brief communication

Excluded volume approximation to protein-solvent interaction
The solvent contact model

François Colonna-Cesari* and Chris Sander‡
*Laboratoire d'Enzymologie Physico-Chimique et Moleculaire, Universite de Paris Sud, F-91405 Orsay, France; and
‡BIOcomputing Programme, European Molecular Biology Laboratory, D-6900 Heidelberg, Federal Republic of Germany

ABSTRACT Important properties of globular proteins, such as the stability of its folded state, depend sensitively on interactions with solvent molecules. Existing methods for estimating these interactions, such as the geometrical surface model, are either physically misleading or too time consuming to be applied routinely in energy calculations. As an alternative, we derive here a simple model for the interactions between protein atoms and solvent atoms in the first hydration layer, the solvent contact model, based on the conservation of the total number of atomic contacts, a consequence of the excluded-volume effect. The model has the conceptual advantage that protein–protein contacts and protein–solvent contacts are treated in the same language and the technical advantage that the solvent term becomes a particularly simple function of interatomic distances. The model allows rapid calculation of any physical property that depends only on the number and type of protein–solvent nearest-neighbor contacts. We propose use of the method in the calculation of protein solvation energies, conformational energy calculations, and molecular dynamics simulations.

INTRODUCTION

Protein-solvent interaction

The protein folding process can be viewed as competition between protein–protein and protein–water contacts. As protein conformation changes, the contacts a protein atom makes with other protein atoms are replaced by contacts with solvent molecules and vice versa. Here our goal is to calculate protein–solvent contacts for an arbitrary conformation of a globular protein as an estimate of protein–solvent interaction energy.

The principal difficulty lies in the uncertainty of the atomic positions of water molecules. Water molecules are more mobile than protein atoms, as they are not covalently attached to the polymer. Although the positions of tightly bound water molecules are known in highly resolved crystal structures of some proteins, we do not yet have a reliable method for calculating the positions of specific water molecules for a given protein conformation; nor do we have a good method for calculating the time-average interaction between a protein and water other than by explicitly averaging over simulated molecular dynamics trajectories.

We propose here to circumvent the problem of unknown water positions by assuming that all empty space near a protein is uniformly filled by solvent: "Non-protein space is solvent space." We are interested in an estimate of the time average of protein–solvent interactions in terms of nearest-neighbor contacts.

MODEL

Solvent contact model

The basic model is very simple (Fig. 1): Assume that a protein atom has a constant total number of nearest-neighbor contacts (volume conservation, excluded volume effect); partition the total number of nearest-neighbor contacts (C) into contacts with other protein atoms (CP) and contacts with solvent (CW); calculate solvent contacts as the difference between the total number of contacts and the number of protein contacts:

$$\alpha \cdot CW[\text{solvent}] = (C[\text{total}] - CP[\text{protein}]), \qquad (1)$$

where the scale factor α reflects the different density of protein atoms and water molecules.

There are a number of different ways of implementing the model, e.g., by calibration on crystallographically determined water positions. In this paper we present a simple, first implementation calibrated on accessible surface area values.

IMPLEMENTATION

Calibration of contact counts

To quantitate the model, we need to define the *protein* contact counts, CP, the *solvent* contact counts, CW, and calibrate their relative scale α. We also need to determine

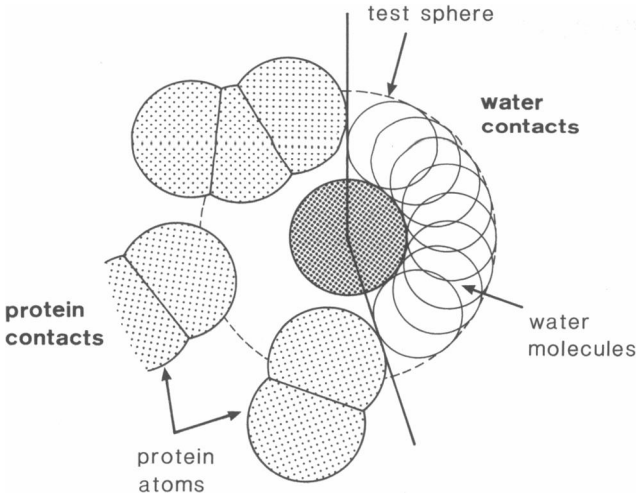

FIGURE 1 Nearest-neighbor sphere around a protein atom: The central atom can make atomic contacts either with other protein atoms or with water molecules, depending on the local conformation of the protein. If the total number of contacts remains approximately constant as the protein conformation is varied, the number of protein–water contacts can be estimated as the complement of protein–protein contacts, i.e., by a simple sum of terms that depend only on atomic distances within the protein.

C(total), the total number of contacts possible for an atom or residue. In the current implementation of the model, we start with the assumption of dense packing of nearest neighbors around a protein atom (nearest neighbors are either covalently linked protein atoms, protein atoms not covalently linked, and/or solvent molecules). We then calculate protein contacts using distance criteria or simple energy estimates, calculate solvent contacts in terms of accessible surface area, and determine the relative scale between protein and solvent contacts, α, using data on known protein structures. The factor α expresses how many protein contacts can replace one water contact.

There are a number of ways in which the concept of contacts can be made quantitative. Here, for protein atom i, we count *protein–protein* contacts (CP) in a shell of thickness of one water diameter outside of the hard sphere radius of the atom, using a rectangular well with linear edge. More precisely, the contact strenght CP(ij) between atom i and atom j is equal to 1.0 if the atoms overlap or just touch, i.e., if the atom–atom distance $d_{ij} \leq r_i + r_j$ (where r_i and r_j are hard sphere or van-der-Waals radii); the contact strength then decreases linearly down to zero with distance until a water molecule can just fit between the two atoms, i.e., until $d_{ij} = r_i + r_j + 2r(H_2O)$. Instead of this linear decrease a sharp cutoff or a Gaussian falloff would also be reasonable. The contact count CP(i) for atom i is the sum of CP(ij) over all neighbors j.

The number of *protein–water* contacts (CW) of protein atom i is defined here as equal to the number of water molecules (NW) associated with atom i in the first hydration shell. The number of such water molecules is estimated by relating V, the first hydration shell volume associated with the atom, to V_0, the volume occupied by a single water molecule. In terms of the surface area S and the thickness t of the shell:

$$CW = NW \approx \frac{V}{V_0} = \frac{S \cdot t}{V_0} = \frac{S}{V_0^{2/3}} = \frac{S}{9.65}, \qquad (2)$$

where the volume of one water molecule is $V_0 = 18 \cdot 10^{24}$ [Å^3/mol]/$6.023 \cdot 10^{23}$ [molecules/mol] = 30 Å^3, where the thickness t is chosen such that $t^3 = V_0$ and where the surface area S is taken to be the solvent accessible surface area in angstroms squared (Lee and Richards, 1971) calculated by numerical integration with a water probe radius of 1.4 Å as in DSSP (Kabsch and Sander, 1983).

A simple way of determining the relative scale of CP and CW is a scatter plot of protein and water contacts for side chains of many residues in known protein structures. The negative slope in the plot (Fig. 2) reflects the fact that a side chain makes either more protein or more water contacts, depending on its environment, and that the sum of the two is approximately constant. This is consistent with the view that the space around an atom is occupied approximately uniformly by protein or solvent atoms and that internal cavities large enough to accomodate water molecules are not empty. The two axis intercepts represent the extreme cases: a side chain totally buried in the interior ($CW = 0$, no water contacts) and a side chain totally exposed ($CP = 0$, no contacts with other side chains, as in an extended peptide G-G-X-G-G). The intercepts provide a best estimate for the approximately conserved total number of contacts possible for one side chain, C[total] (Fig. 2). The scale factor α is the ratio of the two intercepts and quantifies how may protein–protein contacts are equivalent to one protein–water contact. When averaged over all side chain types weighted with the number of atoms per side chain in 70 proteins, its numerical value is:

$$\alpha = \left\langle \frac{CP(CW = 0)}{CW(CP = 0)} \right\rangle = 3.0. \qquad (3)$$

The scatter around the straight line comes from two sources: (*a*) nonuniform packing, as a result of varying side chain conformation of the central residue, varying type of contacting side chain (e.g., rings/aliphatics), varying volume occupation by neighboring main chain or inaccurate coordinates; and (*b*) arbitrariness in distance cutoff. As our current interest is in illustrating the main idea of the contact model, it is left to future work to refine details.

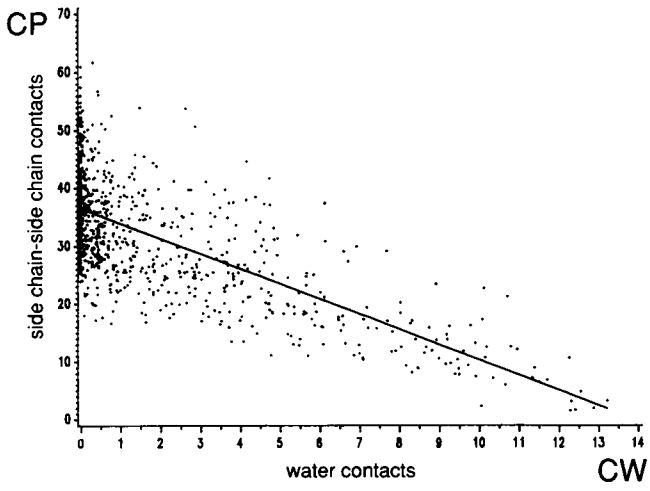

FIGURE 2 Protein/water linear regression: conservation of the total number of contacts for an amino acid side chain (here: ILE) is as good as the straight line fit to this scatter plot of intraprotein side chain contacts (*CP*) versus solvent contacts (*CW*). Each point represents a side chain in a protein of known structure. The deviations from a straight line are caused in part by the fact that surface area, used here to estimate the number of contacting water molecules, is only an approximation to volume occupation. The approximate conservation of the total number of contacts can be used to calibrate the relative strength of intraprotein contacts versus solvent contacts as the ratio of the *CP*-axis and *CW*-axis intercepts (here, for Ile, $CP[CW = 0] = 36.7$, $CW[CP = 0] = 13.9$). The partcular numerical value of this ratio depends on the functional form used to calculate contacts, here a square/linear well, and on atomic parameters (here, all side chain atomic radii 1.8 Å as in Kabsch and Sander [1983]; radius of a water molecule 1.4 Å). Averaged over all side chain types and 70 proteins we get $\alpha = 3.0 \pm 0.2$ ($\pm$ standard error in the slope in regression analysis). Protein Data Bank identifiers of the 70 protein structures used are (Bernstein et. al., 1977):

1ABP, 2ACT, 4ADH, 2ADK, 2ALP,
4APE, 2APP, 1APR, 4ATC, 1AZU,
2B5C, 1BP2, 3C2C, 1CAC, 7CAT,
3CNA, 5CPA, 1CPV, 1CRN, 1CTX,
3CYT, 3DFR, 1ECD, 1EST, 3FAB,
1FDX, 3FXC, 4FXN, 2GCH, 1GCN,
1GPD, 2GRS, 2HHB, 1HIP, 1HMQ,
1INS, 4LDH, 1LH1, 7LYZ, 1LZM,
1MBN, 2MDH, 2MHB, 1MLT, 1NXB,
1OVO, 2PAB, 8PAP, 1PCY, 3PGM,
1PPT, 4PTI, 2PTN, 1REI, 1RHD,
2RHE, 1RNS, 4RXN, 1SBT, 2SGA,
2SNS, 2SOD, 2SSI, 2TAA, 1TIM,
3TLN, 3WGA, 351C, 155C, 156B.

APPLICATION

Estimating the number of water molecules in the first hydration layer

The contact model provides a way of calculating protein–solvent contacts that is as simple as the calculation of protein–protein contacts; the calculation requires nothing more than a simple sum over terms depending only on interatomic distances within the protein. As an example, we use the contact model to calculate an estimate for the total number of water molecules in the first hydration shell of 70 proteins and compare it with an estimate of the same quantity based on calculation of the solvent accessible surface area. The two estimates agree with an average deviation of 8.3%; the correlation coefficient is 0.98 (Fig. 3). Comparison with the precise value of the number of contacting water molecules in the first hydration shell is not possible, as it has not been measured experimentally. The high correlation coefficient demonstrates that both the surface model and the contact model give similar numerical estimates of water contacts.

DICUSSION

Surface model versus contact model

For calculation of surface area, excellent and now classical algorithms have been available for over a decade, either by exact numerical integration (Lee and Richards, 1971; Shrake and Rupley, 1973), by statistical approximation (Wodak and Janin, 1980) or by analytical formulae (Richmond, 1984). Surface area algorithms have been extremely useful in visualizing the solvent accessible surface area in molecular graphics (e.g., Connolly, 1983) and in describing the extent to which solvent exposure decreases during protein folding and during protein–protein association.

The surface model has also been used in making quantitative *estimates of free energy differences* (e.g., Eisenberg and McLachlan, 1986; Cohen et al., 1982; Richmond and Richards, 1978; Chothia and Janin, 1975; Finney, 1975; and reviews of Chothia, 1984, and Richards, 1977). Agreement with experiment in estimating changes in the free energy differences of unfolding due to mutating nonpolar side chains in the protein interior has been good in some cases (T4 lysosyme, Matsumara et al., 1988), less good in others (barnase, Kellis et al., 1989). In any event, there is no *a priori* reason to believe that the description of the molecular surface in terms geometrical contours (in units of angstroms squared) provides in general an adequate approximation to free energies of solvation (in units of kilocalories per mole) and indeed there has been considerable debate on this point (e.g., Tanford, 1979; Karplus, 1980; Gilson and Honig, 1988).

We have therefore explored the contact model as an alternative to the surface model with the ultimate goal of developing better approximations to the calculation of *protein–solvent energies,* a difficult and incompletely solved problem, not to improve the calculation of *protein*

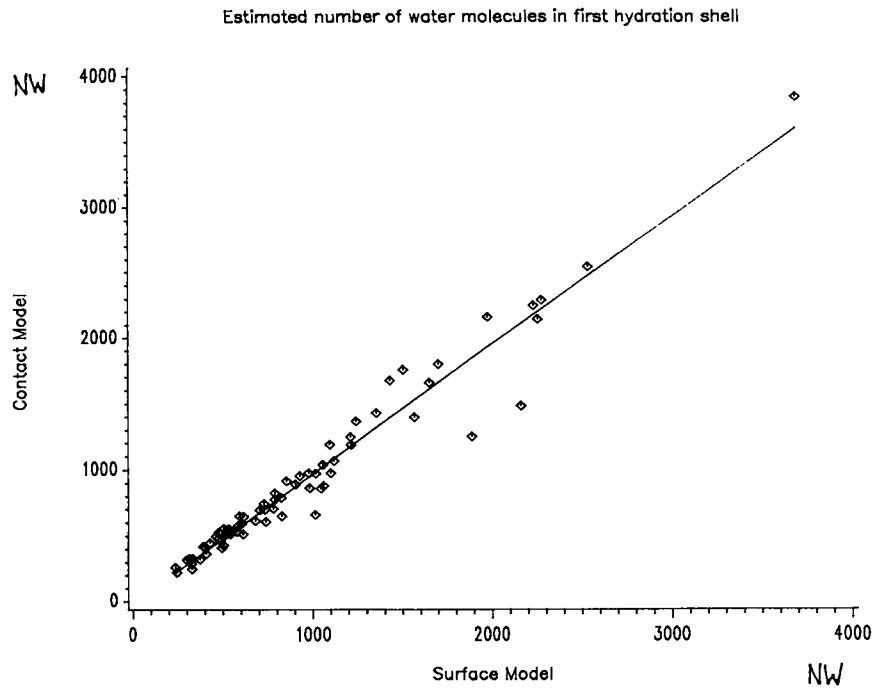

FIGURE 3 Correlation between two approximations to the total number (*NW*) of water molecules in the first hydration shell of proteins: *NW* is estimated either as the maximum number of contacts minus the number of intraprotein contacts (contact model) or as proportional to the classical solvent accessible surface area (surface model). For a typical 300-residue protein, the contact model calculations took 30 s of CPU time compared to 270 s for the surface model, calculated as in Kabsch and Sander (1983). The contact model calculation is more efficient if the goal is not to calculate surface area per se, but rather to develop a physically meaningful parametrization of effective interactions at the protein surface. Note that here all numbers refer to side chains only, i.e., no main chain contacts or main chain surfaces are taken into account. Parameters used in the contact model estimate of protein-water contacts: (*a*) Total number of side chain contacts C[total] for each amino acid type:

```
ALA 9.71,   ARG 70.28,  ANS 38.77,  ASP 41.76,  CYS 19.88,
GLN 48.06,  GLU 53.19,  GLY 6.24,   HIS 65.86,  ILE 36.66,
LEU 36.73,  LYS 47.57,  MET 40.40,  PHE 68.84,  PRO 29.99,
SER 19.37,  THR 28.74,  TYR 77.80,  VAL 27.98,  TRP 95.90.
```

(*b*) Scale factor to convert from solvent contact counts (number of contacting water molecules) to protein contact counts: $\alpha = 3.0$ (see Eqs. 1 and 3). Parameter used in the surface model estimate of protein–water contacts: Scale factor to convert from number of contacting water molecules to surface in $Å^2$: 9.65 (see Eq. 2).

surface areas, a solved problem. We are motivated by the consideration that in molecular physics potential energies (enthalpies) are fundamentally sums over pair interactions. Correspondingly, the basic entity in the contact model are pairs of interacting atoms characterized by the *type of atoms* involved and the exact (for protein–protein contacts) or approximate (for protein–solvent contacts) *distance* between atom centers.

The contact model has the conceptual advantage that both protein–protein and protein–solvent interactions are formulated in identical terms and that those terms relate directly to quantities used in basic molecular physics. It remains to be shown, however, that entropy of solvation, essentially a nonadditive property of the many-particle system, can be approximated adequately by a sum over pair terms.

Surface model and contact model both provide an estimate of the number of protein–water contacts and hence of the strength of protein–solvent interactions when the positions of water molecules are not fixed or not known. The question of which model provides the physically more accurate and practically more useful approximation to protein–solvent energies, including entropic contributions, remains to be answered. Perhaps detailed molecular dynamics simulation in water, averaging over many simulated water configurations at the protein surface, complemented by analysis of crystallographically determined fixed water molecule positions, will provide the answer. Work on comparing free energy estimates based on the sovent contact model with experimental values is in progress.

We have argued that the contact model is conceptually

well suited for estimates of interaction energies while the explicit calculation of solvent accessible surface points (Lee and Richards, 1971) is ideal for geometrical surface representation, especially in computer graphics visualization (Connolly, 1983).

Outlook

The solvent contact model has its main conceptual advantage in that it relies directly on the known positions of protein atoms in estimating the number of water molecules interacting with the protein, without recourse to definition and construction of a geometrical surface. We suggest the model, calibrated appropriately, can be used to approximate a number of different physical effects that depend only on the interaction of the first hydration layer with the protein. In particular, we think (but have not yet proven) that the free energy of protein–solvent interaction can be approximately quantified within the context of the model as a weighted sum over protein–solvent contacts, where for each contact the weight depends on the chemical type of the atoms involved. Such free energy terms can be made an explicitly differentiable function of atomic positions by choosing an appropriate functional form for the change of contact strenght with distance (e.g., Gaussian rather than rectangular/linear). Use of such an approximate protein–solvent interaction would correct a major deficiency of vacuum molecular dynamics and energy minimization calculations and would do so without costly simulation of explicitly positioned water molecules.

We thank the crystallographers for protein coordinates. We thank Cornelius Frömmel, Jacques Haiech, Wolfgang Kabsch, Arthur Lesk, Michael Schaefer, and Michael Scharf for interesting discussions. Part of this work was performed at the Max-Planck-Institute of Medical Research, Abteilung Biophysik, Heidelberg, FRG.

Received for publication 17 April 1989 and in final form 23 January 1990.

REFERENCES

Bernstein, F.C., T. F. Koetzle, G. J. B. Williams, E. F. Meyer, M. D. Brice, J. R. Rogers, O. Kennard, T. Shimanouchi, and Q. Tasumi. 1977. The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* 112:535–542.

Chothia, C. 1984. Principles that determine the structure of proteins. *Annu. Rev. Biochem.* 55:537–572.

Chothia, C., and J. Janin. 1975. Principles of protein-protein recognition. *Nature (Lond.).* 256:705–708.

Cohen, F. E., M. J. E. Sternberg, and W. R. Taylor. 1982. Analysis and prediction of the packing of α-helices against a β-sheet in the tertiary structure of globular proteins. *J. Mol. Biol.* 156:821–862.

Connolly, M. L. 1983. Solvent-accessible surfaces of proteins and nucleic acids. *Science (Wash. DC).* 221:709–713.

Eisenberg, D., and A. D. McLachlan. 1986. Solvation energy in protein folding and binding. *Nature (Lond.).* 319:199–203.

Finney, J. L. 1975. Volume occupation, environment and accessibility in proteins. The problems of the protein surface. *J. Mol. Biol.* 96:721–732.

Gilson, M. K., and B. Honig. 1988. Calculation of the total electrostatic energy of a macromolecular system: solvation energies, binding energies, and conformational analysis. *Proteins: Structure, Function, and Genetics.* 4:7–18.

Kabsch, W., and C. Sander. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers.* 22:2577–2637.

Karplus, M. 1980. Solvent accessibility, protein surfaces, and protein folding (Lesk, A. M., and C. Chothia). Discussion. *Biophys. J.* 32:44–47.

Kellis, J. T., Jr., K. Nyberg, and A. R. Fersht. 1989. Energetics of complementary side-chain packing in a protein hydrophobic core. *Biochemistry.* 28:4914–4922.

Lee, B., and F. M. Richards. 1971. The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* 55:379–400.

Matsumura, M., W. J. Becktel, and B. W. Matthews. 1988. Hydrophobic stabilization in T4 lysozyme determined directly by multiple substitutions of Ile 3. *Nature (Lond.).* 334:406–410.

Richards, F. M. 1977. Areas, volumes, packing, and protein structure. *Annu. Rev. Biophys. Bioeng.* 6:151–176.

Richmond, T. J. 1984. Solvent accessible surface area and excluded volume in proteins. *J. Mol. Biol.* 177:63–89.

Richmond, T. J., and F. M. Richards. 1978. Packing of α-helices: geometrical constraints and contact areas. *J. Mol. Biol.* 119:537–555.

Shrake, A., and J. A. Rupley. 1973. Environment and exposure to solvent of protein atoms. *J. Mol. Biol.* 79:351–371.

Tanford, C. 1979. Interfacial free energy and the hydrophobic effect. *Proc Natl. Acad. Sci. USA.* 76:4175–4176.

Wodak, S. J., and J. Janin. 1980. Analytical approximation to the accessible surface area of proteins. *Proc. Natl. Acad. Sci. USA.* 77:1736–1740.