Cyrus Bench Homology: A highly automated, cloud-enabled, powerful, and user-friendly homology modeling pipeline using RosettaCM

White Paper

Lucas G. Nivón, Ryan Gomoto, and Yifan Song

November, 2017

Author Affiliations: Cyrus Biotechnology, Inc. 500 Union Street Suite 320, Seattle WA 98101.
Corresponding Author: Yifan Song, info@cyrusbio.com or yifan@cyrusbio.com.

Abstract

We describe a new protein structure prediction tool, Cyrus Bench Homology, offering the RosettaCM package on very large, cloud-based computational resources, and presented to users via an easy-to-use web application. Among recent protein structure homology modeling methods, RosettaCM has emerged as the best-performing tool by various metrics in blinded comparisons. However, in practice RosettaCM depends on a variety of external software to complete its standard pipeline for processing homology models and requires large computing resources to achieve top performance. These software and computational requirements have made a full RosettaCM pipeline very difficult to set up outside of large institutions that can provide thousands of CPUs and coordinate software licensing. Cyrus Bench Homology (available at cyrusbio.com) implements the full RosettaCM pipeline for homology prediction, with all software dependencies and computational resources auto-configured. We provide a practical summary of how to use Cyrus Bench Homology, describe the underlying algorithms with an emphasis on recent improvements and key intellectual antecedents, highlight the results of scientific testing, and apply this tool in a variety of examples. Many software tools for homology modeling explicitly aim for accessibility while sacrificing overall accuracy. Cyrus Bench Homology is novel in offering the most accurate structure prediction tool in a very easy-to-use package accessible to any bench scientist.

Table of Contents:

1. Introduction

The RosettaCM homology modeling pipeline uses a variety of bioinformatics tools to identify protein sequence homologs, predict protein secondary structure, and produce initial three-dimensional alignments (“threading”) of the query sequence to various atomic-resolution template structures. The main RosettaCM application uses these inputs to predict the query output structure for portions where

homology is present, and it uses the Rosetta ab initio knowledge-based protein fragment approach to build structure for portions with no homology [Simons]. Final structures are energy minimized using a Monte Carlo (MC) local structure search algorithm optimized for protein structure prediction, filtered and clustered computationally to identify the most high-quality models [Shortle, 1998]. Scoring of structures during these various steps employs multiple different scoring models, both full-atom and coarse-grained, that use both knowledge-based and physically-derived scoring terms.

Any structure prediction algorithm requires sampling methods to generate hypothetical structures and scoring methods to measure how likely a given structure is (or, in the language of physical chemistry, the relative free energy of each structure). Early generations of structure prediction tools are often based entirely on physically-derived molecular dynamics scoring terms (or potential functions) and on molecular mechanics structure sampling. RosettaCM is different from these types of scoring approaches because it adds knowledge-based score terms (e.g. a hydrogen bonding potential derived from x-ray crystal structures). Sampling in RosettaCM is also different because it uses Monte Carlo (MC; randomly chosen) protein backbone and sidechain methods with knowledge-based MC steps (protein backbone fragments and sidechain rotamers), rather than deterministic molecular mechanics sampling using Newton's equations of motion. This type of sampling allows a very large structural space to be sampled very rapidly, in many cases more rapidly than non-MC approaches [Leach].

2. Modeling with Cyrus Bench Homology: Results and Algorithms

Cyrus Bench Homology implements the full RosettaCM pipeline for homology prediction, with all software dependencies and computational resources auto-configured. It includes the full pipeline used in comparisons of RosettaCM with other tools, without any subtractions or omissions, which would compromise performance. This pipeline is the top-performing software in the continuous fully-blinded CAMEO weekly contest and is ranked first in accuracy by most metrics in the bi-annual blinded CASP competition [Song]. Here we summarize comparative results of RosettaCM and describe the underlying algorithms.

RosettaCM results and comparisons

We focus on two blind and public competitions between various protein structure prediction servers, here using results from the Robetta server that runs the same algorithms and scripts as Cyrus Bench Homology, but on different computing hardware. RosettaCM has the top performance by most structure prediction quality metrics (similarity to the experimental crystal structure) in both evaluations.

The first evaluation is the Critical Assessment of protein Structure Prediction (CASP) which takes place every two years, and we focus on the fully-analyzed results from CASP10; recent results from CASP11 are available but the final publications have not yet been published at this time [Song 2013]. The second is Continuous Automated Model EvaluatiOn (CAMEO), which is run continuously. Most new protein structures submitted to the PDB are presented as structure prediction problems to participating servers before the PDB structures are released publicly. CAMEO results therefore present a running tally of comparative server performance updated every week.

In CASP10 RosettaCM had the most first-place (closest to the correct structure by GDT-TS out of all servers) performances of any server (Figure 1a) [Zemla]. Another, more fine-grained metric of the correctness of a structure prediction is the percentage of hydrogen bonds appearing in the crystal structure that are also found in the structure prediction – because many servers might get large portions of the structure correct, this metric distinguishes higher-quality predictions. The predicted structures can be ranked by fraction of hydrogen bonds correctly predicted (Figure 1b), where RosettaCM is far superior to other servers.

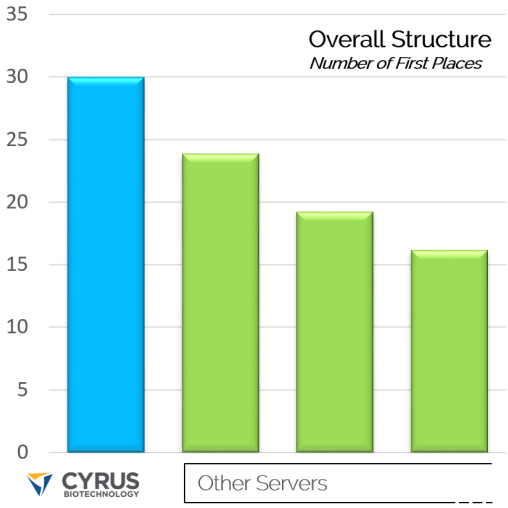


Figure 1a

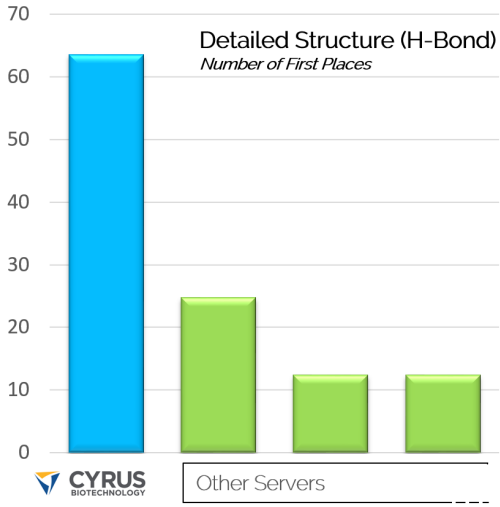


Figure 1b

The second evaluation is the continuous CAMEO comparison. Because it is run continually, we extracted a set of recent results for a sixth month timespan ending October 30, 2015. The average GDT-HA (where higher is better) was calculated for all servers, and RosettaCM was found to have a higher structure quality by this metric than any other server (Figure 2). To put the values in perspective, a 2.5 point increase in GDT HA means that over a 300-residue protein roughly 8 residues are more than 2 Angstroms closer to the correct experimental structure. In practice most of the differences between server predictions are in loops, which are more likely to be near a ligand or protein-protein interaction site.

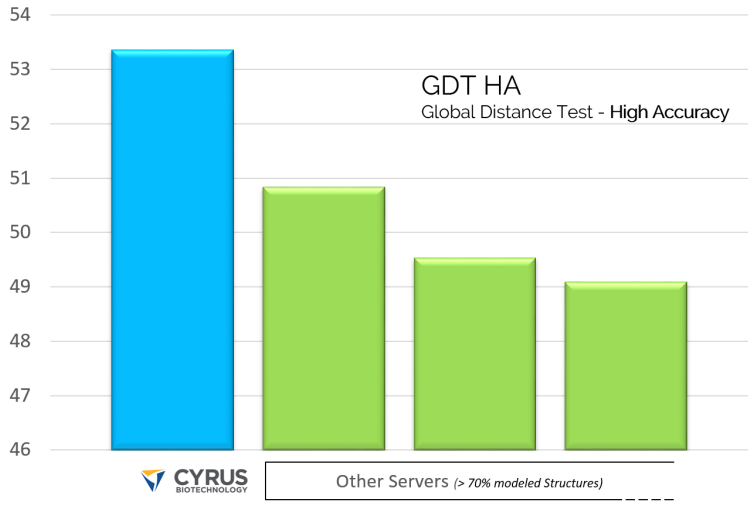


Figure 2

Pipeline Overview

The starting input is a protein sequence of unknown structure, optionally with known similar protein structures that are not included in public databases. The pipeline has three broad steps, all of which are fully automated in Cyrus Bench with minimal user interaction and no setup required:

1. In preparation for a run, a variety of sequence alignment tools are used to identify homologous proteins with structures, a secondary structure is predicted, the input sequence is threaded on to the homologs, and a Rosetta fragment library is custom-created for the input sequence.

2. These are all used as input to RosettaCM, which automatically recombines the various input threaded structures and uses ab initio structure modeling to predict structure of portions that have no homolog structure information.

3. Finally, the hundreds or thousands of RosettaCM outputs are filtered and clustered to give output structures and quantitative analysis of results.

In practice, the use of the most up-to-date RosettaCM software pipeline has been relatively limited because of its complexity, so it is important to outline related practical requirements. First, we note that while it is common to denote this pipeline simply as “Rosetta”, it has a variety of complex software requirements and dependencies - it is part of a much broader software stack, not unlike the common Linux, Apache, MySql, PHP/Python (“LAMP”) Web stack. There are ten distinct steps on a variety of servers, as detailed below in this section. Because of the complexity of setting up so many different inter-operating parts with various licenses and databases, many users employ just portions of the full stack, which will not yield the superior results demonstrated in various public protein structure contests.

The Cyrus Bench Homology backend infrastructure includes all of the dependencies to automate the full RosettaCM stack. In Cyrus Bench, all of these tools run automatically and are auto-configured based on the characteristics of the input sequence for optimal performance. Cyrus Bench is also deployed on a publicly available computing cloud with automatic, rapid provisioning of CPUs as needed. This deployment allows any user of Cyrus Bench to compute arbitrarily many structure predictions in the minimal possible time. Below we describe these steps in greater detail and summarize recent structure prediction results with RosettaCM, especially in comparison with other tools in the CASP and CAMEO evaluations.

Back-end Preparation for RosettaCM in Cyrus Bench Homology

Note that Cyrus Bench includes all software tools described here, with appropriate database setup, software licensing, and correct settings automatically configured. Automatic back-end preparation for a Cyrus Bench Homology run proceeds as follows:

1. HH-suite. This is a suite of protein sequence alignment and analysis tools. The Cyrus Bench Homology pipeline employs the HHpred tool to identify protein homologs with structures, using Hidden Markov Models (HMM) to perform a profile-to-profile search [Hildebrand].

2. BLAST. Basic Local Alignment Search Tool. The very powerful and ubiquitous sequence similarity search program is not directly used by RosettaCM, but is called by other sequence alignment tools. [Altschul, Johnson]

3. SparksX. This tool identifies very deep (low similarity) protein structure homologs that are often not found by other tools, and has been the top fold recognition performer in CASP. It uses a weighted matching of multiple sequence profiles, including ones from multiple sequence alignments, predicted versus actual secondary structures, a knowledge-based score function, depth-dependent sequence profiles, solvent accessible surface area, and dihedral angles. [Yang]

4. Alignment. Previous steps have generated a variety of different alignments to various homologs from a few different tools. This step combines those alignments to give the maximum likelihood of a correct structure prediction and assigns weights to each alignment. These alignment combinations have been tuned on large experimental databases to maximize likelihood of correctness. These scripts are distributed with Rosetta and have been customized in Cyrus Bench.

5. Threading. After earlier steps identify the set of structural homologs, these scripts place the input sequence amino acids into the three-dimensional model of each homolog, “threading” this input sequence through the backbone of each existing structure. These are designed to run on the Cyrus Bench architecture.

6. Constraint generation. RosettaCM uses C-alpha to C-alpha constraints from the threaded structures to guide structural sampling. These constraints are generated from the weighted input alignments. For short alignment gaps, the background distance distribution is used [Thompson, 2011]. If a given template has a gap longer than 50 residues, the contribution from that template at those positions is removed. The scripts generate and format all of the required constraint files. These scripts have been custom-designed for the Cyrus Bench architecture.

7. Fragment picking. This Rosetta application uses the input sequence and secondary structure prediction to identify 3-mer and 9-mer fragments for every position in the protein sequence. These fragments are short protein structures pulled from a pre-filtered version of the Protein Data Bank (PDB). The output fragment files are used as the fragment library input to RosettaCM. Note that this Rosetta application has gone through iterations of updates over the years and Cyrus Biotechnology applies the latest, extensively benchmarked iteration.

Cyrus Bench Homology Algorithms

RosettaCM, as implemented in the Cyrus Bench Homology pipeline, uses the most recent “hybridize”

protocol developed by Song and co-workers [Song]. As described above, inputs to RosettaCM are sequence-based fragments, threaded template homolog structures, and constraints generated from those templates.

RosettaCM proceeds in three major steps:

1. Full length model assembly: Complete models are constructed by recombining fragments from the aligned template structures, with fragments from the fragment library used in unaligned portions. Fragments are assembled using torsion space fragment insertion and by Cartesian space template segment recombination.

2. Local structure optimization and loop closure: Geometrical changes from the templates are explored and gaps between aligned regions or between aligned and unaligned regions are modified. This is carried out by local fragment superposition (larger geometric changes) and by gradient-based energy minimization (smaller changes or closing of small gaps).

3. Full-atom sidechain and backbone refinement: This step uses Rosetta full-atom refinement with rotamer-based sidechain repacking and gradient-based backbone and sidechain minimization runs, iteratively.

Cyrus Bench Homology data collection and analysis

The standard scripts available with Rosetta for result collection, analysis, clustering and data plotting are implemented on the Robetta server, and designed for the BOINC distributed computing system [Anderson]. The Robetta server provides a simple non-GUI web interface to the RosettaCM pipeline, but it is only available to non-commercial users, calculates one job per user at a time, displays results publicly for all to view, and takes weeks to complete most jobs [Kim, 2004]. Cyrus Bench uses a set of re-factored scripts that implement similar algorithms on any public cloud system and is deployable onto nearly any server architecture with an OpenStack software layer (compatible with Google Compute Engine, Amazon Web Services, etc.).

RosettaCM generates hundreds or thousands of output PDB structures with associated Rosetta scoring files, either as “silent files” or as standard PDB format files. Each job has to be tracked and data collected if the job finishes with correct, qualified output – this task is automatically accomplished inside Cyrus Bench from all CPUs, with error checking.

Cyrus Bench filters and clusters the final structures and score terms to take the top 500 output structures (best overall Rosetta all-atom score). It then presents the overall best-scoring structure as the top structure. Then, it presents the best-scoring structure from the next cluster as the next structure, and so on. A standard output reports the top five unique best structures by cluster. Scatter plots of all output structures with various output scores are also prepared by Cyrus Bench (this is the standard method to analyze outputs), so the user can evaluate the convergence of top structures, the variation in total scores produced by RosettaCM, and variation in individual score terms (e.g. Hydrogen bonding) in the entire output set. These data help an expert user to understand the quality of structure predictions from a given run.

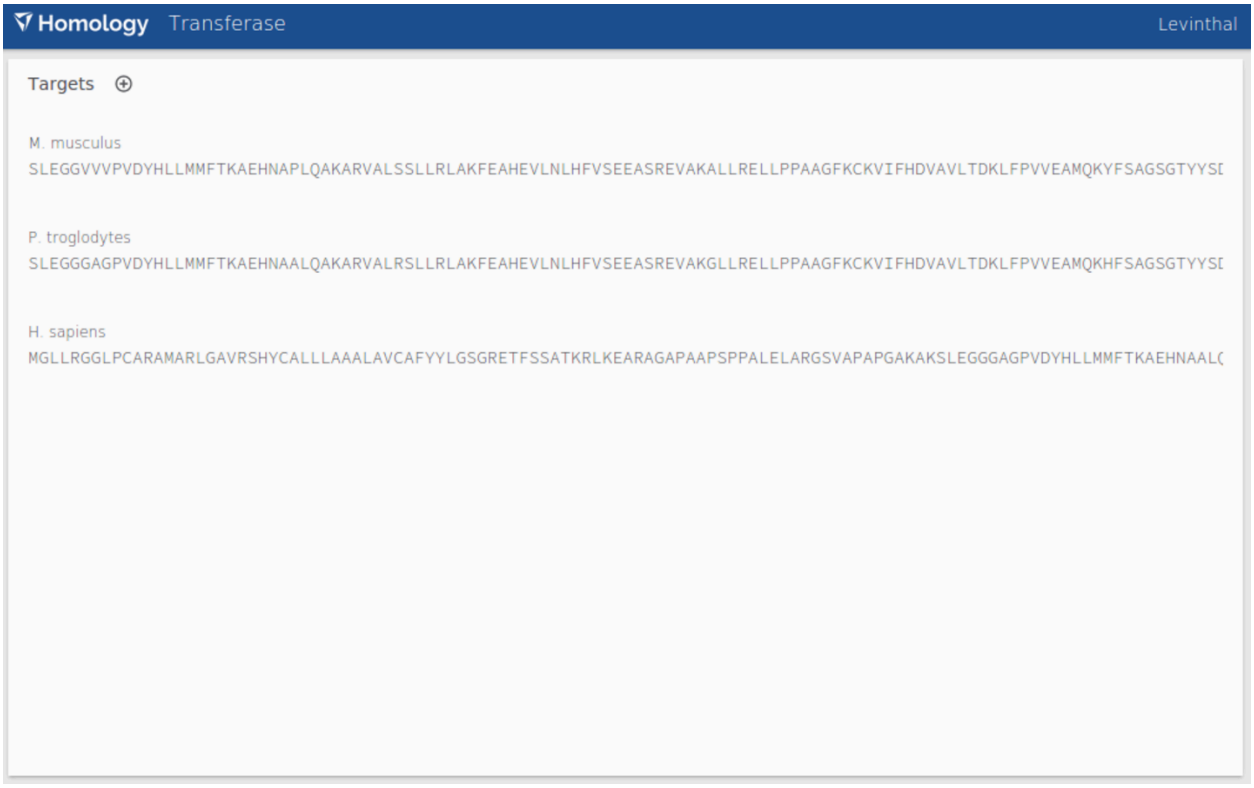
3. Cyrus Bench Homology Tutorial: Model submission, processing and result analysis

Before You Start

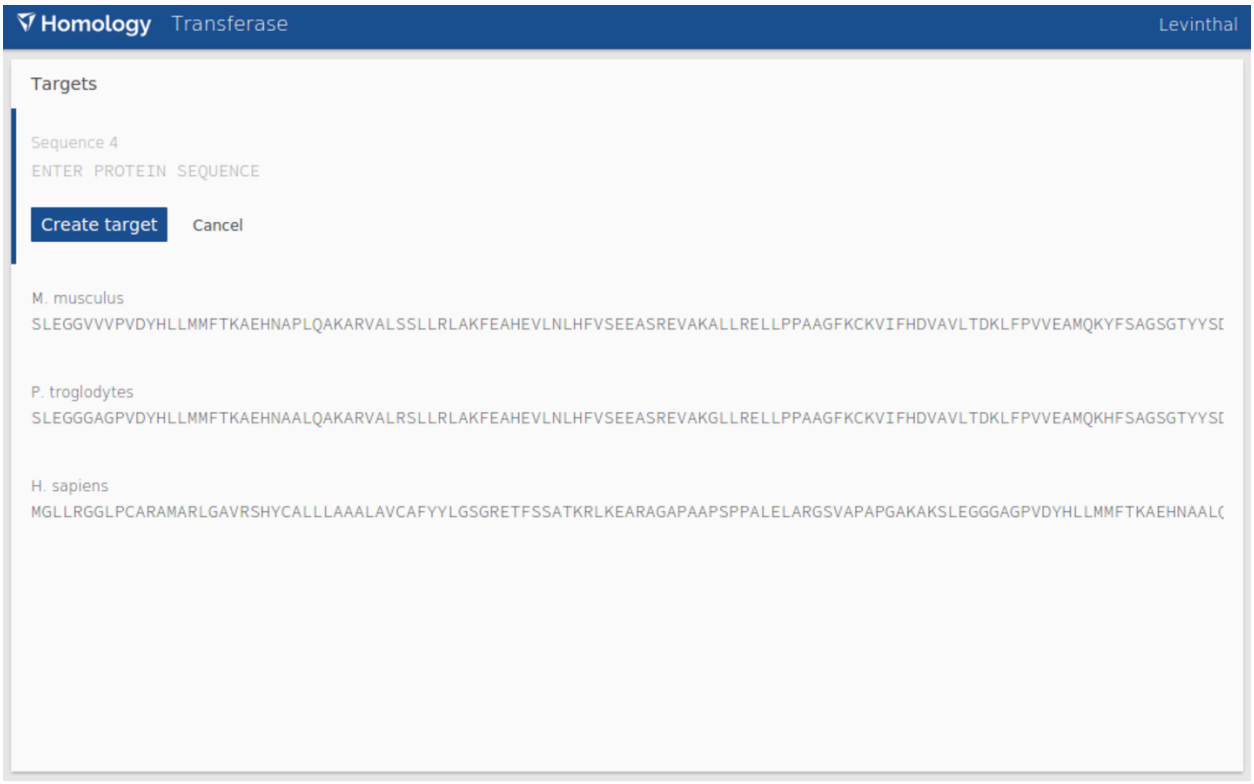
Cyrus Bench is a web-based application and requires no installation. New users sign up to receive a user name and then they set their own password, in a similar fashion to other web servers such as email servers. All computations and data storage are managed inside the Cyrus cloud-based servers. Inputs to Cyrus Bench Homology are raw protein sequences, for example from a FASTA format file.

A Homology Run

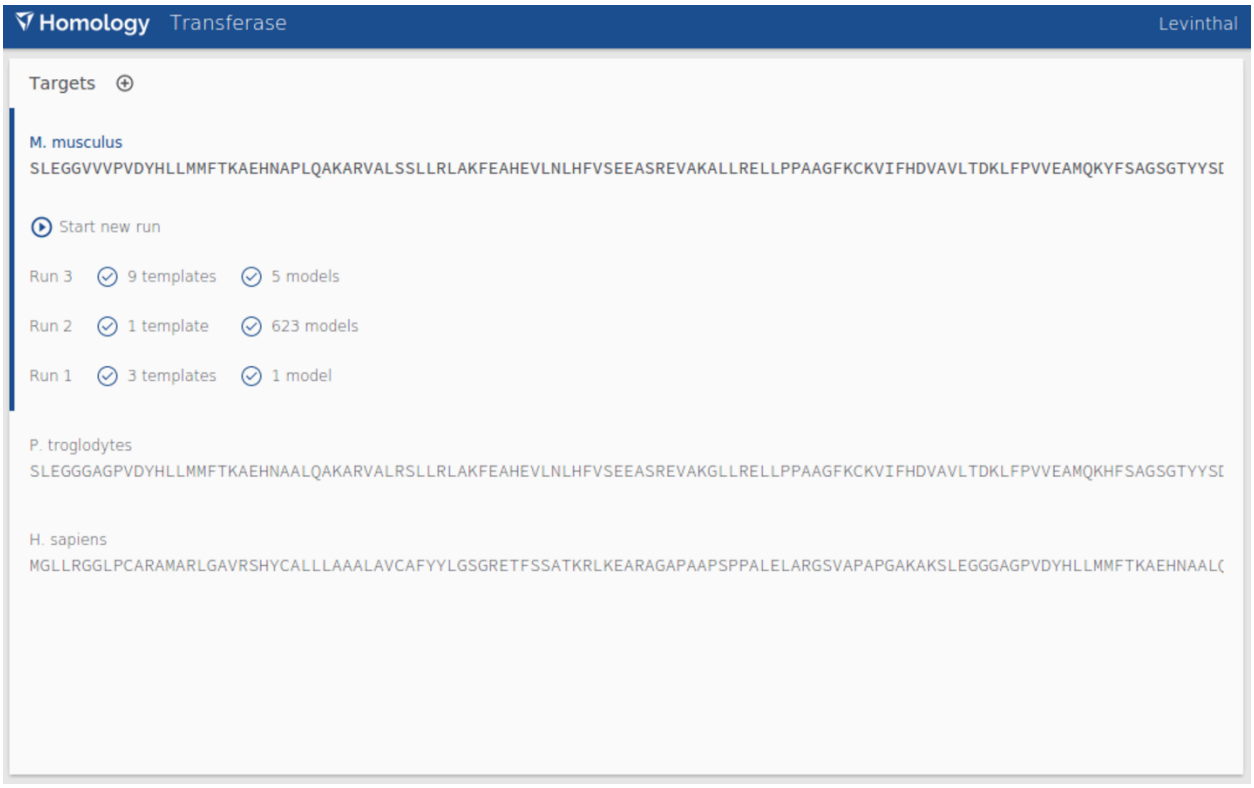
1. Once logged in, the user creates a new session (Figure 1). Each session contains a variety of "Targets" or input sequences; the Targets in a session may or may not be conceptually related to one another - the session concept simply allows users to organize their work prior to exporting results. The example session here is called "Transferase", the user name is "Levinthal" and we can see a few Targets that have already been entered. In this example session the user has already entered three targets before, so we see three targets in the list upon initially logging in.



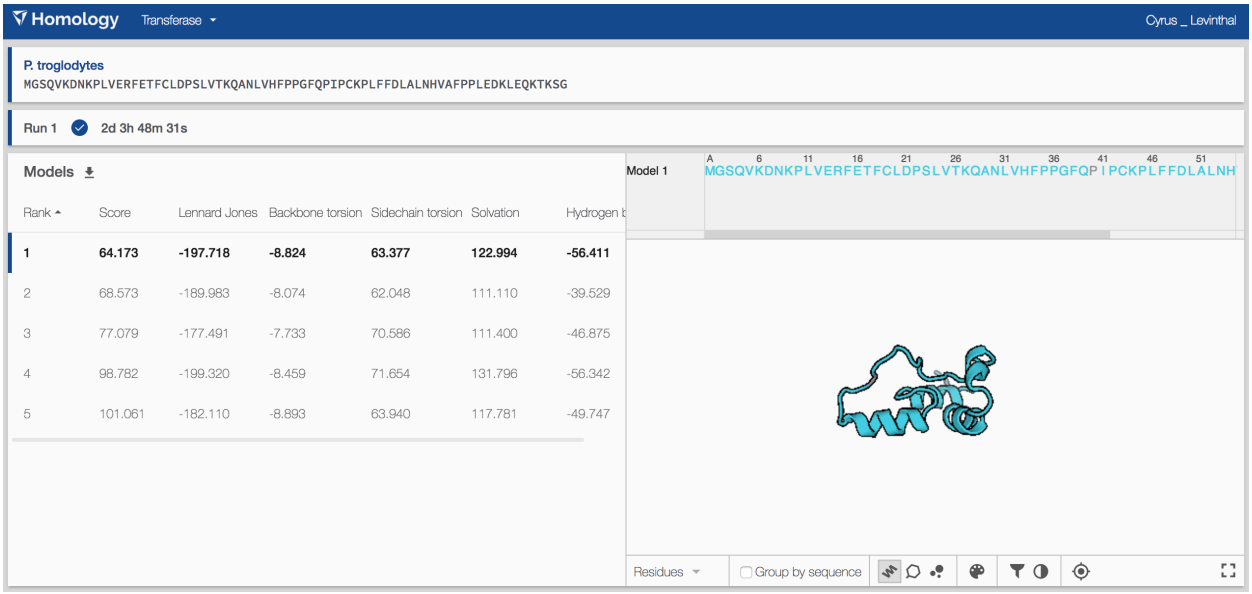
2. A new target is created by simply pressing the "+" button, pasting in the sequence of unknown structure, and clicking "create target" (Figure 2).



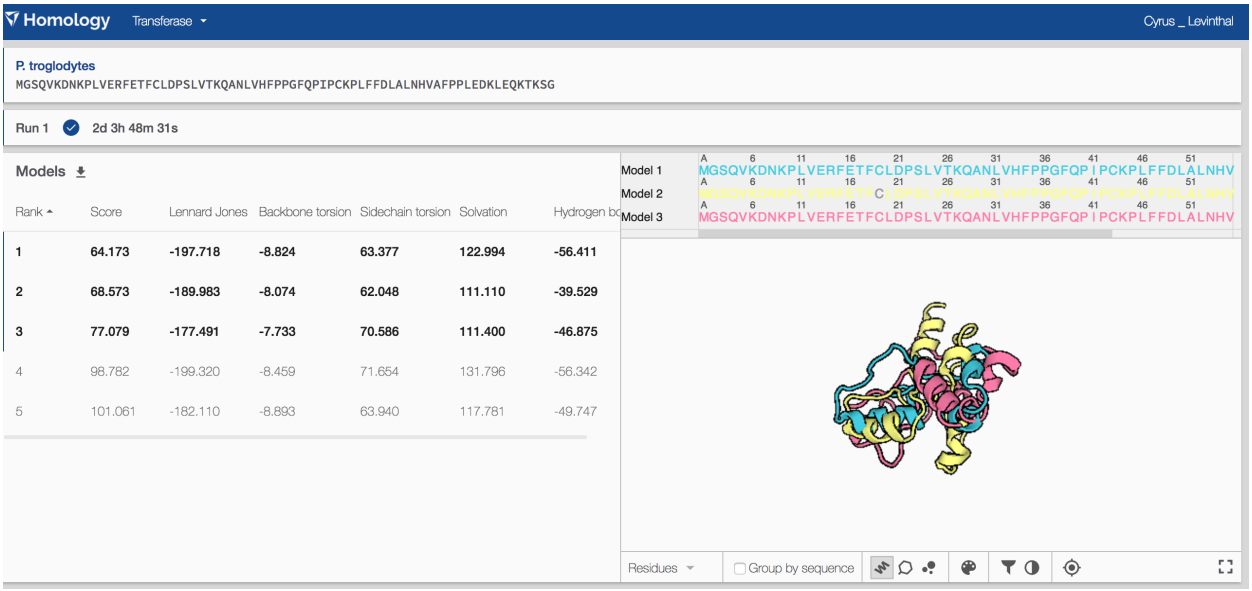
3. This initiates the first few steps of the homology pipeline, performing multiple sequence alignments, identifying templates, and preparing derivative data from those templates. The user can keep track of progress during a run, first noting the number of templates found, and then the number of models generated as RosettaCM proceeds. Here we show a run on a mouse xylosyltransferase sequence (Figure 3) with 9 templates and the default best 5 output models.



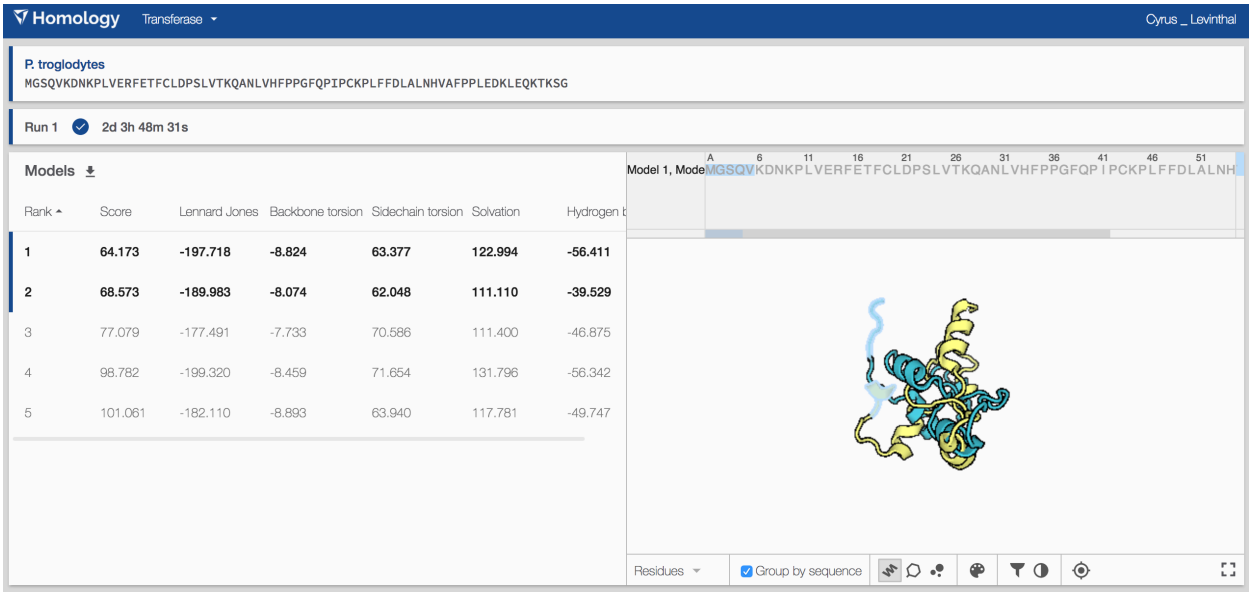
4. Once the run is finished, clicking anywhere on the relevant “run” line brings up the full output viewer (Figure 4). In the lower right, a structure viewer appears showing the default overlay of top template and top model. In the lower left, this run view also brings up the list of top models, based upon a clustering algorithm also used with CASPR and CAMEO. Rank number one is most likely to be closest to the correct model. Scoring categories are displayed which are used to calculate the final Score, lower scores being better than higher ones. Note that the user may decide to re-run the same sequence, removing some of the templates, which would create a new “run” for that Target.



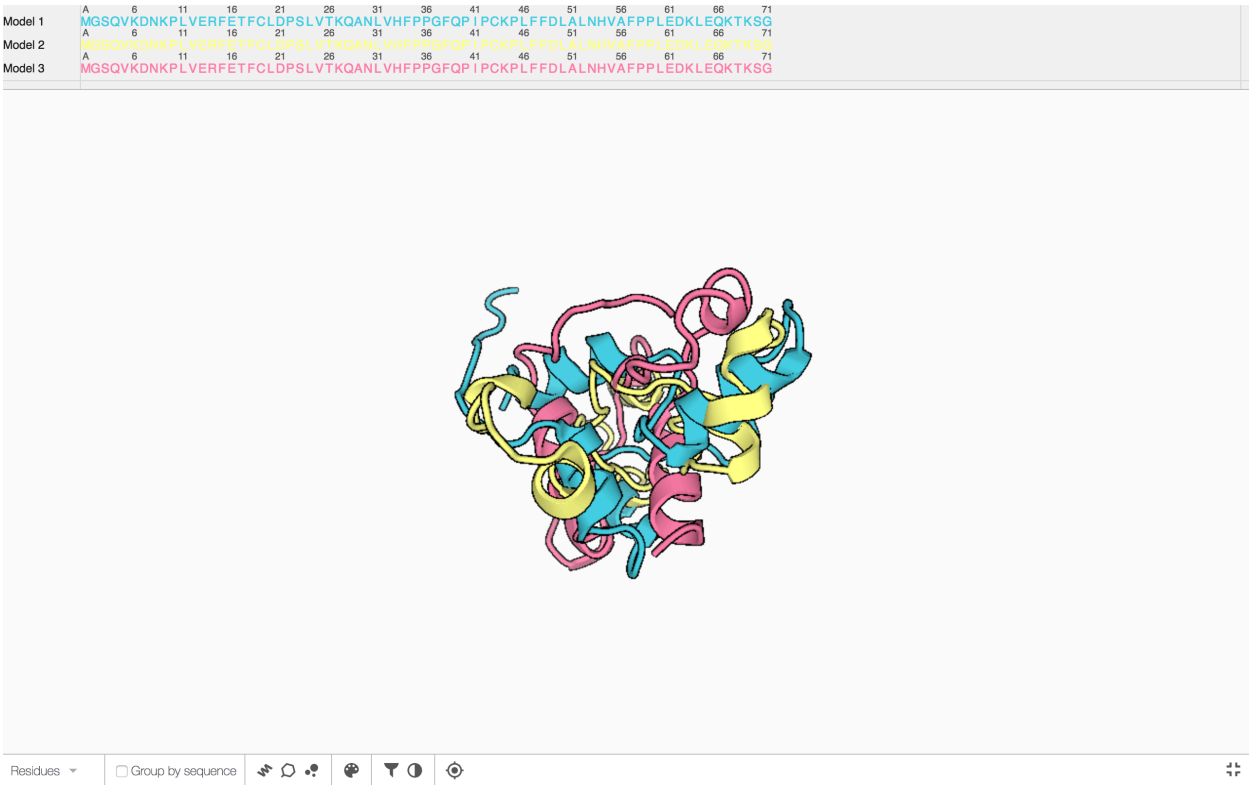
5. The user can also scroll through viewing of the output models, each shown in a different color (Figure 5). The models can be simultaneously displayed in the structure viewer (clicking on a model displays it in the viewer). In the figure, we show the top 3 models. Again, based on the use of this model, users might prefer a given model with local structure similarity over another model.



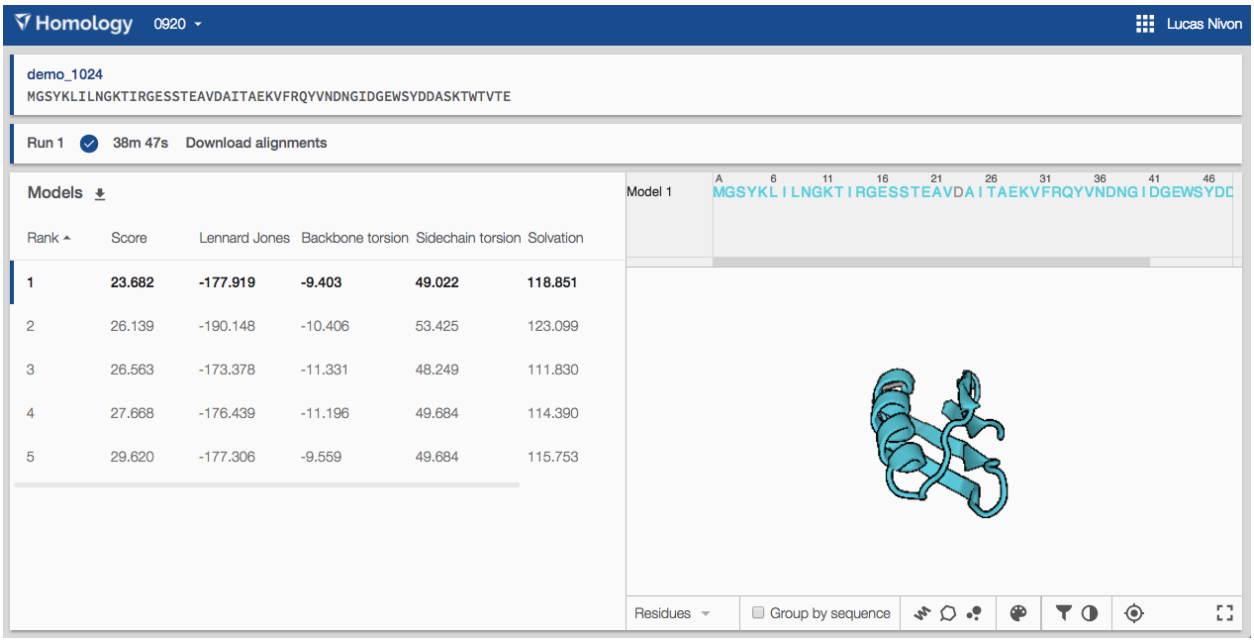
6. The user can also group by sequence, and make selections to see what structure each protein has at a specific position (Figure 6).



7. More fine structural analysis is possible in the zoomed mode, accessed via the zoom button in the lower-right corner of the structure viewer window (Figure 7).



8. More detailed information about the various templates used in modeling is available at the Download Alignments button, which downloads all of the templates aligned to your input sequence in a json file format that is easily editable in a basic text editor or usable as input for your favorite scripting language. See the support page “**How can I analyze output structure quality from HM?”** for more detailed information on interpreting output structures from HM, this document is focused on algorithmic and process detail in HM.



Cyrus Bench Homology is designed to be easy to use with very little training, with most of the interface focused on providing easy structure viewing and comparison for the biochemist.

4. Homology modeling background

Structure prediction with homologs: History

Protein structures are now known for at least one representative of most protein families, and homology modeling can be used to build structures for proteins in these families using the structures available in the PDB or other structures. In the most generic case, a first step for modeling is to identify close sequence homologs with known structures, and a second step is to derive structures for the unknown protein based on one or more of these solved experimental structures. In the second step it is common to align the unknown sequence with the homolog structures in 3-dimensional space, a process known as threading, prior to further structural optimization.

Many very useful methods for comparative modeling have been developed, that incorporate: 1)

Threading on to homologs, 2) Threaded structure optimization with molecular dynamics (MD), 3) Derivation of distance restraints from homology models and optimization via restraint-satisfaction, 4) Use of fragment-based approaches for regions of low homology or no homology. RosettaCM uses most of these concepts, but does not use molecular dynamics, instead using Monte Carlo based sampling over protein backbone and sidechains. RosettaCM is highly tuned on a variety of benchmark test sets to optimize structure prediction, using a careful balance of various approaches to sample and score hypothetical structures.

One widely used tool is the MODELLER program, which uses a set of spatial restraints derived from the homolog structures to build new models that maximally satisfy these restraints [Sali]. More recently, the I-TASSER tool [Xu et al] has been introduced. It combines atomic-resolution information from multiple protein templates. This software also uses de novo structure prediction algorithms to supplement homologous structures for distant targets, using the QUARK tools, and incorporates multiple threading alignments and molecular dynamics along with short structural fragments collected from the PDB.

Structure prediction without homologs: ab initio

The concept of local sequence/structure correlations is critical to understanding algorithms based on protein structure "fragments" – these algorithms form the basis for structure prediction and optimization across many homolog modeling tools including RosettaCM.

Local protein structural motifs, such as helices and beta strands, have been observed since the very first atomic-resolution crystal structures of proteins. Since then, a variety of more specific sequence patterns for more idiosyncratic protein structures (such as tight turns or helix caps) have been developed [Hutchingson & Thornton, 1994]. More generally, it has been hypothesized and verified experimentally that a wide variety of local sequence patterns give recurring three-dimensional structures – early work used cluster analysis on protein sequences to identify local sequence motifs that are found across protein family boundaries [Han and Baker, 1995, 1997]. Bystroff and co-workers used the structures formed by these various local sequence motifs to discover a set of sequence profiles with associated structures [Bystroff].

These local sequence/structure ideas were applied to protein structure prediction by deriving a set of short structural fragments with similar local sequence to the unknown structure, and then combining those fragments in a simulated-annealing protocol. This method was shown to create native-like conformations that could be filtered to correctly identify close-to-native structures using database-derived Bayesian scoring functions [Simons]. This was the first application of the structure fragment concept (based on local structural preferences inherent to a given amino acid sequence) to predict protein structures, and this method forms the basis of ab initio structure prediction in Rosetta. In RosettaCM, a version of this fragment assembly method is used to produce structure predictions for unaligned sequence regions.

5. Applications of Cyrus Bench Homology

The Cyrus Bench Homology application, partially described earlier in this manuscript, is the first in a series of highly automated, integrated protein modeling and design web-based (GUI) applications based on RosettaCM. Over time, Cyrus Biotechnology will offer all the elements of the applications described below in this section for rapid dissemination with minimal setup throughout the academic, commercial and government scientific communities.

The most recent homology modeling pipeline in the Rosetta modeling toolkit has a wide variety of demonstrated uses, including helping to solve difficult or impossible cryo-EM and X-ray crystal structures, higher-quality structure predictions based on co-evolutionary data, and use in subsequent small-molecule drug discovery or protein design. Some of these applications, such as iterative homology modeling during molecular replacement, precede the most recent "hybridize" HM protocol, but the recent improvements are making nearly all applications of the Rosetta HM tools more likely to include higher quality models

[Song].

The modular nature of the Rosetta toolkit, with a core set of highly inter-operable software classes, makes the creation of a wide variety of stable derivative applications efficient from a software engineering point of view [Leaver-Fay Rosetta architecture]. The highly decentralized consortium structure of the RosettaCommons software development community builds upon this engineering modularity to create very rapid diversification and scientific testing of the wide variety of applications described here.

a) Homology modeling for Structure Based Drug Design, small-molecule docking, and enzyme characterization

In small-molecule drug development, homology modeling is often used to build structures of protein targets when X-ray crystal structures are not available. These structures can then be used for rational modifications to the ligand in Structure Based Drug Design (SBDD), or for further computational analysis such as ligand docking, virtual screening, or ligand-fragment-based diversification of molecular structures. In general, the higher the quality (accuracy measured against the true crystal structure) of a homology model, the better the expected results for any of these downstream applications in drug discovery. Therefore, in most cases the best SBDD and other outputs would be expected when using homology models from Cyrus Bench Homology.

Here we highlight a pair of recent examples of this use case to both benchmark this workflow and for a real drug target discovery application. A set of novel homing endonucleases were identified and characterized with bioinformatics data mining followed by high throughput activity screening, with characterization of substrate specificity and kinetics through deep sequencing. The sequence specificities demonstrated by these experiments were partially recapitulated computationally using structure-based models from RosettaCM [Thyme].

A novel fatty acid double-bond hydratase from Lactobacillus plantarum was identified in the membrane fraction of lactic bacteria, with the ability to hydrate linoleic acid to a less toxic form. The hydratase structure was predicted using RosettaCM, identifying two putative binding sites for the linoleic acid substrate [Ortega-Anaya].

b) Solving difficult or impossible X-ray phasing and structure solving problems

The RosettaCM homology modeling protocols can be combined with X-ray diffraction data to solve crystal structures from poor resolution data or to solve structures for otherwise impossible molecular replacement problems [DiMaio Nature 2011]. These tools are designed to work with Phenix, starting with a RosettaCM homology modeling and iteratively rebuilding initial models into density. Future versions of Cyrus Bench Homology will allow a user to upload electron density to run each of these single rounds, and to integrate automatically with Phenix and other X-Ray data analysis tools to automate the entire workflow, removing laborious manual steps and the attendant possibility for error at those steps.

Early work showed that very accurate structures could be obtained from very sparse NMR data sets, by adding the experimental data into RosettaCM to guide structure search and scoring, with gradual improvements in methodology over a few years [Raman, 2010]. This philosophy of combining experimental data with RosettaCM was then applied to X-ray datasets, where electron density maps from molecular replacement solutions for a set of initial models are used to guide structure rebuilding, sidechain packing, and minimization [Das, 2008]. New maps are derived using phase information from the lowest-energy resulting models that are also consistent with X-ray data and automatically chain traced, using R-free to monitor goodness of fit to the raw data [Bruenger, 1992].

More recently, these methods have been tested on 13 real X-ray diffraction data sets that could not be solved in expert crystallography labs, and were still unsolved after applying a set of alternative methods. This combined approach using RosettaCM with Autobuild chain tracing gave high-resolution structures for 8 of the 13 otherwise-impossible structures. This combined iterative method should converge to high-quality structures in roughly 50% of cases where: 1) There is no experimental phase information 2) Data sets better than 3.2 Angstroms resolution, 3) Four or fewer copies in the asymmetric unit, and 4) Identification of homologous structures with greater than 20% sequence identity [diMaio Nature 2011].

Improvements in the underlying RosettaCM algorithms in the more recent “hybridize” approach, particularly around unaligned or missing backbone segments, continue to yield improvements in the success rate and accuracy of structures solved with this method [dimaio Acta, 2013].

It was recently shown that methods can also be used for X-ray fiber diffraction data sets, where data can be obtained for highly oriented fibrillar molecules, but structure determination is very difficult [Potrzebowski 2015]. This modified RosettaCM approach was used to determine the structures of six bacteriophage viruses together with solid-state NMR data, and by solving the structure of a plant virus. Models generated in this way fit well to experimental data and have improved structure quality over other methods [Potrzebowski].

c) Solving difficult or impossible atomic resolution structures from cryo-EM data

Modern cryo-EM data collection methods, including direct electron detection and new data analysis techniques, are making resolution lower than 8 Angstroms accessible. If this data is below 3-3.5 Angstroms, it is of high enough resolution to provide atomic-level-accuracy models [Zhang 2015]. In all other intermediate cases, electron density data from EM can be used along with RosettaCM to produce atomic-resolution models in the 4 to 6 Angstrom electron density resolution range [DiMaio Nat Methods 2015].

The fit between the model and the electron density is introduced into the RosettaCM protocol as a score term, and the homology protocol is performed iteratively with real-space B factor fitting. The quality of final models can be assessed by splitting the input density data into two sets and cross-validating models against the independent test set not used to derive the model (conceptually similar to the free R factor metric in X-ray crystallography) [DiMaio, proteins 2013]. The method using the RosettaCM “hybridize” approach was tested on three systems with 4.5-Angstrom or better resolution. It produced models with atomic-level accuracy for all systems regardless of initial model quality, and outperformed the MDFF method based on molecular dynamics [DiMaio Nat Methods].

Full structure determination of unknown structures requires model building into density at the 3 to 5 Angstrom range. Starting from RosettaCM, a predictive model-building approach with fitting into density is able to construct new backbone where input models are lacking. This modified RosettaCM algorithm gave accurate models for six out of nine test cases with density in the range from 3.3 to 4.8 Angstroms, and produced an almost complete model for a heterodimeric 660 residue protein [Wang, Nature Methods]. Together, the “hybridize” RosettaCM method, new model building algorithms, iterative refinement into cryo-EM density, and cross-validation assessment methods allow the route determination of atomic-resolution structures from cryo-EM density maps in the 3 to 5 Angstrom range.

The iterative cryo-EM method was used to derive density-consistent atomic-level models of Tubulin during various phases of GTP hydrolysis, revealing movements of individual helices upon hydrolysis [Alushin]. This level of detail would not have been accessible with traditional techniques that simply dock homology models into cryo-EM density [Alushin].

d) Homology modeling with RosettaCM for rational enzyme engineering or novel enzyme discovery

Rational enzyme engineering approaches require either a crystal structure with bound substrate/analog, or high-quality homology models. In general, if experimental structures are not available, Cyrus Bench Homology can provide the highest quality models for further rational engineering, especially around the active site, or for computational protein design to alter or improve activity. This approach can be generalized, with massively parallel homology modeling over a set of sequences in an enzyme family, followed by computational ligand docking to identify the most likely proteins with desired activity. This latter approach provides a rapid way to diversity across the activities of an evolutionary family of enzymes, increasing the effective search space in an enzyme engineering project by combining evolutionary information with computational modeling.

A simple application of homology modeling is the creation of a model of an enzyme of unknown structure to analyze enzyme activity. This approach was taken using a slightly simplified version of the RosettaCM protocol to model a toluene monooxygenase. Previous attempts to engineer alkene

monooxygenases to perform a desired ethylene oxidation by traditional protein expression and engineering were unsuccessful. A set of mutant toluene monooxygenases from a screen for a related reaction by Wood and colleagues were tested experimentally, identifying mutants capable of catalyzing oxidation at greater than 99% yield. The homology models for the top mutants reveal that the most activating mutations are at the active site above a diiron center [Carlin].

Homology modeling followed by ligand docking was used to virtually test a large set of 239 bioinformatics-identified putative enzymes for longer-chain alcohol production from sugar [Mak]. The top-scoring ligand docked structures were screened experimentally, identifying a set of ketoacid decarboxylases capable of using a C8 substrate. This is an alternative to re-design of an enzyme with similar substrate specificity using Rosetta enzyme design capabilities, which can also identify an active enzyme for the desired alcohol production [Mak]. The next step in such a procedure would be the combination of these capabilities, discovering plausible natural enzymes by homology modeling with ligand docking, followed by re-engineering of these wild-type enzymes with Rosetta enzyme design.

e) RosettaCM with co-evolutionary data for improved homology models, high-accuracy models without homologs, and protein/protein interface modeling

Protein sequence alignments can be used to construct protein structure models even without any solved x-ray structures of protein homologs [Marks et al, 2011]. These methods use residue-residue co-variation to identify putative pairwise contacts between residues in three-dimensional space. This type of pairwise data can be combined with the RosettaCM homology pipeline to build much higher quality structures than either method alone. In general, surprisingly few contacts, as few as 1 pairwise contact per 12 residues, are required to produce robust predictions of protein topology [Kim, 2014 proteins].

The structure predictions made with RosettaCM and GREMLIN for proteins over 100 amino acids in length lacking any known homologs were the most accurate predictions made in the 20 years of CASP experiments [Ovchinnikov, 2015 elife]. Co-evolutionary methods had been used to correctly infer residue-residue contacts [Marks et al, 2011], and recent methods have shown improved accuracy using a pseudo-likelihood (PLM)-based algorithm implemented in the GREMLIN software package [Balakrishnan et al, 2011, Kamisetty et al, 2013]. The recent no-homolog blind predictions produced models as low as 2.1 Angstroms RMSD for a 245 residue protein, and these structures would have been impossible to solve using either RosettaCM alone (no co-evolutionary data) or the pre-“hybridize” RosettaCM algorithms [Ovchinnikov, 2015, elife]. This method was able to generate converged structure predictions for 58 protein families where no structure has ever been solved [Ovchinnikov, 2015, elife].

The co-variance approach can also be used to predict the structures of protein/protein complexes. It was hypothesized that residue covariance between different proteins in a genome could correctly identify contacts between those proteins [Ovchinnikov, 2014]. The putative contacts identified using GREMLIN in the 50S ribosome and other large protein complexes are nearly always real contacts, provided that the number of aligned sequences is greater than the average length of the two proteins. This method was used along with RosettaCM to predict the structures of 36 bacterial protein complexes, where the number of aligned sequences tends to be far higher than in eukaryotes [Ovchinnikov, 2014].

These co-variance methods will be introduced as a “push button” tool in Cyrus Bench Homology, and can currently be introduced with user-defined restraints. Distance restraints derived from multiple sequence alignments can be generated and imported into Cyrus Bench Homology as a set of user-defined restraints in the current pipeline. Future versions of the software will include the ability to search for co-evolutionary signal and add these constraints to Homology if enough alignments are identified, according to the heuristics described in Ovchinnikov, et al. [Ovchinnikov, 2015, elife].

6. Discussion

RosettaCM has been the most accurate protein structure prediction tool for homology modeling over a number of public contests for many years, including the most recent CASP and CAMEO blinded

evaluations. There are many tools available for homology modeling, and these involve various tradeoffs, which can be broadly categorized in the areas of accuracy and ease-of-use/features.

Cyrus Bench Homology allows for a no-compromise solution that was previously not available, filling an important niche in protein structure prediction.

The table below characterizes the accuracy and ease-of-use for the various existing tools, including Cyrus Bench Homology. For ease-of-use we designate any tool with has a graphical user interface with integrated viewing and analysis as "highest", a simpler graphical interface (but not command-line interface) as high, a command-line interface with one main component as "medium", and a command-line interface with many other dependencies that also have command-line interfaces as "low". Note that some commercially available tools (e.g., Schrodinger and CCG) do not participate in the public blind contests, and it is therefore impossible to objectively assess their accuracy relative to other tools in a peer-reviewed and unbiased way.

	Accuracy	**Ease-of-Use**
Cyrus Bench Homology	Highest	Highest
Rosetta CM on Robetta (*)	Highest	Medium
Rosetta CM	Highest	Low
iTasser	High	Medium
HHPred	Medium	Medium
SwissModel	Medium	High
MODELLER	Low	High

(*) Robetta is for Academic Use Only

References

Alushin, G.M., Lander, G.C., Kellogg, E.H., Zhang, R., Baker, D., and Nogales, E. High-resolution microtubule structures reveal the structural transitions in αβ-tubulin upon GTP hydrolysis. 2014. Cell 157, 1117–1129.

Altschul, S, W Gish, W. Miller, E Myers, D Lipman. Basic local alignment search tool. 1990. Journal of Molecular Biology. 215:403-410.

Anderson, DP. BOINC: A system for public-resource computing and storage. 2004. GRID '04 Proceedings of the 5th IEEE/ACM International Workshop on Grid Computing. Pages 4-10.

Balakrishnan S, Kamisetty H, Carbonell JG, Lee SI, Langmead CJ. 2011. Learning generative models for

protein fold families. Proteins 79:1061–1078.

Bruenger, A. T. Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. 1992. Nature 355:472–475.

Bystroff, C; Baker, D. Prediction of local structure in proteins using a library of sequence-structure motifs. 1998. Journal of Molecular Biology, 281: 565-77.

Carlin, DA, SJ Bertolani, and JB Siegel. Biocatalytic conversion of ethylene to ethylene oxide using an engineered toluene monooxygenase. 2015. Chem. Commun. 51:2283-2285.

Das, R. & Baker, D. Macromolecular modeling with Rosetta. 2008. Annu. Rev. Biochem. 77:363–382.

Di Maio F, TC Terwilliger, RJ Read, A Wlodawer, G Oberderfer, U Wagner, E Valkov, A Alon, D Fass, HL Axelrod, D Das, SM Vorobiev, H Iwa, PR Pokkuluri and D. Baker. 2011. Nature 471: 540-545.

DiMaio, F. Advances in Rosetta structure prediction for difficult molecular-replacement problems. 2013. Acta Crystallographic Section D. 69:2202-2208.

Di Maio, F., J. Zhang, W Chiu and D. Baker. Cryo-EM model validation using independent map reconstructions. 2013 Protein Science 22: 865-868.

DiMaio, F, Y Song, X Li, MJ Brunner, C Xu, V Conticello, E Egelman, TC Marlovits, Y Cheng and David Baker. Atomic-accuracy models from 4.5-Å cryo-electron microscopy data with density-guided iterative local refinement. Nature Methods 12: 361-365.

Han, K. F. & Baker, D. Recurring local sequence motifs in proteins. 1995. J. Mol. Biol. 251:176-187.

Han, K. F., Bystroff, C. & Baker, D. (1997). Three dimensional structures and contexts associated with recurrent amino acid sequence patterns. Protein Sci. 6, 1587-1590.

Hildebrand, A, M Remmert, A Biegert, J Soeding. Fast and accurate automatic structure prediction with HHpred. 2009. Proteins: 77(Suppl 9):128-132.

Johnson, M, I Zaretskaya, Y Raytselis, Y Merezhuk, S McGinns, TL Madden. 2008. NCBI BLAST: a better web interface. Nucleic Acids Res. 36(Web Server issue):W5-9.

Kamisetty H, Ovchinnikov S, Baker D. 2013. Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. Proceedings of the National Academy of Sciences of USA 110:15674–15679.

Kim, DE, D Chivian, D Baker. Protein structure prediction and analysis using the Robetta server. 2004. Nucleic Acids Res. 2004 Jul 1; 32(Web Server issue): W526–W531.

Kim, DE, F. DiMaio, RY Wang, Y Song, and D. Baker. 2014 One contact for every twelve residues allows robust and accurate topology-level protein structure modeling. Proteins: Structure, function and bioinformatics. 82:208-218.

Leach, AR. Molecular Modeling Principles and Applications. 2nd edition. 2001. Essex England: Pearson Education Limited. Chapter 8, Section 13.

Mak, WS, S Tran, R Marcheschi, S Bertolani, J Thompson, D Baker, JC Liao and JB Siegel. Integrative genomic mining for enzyme function to enable engineering of a non-natural biosynthetic pathway. 2015. Nature Communications:6, 10005.

Marks CS, Colwell, LF, Sherida R., Hopf TA, Pagnani A, Zecchina R, Sander C. 2011. Protein 3D structure computed from evolutionary sequence variation. PLOS ONE 6:e28766.

Ortega-Anaya, J and A Hernandez-Santoyo. Functional characterization of a fatty acid double-bond hydratase from lactobacillus platarum and its interaction with biosynthetic membranes. 2015. Biochim. Et Biophys. Acta – Biomembranes. 1848: 3166-3174.

Ovchinnikov S, Kamisetty H, Baker D. 2014. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. eLife 3:e02030.

Ovchinnikov S, L Kinch, H. Park, Y. Liao, J. Pei, DE Kim, H. Kamisetty, NV Grishin, D. Baker. Large-scale determination of previously unsolved protein structures using evolution information. eLIFE 2015;4:e09248.

Potrzebowki, W, and E Andre. Automated determination of fibrillar structures by simultaneous model building and fiber diffraction refinement. 2015. Nature Methods 12:679–684.

Raman, S. OF Lange, P Rossi, M tyka, X Wang, J Aramini, G. Liu, TA Ramelot, A Eletsky, T Szyperski, MA Kennedy, J Prestegard, GT Montelione, D Baker. NMR structure determination for larger proteins using backbone only data. 2010. Science 327:1014–1018.

Sali, A., and Blundell, T.L. Comparative protein modelling by satisfaction of spatial restraints. 1992. J. Mol. Biol. 234:779–815.

Shortle, D, KT Simons, D. Baker. Clustering of low-energy conformations near the native structures of small proteins. Proc. Nat. Acad. Sci. USA. 1998. 95:11158-11162.

Simons, KT, C. Kooperberg, E. Huang and D. Baker. Assembly of protein teriary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. 1997. J. Mol. Biol. 268: 209-225.

Thompson, J, D Baker. Incorporation of evolutionary information into Rosetta comparative modeling. 2011. Proteins 79:2380–2388.

Thyme, SB, Y Song, TJ Brunette, MD Szeto, L Kusak, P Bradley, and D. Baker. Massively parallel determination and modeling of endonuclease substrate specificity. 2014. Nucleic Acids Research. 42:13839:13852.

Xu, D., Zhang, J., Roy, A., and Zhang, Y. Automated protein structure modeling in CASP9 by I-TASSER pipeline combined with QUARK-based ab initio folding and FG-MD-based structure refinement. 2011. Proteins 79:147–160.

Yuedong Yang, Eshel Faraggi, Huiying Zhao, Yaoqi Zhou. Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of the query and corresponding native properties of templates. 2011. Bioinformatics 27:2076-82

Zemla, A, C Venclovas, J Moult and K Fidelis. Processing and analysis of CASP3 protein structure predictions. 1999. Proteins Suppl 3:22-9

Zhang, R, GM Alushin, A Brown, E. Nogales. Mechanistic Origin of Microtubule Dynamic Instability and Its Modulation by EB Proteins. Cell 162: 849–859 (2015).