ELSEVIER

Potential energy functions for protein design

F Edward Boas and Pehr B Harbury

Different potential energy functions have predominated in protein dynamics simulations, protein design calculations, and protein structure prediction. Clearly, the same physics applies in all three cases. The differences in potential energy functions reflect differences in how the calculations are performed. With improvements in computer power and algorithms, the same potential energy function should be applicable to all three problems. In this review, we examine energy functions currently used for protein design, and look to the molecular mechanics field for advances that could be used in the next generation of design algorithms. In particular, we focus on improved models of the hydrophobic effect, polarization and hydrogen bonding.

Addresses
Department of Biochemistry, Stanford University School of Medicine, Beckman B437, 279 Campus Drive West, Stanford, CA 94305-5307, USA

Corresponding author: Harbury, Pehr B (harbury@cmgm.stanford.edu)

Current Opinion in Structural Biology 2007, **17**:199–204

This review comes from a themed issue on
Theory and simulation
Edited by Richard Lavery and Kim A Sharp

Available online 26th March 2007

0959-440X/$ – see front matter

DOI 10.1016/j.sbi.2007.03.006

Introduction

Computational protein design algorithms use models of protein energetics to engineer protein sequences with new functions. This is similar to more established branches of engineering, such as circuit simulation or stability analysis of buildings, whereby accurate computer models are used to evaluate designs before they are built. Protein design provides a rigorous test of the energetic model that is used, because the design algorithm must pick functional sequences out of an astronomically large space of non-functional sequences.

As with any calculation, there is a trade-off between accuracy and speed when modeling or designing proteins. For example, simulation of a one-second dissociation event using a molecular dynamics calculation with explicit water would take ten million years on a typical desktop computer. Protein design algorithms use several strategies to speed up the modeling process. First, protein design algorithms do not simulate kinetics; instead they calculate the energies of a small number of target states (these energies are used as a surrogate for the free energies of conformational neighborhoods). Many fast algorithms exist for optimizing the structure of each target state. Second, protein design calculations do not explicitly model water; instead they use a continuum representation of water. Finally, protein design algorithms generally use less computationally intensive energy functions than molecular mechanics calculations do.

Previous reviews have described potential energy functions (PEFs) used for molecular mechanics simulations [1,2], protein design [3,4] and protein structure prediction [5]. In this review, we compare these energy functions (Figure 1).

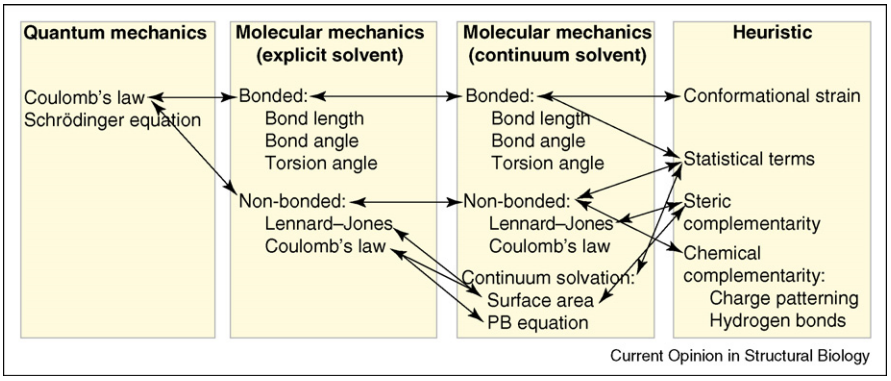
Potential energy functions

Overview

Molecular mechanics potential energy functions (MM-PEFs) incorporate both 'bonded' and 'non-bonded' terms (Figure 2). The bonded terms apply to sets of two to four atoms that are covalently linked, and they serve to constrain bond lengths and angles near their equilibrium values. The bonded terms also include a torsional potential that models the periodic energy barriers encountered during bond rotation. The non-bonded terms consist of the Lennard–Jones (LJ) function (which includes van der Waals attraction and repulsion owing to orbital overlap) and Coulomb's law. The parameters for the bonded and non-bonded terms of an MM-PEF are derived from quantum calculations and from thermodynamic, crystallographic and spectroscopic data on a wide range of systems [1,2]. MM-PEFs have been used predominately to simulate protein folding and dynamics, but are also used to refine X-ray crystal structures.

An alternative type of PEF is the knowledge-based, or statistical, energy function [5,6] (Figure 3). This type of energy function derives from the database of known protein structures. The probabilities that residues appear in specific configurations (such as rotamer conformations or buried versus surface environments) or the probabilities that pairs of residues appear together in a defined relative geometry are calculated. These probabilities are converted into an effective potential energy using the Boltzmann equation: $\Delta G = -RT\ln(p_{obs}/p_{exp})$, where p_{obs} is the probability of seeing a particular structural element and p_{exp} is the expected probability of seeing that structural element by chance [7–9]. The advantage of a knowledge-based energy function is that it can model any behavior seen in known protein crystal structures, even

Figure 1

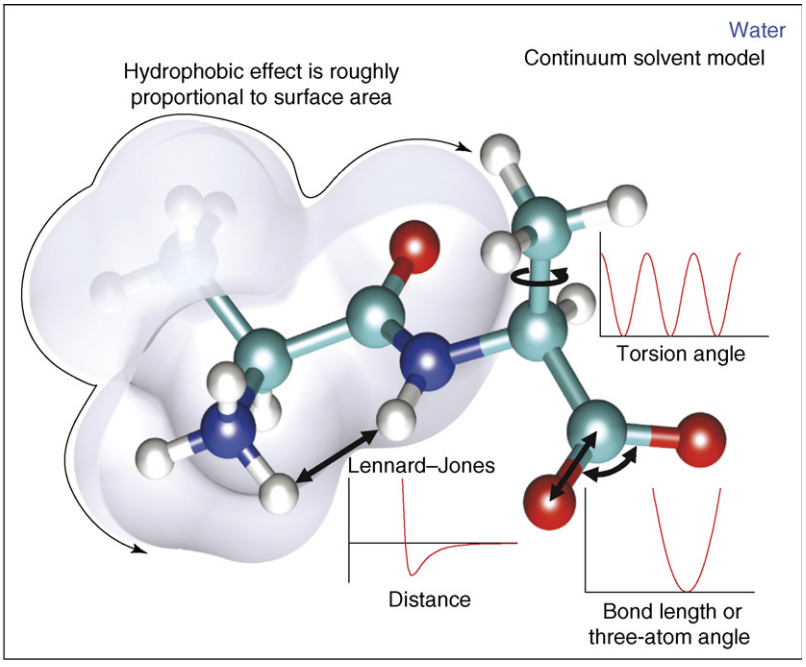


PEFs for evaluating protein conformations range from quantum mechanics, which is accurate but very slow, to more heuristic energy functions that include statistical terms. In between are MM-PEFs, which are the most thoroughly tested model of molecular energetics. Currently, the protein design field uses heuristic energy functions, but the trend is towards using more physically based PEFs.

if a good physical understanding of the behavior does not exist. The disadvantage is that these energy functions are phenomenological and cannot predict new behaviors absent from the training set.

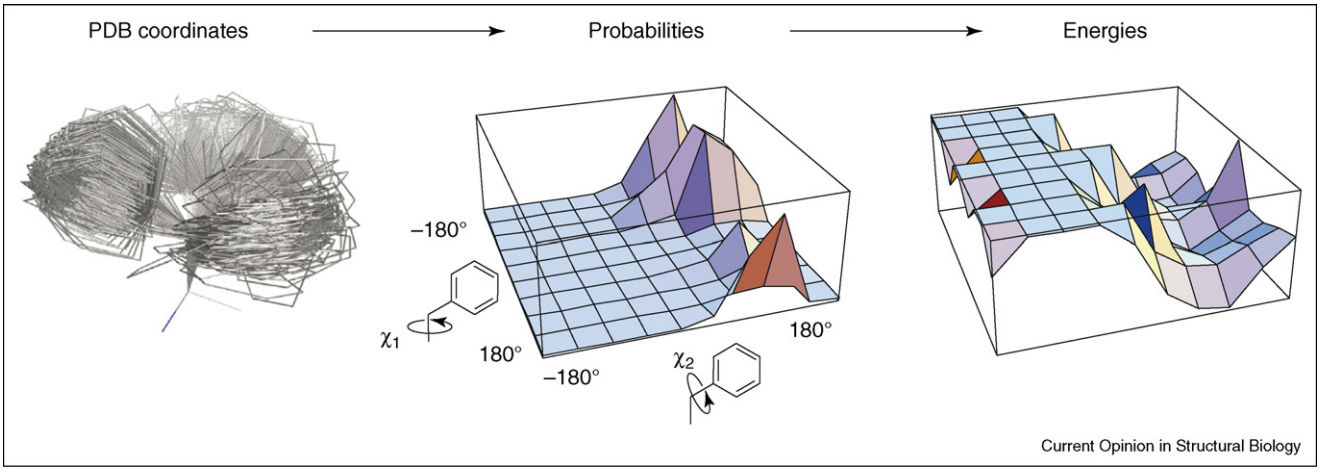
Design energy functions include a combination of MM-PEF, knowledge-based and other terms. In contrast to MM-PEFs, which have become fairly standardized, design potentials vary enormously between laboratories. The various terms are typically calibrated and weighted to optimize performance for one type of prediction, such as experimental binding energy [10,11], or to produce native-like sequences when redesigning natural proteins [12]. By way of illustration, we describe the PEFs used in two recent landmark protein design papers. In the first example, Looger *et al.* [13] redesigned various bacterial periplasmic binding proteins to bind trinitrotoluene, lactate and serotonin. Their energy function included an LJ term (using CHARMM22 parameters [14]) with the repulsive component scaled down to 35%, a Coulombic term with a distance-dependent dielectric constant of 8.0*r* and partial charges from CHARMM22, an explicit hydrogen-

Figure 2



MM-PEF with continuum solvent.

Figure 3



Knowledge-based PEF.

bonding term derived from the DREIDING MM-PEF [15], a surface-area-based solvation term, a knowledge-based rotamer term [16], and a term requiring all hydrogen-bond donors and acceptors to be satisfied. In a subsequent paper, Dwyer *et al.* designed *de novo* triosephosphate isomerase activity into ribose-binding protein [17] using a more accurate electrostatics model that included multiple geometry-dependent dielectric constants [18]. A second example is the 93-residue protein with a new α/β fold designed by Kuhlman *et al.* [12]. Their energy function included an LJ term (with well depths from CHARMM19 and radii fit to match the distribution of distances seen in the PDB), a Lazaridis–Karplus empirical solvation term [19], a knowledge-based hydrogen-bonding term [20], a knowledge-based rotamer term and a knowledge-based pairwise residue interaction term. The scaling factors for each term were adjusted to optimize recovery of native sequences when redesigning a training set of 30 proteins.

Why are MM-PEFs and design PEFs so different, and why do the latter include so many *ad hoc* terms? The basic answer is that design PEFs must compensate for an incomplete simulation of protein behavior: many degrees of freedom are ignored, modeled implicitly or sampled at low resolution. We examine this question term-by-term in the following sections.

Bonded terms

Although it is straightforward to directly use the bonded portion of MM-PEFs to determine the relative energies of different rotamer geometries, design potentials have tended to use fixed rotamer coordinates and knowledge-based rotamer potentials. MM-PEF bonded energies vary greatly with small changes in bond lengths and angles. Thus, these energies are not meaningful unless the structures have first been locally energy minimized (perhaps with dihedral angle restraints).

Lennard–Jones

The LJ function includes a weakly attractive component at long distances (the van der Waals energy) and a strongly repulsive component at short distances. The repulsive component is sensitive to small atomic displacements: the LJ energy of a protein crystal structure can decrease by hundreds of kilocalories per mole upon local energy minimization, despite imperceptible changes in the atomic coordinates.

The discrete rotamer sampling used for protein design calculations inevitably leads to small atomic overlaps, producing large unfavorable LJ energies. In many cases, the overlaps could be eliminated by local minimization, but such minimization cannot be readily incorporated into combinatorial sequence design algorithms. Instead, the functional form of the LJ interaction is almost always softened so that overlaps are less energetically unfavorable. For example, the LJ radii can be scaled down [21], the repulsive component of the LJ energy can be scaled down [13] or the LJ function can be linearly extrapolated below a cutoff distance [12].

Softening the LJ function is based on a presumption that protein cores are reasonably fluid and thus can always rearrange to accommodate small overlaps. This modification, however, always leads to qualitative and quantitative errors in interaction energies. For example, modern MM-PEFs model hydrogen bonds as a combination of an electrostatic interaction and an LJ interaction. When overlaps are allowed, atoms can approach more closely, producing artificially favorable hydrogen-bond energies. In general, changing the LJ parameters in any way will destroy the delicate balance engineered into an MM-PEF. Use of unmodified LJ functions for protein design will require either very high resolution discrete sampling or some form of continuous optimization.

Solvation

Computing the energy of a protein embedded in explicit solvent molecules is time consuming, because the energy must be averaged over many solvent configurations. To speed up calculations, solvent can instead be modeled as a smooth continuous material with a characteristic dielectric constant and surface tension. The solvation energy of such protein continuum-solvent systems is generally separated into two components. The first component is the hydrophobic effect, which accounts for the interfacial free energy of the uncharged protein and the continuum solvent. The second component is the solvation polarization energy, which accounts for the interaction of partial charges in the protein with dipoles and ion clouds induced in the solvent. Charged atoms closer to the protein surface have more favorable solvation energies and smaller apparent charge–charge interactions.

Both the LJ function and Coulomb's law are pairwise factorable, meaning that the total energy can be expressed as a sum of interactions between pairs of atoms without regard to the position of any other atom in the system. This is important because the total energy can then be determined by summing precalculated pairwise interaction energies (required for most rapid structural optimization procedures). Solvation energies, on the other hand, are not inherently pairwise factorable. The interaction between two charges depends on the positions of other atoms, because the other atoms displace solvent and salt.

Hydrophobic effect

The continuum hydrophobic effect has traditionally been modeled as being proportional to the solvent-accessible surface area of a solute [22]. Pairwise factorable approximations of surface area have been developed for use in design calculations [23]. Although widely applied, the surface-area-based model has clear limitations. For example, hydrophobic solutes in water can interact favorably when they are separated by a single layer of water molecules [24]. This type of interaction is completely absent from a surface-area-based energy. Wagoner and Baker [25•] have developed a model of the hydrophobic effect that captures such complex wetting phenomena. It produces energies that are closer to explicit solvent simulations than the surface-area-based energies are. Their energy function includes a term proportional to surface area, a term proportional to volume and a solute–solvent van der Waals term. Adapting this improved model for protein design work will require either the development of a pairwise factorable approximation or the use of a design algorithm that does not require precalculated energies.

Solvent polarization

Solvent polarization is very difficult to simulate quickly and accurately. Consequently, many different empirical models that subsume polarization energies have been used in protein design efforts [16,18,19,26]. These models commonly include a solvation energy for charged atoms based on accessible surface area and a Coulomb's law term with a distance-dependent dielectric constant. The surface-area models disregard the non-zero contributions of fully buried charges to the polarization energy. The distance-dependent dielectric constant scales down Coulomb's law to account for the screening of charge–charge interactions by water. However, it ignores the fact that screening depends on the local environment of each charge.

A more physical approach is to solve the Poisson–Boltzmann (PB) differential equation [27] that describes the relationship between fixed charge and the electric potential in a continuum dielectric environment. Water is assigned a dielectric constant of 80, the protein interior is typically assigned a dielectric constant between 1 and 20, and the molecular surface defines the boundary between protein and solvent. Values of the electric potential on a spatial grid can be obtained using a finite-difference algorithm. Marshall *et al.* [28] describe a pairwise factorable approximation to the PB equation based on summing precalculated energies for single residues and for pairs of residues. This treatment does not take into account rotamer-conformation-dependent changes in the protein–solvent boundary or that solutions to the PB equation are not truly superimposable.

Alternatively, the generalized Born equation [29] provides a fast approximate solution to the PB equation, and it has been used for protein design [30]. Recent improvements to the generalized Born functional form [31,32] yield solvation energies that are comparable to those derived from finite-difference calculations [33].

Explicit water

Continuum solvent models break down when water molecules are tightly bound to proteins. It may be possible, however, to incorporate a handful of explicit water molecules in a continuum solvent calculation. Schymkowitz *et al.* [34] developed a method for predicting the positions of tightly bound water molecules in proteins. Jiang *et al.* [35•] showed how to incorporate water molecules into amino acid rotamers.

Hydrogen bonds

In an MM-PEF, hydrogen bonds are typically modeled as dipole–dipole interactions. The optimal geometry for a dipole–dipole interaction, for example, between the C=O and N–H dipoles in the protein backbone, places all four atoms in a straight line. However, the charge distribution around the carbonyl oxygen adopts a trigonal sp^2 arrangement, which is not spherically symmetrical. The sp^2 lone pair geometry should favor a bent hydrogen bond. Morozov *et al.* [36] showed that the bent geometry is indeed

preferred, according to quantum calculations and crystal structures in the PDB. Using the PDB statistics, they developed a knowledge-based hydrogen-bonding energy function [20,37] and used it to design a new protein [12].

Solute polarization and quantum effects

A widely recognized limitation of MM-PEFs is that they assume fixed atomic charges and do not model environment-dependent rearrangement of charge on a solute. Recently developed polarizable force fields address this limitation by allowing the electric field to induce dipoles at each atom [38,39]. Importantly, solute polarization breaks down the pairwise factorability property of traditional MM-PEFs. MM-PEFs also do not model chemical realities such as lone pairs, bond formation and the partial covalent character of hydrogen bonds. One possible compromise is to model key parts of the protein using quantum mechanics and the rest of the protein using molecular mechanics [40,41].

Conclusions and future directions

The techniques described above have been used to design proteins with a wide variety of new functions. Clark *et al.* [42•] optimized the recombining site of an antibody to increase the ligand affinity and Lazar *et al.* [43] optimized the Fc region of an antibody to bind more tightly to the Fc receptor. Ashworth *et al.* [44••] redesigned an endonuclease to recognize and cut a heterologous DNA sequence. Ambroggio and Kuhlman [45•] designed a protein that reversibly switches between two distinct protein folds with a change in pH or cobalt concentration.

These examples illustrate the diverse range of useful functions already achievable by protein design. As PEFs, search algorithms and computational power continue to improve, protein design should become a standard and general research tool.

Acknowledgements

This work was supported by the National Institutes of Health (GM068126-01 to PBH). FEB was partially supported by a training grant from the National Institutes of General Medical Sciences (5T32 GM07365-28).

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Mackerell AD Jr: **Empirical force fields for biological macromolecules: overview and issues**. *J Comput Chem* 2004, **25**:1584-1604.
2. Jorgensen WL, Tirado-Rives J: **Potential energy functions for atomic-level simulations of water and organic and biomolecular systems**. *Proc Natl Acad Sci USA* 2005, **102**:6665-6670.
3. Gordon DB, Marshall SA, Mayo SL: **Energy functions for protein design**. *Curr Opin Struct Biol* 1999, **9**:509-513.
4. Pokala N, Handel TM: **Review: protein design — where we were, where we are, where we're going**. *J Struct Biol* 2001, **134**:269-281.
5. Lazaridis T, Karplus M: **Effective energy functions for protein structure prediction**. *Curr Opin Struct Biol* 2000, **10**:139-145.
6. Mohanty D, Dominy BN, Kolinski A, Brooks CL III, Skolnick J: **Correlation between knowledge-based and detailed atomic potentials: application to the unfolding of the GCN4 leucine zipper**. *Proteins* 1999, **35**:447-452.
7. Ben-Naim A: **Statistical potentials extracted from protein structures: Are these meaningful potentials?** *J Chem Phys* 1997, **107**:3698-3706.
8. Simons KT, Kooperberg C, Huang E, Baker D: **Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions**. *J Mol Biol* 1997, **268**:209-225.
9. Dehouck Y, Gilis D, Rooman M: **A new generation of statistical potentials for proteins**. *Biophys J* 2006, **90**:4010-4017.
10. Kortemme T, Baker D: **A simple physical model for binding energy hot spots in protein-protein complexes**. *Proc Natl Acad Sci USA* 2002, **99**:14116-14121.
11. Kortemme T, Joachimiak LA, Bullock AN, Schuler AD, Stoddard BL, Baker D: **Computational redesign of protein-protein interaction specificity**. *Nat Struct Mol Biol* 2004, **11**:371-379.
12. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D: **Design of a novel globular protein fold with atomic-level accuracy**. *Science* 2003, **302**:1364-1368.
13. Looger LL, Dwyer MA, Smith JJ, Hellinga HW: **Computational design of receptor and sensor proteins with novel functions**. *Nature* 2003, **423**:185-190.
14. MacKerell AD, Bashford D, Bellott M, Dunbrack RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S *et al.*: **All-atom empirical potential for molecular modeling and dynamics studies of proteins**. *J Phys Chem B* 1998, **102**:3586-3616.
15. Dahiyat BI, Gordon DB, Mayo SL: **Automated design of the surface positions of protein helices**. *Protein Sci* 1997, **6**:1333-1337.
16. Looger LL, Hellinga HW: **Generalized dead-end elimination algorithms make large-scale protein side-chain structure prediction tractable: implications for protein design and structural genomics**. *J Mol Biol* 2001, **307**:429-445.
17. Dwyer MA, Looger LL, Hellinga HW: **Computational design of a biologically active enzyme**. *Science* 2004, **304**:1967-1971.
18. Wisz MS, Hellinga HW: **An empirical model for electrostatic interactions in proteins incorporating multiple geometry-dependent dielectric constants**. *Proteins* 2003, **51**:360-377.
19. Lazaridis T, Karplus M: **Effective energy function for proteins in solution**. *Proteins* 1999, **35**:133-152.
20. Kortemme T, Morozov AV, Baker D: **An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes**. *J Mol Biol* 2003, **326**:1239-1259.
21. Dahiyat BI, Mayo SL: **Probing the role of packing specificity in protein design**. *Proc Natl Acad Sci USA* 1997, **94**:10172-10177.
22. Chothia C: **Hydrophobic bonding and accessible surface area in proteins**. *Nature* 1974, **248**:338-339.
23. Street AG, Mayo SL: **Pairwise calculation of protein solvent-accessible surface areas**. *Fold Des* 1998, **3**:253-258.
24. Choudhury N, Pettitt BM: **On the mechanism of hydrophobic association of nanoscopic solutes**. *J Am Chem Soc* 2005, **127**:3556-3567.
25. • Wagoner JA, Baker NA: **Assessing implicit models for nonpolar mean solvation forces: the importance of dispersion and volume terms**. *Proc Natl Acad Sci USA* 2006, **103**:8331-8336.
The authors developed an alternative to the standard solvent-accessible surface-area model of non-polar solvation energies. Their model accurately predicts non-polar solvation forces from explicit solvent simulations.
26. Eisenberg D, McLachlan AD: **Solvation energy in protein folding and binding**. *Nature* 1986, **319**:199-203.

27. Honig B, Sharp K, Yang AS: **Macroscopic models of aqueous-solutions — biological and chemical applications**. *J Phys Chem* 1993, **97**:1101-1109.

28. Marshall SA, Vizcarra CL, Mayo SL: **One- and two-body decomposable Poisson-Boltzmann methods for protein design calculations**. *Protein Sci* 2005, **14**:1293-1304.

29. Bashford D, Case DA: **Generalized Born models of macromolecular solvation effects**. *Annu Rev Phys Chem* 2000, **51**:129-152.

30. Pokala N, Handel TM: **Energy functions for protein design: adjustment with protein-protein complex affinities, models for the unfolded state, and negative design of solubility and specificity**. *J Mol Biol* 2005, **347**:203-227.

31. Lee MS, Feig M, Salsbury FR Jr, Brooks CL III: **New analytic approximation to the standard molecular volume definition and its application to generalized Born calculations**. *J Comput Chem* 2003, **24**:1348-1356.

32. Yu Z, Jacobson MP, Friesner RA: **What role do surfaces play in GB models? A new-generation of surface-generalized Born model based on a novel Gaussian surface for biomolecules**. *J Comput Chem* 2006, **27**:72-89.

33. Feig M, Onufriev A, Lee MS, Im W, Case DA, Brooks CL III: **Performance comparison of generalized born and Poisson methods in the calculation of electrostatic solvation energies for protein structures**. *J Comput Chem* 2004, **25**:265-284.

34. Schymkowitz JW, Rousseau F, Martins IC, Ferkinghoff-Borg J, Stricher F, Serrano L: **Prediction of water and metal binding sites and their affinities by using the Fold-X force field**. *Proc Natl Acad Sci USA* 2005, **102**:10147-10152.

35. • Jiang L, Kuhlman B, Kortemme T, Baker D: **A ''solvated rotamer'' approach to modeling water-mediated hydrogen bonds at protein-protein interfaces**. *Proteins* 2005, **58**:893-904.

The authors present a new approach to including explicit solvent in protein design calculations. They include solvent atoms in their rotamers and use this approach to recover sidechain identities at single positions in protein interfaces.

36. Morozov AV, Kortemme T, Tsemekhman K, Baker D: **Close agreement between the orientation dependence of hydrogen bonds observed in protein structures and quantum mechanical calculations**. *Proc Natl Acad Sci USA* 2004, **101**:6946-6951.

37. Morozov AV, Kortemme T: **Potential functions for hydrogen bonds in protein structure prediction and design**. *Adv Protein Chem* 2005, **72**:1-38.

38. Friesner RA: **Modeling polarization in proteins and protein-ligand complexes: Methods and preliminary results**. *Adv Protein Chem* 2006, **72**:79-104.

39. Maple JR, Cao YX, Damm WG, Halgren TA, Kaminski GA, Zhang LY, Friesner RA: **A polarizable force field and continuum solvation methodology for modeling of protein-ligand interactions**. *Journal of Chemical Theory and Computation* 2005, **1**:694-715.

40. Friesner RA: ***Ab initio* quantum chemistry: methodology and applications**. *Proc Natl Acad Sci USA* 2005, **102**:6648-6653.

41. Cho AE, Guallar V, Berne BJ, Friesner R: **Importance of accurate charges in molecular docking: quantum mechanical/ molecular mechanical (QM/MM) approach**. *J Comput Chem* 2005, **26**:915-931.

42. • Clark LA, Boriack-Sjodin PA, Eldredge J, Fitch C, Friedman B, Hanf KJ, Jarpe M, Liparoto SF, Li Y, Lugovskoy A *et al.*: **Affinity enhancement of an *in vivo* matured therapeutic antibody using structure-based computational design**. *Protein Sci* 2006, **15**:949-960.

The authors used a variety of different design algorithms to optimize the affinity of an antibody–ligand interaction. The affinity was improved by an order of magnitude and the crystal structure shows that the design makes the predicted contacts.

43. Lazar GA, Dang W, Karki S, Vafa O, Peng JS, Hyun L, Chan C, Chung HS, Eivazi A, Yoder SC *et al.*: **Engineered antibody Fc variants with enhanced effector function**. *Proc Natl Acad Sci USA* 2006, **103**:4005-4010.

44. •• Ashworth J, Havranek JJ, Duarte CM, Sussman D, Monnat RJ Jr, Stoddard BL, Baker D: **Computational redesign of endonuclease DNA binding and cleavage specificity**. *Nature* 2006, **441**:656-659.

In this remarkable paper, the authors redesigned the cleavage specificity of an endonuclease. The redesigned enzyme cleaves the new target sequence and its crystal structure matches the computational prediction.

45. • Ambroggio XI, Kuhlman B: **Computational design of a single amino acid sequence that can switch between two distinct protein folds**. *J Am Chem Soc* 2006, **128**:1154-1161.

The authors developed a design procedure for optimizing a single amino acid sequence for multiple target structures. They used this approach to design a protein that reversibly switches between two distinct protein folds upon a change in pH or cobalt concentration.