Chapter 2: Potential energy functions for protein design

This chapter has been adapted from: Boas FE and Harbury PB. (2007) "Potential energy functions for protein design." *Current Opinion in Structural Biology*. 17: 199-204.

Summary

Different potential energy functions have been used in protein dynamics simulations, protein design calculations, and protein structure prediction. Clearly, the same physics applies in all three cases, so the variation in potential energy functions reflects 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. Recently improved models of polarization, the hydrophobic effect, and hydrogen bonding may be applicable to both molecular mechanics and protein design.

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, where 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 tradeoff 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 require 10 million years on a typical desktop computer. Protein design algorithms use several strategies to speed up the process. First, protein design algorithms do not simulate kinetics, but rather 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, but rather use a continuum representation of water. Finally, protein design algorithms generally use less computationally intensive energy functions than molecular mechanics calculations.

Previous reviews have described potential energy functions (PEFs) used for molecular mechanics simulations,[23,24] protein design,[25,26] and protein structure prediction.[27] In this review, we compare these energy functions (Figure 5). We also describe advances in the molecular mechanics field that could be used in the next generation of design algorithms.

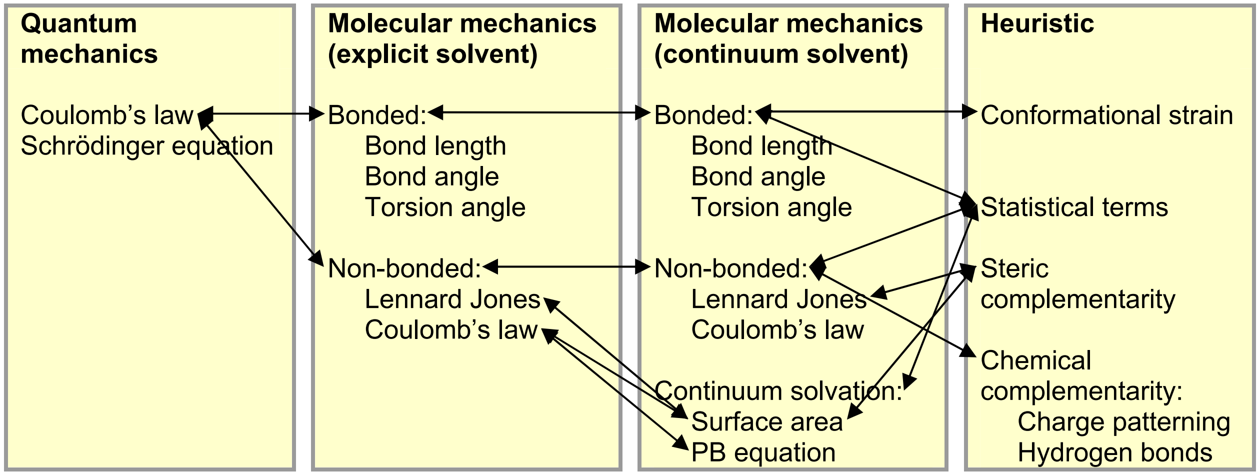

Figure 5. Proteins can be modeled at different levels of detail.

Potential energy functions 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 molecular mechanics potential energy functions, which are the most thoroughly tested models of molecular energetics. Currently, the protein design field uses heuristic energy functions, but the trend is towards using more physically based potential energy functions.

Potential energy functions

Overview

Molecular mechanics potential energy functions (MM-PEFs) incorporate two types of terms: “bonded” and “non-bonded” (Figure 6). The bonded terms apply to sets of 2 to 4 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 function (which includes van der Waals attraction, and repulsion due 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.[23,24] MM-PEF’s have been used predominately to simulate protein folding and dynamics, and are also used to refine X-ray crystal structures.

An alternative type of potential energy function is the knowledge-based, or statistical, energy function[27,28] (Figure 7). 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 vs. surface environments), or the probabilities that pairs of residues appear together in a defined relative geometry is 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 based on chance.[29-31] The advantage of a knowledge-based energy function is that it can model any behavior seen in known protein crystal structures, even if no good physical understanding of the behavior exists. The disadvantage is that these energy functions are phenomenological and can't predict new behaviors absent from the training set.

Design potentials 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 labs. The various terms are typically calibrated and weighted to optimize performance for one type of prediction, such as experimental binding energy,[12,32] or to produce native-like sequences when redesigning natural proteins.[7] By way of illustration, we describe the potential energy functions used in two recent landmark protein design papers. In the first example, Looger *et al.* redesigned various bacterial periplasmic binding proteins to bind trinitrotoluene, lactate, and serotonin.[2] Their energy function included a Lennard-Jones 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 bonding term derived from the DREIDING MM-PEF,[33] a surface area-based solvation term, a knowledge-based rotamer term,[34] 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,[3] using a more accurate electrostatics model that included multiple geometry-dependent dielectric constants.[35] In the second example, Kuhlman *et al.* designed a 93-residue protein with a new α/β fold.[7] Their energy function included an LJ term (with radii fit to match the distribution of distances seen in the PDB, and well depths from CHARMM19), a Lazaridis-Karplus empirical solvation term,[36] a knowledge-based hydrogen bonding term,[37] 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 either ignored, modeled implicitly, or sampled at low resolution. We examine this question term-by-term in the following sections.

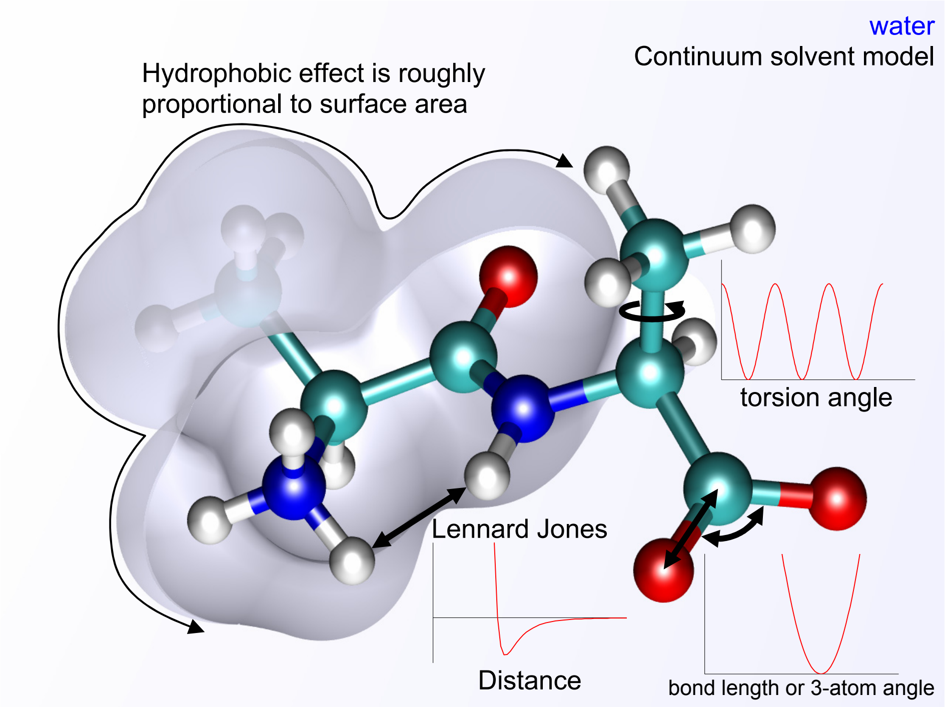
water
Continuum solvent model
Hydrophobic effect is roughly proportional to surface area
torsion angle
Lennard Jones
Distance
bond length or 3-atom angle

Figure 6. Molecular mechanics potential energy function with continuum solvent.

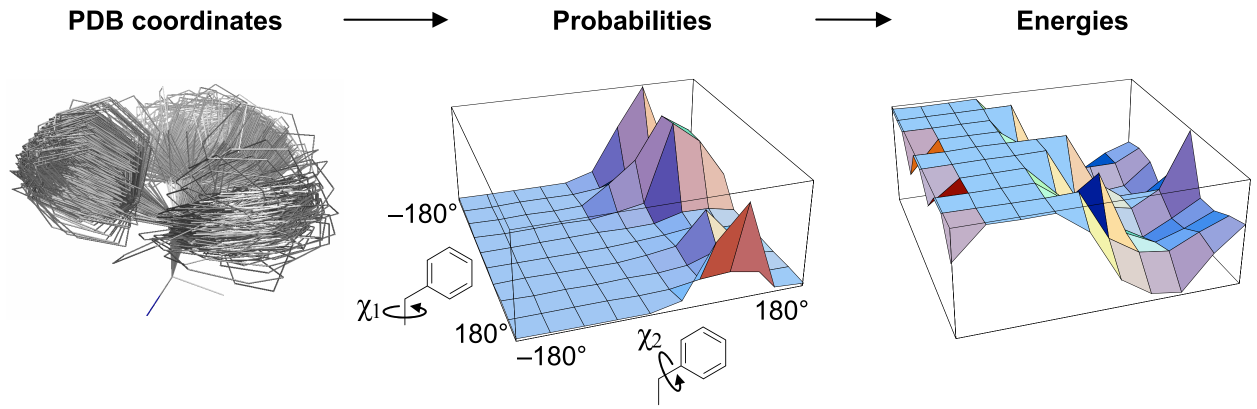

Figure 7. Knowledge-based potential energy function.

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 Lennard-Jones (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 kcal/mol 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 Lennard Jones 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,[38] the repulsive component of the LJ energy can be scaled down,[2] or the LJ function can be linearly extrapolated below a cutoff distance.[7]

Softening the LJ function is based on a presumption that protein cores are reasonably fluid and thus can always rearrange to accommodate small overlaps. However, this modification 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 solvent 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’s 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.[39] Pairwise-factorable approximations of surface area have been developed for use in design calculations.[40] 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.[41] This type of interaction is completely absent from a surface-area based energy. Wagoner and Baker have developed a model[42] of the hydrophobic effect that captures such complex wetting phenomena, and produces energies that are closer to explicit solvent simulations than are surface-area based energies. 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.[34-36,43] 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 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[44] 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.*[45] describe a pairwise-factorable approximation to the PB equation based on summing precalculated energies for single residues and for pairs of residues. However, 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[15] provides a fast approximate solution to the Poisson-Boltzmann equation, and it has been used for protein design.[46] Recent improvements to the generalized Born functional form[47,48] yield solvation energies that are comparable to those derived from finite-difference calculations.[49]

Explicit water

Continuum solvent models break down when water molecules are tightly bound to proteins. However, it may be possible to incorporate a handful of explicit water molecules in a continuum solvent calculation. Schymkowitz *et al.* developed a method for predicting positions of tightly bound water molecules in proteins.[50] Jiang *et al.* show how to incorporate water molecules into amino acid rotamers.[51]

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.* showed that the bent geometry is indeed preferred according to quantum calculations and crystal structures in the PDB.[52] Using the PDB statistics, they developed a knowledge-based hydrogen bonding energy function[37,53] and used it to design a new protein.[7]

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.[54,55] Importantly, solute polarization breaks down the pairwise-factorability property of traditional MM-PEFs. MM-PEFs also do not model chemical realities such as bond formation, partial covalent character of hydrogen bonds, and lone pairs. One possible compromise is to model key parts of the protein using quantum mechanics, and the rest of the protein using molecular mechanics.[56,57]

Reference states

Protein design potentials frequently use implicit reference states. The MM-PEF can only tell the energy difference between different conformations of the same sequence. To compare different sequences, we must subtract the energy of each sequence in an alternative undesired conformation, such as the unfolded or unbound states. These undesired conformations are typically treated implicitly by subtracting a fixed reference energy for each amino acid.

The unfolded and unbound states can also be modeled explicitly. For example, the unfolded state can be modeled using fixed reference energies for each amino acid, plus a random walk model of long range electrostatics.[58] The unbound state can be modeled explicitly using the same structural optimization algorithm used on the bound state.

Modeling the correct reference states is critical to calculating the binding energy of a complex. For example, the binding energy due to a salt bridge or hydrogen bond is the interaction energy of the charges in the bound conformation, relative to their interaction energies with water in the unbound conformation. A typical salt bridge might have a Coulomb interaction energy of 50 kcal/mol, but this is almost completely canceled out by the charges interacting with water in the unbound state. Thus, accurate calculations of the energies of both the bound and unbound structures are needed to calculate accurate binding energies. In many cases, salt bridges are actually destabilizing relative to a hydrophobic interaction[59]: the charges would prefer to interact with water than with each other.

Search algorithms

Three major algorithms have been used to search through sequence and conformational space in protein design. Many variations and hybrid algorithms are possible, but here we describe a typically implementation of each algorithm, and briefly discuss the advantages and disadvantages of each.

The dead-end elimination (DEE) algorithm[34,60] starts with a set of rotamers at each position in the protein, and a precalculated matrix of interaction energies between these rotamers. The algorithm uses a series of filters to eliminate rotamers that provably can not be present in the global energy minimum. Typically, a large fraction of the rotamers can be eliminated, and the remaining rotamers are searched by exhaustive enumeration or Monte Carlo search. The advantage of DEE is that very large sequence and structural spaces can be searched comprehensively. Sequence and structural space are typically searched simultaneously, which requires the use of an implicit reference state. In other words, undesired conformations such as the unfolded or unbound states are typically not modeled explicitly, but rather are treated using fixed reference energies for each amino acid. Thus, for example, dead end elimination does not distinguish between intramolecular and intermolecular interactions, and will propose mutations that stabilize the protein without improving its interaction with the ligand.[61] Reference states could be included if DEE were only used for structural optimization of single sequences, with another procedure used for sequence optimization. This is typically not done because it is much faster to optimize sequence and structure simultaneously.

The mean field algorithm[19,62] also uses a rotamer-based picture with precomputed energy matrices. However, rather than finding a single low energy structure, mean field treats the protein as a probabilistic ensemble. Each rotamer is assigned a probability, and these probabilities are updated iteratively to match the Boltzmann distribution. The final probabilities can be used to calculate the protein's conformational entropy. The mean field algorithm is typically used to optimize the structure of a single sequence, and the sequence optimization is typically done using a genetic algorithm. The advantage of this approach is that undesired conformations such as the bound, unbound, and unfolded states can be modeled explicitly. The use of multiple states allows for stability, affinity, and specificity to be explicitly calculated and optimized.

Monte Carlo methods[63-65] typically start with a single protein structure, and use a set of moves to perturb this structure. If the new structure has a lower energy, then it is accepted. If the new structure has higher energy, then it is accepted with probability $e^{-\Delta E/RT}$, where ΔE is the energy change, and T is the temperature, which is slowly annealed to 0. The advantage of this approach is that there is no need to precompute large energy matrices. Thus, it is CPU-intensive rather than memory-intensive, which better matches today's distributed computing systems. Furthermore, the energy function can include non-pairwise additive terms such as polarization. The Monte Carlo moves can include randomly switching from one rotamer conformation to another, but they can also include non-rotameric moves.

The Baker lab has developed a clever strategy for including backbone flexibility in protein design[7,66]. They alternate between sequence design on a fixed backbone, and structural optimization for a designed sequence.

Conclusions and future directions

The techniques described above have been used to design proteins with a wide variety of new functions. Clark *et al.*[60] optimized the recombining site of an antibody to increase the ligand affinity, and Lazar *et al.*[10] optimized the Fc region of an antibody to bind more tightly to the Fc receptor. Ashworth *et al.*[64] redesigned an endonuclease to recognize and cut a heterologous DNA sequence. Kuhlman *et al.*[65]

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 accessible by protein design. As potential energy functions, 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 P.B.H.). F.E.B. was supported by a training grant from the National Institute of General Medical Sciences (5T32 GM07365-28).