e. Reference States are critical for the application of empirical potentials

Just as in thermodynamics, where specification of the standard state is crucial to the meaning of any value, here the definition of the reference state is similarly critical. In the *quasi-chemical approximation* {Miyazawa & Jernigan 1985 ID: 1010} the random mixing approximation is utilized: the number of contacts between a particular pair of species is taken to be directly proportional to their relative frequencies. Contact formation is described as if it were a chemical reaction, which is a particularly useful device for describing inter-residue interactions in proteins.

In principle, it is straightforward to develop effective contact potentials for many different reference states. The reference state should, however, incorporate as much information as possible that is specific to the application at hand. Three reference states are most common. For the first, the preference of an A-type residue for a B-type residue is compared to that of their self-interactions expressed as

$$A\bullet A + B\bullet B \rightarrow 2\ A\bullet B \qquad \text{(IV.6.9)}$$

The effective contact energy is also referred to as *self contact energy,* in view of the absence of any third molecule/residue contribution other than those from A and B directly interacting. The energy for forming the A•B contact is accounted for in this case by the difference between the energies of the terms on the right hand side and those on the left hand side (the reference state, given in [ ] in eq *IV.6.10*)

$$2\ e_{AB}'(R_C) = 2\ W_{AB}(R_C) - [W_{AA}(R_C) + W_{BB}(R_C)] \qquad \text{(IV.6.10)}$$

where the argument (R_C) indicates that interactions between pairs closer than the cutoff distance R_C are taken into consideration. We note that W_{AB} could equally be replaced by ΔW_{AB}, as the homogeneous contributions would vanish in eq *IV.6.10*. $e_{AB}'(R_C)$ values for all types of residue pairs are given in the Appendix in *Table IV.6.A1*. *Energy* values throughout this section are given in dimensionless RT units, unless otherwise stated. For this reference state, the opposite charge interactions are the most favored pairs and the hydrophobic pairs exhibit quite weak interactions; so this reference state informs us mostly about specificity.

Another more interesting reference state involves desolvation of residues A and B prior to their association, as

$$A\bullet 0 + B\bullet 0 \rightarrow A\bullet B + 0\bullet 0 \qquad \text{(IV.6.11)}$$

where '0' indicates solvent molecules. The corresponding *solvent-mediated* (indicated by superscript 0) effective contact energy is given by

$$e_{AB}^{0}(R_C) = W_{AB}(R_C) + W_{00}(R_C) - W_{A0}(R_C) - W_{B0}(R_C) \qquad \text{(IV.6.12)}$$

The solvent-residue potentials, $W_{A0}(R_C)$ and $W_{B0}(R_C)$, are determined from the number of effective solvent "0" molecules coordinating residue types A and B. $W_{A0}(R_C)$ depends on two quantities: the residue coordination numbers $\langle q_A^X(R_C)\rangle$, computed using eq *IV.6.4*, and the total coordination number $\langle q_A(R_C)\rangle$, which indirectly yields the average number of effective solvent molecules $\langle q_A^0(R_C)\rangle = \langle q_A(R_C)\rangle - \langle q_A^X(R_C)\rangle$. The total coordination number multiplied with N_A gives the total number of contacts, which residues of type A are theoretically expected to make in all structures, if fully coordinated. See eq *IV.6.4*. Using this information, one can extract the theoretical fraction, or probability, of contacts between residues of type A and solvent molecules as

$$P_{A0}(R_C) = 1 - \Sigma_B N_{AB}(R_C) / [\langle q_A(R_C)\rangle N_A] = \langle q_A^0(R_C)\rangle / \langle q_A(R_C)\rangle \qquad \text{(IV.6.13)}$$

which is substituted into the inverse Boltzmann equation

$$W_{A0}(R_C) = - RT \ln [\langle q_A^0(R_C)\rangle / \langle q_A(R_C)\rangle] + \textit{Const} \qquad \text{(IV.6.14)}$$

for calculating the residue-specific solvent-residue potentials. The constant in eq *IV.6.14* and the solvent-solvent interaction potential $W_{00}(R_C)$ contribute a constant amount to each of the contact energies $e_{AB}^0(R_C)$. These can therefore be cast into a single parameter, $W_{00}^*(R_C)$, that shifts the absolute values of the contact potentials without altering the residue-specific preferences. The $W_{00}^*(R_C) = - 3.3$ RT adopted {Bahar & Jernigan 1997 ID: 81} closely reproduces the quasi-chemical approximation results.

Examples of values of $e_{AB}^0(R_C)$ for $R_C = 6.5$ Å, termed e_{AB}^0(broad), for the 'broad' distance range $r \le R_C = 6.5$ Å are given next for some representative residue types: phenylalanine, leucine, alanine, glycine, glutamic acid, and lysine {Miyazawa & Jernigan 1996 ID: 174}.

e_{AB}^0(broad)	F	L	A	G	E	K
F	-7.26	-7.28	-4.81	-4.13	-3.56	-3.36
L		-7.37	-4.91	-4.16	-3.59	-3.37
A			-2.72	-2.31	-1.51	-1.31
G				-2.24	-1.22	-1.15
E					-0.91	-1.80
K						-0.97

The full set of values is given in *Table IV.6.A2*. Here, since full exposure to water is the reference state, quite large negative values are seen for contacts between hydrophobic residues. This corresponds to the well-known strong effect of burying hydrophobic residues to remove them from water. Thus, we can already see that the residue pairs most favored strongly usually will depend upon the reference state utilized.

In a third common reference state, the interactions on the reference side of the "reaction" are taken to be with an *average residue*, "X". This corresponds to the transition

$$A\bullet X + B\bullet X \rightarrow \quad A\bullet B + X\bullet X \qquad \text{(IV.6.15)}$$

The corresponding *residue-mediated* (indicated by superscript X) effective contact energy is

$$e_{AB}^{X}(R_C) = W_{AB}(R_C) + W_{XX}(R_C) - W_{AX}(R_C) - W_{BX}(R_C) \qquad \text{(IV.6.16)}$$

W_{AX} is the average of the W_{AB} values over the twenty different types of residue B, where the contribution of each residue pair is weighted according to its number of occurrences, and W_{xx} is found by further averaging W_{AX} over all residue types A. The residue-mediated contact potentials for the same residue types are:

e_{AB}^{X}(broad)	F	L	A	G	E	K
F	-0.29	-0.26	0.03	0.27	0.44	0.37
L		-0.30	-0.08	0.29	0.46	0.41
A			-0.13	-0.10	0.30	0.23
G				-0.41	0.21	0.01
E					0.12	-1.04
K						-0.48

See *Table IV.6.A3* for the complete set of residue-mediated contact energies in globular proteins.

Although the denatured state is usually poorly characterized, it is quite plausible that it is intermediate between complete exposure of residues to solvent and complete burial. Consequently, we can consider new contact energies in the native state formed as a weighted average over the two energies defined in equations *IV.6.12* and *IV.6.16*. Hence we define a folding potential as a mixture of two fractional contributions

$$E_{AB}(R_C) = f\, e_{AB}^{0}(R_C) + (1\text{-}f)\, e_{AB}^{x}(R_C) \qquad \text{(IV.6.17)}$$

This actually corresponds to a definition of the denatured state as having an initial fraction f of residues A and B exposed to water and the remaining fraction (1-f) randomly buried. Park and Levitt {Park & Levitt 1996 ID: 1111} demonstrated the superiority of such a combination (equivalent to f = 0.5) over either type of energy reference state individually for selecting native conformations.

The solvent-mediated contact potentials might be more appropriate to use at initial stages of folding; whereas the intramolecular contact potentials e^{x} would be more appropriate for portraying interactions between residue pairs buried in the core which have only other residue contacts as alternatives. Thus folding simulations and the potentials also ought to change together in a coordinated way, with less and less water in the reference state as folding proceeds, i.e., by gradually letting $f \rightarrow 0$.

f. Residue-solvent interaction potentials dominate the effective inter-residue contact energies

The residue-solvent and residue-average residue interaction potentials $W_{A0}(R_C)$ and $W_{AX}(R_C)$ are important residue-specific parameters that determine the effective solvent-mediated and residue-mediated inter-residue contact energies, respectively. *In particular,* the solvent-residue interaction potentials $W_{A0}(R_C)$ are rather strong and discriminative.

Let us consider the solvation of residue A, originally located in a folded structure. This process is shown by the scheme

$$A\bullet X + 0 \rightarrow A\bullet 0 + X \qquad \text{(IV.6.17)}$$

and the accompanying *free energy of solvation* can be estimated from

$$\Delta A_A^{sol}(R_C) = W_{A0}(R_C) - W_{AX}(R_C) = -RT \ln [<q_A^0(R_C)> / <q_A^X(R_C)>] \qquad \text{(IV.6.18)}$$

by analogy to *eq IV.6.13*. *Figure IV.6.6* displays, in part *(A)*, the solvation free energies $\Delta A_A^{sol}(R_C)$ for each type of residue, extracted from two sets of structures: a dataset of monomeric proteins (open circles), and a dataset of interfacial regions in multimeric proteins or protein-protein complexes (filled circles). Part *(B)* displays the potentials corresponding to the interaction with 'average residue', $W_{AX}(R_C)$.

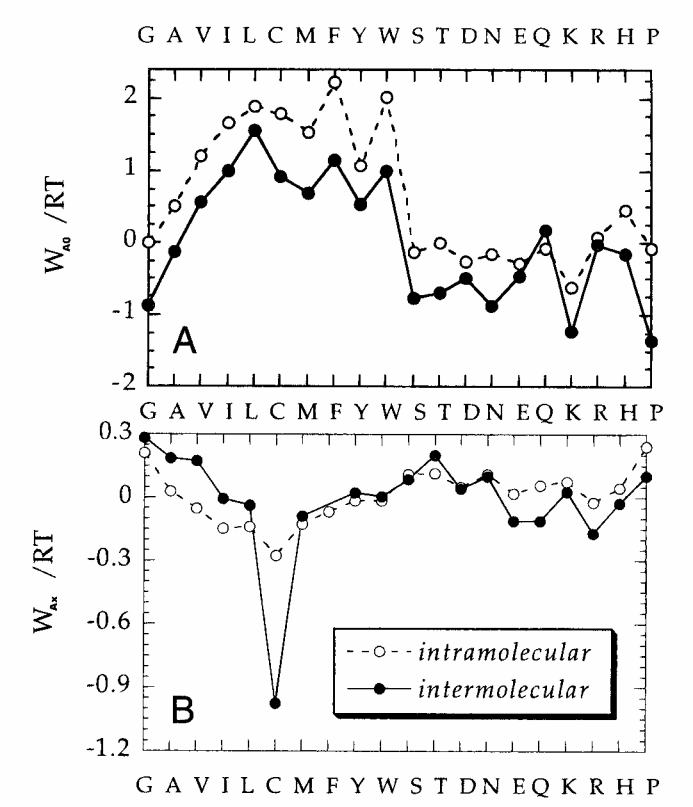

Figure IV.6.6. *Solvation free energies* $\Delta A_A^{sol} = W_{A0} - W_{AX}$ (top) *and potential of mean force* W_{AX} *between residue type A and 'average residue' X in folded structures, shown for each residue type (single letter amino acid names along the abscissa). The filled circles and solid line refer to the inter-molecular inter-residue potentials; these are obtained using residue pairs located at protein-protein interfaces. Open circles and dashed line are for the intramolecular inter-residue potentials, extracted from single chain proteins. (taken from {Keskin, Bahar, et al. 1998 ID: 49})*

A simplified method has been recently adopted by for estimating ΔA_A^{sol}, which yields results almost indistinguishable from those obtained with the more elaborate approach summarized above{Keskin, Bahar, et al. 1998 ID: 49}. ΔA_A^{sol} is estimated from

$$\Delta A_A^{sol} = - RT \ln [f_A^0(R_C) / f_A^X(R_C)] \qquad \text{(IV.6.19)}$$

where $f_A^0(R_C)$ is the fraction of residues of type A among all solvent-exposed residues, and $f_A^X(R_C)$ is the fraction of residues of type A among those completely buried. In this approximation, a given residue A is assumed to be solvent exposed if $q_A^X(R_C) \leq 4$ and completely buried if $q_A^X(R_C) \geq 7$. This simple expression is useful for a rapid, yet physically meaningful estimation of solvation free energies.

i. Empirical solvent-mediated inter-residue potentials hold for both intramolecular and intermolecular contacts

The expressions utilized to define the potentials operating at interfaces resemble those used above for intramolecular cases

$$e_{AB}^{0}(\text{inter}) = W_{AB}^{0}(\text{inter}) + W_{00}(\text{inter}) - W_{A0}(\text{inter}) - W_{B0}(\text{inter}) \qquad \text{(IV.6.20)}$$

and

$$e_{AB}^{x}(\text{inter}) = W_{AB}(\text{inter}) + W_{xx}(\text{inter}) - W_{Ax}(\text{inter}) - W_{Bx}(\text{inter}) \qquad \text{(IV.6.21)}$$

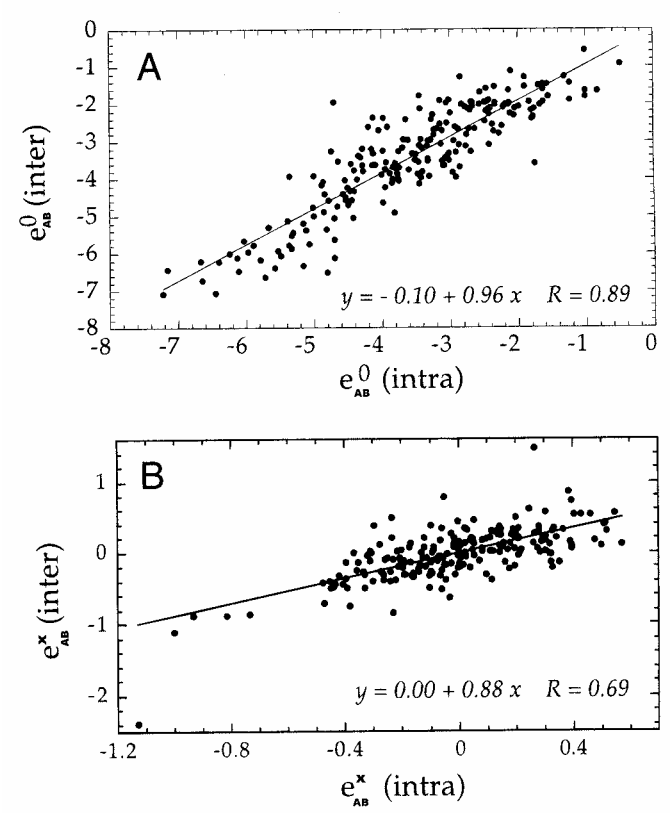

Figure IV.6.7*. Comparison of intermolecular and intramolecular inter-residue contact potentials. Values obtained for interface regions of protein-protein complexes, or multimeric proteins (ordinate) are plotted against those extracted from monomers (abscissa). Parts (A) and (B) are for solvent-mediated (e_{AB}^{0}) and residue-mediated (e_{AB}^{x}) potentials, respectively, shown in RT units. The best fit linear regression line to the data for the 210 distinct pairs, and the corresponding equation and correlation coefficient (R) are shown. (taken from {Keskin, Bahar, et al. 1998 ID: 49})*

The close similarity between the intermolecular and intramolecular solvent-mediated potentials is consistent with recent comparisons of structural motifs at protein-protein interfaces and protein cores {Tsai & Nussinov 1997 ID: 1141}{Tsai, Xu, et al. 1997 ID: 1142}. These studies showed that, although the details can vary, the global features of protein architectural motifs present in the monomers do recur at the interfaces. An important implication of the present results is that the solvent-mediated potentials e_{AB}^{0} may be used with confidence for analyzing both monomeric proteins and protein-protein interfaces, and for providing guidance regarding the energetics of both *folding* and *binding* processes.

j. The dominant role of solvent mediation permits us to express the inter-residue interactions in terms of a reduced set of single-body potentials

Two major conclusions are reached in the preceding sections:

1. The discriminatory ability of residue-residue interactions is strong in the presence of solvent. The solvent-mediated contact energies e_{AB}^{0} vary over a significantly wider range of values than do the e_{AB}^{x} values, suggesting that the solvent plays a major role in inducing residue specificities.

2. Inter-residue interactions at protein-protein interfaces bear a close resemblance to those operating intramolecularly. Perturbations to potentials due to chain connectivity or slight differences in the structural motifs at interfaces and at protein interiors, are small. The same set of energy parameters is therefore valid, to a good approximation, for both the intermolecular and intramolecular regimes, a result that removes reservations about adopting the same force fields for folding and for binding.

k. Reduced set of parameters

The specificity and robustness of W_{A0} values suggest that, to a good approximation, the 20 x 20 solvent-mediated inter-residue potentials may be estimated by using a smaller number of parameters. The idea is to use 20 energy parameters to account for the *single-body* solvation or hydrophobicity characteristics of each of the different types of amino acids, and a few (2 or 3) additional parameters to account for particular two-body (residue-residue) interactions that are more pronounced. An optimization scheme {Keskin, Bahar, et al. 1998 ID: 49}based on the minimization of the difference between database extracted e_{AB}^{0} values, and the approximate values, e_{AB}*, calculated by combining a reduced set of parameters, leads to the set of parameters presented in *Table IV.6.2*. Therein, two types of parameters are introduced: single-body potentials, W_A*, and *two-body* potentials, ΔW_{AB}*. ΔW_{AB}* is taken as zero for all contacts, except for the listed pairs, i.e. pairs of hydrophobic residues [HΦ, HΦ], oppositely charged amino acids [+, -], and disulfide bridges [Cys*, Cys*]. These parameters are used in

$$e_{AB}^* = \Delta W_{AB}^* + W_{00}^* - W_A^* - W_B^* \qquad \text{(IV.6.22)}$$

for estimating the 20x20 solvent-mediated inter-residue contact potentials. Here W_{00}^* is the optimized solvent-solvent interaction parameter. A value of -3.645 RT is assigned to W_{00}^*.

The correlation coefficient between the e_{AB}^* values found from eq *IV.6.15* and the Miyazawa-Jernigan e_{AB}^0 values (Table *IV.6.A2*) is ≅ 0.99. This is a strikingly important result, in that a total of 23 parameters (20 single-body, and 3 two-body potentials) suffices to describe a set of 210 parameters! Importantly, the same set holds both for intermolecular and intramolecular contacts.

Table IV.6.2. Reduced set of energy parameters for calculating inter-residue contact potentials (see eq IV.6.22).

Single-body potentials		***Two-body potentials*** [(a)]		
A	**W_A^*/RT**	**A**	**B**	**$\Delta W_{AB}^*/RT$**
Gly	-0.845	HΦ	HΦ	-0.3
Ala	-0.531	Cys*	Cys*	-1.1
Val	0.633	(+)	(-)	-0.8
Ile	1.087	all other pairs		0.0
Leu	1.502			
Ser	-1.076			
Thr	-0.828			
Asp	-1.302			
Asn	-1.104			
Glu	-1.334			
Gln	-1.038			
Lys	-1.648			
Arg	-1.043			
Cys	0.246			
Met	0.707			
Phe	1.512			
Tyr	0.355			
Trp	0.656			
His	-0.429			
Pro	-0.907			

(a) HΦ for the hydrophobic residues Leu, Val, Ile, Met, Phe, Trp and Cys; (+) for the positively charged residues Arg and Lys, (-) for the negatively charged residues Lys and Glu, and Cys for disulfide bridge forming Cys.*