Pairwise calculation of protein solvent-accessible surface areas

Arthur G Street[1] and Stephen L Mayo[2]

Background: The tractability of many algorithms for determining the energy state of a system depends on the pairwise nature of an energy expression. Some energy terms, such as the standard implementation of the van der Waals potential, satisfy this criterion whereas others do not. One class of important potentials that are not pairwise involves benefits and penalties for burying hydrophobic and/or polar surface areas. It has been found previously that, in some cases, a pairwise approximation to these surface areas correlates with the true surface areas. We set out to generalize the applicability of this approximation.

Results: We develop a pairwise expression with one scalable parameter that closely reproduces both the true buried and the true exposed solvent-accessible surface areas. We then refit our previously published coiled-coil stability data to give solvation parameters of 26 cal/mol Å^2 favoring hydrophobic burial and 100 cal/mol Å^2 opposing polar burial.

Conclusions: An accurate pairwise approximation to calculate exposed and buried protein solvent-accessible surface area is achieved.

Addresses: [1]Division of Physics, Mathematics and Astronomy, California Institute of Technology, MC 147-75, Pasadena, CA 91125, USA. [2]Howard Hughes Medical Institute and Division of Biology, California Institute of Technology, MC 147-75, Pasadena, CA 91125, USA.

Correspondence: Stephen L Mayo
E-mail: steve@mayo.caltech.edu

Key words: dead-end elimination, hydrophobic burial, pairwise surface area, protein design, solvent-accessible surface area

Received: **12 January 1998**
Revisions requested: **16 February 1998**
Revisions received: **16 March 1998**
Accepted: **23 March 1998**

Published: **03 June 1998**
http://biomednet.com/elecref/1359027800300253

Folding & Design 03 June 1998, **3:**253–258

Introduction

Many energy minimization schemes require an energy expression that depends exclusively on the superposition of two-body interactions. Of particular interest to us is the dead-end elimination theorem [1], which allows at most two-body interactions between amino acid sidechain rotamers and the protein backbone (or template) and between pairs of rotamers. Terms that depend on more than two bodies cannot be included. This leads to a general problem of accommodating surface area dependent terms in such energy expressions because the buried and/or exposed surface areas of three or more interacting bodies cannot be calculated exactly as the sum of two body interactions.

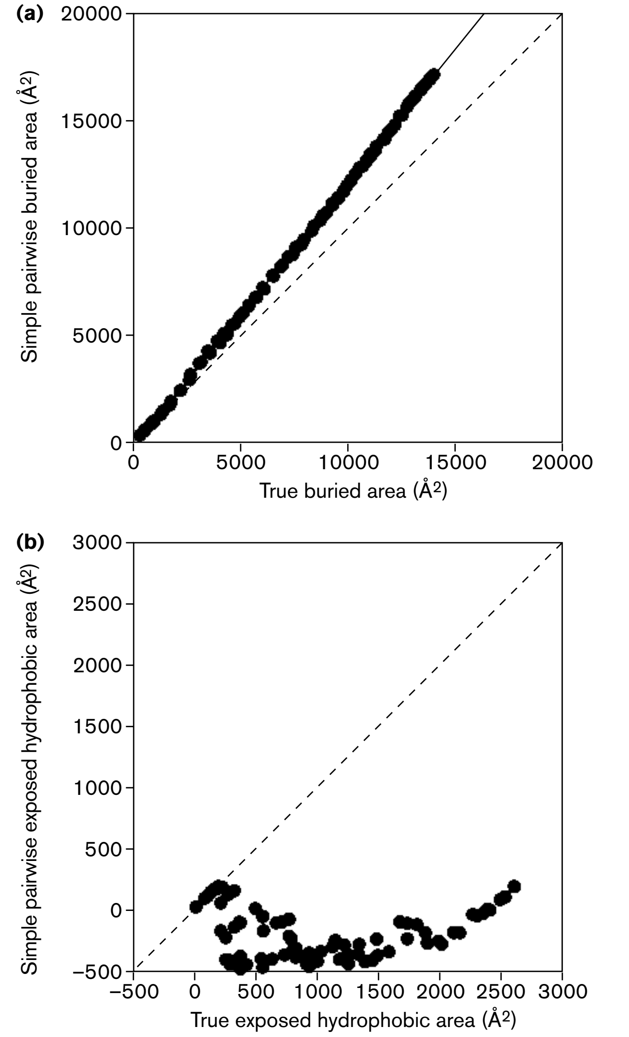
The problem is exacerbated when calculating surface areas using the Lee and Richards' [2] definition of solvent-accessible surface area, in which 1.4 Å is added to every atomic radius before calculation of the area. This increases the number of intersecting atoms and makes an accurate calculation of solvent-accessible surface areas by a two-body method problematic (Figure 1). As Figure 1b shows, a simple two-body method to calculate exposed hydrophobic solvent-accessible surface areas correlates poorly with the true surface areas, and as such limits the use of a simple two-body method in protein design calculations.

A two-body approach has been considered in the context of increasing the speed of calculation of buried hydrophobic surface area for folding studies [3,4] in which the areas of individual atoms or pseudo-atoms were calculated pairwise. These areas were either combined statistically (assuming randomly distributed atoms) or added and scaled and a high correlation with the true Lee and Richards surface areas was found. The use of reduced van der Waals radii to compensate for pairwise over-counting has also been discussed [5,6]. Other (not necessarily pairwise) techniques for calculating surface areas have been reviewed recently [7]. Here, we find empirically that by scaling only the portion of the expression for pairwise area that is subject to over-counting, we can achieve excellent agreement with both the true buried and the true exposed solvent-accessible surface areas.

Results and discussion

The pairwise calculation of surface areas used in this study differs in several key respects from that of our previous work [8]. Here, we include backbone atoms (N, HN, CA, HCA, C and O) in the calculation of surface areas. For each sidechain rotamer r at residue position i with a local tri-peptide backbone $t3$ ([CA, C, O]$_{i-1}$, [N, HN, CA, HCA, C, O]$_i$, [N, HN, CA]$_{i+1}$), we calculate $A^{o}_{i_r t3}$, the exposed area of the rotamer and its backbone in the presence of the local tri-peptide backbone, and $A_{i_r t}$, the exposed area of the rotamer and its backbone in the presence of the entire template t, which is the protein backbone (Figure 2). The difference between $A^{o}_{i_r t3}$ and $A_{i_r t}$ is the total area buried by the template for a rotamer r at residue position i. For each pair of residue positions i and j and for rotamers r and s on i and j, respectively, we calculate $A_{i_r j_s t}$, the exposed area of the rotamer pair in the

Figure 1



Folding & Design

A comparison of true solvent-accessible surface area and the area calculated with the simplest pairwise technique (Equations 1 and 3 with $s = 1$) for subsets of 1mol. **(a)** Buried area. The line of best fit has a slope = 1.24 and a correlation coefficient $R^2 = 1.00$. The differences between calculated and true buried areas are in the range 0–22%. **(b)** Exposed hydrophobic area. The differences between calculated and true areas in the range 0–250% for small areas converge to 100% for areas > 1000 Å². The line of best fit (not shown) has a slope = 0.00 and $R^2 = 0.00$. In each case, a dashed line of slope = 1 is shown.

presence of the entire template. The difference between $A_{i_r j_s t}$ and the sum of $A_{i_r t}$ and $A_{j_s t}$ is the area buried between residues i and j, excluding that area buried by the template. The pairwise approximation to the total buried surface area is:

$$A^{\text{pairwise}}_{\text{buried}} = \sum_i (A^{\text{o}}_{i_r t3} - A_{i_r t}) + s \sum_{i<j} (A_{i_r t} + A_{j_s t} - A_{i_r j_s t}) \quad (1)$$

As shown in Figure 2, the second sum in Equation 1 over-counts the buried area. We have therefore multiplied the second sum by a scale factor s whose value is to be determined empirically. Expected values of s are discussed below.

Noting that the buried and exposed areas should add to the total area:

$$\sum_i A^{\text{o}}_{i_r t3} \quad (2)$$

the solvent-exposed surface area is:

$$A^{\text{pairwise}}_{\text{exposed}} = \sum_i A_{i_r t} - s \sum_{i<j} (\text{A}_{i_r t} + \text{A}_{j_s t} - \text{A}_{i_r j_s t}) \quad (3)$$

The first sum of Equation 3 represents the total exposed area of each rotamer in the context of the protein template ignoring interactions with other rotamers. The second sum of Equation 3 subtracts the buried areas between rotamers and is scaled by the same parameter s as in Equation 1.

Some insight into the expected value of s can be gained from consideration of a close-packed face-centered cubic lattice of spheres of radius r. When the radii are increased from r to R, the surface area on one sphere buried by a neighboring sphere is $2\pi R(R - r)$. We take r to be a carbon radius (1.95 Å), and R is 1.4 Å larger. Then, using:

$$s = \frac{\text{true buried area}}{\text{pairwise buried area}} \quad (4)$$

and noting that each sphere has 12 neighbors, we have:

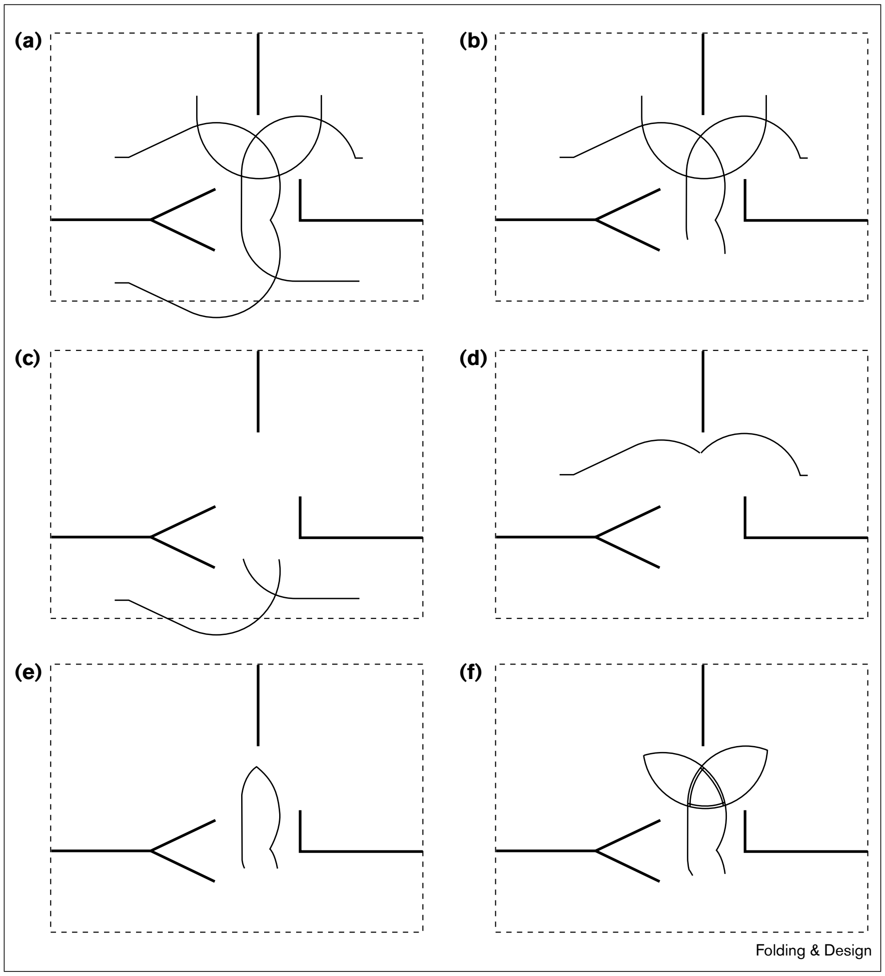
$$s = \frac{4\pi R^2}{12 \times 2\pi R(R - r)} \quad (5)$$

This yields $s = 0.40$. We note that a close-packed face-centered cubic lattice has a packing density of 74% and that protein interiors have a similar packing density, although because many atoms are covalently bonded the close packing is exaggerated [9,10]. We therefore expect $s = 0.40$ to be a lower bound for real protein cores. For non-core residues, where the packing density is lower, we expect a somewhat larger value of s.

We classified residues from ten proteins ranging in size from 54 to 289 residues into core or non-core, as described in the Materials and methods section (Table 1). The classification into core and non-core was made because core residues interact more strongly with one another than do non-core residues. This leads to greater over-counting of the buried surface area for core residues.

Considering the core and non-core cases separately, the value of s that most closely reproduced the true Lee and Richards' surface areas was calculated for the ten proteins. The pairwise approximation very closely matches the true buried surface area (Figure 3). It also performs very well for the exposed hydrophobic surface area of non-core residues (Figure 4b). The calculation of the exposed surface area of the entire core of a protein involves the

Figure 2



Folding & Design

Areas involved in calculating the buried and exposed areas of Equations 1 and 3. The dashed box is the protein template (i.e. the protein backbone), the heavy solid lines correspond to three rotamers at three different residue positions, and the lighter solid lines correspond to surface areas. **(a)** $A^{o}_{i_r t3}$ for each rotamer. **(b)** $A_{i_r t}$ for each rotamer; notice that the template has buried some area from the lower two rotamers. **(c)** $A^{o}_{i_r t3} - A_{i_r t}$ summed over the three residues. The upper residue does not bury any area against the template except that buried in the tri-peptide state $A^{o}_{i_r t3}$. **(d)** $A_{i_r j_s t}$ for one pair of rotamers. **(e)** The area buried between rotamers ($A_{i_r t} + A_{j_s t} - A_{i_r j_s t}$) for the same pair of rotamers as in (d). **(f)** The area buried between rotamers ($A_{i_r t} + A_{j_s t} - A_{i_r j_s t}$) summed over the three pairs of rotamers. The area intersected by all three rotamers (and only that area) is counted twice and is indicated by the double lines. The buried area calculated by Equation 1 is the area buried by the template, represented in (c), plus *s* times the area buried between rotamers, represented in (f). The scaling factor *s* accounts for the overcounting shown by the double lines in (f). The exposed area calculated by Equation 3 is the exposed area in the presence of the template, represented in (b), minus *s* times the area buried between rotamers, represented in (f).

difference of two large and nearly equal areas and is less accurate (Figure 4a); as will be shown, however, when there is a mixture of core and non-core residues, a high accuracy can still be achieved. These calculations indicate that for core residues *s* is 0.42 and for non-core residues *s* is 0.79.

Table 1

Selected proteins, total number of residues and the number of residues in the core and non-core of each protein (glycine and proline were not considered).

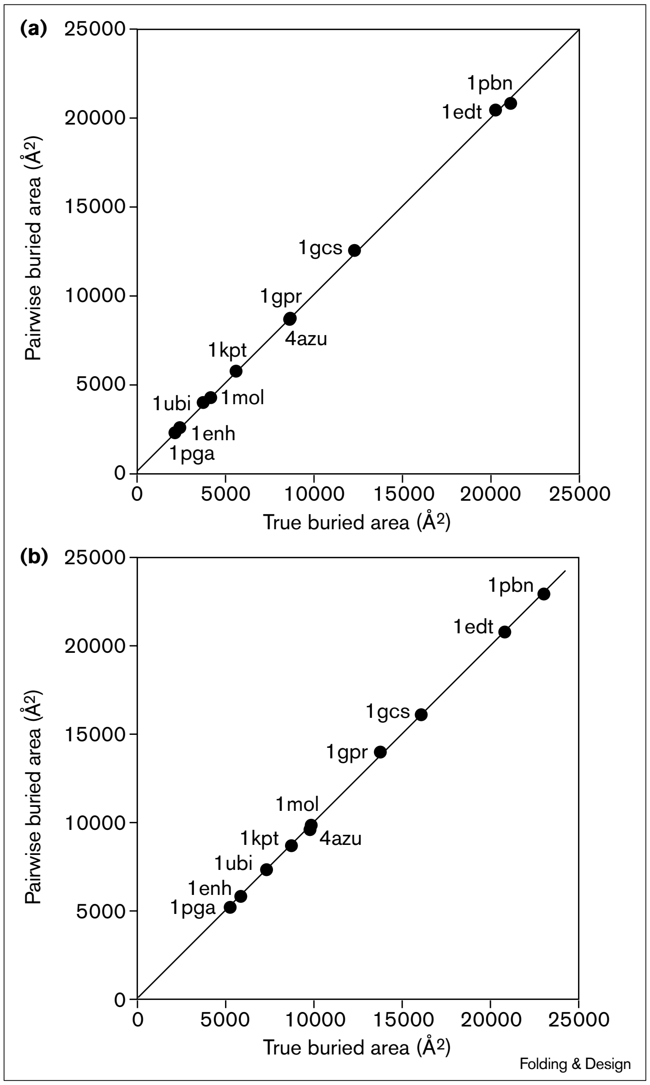
PDB code	Total size	Core size	Non-core size
1enh	54	10	40
1pga	56	10	40
1ubi	76	16	50
1mol	94	19	61
1kpt	105	27	60
4azu-A	128	39	71
1gpr	158	39	89
1gcs	174	53	98
1edt	266	95	133
1pbn	289	96	143

To test whether the classification of residues into core and non-core was sufficient, we examined subsets of interacting residues in the core and non-core positions, and compared the true buried area of each subset with that calculated by Equation 1 (using the above values of *s*). For both subsets of the core and of the non-core, the correlation remained high ($R^2 = 1.00$) indicating that no further classification is necessary (data not shown). (Subsets were generated as follows: given a seed residue, a subset of size two was generated by adding the closest residue; the next closest residue was added for a subset of size three, and this was repeated up to the size of the protein. Additional subsets were generated by selecting different seed residues.)

It remains to apply this approach to calculating the buried or exposed surface areas of an arbitrary selection of interacting core and non-core residues in a protein. When a core residue and a non-core residue interact, we replace Equation 1 with:

$$A^{\text{pairwise}}_{\text{buried}} = \sum_{i} (A^{o}_{i_r t3} - A_{i_r t}) + \sum_{i<j} (s_i \mathrm{A}_{i_r t} + s_j \mathrm{A}_{j_s t} - s_{ij} \mathrm{A}_{i_r j_s t}) \quad (6)$$

Figure 3



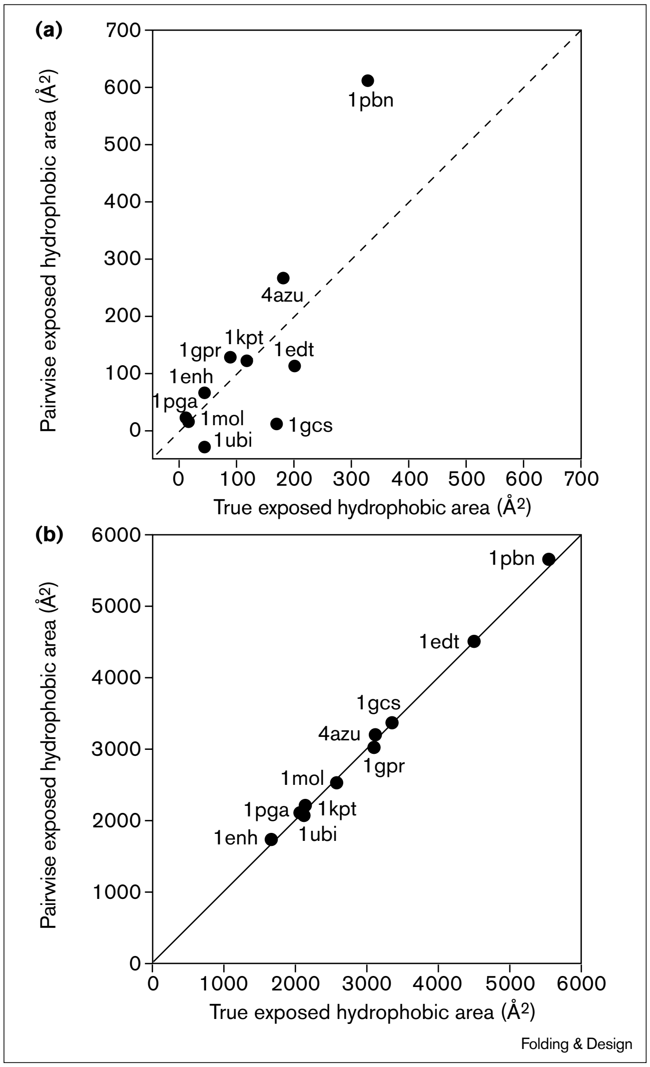
A comparison for ten proteins of the true buried surface area and the pairwise buried surface area calculated using Equation 1. **(a)** Core residues using $s = 0.42$. **(b)** Non-core residues using $s = 0.79$. In each case the correlation coefficient $R^2 = 1.00$. The lines of best fit have slope = 0.99 and slope = 1.00 for (a) and (b), respectively, and differences between calculated and true buried areas are ≤ 2.5%.

and Equation 3 with:

$$A^{\text{pairwise}}_{\text{exposed}} = \sum_{i} A_{i_r t} - \sum_{i<j} \left(s_i \mathrm{A}_{i_r t} + s_j \mathrm{A}_{j_s t} - s_{ij} \mathrm{A}_{i_r j_s t} \right) \qquad (7)$$

where s_i and s_j are the values of s appropriate for residues i and j, respectively, and s_{ij} takes on an intermediate value. Using subsets from the whole of 1pga, the optimal value of s_{ij} was found to be 0.74. This value was then shown to be appropriate for other test proteins (Figure 5). The correlation shown in Figure 5b represents a substantial improvement over that shown in Figure 1b and demonstrates the utility of our approach.

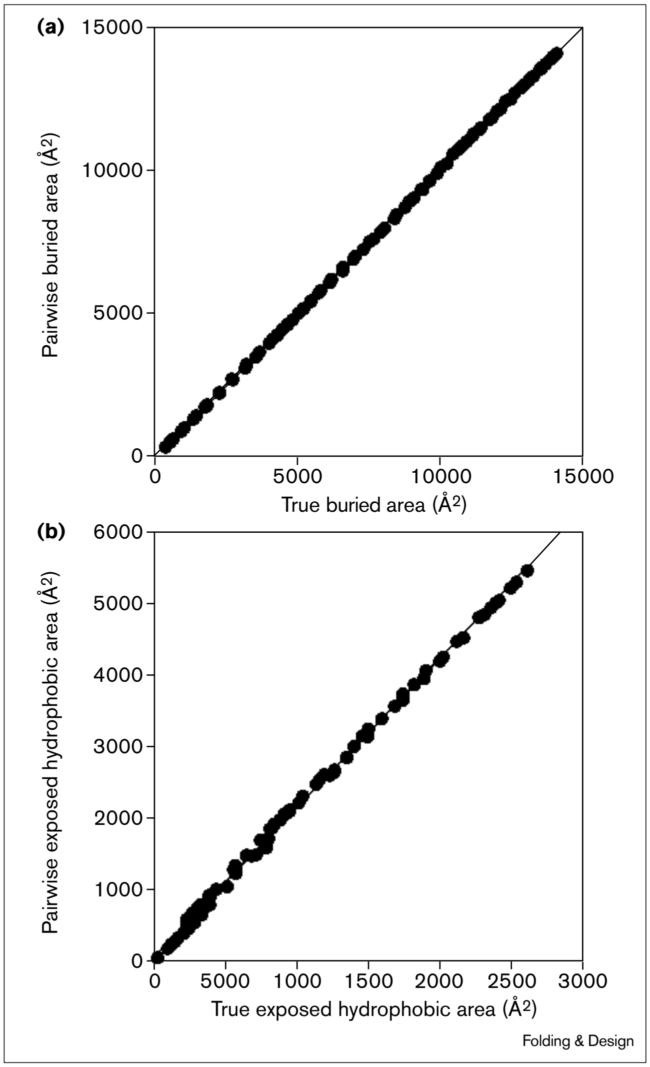
Figure 4



A comparison for ten proteins of the true exposed hydrophobic surface area and the pairwise exposed hydrophobic surface area calculated using Equation 3. **(a)** Core residues using $s = 0.42$, with $R^2 = 0.69$ (a dashed line of slope = 1 is shown for reference). The maximum difference between calculated and true exposed hydrophobic areas is 170%. **(b)** Non-core residues using $s = 0.79$. The line of best fit has slope = 1.02 and a correlation coefficient $R^2 = 1.00$. The maximum difference between calculated and true exposed hydrophobic areas is 5%.

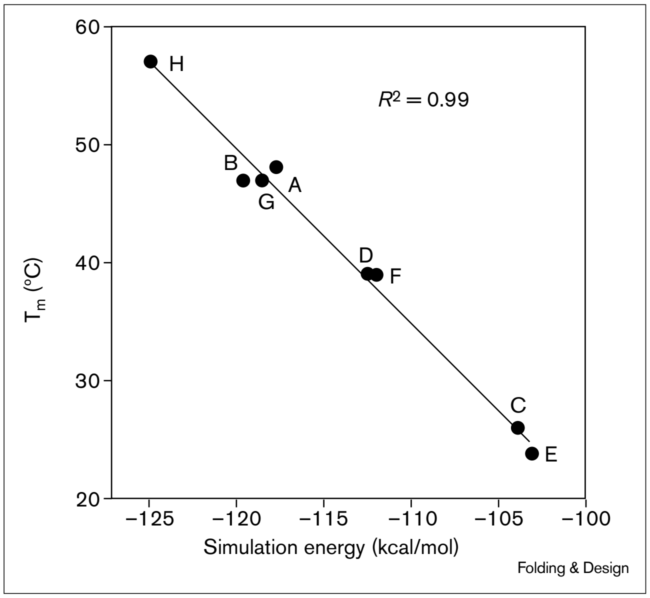
In previous work, we examined the ability of a simple van der Waals potential energy function to predict the thermal stability of a series of coiled coils [8]. We noted a significant improvement in the correlation between calculated stabilities and experimentally measured stabilities when a hydrophobic burial benefit of $\sigma_{\text{np}} A^{\text{np}}_{\text{buried}}$ was included in the calculated energies, where σ_{np} is a hydrophobic solvation parameter whose value was determined to be 23 cal/mol Å^2 and $A^{\text{np}}_{\text{buried}}$ was the calculated buried hydrophobic area. The correlation between calculated energies and experimental

Figure 5



A comparison of true surface area and that calculated with Equations 6 and 7 for subsets of 1mol using $s_{ij} = 0.74$. The subsets are the same as in Figure 1. **(a)** Buried area. The line of best fit has slope = 1.01, a correlation coefficient $R^2 = 1.00$, and a maximum difference between calculated and true buried area of 2%. **(b)** Exposed hydrophobic area. The line of best fit has slope = 1.05 and a correlation coefficient $R^2 = 1.00$, with differences between calculated and true areas of 0–30% for small areas, converging to 5% for areas > 1000 Å^2. These differences represent approximately an order of magnitude improvement over Figure 1.

Figure 6



Correlation between calculated and measured stability for designed coiled coils using buried surface areas calculated using Equation 6 (compare to Figure 5b of [8]). Solvation parameter values are 26 cal/mol Å^2 favoring hydrophobic burial and 100 cal/mol Å^2 opposing polar burial. The labels A–H correspond to proteins PDA-3A–PDA-3H of [8].

melting temperatures was further improved by penalizing polar surface area burial by $\sigma_p A^p_{buried}$, where σ_p is a polar solvation parameter and A^p_{buried} was the calculated buried polar area. The best values of σ_{np} and σ_p were found to be 16 cal/mol Å^2 and 86 cal/mol Å^2, respectively, when both solvation terms were used together. In order to benefit from the more accurate pairwise surface area method in protein design studies, it is necessary to update the values of σ_{np} and σ_p. We use Equation 6 and the values of s described above. Residue 26 of the coiled coil used in the previous study was the only residue determined to be in the core. When only the hydrophobic burial benefit was considered, the best fit value of σ_{np} was determined to be 48 cal/mol Å^2. When both the hydrophobic burial benefit and the polar burial penalty were considered together, the best fit values of σ_{np} and σ_p were determined to be 26 cal/mol Å^2 and 100 cal/mol Å^2, respectively (Figure 6).

By examining a test set of proteins of various sizes, we have determined that the true Lee and Richards' buried and exposed surface areas can be approximated well as a superposition of two-body interactions using Equations 6 and 7, with values for the parameter s that depend on the structural context of each residue. For core residues $s = 0.42$, for non-core positions $s = 0.79$, and for interactions between core and non-core positions $s_{ij} = 0.74$.

Materials and methods

We considered ten representative proteins whose Brookhaven PDB codes [11] are listed in Table 1. The program BIOGRAF (Molecular Simulations Incorporated, San Diego, CA) was used to generate explicit hydrogens on the structures, which were then conjugate gradient minimized for 50 steps using the DREIDING force field [12].

We classified residues as core or non-core using an algorithm that considered the direction of each sidechain's Cα–Cβ vector relative to a surface computed using only the template Cα atoms with a carbon radius of 1.95 Å, a probe radius of 8 Å and no add-on radius. A residue was classified as a core position if both the distance from its Cα atom (along its Cα–Cβ vector) to the surface was > 5.0 Å and the distance from its Cβ atom to the nearest point on the surface was > 2.0 Å [13]. The advantage of such an algorithm is that a knowledge of the amino acid type actually present at each residue position is not necessary.

Surface areas were calculated using the Connolly algorithm with a dot density of 10 $Å^{-2}$ [14], using a probe radius of zero and an add-on radius of 1.4 Å [2] and atomic radii from the DREIDING forcefield [12]. Atoms that contribute to the hydrophobic surface area are carbon, sulfur, and hydrogen atoms attached to carbon and sulfur.

Energy calculations and parameter optimizations for the coiled-coil system were performed as previously described [8].

Acknowledgements

We thank D.B. Gordon and B.I. Dahiyat for helpful discussions and assistance. This work was supported by the Rita Allen Foundation, the David and Lucile Packard Foundation and the Searle Scholars Program/The Chicago Community Trust. A.G.S. is partially supported by the Eleanor Sophia Wood Travelling Scholarship from the University of Sydney, Australia.

References

1. Desmet, J., De Maeyer, M., Hazes, B. & Lasters, I. (1992). The dead-end elimination theorem and its use in protein sidechain positioning. *Nature* **356**, 539-542.
2. Lee, B. & Richards, F.M. (1971). The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **55**, 379-400.
3. Wodak, S.J. & Janin, J. (1980). Analytical approximation to the accessible surface area of proteins. *Proc. Natl Acad. Sci. USA* **77**, 1736-1740.
4. Kurochkina, N. & Lee, B. (1995). Hydrophobic potential by pairwise surface area sum. *Protein Eng.* **8**, 437-442.
5. Hodes, Z.I., Némethy, G. & Scheraga, H.A. (1979). Model for the conformational analysis of hydrated peptides. Effect of hydration on the conformational stability of the terminally blocked residues of the 20 naturally occurring amino acids. *Biopolymers* **18**, 1565-1610.
6. Augspurger, J.D. & Scheraga, H.A. (1996). An efficient, differentiable hydration potential for peptides and proteins. *J. Comp. Chem.* **17**, 1549-1558.
7. Connolly, M.L. (1996). Molecular surfaces: a review. *Network Sci.* **2**, http://www.netsci.org/Issues/1996/Apr/articles.html.
8. Dahiyat, B.I. & Mayo, S.L. (1996). Protein design automation. *Protein Sci.* **5**, 895-903.
9. Creighton, T.E. (1993). *Proteins: Structure and Molecular Properties.* W.H. Freeman and Co., New York.
10. Richards, F.M. & Lim, W.A. (1994). An analysis of packing in the protein folding problem. *Quart. Rev. Biophys.* **26**, 423-498.
11. Bernstein, F.C., *et al.*, & Tasumi, M. (1977). The protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535-542.
12. Mayo, S.L., Olafson, B.D. & Goddard, W.A., III. (1990). Dreiding - a generic force-field for molecular simulations. *J. Phys. Chem.* **94**, 8897-8909.
13. Dahiyat, B.I. & Mayo, S.L. (1997). De novo protein design: fully automated sequence selection. *Science* **278**, 82-87.
14. Connolly, M.L. (1983). Solvent accessible surfaces of proteins and nucleic acids. *Science* **221**, 709-713.

Because *Folding & Design* operates a 'Continuous Publication System' for Research Papers, this paper has been published on the internet before being printed. The paper can be accessed from http://biomednet.com/cbiology/fad – for further information, see the explanation on the contents pages.