

Energy functions for protein design

D Benjamin Gordon*, Shannon A Marshall* and Stephen L Mayo†

Recent successes in protein design have illustrated the promise of computational approaches. These methods rely on energy expressions to evaluate the quality of different amino acid sequences for target protein structures. The force fields optimized for design differ from those typically used in molecular mechanics and molecular dynamics calculations.

Addresses

*Division of Chemistry and Chemical Engineering, California Institute of Technology, MC 147-75, Pasadena, CA 91125, USA

†Howard Hughes Medical Institute, Division of Biology, California Institute of Technology, MC 147-75, Pasadena, CA 91125, USA; e-mail: steve@mayo.caltech.edu

Current Opinion in Structural Biology 1999, **9**:509–513

<http://biomednet.com/elecref/0959440X00900509>

© Elsevier Science Ltd ISSN 0959-440X

Introduction

Computational protein design is a general, closed-loop approach for finding the optimal sequence of amino acids for a desired protein fold [1•]. A potential energy function that represents the dominant factors, as well as the subtleties, of protein stability is used to predict the energy of each possible amino acid sequence for a target protein structure. Current design efforts have used fixed protein backbones as target structures, with two notable exceptions [2,3,4••]. Atomic-level detail is introduced by using statistically significant sidechain conformations, called rotamers [5], to represent the flexibility of each amino acid. A variety of stochastic and deterministic search algorithms [6] are then used to find the optimal combination of amino acid sidechain rotamers for the target structure, as ranked by the potential energy function. Finally, the experimentally determined stability and structure of the designed proteins are analyzed and rational improvements to the potential function are implemented.

The purpose of this review is to discuss the development of protein design force fields and to survey the potential energy terms that have been used thus far. The terms fall into five broad categories. First, we discuss the energies describing the packing among atoms that are not covalently bonded. Nonbonded polar interactions are considered next. We briefly survey internal coordinate energies and finally examine solvation and entropy, which are computed differently than in typical molecular mechanics force fields.

Force-field requirements

Protein design presents a demanding task for a potential energy function. Design potentials must be sensitive to the subtle changes in amino acid identity that are known to perturb the experimental stability of proteins. Design

force fields should not be overly sensitive to small variations in rotamer geometry, however, as discrete rotamers are used to model sidechain conformations. The force field also must be compatible with the computational requirements of protein design. For example, most search algorithms demand that the energy terms be pairwise decomposable and design problems with large combinatorial complexity require energy terms that can be calculated quickly.

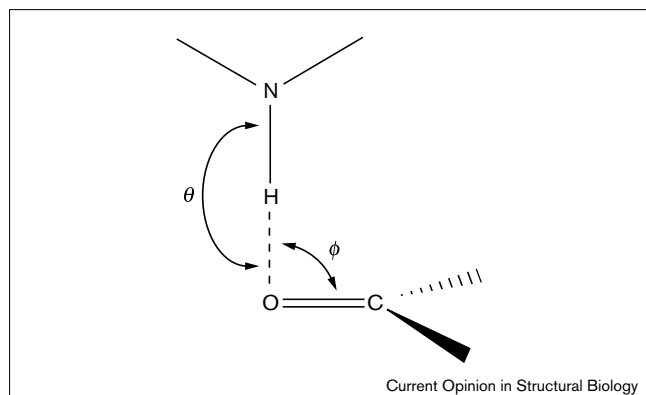
As the energies produced by design potentials are intended to correlate with the free energy of folding, the force field must model the unfolded state, as well as the folded state. Experimental and theoretical studies [7] indicate that unfolded proteins can sometimes have residual structure and mutations may alter the properties of the unfolded state ensemble. In design calculations, however, the unfolded state is commonly assumed to have no residual structure: nonbonded interactions among sidechains are considered to be insignificant, the sidechains are assumed to be fully solvated, all rotamers are modeled as being equally probable and all sequences in the unfolded state are isoenergetic.

Because of the demands posed by protein design, the force fields that are widely used to perform molecular mechanics calculations, such as CHARMM [8], AMBER [9,10] and DREIDING [11], are not necessarily appropriate for design. Similarly, the statistically derived pair potentials that are quite effective in structure compatibility studies [12] do not manifest the structural sensitivity that is necessary for protein design. Instead, new force fields must be developed for protein design that properly balance each factor described by the potential energy function. Over the past few years, the first force fields tailored for design have been constructed. Very few potential energy terms have been used in these force fields, however, and even fewer have been evaluated through a comparison of design predictions and experimental results. Future progress in protein design force fields will be realized by the continued, systematic experimental validation of the terms comprising the potential function.

van der Waals

Packing specificity is critical to protein design. For protein core calculations, which comprise the majority of design studies, a force field that models only packing specificity is sufficient to design well-folded proteins [13–16]. Although packing can be evaluated exclusively using interatomic distance restraints [17], most design programs utilize a van der Waals potential. This potential provides a physical basis for sidechain packing specificity, thereby favoring native-like folded states with well-organized cores and selecting against disordered or molten-globule states. The

Figure 1



An example of a nonphysical hydrogen-bond geometry that can be selected when a hydrogen-bond potential that is dependent only on θ is used for protein design. A more restrictive hydrogen-bond potential, described in Equations (3) through (6), correctly predicts that no favorable interaction is present because $\phi = 90^\circ$.

van der Waals energy, E_{vdW} , is typically calculated using a Lennard–Jones 12–6 expression:

$$E_{vdW} = D_0 \left[\left(\frac{R_0}{R} \right)^{12} - 2 \left(\frac{R_0}{R} \right)^6 \right] \quad (1)$$

The interatomic distance, R , is computed from atomic coordinates. The equilibrium radii, R_0 , and well depths, D_0 , are parameters that are defined within each force field.

Two examinations of van der Waals parameters underscore the need to tune molecular mechanics potential functions to protein design. Lazar and co-workers [16] compared the predictive ability of variations of Hagler and AMBER van der Waals parameters for a set of ubiquitin variants with redesigned cores. United-atom parameters from AMBER95 were markedly superior to the other variations when used in conjunction with a detailed rotamer library. Dahiyat and Mayo [15] generated sequences by systematically varying the scale of the atomic radii, based on the DREIDING parameter set, and by using rotamers with explicit hydrogen atoms. Scaling the radii by a factor of 0.90 achieved the optimal balance between packing specificity and hydrophobic collapse, as represented by a solvation term (discussed below).

Hydrogen bonding

As the majority of computational protein design studies have focused on protein cores, electrostatic and hydrogen-bonding terms have not been as thoroughly validated by experiment. Nevertheless, initial forays have proven these terms to be useful for the design of helical surfaces [18] and for full sequence design [19**].

Hydrogen bonds are typically represented by an angle-dependent, 12–10 hydrogen-bond potential:

$$E_{HB} = D_0 \left[5 \left(\frac{R_0}{R} \right)^{12} - 6 \left(\frac{R_0}{R} \right)^{10} \right] F(\theta) \quad (2)$$

where R_0 is the equilibrium distance, D_0 is the well depth and R is the interatomic distance between the donor and acceptor heavy atoms. The angle-dependence term, $F(\theta)$, is typically $\cos^4\theta$, where θ is the donor–hydrogen–acceptor angle.

We have observed that calculations performed with the above potential will allow rotameric arrangements with nonphysical hydrogen-bond geometries, as shown in Figure 1. To circumvent this problem, we employ more restrictive hybridization-dependent angle-dependence terms that enforce reasonable geometries [18]:

$$\begin{aligned} sp^3 \text{ donor} - sp^3 \text{ acceptor} \quad F &= \cos^2 \theta \cos^2 (\phi - 109.5) \\ \theta > 90^\circ, \phi - 109.5^\circ < 90^\circ \quad (3) \end{aligned}$$

$$\begin{aligned} sp^3 \text{ donor} - sp^2 \text{ acceptor} \quad F &= \cos^2 \theta \cos^2 \phi \\ \phi > 90^\circ \quad (4) \end{aligned}$$

$$sp^2 \text{ donor} - sp^3 \text{ acceptor} \quad F = \cos^4 \theta \quad (5)$$

$$sp^2 \text{ donor} - sp^2 \text{ acceptor} \quad F = \cos^2 \theta \cos^2 (\max[\phi, \varphi]) \quad (6)$$

The angles ϕ and φ refer to the hydrogen–acceptor–base angle (where the base is the atom covalently attached to the acceptor) and the angle between the normals of the planes defined by the six atoms attached to the two sp^2 centers, respectively.

A potential energy term based on the above equations allows only physically reasonable sidechain–sidechain and sidechain–backbone hydrogen bonds. Unfortunately, using a highly restrictive energy term in combination with a discrete rotamer library causes the force field to predict poor energies for some sequences that may actually form good hydrogen-bond interactions.

Electrostatics

The role of electrostatics in protein stability is subject to debate. At moderate temperatures, favorable electrostatic interactions are not thought to be strong enough to compensate for the energy of desolvation [20]. In more extreme conditions, however, salt bridges may stabilize proteins [21,22]. Moreover, electrostatics may play a more significant role in defining the specificity, rather than the stability, of folding and of functional interactions [23–26].

Computational protein design efforts have not yet developed an electrostatic term intended to represent these considerations. Rather, electrostatics are used sparingly, primarily to guard against destabilizing interactions between like-charged residues. The simplest treatment of electrostatic interactions is based on Coulomb's law, which describes the energy of two charges, Q_i and Q_j , separated by distance R in a medium with dielectric constant ϵ :

$$E_{elec} = 322.0637 \left(\frac{Q_i Q_j}{\epsilon R} \right). \quad (7)$$

Our laboratory uses a distance-attenuated version of Coulomb's law, with an effective dielectric constant value of $40R$ and partial atomic charges that give a total coulombic energy of approximately ± 1 kcal/mol for the interaction between juxtaposed charged residues. Thus, electrostatic contributions to the total energy are only significant when charged atoms are in close proximity. In sharp contrast, electrostatic energy is often the largest contributor to the total energy in potentials used for molecular mechanics and dynamics calculations.

Internal coordinate terms

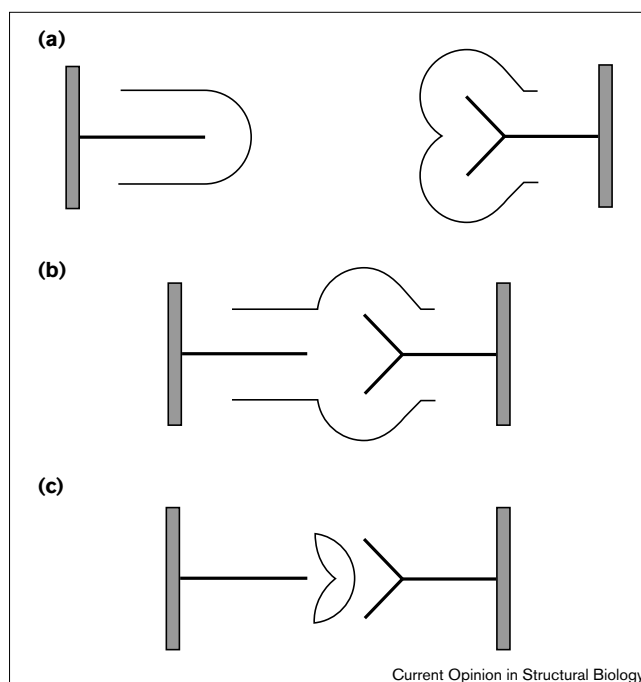
Typical molecular mechanics force fields have terms that evaluate bonds, angles, torsions and inversions among atoms that are covalently attached. These internal coordinate or 'bonded' energies must be considered when generating rotamers or modifying the protein backbone and, in some cases, have been used for protein design [4**,16]. The usefulness of these terms for design, however, has not been rigorously demonstrated. As rotamers derived from the statistical analysis of protein structure databases generally have good internal coordinate energies, many design potential functions do not include them at all.

Solvation

Because the hydrophobic effect drives protein folding [27], modeling solvation effects is critical to a protein design force field. The computational expense of explicitly modeling protein-solvent interactions for all the sequences under consideration is, however, prohibitively expensive. Therefore, several groups have employed approximate methods utilizing octanol-water and gas-water free energy of transfer data for each amino acid [28,29]. The experimentally measured free energies of transfer are correlated with the molecular surface area [30], as shown in Figure 2. These energies are either used directly for residues in the protein core [31] or they are scaled by the change in the solvent-exposed surface area that is associated with protein folding [14,32].

The energy required to transfer a sidechain from a solvated, unfolded protein to a partially or completely desolvated position in the folded protein is not necessarily the same as the transfer energy from water to gas or to a nonpolar solvent. But, the approximate linear relationship

Figure 2



Pairwise calculation of buried surface areas. **(a)** Unfolded or reference-exposed surface areas of two sidechain rotamers. **(b)** Folded, exposed surface area for the rotamer pair. **(c)** Buried surface area for the rotamer pair, calculated by subtracting (b) from (a).

between transfer energy and the change in surface area should be correct for both cases. Dahiyat and Mayo [14] determined the optimal values for polar and nonpolar atomic solvation parameters by fitting them to the experimentally determined stability of designed proteins. The inclusion of a hydrophobic burial benefit and a polar burial penalty in the protein design force field provides a significant improvement in the predictive power compared with a force field with only a van der Waals term.

Two other considerations have affected the formulation of a protein design solvation potential. First, a negative design term that penalizes the exposure of nonpolar surface area is sometimes used [15,33]. Although nonpolar exposure should not destabilize a protein, it can lead to aggregation or misfolding. Therefore, a nonpolar exposure penalty is required to limit the amount of exposed, nonpolar surface area at boundary and surface positions [34*]. Second, many optimization algorithms require that energy terms be pairwise decomposable, but the pairwise calculation of buried surface areas leads to significant overcounting. Street and Mayo [35] have developed a pairwise expression with one scalable parameter that closely reproduces both the true buried area and the true exposed solvent-accessible surface areas.

Entropy

A simple entropy term is sometimes incorporated into protein design potential functions [31,32]. The change in

sidechain entropy upon folding is modeled as the change in the number of rotatable bonds, making the assumption that conformational freedom is completely restricted in the folded state. The unfolded state entropies are calculated either by assuming that all the rotamers are equally populated or by fitting to semi-empirical estimates [36]. The inclusion of an entropy term based on the number of rotatable bonds did not significantly improve the correlation between the predicted and observed stabilities of the GCN4-p1 coiled-coil core [14]. This simple model for entropy may have failed because it neglects residual sidechain entropy in folded proteins and possible residual structure in the unfolded state.

Looking forward

Protein design force fields have been successful, in part, because of their stringency. Restrictive functions, such as the van der Waals and the hybridization-dependent hydrogen-bond potentials, in particular, result in a very high rejection rate and a significant false-negative rate. Fortunately, many design force fields also show a low false-positive rate. Therefore, sequences that are selected in protein design studies tend to fold properly, even though many other equally acceptable sequences are rejected.

As a result of the high false-negative rate, potential functions derived through protein design efforts may not be suitable for folding studies. In order to gain a deeper understanding of the determinants of protein stability, it is therefore important to lower the false-negative rate. Softening the restrictive potentials could result in design models that more accurately describe the fundamental relationship among sequence, structure and stability.

Acknowledgements

We wish to thank AG Street for helpful comments on the manuscript. This work was supported by the Howard Hughes Medical Institute (SLM), the Helen G and Arthur McCallum Foundation (DBG), a National Institutes of Health NRSA training grant and the Caltech Initiative in Computational Molecular Biology program, awarded by the Burroughs Wellcome Fund (SAM).

References and recommended reading

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

- of special interest
- of outstanding interest

1. Street AG, Mayo SL: **Computational protein design.** *Structure* 1999, **7**:R105-R109.

This review surveys the computational principles of the protein design cycle, including residue classification, negative design and discretization of sidechain conformations.

2. Harbury PB, Tidor B, Kim PS: **Repacking protein cores with backbone freedom: structure prediction for coiled coils.** *Proc Natl Acad Sci USA* 1995, **92**:8408-8412.
3. Su A, Mayo SL: **Coupling backbone flexibility and amino acid sequence selection in protein design.** *Protein Sci* 1997, **6**:1701-1707.
4. Harbury PB, Plecs JJ, Tidor B, Alber T, Kim PS: **High-resolution protein design with backbone freedom.** *Science* 1998, **282**:1462-1467.

The *de novo* design of an unnatural fold, a right-handed coiled coil, was accomplished using a computation that incorporates both sidechain and backbone flexibility.

5. Ponder JW, Richards FM: **Tertiary templates for proteins – use of packing criteria in the enumeration of allowed sequences for different structural classes.** *J Mol Biol* 1987, **193**:775-791.
6. Desjarlais JR, Clarke ND: **Computer search algorithms in protein modification and design.** *Curr Opin Struct Biol* 1998, **8**:471-475.
7. Dill KA, Shortle D: **Denatured states of proteins.** *Annu Rev Biochem* 1991, **60**:795-825.
8. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M: **CHARMM: a program for macromolecular energy, minimization, and dynamics calculations.** *J Comput Chem* 1983, **4**:187-217.
9. Weiner SJ, Kollman PA, Case DA, Singh UC, Ghio C, Alagona G, Profeta SJ, Weiner P: **A new force field for molecular mechanical simulation of nucleic acids and proteins.** *J Am Chem Soc* 1984, **106**:765-784.
10. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM Jr, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA: **A second-generation force field for the simulation of proteins, nucleic acids, and organic molecules.** *J Am Chem Soc* 1995, **117**:5179-5197.
11. Mayo SL, Olafson BD, Goddard WA III: **Dreiding – a generic force-field for molecular simulations.** *J Phys Chem* 1990, **94**:8897-8909.
12. Bowie JU, Luthy R, Eisenberg D: **A method to identify protein sequences that fold into a known three-dimensional structure.** *Science* 1991, **253**:164-170.
13. Desjarlais JR, Handel TM: **De novo design of the hydrophobic cores of proteins.** *Protein Sci* 1995, **4**:2006-2018.
14. Dahiyat BI, Mayo SL: **Protein design automation.** *Protein Sci* 1996, **5**:895-903.
15. Dahiyat BI, Mayo SL: **Probing the role of packing specificity in protein design.** *Proc Natl Acad Sci USA* 1997, **94**:10172-10177.
16. Lazar GA, Desjarlais JR, Handel TM: **De novo design of the hydrophobic core of ubiquitin.** *Protein Sci* 1997, **6**:1167-1178.
17. Jiang X, Bishop EJ, Farid RS: **A de novo designed protein with properties that characterize natural hyperthermophilic proteins.** *J Am Chem Soc* 1997, **119**:838-839.
18. Dahiyat BI, Gordon DB, Mayo SL: **Automated design of the surface positions of protein helices.** *Protein Sci* 1997, **6**:1333-1337.
19. Dahiyat BI, Mayo SL: **De novo protein design: fully automated sequence selection.** *Science* 1997, **278**:82-87.
This paper describes the first fully automated design and experimental validation of a novel sequence for an entire protein. The force field employed uses several of the energy functions described in this review.
20. Hendsch ZS, Tidor B: **Do salt bridges stabilize proteins – a continuum electrostatic analysis.** *Protein Sci* 1994, **3**:211-226.
21. Elcock AH: **The stability of salt bridges at high temperatures: implications for hyperthermophilic proteins.** *J Mol Biol* 1998, **284**:489-502.
22. de Bakker PIW, Hunenberber PH, McCammon JA: **Molecular dynamics simulations of the hyperthermophilic protein Sac7d from *Sulfolobus acidocaldarius*: contribution of salt bridges to thermostability.** *J Mol Biol* 1999, **285**:1811-1830.
23. Lumb KJ, Kim PS: **A buried polar interaction imparts structural uniqueness in a designed heterodimeric coiled coil.** *Biochemistry* 1995, **34**:8642-8648.
24. Schneider JP, Lear JD, DeGrado WF: **A designed buried salt bridge in a heterodimeric coiled coil.** *J Am Chem Soc* 1997, **119**:5742-5743.
25. Sindelar CV, Hendsch ZS, Tidor B: **Effects of salt bridges on protein structure and design.** *Protein Sci* 1998, **7**:1898-1914.
26. Spek EJ, Bui AH, Lu M, Kallenbach NR: **Surface salt bridges stabilize the GCN4 leucine zipper.** *Protein Sci* 1998, **7**:2431-2437.
27. Dill KA: **Dominant forces in protein folding.** *Biochemistry* 1990, **29**:7133-7155.
28. Fauchère J-L, Plicska V: **Hydrophobic parameters of amino-acid side-chains from the partitioning of N-acetyl-amino-acid amides.** *Eur J Med Chem* 1983, **18**:369-375.
29. Ooi T, Oobatake M, Nementy G, Scheraga HA: **Accessible surface areas as a measure of the thermodynamic parameters of**

- hydration of peptides. *Proc Natl Acad Sci USA* 1987, **84**:3086-3090.
30. Wesson L, Eisenberg D: **Atomic solvation parameters applied to molecular dynamics of proteins in solution.** *Protein Sci* 1992, **1**:227-235.
31. Hellinga HW, Richards FM: **Optimal sequence selection in proteins of known structure by simulated evolution.** *Proc Natl Acad Sci USA* 1994, **91**:5803-5807.
32. Kono H, Nishiyama M, Tanokura M, Doi J: **Designing the hydrophobic core of *Thermus flavus* malate dehydrogenase based on side-chain packing.** *Protein Eng* 1998, **11**:47-52.
33. Sun S, Brem R, Chan HS, Dill KA: **Designing amino acid sequences to fold with good hydrophobic cores.** *Protein Eng* 1995, **8**:1205-1213.
34. Malakauskas SM, Mayo SL: **Design, structure and stability of a hyperthermophilic protein variant.** *Nat Struct Biol* 1998, **5**:470-475. The authors impart hyperthermal stability to a mesophilic protein by applying principles of computational protein design to amino acid residues positioned at the boundary between the protein core and the surface.
35. Street AG, Mayo SL: **Pairwise calculation of protein solvent accessible surface areas.** *Fold Des* 1998, **3**:253-258.
36. Sternberg MJE, Chickos JS: **Protein side-chain conformational entropy derived from fusion data – comparison with other empirical scales.** *Protein Eng* 1994, **7**:149-155.