PROTEINS: Structure, Function, and Bioinformatics 57:565–576 (2004)

Fast Accurate Evaluation of Protein Solvent Exposure

Naigong Zhang,[1] Chen Zeng,[1*] and Ned S. Wingreen[2]
[1]*Department of Physics, George Washington University, Washington, District of Columbia*
[2]*NEC Laboratories America, Inc., 4 Independence Way, Princeton, New Jersey*

ABSTRACT **Protein solvation energies are often taken to be proportional to solvent-accessible surface areas. Computation of these areas is numerically demanding and may become a bottleneck for folding and design applications. Fast graph-based methods, such as dead-end elimination (DEE), become possible if all energies, including solvation energies, are expressed as single-residue and pair-residue terms. To this end, Street and Mayo originated a pair-residue approximation for solvent-accessible surface areas (Street AG, Mayo SL. Pairwise calculation of protein solvent accessible surface areas. Fold Des 1998;3:253–258). The dominant source of error in this method is the overlapping burial of side-chain surfaces in the protein core. Here we report a new pair-residue approximation, which greatly reduces this overlap error by the use of optimized generic side-chains. We have tested the generic-side-chain method for the ten proteins studied by Street and Mayo and for 377 single-domain proteins from the CATH database (Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH-A hierarchic classification of protein domain structures. Structure 1997;5:1093–1108). With little additional cost in computation, the new method consistently reduces error for total areas and residue-by-residue areas by more than a factor of two. For example, the residue-by-residue error (for buried area) is reduced from 7.42 Å^2 to 3.70 Å^2. This difference translates into a solvation energy difference of ~0.2 kcal/mol per residue, amounting to a reduction in root-mean-square energy error of 2 kcal/mol for a 100 residue chain, a potentially critical difference for both protein folding and design applications. Proteins 2004;57:565–576.**
© 2004 Wiley-Liss, Inc.

Key words: solvent-accessible surface area; pairwise surface area; dead-end elimination; generic side-chains; protein design

INTRODUCTION

Solvation energies play a major role in protein collapse. Moreover, the burial of hydrophobic residues away from the aqueous solvent is a major determinant of specific folding.[2] Accurate solvation energies are needed both for folding simulations and for protein design. In practice, accurate solvation energies depend on accurate calculation of solvent exposure.[3] However, exact evaluation of exposure is numerically demanding and may become a computational bottleneck. Fast optimization of protein conformations and sequences is possible if all energies, including solvation energies, can be expressed as single-residue and pair-residue terms.[4]

Such a pairwise separation naturally leads to a graphic representation, with the protein residues and side-chain rotamers as the vertices of the graph and the pairwise energies as the weights on the edges of the graph. Fast search algorithms based on combinatorial optimization can then be applied to the graph.[5] One important application is *de novo* protein design based on the dead-end-elimination (DEE) algorithm.[4] Very recent DEE studies of sequence redesign on some given protein scaffolds have led to novel enzymes[6] as well as biosensors.[7] A requirement for successful application of the DEE algorithm is the existence of an accurate pairwise approximation for the total protein exposure and the separate burial of polar and nonpolar atoms. A second application of fast graph-based search algorithms is the detection of protein domains by finding the cut with minimum interaction energy out of the exponentially many possible cuts that break the protein into two or more units.[8–10] Here, an accurate pairwise approximation for the exposure and burial of each individual residue within a protein is required.

The difficulty of obtaining an accurate pairwise approximation to surface areas lies in treating multiple-atom, multiple-residue buried areas. For two spheres in contact, there is a simple analytical formula to calculate the buried area (and therefore the exposed area) using the two radii and the separation between the spheres.[10] For proteins, this formula was initially applied to each pair of atoms, and the total exposed and buried areas were estimated via a statistical combination.[10] Street and Mayo[1] significantly improved on this statistical combination of pair-atom areas by calculating pair-residue areas. In the presence of the backbone, two residues were placed at positions i and j, first separately and then together, and the areas buried by the two side-chains were obtained. Because buried areas often overlap, a straightforward addition of pair-residue areas overestimates the total buried area. Consequently, Street and Mayo scaled down the pairwise buried area by

Correspondence to: C. Zeng, Department of Physics, George Washington University, Washington, DC 20052, USA. E-mail: chenz@gwu.edu.

Received 22 September 2003; Revised 23 February 2004; Accepted 26 March 2004

Published online 22 July 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20191

© 2004 WILEY-LISS, INC.

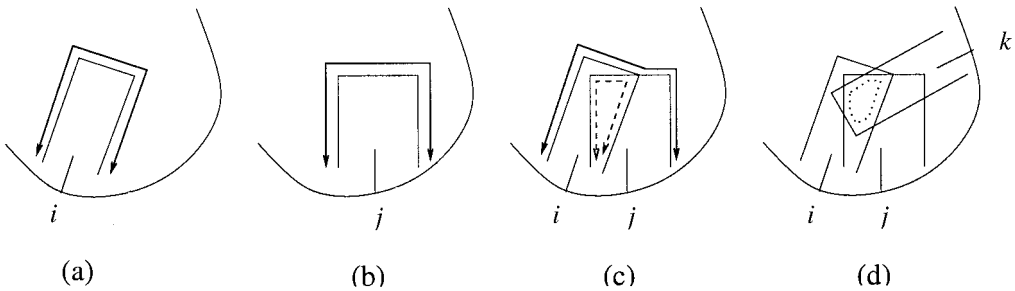

Fig. 1. Schematic reprensentation of the Street and Mayo method for pairwise calculation of the exposed surface area [eq. (2)]. (a) $A_{i,\mathrm{bb}}^{\mathrm{exposed}}$ (indicated by the bold line with arrows) the exposed area of the side-chain at *i* in the presence of the backbone. (b) $A_{j,\mathrm{bb}}^{\mathrm{exposed}}$. (c) The total exposed area of the side-chains *i* and *j* in the presence of the backbone ($A_{ij,\mathrm{bb}}^{\mathrm{exposed}}$ bold line with arrows). The dashed line with arrows shows $A_{i,\mathrm{bb}}^{\mathrm{exposed}} + A_{j,\mathrm{bb}}^{\mathrm{exposed}} - A_{ij,\mathrm{bb}}^{\mathrm{exposed}}$, which is the area buried by the side-chains at *i* and *j*. This buried area is subtracted from $A_{i,\mathrm{bb}}^{\mathrm{exposed}} + A_{j,\mathrm{bb}}^{\mathrm{exposed}}$ in eq. (2) with a scaling factor *s*. (d) The multiply-buried area (dotted line), which would be overcount if $s = 1.0$ was used in eq. (2).

an adjustable factor. They went further and used three pairwise scaling factors for residues classified into core and non-core categories[11]: 0.42 for the interaction of core–core residues, 0.79 for non-core–non-core residues and 0.74 for core–non-core residues.[1] The scaling factor for two core residues was smaller than the scaling factors with non-core residues involved because the overlapping burial effect is stronger in the core. Despite the use of specialized scaling factors, the accuracy of their method was still limited by the effect of overlapping burial of core residues.

In this article, we introduce a method that incorporates the many-residue burial effect directly in the single-residue and pair-residue areas. Instead of calculating these areas in the presence of the backbone alone, we calculate the areas in the presence of the backbone and generic side-chains. These generic side-chains consist of one or a number of spheres and are optimized to approximate the presence of real side-chains. The essential advantage of our method is that much of the overlapping burial effect is automatically accounted for by the generic side-chains. As a result, errors in the area calculations are dramatically reduced, and there is no need to optimize scaling factors; they can be set to one.

Street and Mayo focused on the total buried and exposed areas of the entire protein. We present a surface-area formula for each individual residue. By comparing the pairwise and exact areas residue by residue, the parameters of the generic side-chains were systematically optimized. We tested our method both for the ten-protein set used by Street and Mayo and for the 377 single-segment representative protein domains in the CATH classification (T, or topology, level).* Compared to Street and Mayo's results, our generic-side-chain method greatly improves accuracy for both the total area and the individual residue areas, thus allowing a highly accurate graphic representation of protein solvation energy for use in fast optimization methods.

METHODS

Street and Mayo's Method

Street and Mayo[1] used the following two pairwise formulas to calculate the total buried and exposed surface areas, respectively, of a protein (see Fig. 1 for a schematic representation):

$$A_{\mathrm{pairwise}}^{\mathrm{buried}} = \sum_i (A_{i,\mathrm{GXG}}^{\mathrm{exposed}} - A_{i,\mathrm{bb}}^{\mathrm{exposed}}) + s\sum_{i<j}(A_{i,\mathrm{bb}}^{\mathrm{exposed}} + A_{j,\mathrm{bb}}^{\mathrm{exposed}} - A_{ij,\mathrm{bb}}^{\mathrm{exposed}}) \quad (1)$$

$$A_{\mathrm{pairwise}}^{\mathrm{exposed}} = \sum_i A_{i,\mathrm{bb}}^{\mathrm{exposed}} - s\sum_{i<j}(A_{i,\mathrm{bb}}^{\mathrm{exposed}} + A_{j,\mathrm{bb}}^{\mathrm{exposed}} - A_{ij,\mathrm{bb}}^{\mathrm{exposed}}) \quad (2)$$

Here $A_{i,\mathrm{GXG}}^{\mathrm{exposed}}$ is the exposed surface area of the *i*-th residue in the presence of the local tripeptide. G-G stands for the backbone units of the previous and the following residues and X stands for the *i*-th residue.† $A_{i,\mathrm{bb}}^{\mathrm{exposed}}$ is the exposed surface area of the *i*-th residue in the presence of the entire protein backbone, and $A_{ij,\mathrm{bb}}^{\mathrm{exposed}}$ is the combined exposed surface area of the *i*-th and *j*-th residues in the presence of the backbone. Then $A_{i,\mathrm{GXG}}^{\mathrm{exposed}} - A_{i,\mathrm{bb}}^{\mathrm{exposed}}$ is the area of the *i*-th residue buried by parts of the backbone other than the local tripeptide structure. It is common in calculating protein surface areas to exclude the area buried by the local tripeptide because the burial due to local covalent bonds does not change during folding (see ref. 13). $A_{i,\mathrm{bb}}^{\mathrm{exposed}} + A_{j,\mathrm{bb}}^{\mathrm{exposed}} - A_{ij,\mathrm{bb}}^{\mathrm{exposed}}$ is the area buried by the proximity of the two residues at *i* and *j*, excluding that buried by the backbone [see Fig. 1(c)]. The factor *s*, positive and less than one, scales down the pair-residue buried area to account for the effect of overlapping burial. In Figure 1(d), we schematically show the overlap of three

*We used the 377 CATH classification (Release 2.4) T-level (topology) single-segment representative domains. In this level there is a total of 775 representative domains, of which 663 are single-segment domains; of these 663, we omitted (1) 16 for which the PDB files could not be found in the Protein Data Bank, (2) 6 in which ACE (acetyl group) or PCA (pyroglutamic acid) are included in the domain, (3) 112 for which some atomic coordinates could not be found in the PDB files and (4) 152 for which the domain definition is not consistent.

†For the local tripeptide, Street and Mayo used $[C_\alpha,C,O]_{i-1}$, . . . ,$[N,C_\alpha]_{i+1}$.

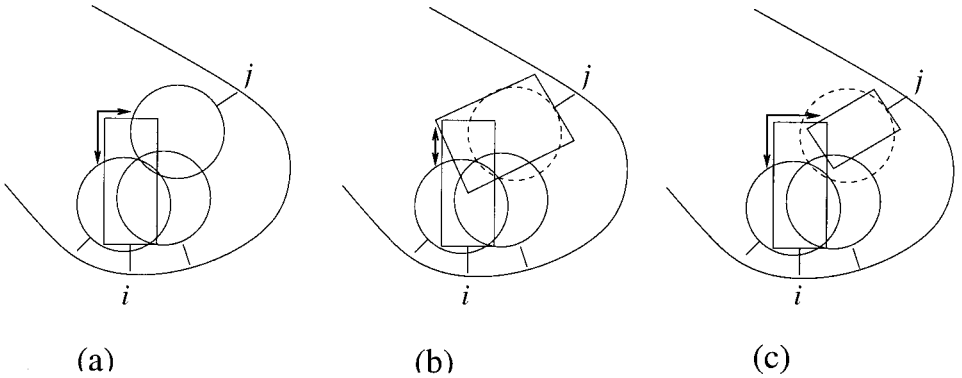

Fig. 2. Schematic representation of our generic-side-chain method for pairwise calculation of exposed surface areas. (a) $A_{i,\mathrm{gs}}^{\mathrm{exposed}}$ (indicated by the bold line with arrows) is the exposed area of the side-chain at *i* in the presence of the backbone and generic side-chains at other positions (generic side-chains are represented by circles); (b,c) $A_{i;j,\mathrm{gs}}^{\mathrm{exposed}}$ (solid line with arrows) is the exposed area of the side-chain at *i* in the presence of the real side-chain *j*, the backbone, and generic side-chains at all other positions. In (b), the real side-chain at *j* is large and covers more area than the generic side-chain at *j*; hence $A_{i;j,\mathrm{gs}}^{\mathrm{exposed}} - A_{i,\mathrm{gs}}^{\mathrm{exposed}}$ is negative. In (c), the real side- chain at *j* is small and covers less area than the generic side-chain; hence $A_{i;j,\mathrm{gs}}^{\mathrm{exposed}}$-$A_{i,\mathrm{gs}}^{\mathrm{exposed}}$ is positive. In eq. (3), $A_{i,\mathrm{gs}}^{\mathrm{exposed}}$ is correct by $s\sum_{i,j\neq i}(A_{i;j,\mathrm{gs}}^{\mathrm{exposed}} - A_{i,\mathrm{gs}}^{\mathrm{exposed}})$. We find that $s = 1.0$ is optimal; *i.e.,* there is no need for a scaling factor *s* because the overlapping burial effect is automatically accounted for by the generic side-chains.

side-chains at *i, j*, and *k*. The multiply-buried area (dotted line) would be overcounted by eqs. (1) and (2) if a unit scaling factor of $s = 1.0$ were used.

In addition, in eqs. (1) and (2), all areas can be separated into contributions from the polar (p) atoms, nitrogen and oxygen, and the nonpolar (n) atoms, carbon and sulfur. The result is six total areas: $A_{\mathrm{pairwise}}^{\mathrm{buried}}$, $A_{\mathrm{pairwise}}^{\mathrm{exposed}}$, $A_{\mathrm{pairwise}}^{(\mathrm{p})\mathrm{buried}}$, $A_{\mathrm{pairwise}}^{(\mathrm{p})\mathrm{exposed}}$, $A_{\mathrm{pairwise}}^{(\mathrm{n})\mathrm{buried}}$, $A_{\mathrm{pairwise}}^{(\mathrm{n})\mathrm{exposed}}$, which can be compared to the six corresponding exact values.

Our Generic Side Chain Method

We have calculated single-residue and pair-residue areas in the presence of generic side-chains in addition to the backbone. We use $A_{i,\mathrm{gs}}^{\mathrm{exposed}}$ to denote the exposed surface area of the *i*-th residue in the presence of the backbone with generic side-chains at all positions other than *i*. Figure 2(a) shows a schematic representation of $A_{i,\mathrm{gs}}^{\mathrm{exposed}}$. To obtain a pairwise formula for the surface area of residue *i* [instead of the entire protein as in eqs. (1) and (2)], we define $A_{i;j,\mathrm{gs}}^{\mathrm{exposed}}$ differently from Street and Mayo's $A_{ij,\mathrm{bb}}^{\mathrm{exposed}}$. Specifically, we define $A_{i;j,\mathrm{gs}}^{\mathrm{exposed}}$ as the exposed surface area of the *i*-th residue itself in the presence of the real side-chain at *j,* the backbone and generic side-chains at all positions other than *i* and *j* [see Fig. 2(b,c)]. (Note that Street and Mayo's $A_{ij,\mathrm{bb}}^{\mathrm{exposed}}$ is the *sum* of the exposed areas of the *i*-th and *j*-th residues.) Then, $A_{i;j,\mathrm{gs}}^{\mathrm{exposed}} - A_{i,\mathrm{gs}}^{\mathrm{exposed}}$ is the correction to the exposed area at *i* due to the presence of the real side-chain at *j* in place of a generic one. When the *j*-th residue is large and covers more area than a generic side-chain at *j,* $A_{i;j,\mathrm{gs}}^{\mathrm{exposed}} - A_{i,\mathrm{gs}}^{\mathrm{exposed}}$ is negative, to reduce the overestimate of the exposed area of residue *i* [see Fig. 2(b)]. On the other hand, when the *j*-th residue is small and covers less area than a generic side-chain at *j,* $A_{i;j,\mathrm{gs}}^{\mathrm{exposed}} - A_{i,\mathrm{gs}}^{\mathrm{exposed}}$ is positive, to increase the underestimate of the exposed area of residue *i* [see Fig. 2(c)]. We therefore obtain, in place of eq. (2), a new exposed area formula for the *i*-th residue:

$$A_{i,\mathrm{pairwise}}^{\mathrm{exposed}} = A_{i,\mathrm{gs}}^{\mathrm{exposed}} + s\sum_{j\neq i}(A_{i;j,\mathrm{gs}}^{\mathrm{exposed}} - A_{i,\mathrm{gs}}^{\mathrm{exposed}}) \qquad (3)$$

where, as explained above, the first term is the single-residue approximation using generic side-chains at all other positions, and the second term sums up all corrections using real side-chains at second positions *j* and generic side-chains at all positions other than *i* and *j*. The scaling factor *s* accounts for any residual effects of overlapping burial; in practice, we find $s = 1.0$ is optimal; *i.e.* there is no systematic overestimate or underestimate of exposed areas. The total exposed area of the protein $A_{\mathrm{pairwise}}^{\mathrm{exposed}}$ is obtained by summing eq. (3) over residues *i*.

Similarly, it is straightforward to write a buried area formula for the *i*-th residue in place of eq. (1):

$$A_{i,\mathrm{pairwise}}^{\mathrm{buried}} = (A_{i,\mathrm{GXG}}^{\mathrm{exposed}} - A_{i,\mathrm{gs}}^{\mathrm{exposed}}) + s\sum_{i,j\neq i}(A_{i,\mathrm{gs}}^{\mathrm{exposed}} - A_{i;j,\mathrm{gs}}^{\mathrm{exposed}}) \qquad (4)$$

where $A_{i,\mathrm{GXG}}^{\mathrm{exposed}} - A_{i,\mathrm{gs}}^{\mathrm{exposed}}$ is the area buried by the backbone and the generic side-chains, but not by the local tripeptide GXG, and $A_{i,\mathrm{gs}}^{\mathrm{exposed}} - A_{i;j,\mathrm{gs}}^{\mathrm{exposed}}$ is the correction to the area buried by the *j*-th residue due to the presence of the real side-chain at *j* in place of a generic one.[††] The scaling factor *s* accounts for the residual effects of overlapping burial for the total buried area; again, we find that a unit scaling factor $s = 1.0$ is optimal. It is convenient to manipulate eq. (4) for the buried area into a form similar to eq. 3 for the exposed area. We define A_i^{total} as the simple sum of the surface areas of all atoms in residue *i* without regard to burial, and $A_{i,\mathrm{GXG}}^{\mathrm{buried}}$ as the area of the *i*-th residue buried by the local tripeptide (GXG). From this definition, we have

[††]For the local tripeptide, we have used $[\mathrm{N},\mathrm{C}_\alpha,\mathrm{C},\mathrm{O}]_{i-1},\ldots,[\mathrm{N},\mathrm{C}_\alpha,\mathrm{C},\mathrm{O}]_{i+1}$ (see ref. 13).

$$A_{i,\mathrm{GXG}}^{\mathrm{exposed}} = A_i^{\mathrm{total}} - A_{i,\mathrm{GXG}}^{\mathrm{buried}}, \quad (5)$$

and we define buried areas, excluding those buried by GXG, as

$$A_{i,\mathrm{gs}}^{\mathrm{buried}} = A_i^{\mathrm{total}} - A_{i,\mathrm{GXG}}^{\mathrm{buried}} - A_{i,\mathrm{gs}}^{\mathrm{exposed}}, \quad (6)$$

$$A_{i:j,\mathrm{gs}}^{\mathrm{buried}} = A_i^{\mathrm{total}} - A_{i,\mathrm{GXG}}^{\mathrm{buried}} - A_{i:j,\mathrm{gs}}^{\mathrm{exposed}} \quad (7)$$

Using eqs. (5), (6) and (7), it is easily verified that our expression for buried area eq. (4) can be written as

$$A_{i,\mathrm{pairwise}}^{\mathrm{buried}} = A_{i,\mathrm{gs}}^{\mathrm{buried}} + s \sum_{j \neq i} (A_{i:j,\mathrm{gs}}^{\mathrm{buried}} - A_{i,\mathrm{gs}}^{\mathrm{buried}}) \quad (8)$$

which is of the same form as the exposed area formula eq. (3). The total buried area of the protein $A_{\mathrm{pairwise}}^{\mathrm{buried}}$ is obtained by summing eq. (8) over residues i. We use eqs. (3) and (8) extensively in our calculations below.

Exact Surface Area Calculations

To calculate the exact surface areas for the i-th residue, we evaluated the buried area of each atom $m^{(i)}$ in this residue, including backbone atoms (N, C_α, C, and O) and side-chain atoms. Hydrogen atoms were not considered in any of our surface calculations. The van der Waals radii of the atoms are 1.5 Å for nitrogen, 1.4 Å for oxygen, 1.85 Å for sulfur, 1.5 Å for carbonyl carbons and 2.0 Å for other carbons. For each atom, 256 evenly distributed dots were placed on the spherical surface.[14] The radius of the sphere is the van der Waals radius plus a water add-on radius of 1.4 Å, representing the radius of a water molecule. For each atom $m^{(i)}$ we find which of the 256 dots are buried by any other atom p of the protein (where the water radius is also added to the radius of each atom p). If p is one of the atoms in the surrounding tripeptide GXG (i.e. p, is one of the N, C_α, C, O backbone atoms of the previous or the next residue or one of the other atoms in residue i), then the dots on the surface of $m^{(i)}$ covered by p are marked as buried by GXG; if not, they are marked as regularly buried. The marking of buried points was completed using the mask method of ref. 15. Thus, after considering all other atoms of the protein, the 256 dots on the surface of atom $m^{(i)}$ were divided into three categories: unmarked, marked as buried by GXG and marked as regularly buried. Counting the dots in each of these three categories gives the three areas, $a_{m^{(i)},\mathrm{exact}}^{\mathrm{exposed}}$, $a_{m^{(i)},\mathrm{GXG}}^{\mathrm{buried}}$ and $a_{m^{(i)},\mathrm{exact}}^{\mathrm{buried}}$ respectively. By summing these three quantities over all atoms $m^{(i)}$ in residue i, we obtain $A_{i,\mathrm{exact}}^{\mathrm{exposed}}$, $A_{i,\mathrm{GXG}}^{\mathrm{buried}}$ and $A_{i,\mathrm{exact}}^{\mathrm{buried}}$. If we sum over only the polar atoms in the residue (nitrogen and oxygen), we obtain $A_{i,\mathrm{exact}}^{(\mathrm{p})\mathrm{exposed}}$, $A_{i,\mathrm{GXG}}^{(\mathrm{p})\mathrm{buried}}$ and $A_{i,\mathrm{exact}}^{(\mathrm{p})\mathrm{buried}}$; if we sum over only the nonpolar atoms (carbon and sulfur), we obtain $A_{i,\mathrm{exact}}^{(\mathrm{n})\mathrm{exposed}}$, $A_{i,\mathrm{GXG}}^{(\mathrm{n})\mathrm{buried}}$ and $A_{i,\mathrm{exact}}^{(\mathrm{n})\mathrm{buried}}$.

Specification of Generic Side-chains

To use our pairwise method to calculate surface areas, we first need to specify generic side-chains. We consider side-chains described by three parameters: n, the number of spheres; d, the distance of the first sphere from C_α in the direction from C_α to C_β and also the separation between the additional spheres; r, the radius of each sphere. Given the atomic coordinates of a protein backbone, virtual C_β atoms are placed at standard glycine-C_β positions.[§] The generic side-chain is composed of n spheres of radius r placed at d, $2d$,. . .,nd from C_α in the direction from C_α to C_β. The optimal values of n, d, and r are reported in Results.

Pairwise Surface Area Calculations

We calculated the single-residue and pair-residue quantities needed for the pairwise method as follows. As in the exact surface-area calculation, for each atom $m^{(i)}$ in residue i, we determined which of the 256 dots are buried by all other atoms in residue i, the entire backbone and the generic side-chains at positions other than i. Note that the water add-on radius of 1.4 Å was also added to the generic side-chain radius r. Again, we obtained three atomic areas, $a_{m^{(i)},\mathrm{gs}}^{\mathrm{exposed}}$, $a_{m^{(i)},\mathrm{GXG}}^{\mathrm{buried}}$ and $a_{m^{(i)},\mathrm{gs}}^{\mathrm{buried}}$, which, when summed over all atoms in residue i, gave the single-residue quantities $A_{i,\mathrm{gs}}^{\mathrm{exposed}}$, $A_{i,\mathrm{GXG}}^{\mathrm{buried}}$ and $A_{i,\mathrm{gs}}^{\mathrm{buried}}$. Note that the area buried by the local tripeptide, $A_{i,\mathrm{GXG}}^{\mathrm{buried}}$, is an exact quantity and does not depend on the generic side-chains. If we sum the other two areas over only the polar atoms in the residue, we obtain $A_{i,\mathrm{gs}}^{(\mathrm{p})\mathrm{exposed}}$ and $A_{i,\mathrm{gs}}^{(\mathrm{p})\mathrm{buried}}$; and if we sum over only the nonpolar atoms, we obtain $A_{i,\mathrm{gs}}^{(\mathrm{n})\mathrm{exposed}}$ and $A_{i,\mathrm{gs}}^{(\mathrm{n})\mathrm{buried}}$. Next, we considered real residues at both i and j. For each atom $m^{(i)}$ in i, we individually considered all other atoms in residue i, all atoms in residue j, the entire backbone, and the generic side-chains at positions other than i and j. We thus obtain $a_{m^{(i)}:j,\mathrm{gs}}^{\mathrm{exposed}}$ and $a_{m^{(i)}:j,\mathrm{gs}}^{\mathrm{buried}}$, which when summed over all atoms in residue i give the pair-residue quantities $A_{i:j,\mathrm{gs}}^{\mathrm{exposed}}$ and $A_{i:j,\mathrm{gs}}^{\mathrm{buried}}$. If we sum over only the polar atoms in the residue, we obtain $A_{i:j,\mathrm{gs}}^{(\mathrm{p})\mathrm{exposed}}$ and $A_{i:j,\mathrm{gs}}^{(\mathrm{p})\mathrm{buried}}$; and if we sum over only the nonpolar atoms, we obtain $A_{i:j,\mathrm{gs}}^{(\mathrm{n})\mathrm{exposed}}$ and $A_{i:j,\mathrm{gs}}^{(\mathrm{n})\mathrm{buried}}$. These single-residue and pair-residue quantities can then be combined in eqs. (3) and (8) to give $A_{i,\mathrm{pairwise}}^{\mathrm{exposed}}$, $A_{i,\mathrm{pairwise}}^{\mathrm{buried}}$, $A_{i,\mathrm{pairwise}}^{(\mathrm{p})\mathrm{exposed}}$, $A_{i,\mathrm{pairwise}}^{(\mathrm{p})\mathrm{buried}}$, $A_{i,\mathrm{pairwise}}^{(\mathrm{n})\mathrm{exposed}}$ and $A_{i,\mathrm{pairwise}}^{(\mathrm{n})\mathrm{buried}}$. When i is summed over all residues, we obtain the six total areas $A_{\mathrm{pairwise}}^{\mathrm{exposed}}$, $A_{\mathrm{pairwise}}^{\mathrm{buried}}$, $A_{\mathrm{pairwise}}^{(\mathrm{p})\mathrm{exposed}}$, $A_{\mathrm{pairwise}}^{(\mathrm{p})\mathrm{buried}}$, $A_{\mathrm{pairwise}}^{(\mathrm{n})\mathrm{exposed}}$ and $A_{\mathrm{pairwise}}^{(\mathrm{n})\mathrm{buried}}$. The choice of scaling factor s is discussed in Results.

If the number of residues in a given protein is N, then for each residue i there are $N - 1$ quantities $A_{i:j,\mathrm{gs}}$ to calculate, corresponding to each $j \neq i$. However, most of the $A_{i:j,\mathrm{gs}}$ are identical to $A_{i,\mathrm{gs}}$ because residue j is not in contact with residue i. In practice, we first calculate the N single-residue quantities $A_{i,\mathrm{gs}}$, and then for each residue j, we check all atoms in i to see whether any of them is in contact with either an atom in j or the generic side-chain at j. If not, we know that $A_{i:j,\mathrm{gs}} = A_{i,\mathrm{gs}}$. This simple check saves a great deal of computation time, as, for proteins of several hundred residues, we find that often less than five percent of the $N(N-1)$ residue pairs are in contact.

[§]The two angles N-C_α-C_β and C_β-C_α-C are both taken to be 110° and the C_α-C_β distance is taken to be 1.53 Å. As far as handedness is concerned, the L-form is used, as in a naturally occurring alanine residue; that is, looking down the H-C_α bond from the hydrogen, the C, C_β, and N atoms are in clockwise order.

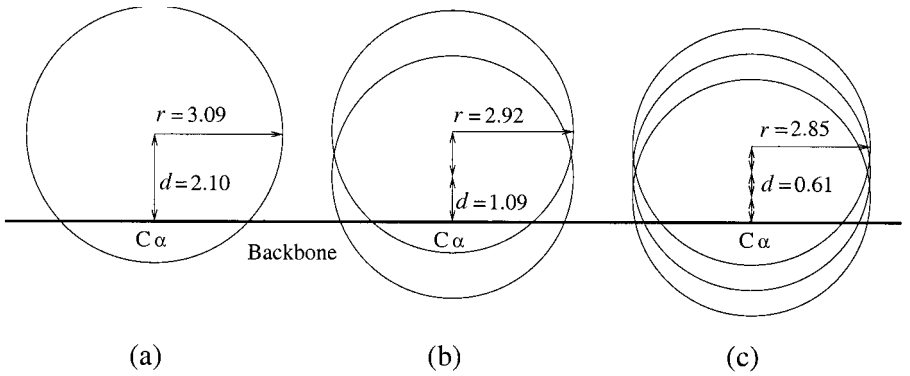

Fig. 3. Generic side-chains drawn to scale with optimized parameters *d* and *r*. *d* is the distance from the C_α atom to the center of the first sphere and the separation between adjacent spheres; *r* is the radius of the generic-side-chain spheres: (a) one sphere ($n = 1$) with $d = 2.10$ Å and $r = 3.09$ Å; (b) two spheres ($n = 2$) with $d = 1.09$ Å and $r = 2.92$ Å; (c) three spheres ($n = 3$) with $d = 0.61$ Å and $r = 2.85$ Å. The thick black line represents the protein backbone.

Measures of Error

To compare the exact and pairwise surface-area results for each protein (or protein domain) of length *N,* we use the following measure of error:

$$\delta A_{\text{protein}}(k) = \frac{1}{N}\sum_{i=1}^{N}[(A_{i,\text{pairwise}}(k) - A_{i,\text{exact}}(k))] \qquad (9)$$

which is the average deviation per residue for the protein *k*. To compare exact and pairwise results on a residue-by-residue basis, we use the following measure of error:

$$\delta A_{\text{residue}}(i) = A_{i,\text{pairwise}} - A_{i,\text{exact}} \qquad (10)$$

which is the deviation for a single residue *i*.

The first formula measures the error for the entire surface of a protein. The accuracy of pairwise results for total area is reflected in the average and standard deviation of $\delta A_{\text{protein}}$ over sets of proteins. The second formula measures the error in the surface area of individual residues. The accuracy of pairwise results for residue-by-residue areas is reflected in the average and standard deviation of $\delta A_{\text{residue}}$ over sets of residues. These two measures can be applied to the six different areas: total buried, polar buried, nonpolar buried, total exposed, polar exposed and nonpolar exposed.

RESULTS

Test Set and Optimization of Generic-Side-chain Parameters

As a test, we first studied the same ten-protein set used by Street and Mayo: (1) 1enh, (2) 1pga, (3) 1ubi, (4) 1mol, (5) 1kpt, (6) 4azu-A, (7) 1gpr, (8) 1gcs, (9) 1edt and (10) 1pbn. In order to compare our results to those of Street and Mayo, we present results for total buried area and nonpolar exposed area. Our generic-side-chain parameters *n, d, r,* and the scaling factor *s* were determined by minimizing

$$\delta A^{(10)} = \sqrt{\frac{1}{10}\sum_{k=1}^{10}[\delta A^{\text{buried}}_{\text{protein}}(k)]^2} \qquad (11)$$

which is the root-mean-square value of the average total protein deviation per residue $\delta A_{\text{protein}}$ for buried area for the ten-protein set. We chose to minimize $\delta A^{\text{buried}}_{\text{protein}}$ rather than $\delta A^{\text{buried}}_{\text{residue}}$ (the residue-by-residue deviation) to make sure that the *total* buried areas are as accurate as possible. Note also that by minimizing $\delta A^{(10)}$, we minimized both the average and the spread of $\delta A^{\text{buried}}_{\text{protein}}$ for the ten-protein set. For $n = 0$, the Street and Mayo method without generic side-chains, we minimized $\delta A^{(10)}$ with respect to one parameter, the scaling factor *s*. For $n = 0$, we also implemented the version of Street and Mayo's method with three different scaling factors: for *i* and *j* both core residues, for *i* and *j* both non-core residues and for *i* or *j* a core residue and the other a non-core residue.[‖] For $n = 1, 2,$ and 3 generic-side-chain spheres, we fixed $s = 1.0$ and minimized $\delta A^{(10)}$ over the spacing *d* and the radius *r* of the spheres. We were able to use a scaling factor $s = 1.0$ because the generic side-chains already account for all systematic effects of overlapping burial.

For $n = 0$, using one parameter *s*, the best result is $\delta A^{(10)}_{\text{best}} = 2.21$ Å^2 for $s = 0.67$. For $n = 0$, using three parameters, the best result is $\delta A^{(10)}_{\text{best}} = 0.70$ Å^2 for $s = 0.55$ (core–core), 0.77 (non-core–non-core) and 0.73 (core–non-core) (the difference from Street and Mayo parameters, 0.42, 0.79, and 0.74, respectively,[1] is due to our use of a slightly modified definition of surface and core residues). For our generic-side-chain method with a fixed $s = 1.0$, we obtained

[‖]To compare to Street and Mayo's three-parameter method, we have used the core–non-core classification scheme of Marshall and Mayo (ref. 11). Namely, at each residue position on the backbone, three spheres with radius 2.0 Å, plus the water add-on radius 1.4 Å, are placed at 1.53 Å, 2 × 1.53 Å and 3 × 1.53 Å from the C_α atom in the C_α-C_β direction. The exposed surface area of each three-sphere side-chain was then calculated in the presence of the backbone and all other three-sphere side-chains. If the surface area of the three-sphere side-chain at a particular site is less than 24 Å^2, then this position is classified as a core position; otherwise it is classified as a non-core position. Note that the Marshall and Mayo method to classify protein positions as core and non-core is a generic side-chain method in the spirit of the work we present here.

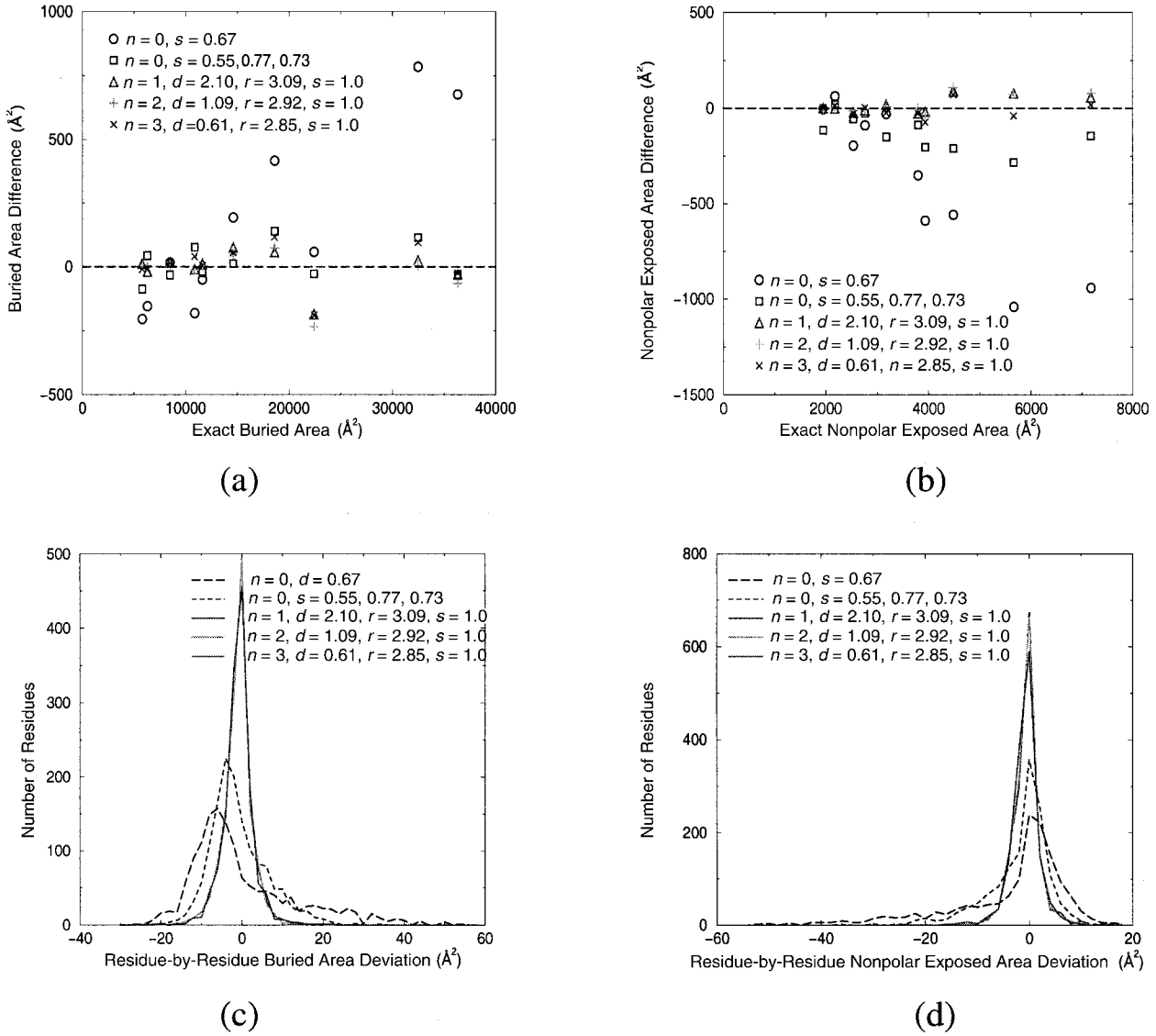

Fig. 4 Surface area results for the ten-protein set: for each protein in the set, (a) the difference between the total pairwise buried area ($A^{\text{buried}}_{\text{pairwise}}$) and the exact buried area ($A^{\text{buried}}_{\text{exact}}$) is plotted against the exact buried area; (b) the difference between the total pairwise nonpolar expose area ($A^{\text{(n)exposed}}_{\text{pairwise}}$) and the exact nonpolar exposed area ($A^{\text{(n)exposed}}_{\text{exact}}$) is plotted against the exact nonpolar exposed area. (c) Distributions of residue-by-residue buried-area deviations $\delta A^{\text{buried}}_{\text{residue}}(i) = A^{\text{buried}}_{i,\text{pairwise}} - A^{\text{buried}}_{i,\text{exact}}$ [eq. (10)] for all residues in the ten-protein set. (d) Distributions of residue-by-residue nonpolar-exposed-area deviations $\delta A^{\text{(n)exposed}}_{\text{residue}}(i) = A^{\text{(n)exposed}}_{i,\text{pairwise}} - A^{\text{(n)exposed}}_{i,\text{exact}}$ for all residues.

$$\delta A^{(10)}_{\text{best}} \begin{array}{ll} = 0.43\,\text{Å}^2 & \text{for } n = 1, \quad d = 2.10\,\text{Å}, \quad r = 3.09\,\text{Å} \\ = 0.47\,\text{Å}^2 & \text{for } n = 2, \quad d = 1.09\,\text{Å}, \quad r = 2.92\,\text{Å} \\ = 0.48\,\text{Å}^2 & \text{for } n = 3, \quad d = 0.61\,\text{Å}, \quad r = 2.85\,\text{Å} \end{array} \tag{12}$$

These are the generic-side-chain parameters used in the following calculations. In Figure 3, we show the three generic side-chains drawn to scale with optimized parameters d and r.

Results for Ten-protein Set and CATH Proteins

Figure 4 shows surface-area results for the ten-protein set of Street and Mayo. For each protein, panel (a) shows the difference between the total buried area using our pairwise method, $A^{\text{buried}}_{\text{pairwise}}$, and the exact value, $A^{\text{buried}}_{\text{exact}}$, versus the exact value, and panel (b) shows the difference between the nonpolar exposed area using the pairwise method, $A^{\text{(n)exposed}}_{\text{pairwise}}$, and the exact value, $A^{\text{(n)exposed}}_{\text{exact}}$, versus the exact value. Total-area results without generic side-chains are shown both for a single scaling factor $s = 0.67$, and for Street and Mayo's three-parameter scaling. For total areas, our new method, with one, two, or three-sphere generic side-chains and no scaling parameter, achieved significantly better results. The bottom two panels in Figure 4 show the distributions of the residue-by-residue area deviation $\delta A_{\text{residue}}$ [eq. (10)] for the ten proteins for (c) total buried and (d) nonpolar exposed area. It is clear that our new method also achieves significantly and consis-

TABLE I. Average and Standard Deviation of $\delta A_{\text{protein}}$ and $\delta A_{\text{residue}}$ (Å^2)

Ten-Protein Set	$\delta A^{\text{buried}}_{\text{protein}}$	$\delta A^{(\text{n})\text{exposed}}_{\text{protein}}$	$\delta A^{\text{buried}}_{\text{residue}}$	$\delta A^{(\text{n})\text{exposed}}_{\text{residue}}$
$n = 0, s = 0.67$	0.10 ± 2.21	-1.96 ± 1.67	1.11 ± 13.2	-2.68 ± 11.4
$n = 0, s = 0.55, 0.77, 0.73$	0.05 ± 0.70	-0.91 ± 0.68	0.13 ± 7.01	-0.90 ± 5.94
$n = 1, d = 2.10$ Å, $r = 3.09$ Å, $s = 1.0$	-0.01 ± 0.43	0.04 ± 0.24	-0.04 ± 3.58	0.11 ± 2.65
$n = 2, d = 1.09$ Å, $r = 2.92$ Å, $s = 1.0$	-0.08 ± 0.47	0.07 ± 0.25	-0.13 ± 3.47	0.15 ± 2.58
$n = 3, d = 0.61$ Å, $r = 2.85$ Å, $s = 1.0$	0.06 ± 0.48	-0.07 ± 0.26	0.06 ± 3.58	-0.06 ± 2.81
CATH Set	$\delta A^{\text{buried}}_{\text{protein}}$	$\delta A^{(\text{n})\text{exposed}}_{\text{protein}}$	$\delta A^{\text{buried}}_{\text{residue}}$	$\delta A^{(\text{n})\text{exposed}}_{\text{residue}}$
$n = 0, s = 0.67$	-2.84 ± 3.21	0.97 ± 2.72	-1.79 ± 11.1	0.15 ± 9.57
$n = 0, s = 0.55, 0.77, 0.73$	-1.43 ± 1.96	0.51 ± 1.63	-1.06 ± 7.42	0.25 ± 6.31
$n = 1, d = 2.10$ Å, $r = 3.09$ Å, $s = 1.0$	0.38 ± 0.58	-0.18 ± 0.42	0.31 ± 3.70	-0.12 ± 2.88
$n = 2, d = 1.09$ Å, $r = 2.92$ Å, $s = 1.0$	0.35 ± 0.59	-0.16 ± 0.43	0.28 ± 3.63	-0.11 ± 2.85
$n = 3, d = 0.61$ Å, $r = 2.85$ Å, $s = 1.0$	0.41 ± 0.62	-0.23 ± 0.45	0.38 ± 3.75	-0.20 ± 3.02

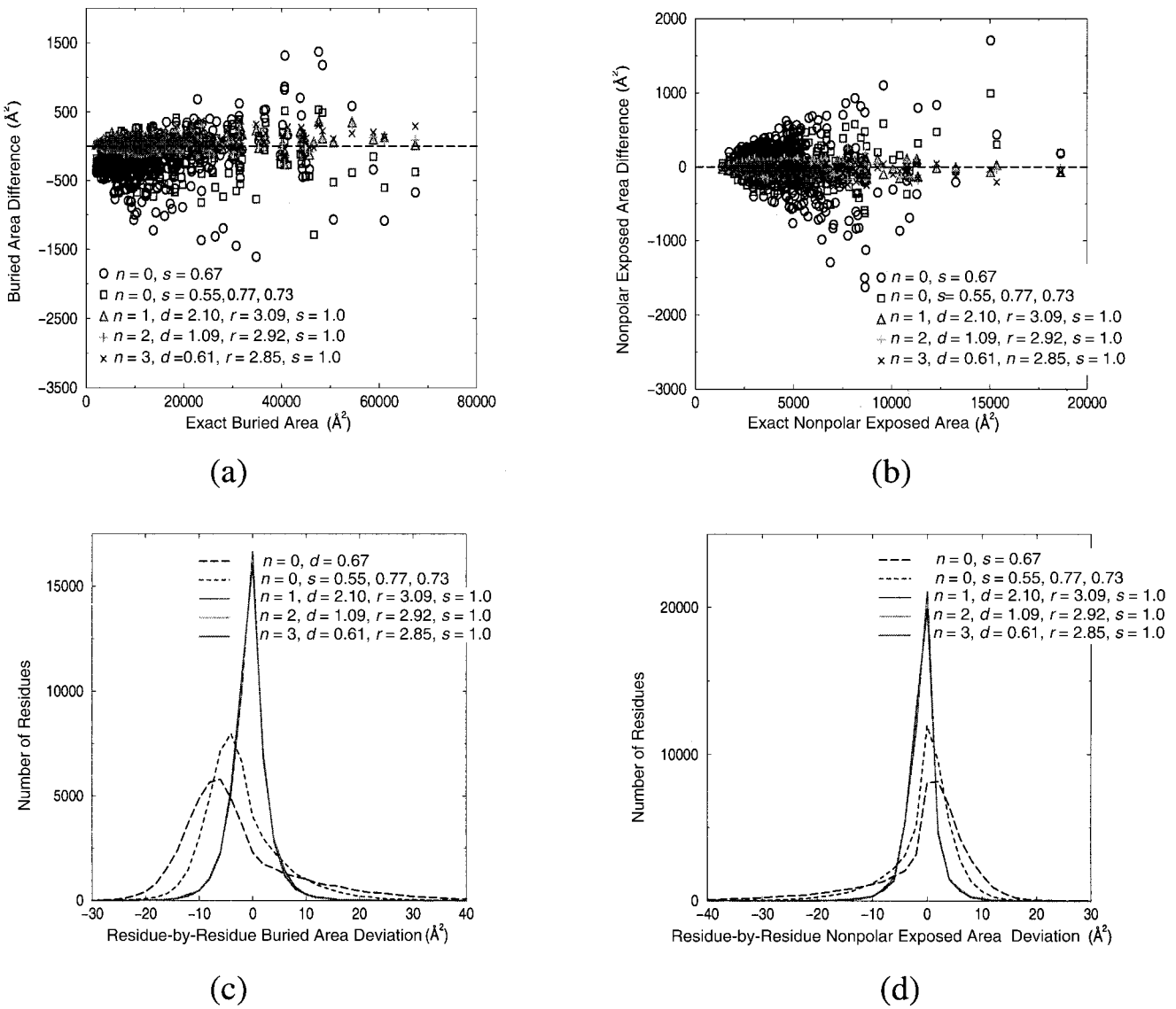

Fig. 5. Surface area results for the 377 CATH proteins: for each protein, (a) the difference between the total pairwise buried area ($A^{\text{buried}}_{\text{pairwise}}$) and the exact buried area ($A^{\text{buried}}_{\text{exact}}$) is plotted against the exact buried area; (b) the difference between the total pairwise nonpolar exposed area ($A^{(\text{n})\text{exposed}}_{\text{pairwise}}$) and the exact nonpolar exposed area ($A^{(\text{n})\text{exposed}}_{\text{exact}}$) is plotted against the exact nonpolar exposed area. (c) Distributions of residue-by-residue buried-area deviations $\delta A^{\text{buried}}_{\text{residue}}(i) = A^{\text{buried}}_{i,\text{pairwise}} - A^{\text{buried}}_{i,\text{exact}}$ for all residues in the 377 CATH proteins. (d) Distributions of residue-by-residue nonpolar-exposed-area deviations $\delta A^{(\text{n})\text{exposed}}_{\text{residue}}(i) = A^{(\text{n})\text{exposed}}_{i,\text{pairwise}} - A^{(\text{n})\text{exposed}}_{i,\text{exact}}$ for all residues.

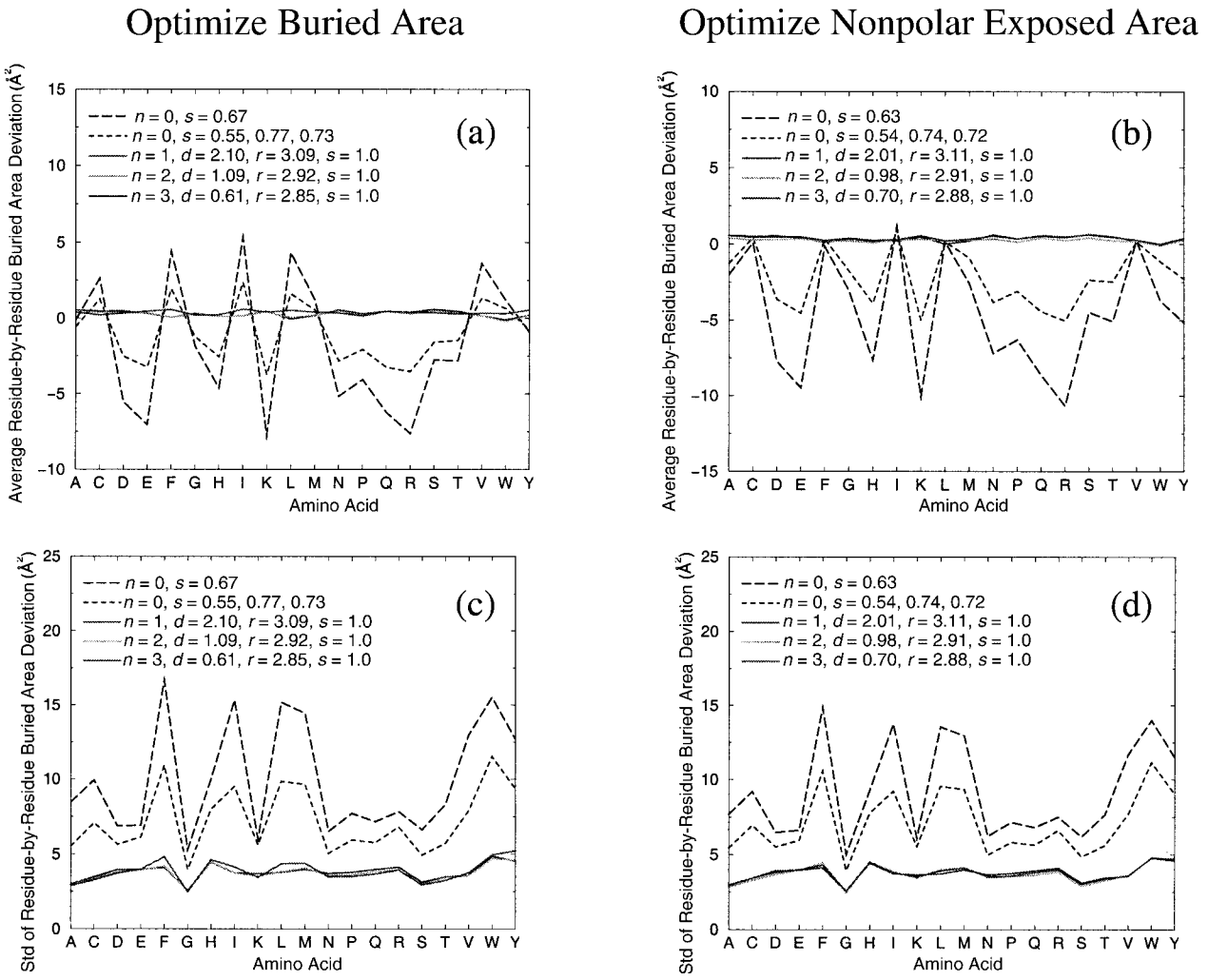

Fig. 6. The average (a, b) and standard deviation (c, d) of the residue-by-residue buried-area deviation for each amino acid type, averaged over the residues of the 377 CATH proteins. The parameters have been optimized for the ten-protein set of Street and Mayo using total buried area in panels (a) and (c) and using total nonpolar exposed area in panels (b) and (d).

tently better results for residue-by-residue surface areas than the methods without generic side-chains. The average and standard deviation results of both $\delta A_{\text{protein}}$ and $\delta A_{\text{residue}}$ for the ten-protein set are given in Table I.

We have further tested our generic side-chain method on a larger, more systematic set of proteins: 377 single-segment representative domains from the CATH database.[12] The results are presented in Figure 5 and Table I. Clearly, as compared to the optimized three-parameter method without generic side-chains, our new, generic-side-chain method achieves significantly better results. In particular, the average total protein deviation per residue for buried area $\delta A^{\text{buried}}_{\text{protein}}$ has been reduced in magnitude from -1.43 Å^2 to 0.38 Å^2, and the spread of this average deviation from protein to protein has been reduced from 1.96 Å^2 to 0.58 Å^2. Strikingly, the spread of the residue-by-residue buried-area deviation $\delta A^{\text{buried}}_{\text{residue}}$ has been reduced from 7.42 Å^2 to 3.70 Å^2. Thus, for individual residues, the use of generic side-chains reduces errors by more than a factor of two. Similar error reductions are also achieved for exposed areas.

In Figure 6, we show (a) the average and (c) the standard deviation of the residue-by-residue buried area deviation for each type of amino acid, averaged over the 377 CATH proteins. The generic-side-chain method significantly improves the accuracy of the area calculation for all 20 amino acid types. For the two methods without generic side-chains, the largest errors are associated with the largest hydrophobic amino acids: phenylalanine (F), isoleucine (I), leucine (L), methionine (M), tryptophan (W), tyrosine (Y) and, for the $s = 0.67$ case, with medium-sized valine (V). These large hydrophobic amino acids are typically found in the protein core, where the overlapping burial effect is most severe. It is precisely for these amino acids that the generic-side-chain method achieves the greatest fractional reduction in error. This clearly demonstrates that the success of the generic-side-chain method is due to its improved treatment of overlapping buried areas.

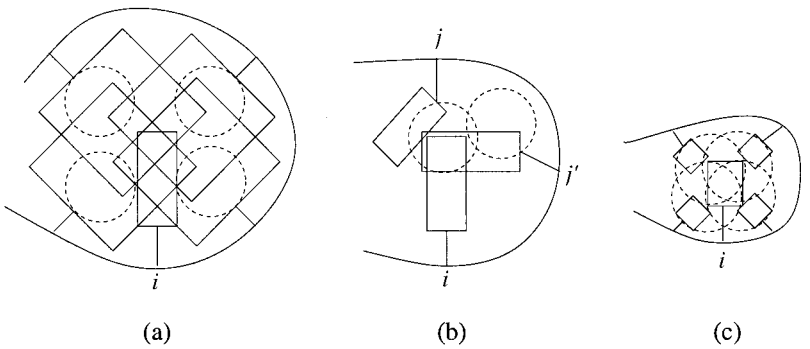

Fig. 7. Schematic illustration of the three sources of large error in the generic-side-chain approximation. (a) Part of the residue at *i* is multiply buried by real side-chains (rectangles) at other sites, but not by the generic side-chains (circles). This results in a large overestimate of the buried area of residue *i*. (b) Part of the residue at *i* is buried by one (and only one) generic side-chain but not by the real side-chain (at *j*), and the same part of *i* is also buried by another real side-chain (at *j'*). This results in an underestimate of the buried area at *i*. (c) The residue at *i* is multiply buried by the generic side-chains at other sites, but not by real side-chains. Like (a), this results in an overestimate of the buried area at *i*.

The success of the generic-side-chain method does not depend sensitively on the choice of area optimized. As described above, we optimized parameters using the total buried area for the ten-protein set of Street and Mayo. For comparison, in Figures 6 (b and d), we show results for parameters chosen to optimize total nonpolar exposed area for the same protein set. The generic-side-chain parameters are very similar, and, more importantly, the residue-by-residue results for buried area are equally good. In contrast, the results without generic side-chains ($n = 0$) depend significantly on the choice of area optimized: in Figure 6(b), when optimizing the nonpolar exposed area, the nonpolar residues (e.g., F, I, L) have small errors but the polar residues (e.g. E, K, R) have large errors; in Figure 6(a), when optimizing total buried area, the polar and nonpolar residues on average have similar errors (but with opposite signs).

As one can see in Figures 4, 5, and 6, the one-sphere generic side-chains do as well as those with two or three spheres. In Figure 3, with the three optimized side-chains drawn to scale, we can see that these three generic side-chains are geometrically similar and thus lead to similar results.

Sources of Error

Can we improve our method further? A straightforward approach would be to use three scaling parameters as Street and Mayo have done, corresponding to core–core, non-core–non-core, and core–non-core interactions.[1] This approach significantly improves results when no generic side-chains are used because the different scaling factors compensate for different amounts of overlapping burial. For example, core residues are more likely to be multiply buried than non-core residues, giving a smaller s for core–core interactions. However, the use of three scaling parameters produced no significant improvement in results for the generic-side-chain method (data not shown). This is understandable because the core–non-core classification is simply a way of estimating the local extent of overlapping burial, and the generic side-chains automatically and very accurately account for multiply-buried areas.

Regarding the sources of error in the generic-side-chain method, we consider a small area a on the surface of the real side-chain at i. There are three situations. In the first situation [Fig. 7(a)], the area a is not buried by any of the generic side-chains at $j \neq i$. Then for the small area we have $A_{i,\mathrm{gs}}^{\mathrm{buried}} = 0$. If the area is buried by a real side-chain at j, we have $A_{i:j,\mathrm{gs}}^{\mathrm{buried}} = a$ and the correction is $A_{i:j,\mathrm{gs}}^{\mathrm{buried}} - A_{i,\mathrm{gs}}^{\mathrm{buried}} = a$ [see eq. (8)]. If the area is multiply buried by n real side-chains, the correction is $A_{i:j,\mathrm{gs}}^{\mathrm{buried}} - A_{i,\mathrm{gs}}^{\mathrm{buried}} = na$, and because the exact result can be at most $A_{i,\mathrm{exact}}^{\mathrm{buried}} = a$ the generic-side-chain method overestimates the buried area by $(n - 1)a$ at this point. In Figure 7(a), we show an extreme case in which the real side-chain at i is not buried by any of the generic side-chains but is multiply buried by many real side-chains. This results in a large overestimate of the buried area (and at the same time a large underestimate of the exposed area).

In the second situation [Fig. 7(b)], the area on the real side-chain at i is buried by one and only one generic side-chain (at j)(i.e., $A_{i,\mathrm{gs}}^{\mathrm{buried}} = a$). Then $A_{i:j',\mathrm{gs}}^{\mathrm{buried}} = a$ for all $j' \neq j$ (because it is always buried by the generic side-chain at j). This means that the correction terms for all j' are zero (*i.e.*, $A_{i:j',\mathrm{gs}}^{\mathrm{buried}} - A_{i,\mathrm{gs}}^{\mathrm{buried}} = 0$). The only non-zero correction occurs if the area is not buried by the real side-chain at j, so that $A_{i:j,\mathrm{gs}}^{\mathrm{buried}} - A_{i,\mathrm{gs}}^{\mathrm{buried}} = -a$. If at the same time the area is buried by a real side-chain at j', the exact result is $A_{i,\mathrm{exact}}^{\mathrm{buried}} = a$, while the pairwise result is zero. This underestimates the buried area by $-a$. Figure 7(b) shows this case: part of the real side-chain at i is buried by one and only one generic side-chain but not by the real side-chain (at j). The area buried by the generic side-chain at j is buried by another real side-chain at j'. This results in an underestimate of the buried area.

In the final situation [Fig. 7(c)], the area on the real side-chain at i is buried by two or more generic side-chains ($A_{i,\mathrm{gs}}^{\mathrm{buried}} = a$). Then because of multiple burial, $A_{i:j,\mathrm{gs}}^{\mathrm{buried}} = a$ and all corrections are zero. If the area is not buried by any of the real side-chains, the true buried area is zero ($A_{i,\mathrm{exact}}^{\mathrm{buried}} = 0$) and there is an overestimate of the buried area by a. In Figure 7(c), we show an extreme case in which the

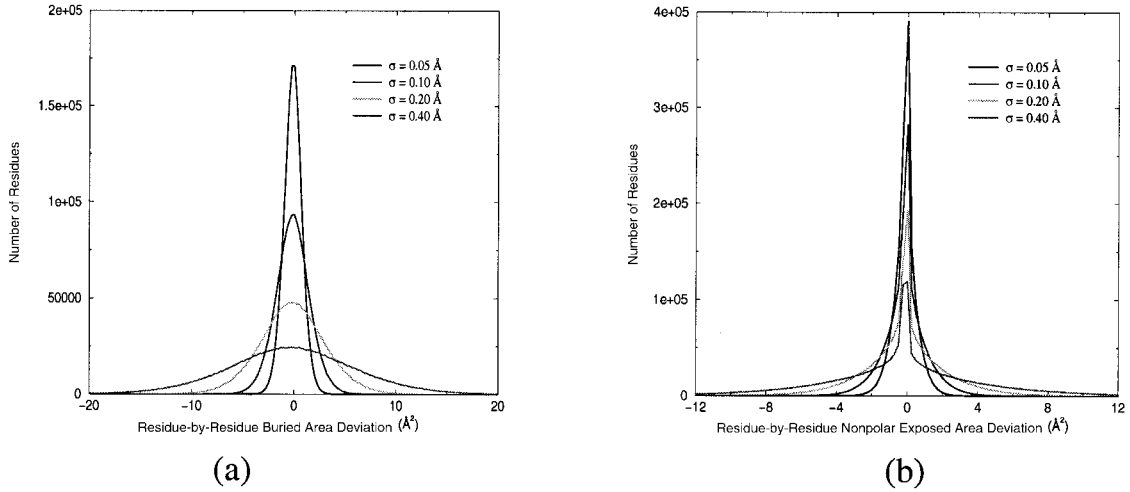

Fig. 8. Distributions about the average for exact residue-by-residue (a) buried area and (b) nonpolar exposed area for 30 randomly perturbed variants of each of the 377 CATH proteins. In each variant, all atoms deviate from their PDB coordinates following a Gaussian distribution with zero mean and (radial) root-mean-square distance σ. For σ = 0.05, 0.10, 0.20, 0.40 Å, respectively, the standard deviations for buried area are 0.89, 1.66, 3.24, 6.36 Å and those for nonpolar exposed area are 0.63, 1.19, 2.33, 4.67 Å^2. To accurately measure small area differences between original and perturbed structures, in this calculation we used 2048 dots per atom with 20 random rotation frames.

entire real side-chain at i is multiply buried by generic side-chains but not by the real side-chains. This results in an overestimate of the buried area of residue i.

In principle, the above types of errors could be substantially reduced by considering three-residue corrections to the buried areas, *i.e.*, by calculating the exposed surface of residue i in the presence of real side-chains at j and j', and generic side-chains at all other positions. However, these three-body terms would add significantly to the number of calculations required, and would not permit fast graph-based optimization techniques such as DEE.[4,6,7]

Effect of Coordinate Changes on Surface Area

Because the coordinates of protein atoms obtained by experimental techniques such as X-ray crystallography and nuclear magnetic resonance (NMR) have uncertainties, it is natural to ask whether the improvement in pairwise surface area calculation obtained here is significant when compared to the inherent surface area uncertainties. For the 377 CATH protein domains,[12] we perturbed the position of each atom from its PDB coordinate by a three-dimensional Gaussian distribution with zero mean and radial root-mean-square distance σ.** In this way, for each protein, we obtained 30 perturbed structures, and we calculated for each residue the deviations of the exact surface area from the mean.*** Fig. 8 shows the distributions of the exact residue-by-residue total buried area (a) and the nonpolar exposed area (b) for σ = 0.05, 0.10, 0.20, 0.40 Å. In Figure 8(b), a large number of residues show no change in nonpolar exposed area (the sharp peak at zero). This is because when the residues are buried in the core, their exposed areas are often zero, and small random perturbations in neighboring atomic coordinates do not change this situation. In Figure 9, we show that the standard deviations of the areas in Figure 8 are proportional to the root-mean-square coordinate perturbation σ. From Figure 9, when the root-mean-square perturbation distance is 0.1 Å, the standard deviation for buried areas is 1.66 Å^2, while that for nonpolar exposed areas is 1.19 Å^2. Earlier we showed that for the CATH set our generic side-chain method reduces the error in total buried area from 7.42 Å^2 (using Street and Mayo's method) to 3.70 Å^2, and reduces the error in nonpolar exposed area from 6.31 Å^2 to 2.88 Å^2. The improvements obtained by our method are therefore larger than the deviations in surface area with small changes in coordinates and are therefore important even in light of coordinate uncertainties.

DISCUSSION AND CONCLUSIONS

The stable folding of proteins depends largely on the burial of hydrophobic residues and the exposure of polar and charged residues. The solvation energy for a residue depends on atomic polarities and charges, and on solvent-

**If a distribution of points is generated in three dimensions, such that each coordinate (x, y, z) has a normalized Gaussian distribution with zero mean and root-mean-square deviation σ′, then the radial distance of the points from the origin r has the distribution $r^2 \exp(-r^2/2\sigma'^2)$ (ignoring the normalization constant) and the root-mean-square radial distance is $\sqrt{\langle r^2 \rangle} = \sigma = \sqrt{3}\sigma'$. Therefore, when we perturb each coordinate of an atom following a Gaussian distribution such that the root-mean-square radial distance is σ = 0.10 Å, the root-mean-square rms deviation for each coordinate is $\sigma' = \sigma/\sqrt{3} = 0.058$ Å.

***To accurately measure small area differences between original and perturbed structures, we used 2048 dots per atom in the dot-surface method and averaged surfaces area results from 20 random rotation frames. This was especially necessary when the perturbation (σ) was small and errors due to the discreteness of the dots became important.

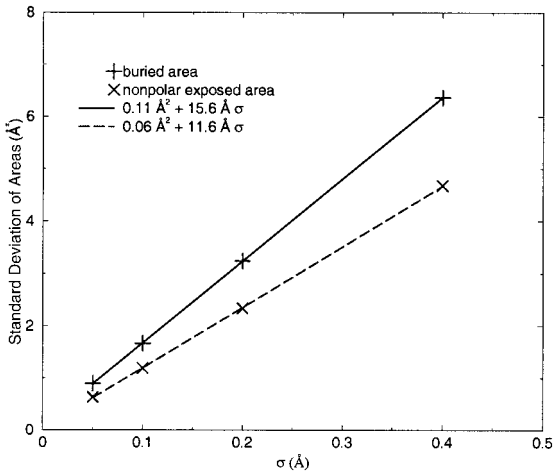

Fig. 9. Standard deviations of the exact residue-by-residue buried and nonpolar exposed areas about the average for 30 randomly perturbed variants of each of the 377 CATH proteins as functions of Gaussian coordinate perturbation rms distance σ. The lines are linear-regression results.

exposed surface areas. By assuming a linear relation between changes in solvation energy and changes in surface area, atom-dependent solvation energies have been derived from experimental free energies of amino acids[16] and proteins.[17] These empirical solvation energies form an essential component of energy functions for protein design,[3] which puts a premium on accurate calculation of solvent-exposed areas. However, for applications such as side-chain and rotamer optimization for a fixed backbone, the combinatorial explosion of configurations makes exact calculation of surface areas impractical. Street and Mayo's pairwise approximation for surface areas has allowed the use of fast search algorithms such as DEE for design optimization.

The generic side-chain method presented here is a consistently more accurate alternative to the Street and Mayo algorithm. Tested on a set of 377 proteins from the CATH database we found a reduction in the spread of the residue-by-residue error from 7.42 Å^2 to 3.70 Å^2. One way to characterize this improvement is in terms of the increased size of proteins that can be designed with confidence. Using Street and Mayo's[1] estimate of solvation energies, 26 cal mol^{-1} Å^{-2} favoring hydrophobic burial and 100 cal mol^{-1} Å^{-2} opposing polar burial, and assuming a 50-50 split between hydrophobic and polar residues, our method reduces per residue root-mean-square energy errors by 0.2 kcal/mol. The accumulation of independent errors begins to exceed a tolerable error margin of $\sim$2 kcal/mol for proteins of 25 residues using Street and Mayo's method, and for proteins of 100 residues using our generic-side-chain method.

An important result of our study is that there is, in effect, only one universal generic side-chain. Specifically, we found very similar side-chain parameters, and nearly identical results, whether we optimized total buried area or total nonpolar exposed area (see Fig. 6). Moreover, geometrically similar generic side-chains were obtained using one, two, or three spheres, as shown in Figure 3, and all three produced equally good results. Therefore, we recommend the use of generic side-chains each consisting of a single sphere, *e.g.* of radius $r = 3.09$ Å centered a distance $d = 2.10$ Å from the C_α atom.

In contrast, the results of the Street and Mayo methods without generic side-chains are not universal. With either one or three scaling parameters, the results for residue-by-residue buried area depend on the choice of whether to optimize total buried area or total nonpolar exposed area, as shown in Figure 6(a and b). For the methods without generic side-chains, the results also depend on the set of proteins used for optimization. For example, the scaling parameters obtained by optimizing total buried area for Street and Mayo's ten-protein set led to a relatively large error for buried area ($\delta A^{\text{buried}}_{\text{protein}} = -1.43 \pm 1.96$ Å^2) when applied to the larger set of 377 CATH proteins.

We speculate that the success of our generic-side-chain approach, and the existence of a universal generic side-chain, reflect a nearly homogeneous mixing of side-chains of different sizes in natural proteins. Otherwise, significant segregation of large and small side-chains between surface and core would yield different generic side-chains when optimizing buried or exposed areas.

A final note is that our generic-side-chain method not only improves the accuracy of the pairwise surface-area calculation but is likely to improve the performance of combinatorial optimization methods such as the dead-end-elimination algorithm. This is because our method shifts the emphasis of the surface area calculation from two-body side-chain-side-chain terms to one-body side-chain-generic side-chain terms. As noted in ref. 18, potential functions that emphasize one-body terms perform better in optimization. This improvement will be investigated quantitatively in a separate publication.

ACKNOWLEDGMENTS

N.G.Z. and C.Z. were supported by NSF Grants DMR-0094176 and DMR-0313129. We thank the editorial manager and the reviewers for helpful comments.

REFERENCES

1. Street AG, Mayo SL. Pairwise calculation of protein solvent accessible surface areas. Fold Des 1998;3:253–258.
2. Dill KA. Dominant forces in protein folding. Biochemistry 1990;29: 7133–7155.
3. Gordon DB, Marshall SA, Mayo SL. Energy function for protein design. Curr Opin Struct Biol 1999;9:509–513.
4. Desmet J, Maeyer MD, Hazes B, Lasters I. The dead-end elimination theorem and its use in protein side-chain positioning. Nature 1992;356:539–542.
5. MoretBME, Shapiro HD. Algorithms from P to NP. Redwood City: Benjamin 1991. Nemhauser GL, Rinnooykan AHG, Todd MJ, eds. Optimization. Amsterdam: North-Holland 1989.
6. Bolon DN, Mayo S. Enzyme-like proteins by computational design. Proc Natl Acad Sci 2001;98:14272–14279.
7. Looger LL, Dwyer MA, Smith JJ, Hellinga HW. Computational design of receptor and sensor proteins with novel functions. Nature 2003;423:185–190.
8. Wodak SJ, Janin J. Location of structural domains in proteins. Biochemistry 1981;20:6544–6552.
9. Wernisch L, Hunting M, Wodak SJ. Indentification of structural domains in proteins by a graph heuristic. Proteins 1999;35:338–352.
10. Wodak SJ, Janin J. Analytical approximation to the accessible surface area of proteins. Proc Natl Acad Sci 1980;77:1736–1740.

11. Marshall SA, Mayo SL. Achieving stability and conformational specificity in designed proteins via binary patterning. J Mol Biol 2001;305:619–631.
12. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH - A hierarchic classification of protein domain structures. Structure 1997;5:1093–1108.
13. Kurochkina N, Lee B. Hydrophobic potential by pairwise surface area sum. Prot Eng 1995;8:437–442.
14. Shrake A, Rupley JA. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. J Mol Biol 1973;79:351–371.
15. LeGrand SM, Merz KM. A rapid analytic approximation to molecular surface area via the use of boolean logic and look-up tables. J Comp Chem 1993;14:349–352.
16. Wesson L, Eisenberg D. Atomic solvation parameters applied to molecular dynamics of proteins in solution. Protein Sci 1992;1:227–235.
17. Dahiyat, BI, Mayo SL. Protein design automation. Protein Sci 1996;5:895–903.
18. Gordon DB, Hom GK, Mayo SL, Pierce NA. Exact rotamer optimization for protein design. J Comput Chem 2003;24:232–243.