6. Empirical Potentials derived from structures

a. Why is there a need for empirical potentials?

Empirical potentials simply provide a way to summarize the interactions seen in proteins. They have the major utilities of providing a quantitative estimate of inter-residue interactions in folded proteins and permitting the rapid scoring of different conformations in order to identify the most probable sequence-structure matches. They are based on a library of structures, usually a non-redundant set of crystal structures. As such, they include averages over many factors, but especially over variable solvent conditions. The resultant functions reflect principally the strength of the hydrophobic effect, but can also tell us about the specificities manifested by polar interactions.

There is a long history to the use of such adjustable parameters to reflect the effective interactions for various conformations. These parameters are usually chosen to deal with specific situations; for instance, the effect of environment on the relative populations of rotational isomers {Mizushima 1954 ID: 388}. In the case of polymer modeling, these were cast in the form of parameters representing interaction strengths between pairs of atomic groups at close approach. Typically these have been evaluated by fitting physical measurements such as the spectra of small molecules, to give rotational isomer populations, or at the other extreme, data for polymeric chains, to match their overall dimensions. Much of this approach to utilize adjustable energy parameters has been summarized in the two monographs by Flory {Flory 1969 ID: 460} and by Mattice and Suter {Mattice & Suter 1994 ID: 493}.

Admittedly, the circumstances prevailing in proteins are substantially more complex than in simple repeating polymers because many more parameters are required to account for the greater atomic complexity. Such an evaluation of parameters is feasible only if it is possible to make some reduction in the number of interacting entities, either by coarse-graining, i.e., considering only the 20 residue types instead of atoms, or alternatively by grouping atoms into a limited number of atom types, usually fewer than 20 types.

In addition, it could be imagined that different effective potentials might be operative at different stages of folding since the environment changes in the course of folding. For instance, at early stages the state more nearly resembles solvated conditions; whereas at later stages in a compact state, a high density of residue contacts would obtain, more similar to the crystalline state. This would affect the effective individual conformational preferences in significant ways. Such dependences can be dealt with effectively by defining different reference states, i.e. by considering interacting pairs to form from a defined starting state specific to each case. Indeed specifying the reference state is the most critical aspect for the successful application of these empirical potentials.

Developing simplified representations of protein structure and statistical potentials suitable to represent interactions at lower levels of resolution is motivated by three considerations, two theoretical and one experimental. From a theoretical point of view, full atomic

description of protein structure, even though more informative, still has some limited applicability, in view of the time and length scales of many observed phenomena. It is impossible with the present computational technology to explore all stages of folding and all motions in full atomic detail, within reasonable computation time. Low resolution methods based on coarse-grained models and parameters appear to be the most practical approach for unraveling complex issues in protein folding and recognition {Levitt 1976 ID: 1009}{Jernigan 1992 ID: 1008}. In addition to computational limitations, a further point is the fact that atomic semi-empirical potentials as commonly used discriminate poorly between correctly folded and misfolded conformations {Novotny, Bruccoleri, et al. 1984 ID: 282}{Novotny, Rashin, et al. 1988 ID: 199}{Wang, Zhang, et al. 1995 ID: 23}, principally because of their inadequacies in accounting for solvent effects. Typically these potentials do not perform well in selecting native folds, although many improvements are underway. Many studies have demonstrated the requirement for a sufficiently pronounced energy minimum or "energy gap" {Shakhnovich 1994 ID: 980}{Sali, Shakhnovich, et al. 1994 ID: 1016}. The lack of a suitable potential function, rather than the design of a folding algorithm, has been suggested {Sali, Shakhnovich, et al. 1994 ID: 955} to be the major bottleneck to structure prediction. The neighborhood of the native state needs to have only relatively low potential energy barriers, in order for the protein to fold into its native state {Karplus & Sali 1995 ID: 1021}. Conventional atomic potentials do not usually behave this way. A lower resolution description can reduce the number of states and also smooth the barriers between local minima. These are important practical considerations for simulations. From the experimental point of view, extremely large numbers of high resolution structures are now available from X-ray crystallography, offering a wealth of information. These structures may be used for extracting the so-called *knowledge-based potentials of mean force*, and can yield effective free energies associated with the various inter-residue contacts in globular proteins.

There are many possible applications of empirical potential functions: 1) to discriminate good protein folds from bad ones, 2) similarly, to thread a given sequence through many different structures and choose the best, 3) to thread inversely by choosing the best sequence for a fixed structure, 4) to inform us about various aspects of protein structures, and in particular, improve our understanding of the dominant interactions stabilizing native structures, and 5) to bind two proteins together in the best arrangement out of the many mutual arrangements between the two rigid proteins. There will be a particular emphasis on item four in this section. In principle, the same set of potential functions can be utilized for any of the first four purposes. The principal obstacle to this is the proper definition of reference states.

b. The extraction of empirical potentials relies on the applicability of the Inverse Boltzmann Principle

The practice of extracting knowledge-based potentials for proteins was initiated by Tanaka and Scheraga {Tanaka & Scheraga 1976 ID: 1029}. A more rigorous determination of the effective inter-residue contact potentials, including both solvent and size effects, was subsequently made by Miyazawa and Jernigan {Miyazawa & Jernigan 1985 ID: 1010}.

The Miyazawa-Jernigan contact potentials consist of 210 parameters (for a given reference state) usually organized into 20 x 20 symmetric matrices, each element of which corresponds to a given pair of residues. Among the potentials developed in different ways and widely used in the literature are those of: Eisenberg and coworkers {Lüthy, Bowie, et al. 1992 ID: 1070} who evaluated the environment of each amino acid in each structure on the basis of several properties in the so-called *Profile method*, and Sippl {Sippl 1990 ID: 1001} who expressed the inter-residue interactions in terms of distance-dependent potentials.

The basic assumption adopted in extracting empirical interaction potentials in these and many other similar studies is the so-called *Inverse Boltzmann principle*. According to the Boltzmann principle, the probability of occurrence of a given conformational state of energy E scales with the *Boltzmann factor* exp{-E /RT}, where R is the gas constant (1.987×10^{-3} $kcal.mol^{-1}K^{-1}$) and T is the absolute temperature ($RT \approx 0.6$ kcal/mol near room temperature). The probability, or frequency of occurrence of a given state, or a given interaction, can thus be calculated given the energy of that interaction. The inverse Boltzmann law, on the other hand, calculates the energies from the probabilities, -or more precisely from the natural logarithm of the observed frequencies. In the application of the inverse Boltzmann law to the extraction of inter-residue potentials from databank structures, a large ensemble of non homologous known structures is considered; and the frequencies of interactions between all pairs of amino acids are analyzed. In a strict sense, the applicability of the inverse Boltzmann law depends on the accessibility of the complete set of conformational states. The observed ensemble of database structures does not conform with this requirement {Thomas & Dill 1996 ID: 956}. Yet, the extracted energy data have been extensively tested in theoretical analyses and simulations, and have proven to be sufficiently robust and discriminative for recognizing the correct sequence-structure matches (see below), which lends support to their use in theoretical and numerical studies of protein structures.

c. How to use a simple model to extract potentials?

In general, the residue-specific *potential of mean force* between two interaction sites, or a pair of residues of types A and B, located within a distance range $r \pm \Delta r$ is given by the Boltzmann relation

$$\Delta W_{AB}(r) = - RT \ln [p_{AB}(r \pm \Delta r)/p_{XX}(r \pm \Delta r)] = W_{AB}(r) - W_{XX}(r) \qquad (IV.6.1)$$

where $p_{AB}(r \pm \Delta r)$ is the probability of finding the specific pair [A, B] at a separation in the range $r \pm \Delta r$, and $p_{XX}(r \pm \Delta r)$ is the reference probability, for the occurrences of all residue types X, R is the gas constant and T is the absolute temperature. $\Delta W_{AB}(r)$ also represents the *free energy change* associated with the formation of the contact pair A-B. $W_{XX}(r)$ may be viewed as the average potential existing between all types of residue pairs, typical of inter-residue interactions in folded structures

This expression already depends on several assumptions: (i) the choice of the reference state, (ii) the applicability of Boltzmann statistics, (iii) the choice of interaction sites in each residue, and (iv) the discretization of conformational space into intervals of width $2\Delta r$. Two other fundamental issues implicit in *eq IV.6.1* are: (v) the potential energy is expressed only as a sum of pair-wise interactions, ignoring many body terms, and (vi) each pair of specific residue types A, B is assumed to behave independently of all others, regardless of the chain connectivity, i.e., ignoring constraints imposed by specific sequential neighbors, and context or environmental conditions. Finally, the experimental data are assumed to be sufficiently large to provide a truly representative sample of all possible interacting forms, including, for example, all orientations. And indeed this averaging over a wide range of interacting conformations causes some loss of specificity that is inherent to coarse-grained models.

This approach is typically simplified further by summing the sampled counts over all distances up to a certain limiting value, R_C. This limit should be chosen so that most pairs within this distance are actually interacting, all pairs of residues further apart are assumed not to interact. An appropriate basis for selecting this cutoff distance is to include the first coordination sphere of residues that can be selected by analyzing the distribution of residue pairs in known structures. In this case eq *IV.6.1* becomes

$$\Delta W_{AB}(R_C) = - RT \ln [p_{AB}(r \le R_C) / p_{XX}(r \le R_C)] \qquad \text{(IV.6.2)}$$

Figure IV.6.1 shows that within the range 6.5 – 7.0 Å can be found an appropriate value of R_C to include the first shell of neighbors. On this basis the general approach followed for data collection of interacting residues is shown in *Figure IV.6.2*. The main procedure is to count the number of pairs of different types in the neighborhood of a central residue of a given type within the coordination sphere of radius R_C. This procedure is repeated for all residues and all proteins. A major assumption is that the empty sites in the neighborhood of the examined residue are filled by *effective solvent molecules*, or by a hypothetical group of water molecules, whose volume is equal to that of an average residue, as illustrated in *Figure IV.6.2*.

Evaluation of the precise number of solvent molecules coordinating a given residue is critically important, inasmuch as solvation, or the hydrophobic effect, is an essential property that dominates the effective the inter-residue interactions. Let us consider for example residues of type A. Let the total number of residues of type A in the dataset of structures be N_A, and the total number of pairs, summed over all non-bonded neighbors (residue 'X' and solvent '0') be $N_{AX} + N_{A0}$. It is possible to define two mean coordination numbers for each residue type: the *total coordination number*, $\langle q_A \rangle$, given by

$$\langle q_A \rangle = (N_{AX} + N_{A0}) / N_A \qquad \text{(IV.6.3)}$$

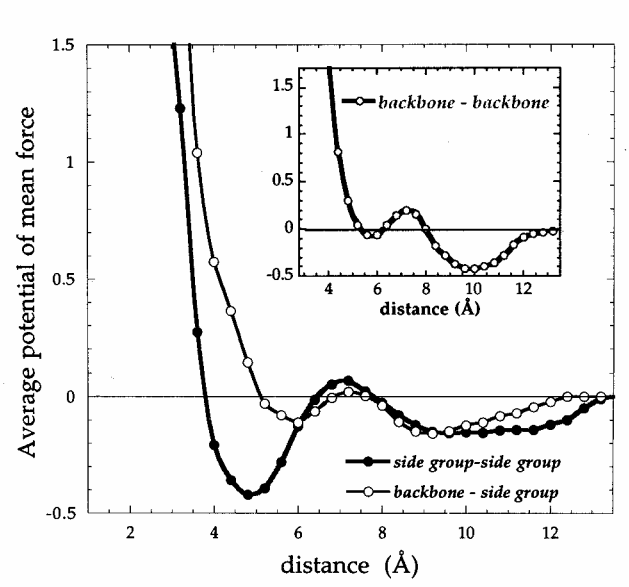

Figure IV.6.1. *Potential of mean force $W_{XX}(r)$ between side chain-side chain (S-S), side chain-backbone (S-B) and backbone-backbone (B-B) sites of interaction, obtained from 150 X-ray elucidated protein structures {Jernigan & Bahar 1996 ID: 161}{Bahar & Jernigan 1997 ID: 81}. Backbone interaction centers are C^{α} atoms. Side chains sites are defined by a group of terminal atoms selected to be specific. All (S-S), (S-B) and (B-B) pairs separated respectively by three or more, four or more, and five or more residues along the sequence are considered, yielding residue non-specific potentials characteristic of compact globular structures. Multiple minima correspond to consecutive coordination shells. A value of $R_C = 7.0$ Å is indicated as an upper bound for including all neighbors located within a first coordination shell. Figure taken from {Jernigan & Bahar 1996 ID: 161}.*

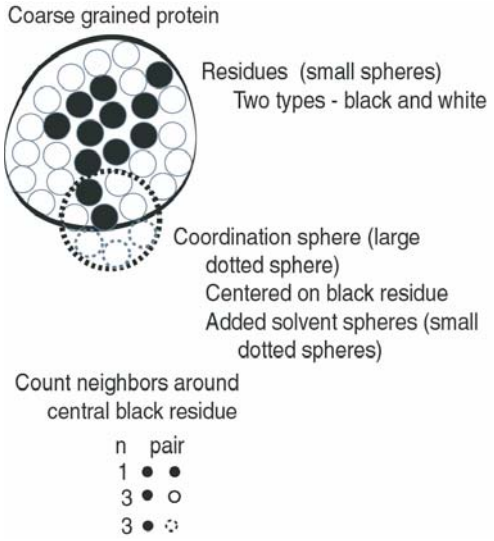

Figure IV.6.2. *Schema for collecting interacting residue pair data from coarse grained protein structures.*

and the *residue coordination number*

$$<q_A^X> = N_{AX} / N_A = \Sigma_B N_{AB} / N_A \qquad \text{(IV.6.4)}$$

which is given essentially by the average number of contacts residues of type A make with other residues, only. N_{AB} designates all contacts of type A-B, and the summation in eq IV.6.4 is performed over all 20 types of contacts, A-X. Clearly, the number of effective solvent molecules coordinating A is $<q_A^0> = <q_A> - <q_A^X> = N_{A0} / N_A$. The coordination numbers depend on R_C, and it is better to designate them as $<q_A(R_C)>$, $<q_A^X(R_C)>$, and $<q_A^0(R_C)>$. Evaluation of $<q_A^X(R_C)>$ is straightforward. It suffices to count all residues of type A and all their inter-residue contacts in the examined dataset of structures. Calculation of $<q_A>$, on the other hand, is more difficult, because database structures do not usually contain solvent molecules. An indirect method is to use the mean field approximation {Miyazawa & Jernigan 1985 ID: 1010}

$$<q_A(R_C)> = (4\pi R_C^3/3 - V_A) / V_X^A(R_C) - q_A^b(R_C) \qquad \text{(IV.6.5)}$$

Here V_A is the volume occupied by residue, and which is excluded to all its neighbors, $V_X^A(R_C)$ is the average volume of all residues X located within R_C around A, to be determined from a statistical examination of databank structures, and $q_A^b(R_C)$ is the average number of *b*onded (first neighbors along the sequence) residues located within R_C. See Table *IV.6.1* for the values of V_A, $q_A(R_C)$, $V_X^A(R_C)$, and $q_A^b(R_C)$ for R_C = 6.5 Å, obtained using the side chain centroid as the interaction site for each residue.

Equation *IV.6.2* provides information on the 20 x 20 contact potentials operating over a given distance range $r < R_C$. On the other hand, it is usually useful to consider the distance-dependence of the potentials. Park and Levitt implemented an approximate distance-dependence in a way similar to that of Wallqvist and Ullner {Wallqvist & Ullner 1994 ID:966}, which appears to aid in discrimating for native conformations {Park & Levitt 1996 ID: 1111}. Likewise, Crippen & Maiorov who fit large numbers of parameters to structures used a smoothing over distance {Crippen & Maiorov 1994 ID: 1138}]. For a more rigorous examination of the distance dependence, one can resort to *radial distribution functions* g(r), also called *pair correlation functions* {Hansen & McDonalds 2000 ID: 1143}{Ben-Naim 1992 ID: 1084}.

Table IV.6.1. Coordination data for residues in globular proteins (R_C = 6.5 Å)

Residue type[(a)]	V_X^A [(c)]	q_A^b [(c)]	$\langle q_A \rangle$ [(d)]	$\langle q_A^X \rangle$	[(e)] Interaction site[(e)]
GLY	133.1	1.75	6.284	2.88	C^{α}
ALA	136.5	1.46	6.334	3.57	C^{β}
SER	132.9	1.33	6.582	2.71	O^{γ}
CYS	133.9	1.15	6.646	3.67	S^{γ}
THR	133.6	1.19	6.486	2.88	O^{γ}
ASP	134.7	1.00	6.487	2.42	$O^{\gamma 1}$,$O^{\gamma 2}$
PRO	140.6	1.45	5.858	2.60	C^{β},C^{γ},C^{δ}
ASN	137.1	0.90	6.574	2.57	$O^{\gamma 1}$,$N^{\delta 2}$
VAL	140.7	1.07	6.155	4.52	$C^{\gamma 1}$,$C^{\gamma 2}$
GLU	142.4	0.72	6.235	2.35	$O^{\varepsilon 1}$,$O^{\varepsilon 2}$
GLN	136.2	0.74	6.469	2.71	$O^{\varepsilon 1}$,$N^{\varepsilon 3}$
HIS	142.1	0.73	6.241	3.43	C^{γ},$N^{\delta 1}$,$C^{\delta 2}$,$C^{\varepsilon 1}$,$N^{\varepsilon 2}$
LEU	144.1	0.74	6.087	5.01	$C^{\gamma 1}$,$C^{\gamma 2}$
ILE	141.2	0.92	6.042	4.84	$C^{\delta 1}$
MET	146.9	0.59	6.137	4.75	S^{δ}
LYS	138.1	0.54	6.569	1.92	N^{ζ}
ARG	140.8	0.35	6.318	2.90	N^{ε},NH^1,NH^2
PHE	147.0	0.56	5.870	4.99	aromatic C's
TYR	143.3	0.54	6.037	4.11	aromatic C's, OH
TRP	144.7	0.45	5.793	4.65	all aromatic atoms
X	138.5	1.01	6.281		
0	139.6	0.0	7.161		

[(a)] X stands for an average residue, and 0 for effective solvent molecule, [(b)]The volumes V, except for ARG, have been taken from Table 2 of {Chothia & Janin 1975 ID: 1144}, and V for ARG from Table 4 of {Chothia & Janin 1975 ID: 1145}; [(c)] from {Miyazawa & Jernigan 1985 ID: 1010} [(d)] from {Miyazawa & Jernigan 1996 ID: 174}) [(e)] side chain representation in {Bahar & Jernigan 1996 ID: 83}

Radial distribution functions multiplied by mean densities represent the effective densities as a function of position with respect to an investigated central site (see Figures *IV.6.3*). In a sense, they correct for the local fluctuations in densities in the neighborhood of a central site.

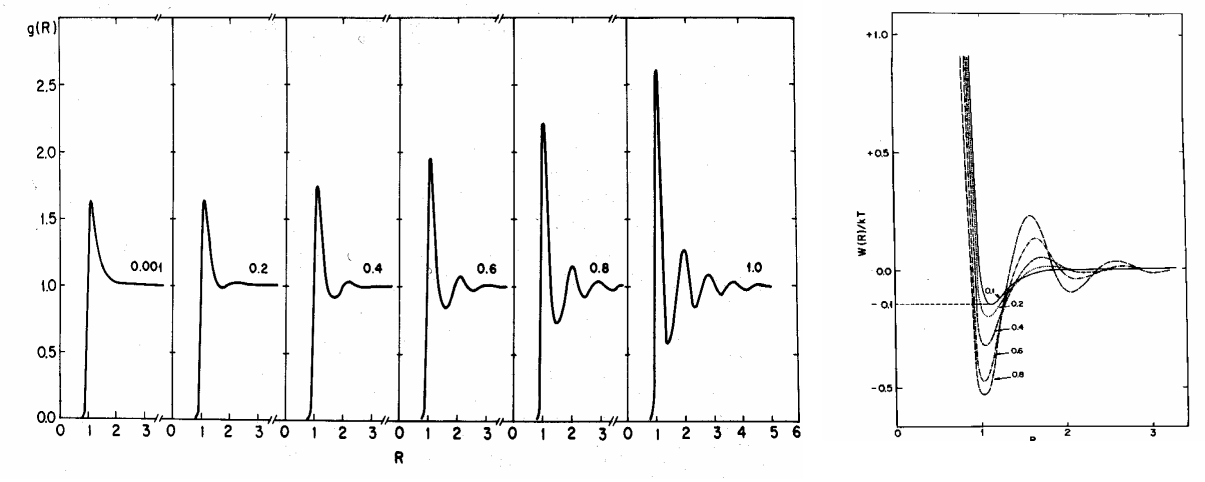

Figure IV.6.3. *Pair radial distribution functions as a function of density (left) and corresponding potentials of mean force (right). Note that g(r) = 0 at small separation due to the volume exclusion effect, reaches a maximum at the optimal (lowest energy) interaction distance, and approaches unity at long distances. Labels in the figures refer to varying density parameters. At low densities a single peak is observed, whereas with increasing density (moving towards the right) multiple peaks appear, indicative of successive shells of coordination around the central molecule. (from {Ben-Naim 1992 ID: 1084} Figures 5.6 and 5.8). Note that the inter-residue potential of mean force exhibits two minima, conforming to the first and second coordination shells of residues in the dense environment of protein interiors.*

Radial distribution functions differ from directly counted frequencies, being normalized with respect to radial distance. The actual observed number $N_{AB}(r \pm \Delta r)$ of neighbors of B type, located in a spherical shell of thickness Δr centered about A type residues is normalized by dividing it by the volume of that shell, $4\pi r^2 \Delta r$. This type of normalization avoids overweighting of neighbors in distant, larger volume elements. As r increases, g(r) approaches unity, or the product of the mole fractions of interacting species for ideal mixtures. The associated potential of mean force {McQuarrie 1976 …}{Ben-Naim 1992 ID: 1084} (or free energy) W(r) = - RT ln g(r) thus vanishes at large separations *(Figure VI.6.3*), following normalization for composition. In globular proteins, g(r) approaches unity, and $W_{AB}(r)$ becomes negligibly small beyond ~ 13 Å.

Distance-dependent residue-specific potentials of mean force have been derived for globular proteins using {Bahar & Jernigan 1997 ID: 81}

$$\Delta W_{AB}(r) = - RT \ln [g_{AB}(r) / g_{XX}(r)] \qquad \text{(IV.6.6)}$$

where $g_{AB}(r)$ is the effective radial distribution

$$g_{AB}(r) = \langle q_A^X \rangle \langle q_B^X \rangle N_{AB}(r) (4\pi r^2)^{-1} / \int N_{AB}(r) (4\pi r^2)^{-1} dr \qquad \text{(IV.6.7)}$$

and $g_{XX}(r)$ is its counterpart for all types of inter-residue contacts. The normalization integral in the denominator is taken in the range $0 \le r \le 13$ Å. Multiplication by $\langle q_A^X \rangle$ takes account of the differences in the coordination numbers of the individual residues. Integration of the radial distributions in the limits $0 \le r \le R_C$ leads to potentials of mean force $\Delta W_{AB}(R_C)$ representative of the residue-specific contacts occurring in the range $r \le R_C$ as

$$\Delta W_{AB}(R_C) = - RT \ln [\int g_{AB}(r) dr / \int g_{XX}(r) dr] \qquad \text{(IV.6.8)}$$

Radial distribution functions have proven to be useful in deriving empirical energies not only for inter-residue interactions in proteins, but also for protein-ligand interactions at the atomic level (Muegge & Martin, 1999).

d. Databank structures reveal significant differences in the coordination numbers and optimal interaction distances of hydrophobic and hydophilic residues

While viewing the distributions of positions for residues it is interesting to note two differences in the behavior of hydrophobic (H) and hydrophilic (or polar (P)) residues. As q_A^X values in Table *IV.6.1* also indicate, hydrophobic residues as a group usually have more neighbors than do hydrophilic ones, i.e., hydrophobic residues are more buried in globular proteins than are hydrophilic residues. See *Figure IV.6.4*. Furthermore there are significant differences in the behavior of the two categories of residues in terms of their typical interaction distances. Hydrophilic residues usually have more specific interactions at closer approach than do hydrophobic ones. Some examples are shown in *Figure IV.6.5*. The ordinates are the potentials of mean force, $\Delta W_{AB}(r)$, extracted from databank structures {Jernigan & Bahar 1996 ID: 161}. We see that the optimal interaction distance shifts from around 5 Å for H-H and H-P pairs (top) to approximately 3 Å in the case of P-P interactions between polar and charged groups (bottom). Furthermore, the P-P interactions, although operating over a narrower distance range, exhibit sharper energy minima. As is well known, hydrophilic pairs are the most specific in their interactions, in part because of the hydrogen bond forming potential of their polar groups.

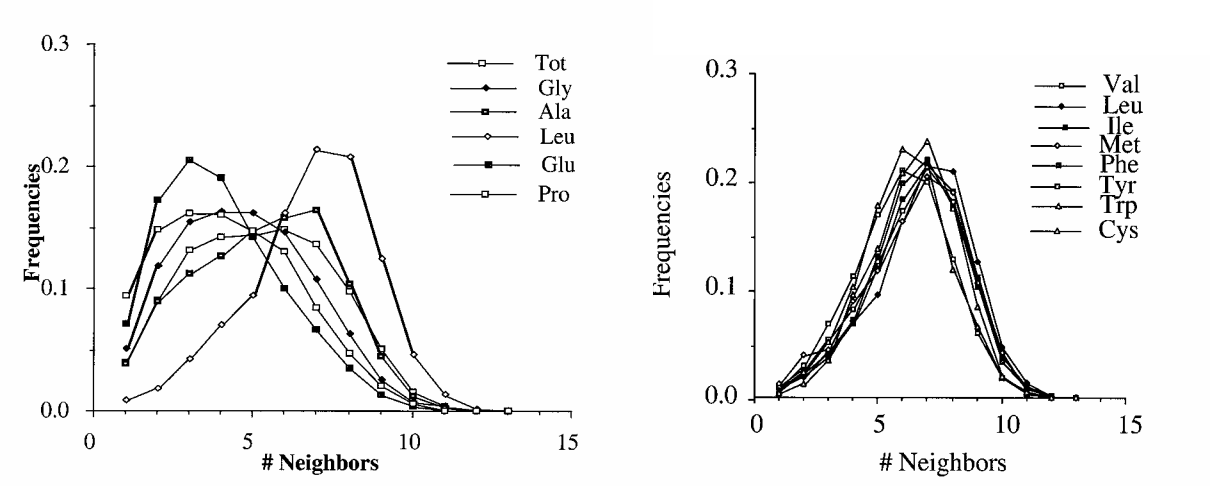

Figure IV.6.4. *Distributions of the numbers of neighboring residues within 6.5 A of the named residue. The curves reflect the different circumstances found in globular proteins where hydrophobic residues are usually buried and hydrophilic ones are usually located more on the surface. The mean value of the distributions yield the residue coordination numbers* $\langle q_A^X \rangle$ *plus bonded neighbors* $q_A^b(R_C)$ *(Modified from {Miyazawa & Jernigan 1996 ID: 174})*

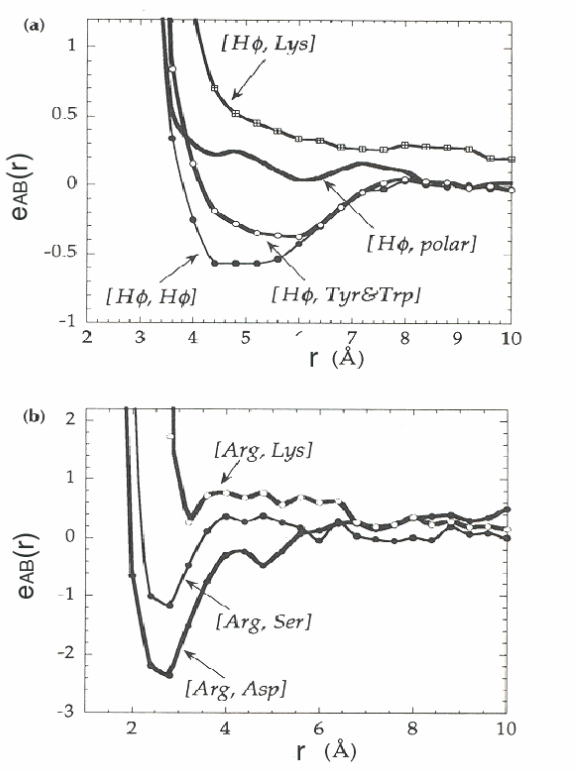

Figure IV.6.5. *$\Delta W_{AB}(r)$ for specific pairs of residues. HΦ refers to the average behavior of the hydophobic residues Leu, Val, Ile, Phe and Met and 'polar' residues include Asn, Gln, Ser, Thr and His. Note the change in the position and width of minima between parts (a) and (b). (from {Jernigan & Bahar 1996 ID: 161}.) See Table IV.6.1, last column, for side chain atoms used for defining the residue interaction sites.*