Nucleic Acids Research, 2006, Vol. 34, Web Server issue ***W235–W238***
doi:10.1093/nar/gkl163

RosettaDesign server for protein design

Yi Liu and Brian Kuhlman*

Department of Biochemistry and Biophysics, University of North Carolina, Chapel Hill, NC 27599, USA

Received January 30, 2006; Revised February 22, 2006; Accepted March 20, 2006

ABSTRACT

The RosettaDesign server identifies low energy amino acid sequences for target protein structures (http://rosettadesign.med.unc.edu). The client provides the backbone coordinates of the target structure and specifies which residues to design. The server returns to the client the sequences, coordinates and energies of the designed proteins. The simulations are performed using the design module of the Rosetta program (RosettaDesign). RosettaDesign uses Monte Carlo optimization with simulated annealing to search for amino acids that pack well on the target structure and satisfy hydrogen bonding potential. RosettaDesign has been experimentally validated and has been used previously to stabilize naturally occurring proteins and design a novel protein structure.

INTRODUCTION

Recently, there have been many successes in the area of computational protein design. Protein design software has been used to stabilize naturally occurring proteins, perturb protein binding specificity, design novel biosensors and enzymes and create novel protein structures [for a review see (1–3)]. In most cases, these studies have been performed by laboratories that specialize in computational design and have direct access to the software and the source code (4–6). To make this technology more accessible to the large number of molecular biology laboratories that regularly use amino acid mutagenesis to probe protein structure and function, we have established a web server for protein design that uses the design module of the Rosetta program (Rosetta Design) (7,8).

Given a target protein structure or complex, RosettaDesign searches for amino acid sequences that pack well, bury their hydrophobic atoms and satisfy the hydrogen bonding potential of polar atoms. RosettaDesign has been parameterized to return sequences with amino acid frequencies comparable to those found in naturally occurring proteins, and to partition the hydrophobic and polar residues between the surface and the core at naturally occurring frequencies. In general, when redesigning a naturally occurring protein ~65% of the residues will mutate. As expected, more sequence variability is seen on the surface of the protein where there are fewer packing constraints. In the core of the protein 45% of the residues mutate on average. RosettaDesign has been experimentally validated. It has been used to stabilize naturally occurring proteins (9), enhance protein binding affinities (10), design a protein that can switch between 2-folds (11) and create a protein with a novel structure (12).

METHOD USED

The RosettaDesign server uses the design module of the Rosetta program to perform fixed backbone protein design simulations. The algorithm has been described previously (7,8). Like other protein design programs, RosettaDesign has two primary components: an energy function for evaluating the relative favorability of a sequence and an optimization procedure for searching through sequence space. All atoms in the protein, including hydrogen, are explicitly modeled. The energy function consists of (i) a Lennard–Jones potential that favors close packed residues, (ii) the Lazaridis–Karplus implicit solvation model which favors hydrophobic amino acids in the interior of proteins and polar amino acids on the surface (13), (iii) an explicit orientation dependent hydrogen bonding term (14), (iv) torsion potentials derived from the PDB (15), (v) a unique reference value for each amino acid type and (vi) electrostatic interactions between charged residues are modeled by an additional term that is based on the probability of seeing two amino acid types near each other in the PDB (16). This is a relatively weak term in the energy function.

To simplify the optimization procedure and favor low energy designs, amino acid side chains are only allowed to adopt a discrete set of favorable conformations, typically referred to as rotamers. RosettaDesign uses Dunbrack's backbone dependent rotamer library (15). To allow for relaxation away from the most preferred side chain conformations, additional rotamers are created for buried residues by varying chi1 and chi2 one standard deviation (~10°) away from the most preferred values. Rotamers are also created for the alternate positions hydrogen can adopt on serine, threonine and tyrosine. To find low energy sequences, RosettaDesign

*To whom correspondence should be addressed. Tel: +1 919 843 0188; Fax: +1 919 966 2852; Email: bkuhlman@email.unc.edu

uses Monte Carlo optimization with simulated annealing. Starting from a random sequence, single amino acid substitutions or rotamer switches are accepted based on the Metropolis criterion. The simulation starts at a very high temperature where almost all substitutions are accepted and finishes at 0°. Approximately 1 million rotamer substitutions are attempted per 100 residues being varied. Independent simulations in which every residue in the protein is allowed to vary generally converge to sequences that are 70–80% identical to each other.

SERVICES

Protein design

The RosettaDesign server returns low energy sequences for target protein structures. The protein backbone remains fixed during the simulation.

Side chain conformation prediction

Given a protein structure and sequence, the RosettaDesign server can be used to predict the lowest energy conformations of the side chains.

INPUTS, OUTPUTS AND JOB OPTIONS

Registration

To receive results via email users must register. Alternatively, users can access the web server as a 'guest'. In this case they must return to the web site to retrieve results.

Input files

PDB file: users must submit a file with the atomic coordinates of the protein that will be the template for design. The coordinates must be in PDB format. There can be gaps in the structure, but each residue must have a complete set of backbone heavy atoms—*N'*, *C'*, *O* and C_α. The residues can be missing side chain atoms.

Resfile: the resfile specifies which sequence positions will be varied, and which amino acids will be considered at each position. Users can also request that the native amino acid be kept at a particular sequence position, but allow the side chain to adopt a new conformation. The resfile can be created on the web site using point-and-click operations (Figure 1) or a user can upload his or her own resfile. The server will check the integrity of the uploaded resfile to ensure the correct format. The resfile created on the web site with point-and-click operations can also be saved for future use. A full description of the format for a resfile is provided in the documentation section of the web site.

Please choose which residues you would like to redesign. At this time, a maximum of 200 residues can be varied in a single simulation.
When you are finished selecting your residues, hit the submit job button at the bottom of the page. Back to resubmit job

Main page
Documentation
Register
Log in
Queue
Submit job
Log out

Chain:	Start:	End:	POLAR
a	6	10	select help

Chain ID	Residue	FIXED	REPACK	POLAR	APOLAR	ALLDESIGN	Choose Amino Acid
A	1	●	○	○	○	○	choose
A	2	○	○	○	●	○	choose
A	3	○	○	○	○	○	choose LPM
A	4	●	○	○	○	○	choose
A	5	●	○	○	○	○	choose
A	6	○	○	●	○	○	choose
A	7	○	○	●	○	○	choose
A	8	○	○	●	○	○	choose
A	9	○	○	●	○	○	choose
A	10	○	○	●	○	○	choose
A	11	●	○	○	○	○	choose
A	12	●	○	○	○	○	choose
A	13	●	○	○	○	○	choose

Figure 1. Interface for choosing which sequence positions to vary.

Job options

Users can choose either to redesign the whole protein with all 20 amino acids considered at each sequence position, or to redesign part of the target protein as specified in a resfile. Because RosettaDesign uses a stochastic sampling algorithm to identify low energy sequences, different simulations will not necessarily give identical results. Users can choose to repeat the same simulation up to 10 times with a single job submission.

Output files

The simulation results are compressed as a zip file that unzips into three files: a log file indicating what commands were used for the simulation, a text file with a list of the mutations that were made, and a third file that provides the coordinates in PDB format along with the energies of the redesigned protein. If a run does not finish, the server will email the user the suspected reason for failure.

There are three sections of the PDB file pertaining to the energy of the redesigned protein:

The first part is a list of scores. Except for the reference energies, a lower score is better. The second section of energies is a table with the energy of each residue in the protein (Table 1). In the cases in which an energy depends on two atoms in separate residues (for instance the Lennard–Jones energy), half of the energy is assigned to each residue. The third section is a table of measured energies—expected energies. Expected energies are derived by calculating the average energies of the different amino acids as a function of buriedness in a large set of proteins from the PDB. For instance, in the PDB leucines with 20 neighbors (residues within 10 Å) have an average Lennard–Jones score of −3.79 kcal/mol. If a leucine in the redesigned protein has 20 neighbors and has a Lennard–Jones energy of −4.2 kcal/mol, then it indicates that leucine is more tightly packed than the average leucine in the PDB. In general, we have found this table especially useful in the design of new protein structures, as it allows one to estimate how much the designed protein resembles proteins found in nature.

Table 1. The scores relevant to protein design

Energy type	Description
Total	The total score for the designed protein (lower is better)
LJatr	Attractive portion of the Lennard–Jones potential (rewards close contacts)
LJrep	Repulsive portion of the Lennard–Jones potential (penalizes overlaps)
LKsol	Lazaridis–Karplus solvation model (penalizes buried polars) (13)
Erot	−lnP(rotlaa,phi,psi), internal energy of side chain rotamers as derived from Dunbrack's statistics (15)
Eintra	Intra-residue steric clashes
Ehbnd	Kortemme hydrogen bonding potential (14)
Epair	Pair score based on the probability of seeing two amino acids near each other in the PDB (favors salt bridges) (16)
Eaa_phipsi	−lnP(aalphi,psi), amino acid phi,psi preferences
Hb_sc	Sidechain-sidechain and sidechain-backbone hydrogen bond energy
Hb_srbb	Backbone-backbone hbonds close in primary sequence
Hb_lrbb	Backbone-backbone hbonds distant in primary sequence
Eref	Reference energy derived from amino acid composition
Egb	Generalized born solvation energy (this is not used by the server)
Eh2o, Eh2o_hb	Energies from explicit waters (this is not used by the server)
Ecst	Constraint energies (this is not used by the server)
Eres	Total energy for the residue (lower is better)
SASApack	SASApack is related to the void volume in a protein. Surface areas are computed with a 1.4 Å probe and 0.5 Å probe and the difference (ASA_0.5 - ASA_1.4) is compared to the expected difference for a particular residue type in a particular environment. A negative value is favorable and indicates that the residue is more tightly packed than is seen in average pdb files.

SERVER PERFORMANCE

There have been 3000 jobs submitted by more than 320 clients to the RosettaDesign web server since March 2005. The server can accept proteins as large as 1000 residues and can redesign up to 200 residues in one simulation. The web site is set up as an apache server with a daemon that automatically invokes the Rosetta++ executable with the users input file and options obtained from the web interface. The user's input files, job options, and the results are recorded in a mySQL database via a php-http module. A maximum of two jobs can be run at the same time. The daemon checks the mySQL database for pending jobs every minute. For proteins between 100 and 200 residues, the simulation typically finishes in 5–30 min.

Accuracy of the RosettaDesign server

In a large scale test of RosettaDesign, the program was used to completely redesign nine naturally occurring proteins (9). The redesigned sequences were on average 35% identical to the wild-type sequence. Five out of the nine proteins were well-folded as evidenced by NMR and thermal and chemical denaturation experiments. All five of the well-folded proteins had higher thermal unfolding midpoints than the wild-type sequence. RosettaDesign has also been used to redesign small regions of a protein to increase protein stability or binding affinities (10,17,18). In many of these cases, lower free energies were obtained by building additional hydrophobic interactions. RosettaDesign has had less success with creating buried hydrogen bond networks. This is presumably because hydrogen bonds are very sensitive to small changes in distance and orientation, and desolvation penalties are difficult to calculate accurately.

Because the RosettaDesign energy function favors like amino acids being near other (polars with polars, hydrophobics with hydrophobics) it will in some cases design large patches of hydrophobic amino acids on the surface of a protein. Although this may be favorable for protein stability, it can lead to aggregation of the protein. In this event, the user can force a small set of residues in the center of the patch to be polar, and this in general will encourage RosettaDesign to put polar residues at the neighboring positions as well.

Possible uses for the RosettaDesign server

Over the last 10 years protein design software has been applied to a large number of interesting problems. Several laboratories have used sequence optimization algorithms to explore the

size and characteristics of sequence space compatible with a particular fold. In a few cases, this information has been used to help detect remote homologs (19,20). In general, protein structures and complexes can be stabilized by identifying mutations that increase buried hydrophobic surface area. Towards this end, the RosettaDesign server can be used to search for holes in proteins that can be filled with larger hydrophobic residues, or partially buried polar residues that can be replaced by hydrophobic residues.

RosettaDesign can be used to search for second-site suppressor mutations. In this scenario, the user has a priori knowledge of a mutation that destabilizes a protein or protein–protein complex. Using a resfile, the user can force the destabilizing mutation and use RosettaDesign to search for mutations that will compensate for the first mutation. A similar approach was recently used to redesign a protein–protein interface so that the redesigned proteins still bind each other, but no longer bind their other naturally occurring binding partners (21). These types of redesigns are useful for probing signal transduction pathways.

In cases where a protein can adopt multiple conformations, RosettaDesign can be used to identify sequences that are specifically optimized for one of the conformations. Mayo and colleagues used this approach to increase the affinity between a receptor protein and its ligand (22). More ambitiously RosettaDesign can be used to help design new protein structures or portions of proteins. In this case, the user must supply the backbone coordinates of the target structure. The challenge is that many arbitrarily chosen protein backbones will not be designable. This is generally reflected in poor LJatr and SASApack (see Table 1) values for the redesigned protein. In the future, as our computational resources grow, we plan to modify the RosettaDesign server so that the backbone coordinates and the sequence can be optimized simultaneously to allow for tight packing between side chains.

ACKNOWLEDGEMENTS

This work was supported by a grant from the NIH (1RO1 GM073151-01). Funding to pay the Open Access publication charges for this article was provided by the NIH.

Conflict of interest statement. None declared.

REFERENCES

1. Pokala,N. and Handel,T.M. (2001) Review: protein design–where we were, where we are, where we're going. *J. Struct. Biol*., **134**, 269–281.
2. Park,S., Yang,X. and Saven,J.G. (2004) Advances in computational protein design. *Curr. Opin. Struct. Biol*., **14**, 487–494.
3. Butterfoss,G.L. and Kuhlman,B. (2005) Computer-based design of novel protein structures. *Annu. Rev. Biophys. Biomol. Struct*., **35**, 49–65.
4. Dahiyat,B.I. and Mayo,S.L. (1997) De novo protein design: fully automated sequence selection. *Science*, **278**, 82–87.
5. Dwyer,M.A. and Hellinga,H.W. (2004) Periplasmic binding proteins: a versatile superfamily for protein engineering. *Curr. Opin. Struct. Biol*., **14**, 495–504.
6. Harbury,P.B., Plecs,J.J., Tidor,B., Alber,T. and Kim,P.S. (1998) High-resolution protein design with backbone freedom. *Science*, **282**, 1462–1467.
7. Rohl,C.A., Strauss,C.E., Misura,K.M. and Baker,D. (2004) Protein structure prediction using Rosetta. *Meth. Enzymol*., **383**, 66–93.
8. Kuhlman,B. and Baker,D. (2000) Native protein sequences are close to optimal for their structures. *Proc. Natl Acad. Sci. USA*, **97**, 10383–10388.
9. Dantas,G., Kuhlman,B., Callender,D., Wong,M. and Baker,D. (2003) A large scale test of computational protein design: folding and stability of nine completely redesigned globular proteins. *J. Mol. Biol*., **332**, 449–460.
10. Eletr,Z.M., Huang,D.T., Duda,D.M., Schulman,B.A. and Kuhlman,B. (2005) E2 conjugating enzymes must disengage from their E1 enzymes before E3-dependent ubiquitin and ubiquitin-like transfer. *Nature Struct. Mol. Biol*., **12**, 933–934.
11. Ambroggio,X.I. and Kuhlman,B. (2006) Computational design of a single amino acid sequence that can switch between two distinct protein folds. *J. Am. Chem. Soc*., **128**, 1154–1161.
12. 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.
13. Lazaridis,T. and Karplus,M. (1999) Effective energy function for proteins in solution. *Proteins*, **35**, 133–152.
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. Dunbrack,R.L.Jr and , Cohen,F.E. (1997) Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci*., **6**, 1661–1681.
16. Simons,K.T., Ruczinski,I., Kooperberg,C., Fox,B.A., Bystroff,C. and Baker,D. (1999) Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins*, **34**, 82–95.
17. Korkegian,A., Black,M.E., Baker,D. and Stoddard,B.L. (2005) Computational thermostabilization of an enzyme. *Science*, **308**, 857–860.
18. Nauli,S., Kuhlman,B. and Baker,D. (2001) Computer-based redesign of a protein folding pathway. *Nature Struct. Biol*., **8**, 602–605.
19. Pei,J., Dokholyan,N.V., Shakhnovich,E.I. and Grishin,N.V. (2003) Using protein design for homology detection and active site searches. *Proc. Natl Acad. Sci. USA*, **100**, 11361–11366.
20. Saunders,C.T. and Baker,D. (2005) Recapitulation of protein family divergence using flexible backbone protein design. *J. Mol. Biol*., **346**, 631–644.
21. Kortemme,T., Joachimiak,L.A., Bullock,A.N., Schuler,A.D., Stoddard,B.L. and Baker,D. (2004) Computational redesign of protein–protein interaction specificity. *Nature Struct. Mol. Biol*., **11**, 371–379.
22. Shimaoka,M., Shifman,J.M., Jing,H., Takagi,J., Mayo,S.L. and Springer,T.A. (2000) Computational design of an integrin I domain stabilized in the open high affinity conformation. *Nature Struct. Biol*., **7**, 674–678.