An Amino Acid Has Two Sides: A New 2D Measure Provides a Different View of Solvent Exposure

Thomas Hamelryck*
Bioinformatics Center, University of Copenhagen, Copenhagen, Denmark

ABSTRACT **The concept of amino acid solvent exposure is crucial for understanding and predicting various aspects of protein structure and function. The traditional measures of solvent exposure however suffer from various shortcomings, like for example the inability to distinguish exposed, partly exposed, buried, and deeply buried residues. This article introduces a new measure of solvent exposure called Half-Sphere Exposure that addresses many of the shortcomings of other methods. The new measure outperforms other measures with respect to correlation with protein stability, conservation among fold homologs, amino acid-type dependency and interpretation. The measure consists of the number of Cα atoms in two half spheres around a residue's Cα atom. Conceptually, one of the half spheres corresponds to the side chain's neighborhood, the other half sphere being in the opposite direction. We show here that the two half spheres correspond to two regions around an amino acid that are surprisingly distinct in terms of geometry and energy. This aspect of protein structure introduced here forms the basis of the Half-Sphere Exposure measure. The results strongly suggest that in many respects, a 2D measure is inherently much better suited to describe solvent exposure than the traditional 1D measures. Importantly, Half-Sphere Exposure can be calculated from the Cα atom coordinates only, which abolishes the need for a full-atom model to calculate solvent exposure. Hence, the measure can be used in protein structure prediction methods that are based on various simplified models. Half-Sphere Exposure has great potential for use in protein structure prediction and analysis. Proteins 2005;59:38–48.** © 2005 Wiley-Liss, Inc.

Key words: protein structure; coordination number; residue depth; accessible solvent area; solvent exposure; simplifed models; protein structure prediction

INTRODUCTION

The concept of amino acid solvent exposure plays a crucial role in understanding, analyzing and predicting protein structure, folding, function and interactions.[1–10] For this reason, a good measure of solvent exposure is of enormous importance. Here, we analyze strengths and weaknesses of some of the most used solvent-exposure measures (accessible surface area, coordination number, residue depth), and subsequently introduce and analyze a new 2D measure that efficiently addresses many conceptual and practical problems associated with the currently used approaches. We justify the new measure by showing that an amino acid's neighborhood in a protein can be subdivided in two distinct regions with different properties in terms of geometry and energy.

The solvent-accessible surface area (ASA), introduced in 1971 by Lee and Richards[11] and subsequently refined by Greer[12] and Connolly,[13] is undoubtedly the most widely used approach to measure solvent exposure today. The ASA of a given residue is calculated as the surface that is accessible to a ball with a certain radius (typically 1.4 or 1.5 Å). New methods offer important speed improvements over the classic Connolly algorithm,[14] but typically the ASA method is too slow for use in many applications without heuristics.[6,15,16]

In order to compare the solvent exposure of residues of different size (for example Leu and Met) the relative solvent accessible surface area (rASA) is often used. This is simply the ASA of a residue divided by the maximum ASA for that residue type.[17] In practice, comparison of rASA values is still difficult for residues of very different size (for example, Arg and Gly).

A well-known limitation of ASA is that it does not provide any information for completely buried residues: it is impossible to distinguish a deeply buried residue from a residue buried just below the surface.

One solution for this problem is using the distance to the solvent accessible surface, or, in other words, the depth, as a way to measure solvent exposure.[18–20] Atom depth is defined as the distance between a given atom and the nearest point on the solvent-accessible surface. Residue depth (RD) is the average atom depth of a residue's atoms. Residue depth has some attractive features, but we show here that RD is not a sensitive measure with respect to partly buried residues, this in contrast to the rASA. In

Abbreviations: ASA, accessible surface area of a residue; CN, coordination number; DSSPa, ASA as calculated by DSSP; DSSPr, rASA calculated from DSSP; HSE, Half-Sphere Exposure; rASA, relative acessible surface area; RD, residue depth; RDα, atom depth of a residue's Cα atom; ΔASA, difference between the maximum ASA and the ASA.

*Correspondence to: Thomas Hamelryck, Bioinformatics Center, University of Copenhagen, Universitetsparken 15, Bygning 10, 2100 Copenhagen, Denmark. E-mail: thamelry@binf.ku.dk

Received 12 August 2004; Accepted 14 September 2004

Published online 1 February 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20379

addition, it remains difficult to compare RD values for residues of different size. Since calculating RD implies calculating the solvent accessible surface, the method suffers from at least the same computational burden as the ASA/rASA method.

A third, widely used way to measure solvent exposure is the coordination number (CN). [2,21] This is simply the number of Cα atoms within a sphere around the Cα atom of a residue. The advantages of this measure are that it is easy to implement, fast to compute, and conceptually simple to interpret. Moreover, contrary to RD and ASA, a full-atom model of a protein is not needed. For these reasons, structure prediction methods that make use of simplified models (typically Cα atoms only[8]) often model solvent exposure using the CN. For example, CN is the method of choice for the successful ROSETTA ab initio structure prediction method.[2] As can be expected, the CN provides only a crude, rather insensitive description of a residue's solvent exposure as compared to rASA or RD. On the other hand, CN values can be directly compared for residues of different sizes, since it essentially measures a residue's local environment.

In addition to the issues mentioned above, some additional factors are important. Solvent exposure, as for example measured by RD[18] and ASA,[22] is expected to be correlated with the stability of mutants. Also, a measure of solvent exposure is expected to be reasonably conserved among proteins with similar folds, which is important for its potential use in solvent-exposure prediction methods.[17,23]

We address the following two questions: how to construct a measure that combines the best features of each of the above mentioned solvent-exposure measures, and, what view of solvent exposure does such a superior measure offer? Ideally, we would like to have a measure that:

- Is easy to implement and speed efficient.
- Does not require a full atom model.
- Distinguishes shallowly and deeply buried residues well.
- Distinguishes buried, partially buried and exposed residues well.
- Allows easy comparison of solvent exposure, regardless of residue size.
- Correlates reasonably well with the stability of mutants.
- Is reasonably well conserved for members of a given fold.

In this article we introduce Half-Sphere Exposure (HSE), a 2D measure of a residue's solvent exposure. This measure is nearly as easy to compute as the CN, but outperforms the classic ASA and RD measures in many respects (for example conservation, correlation with stability of mutants, sensitivity, dependence on a full atom model). We provide two, in essence equivalent ways to compute HSE, depending on whether information is available about Cα and Cβ positions (HSEβ) or only about the Cα positions (HSEα).

METHODS

Database Construction

For the construction of a representative and non-redundant set of crystal structures we used the PFam database (http://pfam.wustl.edu/).[24] All PFam protein families with representatives in the PDB were initially selected. Since many PFam families have more than one representative in the PDB database, we used a method to select a "best" representative. This was done using the PDBSelect database,[25] which ranks all PDB files according to a set of quality parameters (file pdb_select. 2002_Apr.90 from http://www.cmbi.kun.nl/gv/pdbsel/). For each PFam family, the crystal structure (NMR structures were not considered) with the highest PDBSelect score was selected as representative, and the relevant residues were written out to a separate PDB file, discarding hydrogen atoms, waters, and ligands. In total, 144,258 residues were used from 985 structures (see the section "Supplementary Data" below).

For the evaluation of conservation, the SABmark 1.63 Twilight Zone database was used (http://bioinformatics.vub.ac.be/databases/databases.html).[26] This database consists of aligned structures falling in 236 folds, where all aligned sequence pairs have a BLAST E-value of at least 1. For each of the fold groups, one pair of aligned representative structures was chosen randomly. In total, 230 pairs were used (see the section "Supplementary Data" below).

Calculation of Residue Depth and Accessible Surface Area

Solvent accessible surface area (ASA, in Å^2) was calculated using the program DSSP.[27] The relative accessible surface area (rASA), i.e., the ASA divided by the maximum ASA of a residue of that type, was calculated using the values given by Rost and Sander.[17] For the calculation of residue depth (RD) and Cα atom depth (RDα),[18–20] the solvent-accessible surface of a protein was determined with the program MSMS.[14] The program was run with default atom radius parameters and a sphere radius of 1.5 Å (the MSMS default radius). MSMS writes out a list of vertices that represent the solvent-accessible surface. Atom depth was then calculated as the distance between the atom and its closest vertex. Residue depth was calculated as the average depth of a residue's atoms, hydrogen atoms excluded.

Calculation of HSEα and HSEβ

The overall computation of the Half-Sphere Exposure (HSE) measures is explained in the "Results and Discussion" section below. Coordination number (CN) and HSE were calculated using our program `hsexpo` (see section on Implementation below). For the calculation of these measures, a sphere radius of 13 Å was used. The `hsexpo` program calculates the HSE values of a residue by first finding all Cα atoms within 13 Å of the residue's Cα atom, and then applying a specific rotation to these atoms. The rotation is chosen so that it aligns the Cα-Cβ (for HSEβ) or Cα-pseudo Cβ (for HSEα) vector with the Z-axis. The

HSEu and HSEd values are the number of atoms with strictly positive or negative Z-coordinates, respectively.

Thermodynamic Data

The Protherm database[28] provided $\Delta\Delta G$ values of Val/Ile/Leu to Ala point mutants, where $\Delta\Delta G$ is defined as the free energy of unfolding obtained with the equation $\Delta\Delta G = \Delta T_m \Delta S$ in the case of thermal denaturation. All mutations for which data was available on PDB structure identifier, publication reference, mutation and $\Delta\Delta G$ were used. If multiple $\Delta\Delta G$ values occurred for the same residue, we picked the value corresponding to the most recent reference. In total, the dataset contained 91 point mutants. For a list of the mutants used and their references see the section "Supplementary Data" below.

Gnuplot's fit command produced the linear least-squares fit to the thermodynamic data in Figure 12. ΔASA was calculated by subtracting the ASA of a residue from the maximum ASA for that residue type.[17]

Implementation

We implemented a program (`hsexpo`) and library module (`Bio.PDB.HSExposure`) to calculate Half-Sphere Exposure as part of the Biopython toolkit (http://www.biopython.org),[29] a set of freely available (under the Biopython license) python modules for bioinformatics.The program is based on Biopython's structural bioinformatics library module `Bio.PDB`.[30] The `hsexpo` program is found in `Scripts/Structure`.

The `hsexpo` program calculates HSE (α and β), CN, RD, and RDα for all amino acid residues in a PDB file. For the residue depth calculations the MSMS[14] program needs to be installed. In addition, hsexpo also provides an interface to the DSSP[27] program. Optionally, the program writes out a PDB file with the calculated solvent exposure placed in the file's temperature factor records for easy visualization (for example in PyMol, http://pymol.sourceforge.net/).

Supplementary Data

PDB files of the structure of FRIL[31,32] are available, each with one of the discussed solvent-exposure measures (see the section "Sensitivity" below) in the temperature factor field of the atom record. The files and their corresponding solvent-exposure measures are:

- FRIL_CN.pdb: CN
- FRIL_DSSPa.pdb: DSSPa
- FRIL_DSSPr.pdb: DSSPr
- FRIL_HSEau.pdb: HSEαu
- FRIL_HSEbu.pdb: HSEβu
- FRIL_HSEad.pdb: HSEαd
- FRIL_HSEbd.pdb: HSEβd
- FRIL_RDa.pdb: RDα,
- FRIL_RD.pdb: RD

The file structures.txt contains a list of PDB files and corresponding chain sections that were used to calculate the various histograms and statistics. The file `sab_pairs.txt` contains a list of structure pairs from the SABmark 1.63 Twilight Zone database[26] that were used in the section "Conservation" below. The file `thermo.pdf` contains the list of point mutants and their references that were used in the section "Correlation with the Stability of Mutants" below.

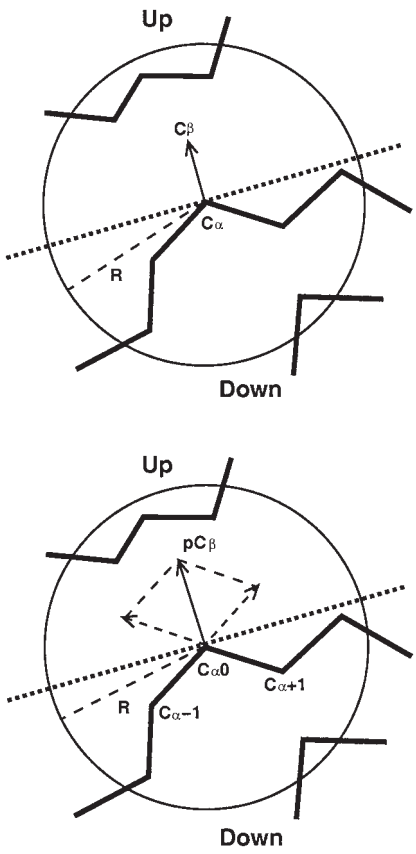

Fig. 1. Computation of HSEβ (top) and HSEα (bottom). The dotted line indicates the position of the plane that divides the sphere with radius R around the Cα atom. The thick black lines represent a part of the Cα trace of the protein.

RESULTS AND DISCUSSION

The HSEβ Measure

In this section, we explain the calculation of the HSE measure from a full atom model, or from a model for which at least the Cα and Cβ positions are known. We call this variant of the HSE measure HSEβ, because it relies on the presence of both Cα and Cβ positions. In the next section, we will describe how to proceed in case only Cα positions are available (for example, in the case of structure prediction using Cα-only models or a low-resolution structure).

The HSEβ calculation is outlined in Figure 1 (top). The first step in the calculation identifies all Cα atoms within a sphere of a certain radius around the residue's Cα. The second step constructs a plane that is perpendicular to the Cα-Cβ vector and runs through the residue's Cα atom. This plane divides the sphere in two equal halves. These halves are labeled "up" in the direction of the Cα-Cβ vector, and "down" in the opposite direction, as indicated in Figure 1. The two dimensional HSEβ measure consists

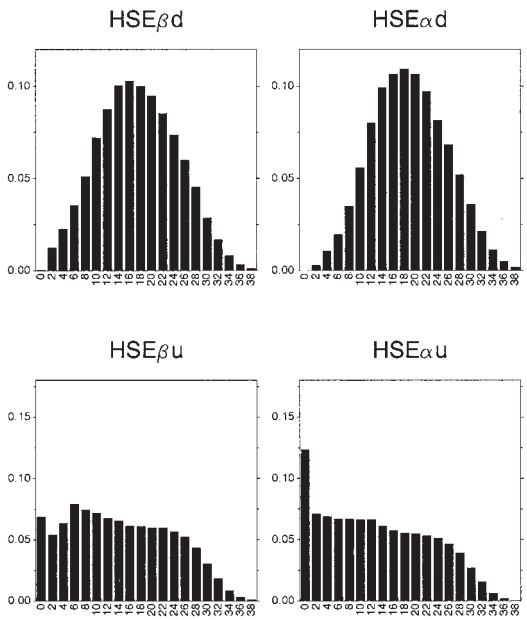

Fig. 2. Histograms for HSEβu, HSEβd, HSEαu and HSEαd. The bin width is 2. The start of each bin is indicated on the x-axis.

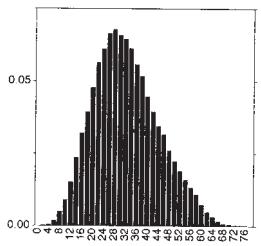

Fig. 3. Histogram for the CN measure. The bin width is 2. The start of the bins is indicated on the x-axis. Only one in two bins is labeled for clarity.

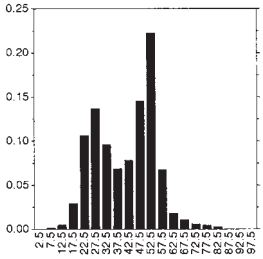

Fig. 4. Histogram of the angle between the pCβ-Cα and the Cβ-Cα vectors for all non-Gly residues. The bin width is 5°. The center of each bin is indicated on the x-axis.

of the number of Cα atoms in the upper (HSEβu) and lower half sphere (HSEβd). For example, the HSEβ measure of the residue in Figure 1 is (3, 5).

In the case of Gly, which obviously lacks a Cβ atom, we construct a pseudo-Cβ atom by rotating the N atom over −120° along the Cα-C axis.

The choice of the sphere radius is a compromise between two demands. A radius that is too small misses residue pairs that are obviously shielding each other from the solvent. A radius that is too large includes irrelevant residue pairs. Based on visual inspection of protein structures, 13 Å is a good compromise, and all results described in this article were obtained using this radius. For use of the measure in solvent-exposure prediction, a radius could be selected that optimizes the predictability of the measure.[33]

We calculated histograms for HSEβu and HSEβd from a set of 985 structures, each representing a protein family (see Methods). The histograms (Fig. 2) show very different distributions, indicating that the described division of a residue's spherical neighborhood indeed captures regions with different properties.

The CN is simply the sum of the HSEβu and HSEβd pair. The histogram of the CN (Fig. 3), is thus the result of the sum of two very different distributions.

The HSEβ measure can be interpreted in the following way: HSEβu expresses the degree of solvent exposure in the direction of a residue's side chain, while HSEβd gives information about the degree of solvent exposure in the opposite direction. The latter direction corresponds to the direction in which the main chain atoms of a residue are less shielded from the solvent by the residue's side chain atoms.

The HSEα Measure

The calculation of HSEβ requires a full atom model, or at least a model for which the coordinates of the Cα atoms and the directions of Cα-Cβ vectors are known. Hence, the measure cannot be calculated directly from a Cα-only model.

Fortunately, it is also possible to derive the general direction of the side chain from the Cα coordinates only. We call this Cα-only version of the HSE measure HSEα. The calculation of HSEα is based on the fact that the approximate direction of the side chain can be inferred from the Cα coordinates, as suggested by Raghunathan and Jernigan.[34] Figure 1 (bottom) illustrates the calculation of HSEα.

The only difference with the calculation of HSEβ is the use of a pseudo-Cβ (pCβ) atom instead of the Cβ atom. The pCβ atom position is calculated using the coordinates of the Cα atom of the considered residue ($C\alpha_0$), and the Cα coordinates of the preceding ($C\alpha_{-1}$) and following ($C\alpha_{+1}$) residues. The Cα-pCβ vector, which approximates the Cα-Cβ vector, is calculated by adding the $C\alpha_{-1}$-$C\alpha_0$ and $C\alpha_{+1}$-$C\alpha_0$ vectors.

To assess the difference in orientation between the Cα-pCβ and Cα-Cβ vectors, we calculated the histogram of the angle between the two vectors for all residues, excluding Gly residues (see Fig. 4).

The histogram shows two clear peaks: the left peak is mainly due due to residues in β-sheets, the right peak to residues in α-helices. The histogram shows that the angle between the two vectors is below 55° for the great majority of residues. Hence, the Cα-pCβ and Cα-Cβ vectors coincide quite well and can thus be used to obtain similar estimates of the general position of the side chain.

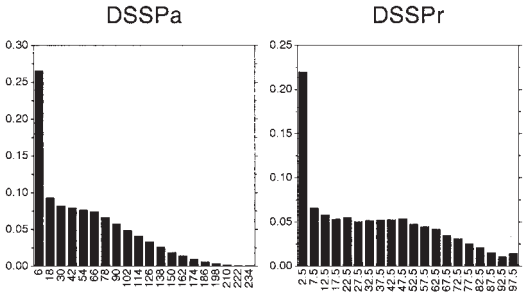

Fig. 5. Histograms of the DSSPa (left) and DSSPr (right) values. The bin widths are 6 for DSSPa and 2.5 for DSSPr. The center of each bin is indicated on the x-axis.

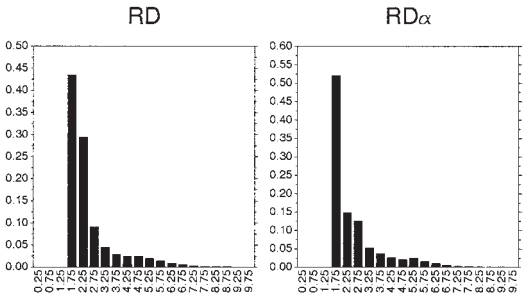

Fig. 6. Histograms of the RD (left) and RDα (right) values. The bin widths are 0.25 in both cases. The center of each bin is indicated on the x-axis.

The HSEα and HSEβ histograms (Fig. 2) are quite similar, with the exception of a peak for HSEαu = 0 that is absent for HSEβu. Apparently, the Cα-pCβ vector more often points in a direction with a near-zero Cα count than the Cα-Cβ vector. As is the case for HSEβ, the CN is obtained as the sum of HSEαu and HSEαd. Both HSEα and HSEβ can be efficiently implemented using elementary vector and matrix computations (see the sections "Calculation of HSEα and HSEβ" and "Implementation," above).

Comparison With Other Solvent-Exposure Measures

We constructed histograms for ASA, rASA, RD, and the atom depth of the Cα atom (called RDα) for the same residues as in the section "The HSEβ Measure" above in order to compare them with the HSEα and HSEβ histograms. In addition, we also determined the correlation between the various solvent-exposure measures. In this and the following sections, we use DSSPa and DSSPr to refer to the ASA and rASA values as calculated by DSSP.[27]

Both the DSSPr/DSSPa (Fig. 5) and RD/RDα (Fig. 6) histograms show a distinct peak: for DSSPa and DSSPr at minimum solvent exposure, and for RD and RDα at maximum solvent exposure. Strikingly, RD/RDα and ASA/rASA give two different views of solvent exposure: according to the former most residues are exposed, while to the latter most residues are buried! The HSE measures give a more balanced picture, with only a relatively small peak on the exposed side for HSEαu (Fig. 2).

To obtain an idea of the interdependencies of the various measures, we calculated the correlation coefficient between all measure pairs (Table I). HSEα and HSEβ are strongly correlated, which is expected since HSEα is meant as an approximation of HSEβ.

More surprising is the lack of correlation between HSEd and HSEu (for both HSEα and HSEβ), meaning that the number of atoms in the upper half-sphere is uncorrelated with the number of atoms in the lower half-sphere.

As expected, RD and DSSPa are most strongly correlated with RDα and DSSPr, respectively. Of all other measures, HSEαu and HSEβu show the best correlation with DSSPr. Similarly, HSEαu, HSEβu and CN (which is simply the sum of HSEu and HSEd) show the best correlation with RD.

Hence, based on these correlations, one might hope that the HSE measure indeed combines the best features of both RD and rASA. In the next sections, we will show that this is indeed to a great extent the case.

Sensitivity

Here we compare the sensitivity of the various measures. By "sensitivity," we mean the capability of measuring a wide range of solvent-exposure conditions (i.e. fully buried, side chain exposed,. . .) in a meaningful and informative way.

In order to illustrate the properties of HSE compared to CN, DSSPr and RD, we calculated these four measures for all residues of a monomer of FRIL (pdb identifier 1QMO), a sugar-binding protein from the legume lectin family.[31,32] The FRIL monomer is a β-sandwich consisting of three layers: an exposed front β-sheet, a solvent-shielded back β-sheet, and a layer of loops that pack against the back β-sheet.

Figures 7 and 8 show the results as color-coded cartoon representations of the FRIL structure (PDB files of the FRIL monomer with the calculated solvent-exposure measures in the temperature factor field are available; see the section "Supplementary Data" above).

The HSEβu measure (Fig. 7, top) readily distinguishes the residues with exposed and buried side chains in the front β-sheet. In addition, it also identifies residues in the back β-sheet whose side chains are deeply buried.

The HSEβd measure (Fig. 7, middle) provides information that is complementary to HSEβu. For residues in the front β-sheet with their side chains towards the solvent, the HSEβd values are high (main chain buried), while the HSEβu values are low (side chain exposed). Residues in the front sheet with their side chains pointing towards the hydrophobic core have high HSEβu (side chain buried) and low HSEβd values (main chain exposed). The residues in the fully buried back β-sheet have high HSEβu and HSEβd values (fully buried).

The CN measure (Fig. 7, bottom) distinguishes readily between deeply buried, intermediate, and exposed residues, but does not capture the characteristic pattern of alternating buried/exposed positions in the solvent exposed front β-sheet. The CN measure is essentially independent of the side-chain orientation: the number of Cα atoms

TABLE I. Correlation Coefficients Between the Various Exposure Measures

	HSEαu	HSEβu	HSEαd	HSEβd	CN	DSSPa	DSSPr	RD	RDα
HSEαu	1.00	0.88	0.03	0.22	0.82	−0.75	−0.82	0.64	0.59
HSEβu		1.00	0.15	0.03	0.80	−0.75	−0.82	0.63	0.53
HSEαd			1.00	0.79	0.59	−0.14	−0.17	0.32	0.32
HSEβd				1.00	0.63	−0.17	−0.20	0.34	0.40
CN					1.00	−0.68	−0.76	0.70	0.66
DSSPa						1.00	0.93	−0.54	−0.47
DSSPr							1.00	−0.60	−0.52
RD								1.00	0.94
RDα									1.00

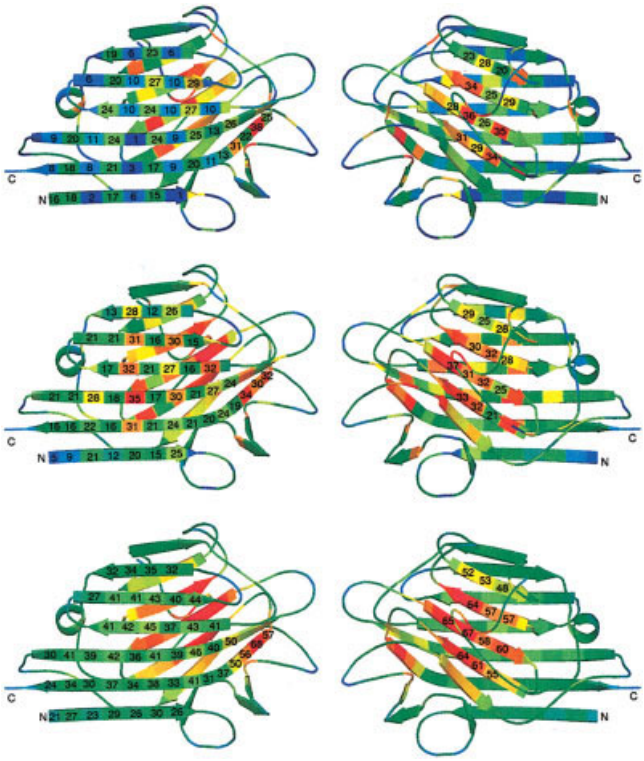

Fig. 7. Solvent exposure of FRIL's β-sheets as measured by HSEβu (top), HSEβd (middle) and CN (bottom). The left view shows the front sheet, and the right view shows the back sheet and the loop layer on top of it. The residue's colors vary from red (buried) over yellow and green to blue (exposed). The labels on the residues represent the actual values of the measures. The labels "N" and "C" indicate the N- and C-terminus. Figures 7 and 8 were made with PyMol (http://pymol.sourceforge.net/).

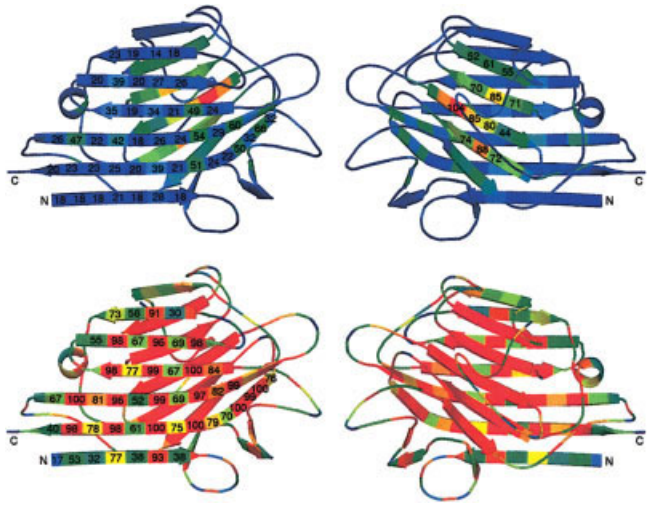

Fig. 8. Solvent exposure of FRIL's β-sheets as measured by RD (top) and percentage buried surface (1-DSSPr, bottom). The labels on the residues represent the actual values of the measures. In the case of RD, we used 10 × RD rounded to the closest integer. The unlabeled residues of the back sheet (right) in the bottom figure are all completely buried according to DSSP (1-DSSPr > 97%). Other details as in Figure 7. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

within a sphere is not sufficient to distinguish the buried from the exposed residues in a β-sheet.

The RD measure (Fig. 8, top) clearly does not capture this exposed/buried pattern either, since residues in the front β-sheet mostly show up as exposed. On the other hand, RD readily identifies deeply buried residues, similarly to the CN measure. This is expected, since recognizing deeply buried residues was one of the original goals of RD.

The DSSPr measure (Fig. 8, bottom) is much better at identifying the buried/exposed pattern in the front β-sheet. However, DSSPr does not distinguish shallowly from deeply buried residues. The buried residues in the front β-sheet, which are close to the surface, have the same value as the deeply buried residues in the solvent shielded back β-sheet.

It is clear that the conclusions from the inspections of the histograms of HSE, ASA, and RD are confirmed in Figure 8: according to the RD measure most residues are exposed, while according to the DSSPr measure most residues are buried. The HSEβ measure (Fig. 7, top and middle) shows a more balanced picture, without overemphasizing exposure or burial.

Unlike any of the other solvent-exposure measures, HSEβ adopts meaningful and informative values for a wide range of solvent-exposure conditions. In this respect, its information content combines that of CN, RD, and ASA. The conclusions for the HSEα measure are essentially identical (see the section "Supplementary Data" above).

Amino Acid Dependency

One of the advantages of the HSE measure is that it is easier to compare HSE values for amino acids of different size in a relevant way. This is because HSE,

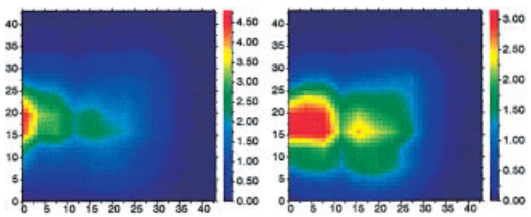

Fig. 9. Two-dimensional histograms of HSEα (left) and HSEβ (right) values for all residues. In both cases, HSEu is along the x-axis and HSEd is along the y-axis. The values of the color legend to the right refer to the percentage of residues per bin. The bin width is two. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

like the CN measure, describes a residue's environment, rather than a quantity that is related to the residue's size (as in the case of RD and ASA based measures). In this section, we compare the HSE distributions for the 20 amino acids.

For the calculation of the HSEα and HSEβ histograms for the various amino acids (Figs. 9–11), we used the same dataset as in the section "The HSEβ Measure" above. Figure 9 shows the histograms of the HSEβ and HSEα values for all residues.

The HSEβ and HSEα histograms of the aliphatic residues (Ala, Ile, Leu, Val; Figs. 10—11, top row) point out an interesting difference between Ala and the others. As expected, Leu and Ile produce very similar histograms, with a

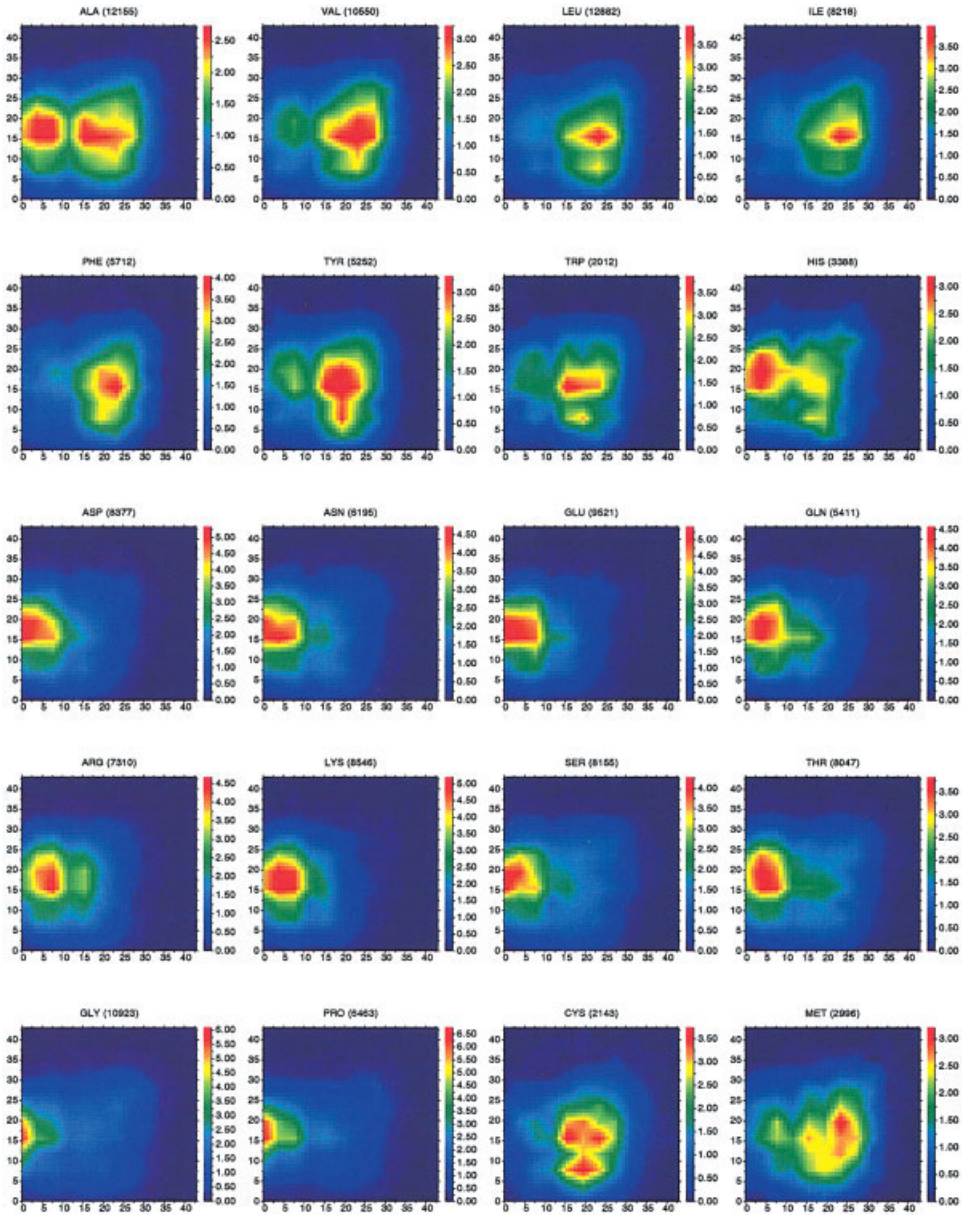

Fig. 10. Two-dimensional histograms of HSEβ for all 20 amino acids. The number of residues used in the construction of the histogram is shown in parentheses at the top. Other details as in Figure 9.

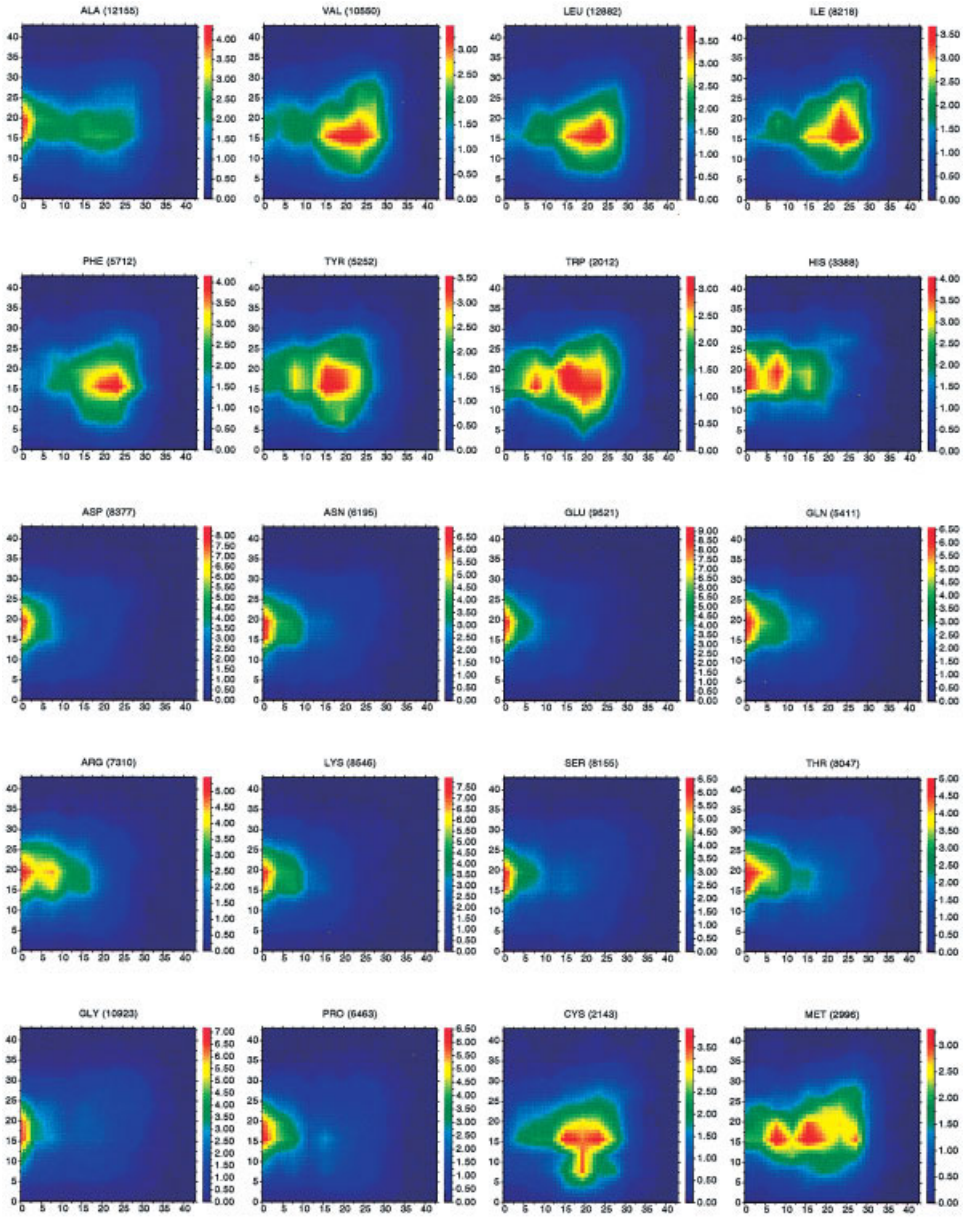

Fig. 11. Two-dimensional histograms of HSEα for all 20 amino acids. Details as in Figure 10. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

preference for high values of HSEu. On the other hand, Ala has access to a much broader range of HSEu values, in contrast to Val, Ile and Leu. Hence, it is quite clear that Ala stands apart from the other aliphatic residues. Another striking observation is that Ala's histogram strongly resembles the overall histogram (Fig. 9). In this view, Ala seems to be the amino acid with "average" behavior with respect to solvent exposure as measured by HSEβ.

The HSEβ and HSEα histograms of the aromatic residues and His (Figs. 10–11, second row) clearly show that His stands apart from Phe, Trp and Tyr. His has a clear preference for lower HSEu values, but also has access to higher HSEu values. His is thus confirmed as an amino acid with a dual hydrophilic/hydrophobic character.

The hydrophilic residues have very similar histograms (Figs. 10–11, third and fourth row). They have a clear preference for low HSEu values, and adopt a more narrow range of HSEd values than the hydrophobic residues.

Gly and Pro both clearly emerge as exposed residues (Figs. 10–11, bottom row), in tune with the fact that they often appear in exposed coil regions. Gly and Pro prefer lower HSEβu values than the classical hydrophylic residues, while the histograms of Gly, Pro, and the hydrophylic residues are very similar for HSEα.

Cys prefers high HSEu values, combined with a wide range of HSEd values (Figs. 10–11, bottom row). Met can adopt a range of HSEu and HSEd values. Overall the histograms of Cys and especially Met are substan-

TABLE II. Conservation of the Various Solvent Exposure Measures[†]

HSEuα	HSEuβ	HSEdα	HSEdβ	CN	DSSPa	DSSPr	RD	RDα
0.71	0.69	0.61	0.64	0.72	0.53	0.61	0.62	0.58

[†]Calculated as a correlation coefficient, see text.

tially different from those of the other hydrophobic amino acids.

Despite the fact that HSEu shows essentially no correlation with HSEd (see the section "Comparison With Other Solvent-Exposure Measures"), the 2D histograms discussed here show that the different amino acids do have strong preferences for specific (HSEu, HSEd) ranges. Solely based on the shape of their HSE histograms, the amino acids can be subdivided in the following four groups: 1. side chain buried (Cys, Ile, Leu, Met, Phe, Trp, Tyr, Val), 2. side chain exposed (Arg, Asn, Asp, Gln, Glu, Lys, Ser, Thr, Gly, Pro), 3. Ala and 4. His. Within the side chain buried group, Cys and Met differ most from the average group profile. In general, the histogram profiles are more uniform for the side chain exposed than for the side chain buried residues.

Conservation

Following Rost and Sander,[17] we evaluated the conservation of the different solvent-exposure measures by calculating the correlation coefficient of their values for equivalent residues in a dataset of related structures. We used a database that consists of superimposed structures with the same fold, but without a probable common evolutionary origin (see Methods). Hence, this dataset would be considered "hard" from a structure prediction point of view. The correlation coefficients for the various measures are shown in Table II.

Three measures based on counting Cα neighbors (HSEαu, HSEβu and CN) are most conserved, with values around 0.70. The DSSPr, RD and RDα measures all show substantially lower correlation coefficients, with values around 0.60. The correlation coefficient for DSSPr reported by Rost and Sander (0.77) is considerably higher, probably due to the fact that it was calculated from a dataset of homologous protein pairs. Finally, DSSPa is least conserved (0.53), which is expected since this measure is not corrected for amino acid size. Interestingly, HSEu is more conserved than HSEd, probably due to conserved side chain dependent interactions. HSEα and HSEβ are essentially equally conserved.

Correlation With the Stability of Mutants

Accessible surface area and (as recently shown) residue depth[18,22,35–37] are correlated with the stability of mutants. For cavity creating mutants of hydrophobic residues, the change in protein stability (as measured by $\Delta\Delta G$) is correlated with the change in buried accessible surface area between wild type and mutant, and with the sum of the atom depths of the atoms that are deleted upon mutation. In addition, the CN is correlated to the $\Delta\Delta G$ values of Ile/Leu/Val to Ala mutants.[36,38]

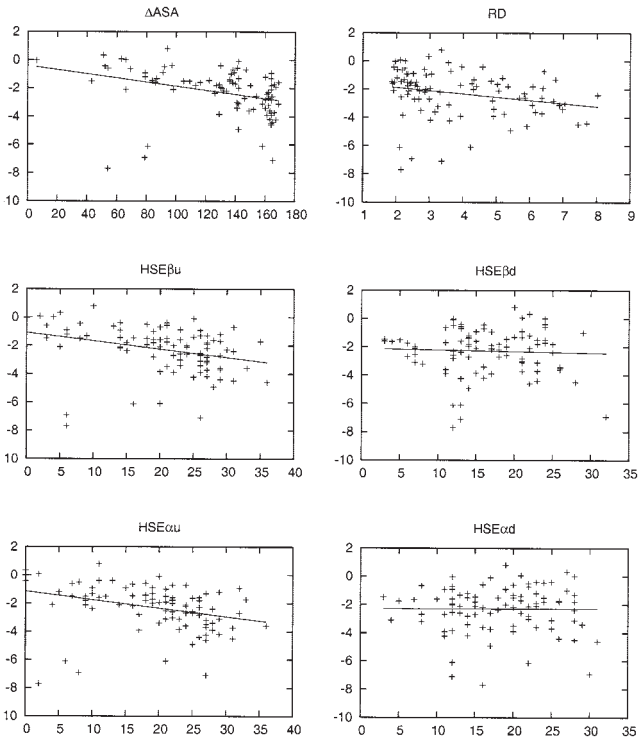

Fig. 12. $\Delta\Delta G$ (y-axis, kcal/mol) versus different measures of solvent exposure (x-axis). The line is the linear least-squares fit to the data (including the $\Delta\Delta G$ values below −6 kcal/mol).

We investigated the correlation between HSE and protein stability using 91 $\Delta\Delta G$ values of Val/Ile/Leu to Ala point mutants, and compared it with the correlation using RD, CN and loss of solvent exposed surface area (ΔASA). Figure 12 shows plots of ΔASA, RD, and HSE versus $\Delta\Delta G$, together with the linear least-squares fit to the data. Table III shows the correlation coefficients between the various measures and $\Delta\Delta G$.

HSEαu, HSEβu, CN and ΔASA show about the same correlation with $\Delta\Delta G$. RD correlates significantly worse, while HSEαd and HSEβd are essentially uncorrelated. The low correlation coefficients of the HSEd measure shows again that the half-sphere construction apparently separates two regions around a residue that are profoundly different, this both in geometric and energetic terms.

In all plots, the same five mutants with $\Delta\Delta G$ below −6.0 kcal/mol appeared to be outliers. Their $\Delta\Delta G$ values are also considerable lower than what is typical.[36,38] Hence, we also calculated the correlation coefficients with these values omitted (Table III). For all measures, except HSEβd and HSEαd which remain uncorrelated, the correlation increased significantly. Both HSEαu and HSEβu values now correlate significantly better with $\Delta\Delta G$ than the CN.

TABLE III. Correlation Coefficients of the Various Solvent Exposure Measures and ΔΔ *G* of 91 Ile/Leu/Val to Ala Point Mutants for All the Data (Top Row) and with Five Outliers (ΔΔ *G* < −6 kcal/mol) Removed (Bottom Row)

	HSEβu	HSEβd	HSEαu	HSEαd	CN	RD	ΔASA
All data	−0.30	−0.04	−0.31	0.00	−0.29	−0.23	−0.34
> −6.0 kcal/mol	−0.58	−0.06	−0.62	−0.01	−0.54	−0.45	−0.62

HSEαu and HSEβu have about the same correlation coefficient as ΔASA (approximately −0.60). This is surprising, since HSEα solely uses information on Cα positions, while ΔASA requires a full atom model.

In conclusion, HSEαu, HSEβu and ΔASA correlate equally well with ΔΔ*G*, with a correlation coefficient of approximately −0.60. The CN and RD measures correlate significantly less, with correlation coefficients of −0.54 and −0.45 respectively. HSEαd and HSEβd, which measure the number of neighboring residues in the direction opposite the side chain, are essentially uncorrelated.

The plots in Figure 12 also illustrate a feature that was already mentioned in the section "Comparison With Other Solvent-Exposure Measures" above: ASA tends to measure residues as buried, while RD tends to measure residues as exposed. In contrast, HSEα and HSEβ values show a more uniform distribution of values.

CONCLUSIONS

We have shown that HSE fulfills the demands outlined in the introduction: it is easy to implement and (in principle) fast to compute, deals with a wide range of solvent-exposure conditions, is well correlated with protein stability and is considerably more conserved than most other measures. Various examples also show that HSE has many features that are desirable in a solvent-exposure measure (for example, capturing the alternating exposed/buried pattern in a β-sheet, distinguishing between exposed, partially buried and deeply buried residues), and makes relevant comparison possible between amino acids regardless of differences in size (see the section "Amino Acid Dependency" above).

Rost and Sander[17] suggested the possibility of developing a new descriptor of solvent exposure that is better conserved than the classical rASA measure. This hypothetical measure would obviously benefit solvent-exposure prediction methods, on the condition that it is also at least as "informative" as the rASA measure. Based on the results described here, the HSE measure seems to be the best choice for use in these methods, since it is (comparatively) well conserved and more "informative" than the CN, RD, and rASA measures.

Analysis of histograms of HSE values (Fig. 2) and the correlation with the stability of mutants prove that the half-sphere construction separates two fundamentally different regions around an amino acid, both in geometric and energetic terms.

The two-dimensional HSE histograms (Figs. 9–11) show a high dependence on amino acid type. Depending on the overall shape of the histograms, the amino acids can be subdivided in four groups: 1. side chain buried, 2. side chain exposed, 3. Ala, and 4. His. Ala emerges as the "average" residue, unlike the other aliphatic residues.

Importantly, calculation of the HSE does not require a full atom model. Hence, it can be used with simplified protein models, for example Cα-only models, or models that use the Cα position plus a single center representing the side chain. The use of these models has become very wide spread in fold recognition, structure prediction and protein folding simulations.[8,10] The HSE approach allows these methods to deal with solvent exposure in a more sophisticated way than current methods (typically based on the CN[2]), without imposing a large speed penalty.

The fact that HSE does not depend on full atomic detail is of course also its weakness. In the analysis of mutants or ligand binding for example, one often wants to analyze subtle features on an atomic scale. In those cases, rASA and RD are definitely the preferred methods.

We suggest that HSE is an excellent target for solvent-exposure prediction. Most methods developed since the pioneering work of Rost and Sander[17] make use of the rASA, typically calculated by DSSP.[39–44] Few studies exist that try to predict solvent-exposure measures other than rASA.

The method of Pollastri et al.[21] predicts the coordination number, using various radii between 6 and 12 Å. Karchin et al.[33] evaluated a number of coordination number variants with respect to conservation and predictability. In general, measures based on neighborhood counts were both more conserved and more predictable. The most conserved, most predictable, and best performing measures were based on counting neighbors around the Cβ atom (using radii of 16 and 14 Å). This can be understood in view of the results presented in this article: using the Cβ as a center instead of the Cα atom, the CN becomes biased towards the HSEβu measure.

In this light and given the properties of the HSE measure discussed in this article, it is not unlikely that a method based on HSEβ or HSEα would perform even better. We are planning to explore the use of HSE for the prediction of solvent exposure and protein structure in general in the future.

ACKNOWLEDGMENTS

This work was supported by the Lundbeckfond (http://www.lundbeckfonden.dk/). I thank Dr. Bente Vestergaard and Wouter Boomsma for critically reading the article and the Biopython community and developers for their efforts.

REFERENCES

1. Gilis D, Rooman M. Predicting protein stability changes upon mutation using database-derived potentials: solvent accessibility

determines the importance of local versus non-local interactions along the sequence. J Mol Biol 1997;272:276–290.
2. Simons K, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 1997;268:209–225.
3. Rice D, Eisenberg D. A 3D-1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. J Mol Biol 1997;267:1026–1038.
4. Rost B. Better 1D predictions by experts with machines. Proteins 1997;Suppl 1:192–197.
5. Shi J, Blundell T, Mizuguchi K. FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J Mol Biol 2001;310: 243–257.
6. Guerois R, Nielsen J, Serrano L. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. J Mol Biol 2002;320:369–387.
7. Bartlett G, Porter C, Borkakoti N, Thornton J. Analysis of catalytic residues in enzyme active sites. J Mol Biol 2002;324:105–121.
8. Head-Gordon T, Brown S. Minimalist models for protein folding and design. Curr Opin Struct Biol 2003;13:160–167.
9. Rost B, Liu J. The PredictProtein server. Nucleic Acids Res 2003;31:3300–3304.
10. Buchete N, Straub J, Thirumalai D. Development of novel statistical potentials for protein fold recognition. Curr Opin Struct Biol 2004;14:225–232.
11. Lee B, Richards F. The interpretation of protein structures: estimation of static accessibility. J Mol Biol 1971;55:379–400.
12. Greer J, Bush B. Macromolecular shape and surface maps by solvent exclusion. Proc Natl Acad Sci USA 1978;75:303–307.
13. Connolly M. Solvent-accessible surfaces of proteins and nucleic acids. Science 1983;221:709–713.
14. Sanner M, Olson A, Spehner J. Reduced surface: an efficient way to compute molecular surfaces. Biopolymers 1996;38:305–320.
15. Fraternali F, Cavallo L. Parameter optimized surfaces (POPS): analysis of key interactions and conformational changes in the ribosome. Nucleic Acids Res 2002;30:2950–2960.
16. Cavallo L, Kleinjung J, Fraternali F. POPS: a fast algorithm for solvent accessible surface areas at atomic and residue level. Nucleic Acids Res 2003;31:3364–3366.
17. Rost B, Sander C. Conservation and prediction of solvent accessibility in protein families. Proteins 1994;20:216–226.
18. Chakravarty S, Varadarajan R. Residue depth: a novel parameter for the analysis of protein structure and stability. Structure Fold Des 1999;7:723–732.
19. Pintar A, Carugo O, Pongor S. Atom depth in protein structure and function. Trends Biochem Sci 2003;28:593–597.
20. Pintar A, Carugo O, Pongor S. Atom depth as a descriptor of the protein interior. Biophys J 2003;84:2553–2561.
21. Pollastri G, Baldi P, Fariselli P, Casadio R. Prediction of coordination number and relative solvent accessibility in proteins. Proteins 2002;47:142–153.
22. Matsumura M, Becktel W, Matthews B. Hydrophobic stabilization in T4 lysozyme determined directly by multiple substitutions of Ile 3. Nature 1988;334:406–410.
23. Russell R, Barton G. Structural features can be unconserved in proteins with similar folds. An analysis of side-chain to side-chain contacts secondary structure and accessibility. J Mol Biol 1994;244: 332–350.
24. Bateman A, Birney E, Durbin R, Eddy S, Howe K, Sonnhammer E. The Pfam protein families database. Nucleic Acids Res 2000;28: 263–266.
25. Hobohm U, Sander C. Enlarged representative set of protein structures. Protein Sci 1994;3:522–524.
26. Van Walle I, Lasters I, Wyns L. Align-m—a new algorithm for multiple alignment of highly divergent sequences. Bioinformatics 2004;20:1428–1435.
27. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
28. Bava K, Gromiha M, Uedaira H, Kitajima K, Sarai A. ProTherm, version 4.0: thermodynamic database for proteins and mutants. Nucleic Acids Res 2004;32 Database issue:D120–121.
29. Chapman B, Chang J. Biopython: python tools for computational biology. ACM SIGBIO Newsletter 2000;20:15–19.
30. Hamelryck T, Manderick B. PDB file parser and structure class implemented in Python. Bioinformatics 2003;19:2308–2310.
31. Loris R, Hamelryck T, Bouckaert J, Wyns L. Legume lectin structure. Biochim Biophys Acta 1998;1383:9–36.
32. Hamelryck T, Moore J, Chrispeels M, Loris R, Wyns L. The role of weak protein–protein interactions in multivalent lectin-carbohydrate binding: crystal structure of cross-linked FRIL. J Mol Biol 2000;299:875–883.
33. Karchin R, Cline M, Karplus K. Evaluation of local structure alphabets based on residue burial. Proteins 2004;55:508–518.
34. Raghunathan G, Jernigan R. Ideal architecture of residue packing and its observation in protein structures. Protein Sci 1997;6:2072–2083.
35. Serrano L, Kellis J, Cann P, Matouschek A, Fersht A. The folding of an enzyme. II. Substructure of barnase and the contribution of different interactions to protein stability. J Mol Biol 1992;224:783–804.
36. Jackson S, Moracci M, elMasry N, Johnson C, Fersht A. Effect of cavity-creating mutations in the hydrophobic core of chymotrypsin inhibitor 2. Biochemistry 1993;32:11259–11269.
37. Ratnaparkhi G, Varadarajan R. Thermodynamic and structural studies of cavity formation in proteins suggest that loss of packing interactions rather than the hydrophobic effect dominates the observed energetics. Biochemistry 2000;39:12365–12374.
38. Otzen D, Rheinnecker M, Fersht A. Structural factors contributing to the hydrophobic effect: the partly exposed hydrophobic minicore in chymotrypsin inhibitor 2. Biochemistry 1995;34:13051–13058.
39. Richardson C, Barlow D. The bottom line for prediction of residue solvent accessibility. Protein Eng 1999;12:1051–1054.
40. Yuan Z, Burrage K, Mattick J. Prediction of protein solvent accessibility using support vector machines. Proteins 2002;48:566–570.
41. Ahmad S, Gromiha M. NETASA: neural network based prediction of solvent accessibility. Bioinformatics 2002;18:819–824.
42. Gianese G, Bossa F, Pascarella S. Improvement in prediction of solvent accessibility by probability profiles. Protein Eng 2003;16: 987–992.
43. Ahmad S, Gromiha M, Sarai A. RVP-net: online prediction of real valued accessible surface area of proteins from single sequences. Bioinformatics 2003;19:1849–1851.
44. Kim H, Park H. Prediction of protein relative solvent accessibility with support vector machines and long-range interaction 3D local descriptor. Proteins 2004;54:557–562.