RosettaAntibody: antibody variable region homology modeling server

Aroop Sircar[1], Eric T. Kim[2] and Jeffrey J. Gray[1,3,4,]*

[1]Department of Chemical and Biomolecular Engineering, [2]Department of Biomedical Engineering and Department of Computer Science, [3]Program in Molecular Biophysics and [4]Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University, 3400 N. Charles Street, Baltimore, MD 21218, USA

ABSTRACT

The RosettaAntibody server (http://antibody.graylab.jhu.edu) predicts the structure of an antibody variable region given the amino-acid sequences of the respective light and heavy chains. In an initial stage, the server identifies and displays the most sequence homologous template structures for the light and heavy framework regions and each of the complementarity determining region (CDR) loops. Subsequently, the most homologous templates are assembled into a side-chain optimized crude model, and the server returns a picture and coordinate file. For users requesting a high-resolution model, the server executes the full RosettaAntibody protocol which additionally models the hyper-variable CDR H3 loop. The high-resolution protocol also relieves steric clashes by optimizing the CDR backbone torsion angles and by simultaneously perturbing the relative orientation of the light and heavy chains. RosettaAntibody generates 2000 independent structures, and the server returns pictures, coordinate files, and detailed scoring information for the 10 top-scoring models. The 10 models enable users to use rational judgment in choosing the best model or to use the set as an ensemble for further studies such as docking. The high-resolution models generated by RosettaAntibody have been used for the successful prediction of antibody–antigen complex structures.

INTRODUCTION

Therapeutic monoclonal antibodies are a genre of biopharmaceuticals which has benefitted healthcare in various fields from oncology to immune and inflammatory disorders. Development of successful novel therapeutic antibodies requires understanding of drug and disease mechanisms and the ability to stabilize, affinity mature, and humanize antibodies. Antibody structures can help overcome these challenges by providing atomic level insights into structure–function relationships and the antibody–antigen interaction [e.g. see refs. (1–4)]. However, experimental techniques for obtaining antibody structures, like X-ray crystallography and nuclear magnetic resonance, are laborious, time consuming and costly. Computational antibody structure prediction provides a fast and inexpensive route to obtain structures, including those which are not obtainable otherwise.

Two antibody variable region (F_V) modeling servers are available on the Internet: the Web Antibody Modeling (WAM) (5) and Prediction of Immunoglobulin Structure (PIGS) (6) servers. WAM can require several days to output one antibody model in response to a submitted query sequence. No information on templates used for modeling the antibody is provided. Furthermore, antibody structures predicted with WAM have internal clashes and their inaccuracies can confound computational docking (2,7). The PIGS server returns an antibody model in about a minute and displays the antibody crystal structures that it selects as templates. The PIGS models are generated by grafting complementarity determining region (CDR) loops onto selected framework templates, even for the hyper-variable and non-canonical CDR H3 loop. Accurate CDR H3 predictions would only be expected when a similar CDR H3 loop is present in the database, which is unlikely for novel antibody sequences. The existing servers do not provide high-resolution refinement of antibody structures and do not consider thermodynamics during modeling.

RosettaAntibody (7) is a homology modeling program within the Rosetta suite (8) for predicting high-resolution antibody F_V structures. The prediction includes modeling CDR H3 loop conformations, and it uses a simple free energy function to relieve steric clashes by simultaneously optimizing the CDR loop backbone dihedral angles,

the relative orientation of the light (V_L) and heavy (V_H) chains, and the side chain conformations. A crude model where all the CDRs are grafted from template structures can be provided in a few minutes, and high-resolution models can be generated by the full RosettaAntibody protocol running on a cluster of computers in about a day (sometimes users may need to wait for other jobs in the queue). The 10 top-scoring RosettaAntibody models can be used in docking techniques such as EnsembleDock (9) that can select binding competent conformers during docking.

A few limitations of RosettaAntibody have been that (i) execution of multiple scripts for the identification of templates can be complex, (ii) finding the template structures that have been used in modeling can be challenging given the large number of intermediate files that are generated (iii) the Rosetta command-line interface can be difficult to use and (iv) it requires significant computational time to generate all-atom models, requiring a cluster of computers. To overcome these limitations and to make the high-resolution modeling available to a broader community, we have developed the RosettaAntibody server (http://antibody.graylab.jhu.edu), where the interface is simple and modest computing resources are provided.

PROCESSING METHOD

RosettaAntibody predicts the structure of the F_V region in two stages. The first stage identifies the CDR loops and the framework regions in the input sequences, chooses the most sequence homologous templates for each respective segment, grafts the template CDRs onto template frameworks, and finally optimizes the side chains of all the residues in the assembled model. The crude model generated in the first stage is used as an input to the second stage.

The second stage of RosettaAntibody is a multi-start, multi-scale Monte Carlo-plus-minimization algorithm that generates two thousand candidate structures. The second stage of the algorithm is split into a low-resolution and a high-resolution phase. The low-resolution phase represents side chains as single pseudo-atoms (10) and generates candidate CDR H3 loop conformations via fragment assembly and cyclic coordinate descent (11) in a Monte Carlo loop. Scoring in the low-resolution phase favors nonlocal properties of native protein structures such as hydrophobic burial, compactness, pairing of β-strands and closure of the chain gap during loop building (12). The high-resolution phase iteratively performs the following: (i) optimizes side chains via rotamer packing and continuous minimization (13), (ii) perturbs CDR backbone torsion angles and the relative orientation of the light and heavy chains, and (iii) uses gradient-based minimization over the CDR torsion angles and the light chain-heavy chain displacement. The high-resolution energy function includes van der Waals energy, orientation-dependent hydrogen bonding (14), implicit Gaussian solvation (15), side-chain rotamer propensities (16), and a low-weighted distance-dependent

dielectric electrostatic energy (17). Complete methodological details are provided in ref. (7).

INPUTS AND OUTPUTS

Input

Amino-acid sequences of the light and heavy chains of the F_V region are submitted to the server by either pasting the sequences in the appropriate field or by uploading them as two separate FASTA formatted files. Since the RosettaAntibody server models only the F_V region of the antibody, F_C and leader sequences should be truncated from the input sequence prior to submission. To uniquely identify each job, the user must specify the name of the antibody and can optionally specify the user's name and an email address for notification when the modeling task has finished.

Output

Figure 1 shows a representative output page from the RosettaAntibody server. The top of the page summarizes the details of the respective job, e.g. the name of the antibody, the name of the user (if provided), and the date and time of submission, execution and completion. A chart shows the boundaries of the CDR loops and the framework regions. Next, a table displays the most sequence-homologous template [identified by BLAST (18)] for each antibody segment (V_H and V_L frameworks, CDRs L1, L2, L3, H1, H2 and H3). In a series of selectable panes, the table also displays the top seven templates for each antibody segment. A picture of the crude F_V model, formed by joining the top templates, is shown next on the output page with a link for downloading a file containing the coordinates and various energies. Residue numbering of all models generated by RosettaAntibody follows Chothia's antibody numbering scheme (19).

If requested, the next section of the page will show the 10 top-scoring high-resolution structures in rank order by energy. Each model output file includes the scoring data of individual energy terms (van der Waals, solvation, hydrogen bonding energies, etc.) for the whole F_V model as well as residue-by-residue breakdowns. Finally, the web page shows a picture of a superposition of the 10 top-scoring structures from the perspective of an antigen, demonstrating the differences in the different models (figure inset).

Any difficulties in processing the sequence would be shown in a 'warnings' section. Warnings sometimes arise to indicate poor matching of template structures or broken predicted CDR H3 loop conformations, resulting in a lower confidence model. Additional explanations of the warnings are provided in the website documentation. The documentation page also explains all output in detail, including a description of the scoring terms found in the coordinate files.

SYSTEM ARCHITECTURE

The RosettaAntibody server has a front-end web process which interfaces with the computation daemon and engine. The front-end, implemented in Python using

Figure 1. Sample results page provided by the server for monoclonal antibody 14B7 (49). Inset shows the overlay of the 10 top-scoring RosettaAntibody models which appears further down the page (antibody framework regions, gray; light chain CDRs, green; CDRs H1 and H2, blue; CDR H3, red). Web page images are generated with MolScript (50).

Django (http://www.djangoproject.com), provides results upon request for users and enters modeling tasks into a MySQL database once an input file is submitted. A back-end daemon pulls tasks from the queue in the MySQL database, translates the modeling task into Perl wrapper scripts that detect the different segments of the antibody variable region, runs BLAST to detect templates, specifies a Rosetta++ command-line, and finally submits a job to a Condor (http://www.cs.wisc.edu/condor) queue. Condor runs the job on our 188-processor Linux cluster at Johns Hopkins University, as time is available. The back-end daemon periodically detects the status of the job to report, and eventually enters the complete set of results into the MySQL database.
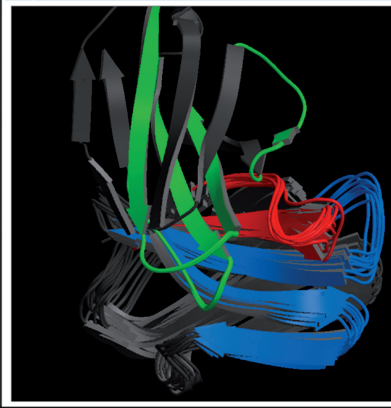
SERVER PERFORMANCE

Since the RosettaAntibody web server opened in December of 2007, over 100 individuals have used the web server for more than 400 modeling jobs. Jobs typically require about 1700 processor-hours, and results are typically complete within a day of submission, although the time will vary with the current cluster load and server queue. The website is free and open to all users with no login requirement.

Validation of the RosettaAntibody server algorithm

In a large scale test of RosettaAntibody, the program was used to recover the native crystal structures of 54 antibodies (7). To simulate blind prediction, when database information was used in modeling, only nonrelated (less than 90% sequence identity) antibody structures in the Protein Data Bank (20) were used. For the best ranked model of each target, the median root mean square deviation (rmsd) of the antigen binding pocket comprising of all the CDR residues was 1.5 Å, and 80% of the targets had an rmsd lower than 2.0 Å.

The loop modeling capabilities of RosettaAntibody were tested by *ab initio* modeling of the CDR H3 loop. The CDR H3 loop is composed of residues 95–102 of the heavy chain [Chothia numbering (19)]. The median backbone heavy atom global rmsd of the CDR H3 loop prediction for the best ranked model was 1.6, 1.9, 2.4, 3.1 and 6.0 Å, respectively, for very short (4–6 residues), short (7–9 residues), medium (10–11 residues), long (12–14 residues) and very long (17–22 residues) loops.

Finally, a practical measure of the accuracy of the antibody structures is their utility for docking to antigens. While the inclusion of the RosettaAntibody refinement steps had a small effect on homology modeling rmsds (other than CDR H3), refinement was critical for achieving docking accuracy (7). When the set of 10 top-scoring RosettaAntibody F_V homology models was used in local ensemble docking to antigen, a moderate-to-high accuracy docking prediction [rated by Critical Assessment of PRediction of Interactions criteria (21)] was achieved in 7 of 15 targets (7).

In a comparison of WAM and RosettaAntibody (7), for some antibodies, the CDR H3 predicted by WAM was closer to the native structure than that of the top-scoring model produced by RosettaAntibody. However, there was typically a more accurate structure among the 10 top-scoring RosettaAntibody models. Furthermore, antibody–antigen docking simulations starting with RosettaAntibody F_V models consistently resulted in more accurate docking predictions than those obtained by starting with WAM generated models or unrefined RosettaAntibody models (7).

Potential uses of the RosettaAntibody server

Antibody structures can be used to guide rational efforts to enhance stability (22,23) or to humanize sequences to minimize immunological response (24,25). Antibody structures can also be used for docking to their antigens, either for epitope mapping (26) or for high-resolution refinement (27). For example, we docked models of monoclonal antibody 14B7 to the anthrax toxin protective antigen (2). The models helped us form hypotheses about the mechanism of affinity maturation of several variants of 14B7. Several other instances of docking antibody homology models are present in the literature (28–30). Docking calculations can be done on several publicly available servers (31–38) including the RosettaDock Server (local docking only for high-resolution refinement, http://rosettadock.graylab.jhu.edu) (39).

Docking of homology models is necessarily less accurate than docking of crystal structures. Experimental information can be used to mitigate errors. For example, we used computational mutagenesis and hotspot analysis to evaluate models of epidermal growth factor receptor binding to mAb 806 (1). In recognition of the errors present in homology models, RosettaAntibody provides 10 alternate low-energy structures. There are several new docking methods which can use multiple input structures for one of the docking partners (40–43). Our EnsembleDock program (9) can improve low-energy docking solutions, and sometimes the low-energy docking solution is formed by the component homology structure that is closest to the crystal structure.

FUTURE DIRECTIONS

Accurate loop modeling remains one of the central challenges of antibody modeling. Thus, future improvements might be made as better and more efficient loop modeling algorithms [e.g. kinematic loop closure (44) or hierarchical local optimization (45,46)] become available. Predictions might also be improved by inclusion of NMR constraints to bias simulations [e.g. (47)]. Some researchers are pursuing therapeutics based 'heavy chain only' (VHH) antibodies discovered in the blood of camelids (48). VHHs are also easy to clone and express, and their structure might be amendable to prediction, although tests are required to assess the use of standard antibody database. Finally, we are currently developing flexible backbone antibody docking techniques which exploit the same antibody structural modeling tools as the server. These induced-fit antibody docking techniques may additionally help overcome homology modeling errors as predicted structures are used for high-resolution applications.

SUPPLEMENTARY DATA

ACKNOWLEDGEMENTS

FUNDING

REFERENCES

1. Sivasubramanian,A., Chao,G., Pressler,H.M., Wittrup,K.D. and Gray,J.J. (2006) Structural model of the mAb 806-EGFR complex using computational docking followed by computational and experimental mutagenesis. *Structure*, **14**, 401–414.
2. Sivasubramanian,A., Maynard,J.A. and Gray,J.J. (2008) Modeling the structure of mAb 14B7 bound to the anthrax protective antigen. *Proteins*, **70**, 218–230.
3. Lippow,S.M., Wittrup,K.D. and Tidor,B. (2007) Computational design of antibody-affinity improvement beyond in vivo maturation. *Nat. Biotechnol.*, **25**, 1171–1176.
4. Mariuzza,R.A., Phillips,S.E. and Poljak,R.J. (1987) The structural basis of antigen-antibody recognition. *Annu. Rev. Biophys. Biophys. Chem.*, **16**, 139–159.
5. Whitelegg,N.R. and Rees,A.R. (2000) WAM: an improved algorithm for modelling antibodies on the WEB. *Protein Eng.*, **13**, 819–824.
6. Marcatili,P., Rosi,A. and Tramontano,A. (2008) PIGS: automatic prediction of antibody structures. *Bioinformatics*, **24**, 1953–1954.
7. Sivasubramanian,A., Sircar,A., Chaudhury,S. and Gray,J.J. (2009) Toward high-resolution homology modeling of antibody Fv regions and application to antibody-antigen docking. *Proteins*, **74**, 497–514.
8. Das,R. and Baker,D. (2008) Macromolecular modeling with rosetta. *Annu. Rev. Biochem.*, **77**, 363–382.
9. Chaudhury,S. and Gray,J.J. (2008) Conformer selection and induced fit in flexible backbone protein-protein docking using computational and NMR ensembles. *J. Mol. Biol.*, **381**, 1068–1087.
10. 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.
11. Canutescu,A.A. and Dunbrack,R.L. Jr. (2003) Cyclic coordinate descent: a robotics algorithm for protein loop closure. *Protein Sci.*, **12**, 963–972.
12. Rohl,C.A., Strauss,C.E., Misura,K.M. and Baker,D. (2004) Protein structure prediction using Rosetta. *Methods Enzymol.*, **383**, 66–93.
13. Wang,C., Schueler-Furman,O. and Baker,D. (2005) Improved side-chain modeling for protein-protein docking. *Protein Sci.*, **14**, 1328–1339.
14. Kortemme,T., Morozov,A.V. and Baker,D. (2003) An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. *J. Mol. Biol.*, **326**, 1239–1259.
15. Lazaridis,T. and Karplus,M. (1999) Effective energy function for proteins in solution. *Proteins*, **35**, 133–152.
16. Dunbrack,R.L. Jr. and Cohen,F.E. (1997) Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci.*, **6**, 1661–1681.
17. Warshel,A., Russell,S.T. and Churg,A.K. (1984) Macroscopic models for studies of electrostatic interactions in proteins: limitations and applicability. *Proc. Natl Acad. Sci. USA*, **81**, 4785–4789.
18. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
19. Al-Lazikani,B., Lesk,A.M. and Chothia,C. (1997) Standard conformations for the canonical structures of immunoglobulins. *J. Mol. Biol.*, **273**, 927–948.
20. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
21. Mendez,R., Leplae,R., Lensink,M.F. and Wodak,S.J. (2005) Assessment of CAPRI predictions in rounds 3-5 shows progress in docking procedures. *Proteins*, **60**, 150–169.
22. Honegger,A., Malebranche,A.D., Rothlisberger,D. and Pluckthun,A. (2009) The influence of the framework core residues on the biophysical properties of immunoglobulin heavy chain variable domains. *Protein Eng. Des. Sel.*, **22**, 121–134.
23. Clark,L.A., Boriack-Sjodin,P.A., Eldredge,J., Fitch,C., Friedman,B., Hanf,K.J., Jarpe,M., Liparoto,S.F., Li,Y., Lugovskoy,A. *et al.* (2006) Affinity enhancement of an in vivo matured therapeutic antibody using structure-based computational design. *Protein Sci.*, **15**, 949–960.
24. Schlapschy,M., Fogarasi,M., Gruber,H., Gresch,O., Schafer,C., Aguib,Y. and Skerra,A. (2008) Functional humanization of an anti-CD16 Fab fragment: obstacles of switching from murine λ to human λ or κ light chains. *Protein Eng. Des. Sel.*, **22**, 175–188.
25. Jones,P.T., Dear,P.H., Foote,J., Neuberger,M.S. and Winter,G. (1986) Replacing the complementarity-determining regions in a human antibody with those from a mouse. *Nature*, **321**, 522–525.
26. Ponomarenko,J.V. and Bourne,P.E. (2007) Antibody-protein interactions: benchmark datasets and prediction tools evaluation. *BMC Struct. Biol.*, **7**, 64.
27. Gray,J.J. (2006) High-resolution protein-protein docking. *Curr. Opin. Struct. Biol.*, **16**, 183–193.
28. Autore,F., Melchiorre,S., Kleinjung,J., Morgan,W.D. and Fraternali,F. (2007) Interaction of malaria parasite-inhibitory antibodies with the merozoite surface protein MSP1(19) by computational docking. *Proteins*, **66**, 513–527.
29. Bracci,L., Pini,A., Bernini,A., Lelli,B., Ricci,C., Scarselli,M., Niccolai,N. and Neri,P. (2003) Biochemical filtering of a protein-protein docking simulation identifies the structure of a complex between a recombinant antibody fragment and alpha-bungarotoxin. *Biochem. J.*, **371**, 423–427.
30. McKinney,B.A., Kalleewaard,N.L., Crowe,J.E. Jr. and Meiler,J. (2007) Using the natural evolution of a rotavirus-specific human monoclonal antibody to predict the complex topography of a viral antigenic site. *Immunome Res.*, **3**, 8.
31. Comeau,S.R., Gatchell,D.W., Vajda,S. and Camacho,C.J. (2004) ClusPro: an automated docking and discrimination method for the prediction of protein complexes. *Bioinformatics*, **20**, 45–50.
32. Tovchigrechko,A. and Vakser,I.A. (2006) GRAMM-X public web server for protein-protein docking. *Nucleic Acids Res.*, **34**, W310–W314.
33. Chen,R., Li,L. and Weng,Z. (2003) ZDOCK: an initial-stage protein-docking algorithm. *Proteins*, **52**, 80–87.
34. Schneidman-Duhovny,D., Inbar,Y., Nussinov,R. and Wolfson,H.J. (2005) PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic Acids Res.*, **33**, W363–W367.
35. Dominguez,C., Boelens,R. and Bonvin,A.M. (2003) HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.*, **125**, 1731–1737.
36. Mashiach,E., Schneidman-Duhovny,D., Andrusier,N., Nussinov,R. and Wolfson,H.J. (2008) FireDock: a web server for fast interaction refinement in molecular docking. *Nucleic Acids Res.*, **36**, W229–W232.
37. Lesk,V.I. and Sternberg,M.J. (2008) 3D-Garden: a system for modelling protein-protein complexes based on conformational refinement of ensembles generated with the marching cubes algorithm. *Bioinformatics*, **24**, 1137–1144.
38. Terashi,G., Takeda-Shitaka,M., Takaya,D., Komatsu,K. and Umeyama,H. (2005) Searching for protein-protein interaction sites

and docking by the methods of molecular dynamics, grid scoring, and the pairwise interaction potential of amino acid residues. *Proteins*, **60**, 289–295.

39. Lyskov,S. and Gray,J.J. (2008) The RosettaDock server for local protein-protein docking. *Nucleic Acids Res.*, **36**, W233–W238.

40. Grunberg,R., Leckner,J. and Nilges,M. (2004) Complementarity of structure ensembles in protein-protein binding. *Structure*, **12**, 2125–2136.

41. Smith,G.R., Sternberg,M.J. and Bates,P.A. (2005) The relationship between the flexibility of proteins and their conformational states on forming protein-protein complexes with an application to protein-protein docking. *J. Mol. Biol.*, **347**, 1077–1101.

42. Bastard,K., Prevost,C. and Zacharias,M. (2006) Accounting for loop flexibility during protein-protein docking. *Proteins*, **62**, 956–969.

43. Krol,M., Chaleil,R.A., Tournier,A.L. and Bates,P.A. (2007) Implicit flexibility in protein docking: cross-docking and local refinement. *Proteins*, **69**, 750–757.

44. Coutsias,E.A., Seok,C., Jacobson,M.P. and Dill,K.A. (2004) A kinematic view of loop closure. *J. Comput. Chem.*, **25**, 510–528.

45. Jacobson,M.P., Pincus,D.L., Rapp,C.S., Day,T.J., Honig,B., Shaw,D.E. and Friesner,R.A. (2004) A hierarchical approach to all-atom protein loop prediction. *Proteins*, **55**, 351–367.

46. Zhu,K., Pincus,D.L., Zhao,S. and Friesner,R.A. (2006) Long loop prediction using the protein local optimization program. *Proteins*, **65**, 438–452.

47. Meiler,J. and Baker,D. (2003) Rapid protein fold determination using unassigned NMR data. *Proc. Natl Acad. Sci. USA*, **100**, 15404–15409.

48. Roovers,R.C., van Dongen,G.A. and en Henegouwen,P.M. (2007) Nanobodies in therapeutic applications. *Curr. Opin. Mol. Ther.*, **9**, 327–335.

49. Maynard,J.A., Maassen,C.B., Leppla,S.H., Brasky,K., Patterson,J.L., Iverson,B.L. and Georgiou,G. (2002) Protection against anthrax toxin by recombinant antibody fragments correlates with antigen affinity. *Nat. Biotechnol.*, **20**, 597–601.

50. Kraulis,P.J. (1991) MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Cryst.*, **24**, 946–950.