function, Rosetta is unable to exhaustively sample all possible structural space due to computational time restraints. Fourth, many Rosetta protocols are optimized for local resampling and require a starting model, which may not exist for some systems.

Evaluating Interfaces. Some biological applications of Rosetta focus on improving, creating, or otherwise altering a well-defined protein−protein, protein−small molecule, or protein−DNA interface. These protocols typically inhabit a much smaller search space and in some cases rely solely on rigid-body optimization to generate a desired interaction.[48] In these instances, a series of specific interactions is evaluated, and the widely used "score vs RMSD" plot (see Figure 5h for an example of a score vs RMSD plot) is repurposed to look at small changes at the interface; here, plotting the "interface score" against the "interface RMSD" prevents small, meaningful changes from being lost in the larger fluctuations when scoring the entire model or computing the RMSD over all atoms. Additionally, analytical tools like the Interface Analyzer provide a series of useful calculations that include binding energy,[49] shape complementarity, the number of buried, unsatisfied hydrogen bonds, and the solvent accessible area buried at the interface. These metrics can be used in conjunction with RosettaHoles[50] to generate a packing statistic score for the interface.

DE NOVO STRUCTURE PREDICTION

De novo protein structure prediction is one of the greatest remaining challenges in computational structural biology. This process models the tertiary structure of a protein from its primary amino acid sequence. Importantly, *de novo* modeling differs from template-based or comparative protein modeling in that structural predictions are not based upon a known homologous structure. To address the challenge of predicting a protein's structure *de novo*, Rosetta uses short peptide "fragments" to assemble a complete protein structure.

The Rosetta *de novo* protein folding algorithm continues to follow the steps described in our previous review.[38] Briefly, short peptide fragments of known protein structures are obtained from the PDB and are inserted into an extended-chain protein following a Monte Carlo strategy.[1] In that sense, Rosetta *de novo* protein folding is not truly *de novo*; it combines a very large number of small templates. The hypothesis is that while not every protein fold is yet represented in the PDB, the conformation of small peptide fragments is exhaustively sampled. These peptide fragments are used to alter the backbone conformation of the extended-chain protein, folding it toward a low-energy tertiary structure. The process is repeated to create an ensemble of models. Finally, these low-resolution models can be filtered on the basis of pass/fail criteria provided by the user. These models can be clustered, and an energy minimization step applied to refine an all-atom model with the high-resolution energy function.

Generating Peptide Fragments. *De novo* protein folding relies on the assembly of short peptide fragments, usually generated as a preprocessing step. First, the primary protein sequence is used to generate secondary structure predictions. Next, the sequence, secondary structure predictions, and NMR data (if available) are used to pick candidate three- and nine-amino acid fragments from the PDB. Finally, these candidate fragments are scored, and the best N fragments are written to a fragment library file. The ROBETTA Web server (http://robetta.bakerlab.org) is available for noncommercial use and allows users to generate fragment libraries using a simple interface.[51] Additionally, Gront et al. have developed the

DOI: 10.1021/acs.biochem.6b00444
Biochemistry 2016, 55, 4748−4763

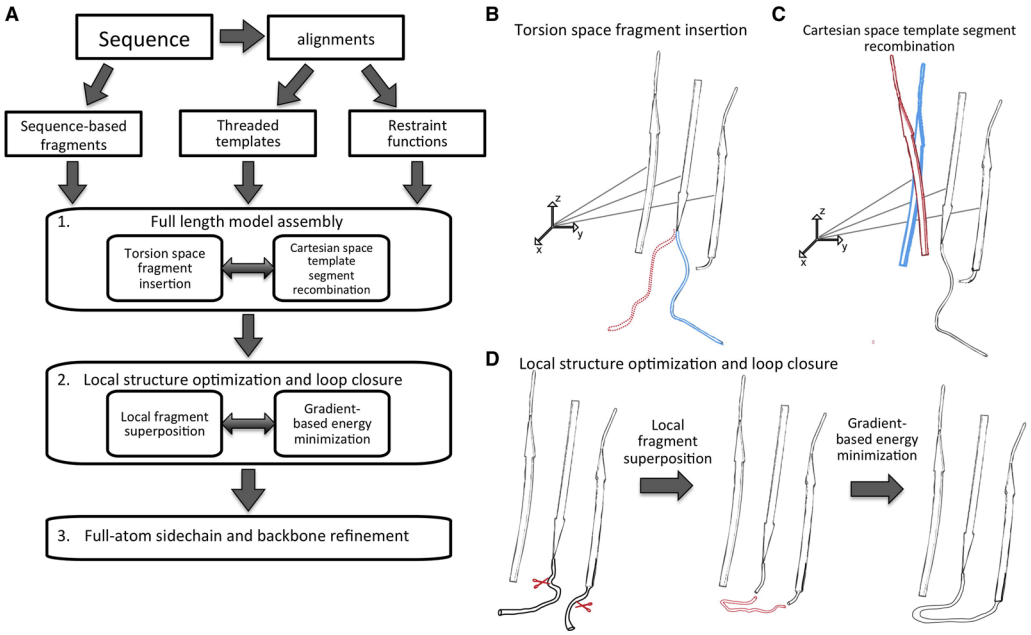

Figure 1. Multitemplate comparative modeling with Rosetta. (A) General workflow of the RosettaCM protocol. (B) Fragment insertion (blue, before insertion; red, after insertion). (C) Recombination of template segments. (D) Fragment insertion and minimization for loop closure. Reprinted with permission from ref 5. Copyright 2013 Elsevier.

FragmentPicker that provides users with total control over the fragment picking protocol.[52]

TopologyBroker. The TopologyBroker,[53] a tool that allows for more complex simulations, is an improvement added to Rosetta since our last review. The conformational space searched during a Rosetta *de novo* modeling simulation is vast, and successful searches often integrate prior knowledge with sampling. In *de novo* protein folding, this prior knowledge may be in the form of β-strand pairing constraints or the formation of a rigid chunk of the target fold based on a structurally homologous domain. Previously, protocol developers were restricted to a sequential sampling approach in which Rosetta could readily violate one set of these constraints while sampling to satisfy the other. The TopologyBroker was developed to create a consensus sampling approach that satisfies all of the requested constraints without requiring additional code development for each unique system; instead, the Broker provides an Application Program Interface (API) that allows for plug-and-play applications to generate complex sampling strategies.

Benchmarking *De Novo*. The *de novo* modeling capabilities of the object-oriented Rosetta software suite ("Rosetta3") were assessed in the CASP8 (Critical Assessment of protein Structure Prediction) experiment.[3] For 13 targets in the assessment, no homologous templates were identified and Rosetta's *de novo* modeling protocol was used to predict the structure of these targets. Following the observation that Rosetta *de novo* structural predictions are sometimes improved by using nonstandard fragment sizes, a range of fragment lengths were used when modeling the CASP8 targets. Longer fragment lengths were found to improve modeling of α-helical proteins, while shorter fragment lengths mainly improved modeling of β-strand proteins.

Limitations of *De Novo*. Because *de novo* structure prediction is such a powerful tool and yet such a complex challenge, it is critically important to understand the limitations of the algorithm. Rosetta performs well at folding small, globular, soluble proteins as well as small, simple membrane proteins containing 80−100 residues. However, large and complex proteins present additional difficulties that are not easily overcome by *de novo* techniques alone. Instead, users must incorporate other biochemical information to obtain nativelike models. Ongoing work shows that the incorporation of residue−residue co-evolution information can significantly improve the prediction accuracy during *de novo* modeling trails.[2] Other techniques such as homology modeling and using experimental constraints are discussed below.

Furthermore, because *de novo* structural prediction will sample many potential protein folds, it is necessary to generate large numbers of models (>10000) to adequately sample the conformational space. Extensive computational resources are needed to generate this number of models, and the use of distributed computational methods (such as computational clusters) is recommended. An example tutorial for the *de novo* prediction of a protein structure with Rosetta is included in the Supporting Information. This tutorial, protein_folding, provides an outline for a basic *de novo* protocol. Structural prediction of a soluble protein is described, both with and without the application of experimentally derived restraints. Also, a brief review of model analysis is covered. Instructions on how to run a membrane protein *de novo* protocol are included in a subfolder of the protein_folding tutorial.

■ COMPARATIVE MODELING

Comparative modeling differs from *de novo* methods in that it utilizes a known protein structure as the starting scaffold or template for structural prediction. If the template structure is a homologous protein, one speaks often of "homology modeling". Comparative modeling is a useful strategy for predicting protein structure and function when experimental methods fail or would be too resource intensive to employ. It increases the probability of obtaining realistic conformational predictions, especially when the target, or desired protein, is greater than 150 amino acids in length and/or adopts a

DOI: 10.1021/acs.biochem.6b00444
Biochemistry 2016, 55, 4748−4763

complex tertiary fold. However, it requires that a related, often homologous, structure has been determined experimentally; this is termed the template. Ideally, the sequence identity between the target and the template is >30%, although proteins with lower sequence identity may still be used for comparative modeling when their tertiary fold is conserved.

The latter case will be examined within the tutorial provided with the Supporting Information. This tutorial, rosetta_cm, outlines the basic steps necessary for comparative modeling in Rosetta. The tutorial focuses on the use of RosettaRelax and RosettaMembrane, as well as information for implementing basic restraints.

Over the past several years, comparative modeling in Rosetta has incorporated many improvements, specifically the use of multiple templates and a specific low-resolution scoring functions.[5] Previously published protocols of comparative modeling with Rosetta suggested using multiple templates to obtain diversity and flexibility.[4] However, models were built on individual templates. The new RosettaCM protocol allows for integration of multiple templates with *de novo* fragments into a single structural model of the protein.[5] Hence, this multitemplate, multistage protocol samples a broader structural landscape and can select well-scoring subtemplates for different regions of the protein to be modeled.

A highly detailed description of RosettaCM design, sampling, and scoring has previously been published.[5] Users are encouraged to refer to this work for a comprehensive assessment of RosettaCM applications, considerations, and caveats. Herein, we will briefly describe features of RosettaCM as they apply to the protocol presented.

Starting Templates. Before utilizing RosettaCM, starting templates must be identified through remote homologue detection methods such as PSIBLAST.[54] When homologues are not found using sequence-based methods, three-dimensional (3D) fold recognition software may be used to obtain suitable templates. As with other modeling software, RosettaCM performance improves with higher sequence similarity and identity.

Three Stages of Multitemplate Comparative Modeling. Multitemplate RosettaCM is a three-stage process in which the best scoring model from each stage is utilized as the input for the following step (Figure 1). The output of stage 1 is a full-length, assembled model that is generally correct in topology. However, segment boundaries where templates are mended can be suboptimal in geometry and energetically frustrated. To resolve these energetic frustrations and to explore the conformational space around this starting model, stage 2 of RosettaCM iteratively improves local environments through a series of fragment insertions, side-chain rotamer sampling, and gradient-based energy minimization of the entire structure using a RosettaCM-specific low-resolution energy function. The best model from this cycle is then moved to stage 3 for a final round of all-atom refinement that improves side-chain geometries, backbone conformations, and packing density before converging on a final output model.

Modeling Loops. In previous Rosetta comparative modeling protocols, a user-defined, "loop" closure step was required to remove chain breaks, reconcile long unstructured coils, or rebuild regions of low sequence similarity (all of which are defined as "loops" within the Rosetta framework). Two different algorithms are available: Cyclic Coordinate Descent (CCD) and Kinematic Loop Closure (KIC). Briefly, CCD quickly closes roughly 99% of loops utilizing a robotics-inspired iterative approach to manipulate dihedral angles of three residue backbone atoms between user-specified C-terminal and N-terminal anchor points. The second loop building algorithm, KIC, explicitly determines all possible combinations of torsion angles within the defined segment using polynomial resultants.[55] While being slower than CCD, KIC determines more accurate loop structures, provided the anchor points are optimally set. Both algorithms within Rosetta can be used in conjunction with fragments derived from the PDB to build regions of missing electron density, poor homology, or backbone gaps.

Unlike the single-template loop building application, comparative modeling with multiple templates closes chain breaks and rebuilds loops internally during stage 2. *De novo* fragment insertions are encouraged in regions of weak backbone geometry, while template-based fragment insertions anneal chain breaks and low-electron density regions. Additional smoothing occurs with the RosettaCM-specific scoring function. This internal step removes the need for additional loop closures by the user. However, it is encouraged for the user to critically examine all output models to validate structural accuracy.

PROTEIN–PROTEIN DOCKING

Determining the optimal binding orientation and interface of two or more protein binding partners has many biological and pharmaceutical applications, yet determining the structure of protein–protein complexes by biochemical techniques is slow and laborious. RosettaDock is a useful tool for computationally predicting protein–protein interactions by employing an algorithm that simulates a biophysical encounter of two or more binding partners and optimizes the conformation of the bound state. The RosettaDock algorithm includes a multiscale, Monte Carlo-based docking algorithm that begins with a centroid-mode stage to identify docking poses, followed by an all-atom refinement stage to optimize rigid-body position and side-chain conformations.[12]

Global versus Local Docking. The initial pose for docking is determined by either global docking or local perturbation. Global docking randomly orients one of the two binding partners in relation to the other to determine an initial binding interface. This is useful when there is no biological or structural evidence to suggest a starting pose. Local perturbation allows the user to define a general starting pose for the binding partners when prior experimental knowledge exists; this initial placement greatly decreases the conformational search space and improves the sampling density close to the starting pose, although this may bias models toward the starting conformation. The tutorial included in the protein–protein_docking folder illustrates one application of Rosetta protein–protein local docking by using two known binding positions of the CR6261 antibody to influenza antigen hemagglutinin (HA) subtypes H1 and H5.

Low-Resolution versus High-Resolution Docking. The full RosettaDock algorithm begins with low-resolution docking. The first step involves rigid-body movements of the binding partners that rotate and translate in relation to one another.[13] The score function is used to achieve a threshold acceptance rate of rigid-body moves.[12] A high-resolution docking mode follows in which the lowest-energy structures and/or largest clusters assessed from the centroid-mode stage are selected for high-resolution refinement. Centroid pseudoatoms are replaced

DOI: 10.1021/acs.biochem.6b00444
Biochemistry 2016, 55, 4748–4763

with all-atom side chains in their initial unbound conformations followed by additional fine-grained rigid-body docking.

Improvements to RosettaDock. The addition of RosettaScripts and PyRosetta to Rosetta now gives users the flexibility to modularize the centroid mode and all-atom mode of RosettaDock to suit case-specific applications. This was done by splitting RosettaDock into three major classes: DockingProtocol, DockingLowRes, and DockingHighRes.[13] The increase in flexibility showed an only marginal increase in successful predictions; however, it is particularly adept at predicting antibody−antigen complexes.[13] The modularization of RosettaDock has allowed users to also incorporate additional features within their docking protocols, including additional parameters for nonprotein moieties and protonation states,[49,56,57] flexible peptide-chain docking using FlexPepDock,[14] and *de novo* peptide docking.[15]

FlexPepDock. The FlexPepDock *de novo* docking algorithm is similar to the RosettaDock algorithm in that it begins with sampling rigid-body moves from the initial protein−peptide complex. Although not included in the tutorial, this step also includes iterative peptide fragment insertions and random moves of the peptide backbone using decreasing simulated temperature weights. Next, the low-resolution model is improved using an all-atom refinement stage by peptide side-chain placement optimization using a Monte Carlo search of "small" and "shear" moves described by Rohl et al.[1] Each round of refinement also includes a decreasing repulsive van der Waals weight term and an increasing attractive van der Waals term to allow a greater degree of perturbation within the binding pocket without causing the peptide and protein to separate during energy minimization. The FlexPepDock *de novo* benchmark demonstrated that the protocol produces near-native models with 86% accuracy (Figure 2).[15]

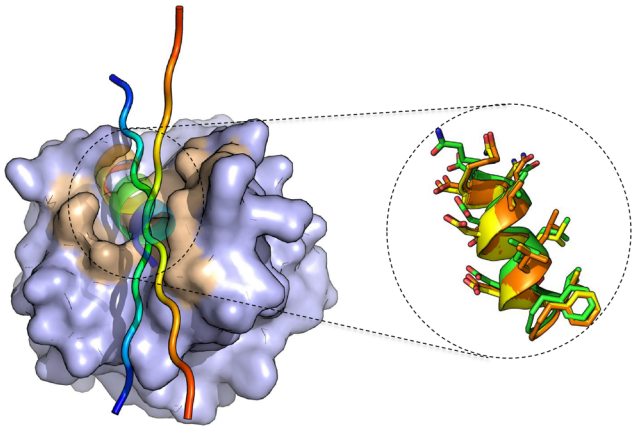

Figure 2. Protein−peptide interface prediction using FlexPepDock *ab initio*. Structure prediction of the Che-Z-derived peptide bound to CheY (PDB entry 2FMF) from two opposite starting orientations converges onto the same final conformation resembling the structure of the native peptide. The left panel is a general view of the CheY receptor (gray; interface residues colored light brown), the two initial, extended peptide conformations (rainbow cartoons), and the final helical peptide conformation (rainbow, transparent cartoon). The right panel is a detailed atomic view of the top FlexPepDock *ab initio* predictions from two simulations (yellow and orange) and the native peptide conformation (green). Reprinted from ref 15.

PROTEIN−SMALL MOLECULE DOCKING

Protein−small molecule docking aims to capture the binding interactions between a protein and a small molecule. This includes recapitulating the binding pose and quantifying the interaction strength. RosettaLigand, Rosetta's protein−small molecule docking protocol, is designed to consider both protein and small molecule flexibility.[16,17] It uses a two-phase docking approach similar to Rosetta's protein−protein docking: a low-resolution phase of rapid sampling based on shape complementarity followed by a high-resolution phase of Monte Carlo minimization of side-chain rotamers and small molecule conformers. The models undergo a final gradient minimization of the protein and molecule torsion degrees of freedom before they are output along with an interface score as a proxy for binding free energy. A small molecule docking tutorial (ligand_docking) included in the Supporting Information demonstrates this optimized protocol.

Improvements to RosettaLigand. In contrast to the previously published RosettaLigand protocol,[38,58] this tutorial replaces the independent translation/rotation low-resolution sampling steps with the new Transform algorithm.[18] The Transform algorithm couples translational, rotational, and conformational sampling into a single Monte Carlo process. In a benchmark case, the Transform algorithm demonstrated a 10−15% improvement in docking success rate and an effective 30-fold speed increase over the classical methods.[18] The improved search time permits the use of RosettaLigand for screening medium-sized small molecule libraries, protocols for which are found in the Supporting Information. For screening work with much larger libraries, Rosetta's Docking Approach using Ray-Casting (DARC) is a GPU-accelerated method demonstrated to be successful for protein−protein interface small molecules.[59] It should also be noted that screening applications use a simplified scoring function because of the computational complexity of fully flexible protein high-resolution refinement.

Customizable Small Molecule Docking Protocols. The RosettaLigand protocol can now be customized through the RosettaScripts XML interface, allowing for greater flexibility of use.[58] Additional features now include docking with explicit interface water molecules, which demonstrated 56% recovery for failed docking cases across a CSAR (Community Structure−Activity Resource) benchmark of 341 diverse structures.[60] Design of interfaces can now be incorporated into a single step for the docking and design of protein−small molecule binding pockets.[61] These RosettaScripts-based protocols have also been used to predict absolute binding energies for HIV-1 protease−inhibitor complexes with an R value of 0.71.[62]

Research questions often focus on small molecules binding to a target protein without an experimentally determined structure. Such cases require first building models of the receptor using *de novo* Rosetta, RosettaCM, or similar protein modeling protocols. When docking small molecules into protein models, Kaufmann and Meiler observed a nativelike binding pose among the top 10 scoring comparative models for 21 of 30 test cases.[63] Furthermore, docking results were significantly better in cases utilizing protein templates containing a small molecule of similar chemotype compared to templates with dissimilar small molecules or proteins in the apo state. A full Rosetta protocol linking comparative modeling and small molecule docking is available in ref 4. Combs et al.

DOI: 10.1021/acs.biochem.6b00444
Biochemistry 2016, 55, 4748−4763

utilized the previously discussed independent translation/rotation low-resolution sampling but can be easily modified to the new Transform sampling.

Small Molecule Docking in Membrane Proteins. Because of their biological importance and the challenges of experimentally determining their structures, membrane proteins are particularly attractive targets for the comparative model docking strategy. While the comparative modeling portion may be handled in a membrane environment, to date, Rosetta handles small molecule docking in a soluble environment. Nguyen et al. demonstrated the applicability of the soluble simplification for G protein-coupled receptors (GPCRs).[64] RosettaLigand sampled near-native poses when docking small molecules into comparative models of GPCRs, but selecting correct small molecule poses by Rosetta score alone remains challenging (Figure 3). The use of templates

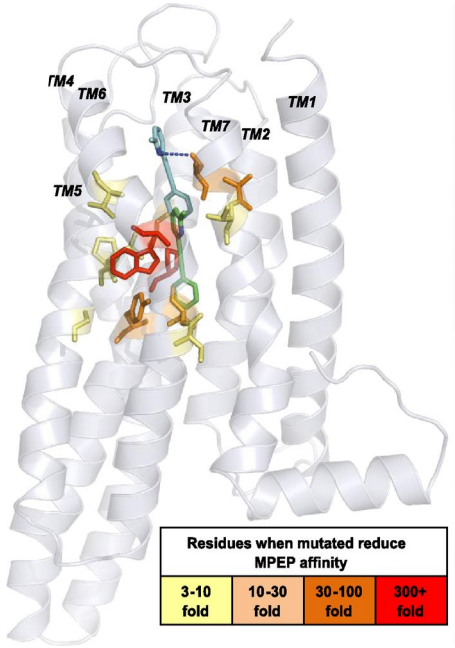

Figure 3. Application of RosettaLigand docking of negative allosteric modulator MPEP into a comparative model of the mGlu5 transmembrane domain. The predicted lowest-energy MPEP docking position (cyan) is close to residues demonstrating a change in MPEP modulations upon mutation (yellow to red). Reprinted with permission from ref 108. Copyright 2013 American Society for Pharmacology and Experimental Therapeutics.

with high sequence identity, knowledge-based binding pocket filters, and experimental contacts are recommended methods for improving accuracy. Additional algorithm development and benchmarking are being pursued to fully integrate RosettaLigand with the RosettaMembrane framework.[23]

INCORPORATING EXPERIMENTAL DATA

While Rosetta can sample near-native structures in a variety of situations, knowledge of limited experimental information can guide sampling and discriminate conformations inconsistent with experimental data, allowing more accurate determination of structures with less sampling. The incorporation of experimental data most commonly takes the form of modifications to the energy function. Addition of experiment-based scoring terms can make the energy landscape less rugged, allowing Rosetta sampling to more rapidly converge on relevant conformations.

For the incorporation of such information, Rosetta has a flexible restraint system (termed "constraints" in Rosetta parlance). Rosetta constraints have a two-part organization: specification of structural measurements such as distances or angles and a function that converts the measurement into an energetic penalty. A wide variety of measurements and functional transformations are currently available within Rosetta, and these can be freely mixed and matched according to the particular use case. There are also built-in tools for incorporating experimental data, allowing users to select only the best of a set of potentially inaccurate restraints. The flexibility of these restraints allows them to be applied in a diversity of situations, from incorporation of nuclear Overhauser enhancement (NOE) distances from NMR spectroscopy[65] to the use of mass spectrometry cross-linking information[66] to the use of custom potentials derived from probability distributions matching EPR/DEER measurements.[67,68]

Although the constraint system provides flexibility when incorporating experimental data for most Rosetta protocols, other experimental data types may reflect more complex structural parameters and require specialized scoring terms. Residual dipolar couplings,[69] pseudocontact shifts,[70] and small-angle X-ray scattering[71] have all been incorporated into Rosetta using specialized score terms, as have several techniques for working with electron microscopy (EM)- and X-ray-based electron density.[37,72,73] An example tutorial for using X-ray crystallography data and electron density maps with Rosetta, structure_refinement, is provided in the Supporting Information.

Improvements in image data analysis and electron detectors have led to advances in electron microscopy, producing electron density maps at resolutions as high as 3 Å for complex molecular machines. However, model building into these near-atomic resolution electron density maps is still difficult and error prone. DiMaio et al. have developed methods in Rosetta that incorporate medium- to high-resolution (3−5 Å) cryo-EM maps for density-guided structure determination and structure refinement.[72−74]

Protein Structure Prediction with Cryo-EM Restraints. This method takes advantage of near-atomic-resolution cryo-EM density maps for protein structure prediction. Using this method, highly accurate models of proteins up to 660 amino acids in length can be determined without homologous structures. This method includes density-traced backbone conformation and side-chain density agreement for sequence assignment during structure prediction. Structure determination starts with obtaining nine-residue fragments centered on each amino acid in the sequence using the Fragment Picker as mentioned previously in the *de novo* folding section. These fragments are then docked into the electron density map using a translational and rotational search to identify possible fragment placement. To further refine these placements, side-chain information is used to identify fragments with physically realistic side-chain conformations consistent with the experimental data. Finally, the largest mutually consistent subset of fragment placements is selected. A subset of placements is scored with a low-resolution score function that evaluates their pairwise consistency. Monte Carlo-simulated annealing finds a subset of fragment placements optimizing this score function. This assignment will not necessarily assign a position for each residue. This process is conducted iteratively until 70% of the

DOI: 10.1021/acs.biochem.6b00444
Biochemistry 2016, 55, 4748−4763

sequence has been assigned a backbone conformation. For consecutive iterations, the portion of the density map already covered in the previous step is excluded from fragment placements. Finally, Rosetta loop modeling and an all-atom refinement step, both guided by the cryo-EM density map, fill in any missing regions in the model.

Cryo-EM-restrained protein structure prediction[72] yielded models within 2.0−3.1 Å all-atom RMSD compared to experimentally determined structures. Structure determination of proteins rich in β-sheets is challenging for this method because of the conformational variability of the structure. Medium-resolution (4.8 Å) density maps provide another challenge during partial structure building for this method.

Density-Guided Iterative Local Refinement. This structure refinement protocol[74] includes techniques from X-ray crystallographic refinement, *de novo* structure prediction, segment rebuilding, and all-atom refinement from comparative modeling in Rosetta to predict models of proteins at atomic-level accuracy starting from a low-resolution model (with the correct topology). Like comparative modeling techniques, backbone fragments are inserted onto a template structure via superposition and minimization to close the peptide bonds. In density-guided structure rebuilding, before the peptide bonds are closed the fragments are optimized to fit the density after superposition. The backbone fragments that do not fit into this density are replaced by backbone fragments derived from the PDB. Peptide bond, backbone, and side-chain geometries are maintained during this step by coordinate constraints at the fragment end points and Ramachandran and rotameric constraints, respectively. This density-guided rebuilding step is followed by alternative refinement of model coordinates and atomic *B* factors until a good correlation is obtained between the model and the density map. Finally, the quality of the refined model is evaluated using all-atom energies as well as agreement with the experimental data, using the Fourier Shell Correlation between the model and map.

With homologues as starting points, the structure of the 20S proteasome, periplasmic domains PrgH and PrgK of the needle complex, and a peptide fiber assembly were refined using this method.[74] The accuracy of the refined models was tested against the quality (sequence identity) of the starting model, the number of images used for the reconstruction of the map, and the resolution of the density map. Density-guided iterative local rebuilding generated >75% accurate models for maps up to 4.4 Å resolution and less accurate models for maps with a resolution lower than 5 Å. This suggests that to successfully refine a model, the helix pitch, individual β-strands, and some of the aromatic side chains should be partially visible in the density map.

Among many applications, the density-guided iterative local rebuilding technique for structure determination in Rosetta has been used to determine structures of the peroxisomal Pex1/Pex6 ATPase complex with a unique double-ring,[75] type VI secretion system contractile sheath in *Vibrio cholerae*,[76] and SIRV2 virion that infects the *Sulfolobus islandicus* hyperthermophilic acidophile.[77]

Refinement with Phenix and Rosetta. Phenix[78] is state-of-the art X-ray refinement software used to determine crystal structures of biomolecules. The Rosetta structure modeling methodology has been combined with the Phenix refinement method to improve structure determination at low and high resolutions. Phenix benefits from the detailed all-atom force field and more effective conformational search and minimization procedures that exist in Rosetta. The Phenix.Rosetta refinement approach[73] utilizes Phenix for bulk solvent correction to calculate electron density maps and refine atomic *B* factors while the Rosetta force-field, minimization, and sampling techniques are used to optimize the model geometry. This method includes alternative real and reciprocal space refinement to improve model structure. Rosetta force-field constrains the refinement to physically plausible conformations, and density maps restrain the Rosetta side-chain and backbone sampling during refinement.

The Phenix.Rosetta refinement method was tested against conventional refinement in Phenix, CNS,[79] and REFMAC.[80] On 26 models with density map resolution ranging from 3.0 to 4.5 Å, Phenix.Rosetta refinement generated models with superior geometry in terms of free *R* factor, MolProbity score, and RMSD compared to that of the published structures.[73]

Phenix.Rosetta refinement has been successfully adapted to determine structures of the flavin binding center of the NqrC subunit of sodium-translocating NADH:quinone oxidoreductase,[81] the full-length protein and regulatory domain of *Pseudomonas aeruginosa* OxyR,[82] the apo-TrmBL2 structure to understand nonspecific binding of DNA by TrmBL2,[83] and the $\alpha\beta$ T cell antigen receptor (TCR)−CD1a complex.[84]

Phenix.mr_rosetta is another model rebuilding technique that integrates structure modeling tools from Rosetta with crystallographic structure determination tools in Phenix.[85] This technique can be used to determine challenging structures for which simple molecular replacement procedures usually fail, when starting models are based on a remote homologue with <30% sequence identity.[86,87] The phenix.mr_rosetta algorithms allow users to identify suitable templates and refine them with Rosetta before performing molecular replacement and then rebuilding the models with Rosetta and Phenix autobuilding tools. Electron density map-guided energy optimization, combinatorial side-chain packing, and torsional space minimization are used to improve molecular replacement models before applying crystallographic model building techniques. Phenix.mr_rosetta allows rapid structure determination without experimental phase information given the availability of homologues structures with >20% sequence identity, diffraction data sets of better than 3.2 Å resolution, and four or fewer copies in the asymmetric unit cell.

PROTEIN DESIGN

Inverse Folding Problem. Protein design is a unique protocol in that instead of finding the optimal conformation of a particular sequence, it aims to determine an optimal sequence for a given conformation. For this reason, it is often termed the "inverse protein folding problem".[38] Generally, there are two main design strategies: design for stability and design for function. The stability protocol considers the entire protein for design, and the score terms of interest are generally focused on improved packing. The design for function protocol is usually a localized design, centered on a specific region, domain, pocket, etc., of a protein with a focused energy function that governs precise interactions, such as electrostatics or hydrogen bonding.

Protein design involves iterative optimization of sequence and structure. During the fixed backbone side-chain optimization step, sequence space is sampled simultaneously with side-chain conformational space using Monte Carlo-simulated annealing by exchanging all possible amino acids at user-specified designable positions while evaluating the predicted

DOI: 10.1021/acs.biochem.6b00444
Biochemistry 2016, 55, 4748−4763

energy.[6] This is followed by flexible backbone minimization to optimize the model. The first successful use of *de novo* RosettaDesign produced a sequence for a fold not seen in the PDB.[6] The experimentally determined structure had an RMSD of 1.1 Å from the computationally design model. An example tutorial for protein design, protein_design, is provided in the Supporting Information.

Design for Stability. Protein stability can be affected by a single-point mutation. Kellogg et al. evaluated several protocols with varying levels of flexibility and sampling and determined one method in particular to be useful for single-point mutations.[7] This method was made into the application ddg_monomer. When ddg_monomer was tested on a set of 1210 single-point mutants from the ProTherm database, the correlation of predicted ddGs to experimental ddGs was 0.69 while the stability classification accuracy was 0.72.

While ddg_monomer is a tool for predicting how a single-point mutation affects the stability of a protein, RosettaVIP (void in packing) is a design strategy that has been developed to identify single-point mutations that could improve the stability of a protein.[9] When Borgo et al. fully designed proteins, they found that the hydrophobic cores of the designed models were poorly packed when compared to their respective native proteins. RosettaVIP was able to identify packing deficiencies and sample a much smaller sequence space to fill the void in packing, resulting in a more stable design.

Design for Functionality. In addition to stabilizing monomeric proteins, RosettaDesign can be used to design interfaces between proteins. Fleishman et al. established a dock design protocol that optimizes the sequence of a protein to bind a surface patch of a target protein during design. Docking was used to optimize the positioning of the interacting proteins at the interface. Experimentally determined structures had an interface very similar to those of the designed models.[10]

Other types of interfaces of interest for design applications are protein−small molecule interfaces. Tinberg et al.[11] provided a great example of using RosettaDesign to design for affinity as well as stability (Figure 4). First, RosettaMatch[88] was used to find a stable scaffold for design for binding a particular small molecule. Next, RosettaDesign was used to maximize the binding affinity between the protein and small molecule. Finally, a second round of design was used to minimize destabilization due to mutagenesis in the first round. To ensure these mutations were meaningful, design was guided by a multiple-sequence alignment. The resulting most energetically favorable model was the highest-affinity binder in experimental studies and had a cocrystal structure that agreed with the computational model.

Most design algorithms in Rosetta are performed while considering a single fixed backbone structure. Recently, efforts to consider several structures during the design process have been undertaken to tackle more difficult design problems. A generalized multistate design protocol was introduced in 2011[8] to help in cases in which design should occur to satisfy multiple conformations or to design specificity toward one state and negative design against other states. Willis et al.[89] showed that RosettaMultistateDesign was capable of predicting residues that were important for polyspecificity when designing the heavy-chain variable region of an antibody. Sevy et al. introduced a new approach to multistate design that accelerates the process of multistate design by reducing the sequence search space,[90] allowing more complex backbone movements to be incorporated into a design protocol.

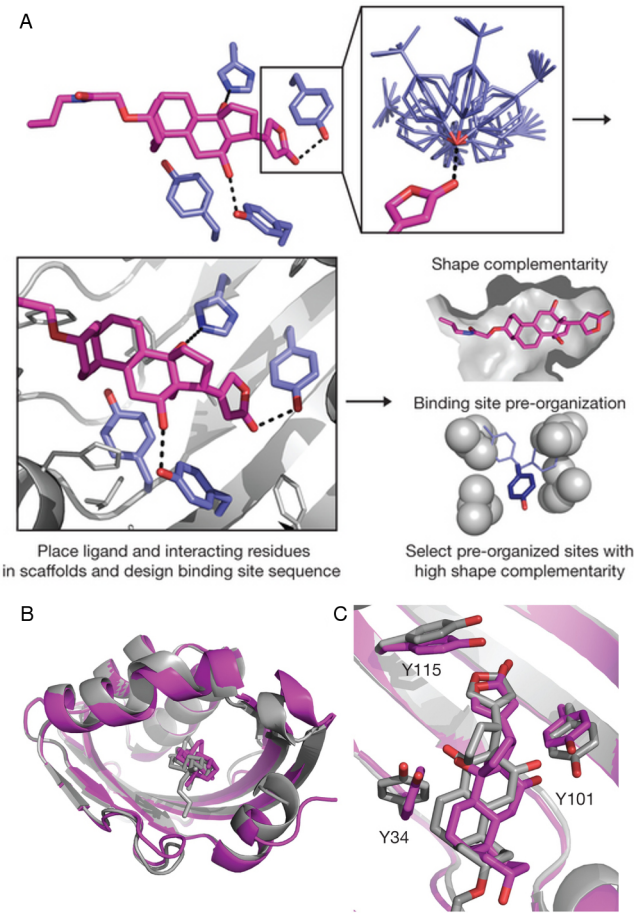

Figure 4. Design of protein−ligand interactions for high affinity and selectivity. (A) The design approach involved specifying binding interactions between the protein and ligand followed by design of the binding site. Finally, only designs in which shape complementarity was better than what is seen in native complexes were selected for experimental characterization. (B) Design crystal structure (purple) and computational model (gray) of the protein−ligand complex resulting from design for high affinity and selectivity. The RMSD was 0.54 Å, while the bound form (C) had an RMSD of 0.99 Å. Reprinted with permission from ref 11. Copyright 2013 Macmillan Publishers Ltd.

■ ADDITIONAL ROSETTA METHODS

Symmetry. Previously, Rosetta2 was limited in its ability to model large symmetric complexes.[24] In 2011, DiMaio et al. introduced a new mode in Rosetta to model symmetric proteins.[25] This allowed protocols to sample and score large, symmetric complexes much more quickly and with less memory usage as this approach samples only symmetric degrees of freedom, greatly reducing the search space. The underlying assumption, however, is that the interactions between all subunits are symmetric. The current implementation of RosettaSymmetry can create complex symmetric assemblies through the use of a symmetry definition file for a symmetric or nearly symmetric structure from the PDB. In the case of *de novo* folding, a symmetry definition file must be generated from scratch.

Membrane. RosettaMembrane has been the method used to model helical transmembrane proteins for several years. RosettaMembrane consists of both low-resolution[35] and high-resolution[91] scoring functions that were developed to describe

DOI: 10.1021/acs.biochem.6b00444
Biochemistry 2016, 55, 4748−4763

how the protein interacts with the membrane environment. Recently, RosettaMP, a new framework for modeling membrane proteins in Rosetta, was developed to facilitate communication between model sampling and scoring.[23] Work is ongoing to adapt existing protocols to be compatible with RosettaMP.

Noncanonical Amino Acids and Noncanonical Backbones. Rosetta was initially developed to predict the three-dimensional structure of proteins using the 20 canonical amino acids. However, the expansion to include noncanonical amino acids (NCAAs) and noncanonical backbones (NCBs) is important, as they allow for the flexibility to create more precise interactions between proteins,[92] metal ions,[93] or antigens.[94] While the expansion to include more diverse structures is critical, the addition is nontrivial.

The addition of NCAAs requires the modification of both the scoring function and how the space is explored. These hurdles, however, are not easy to clear, as Rosetta is built on a foundation of knowledge-based components within its scoring function. Most of these knowledge-based score terms come from published protein structures, and few NCAAs have a statistically relevant representation in the PDB. Therefore, developers need to rework key components of the Rosetta scoring function.[21] All score terms were then reweighted to account for the changes in the score terms. Along with the new score terms, the authors created rotamer libraries for 114 NCAAs, as well as a tool, MakeRotLib, for creating rotamers for user-supplied NCAAs.

An effort was also undertaken to add noncanonical backbones to Rosetta, and in the initial attempt, five new backbones were added.[22] The first hurdle in the addition of an NCB is defining what a "residue" is. In Rosetta,3[26] the "residue" became the central object; therefore, with NCBs, a repeating subunit must be defined. Additionally, new backbone sampling movers must be created, or the backbone must be fixed, as the NCB will have flexibilities different from those of a linear chain of three singly bonded atoms. The final key point in the addition of NCBs is the creation of new rotamer libraries for the side chain. Even if the side-chain atoms are identical to those of a canonical side chain, the chemical change in the backbone will cause different flexibilities, due to sterics or electrostatics. A peptoid (a backbone structure identical to the canonical backbone, with the only change being the side-chain branches from the nitrogen instead of the α-carbon) rotamer generator has been created[95] for users to create rotamers for their own side chains. However, care must be taken when creating rotamers for a blended backbone system.[96]

The main considerations for a user attempting to use NCAAs and/or NCBs in Rosetta are understanding the chemical properties of their side chain and/or backbone and properly representing this knowledge in Rosetta. The correct score terms need to be used, as the standard knowledge-based score terms will not apply. An appropriate rotamer library and/or mover must be added to allow for proper sampling of the protein landscape. Finally, the user must understand that because work on NCAAs and NCBs is still limited, novel score terms or sampling methods may be required.

RNA. Structural predictions of RNA molecules require confronting the same challenges as protein modeling: sampling the conformational space of the heteropolymer and accurately scoring different conformations. Rosetta applies the same strategies developed for modeling proteins to address these challenges in nucleic acids.[19] An assembly of fragments that have been observed in known RNA structures is used to produce nativelike tertiary models. This procedure also captures sequence-dependent local conformational biases. A centroid-based "low-resolution" scoring function is used for selection of initial models. It is knowledge-based (statistically derived from frequencies observed in known crystal structures) and includes scoring terms for base pairing, base stacking, and compactness and terms for maintaining coplanarity and disfavoring steric clashes. In current protocols, models are subsequently refined using a full-atom physics-based energy potential and a Stepwise Ansatz.[20] This protocol is termed FARFAR, Fragment Assembly of RNA with Full-Atom Refinement, and is available using the distributed Rosetta suite and the ROSIE Web server.[46]

The Stepwise Ansatz (Figure 5) has been benchmarked on loops and hairpins up to 10 nucleotides in length. For larger structures, additional information is needed to restrict the conformational sampling to a tractable amount. This

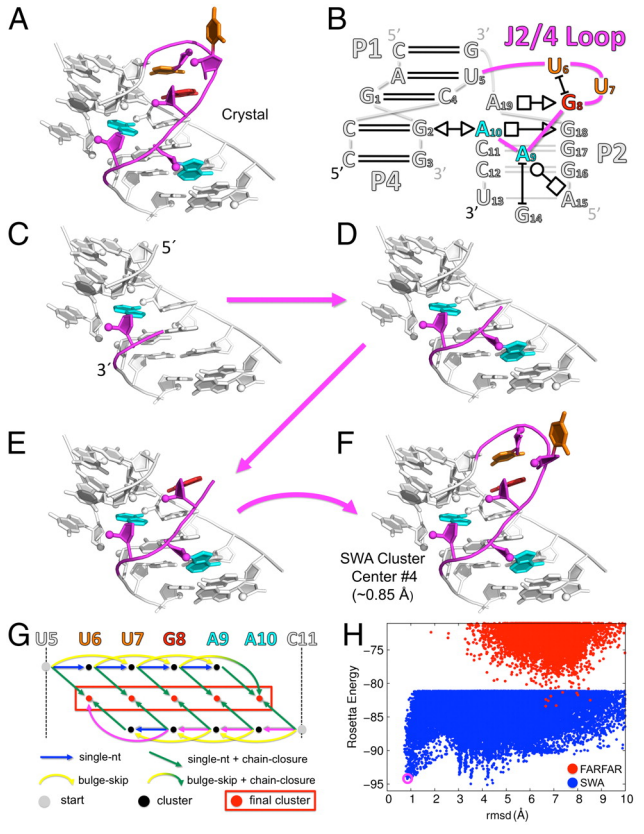

Figure 5. Stepwise assembly (SWA) structure modeling method for RNA. Illustration of the J2/4 loop from the three-way junction of a TPP-sensing riboswitch (PDB entry 3DV2). (A) Crystallographic conformation of the five-nucleotide loop (colored). (B) Schematic of the three-way junction. (C−F) The loop is built in a stepwise manner, starting from the 3′ end. (G) A directed acyclic graph recursively covers all possible build-up paths. The steps shown in panels C−F are colored magenta. Gray vertices correspond to the starting point with none of the loop nucleotides built. Black vertices are partially built subregions. Red vertices correspond to the ending points with the loop completely built. (H) Energy vs RMSD from the crystal for models generated by SWA (blue points) and by the prior method (FARFAR, red points). The SWA fourth lowest-energy cluster center (purple circle) is within atomic accuracy of the crystallographic model (0.85 Å RMSD). Reprinted with permission from ref 109. Copyright 2011 National Academy of Sciences.

DOI: 10.1021/acs.biochem.6b00444
Biochemistry 2016, 55, 4748−4763

information can come from both predictions and experimental data. Secondary structure predictions can be made using algorithms that take into account structures of homologous sequence. Chemical mapping experiments provide useful reactivity data that help assign base pairing status to each nucleotide. Multidimensional chemical mapping, such as "M2, mutate-and-map"[97] and Multiplexed -OH Cleavage Analysis by paired-end sequencing (MOHCA-seq),[98] can provide specific pairwise proximity information. Additional efficiency is gained by the preassembly of helical structures as input to the fragment assembly step. These methods have performed well in the recent blind prediction experiments called "RNA-Puzzles".[99]

The same techniques used for structure prediction can also be applied to structure refinement, to improve the quality of RNA crystallographic models in the presence of X-ray data. This procedure has been implemented in the "Enumerative Real-space Refinement ASsisted by Electron density under Rosetta" ERRASER-Phenix pipeline[100] and was demonstrated to improve the geometrical parameters and model quality of 24 RNA-containing structures in the PDB, including small pseudoknots and large ribosomal subunits.[101]

NMR structure determination of proteins or nucleic acids typically relies on a large number of NOE measurements to derive distance constraints for structure calculations. Using a relatively small number of measurements of only ^{1}H chemical shift values, CS-Rosetta-RNA was demonstrated to provide sufficient information to determine the structures of 23 noncanonical RNA motifs at high resolution.[102] This functionality is also available on the ROSIE Web server.[46]

RNA Design. RNA can be designed using Rosetta's RNA Redesign algorithm. It performs fixed backbone design on 3D RNA structures to produce sequences that best stabilize a given 3D conformation.[103] The success rate for a benchmark set of 15 RNA crystal structures was 45% sequence recovery overall and 65% sequence recovery for noncanonical sequences (not Watson–Crick or G-U). Finally, the algorithm was able to predict a sequence that would increase the thermostability of domain IV of the signal recognition particle.

CONCLUSIONS

The Rosetta software suite represents a compilation of computational tools aimed at obtaining physically relevant structural models of proteins, RNA, and small molecule interactions. Herein, we presented a general outline of updated Rosetta applications, protocols, frameworks, and functionalities with the aim of improving user success. All protocols are generalizable and can be applied to an extended list of biological queries that other structure-determining methods may not be able address.

Improvements to the variety of Rosetta interfaces (RosettaScript, PyRosetta, and many web interfaces) allow the user a high degree of flexibility and personalization for each specific structural problem, as well as providing a previously unavailable entry point for novice users.

The current, default Rosetta score function (*talaris2014*) has been optimized and improved with new score terms as well as reweighted knowledge- and physics-based potentials. Rosetta also incorporates a new release of the Dunbrack rotamer library.[33]

De novo structure prediction has greatly improved with the implementation of the TopologyBroker, which was developed to create consensus sampling that satisfies all user-requested constraints without requiring additional code development for each unique system. Recent progress in comparative modeling applications has broadened the possible conformational search space by incorporating multiple starting templates. Protocols for protein–protein docking now include flexibility to modularize the coarse-grained and high-resolution modes of RosettaDock, giving the user more freedom to incorporate additional features in the docking process while narrowing the computational search space. Improvements in protein–small molecule docking utilize an improved *Transform* algorithm that increases both the speed and quality of this tool in obtaining more nativelike conformations. Likewise, the flexibility in incorporating experimentally derived constraints for most protocols has also greatly improved. To tackle the challenge of the inverse folding problem, new implementations of multistate design permit users to optimize sequences while considering several structures simultaneously.

Continuous developments in Rosetta have enhanced its utility by adding functionality to model proteins embedded in the membrane, expansion into nontraditional protein modeling by adding noncanonical amino acids, noncanonical backbones, and nucleic acids, and adding the ability to model ever-larger proteins by the addition of symmetry.

Installation and Licensing. The Rosetta licenses are available at http://www.rosettacommons.org/software free of charge for academic and governmental laboratories. Rosetta is compatible with most Unix-based operating systems and is distributed as source code. A user manual describing compilation, installation, and usage for the current release can be found at http://www.rosettacommons.org/docs/latest/. Demos and tutorials for additional Rosetta protocols can be found at http://www.rosettacommons.org/demos/latest/. Interested developers can join the RosettaCommons organization to contribute to the Rosetta software package.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.biochem.6b00444.

Step-by-step tutorials for six of the protocols discussed in the paper (*de novo* folding, comparative modeling, protein–protein docking, protein–ligand docking, working with electron density, and protein design) (ZIP)

AUTHOR INFORMATION

Corresponding Author

*Department of Chemistry, Vanderbilt University, 7330 Stevenson Center, Station B 351822, Nashville, TN 37235. E-mail: rocco.moretti@vanderbilt.edu. Telephone: +1 (615) 936-6594.

Author Contributions

B.J.B., A.C., A.M.D., J.A.F., D.F., A.D.L., B.K.M., A.K.S., M.F.S., and A.M.S. contributed equally to this work.

Funding

This work was supported by a grant from the National Institutes of Health (NIH) (R01GM073151) and the RosettaCommons. Work in the Meiler laboratory is supported by the NIH (R01 GM080403, R01 GM099842, R01 DK097376, R01 HL122010, R01 GM073151, and U19 AI117905) and the National Science Foundation (CHE 1305874).

DOI: 10.1021/acs.biochem.6b00444
Biochemistry 2016, 55, 4748–4763

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors acknowledge Ray Y.-R. Wang and Sam DeLuca for assistance in preparing the tutorials. The table of contents graphic incorporates a CC-BY-SA 4.0 licensed photo by Hans Hillewaert.

■ REFERENCES

(1) Rohl, C. A., Strauss, C. E., Misura, K. M., and Baker, D. (2004) Protein structure prediction using Rosetta. *Methods Enzymol. 383*, 66−93.

(2) Ovchinnikov, S., Kinch, L., Park, H., Liao, Y., Pei, J., Kim, D. E., Kamisetty, H., Grishin, N. V., and Baker, D. (2015) Large-scale determination of previously unsolved protein structures using evolutionary information. *eLife 4*, e09248.

(3) Raman, S., Vernon, R., Thompson, J., Tyka, M., Sadreyev, R., Pei, J., Kim, D., Kellogg, E., DiMaio, F., Lange, O., Kinch, L., Sheffler, W., Kim, B. H., Das, R., Grishin, N. V., and Baker, D. (2009) Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins: Struct., Funct., Genet. 77* (Suppl. 9), 89−99.

(4) Combs, S. A., Deluca, S. L., Deluca, S. H., Lemmon, G. H., Nannemann, D. P., Nguyen, E. D., Willis, J. R., Sheehan, J. H., and Meiler, J. (2013) Small-molecule ligand docking into comparative models with Rosetta. *Nat. Protoc. 8*, 1277−1298.

(5) Song, Y., DiMaio, F., Wang, R. Y., Kim, D., Miles, C., Brunette, T., Thompson, J., and Baker, D. (2013) High-resolution comparative modeling with RosettaCM. *Structure 21*, 1735−1742.

(6) Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L., and Baker, D. (2003) Design of a novel globular protein fold with atomic-level accuracy. *Science 302*, 1364−1368.

(7) Kellogg, E. H., Leaver-Fay, A., and Baker, D. (2011) Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins: Struct., Funct., Genet. 79*, 830−838.

(8) Leaver-Fay, A., Jacak, R., Stranges, P. B., and Kuhlman, B. (2011) A generic program for multistate protein design. *PLoS One 6*, e20937.

(9) Borgo, B., and Havranek, J. J. (2012) Automated selection of stabilizing mutations in designed and natural proteins. *Proc. Natl. Acad. Sci. U. S. A. 109*, 1494−1499.

(10) Fleishman, S. J., Whitehead, T. A., Ekiert, D. C., Dreyfus, C., Corn, J. E., Strauch, E. M., Wilson, I. A., and Baker, D. (2011) Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science 332*, 816−821.

(11) Tinberg, C. E., Khare, S. D., Dou, J., Doyle, L., Nelson, J. W., Schena, A., Jankowski, W., Kalodimos, C. G., Johnsson, K., Stoddard, B. L., and Baker, D. (2013) Computational design of ligand-binding proteins with high affinity and selectivity. *Nature 501*, 212−216.

(12) Gray, J. J., Moughon, S., Wang, C., Schueler-Furman, O., Kuhlman, B., Rohl, C. A., and Baker, D. (2003) Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J. Mol. Biol. 331*, 281−299.

(13) Chaudhury, S., Berrondo, M., Weitzner, B. D., Muthu, P., Bergman, H., and Gray, J. J. (2011) Benchmarking and analysis of protein docking performance in Rosetta v3.2. *PLoS One 6*, e22477.

(14) Raveh, B., London, N., and Schueler-Furman, O. (2010) Sub-angstrom modeling of complexes between flexible peptides and globular proteins. *Proteins: Struct., Funct., Genet. 78*, 2029−2040.

(15) Raveh, B., London, N., Zimmerman, L., and Schueler-Furman, O. (2011) Rosetta FlexPepDock ab-initio: simultaneous folding, docking and refinement of peptides onto their receptors. *PLoS One 6*, e18934.

(16) Meiler, J., and Baker, D. (2006) ROSETTALIGAND: protein-small molecule docking with full side-chain flexibility. *Proteins: Struct., Funct., Genet. 65*, 538−548.

(17) Davis, I. W., and Baker, D. (2009) RosettaLigand docking with full ligand and receptor flexibility. *J. Mol. Biol. 385*, 381−392.

(18) DeLuca, S., Khar, K., and Meiler, J. (2015) Fully Flexible Docking of Medium Sized Ligand Libraries with RosettaLigand. *PLoS One 10*, e0132508.

(19) Das, R., and Baker, D. (2007) Automated de novo prediction of native-like RNA tertiary structures. *Proc. Natl. Acad. Sci. U. S. A. 104*, 14664−14669.

(20) Das, R. (2013) Atomic-accuracy prediction of protein loop structures through an RNA-inspired Ansatz. *PLoS One 8*, e74830.

(21) Renfrew, P. D., Choi, E. J., Bonneau, R., and Kuhlman, B. (2012) Incorporation of noncanonical amino acids into Rosetta and use in computational protein-peptide interface design. *PLoS One 7*, e32637.

(22) Drew, K., Renfrew, P. D., Craven, T. W., Butterfoss, G. L., Chou, F. C., Lyskov, S., Bullock, B. N., Watkins, A., Labonte, J. W., Pacella, M., Kilambi, K. P., Leaver-Fay, A., Kuhlman, B., Gray, J. J., Bradley, P., Kirshenbaum, K., Arora, P. S., Das, R., and Bonneau, R. (2013) Adding diverse noncanonical backbones to rosetta: enabling peptidomimetic design. *PLoS One 8*, e67051.

(23) Alford, R. F., Koehler Leman, J., Weitzner, B. D., Duran, A. M., Tilley, D. C., Elazar, A., and Gray, J. J. (2015) An Integrated Framework Advancing Membrane Protein Modeling and Design. *PLoS Comput. Biol. 11*, e1004398.

(24) André, I., Bradley, P., Wang, C., and Baker, D. (2007) Prediction of the structure of symmetrical protein assemblies. *Proc. Natl. Acad. Sci. U. S. A. 104*, 17656−17661.

(25) DiMaio, F., Leaver-Fay, A., Bradley, P., Baker, D., and André, I. (2011) Modeling symmetric macromolecular structures in Rosetta3. *PLoS One 6*, e20450.

(26) Leaver-Fay, A., Tyka, M., Lewis, S. M., Lange, O. F., Thompson, J., Jacak, R., Kaufman, K., Renfrew, P. D., Smith, C. A., Sheffler, W., Davis, I. W., Cooper, S., Treuille, A., Mandell, D. J., Richter, F., Ban, Y. E., Fleishman, S. J., Corn, J. E., Kim, D. E., Lyskov, S., Berrondo, M., Mentzer, S., Popović, Z., Havranek, J. J., Karanicolas, J., Das, R., Meiler, J., Kortemme, T., Gray, J. J., Kuhlman, B., Baker, D., and Bradley, P. (2011) ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol. 487*, 545−574.

(27) Rohl, C. A., Strauss, C. E., Chivian, D., and Baker, D. (2004) Modeling structurally variable regions in homologous proteins with rosetta. *Proteins: Struct., Funct., Genet. 55*, 656−677.

(28) Simons, K. T., Kooperberg, C., Huang, E., and Baker, D. (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol. 268*, 209−225.

(29) Dunbrack, R. L., and Karplus, M. (1993) Backbone-dependent rotamer library for proteins. Application to side-chain prediction. *J. Mol. Biol. 230*, 543−574.

(30) Kuhlman, B., and Baker, D. (2000) Native protein sequences are close to optimal for their structures. *Proc. Natl. Acad. Sci. U. S. A. 97*, 10383−10388.

(31) Song, Y., Tyka, M., Leaver-Fay, A., Thompson, J., and Baker, D. (2011) Structure-guided forcefield optimization. *Proteins: Struct., Funct., Genet. 79*, 1898−1909.

(32) Leaver-Fay, A., O'Meara, M. J., Tyka, M., Jacak, R., Song, Y., Kellogg, E. H., Thompson, J., Davis, I. W., Pache, R. A., Lyskov, S., Gray, J. J., Kortemme, T., Richardson, J. S., Havranek, J. J., Snoeyink, J., Baker, D., and Kuhlman, B. (2013) Scientific benchmarks for guiding macromolecular energy function improvement. *Methods Enzymol. 523*, 109−143.

(33) Shapovalov, M. V., and Dunbrack, R. L., Jr. (2011) A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure 19*, 844−858.

(34) Bazzoli, A., Kelow, S. P., and Karanicolas, J. (2015) Enhancements to the Rosetta Energy Function Enable Improved Identification of Small Molecules that Inhibit Protein-Protein Interactions. *PLoS One 10*, e0140359.

DOI: 10.1021/acs.biochem.6b00444
Biochemistry 2016, 55, 4748−4763

(35) Yarov-Yarovoy, V., Schonbrun, J., and Baker, D. (2006) Multipass membrane protein structure prediction using Rosetta. *Proteins: Struct., Funct., Genet. 62*, 1010−1025.
(36) Fleishman, S. J., Leaver-Fay, A., Corn, J. E., Strauch, E. M., Khare, S. D., Koga, N., Ashworth, J., Murphy, P., Richter, F., Lemmon, G., Meiler, J., and Baker, D. (2011) RosettaScripts: a scripting language interface to the Rosetta macromolecular modeling suite. *PLoS One 6*, e20161.
(37) O'Meara, M. J., Leaver-Fay, A., Tyka, M., Stein, A., Houlihan, K., DiMaio, F., Bradley, P., Kortemme, T., Baker, D., Snoeyink, J., and Kuhlman, B. (2015) A Combined Covalent-Electrostatic Model of Hydrogen Bonding Improves Structure Prediction with Rosetta. *J. Chem. Theory Comput. 11*, 609−622.
(38) Kaufmann, K. W., Lemmon, G. H., Deluca, S. L., Sheehan, J. H., and Meiler, J. (2010) Practically useful: what the Rosetta protein modeling suite can do for you. *Biochemistry 49*, 2987−2998.
(39) Chaudhury, S., Lyskov, S., and Gray, J. J. (2010) PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics 26*, 689−691.
(40) Schenkelberg, C. D., and Bystroff, C. (2015) InteractiveROSETTA: a graphical user interface for the PyRosetta protein modeling suite. *Bioinformatics 31*, 4023−4025.
(41) Baugh, E. H., Lyskov, S., Weitzner, B. D., and Gray, J. J. (2011) Real-time PyMOL visualization for Rosetta and PyRosetta. *PLoS One 6*, e21931.
(42) Thornburg, N. J., Nannemann, D. P., Blum, D. L., Belser, J. A., Tumpey, T. M., Deshpande, S., Fritz, G. A., Sapparapu, G., Krause, J. C., Lee, J. H., Ward, A. B., Lee, D. E., Li, S., Winarski, K. L., Spiller, B. W., Meiler, J., and Crowe, J. E., Jr. (2013) Human antibodies that neutralize respiratory droplet transmissible H5N1 influenza viruses. *J. Clin. Invest. 123*, 4405−4409.
(43) Jardine, J., Julien, J. P., Menis, S., Ota, T., Kalyuzhniy, O., McGuire, A., Sok, D., Huang, P. S., MacPherson, S., Jones, M., Nieusma, T., Mathison, J., Baker, D., Ward, A. B., Burton, D. R., Stamatatos, L., Nemazee, D., Wilson, I. A., and Schief, W. R. (2013) Rational HIV immunogen design to target specific germline B cell receptors. *Science 340*, 711−716.
(44) *PyMOL Molecular Graphics System* (2015) Schrödinger, LLC, Portland, OR.
(45) Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and de Hoon, M. J. (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics 25*, 1422−1423.
(46) Lyskov, S., Chou, F. C., Conchúir, S., Der, B. S., Drew, K., Kuroda, D., Xu, J., Weitzner, B. D., Renfrew, P. D., Sripakdeevong, P., Borgo, B., Havranek, J. J., Kuhlman, B., Kortemme, T., Bonneau, R., Gray, J. J., and Das, R. (2013) Serverification of molecular modeling applications: the Rosetta Online Server that Includes Everyone (ROSIE). *PLoS One 8*, e63906.
(47) Adolf-Bryfogle, J., and Dunbrack, R. L., Jr. (2013) The PyRosetta Toolkit: a graphical user interface for the Rosetta software suite. *PLoS One 8*, e66856.
(48) Lewis, S. M., Wu, X., Pustilnik, A., Sereno, A., Huang, F., Rick, H. L., Guntas, G., Leaver-Fay, A., Smith, E. M., Ho, C., Hansen-Estruch, C., Chamberlain, A. K., Truhlar, S. M., Conner, E. M., Atwell, S., Kuhlman, B., and Demarest, S. J. (2014) Generation of bispecific IgG antibodies by structure-based design of an orthogonal Fab interface. *Nat. Biotechnol. 32*, 191−198.
(49) Kilambi, K. P., Pacella, M. S., Xu, J., Labonte, J. W., Porter, J. R., Muthu, P., Drew, K., Kuroda, D., Schueler-Furman, O., Bonneau, R., and Gray, J. J. (2013) Extending RosettaDock with water, sugar, and pH for prediction of complex structures and affinities for CAPRI rounds 20−27. *Proteins: Struct., Funct., Genet. 81*, 2201−2209.
(50) Sheffler, W., and Baker, D. (2009) RosettaHoles: rapid assessment of protein core packing for structure prediction, refinement, design, and validation. *Protein Sci. 18*, 229−239.
(51) Kim, D. E., Chivian, D., and Baker, D. (2004) Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res. 32*, W526−531.
(52) Gront, D., Kulp, D. W., Vernon, R. M., Strauss, C. E., and Baker, D. (2011) Generalized fragment picking in Rosetta: design, protocols and applications. *PLoS One 6*, e23294.
(53) Porter, J. R., Weitzner, B. D., and Lange, O. F. (2015) A Framework to Simplify Combined Sampling Strategies in Rosetta. *PLoS One 10*, e0138220.
(54) Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res. 25*, 3389−3402.
(55) Mandell, D. J., Coutsias, E. A., and Kortemme, T. (2009) Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. *Nat. Methods 6*, 551−552.
(56) Guilhot-Gaudeffroy, A., Froidevaux, C., Azé, J., and Bernauer, J. (2014) Protein-RNA complexes and efficient automatic docking: expanding RosettaDock possibilities. *PLoS One 9*, e108928.
(57) Kilambi, K. P., Reddy, K., and Gray, J. J. (2014) Protein-protein docking with dynamic residue protonation states. *PLoS Comput. Biol. 10*, e1004018.
(58) Lemmon, G., and Meiler, J. (2012) Rosetta Ligand docking with flexible XML protocols. *Methods Mol. Biol. 819*, 143−155.
(59) Gowthaman, R., Miller, S. A., Rogers, S., Khowsathit, J., Lan, L., Bai, N., Johnson, D. K., Liu, C., Xu, L., Anbanandam, A., Aubé, J., Roy, A., and Karanicolas, J. (2016) DARC: Mapping Surface Topography by Ray-Casting for Effective Virtual Screening at Protein Interaction Sites. *J. Med. Chem. 59*, 4152−4170.
(60) Lemmon, G., and Meiler, J. (2013) Towards ligand docking including explicit interface water molecules. *PLoS One 8*, e67536.
(61) Allison, B., Combs, S., DeLuca, S., Lemmon, G., Mizoue, L., and Meiler, J. (2014) Computational design of protein-small molecule interfaces. *J. Struct. Biol. 185*, 193−202.
(62) Lemmon, G., Kaufmann, K., and Meiler, J. (2012) Prediction of HIV-1 protease/inhibitor affinity using RosettaLigand. *Chem. Biol. Drug Des. 79*, 888−896.
(63) Kaufmann, K. W., and Meiler, J. (2012) Using RosettaLigand for small molecule docking into comparative models. *PLoS One 7*, e50769.
(64) Nguyen, E. D., Norn, C., Frimurer, T. M., and Meiler, J. (2013) Assessment and challenges of ligand docking into comparative models of G-protein coupled receptors. *PLoS One 8*, e67302.
(65) Zhang, Z., Porter, J., Tripsianes, K., and Lange, O. F. (2014) Robust and highly accurate automatic NOESY assignment and structure determination with Rosetta. *J. Biomol. NMR 59*, 135−145.
(66) Kahraman, A., Herzog, F., Leitner, A., Rosenberger, G., Aebersold, R., and Malmström, L. (2013) Cross-link guided molecular modeling with ROSETTA. *PLoS One 8*, e73411.
(67) Alexander, N. S., Stein, R. A., Koteiche, H. A., Kaufmann, K. W., McHaourab, H. S., and Meiler, J. (2013) RosettaEPR: rotamer library for spin label structure and dynamics. *PLoS One 8*, e72851.
(68) Hirst, S. J., Alexander, N., McHaourab, H. S., and Meiler, J. (2011) RosettaEPR: an integrated tool for protein structure determination from sparse EPR data. *J. Struct. Biol. 173*, 506−514.
(69) Sgourakis, N. G., Lange, O. F., DiMaio, F., André, I., Fitzkee, N. C., Rossi, P., Montelione, G. T., Bax, A., and Baker, D. (2011) Determination of the structures of symmetric protein oligomers from NMR chemical shifts and residual dipolar couplings. *J. Am. Chem. Soc. 133*, 6288−6298.
(70) Schmitz, C., Vernon, R., Otting, G., Baker, D., and Huber, T. (2012) Protein structure determination from pseudocontact shifts using ROSETTA. *J. Mol. Biol. 416*, 668−677.
(71) Rossi, P., Shi, L., Liu, G., Barbieri, C. M., Lee, H. W., Grant, T. D., Luft, J. R., Xiao, R., Acton, T. B., Snell, E. H., Montelione, G. T., Baker, D., Lange, O. F., and Sgourakis, N. G. (2015) A hybrid NMR/SAXS-based approach for discriminating oligomeric protein interfaces using Rosetta. *Proteins: Struct., Funct., Genet. 83*, 309−317.
(72) Wang, R. Y., Kudryashev, M., Li, X., Egelman, E. H., Basler, M., Cheng, Y., Baker, D., and DiMaio, F. (2015) De novo protein

DOI: 10.1021/acs.biochem.6b00444
Biochemistry 2016, 55, 4748−4763

structure determination from near-atomic-resolution cryo-EM maps. *Nat. Methods 12*, 335−338.
(73) DiMaio, F., Echols, N., Headd, J. J., Terwilliger, T. C., Adams, P. D., and Baker, D. (2013) Improved low-resolution crystallographic refinement with Phenix and Rosetta. *Nat. Methods 10*, 1102−1104.
(74) DiMaio, F., Song, Y., Li, X., Brunner, M. J., Xu, C., Conticello, V., Egelman, E., Marlovits, T. C., Cheng, Y., and Baker, D. (2015) Atomic-accuracy models from 4.5-Å cryo-electron microscopy data with density-guided iterative local refinement. *Nat. Methods 12*, 361−365.
(75) Blok, N. B., Tan, D., Wang, R. Y., Penczek, P. A., Baker, D., DiMaio, F., Rapoport, T. A., and Walz, T. (2015) Unique double-ring structure of the peroxisomal Pex1/Pex6 ATPase complex revealed by cryo-electron microscopy. *Proc. Natl. Acad. Sci. U. S. A. 112*, E4017−4025.
(76) Kudryashev, M., Wang, R. Y., Brackmann, M., Scherer, S., Maier, T., Baker, D., DiMaio, F., Stahlberg, H., Egelman, E. H., and Basler, M. (2015) Structure of the type VI secretion system contractile sheath. *Cell 160*, 952−962.
(77) DiMaio, F., Yu, X., Rensen, E., Krupovic, M., Prangishvili, D., and Egelman, E. H. (2015) Virology. A virus that infects a hyperthermophile encapsidates A-form DNA. *Science 348*, 914−917.
(78) Adams, P. D., Afonine, P. V., Bunkóczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L. W., Kapral, G. J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwilliger, T. C., and Zwart, P. H. (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr., Sect. D: Biol. Crystallogr. 66*, 213−221.
(79) Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J. S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T., and Warren, G. L. (1998) Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr., Sect. D: Biol. Crystallogr. 54*, 905−921.
(80) Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F., and Vagin, A. A. (2011) REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr., Sect. D: Biol. Crystallogr. 67*, 355−367.
(81) Borshchevskiy, V., Round, E., Bertsova, Y., Polovinkin, V., Gushchin, I., Ishchenko, A., Kovalev, K., Mishin, A., Kachalova, G., Popov, A., Bogachev, A., and Gordeliy, V. (2015) Structural and functional investigation of flavin binding center of the NqrC subunit of sodium-translocating NADH:quinone oxidoreductase from Vibrio harveyi. *PLoS One 10*, e0118548.
(82) Jo, I., Chung, I. Y., Bae, H. W., Kim, J. S., Song, S., Cho, Y. H., and Ha, N. C. (2015) Structural details of the OxyR peroxide-sensing mechanism. *Proc. Natl. Acad. Sci. U. S. A. 112*, 6443−6448.
(83) Ahmad, M. U., Waege, I., Hausner, W., Thomm, M., Boos, W., Diederichs, K., and Welte, W. (2015) Structural Insights into Nonspecific Binding of DNA by TrmBL2, an Archaeal Chromatin Protein. *J. Mol. Biol. 427*, 3216−3229.
(84) Birkinshaw, R. W., Pellicci, D. G., Cheng, T. Y., Keller, A. N., Sandoval-Romero, M., Gras, S., de Jong, A., Uldrich, A. P., Moody, D. B., Godfrey, D. I., and Rossjohn, J. (2015) $\alpha\beta$ T cell antigen receptor recognition of CD1a presenting self lipid ligands. *Nat. Immunol. 16*, 258−266.
(85) Terwilliger, T. C., Dimaio, F., Read, R. J., Baker, D., Bunkoczi, G., Adams, P. D., Grosse-Kunstleve, R. W., Afonine, P. V., and Echols, N. (2012) phenix.mr_rosetta: molecular replacement and model rebuilding with Phenix and Rosetta. *J. Struct. Funct. Genomics 13*, 81−90.
(86) DiMaio, F., Terwilliger, T. C., Read, R. J., Wlodawer, A., Oberdorfer, G., Wagner, U., Valkov, E., Alon, A., Fass, D., Axelrod, H. L., Das, D., Vorobiev, S. M., Iwai, H., Pokkuluri, P. R., and Baker, D. (2011) Improved molecular replacement by density- and energy-guided protein structure optimization. *Nature 473*, 540−543.
(87) DiMaio, F., Tyka, M. D., Baker, M. L., Chiu, W., and Baker, D. (2009) Refinement of protein structures into low-resolution density maps using rosetta. *J. Mol. Biol. 392*, 181−190.
(88) Zanghellini, A., Jiang, L., Wollacott, A. M., Cheng, G., Meiler, J., Althoff, E. A., Röthlisberger, D., and Baker, D. (2006) New algorithms and an in silico benchmark for computational enzyme design. *Protein Sci. 15*, 2785−2794.
(89) Willis, J. R., Briney, B. S., DeLuca, S. L., Crowe, J. E., and Meiler, J. (2013) Human germline antibody gene segments encode polyspecific antibodies. *PLoS Comput. Biol. 9*, e1003045.
(90) Sevy, A. M., Jacobs, T. M., Crowe, J. E., and Meiler, J. (2015) Design of Protein Multi-specificity Using an Independent Sequence Search Reduces the Barrier to Low Energy Sequences. *PLoS Comput. Biol. 11*, e1004300.
(91) Barth, P., Schonbrun, J., and Baker, D. (2007) Toward high-resolution prediction and design of transmembrane helical protein structures. *Proc. Natl. Acad. Sci. U. S. A. 104*, 15682−15687.
(92) Sievers, S. A., Karanicolas, J., Chang, H. W., Zhao, A., Jiang, L., Zirafi, O., Stevens, J. T., Münch, J., Baker, D., and Eisenberg, D. (2011) Structure-based design of non-natural amino-acid inhibitors of amyloid fibril formation. *Nature 475*, 96−100.
(93) Mills, J. H., Khare, S. D., Bolduc, J. M., Forouhar, F., Mulligan, V. K., Lew, S., Seetharaman, J., Tong, L., Stoddard, B. L., and Baker, D. (2013) Computational design of an unnatural amino acid dependent metalloprotein with atomic level accuracy. *J. Am. Chem. Soc. 135*, 13393−13399.
(94) Xu, J., Tack, D., Hughes, R. A., Ellington, A. D., and Gray, J. J. (2014) Structure-based non-canonical amino acid design to covalently crosslink an antibody-antigen complex. *J. Struct. Biol. 185*, 215−222.
(95) Renfrew, P. D., Craven, T. W., Butterfoss, G. L., Kirshenbaum, K., and Bonneau, R. (2014) A rotamer library to enable modeling and design of peptoid foldamers. *J. Am. Chem. Soc. 136*, 8772−8782.
(96) Butterfoss, G. L., Drew, K., Renfrew, P. D., Kirshenbaum, K., and Bonneau, R. (2014) Conformational preferences of peptide-peptoid hybrid oligomers. *Biopolymers 102*, 369−378.
(97) Cordero, P., Kladwang, W., VanLang, C. C., and Das, R. (2014) The mutate-and-map protocol for inferring base pairs in structured RNA. *Methods Mol. Biol. 1086*, 53−77.
(98) Cheng, C., Chou, F.-C., Kladwang, W., Tian, S., Cordero, P., and Das, R. (2014) MOHCA-seq: RNA 3D models from single multiplexed proximity-mapping experiments. *bioRxiv*, 004556.
(99) Miao, Z., Adamiak, R. W., Blanchet, M. F., Boniecki, M., Bujnicki, J. M., Chen, S. J., Cheng, C., Chojnowski, G., Chou, F. C., Cordero, P., Cruz, J. A., Ferré-D'Amaré, A. R., Das, R., Ding, F., Dokholyan, N. V., Dunin-Horkawicz, S., Kladwang, W., Krokhotin, A., Lach, G., Magnus, M., Major, F., Mann, T. H., Masquida, B., Matelska, D., Meyer, M., Peselis, A., Popenda, M., Purzycka, K. J., Serganov, A., Stasiewicz, J., Szachniuk, M., Tandon, A., Tian, S., Wang, J., Xiao, Y., Xu, X., Zhang, J., Zhao, P., Zok, T., and Westhof, E. (2015) RNA-Puzzles Round II: assessment of RNA structure prediction programs applied to three large RNA structures. *RNA 21*, 1066−1084.
(100) Chou, F. C., Echols, N., Terwilliger, T. C., and Das, R. (2016) RNA Structure Refinement Using the ERRASER-Phenix Pipeline. *Methods Mol. Biol. 1320*, 269−282.
(101) Chou, F. C., Sripakdeevong, P., Dibrov, S. M., Hermann, T., and Das, R. (2012) Correcting pervasive errors in RNA crystallography through enumerative structure prediction. *Nat. Methods 10*, 74−76.
(102) Sripakdeevong, P., Cevec, M., Chang, A. T., Erat, M. C., Ziegeler, M., Zhao, Q., Fox, G. E., Gao, X., Kennedy, S. D., Kierzek, R., Nikonowicz, E. P., Schwalbe, H., Sigel, R. K., Turner, D. H., and Das, R. (2014) Structure determination of noncanonical RNA motifs guided by ^{1}H NMR chemical shifts. *Nat. Methods 11*, 413−416.
(103) Das, R., Karanicolas, J., and Baker, D. (2010) Atomic accuracy in predicting and designing noncanonical RNA structure. *Nat. Methods 7*, 291−294.
(104) Liu, Y., and Kuhlman, B. (2006) RosettaDesign server for protein design. *Nucleic Acids Res. 34*, W235−238.

DOI: 10.1021/acs.biochem.6b00444
Biochemistry 2016, 55, 4748−4763

(105) London, N., Raveh, B., Cohen, E., Fathi, G., and Schueler-Furman, O. (2011) Rosetta FlexPepDock web server–high resolution modeling of peptide-protein interactions. *Nucleic Acids Res. 39*, W249–253.
(106) Lauck, F., Smith, C. A., Friedland, G. F., Humphris, E. L., and Kortemme, T. (2010) RosettaBackrub–a web server for flexible backbone protein structure modeling and design. *Nucleic Acids Res. 38*, W569–575.
(107) London, N., and Schueler-Furman, O. (2008) FunHunt: model selection based on energy landscape characteristics. *Biochem. Soc. Trans. 36*, 1418–1421.
(108) Gregory, K. J., Nguyen, E. D., Reiff, S. D., Squire, E. F., Stauffer, S. R., Lindsley, C. W., Meiler, J., and Conn, P. J. (2013) Probing the metabotropic glutamate receptor 5 (mGlu(5)) positive allosteric modulator (PAM) binding pocket: discovery of point mutations that engender a "molecular switch" in PAM pharmacology. *Mol. Pharmacol. 83*, 991–1006.
(109) Sripakdeevong, P., Kladwang, W., and Das, R. (2011) An enumerative stepwise ansatz enables atomic-accuracy RNA loop modeling. *Proc. Natl. Acad. Sci. U. S. A. 108*, 20573–20578.

DOI: 10.1021/acs.biochem.6b00444
Biochemistry 2016, 55, 4748–4763