PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Quantitative expression of protein heterogeneity: Response of amino acid side chains to their local environment

Debashree Bandyopadhyay and Ernest L. Mehler*

Department of Physiology and Biophysics, Weill Medical College of Cornell University, New York, New York 10021

ABSTRACT

A general method has been developed to characterize the hydrophobicity or hydrophilicity of the microenvironment (MENV), in which a given amino acid side chain is immersed, by calculating a quantitative property descriptor (QPD) based on the relative (to water) hydrophobicity of the MENV. Values of the QPD were calculated for a test set of 733 proteins to analyze the modulating effects on amino acid residue properties by the MENV in which they are imbedded. The QPD values and solvent accessibility were used to derive a partitioning of residues based on the MENV hydrophobicities. From this partitioning, a new hydrophobicity scale was developed, entirely in the context of protein structure, where amino acid residues are immersed in one or more "MENVpockets." Thus, the partitioning is based on the residues "sampling" a large number of "solvents" (MENVs) that represent a very large range of hydrophobicity values. It was found that the hydrophobicity of around 80% of amino acid side chains and their MENV are complementary to each other, but for about 20%, the MENV and their imbedded residue can be considered as mismatched. Many of these mismatches could be rationalized in terms of the structural stability of the protein and/or the involvement of the imbedded residue in function. The analysis also indicated a remarkable conservation of local environments around highly conserved active site residues that have similar functions across protein families, but where members have relatively low sequence homology. Thus, quantitative evaluation of this QPD is suggested, here, as a tool for structure–function prediction, analysis, and parameter development for the calculation of properties in proteins.

Proteins 2008; 72:646–659.

Key words: hydrophobic matching; protein heterogeneity; microenvironments; hydrophobicity of local environments; Rekker hydrophobic fragmental constants.

INTRODUCTION

The heterogeneity of protein interiors is due to the differences in the properties of the organic functionalities making up amino acid side chains, which range from hydrophobic residues through polar to titratable residues that are usually charged at physiological pH. This heterogeneity has long been recognized as an inherent characteristic of proteins and a key determinant of their properties. The usual approach to describe this internal multiformity has been the development of hydrophobicity scales (HScls)[1–4] for amino acid residues. This type of characterization of hydrophobicity has been of major importance in protein folding studies,[5–7] but it is a static, externally derived description of hydrophobicity. It does not express the structural attribute that every amino acid in a protein is imbedded in a local environment defined by the local protein architecture and the surrounding solvent.

That such a description might be useful was first recognized by Ponnuswamy,[8] who proposed a scheme for characterizing local environment based on amino acid HScls. That work provided early evidence for the lack of correlation between extent of burial and other properties. Another approach to define descriptors of local environment was developed by Eisenberg and collaborators for identifying structural homology in cases of little or no sequence homology.[9] The method was applied to human bactericidal/permeability-increasing protein.[10] In another approach, Kellogg and collaborators defined hydrophobic fields to characterize the local environment around atoms in organic molecules and proteins. Their approach used the hydrophobic atom constants developed by Leo, Hanch, and Abraham[11–14] to quantify the hydrophobic fields, which were then used for empirical studies of protein properties, QSAR and virtual screening of biological targets.[15–17] A quantitative descriptor of local environment was used earlier to

The Supplementary Material referred to in this article can be found online at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/

Abbreviations: MENV, microenvironment; GPCR, G-protein coupled receptor; QPD, quantitative property descriptor; HScl, hydrophobicity scale.

Grant sponsor: NIH; Grant numbers: R01-DA015170, PO1 GM066531.

*Correspondence to: Ernest L. Mehler, Department of Physiology and Biophysics, Weill Medical College of Cornell University, 1300, York Avenue, New York, NY 10021.
E-mail: elm2020@med.cornell.edu

Received 14 June 2007; Revised 21 November 2007; Accepted 6 December 2007
Published online 4 February 2008 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.21958

improve the prediction of pKa values of titratable residues that were deeply buried in hydrophobic local environments in the protein interior.[18,19] Titratable residues buried in such local environments often exhibited very large shifts in their solution pKa values and were particularly problematic to predict with reasonable reliability.

Detailed experimental understanding of local environments is limited to cases where intrinsic or extrinsic fluorescence quenching by effective quenchers can be measured as a function of the microenvironments (MENVs) of the protein.[20] Local environmental changes in several proteins have been monitored by fluorescence quenching methods.[21–23]

The main result of this article is the development of a quantitative property descriptor (QPD) that expresses the properties of the local environment around an amino acid side chain in terms of a hydrophobicity descriptor. Subsequently, the modulating effect of these environments on the properties of their imbedded residues is discussed. It should be emphasized that while this QPD is discussed here in the context of proteins, it is easily generalized to nucleic acids, ligands, or other molecules that are biologically important in the system under study. This is in contrast to most approaches that are primarily defined for proteins and are not easily extendible. It is also important to note the QPD that is defined describes the local environment, not the imbedded amino acid side chain. Because of this, it is important to show that this quantity is physically reasonable, which is accomplished here by using it to develop a HScl for amino acid side chains which is shown to correlate reasonably well with earlier scales, in particular with experimentally derived scales.[24,25] Nevertheless, because of the uniqueness of the definition of this QPD, it has the potential for providing new insights as will be discussed in the latter parts of this report.

The fact that the QPD can be used to develop a Hscl suggests that in most cases the hydrophobicity between local environment and its imbedded residue is reasonably well matched. Nevertheless, the analysis shows that there are differing protein architectures that produce mismatches and that residues imbedded in such mismatched MENV are often involved in function or structural stabilization of the active site region. These will be discussed in the final sections of the article, where some consequences of mismatching are explored and the conservation of the hydrophobicity of local environment is analyzed for a number of properties such as structural homology of functional residues with little overall sequence homology.

FORMULATION

Quantitative determination of the MENVs

There are a large number of HScls available for evaluating the hydrophobicity of amino acid residues.[1] Here, the hydrophobicity of the residue imbedded in an MENV is not required, but instead it is a quantitative measure of the hydrophobicity/hydrophilicity (Hpy) of the MENV that we seek. Once such a quantitative measure is available the response of the residue to changes in the local environment can be evaluated. Note that the abbreviation, Hpy, refers to the generic properties of hydrophobicity and hydrophilicity of MENVs as well as their quantitative values. The calculation of Hpy values for MENVs is most conveniently based on an atom or small fragments description because this allows the Hpy to be defined in a simple, distance dependent way. Moreover, all the atoms of a side chain do not necessarily belong to a given MENV; also, an atom/fragment based scale can account for the fact that certain side chains are best described by more than one Hpy value along their entire length. A further important advantage of a fragment-based scale is that the presence of coenzymes, prosthetic groups, ligands, or nucleic acids can all be accounted for with a consistent scale, which is not the case for most scales proposed in the literature. These considerations suggest that the Rekker fragmental hydrophobic constants[26,27] would be most suitable for the present development. The Rekker fragmental constants were developed on the basis of experimental water–octanol partition coefficients and provided a set of atomic or fragment parameters to calculate the partition coefficients of candidate drug molecules. Importantly, Rekker's approach contains no information from any aspect of protein structure or composition.

The Rekker scale was the first fragmental system developed and has been widely used in the pharmaceutical industry as a means to calculate partition coefficients of potential drug candidates. Other fragmental systems have been proposed including two that are based on functional fragments[28,29] and one that is atom based.[30] An earlier comparison indicated that the Rekker system was one of the most accurate available.[31] Here, we use it *not* to obtain the Hpy of a particular residue, but to obtain this quantity for the local protein environment around an imbedded side chain. This is done in three steps as follows:

i. identify fragments of the residues that are in the neighborhood of the imbedded residue (Figure 1),
ii. assign the Rekker coefficients to these fragments, and
iii. sum them to determining the local Hpy of the region around the imbedded residue.

Decomposition of amino acid side chains into fragments with different chemical properties evolved from the presence of hydrophobic (carbon and sulfur) or polar and charged (oxygen or nitrogen) in the same amino acid side chain.[32] Adjacencies of atoms of a particular type define patches with different properties, and the

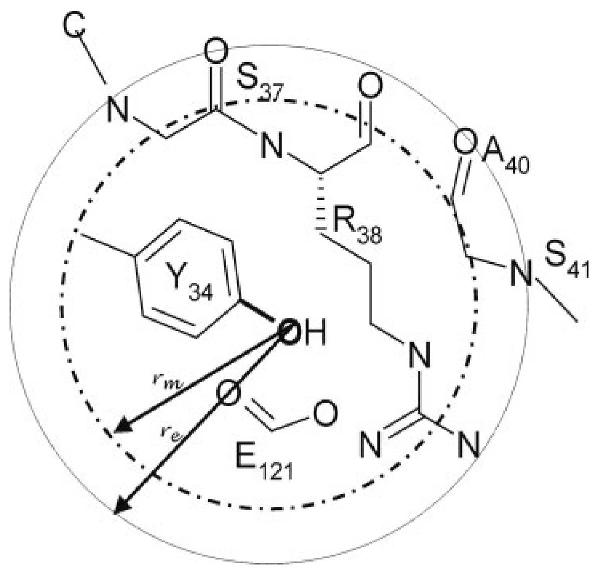

Figure 1

Schematic representation of the microenvironment: Shown are atoms from amino acid residues found within the MENV around the O of the Y_PH group of Tyrosine34 (pdb access no. 1ayo). The atoms with $r \leq r_m$ are counted with a weight of 1. The atoms inside the shell, i.e., $r_e \geq r \geq r_{rm}$, make a contribution < 1, according to a switching function.[37] Atoms with $r \geq r_e$ do not contribute (see Methods).

decomposition used here is based on fragments that characterize different organic functionalities in a given side chain. The definitions of the chemical groups are included in Table I and schematically by Figure S1 in the supplementary materials; it is noted that nine side chains are made up of two or more groups.

The MENV around any chemical group has been constructed to include the atoms within the first van der Waals' (vdW) interaction shell[33] and is presented schematically in Figure 1. The MENV around any atom of the imbedded side chain/chemical group is defined by two variables, r_m and r_e, that allow for a continuous scaling of an atom's or molecular fragment's contribution from its full value for $r \leq r_\mathrm{m}$ to zero for $r \geq r_\mathrm{e}$ (see Fig. 1). Advised by the formula used to calculate logP from the fragmental hydrophobic constants,[27] Hpy_A for the MENV of the imbedded group A is calculated from the formula

$$\mathrm{Hpy}_\mathrm{A} = \sum_{a}^{N_\mathrm{A}} \sum_{Bb}^{N_\mathrm{B}} \max\{d_\mathrm{b}(r_\mathrm{ab})\} F_\mathrm{b} \quad (\mathrm{B} \neq \mathrm{A}) \qquad (1)$$

where A and B refer to side chains or chemical groups constituting the side chain, N_A and N_B are the number of atoms in A and B, respectively, and r_ab is the distance between atoms $a \in A$ and $b \in \mathrm{B}$. F_b is the Rekker hydrophobic fragmental constant[27] of atom b in the MENV of the imbedded group A. The explicit values of F_b were reported in Table II of reference.[18] These quantities were also used for the pKa calculations, and for the sake of simplicity and uniformity they will also be used here. Max$\{d\}$ scales F_b with a value depending on the distance of atom b from each atom a of A, as defined by Eq. (2):

$$d_\mathrm{b}(r_\mathrm{ab}) = \begin{cases} 1 & \text{for } r_\mathrm{ab} \leq r_\mathrm{m} \\ \dfrac{(r_\mathrm{e} - r)^2 (r_\mathrm{e} + 2r - 3r_\mathrm{m})}{(r_\mathrm{e} - r_\mathrm{m})^3} & \text{for } r_\mathrm{m} < r_\mathrm{ab} < r_\mathrm{e} \\ 0 & \text{for } r_\mathrm{ab} \geq r_\mathrm{e} \end{cases} \qquad (2)$$

where the CHARMM switching function[37] has been adapted to provide a continuous reduction of the scaling factor, d_b, from one to zero in the shell region ($r_m < r < r_e$), as well as continuous first and second derivatives.

The values of r_m were estimated by extending the vdW radii of the different atom types, as defined in CHARMM,[37] by the vdW radius of the C_α atom (2.275 Å) to ensure that all the atoms in the first shell are included. For polar atoms, (N and O), an additional extension of 0.2 Å is added to account for the presence

Table I

Summary of Statistics for Test Set of 733 Proteins

Residue	No. of aa	Groups[a,b]	$>\lvert\sigma\rvert$[c]	% $>\sigma$	$>\lvert 3\sigma\rvert$[d]
Cys	2114	TO	214	10.1	25
Cystine	701	TD	65	9.3	1
Ile	12,687	HO	1664	13.1	174
Leu	19,435	HO	2638	13.6	127
Phe	9060	RS, H1	1068	11.8	79
Val	15,848	HO	2308	14.5	108
Trp	3266	RS, H1	930	28.5	25
Tyr	7875	RS, PH, H1	983, 528	12.5, 6.7	39, 0
Met	4511	TE, H1	712	15.8	0 (16[e])
Ala	17,779	HO	3553	20.0	0
Pro	10,084	H2	1698	16.8	0
His	5211	HS	493	9.4	0
Thr	12,470	OL, HO	1345	10.8	1
Ser	13,386	OL	1828	13.7	0
Arg	10,731	GS, H2	1448	13.5	0
Gln	8586	AD, H1	1285	15.0	2
Asn	9975	AD	1466	14.7	1
Asp	12,946	CO	1706	13.2	2
Glu	14,422	CO, H1	2033	14.1	15
Lys	13,113	AS, H2	1938	14.8	13

[a] The chemical groups comprising each side chain and the residues where they appear are defined as follows (functionality, symbol: residues where it appears): terminal aliphatic chain, HO: Ile, Leu, Val, Ala, Thr; connecting $\mathrm{CH_2}$, H1: Phe, Tyr, Trp, Met, Glu, Gln; connecting $\mathrm{C_nH_{2n}}$ ($n > 1$), H2: Lys, Arg, Pro; aromatic ring, RS: Phe, Trp, Tyr; amide, AD: Asn, Gln; carboxyl, CO: Asp, Glu; thiol ($\mathrm{CH_2SH}$), TO: cys; disulfide bridge ($\mathrm{CH_2S\text{-}SCH_2}$), TD: cystine; thioether, TE: Met; guanidinium, GS: Arg; ammonium: AS, Lys.

[b] Numbers and percentages of mismatches ($\lvert\sigma\rvert$) and extreme mismatches ($>\lvert 3\sigma\rvert$) are given only for the functionally most important chemical groups, excluding H1; see text for further discussion of the entries in this table.

[c] Mismatched chemical groups having ζ and rHpy values that deviate from the mean $>1\sigma$.

[d] Extreme mismatches deviate from the mean $>3\sigma$.

[e] Number of totally solvent exposed ($\zeta = 0$) Met residues with rHpy = 1.0.

Table II
Residue Hydrophobicity Scale-Based on Eq. (4) Partitioning Compared with Previously Published Hydrophobicity Scales

AA	TW[a]	Faupl[24]	Abodr[25]	Rose[34]	Ponnu[8]	Mijer[35]	KyteDo[36]	White[4]	Eisen[2]
C	1.15	1.54	[b]	0.91	14.93	7.93	2.5	−0.02	0.38
I	0.97	1.8	9.3	0.88	14.77	8.83	4.5	−1.12	1.90
L	0.87	1.7	10.0	0.85	14.10	8.47	3.8	−1.25	1.90
F	0.85	1.79	9.6	0.88	13.43	9.03	2.8	−1.71	2.30
V	0.83	1.22	8.5	0.86	15.07	7.73	4.2	−0.46	1.50
W	0.67	2.25	9.2	0.85	12.95	7.66	−0.9	−2.09	2.60
Y	0.60	0.96	8.0	0.76	13.29	5.89	−1.3	−0.71	1.60
M	0.54	1.23	8.7	0.85	14.33	8.95	1.9	−0.67	2.40
A	0.33	0.31	5.1	0.74	12.28	5.33	1.8	0.50	0.67
P	0.32	0.72	4.9	0.64	11.19	3.87	−1.6	0.14	1.20
H	0.25	0.13	1.6	0.78	12.84	5.1	−3.2	2.33	0.64
T	0.21	0.26	3.5	0.70	11.65	4.49	−0.7	0.25	0.52
S	0.05	−0.04	3.1	0.66	11.26	4.09	−0.8	0.46	0.01
R	−0.01	−1.01	2.0	0.64	11.49	4.18	−4.5	1.81	−2.10
Q	−0.05	−0.22	1.4	0.62	11.28	3.87	−3.5	0.77	−0.22
N	−0.07	−0.6	0.6	0.63	11.00	3.71	−3.5	0.85	−0.6
D	−0.22	−0.77	0.7	0.62	10.97	3.59	−3.5	3.64	−1.2
E	−0.24	−0.64	1.8	0.62	11.19	3.65	−3.5	3.63	−0.76
K	−0.40	−0.99	1.3	0.52	10.8	2.95	−3.9	2.8	−0.57
r[c]		0.93	0.94	0.95	0.92	0.91	0.88	−0.82	0.77

[a] TW: This work.
[b] Value not reported by author.
[c] Correlation coefficients between the new scale and the others.

of the polar hydrogen. Thus, the MENV is divided into three regions separated by r_m and r_e (see Fig. 1). The values of r_m are as follows:

$$\text{i.}\quad r_m = r_a + 2.275\ \text{Å}$$

where r_a is the vdW radius of atom a belonging to the imbedded group A, and

$$\text{ii.}\quad r_e = r_m + 0.2\ \text{Å}$$

is the extended radius that defines the distance from each atom, a ($\in A$), where the contribution of atom b ($\in B$) goes to 0. The width of this extension was chosen to be narrow to avoid contributions beyond the first vdW interaction shell.

The total contribution, Hpy, calculated from all the atoms of the protein located in the MENV around the imbedded chemical group does not consider the relative position of the fragment inside the protein, i.e., the fraction exposed to the solvent. To account for the contributions from both the protein and solvent the total hydrophobicity index (THpy), is defined by

$$\text{THpy} = \zeta\text{Hpy} + (1-\zeta)\text{Hpy}^s \qquad (3)$$

where ζ is the fraction of the side chain or fragment buried in the protein, Hpy^s is the contribution to THpy from the solvent, and the buried fraction, ζ, is calculated using GEPOL93.[38] THpy is a size extensive descriptor, and to eliminate this dependency, it is normalized in the form rHpy = THpy/Hpy^s. The QPD, rHpy, is a relative measure of hydrophobicity (or hydrophilicity) of the MENV around any chemical group of the imbedded side chain. The value of rHpy ranges from ∼1, indicating a hydrophilic MENV approximately equivalent to water, to about −0.4, indicating an extremely hydrophobic MENV. It is noted that rHpy is a nonlocal descriptor while a Hscl assigns values to amino acids that are local descriptors.

To explore the role of these QPDs in modulating the properties of the imbedded side chains, rHpy values were calculated for 309,459 chemical groups in 733 proteins with a total of 204,200 amino acid residues (see Table I and supplementary materials for a list of the proteins in the data set). Only crystal structures of proteins with resolution better than 2.0 Å were selected from the PDB[39] and with sequence homology less than 25%.[40] Solvent, ions, and other hetero atoms were eliminated from the PDB structures and all hydrogen atoms were placed using the CHARMM[37] HBUILD command. No additional optimization of the H positions was carried out.

Determination of Hpy^s

The contribution of the solvent to THpy, Eq. (3), was determined for all chemical groups comprising the side chains as previously described for the titratable organic functionalities.[18] Each residue was capped with neutral fragments at both termini and immersed in a water droplet of about 415 TIP3 molcules. Simulations were carried out by heating these systems to 300 K and then running the systems for 200–300 ps until they appeared to be equilibrated. From the last 30 ps, 30 coordinate sets were extracted for each residue and the Hpy^s values were calculated for each chemical group (or side chain) using the above protocol and the water values of F_b.[18] From the 30 values for each chemical group, means were computed and the results taken as the solvent contribution for the given fragment. The results are given in Table SII of the supplementary material.

RESULTS AND DISCUSSION

Protein-based partitioning of amino acid side chains and chemical groups

To develop an amino acid Hscl based on the Hpy of the MENVs, it is assumed that in most cases the hydrophobicity of the imbedded amino acid more or less matches the hydrophobicity of its MENV. Thus, they are considered as "solvents" characterized by rHpy, and the residues of the protein are immersed to a greater or lesser degree in one or another of these local solvents. The working hypothesis is that the distribution of the side chains in different MENVs can be used to develop a partitioning based on their placement (ζ) and the relative hydrophobicity values (rHpy).

To test this hypothesis, each residue is associated with a point in a 2-D space defined by its values of ζ, and rHpy and a K-mean algorithm[41] is used to partition the residues into two classes. This algorithm associates all the data points with one of the two self-consistent centers. Carrying out the iterative procedure[42] it was found that one center was located deep inside the proteins (ζ = 0.90, rHpy = 0.20) while the other one was located near the surface (ζ = 0.39, rHpy = 0.68). It is also noted that various initial guesses of the locations of the two centroids were tried, but they all converged to the same final position given above. About 70% of the residues cluster around the former while the remaining 30% cluster around the latter center. This type of distribution is similar to the findings of a previous study where 69% of all groups from amino acids, including main chains, were at least 90% buried and the remaining 31% of groups were distributed between 0 and 90% burial.[32] This finding shows that the partitioning of the rHpy values correctly reproduces the well known structural fact that the large majority of residues are buried and hydrophobic.[43] However, to construct an Hscl for amino acids, the partitioning of an imbedded residue must also approximately match the Hpy of the MENV so that the partitioning can differ substantially from that of the overall distribution. Figure 2 shows the distribution for Lys around its two centroids with locations shown by the two squares. It is noted that the values converge to a single point because for complete solvent exposure the rHpy value is one, the water value. The plot consists of more that 13,000 points and 9368 of these are associated with the centroid of high solvent exposure and high rHpy value ($\zeta \sim 0.2$, rHpy $\sim$ 0.8), while the remaining 3745 cluster around the other centroid. These results indicate that the populations of the two classes forming the partitioning correctly reflect the polarity of the Lys side chain. Perhaps, the most surprising feature shown in the plot is the relatively large number of cases where the lysine side chain is deeply buried in a hydrophobic MENV. Analysis of the partitioning obtained from the other residues showed that in all cases the k-means algorithm used here gave results corresponding to the polar (or apolar) nature of the side chains. It also is seen that the partitioning does not yield two truly disjoint classes and a few points appear to be misassigned. In the present case, this misclassification involves at most around 100 points at or near the boundary between the two classes and therefore can be safely neglected. Nevertheless, misclassification using the k-means algorithm is an important issue and further discussion, especially on minimizing it, can be found in Ref. 44.

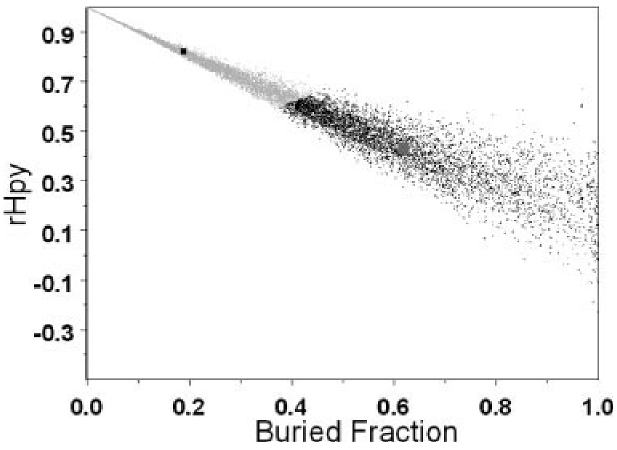

Figure 2

K-means algorithm distribution into a hydrophilic and hydrophobic class of the ammonium group of 13,113 Lys occurring in the 733 protein data base. The hydrophilic population of 9368 points clustered around the hydrophilic centroid (low buried fraction, high rHpy) is colored gray and the hydrophobic population of 3745 points clustered around the hydrophobic centroid (high buried fraction, low rHpy) is colored black. Location of the two centroids are marked by squares.

The population around the deeply buried center can be termed as hydrophobic and that around the exposed center as hydrophilic. The ratio of the populations from these two classes for the ith amino acid side chain gives a partitioning defined by

$$P_{\rm i} = N^{i}_{\rm hydrophobic}/N^{i}_{\rm hydrophilic} \tag{4}$$

that can be used to construct a HScl for the amino acid side chains based on the log(P_i) values defined by Eq. (4). It is given in Table II along with several other HScls. As shown in Table II, the new scale developed in this work correlates well with a number of previously proposed HScls[2,4,8,24,25,34–36] derived from various experimental and statistical methods with correlation coefficients ranging from 0.77 to 0.95 (see Table II). In particular, the scale correlates well with the two experimental scales[24,25] and also reasonably well with the experimental ΔG scale of White and collaborators.[4] The poorest correlation is with the Eisen scale[2] based on calculated ΔG values. However, it was already noted by Cornette *et al.*[1] that scales based on partition coefficients do not correlate as well with scales based on free energies as with each other. Cysteine was found to be the most hydrophobic residue according to the scale developed here, which agrees with the scale based on mean fractional area loss of amino acids by Rose and collaborators[45] as well as several others reported in Ref. 1. The most hydrophilic residue was found to be lysine in agreement with many pre-existing scales.

The result that the Hscl developed here correlates so well with many earlier scales is really quite striking. Almost all earlier scales are calibrated using one or another property of amino acid residues, e.g., octanol/water logP, ΔG, etc. In contrast, the MENVs are quantified using the Rekker Fragmental Constants,[27] quantities that were determined without any reference to proteins or amino acids. To construct the scale each side chain essentially "samples" a very large set of "solvents," which span a broad range of hydrophobicity values. Thus, the scale is determined completely in the context of proteins using parameters that are completely independent of proteins and their constituent elements. In addition, the scale does not involve the interaction of the residues with any solvent, but only with atoms of the protein in its MENV. Because of this, uncertainties in the values of polar side chains due to nonhydrophobic interactions with a particular solvent,[46] e.g., hydrogen bonding, are eliminated.

It is of interest to analyze the partitioning of the organic functionalities comprising the residue side chains, and to that end a group scale has been constructed, in exactly the same way as the residue scale; the results are presented in Table III. Note that Aboderin[25] (chromatography-based scale) had proposed a similar group based scale some time ago and those results are also given in Table III. The correlation between the two scales is quite good (0.9), although four values are missing in the Aboderin scale.[25] According to the group scale, the disulfide fragment in Cystine (TD) is more hydrophobic than the thiol group of Cysteine (TO). Free energies of transfer determined from cyclohexane/water partitioning for cysteine and cystine analogs have shown that burial of a disulfide is favored over the burial of two cysteines by 0.5 kcal/mol indicating less polarity of disulfide bonds compared to two cysteines.[47] The corresponding free energy difference calculated from the values in Table III for transferring TD and TO groups from protein interior to water was found to be 0.35 kcal/mol, in reasonable agreement with the aforementioned value. In another example, the octanol-to-water transfer free energies for charged side chains of lysine was about 1 kcal/mol lower than that of Arginine,[4] which can be compared with the free energy difference of 0.7 Kcal/mol calculated from Table III.

Table III
Hydrophobicity Scale for Chemical Groups in Amino Acid Side Chains

Group[a]	logP	ζ[b]	rHpy[b]	Abodr[c]
TD	1.31	0.92 (0.12)	0.21 (0.16)	[d]
TO	1.06	0.90 (0.17)	0.19 (0.20)	[d]
RS	0.82	0.84 (0.20)	0.17 (0.24)	4.5
HO	0.60	0.81 (0.25)	0.24 (0.27)	[d]
TE	0.59	0.81 (0.24)	0.21 (0.28)	0.6
PH	0.36	0.73 (0.28)	0.26 (0.35)	−1.6
H1	0.30	0.77 (0.29)	0.44 (0.26)	1.0
HS	0.25	0.70 (0.25)	0.38 (0.23)	−3.5
OL	0.10	0.68 (0.28)	0.45 (0.24)	−2.0
H2	−0.03	0.65 (0.25)	0.46 (0.22)	[d]
AD	−0.13	0.59 (0.27)	0.49 (0.24)	−4.6
GS	−0.19	0.55 (0.26)	0.49 (0.23)	−5.1
CO	−0.32	0.53 (0.26)	0.54 (0.23)	−4.4
AS	−0.70	0.38 (0.26)	0.65 (0.23)	−6.8

[a] See Figure 2 for group definitions.
[b] Mean and (standard deviation).
[c] Correlation coefficient between present scale and Abodr Scale[25] is 0.94.
[d] Values not reported by the author.

Insight into the effects on the properties of particular side chains with two functionalities can be obtained by comparing group with side chain logP values. First note that the effect of H1 seems to be small (compare logP values of N and Q, or D and E in Table II). In contrast, the three residues with aromatic rings exhibit substantial differences. In order of decreasing hydrophobicity our scale indicates F > W > Y, but there is little agreement on this ordering between the different scales. The ordering of the aromatic residues in our scale also indicates the increasing amphiphilic character of the MENV from F to Y residues due to the addition of more polar groups in the side chains. In spite of the fact that the phenol group in Tyr (Y_PH) is polar, it has quite a large population in solvent buried-low rHpy regions (see next section). This observation suggests that adjacent groups of a side chain affect each other's MENV because of restraints imposed by their being covalently bound. The presence of Y_PH groups deep inside the protein can be rationalized on the basis of multiple hydrogen bond formation and the consequent stabilization of the structure.[48] Comparison of the logP values of Thr and Ser indicates that the addition of the methyl group (HO) has a large effect on the logP value. It is also of interest that, in contrast to most scales, the logP value of Arg indicates that it is equally partitioned between the two centroids determined by the K-mean algorithm. Comparison of the residue and chemical group logP values of polar groups connected to $C\alpha$ by one or more CH_2 groups indicates that the latter have a fairly systematic effect on the logP values of the corresponding residues. Thus, comparing the logP values of GS and AS in Table III with the logP values of Arg and Lys in Table II shows that the $(CH_2)_3$ group in Lys and the $(CH_2)_2$ group in Arg shift the AS and GS values by 0.3 and 0.2 logP units, respectively. The log P values of the CO group of Glu and the AD group of Gln are −0.37 and −0.18, respectively, so that the residue values in Table II indicate that the single CH_2 group shift the logP values by about 0.1. The above findings suggest that the scale is physically reasonable supporting the validity of rHpy as a QPD. Some of the potential applications of this QPD are discussed in the following sections.

Population distribution of chemical groups in ζ-rHpy space

The successful construction of a HScl based on protein MENVs supports the well accepted hypothesis that on the average, matching the hydrophobicity of the side chain and the local environment is energetically favorable.[49] HScls provide an essentially thermodynamic description of the distribution of amino acid residues in the native structures of the proteins. The detailed probability distributions of the individual residues in the proteins can be obtained from histogram plots of the group populations as a function of ζ and rHpy. The normalized histogram plots in ζ-rHpy space for all the groups are presented in Figure 3 (the group plots for each residue, where a given group appears in more than one residue, are available as supplementary materials). The hydrophobic groups, TD, TO, RS, HO, and TE [Fig. 3(A–E)] are typically characterized by very high populations at the buried region with low or negative rHpy values and small populations in the more hydrophilic regions. The Y_PH (phenolic group of tyr) group [Fig. 3(F)] exhibits a smaller population in the buried-low rHpy regions, which is compensated by a significant population in the hydrophilic region. The amphiphilic groups, i.e., HS, OL, and H2 [Fig. 3(G–I)] exhibit a more uniform distribution, while groups such as GS, AD, CO, and AS [Fig. 3(J–M)] exhibit high populations in the hydrophilic regions (low values of ζ and high values of rHpy). There also appear to be distinct regions of the ζ, rHpy distribution that are prohibited to all the groups. These are characteristic of globular proteins that may not be conserved for transmembrane proteins.

Distributions of the same group from different side chains often are quite similar, but in some cases there can be substantial differences. The distribution of H1 groups mimics that of the adjacent side chain functionality in different amino acids. Therefore, the distribution of H1 groups connecting the aromatic side chains to the main chain in Phe, Tyr, and Trp (not shown) differ considerably from those in the polar and charged amino acids, Gln and Glu (not shown). The HO (terminal aliphatic) groups, excluding Ala and Thr, have the distribution typical for hydrophobic groups [Fig. 3(D)]. In contrast, the distribution of HO of Ala and Thr are much more diffuse (see Fig. 4) most likely because exposure of a single methyl group to solvent is less disfavored than larger aliphatic groups.[32]

Structure–function consequences of residue-MENV mismatch

The distributions of the groups shown in Figure 3 indicate that for the most part the hydrophobicities of the MENVs and their imbedded side chains are reasonably well matched. However, it is also seen that the distributions of almost all the groups contain some highly mismatched residue-MENV pairs as indicated by hydrophobic residues found in regions with high rHpy values or polar and titratable groups found in deeply buried regions with hydrophobic rHpy values. The wide range in the distribution of the amino acid residues ensures that a significant number of chemical groups in proteins will be mismatched. For the purpose of the following analysis, a chemical group where both ζ and rHpy deviate less than one standard deviation (σ) from the mean (see Table III) are considered here to be matched to their MENVs. Chemical groups outside this matched region comprise 10–20% of the total population in each group (Table I). Trp is an exception because of the large indole ring that enables this side chain to contact all amino acid residues with relatively high propensities.[50] From statistical studies, Trp was found to be the highest ranking conserved residue at the binding sites of protein–protein interactions.[51] Another contribution to the large percentage of mismatches for the Trp side chain may be its ability to participate in (albeit weak) hydrogen bonds. Since the H-bonding partner will usually be hydrophilic there is a greater chance for the MENV of Trp to be hydrophilic, contributing to the observed large number of mismatches.

Hydrophobic residues

In the sample of 733 proteins, 214 mismatched Cys were found. This number would be even greater if slightly lower resolution crystal structures were also included, e.g., exposed thiol groups of Cys in crystal structures of iron–sulfur clusters in thioredoxin.[52] These solvent exposed —SH groups are often found at the active sites in coordination with metal ions, e.g., Zn[53] or Fe, or present as a part of the heme group coordinating to the metal ion.[54,55] In addition, these groups form disulfide bonds upon dimerization or higher order oligomerization, which is a crucial step in many biological processes.[56,57] The mismatch of most of the hydrophobic side chains is due to their being almost entirely solvent exposed, residing at the protein surface. The rHpy values of these side chains is ∼1, and they constitute potential sites for protein–protein interactions as has been well documented.[51,58–62] However, there are also a

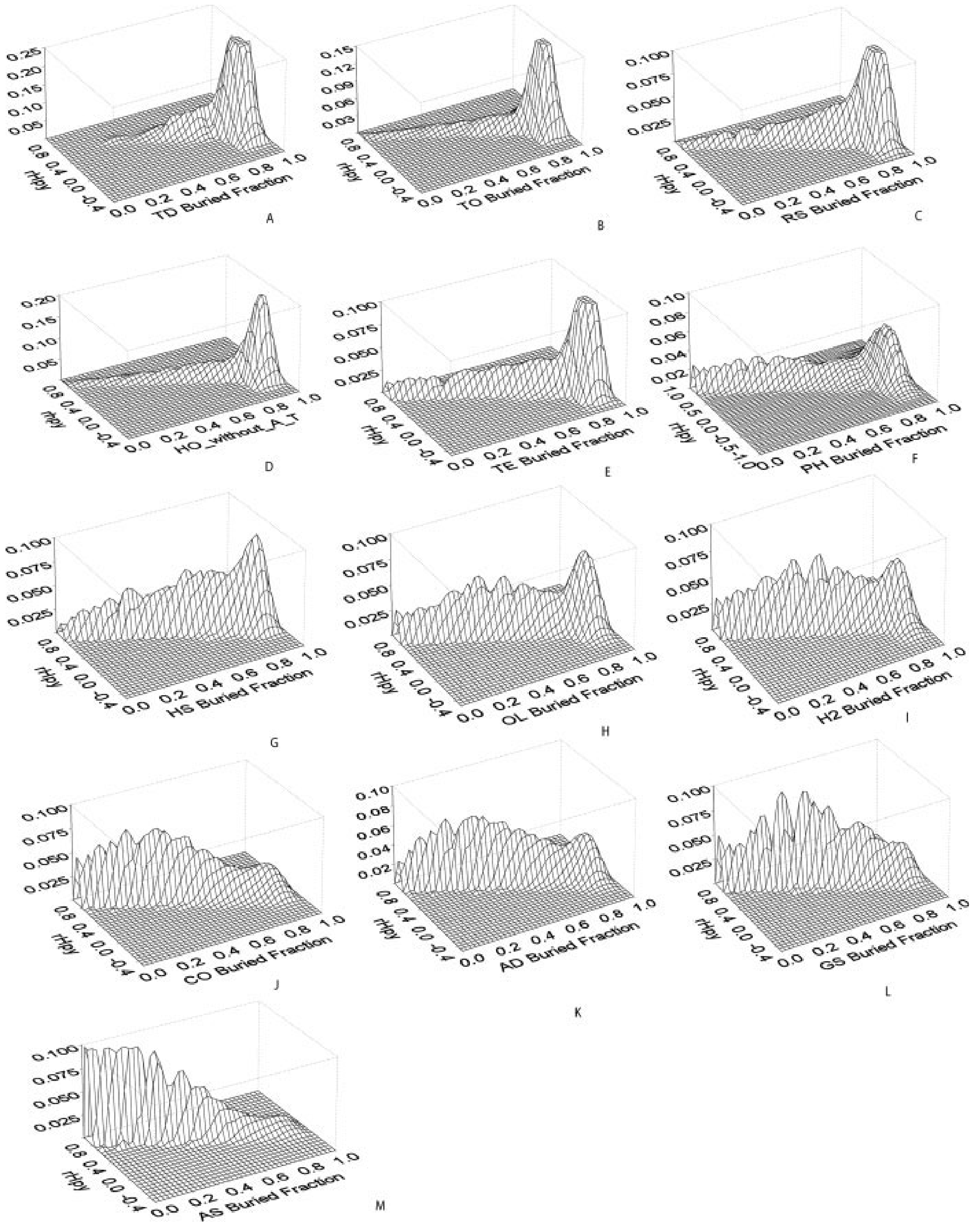

Figure 3

Histogram plots in ζ-rHpy space for the chemical groups comprising the amino acid side chains. The z-axis shows the normalized populations in each bin with respect to the total population for the given fragment. The plots are arranged from the most hydrophobic (TD) to the most hydrophilic (AS) (see also Table III). The hydrophobic groups (A–E) consist of TD, the disulfide bridge; TO, the thiol group; RS, the aromatic group; HO, the terminal aliphatic group; and TE, the thioether group. PH (F) is the phenolic group and the amphiphilic groups (G–I) comprise HS, the indole ring; OL, the aliphatic alcohol group; and H2. The final charged and polar groups are CO, the carboxyl group; AD, the amide group; GS, the guanidine group; and AS, the ammonium group. Each plot contains the contributions to the group from all residues containing the given group except for the HO group (D) which excludes the contributions from Ala and Thr (see text and Fig. 4). The side chains that contain a given group are given in Table I, footnote a.

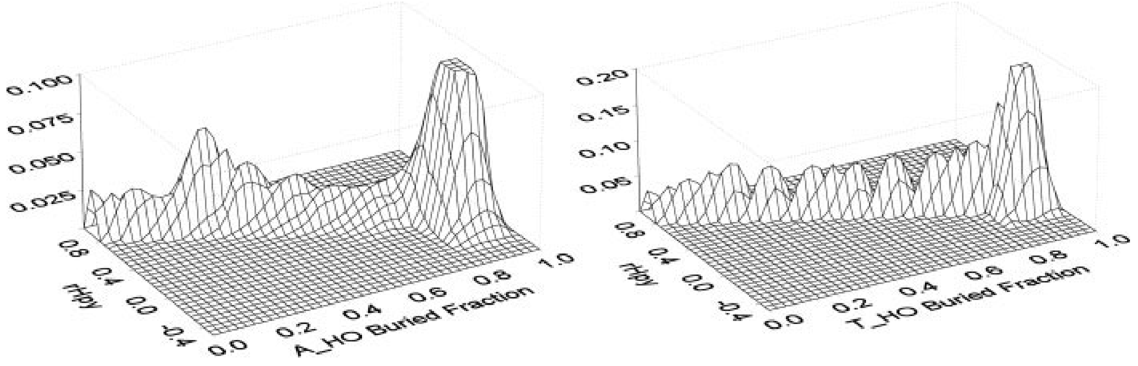

Figure 4

Histograms in ζ-rHpy space for the terminal aliphatic groups (HO) of Ala and Thr, which are shown separately because of their different distributions compared to the other residues with HO side chains (see text and Fig. 3).

few that are more deeply buried in the protein so that hydrophilic residues in their MENV contribute to the mismatch. For mismatched, hydrophobic residues with deviation $>2\sigma$, 80 (out of 5578) were found that were at least partially buried in the protein ($\zeta > 0.7$), but are imbedded in hydrophilic MENVs (rHpy > 0.3–0.4). It is noted that it was previously found[18] that when the degree of burial is < 0.7 the local environment is controlled by the aqueous solvent. In addition, rHpy values > 0.3–0.4 are already fairly hydrophilic in their effects on the imbedded residue. One interesting case is a leucine in *Thermotoga maritima* spermidine synthase. The residue is at the C-terminus of the protein with $\zeta = 0.97$ and rHpy $= 0.63$. No function has been identified for this residue but the system is the target of antimalaria drugs, so that perturbed, but conserved residues may be targets for nonactive site inhibitors. Interestingly, the sequence position is highly conserved (identity or homology) in other members of this family, many of which are involved in diseases. Another interesting case of deeply buried hydrophobic residues with relatively hydrophilic MENVs is met in the G-protein coupled receptors (see below).

Titratable residues

The response of titratable groups buried in mismatched (hydrophobic) MENVs is probably the most studied response of amino acids imbedded in mismatched MENVs.[19,63] We have calculated the pKa of the titratable residues in all the proteins in the data set of 733 proteins using the MM_SCP approach,[18] and the results from these calculations will be reported elsewhere. Mismatched MENVs around titratable residues often induce large pKa shifts compared to their normal values primarily due to the unfavorable solvation energies.[19] All the highly mismatched carboxylic acids (deviation $> 2\sigma$) yield upward pKa shifts by 1 to 4 pH units (solvent values for Asp and Glu are 4 and 4.4, respectively). This observation indicates that the carboxylate groups buried in hydrophobic MENVs often are protonated at physiological pH or in the cellular compartment where they are active. Interestingly, it was also found that 13 Lys residues out of our entire dataset have pKa values less than physiological pH, i.e., downward pKa shifts of more than 3 pH unit. Nine of these Lys residues are directly involved in biological function. The accuracy of the above calculated pKa values was verified by comparing calculated pKa values with experimentally known values where available.[64] Several proteins in the data set {lysozyme,[65] CO-Sperm Whale Myoglobin,[66] Subtilysin[67] and Barnase[68]} contain titratable residues for which the pKa have been measured; the RMS error for 32 values was found to be 0.8; it is noted that results have been reported for a larger subset of measured pKa where the RMS error was found to be around 0.5.[18]

Conservation of MENVs around structurally conserved functional amino acids

The observation that most chemical groups are matched to their MENVs suggests that the local environment around structurally conserved functional residues should also be conserved across protein families, independently of their sequence homology. This hypothesis was tested on several examples.

Adenylyl cyclases

The common metal binding sites of these proteins, obtained from mammalian (PDB: 1cjk),[69] bacterial (PDB: 1y11)[70] and pathogenic anthrax (PDB: 1k90),[71] and from T7 DNA polymerase (PDB: 1t7p) and Klenow fragment editing complex (PDB: 1kfd),[72] have pairwise sequence homology ranging from 20 to 40% (as determined by ClustalW[73]) (Table IV). They all have two

Table IV

RMSD (Å) of (A) two coordinated Asp at metal binding sites and (B) other catalytic groups in active site from adenylyl cyclases in different species and DNA polymerases

Species specification	Bacterial[74]	Mammalian	Anthrax	T7 DNA pol
A[a]				
Mammalian[75]	1.01			
Anthrax[76]	1.08	0.82		
T7 DNA pol[77]	0.74	1.24	0.90	
Klenow frag[77]	1.46	1.42	1.51	1.68
B[b]				
Mammalian	1.47			
Anthrax	7.43	7.06		

[a]Atoms superimposed in **A** are C_β-$C_\gamma O_{\sigma1} O_{\sigma2}$ from two D_CO group.
[b]Chemical groups in catalytic residues superimposed in **B** are K_AS, D_CO, N_AD and R_GS, respectively.

strictly conserved aspartic acids as part of their metal binding sites. The superposition of the titratable moieties (CO, see Table I) of the conserved aspartates yield low RMSD values of the superimposed carboxylate groups between 0.7 and 1.7 Å (Table IV(A)), indicating that the structures of this component of the metal binding site range from similar to nearly identical. The superposition of the functional groups in other highly conserved[78] catalytic residues of the three adenylyl cyclases yield small RMSD between the mammalian and bacterial adenylyl cyclases (Table IV(B)), but the RMSD between the catalytic residues of anthrax adenylyl cyclase and the bacterial or mammalian adenylyl cyclases are quite large (Table IV(B)).

According to our working hypothesis the MENV descriptors of these two carboxylate groups are also expected to show very little variation. To show this the means of the ζ and rHpy values have been calculated and their standard deviations are plotted in Figure 5. Although there is some variation in the σ values of the different sites, they are substantially smaller than the σ values of all residues (right most bars) showing that the MENVs of the active residues are well conserved.

Serine proteases

The second example of MENV conservation comes from the His-Ser-Asp catalytic triad of six serine protease families obtained from different species: bovine trypsin,[79] human hepsin, FVIIA-STF[k] and thrombin,[75] Hepatitis-C NS3 Chymotrypsin-like[74] and Dengue viruses NS3 trypsin-like,[76] with sequence identities of 30–40% (as determined by ClustalW[73]). The side chains of the His-Ser pair participate directly in the catalytic process of cleavage in all serine proteases while the Asp stabilizes the reaction intermediate in different ways[75] depending upon the nature of the substrates. The σ values of ζ and rHpy for all three residues are given in Figure 6, and are seen to be substantially smaller than the overall average value. Moreover, Asp exhibits a somewhat larger deviation in rHpy than the HS and OL groups (see Fig. 6), which may reflect its more varied function.

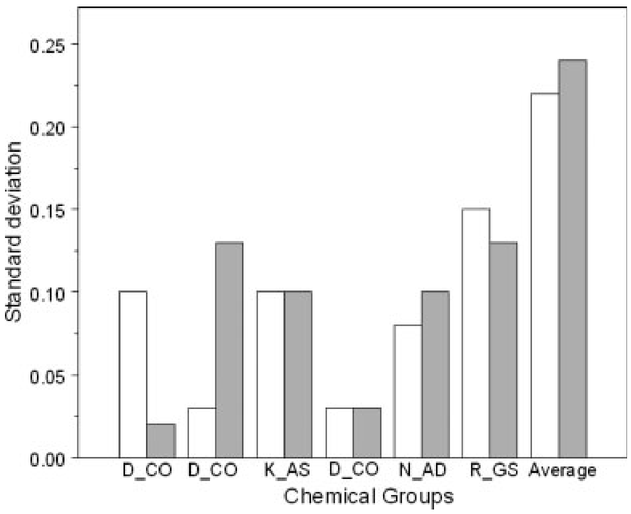

Figure 5

Standard deviations (σ) of the ζ and rHpy mean values in two metal coordinating D_CO groups (left) and other functionally important groups at the active site of adenylyl cyclases and DNA polymerase enzymes. The light bars represent σ values of ζ and the dark bars indicate σ values of rHpy.

Immunoglobulin fold

Another example considers one conserved Tryptophan residue in five different protein families {Tenasin,[77]

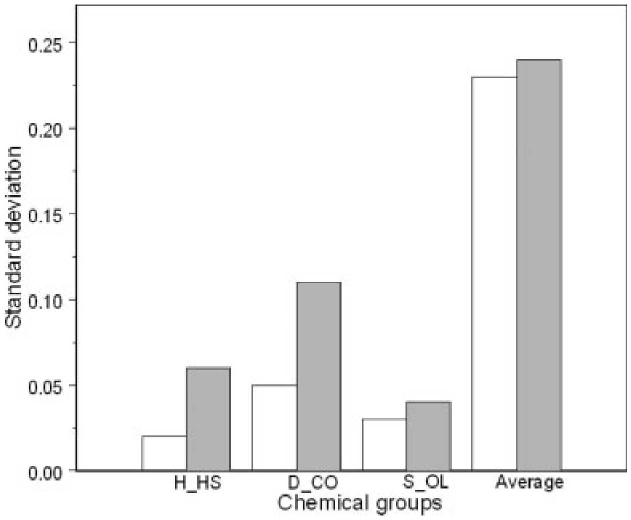

Figure 6

σ values of the means of ζ and rHpy in the catalytic triads of several serine proteases. See Figure 5 for bar code.

Table V
The ζ and rHpy Values of Chemical Groups of the Most Conserved TMH Residue Obtained from Members of Class I Rhodopsin Family GPCRs

	1GZM[b,85]		1L9H[b,84]		$5HT_2A$[90]		D_2R[88]		D_4R[89]	
Chemical groups[a]	ζ	rHpy	Z	rHpy	ζ	rHpy	ζ	rHpy	ζ	rHpy
N_AD (1.50)	1.0	0.04	1.0	0.15	1.0	0.04	1.0	0.19	1.0	0.17
D_CO (2.50)	0.96	0.29	0.96	0.40	0.99	0.38	1.0	0.31	0.99	0.41
R_H2 (3.50)	0.96	0.14	0.95	0.31	1.0	0.11	1.0	0.15	0.85	0.20
R_GS	0.95	0.13	0.96	0.09	0.80	0.23	0.97	0.10	0.92	0.11
W_RS (4.50)	0.78	0.25	0.71	0.30	0.43	0.54	0.76	0.30	0.58	0.59
P_H2 (5.50)	1.0	0.44	1.0	0.05	0.91	0.20	0.98	0.25	0.83	0.38
P_H2 (6.50)	0.53	0.54	0.58	0.54	0.53	0.56	0.50	0.60	0.55	0.49
P_H2 (7.50)	1.0,	0.27	1.0	0.37	0.98	0.30	0.90	0.46	0.89	0.41

[a]Conserved residue numbers in different transmembrane helices[93] is given in parenthesis.
[b]Crystal structures of rhodopsin with different space group symmetry.

CD8,[77] KB5-c20 t-cell receptor,[80] myelin p0 protein fragment,[81] and CD4[82]}, which structurally fits the immunoglobulin fold of those proteins.[83] These proteins mainly utilize hydrophobic and aromatic stacking interactions to stabilize the position of the conserved Trp in their respective structures, which is reflected in the small σ values of 0.02 and 0.08 of the mean ζ and rHpy values, respectively, calculated for the aromatic rings (RS) in this conserved Trp (the σ values for all chemical groups from the above mentioned proteins are 0.23 and 0.25, respectfully).

Most conserved residues in the rhodopsin family of G-protein coupled receptors (GPCR)

Results from the above sections support the hypothesis that MENVs around conserved, functional chemical groups of amino acids are also conserved, perhaps because evolution has found local structural arrangements that are optimal or provide favorable molecular orientations for a chemical reaction to occur. The MENV conservation hypothesis was applied to the most conserved residue in each of the seven transmembrane helices in two rhodopsin crystal structures[84–87] as well as model built structures of the dopamine receptors, D_2R[88] and D_4R[89] and the serotonin receptor, $5HT_{2A}R$.[90] Bovine rhodopsin shares about 19% overall sequence identity with the D_2 receptor and 25% in the TM portion.[91] The D_2 and D_4 receptors share 53% identity in the TMHs.[92] The ζ and rHpy values were calculated for the appropriate residues in the transmembrane helices, 1.50, 2.50…to 7.50 (for the numbering scheme see Ref. 93 which adopts the convention that the most conserved residue in the *n*th TMH is labeled n.50). The results reported in Table V show that the residues 1.50, 2.50, 3.50, 5.50, and 7.50 are deeply buried while 4.50 and 6.50 are more solvent exposed. The rHpy values range from hydrophobic, e.g., 1.50, to hydrophilic, e.g., 4.50 and 6.50. Interestingly Table V shows that in spite of the high degree of burial, the MENV of the three polar residues tend to be hydrophobic with smaller rHpy values, while the MENV of the 4 hydrophobic residues tend to be more hydrophylic with larger rHpy values. Whether or not this intriguing result is representative for membrane spanning proteins or a peculiarity of these GPCR remains to be clarified.

The means and σ values have been calculated for the side chains (major chemical groups) of the most conserved residues and the latter are presented in Figure 7, which shows that the variability of rHpy of the n.50 residues of helices 1, 2, 3, 6, and 7, is much less than the overall average value, while the MENVs around 4.50 and 5.50 are somewhat more variable. These two helices are connected by loop *el-2* that also forms a disulfide bond

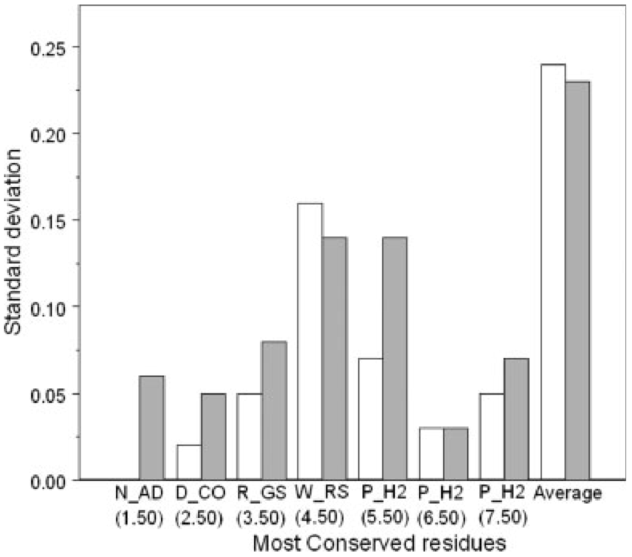

Figure 7
σ values of the means of ζ and rHpy in the most conserved residues of the trans-membrane helices of several GPCRs in the rhodopsin family (see Table V for details, and Fig. 5 for bar code).

to a Cys in TMH3. Since the position of the loop Cys relative to the *el-2* N- and C-termini is highly variable in different GPCRs, the higher σ values for 4.50 and 5.50 may reflect the changes in relative conformation of the two helices required to satisfy the additional constraint of the disulfide bridge. Nevertheless, for all helices the σ values are substantially less than the reference value.

CONCLUSIONS

The quantitative description of the MENVs reveals a complex mosaic of dielectric regions that clearly exhibit the heterogeneity of proteins so that this QPD may add an additional dimension in the study of structure and function at the molecular level. The MENVs are found, in almost all cases to be more hydrophobic than water, but perhaps less so than commonly thought. An important advantage of the HScl proposed here is its internal consistency because of its independence from nonprotein solvents, as discussed in a previous section. It is also clear that the present scale supports the notion of matching between the hydrophobicity of the residue-MENV pairs. However, from the probability distributions (see Fig. 3) of the chemical groups, it also was found that mismatch between the residue and MENV is quite common, and that the number of mismatches was 10–20% of the total number of residues for most groups. Residues imbedded in such mismatched MENV's often are structurally and/or functionally important, so that within the context of matching, the existence of mismatched residue-MENV pairs expresses the ability of evolution to create specialized protein architectures that modify the properties of imbedded residues (sometimes quite strongly) to meet some structural or functional need. It also should be noted that the ubiquituousness of mismatched pairs may affect statistical methods in unforeseeable ways since these are usually based on averages that makes it difficult to account for outliers because of inherent difficulties to predict, a priori, the mismatched pairs present in any particular protein.[94]

The availability of a protein based HScl taken together with a quantitative description of the Hpy of the local environment suggests the possibility of developing a high resolution analysis of the nonpolar contribution to the free energy. In this approach the free energy of a group, *A*, could be expressed as $G_A = -H_A \mathrm{rHpy}^A$,[95] where H_A is the HScl value of group *A*. Note that the minus sign is used so that matched Hpy values are favorable, while mismatched Hpy are unfavorable. The free energy of transferring the group from state *I* to state *J* would then be $\Delta G_A = H_A(\mathrm{rHpy}_I^A - \mathrm{rHpy}_J^A)$, where the transfer from *I* to *J* represents any change of state of group *A*, e.g., solvent to protein, monomer to dimer, etc. Because the evaluation of Hpy values is computationally very fast, such terms could be used to study energy contributions arising from changes in individual MENVs, e.g., on dimer formation, or be added to force fields to represent the hydrophobicity contribution to the free energy.

In this work the notion that side chains may consist of more than one chemical functionality has been emphasized to some extent because of its potential impact for developing course grained models. If the identification of chemical groups according to their functionality in side chains is physically meaningful as suggested by the analysis presented in this article, then, instead of representing each side chain by an arbitrary number of beads it would be reasonable to represent each chemical group by a single bead. Thus, the side chain of, say, Ile would be represented by one bead, whereas two beads would be assigned to Thr and similarly for the other side chains. Such a variable assignment has the advantage that a more physically realistic parameterization of the beads should be possible. Finally, another intriguing possibility for effecting altered function of a protein might be realized by mutating residues in the MENV of a given functional residue. Such an approach may allow effecting functional changes without undue disturbance of structure and stability often associated with mutating active, highly conserved residues.

ACKNOWLEDGMENTS

Computational support was provided by the National Science Foundation Terascale Computing System at the Pittsburgh Supercomputing Center. The authors also acknowledge access to the computer facilities at the Institute of Computational Biomedicine of Weill Medical College of Cornell University.

REFERENCES

1. Cornette JL, Cease KB, Margalit H, Spouge JL, Berzofsky JA, DeLisi C. Hydrophobicity scale and computational techniques for detecting amphipathic structures in proteins. J Mol Biol 1987;195:659–685.
2. Eisenberg D, McLachlan AD. Solvation energy in protein folding and binding. Nature 1986;319:199–203.
3. Wilce MCJ, Aguilar MI, Hearn MTW. Physicochemical basis of amino-acid hydrophobicity scales—evaluation of 4 new scales of amino-acid hydrophobicity coefficients derived from rp-hplc of peptides. Anal Chem 1995;67(7):1210–1219.
4. Wimley CW, Creamer TP, White SH. Solvation energies of amino acid side chains and backbone in a family of host-guest pentapeptides. Biochemistry 1996;35:5109–5124.
5. Dill KA. Dominant forces in protein folding. Biochemistry 1990; 29(31):7133–7155.
6. Gruebele M. Downhill protein folding: evolution meets physics. CR Biol 2005;328:701–712.
7. Mehdi Y, Burge CB, Kardar M. Untangling influences of hydrophobicity on protein sequences and structures proteins: Struct Funct Bioinformatics 2006;62(4):1101–1106.
8. Ponnuswamy PK, Prabhakaran M, Manavalan P. Hydrophobic packing and spatial arrangement of amino acid residues in globular proteins. Biochim Biophys Acta 1980;623:301–316.
9. Bowie JU, Luthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. Science 1991;253:164–170.

10. Kleiger G, Beamer LJ, Grothe R, Mallick P, Eisenberg D. The 1.7 Acrystal structure of BPI: a study of how two dissimilar amino acid sequences can adopt the same fold. J Mol Biol 2000;299(4):1019–1034.
11. Leo A, Jow PYC, Silipo C, Hansch C. Calculation of hydrophobic constant (logP) from π and f-Constants. J Med Chem 1975;18:865–868.
12. Abraham DJ, Leo AJ. Extension of the fragment method to calculate amino acid zwitterion and side chain partition coefficients. Proteins 1987;2:130–152.
13. Kellogg GE, Semus SF, Abraham DJ. HINT: a new method of empirical hydrophobic field calculation for CoMFA. J Comput Aided Mol Des 1991;5(0920-654X (Print)):545–552.
14. Abraham DJaK GE. The effect of physical organic properties on hydrophobic fields. J Comput Aided Mol Des 1994;8:41–49.
15. Spyrakis F, Cozzini P, Bertoli C, Marabotti A, Kellogg GE, Mozzarelli A. Energetics of the protein-DNA-water interaction. BMC Struct Biol 2007;7:4.
16. Kellogg GE, Fornabaio M, Chen DL, Abraharn DJ, Spyraki SF, Cozzini P, Mozzarelli A. Tools for building a comprehensive modeling system for virtual screening under real biological conditions: the computational titration algorithm. J Mol Graph Model 2006;24(6): 434–439.
17. Kellogg GE, Phatak S, Nicholls AJ, Grant A. Validation of Poisson-Boltzmann electrostatic potential fields in 3D QSAR: a CoMFA study on multiple datasets. QSAR Comb Sci 2003;22(9/10):959–964.
18. Mehler EL, Guarnieri F. A self-consistent microenvironment modulated screened coulomb potential approximation to calculate ph dependent electrostatic effects in proteins. Biophysics J 1999;77:3–22.
19. Mehler EL, Fuxreiter M, Simon I, Garcia-Moreno EB. The role of hydrophobic microenvironment in modulating pka shifts in proteins. Proteins 2002;48:283.
20. Lakowicz JR. Principles of fluoresence spectroscopy. New York: Plenum; 1983.
21. Haque ME, Ray S, Chakrabarti A. Polarity estimate of the hydrophobic binding sites in erythroid spectrin: a study by pyrene fluorescence. J Fluoresc 2000;V10(1):1–6.
22. Ray S, Bhattacharyya M, Chakrabarti A. Conformational study of spectrin in presence of submolar concentrations of denaturants. J Fluoresc 2005;V15(1):61–70.
23. Mariya K, Alexander K. Conformational flexibility of cytokine-like C-module of tyrosyl-tRNA synthetase monitored by Trp144 intrinsic fluorescence. J Fluoresc 2006;V16(5):705–711.
24. Fauchere JL Pliska V. Hydrophobic parameters p of amino acid side chains from the partitioning of N-acetyl-amino acid amides Eur J Med Chem 1983;18:369–375.
25. Aboderin AA. An empirical hydrophobicity scale for [alpha]-amino-acids and some of its applications. Int J Biochem 1971;2(11): 537–544.
26. Rekker RF. The hydrophobic fragmental constant: an extension to a 1000 data point set. Eur J Med Chem 1979;14:479–488.
27. Rekker RF. The hydrophobic fragmental constant. Nauta WT, Rekker RF, editors. Amsterdam: Elsevier; 1977.
28. Dinner AR. Local deformations of polymers with nonplanar rigid main-chain internal coordinates. J Comput Chem 2000;21(13): 1132–1144.
29. Suzuki T, Kudo Y. Automatic Log P estimation based on combined additive modeling methods. J Comput Aided Mol Des 1990;4:155–198.
30. Ghose AK, Crippen GM. Atomic physicochemical parameters for three-dimensional structure-directed quantitative structure-activity relationships, I. Partition coefficients as a measure of hydrophobicity. J Comp Chem 1986;7:565–577.
31. Mannhold R, Rekker RF, Sonntag C, ter Laak AM, Dross K, Polymeropouos EE. Comparative evaluation of the predictive power of calculation procedures for molecular lipophilicity. J Pharm Sci 1995; 84:1410–1419.
32. Lesser GJ, Rose GD. Hydrophobicity of amino-acid subgroups in proteins. Proteins 1990;8(1):6–13.
33. Richards FM. Areas. Volumes packing, and protein structure. Annu Rev Biophys Bioeng 1977;6(1):151–176.
34. Rose GD, Geselowitz AR, Lesser GJ, Lee RH, Zehfus MH. Hydrophobicity of amino acid residues in globular proteins. Science 1985;229(4716):834–838.
35. Sanzo M, Jernigan RL. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules 1985;18(3):534–552.
36. Kyte J, Doolittle RF. A simple model for displaying the hydropathic character of a protein. J Mol Biol 1982;157:105–132.
37. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: a program for macromolecular energy, minimization and dynamics calculations. J Comp Chem 1983;4: 187–217.
38. Pascual-Ahuir JL, Silla E, Tunon I. GEPOL: an improved description of molecular surfaces, III. A new algorithm for computation of a solvent-excluding surface. J Comput Chem 1994;15:1127–1138.
39. Berman HM, Bhat TN, Bourne PE, Feng Z, Gilliland G, Weissig H, Westbrook J. The Protein Data Bank and the challenge of structural genomics. Nat Struct Biol 2000;7 Suppl:957–959.
40. Hobohm U, Sander C. Enlarged representative set of protein structures. Protein Sci 1994;3:522.
41. MacQueen JB. Some methods for classification and analysis of multivariate observations. Berkeley, USA: University of California Press; 1967. pp 281–297.
42. Arthur D, Vassilvitskii S.How slow is the k-means method? Proceedings of the 2006 Symposium on Computational Geometry (SoCG) June 5–7, 2006; Sedona, AR.
43. Janin J. Surface and inside volumes in globular proteins. Nature 1979;277(5696):491–492.
44. Crawley MJ. The R book. West Sussex: Wiley; 2007.
45. Rose GD, Gierasch LM, Smith JA. Turns in peptides and proteins. Adv Prot Chem 1985;37:1–105.
46. Karplus PA. Hydrophobicity regained. Protein Sci 1997;6(6):1302–1307.
47. Saunders AJ, Young GB, Pielak GJ. Polarity of disulfide bonds. Protein Sci 1993;2(7):1183–1184.
48. Pace CN, Horn G, Hebert EJ, Bechert J, Shaw K, Urbanikova L, Scholtz JM, Sevcik J. Tyrosine hydrogen bonds make a large contribution to protein stability. J Mol Biol 2001;312(2):393–404.
49. Southall NT, Dill KA, Haymet ADJ. A view of the hydrophobic effect. J Phys Chem B 2002;106(10):2812–2812.
50. Samanta U, Pal D, Chakrabarti P. Environment of tryptophan side chains in proteins. Proteins 2000;38(3):288–300.
51. Ma B, Elkayam T, Wolfson H, Nussinov R. Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. PNAS 2003;100(10):5772–5777.
52. Collet J-F, Peisach D, Bardwell JCA, Xu Z. The crystal structure of TrxA(CACA): insights into the formation of a [2Fe-2S] iron-sulfur cluster in an *Escherichia coli* thioredoxin mutant. Protein Sci 2005; 14(7):1863–1869.
53. Blom NS, Tetreault S, Coulombe R, Sygusch J. Novel active site in *Escherichia coli* fructose 1,6-bisphosphate aldolase. Nat Struct Mol Biol 1996;3(10):856–862.
54. Ozawa K, Takayama Y, Yasukawa F, Ohmura T, Cusanovich MA, Tomimoto Y, Ogata H, Higuchi Y, Akutsu H. Role of the Aromatic Ring of Tyr43 in tetraheme cytochrome c3 from *Desulfovibrio vulgaris* Miyazaki F. Biophys J 2003;85(5):3367–3374.
55. Pokkuluri PR, Londer YY, Duke NEC, Long WC, Schiffer M. Family of cytochrome c_7-type proteins from *Geobacter sulfurreducens*: structure of one cytochrome c_7 at 1.45 Å Resolution. Biochemistry 2004; 43(4):849–859.
56. Patel SD, Rajala MW, Rossetti L, Scherer PE, Shapiro L. Disulfide-dependent multimeric assembly of resistin family hormones. Science 2004;304(5674):1154–1158.

57. Bieger B, Essen L-O. Structural analysis of adenylate cyclases from *Trypanosoma brucei* in their monomeric state. *EMBO J* 2001;20(3):433–445.
58. Anashkina A, Kuznetsov E, Esipova N, Tumanyan V. Comprehensive statistical analysis of residues interaction specificity at protein-protein interfaces. Proteins 2007;67:1060–1077.
59. Conte LL, Chothia C, Janin J. The atomic structure of ptotein-protein recognition sites. J Mol Biol 1999;285:2177–2198.
60. Jones S, Thornton JM. Analysis of protein-protein interaction sites using surface patches. J Mol Biol 1997;272(1):121–132.
61. Nooren IMA, Thornton JM. Diversity of protein-protein interactions. EMBO J 2003;22:3486–3492.
62. José LJ. Does structural and chemical divergence play a role in precluding undesirable protein interactions? Proteins 2005;59(4):757–764.
63. Schutz CN, Warshel A. What are the dielectric “constants” of proteins and how to validate electrostatic models? Proteins 2001;44:400–417.
64. Wisz MS, Hellinga HW. An empirical model for electrostatic interactions in proteins incorporating multiple geometry-dependent dielectric constants. Proteins 2003;51:360–377.
65. Walsh MA, Schneider TR, Sieker LC, Dauter Z, Lamzin VS, Wilson KS. Refinement of triclinic hen egg-white lysozyme at atomic resolution Acta Cryst D 1998;54:522–546.
66. Chu K, Berendzen J, Sweet RM, Schlichting I. Crystal structures of myoglobin-ligand complexes at near-atomic resolution. Biophys J 1999;77(4):2153–2174.
67. Kuhn P, Knapp M, Soltis SM, Ganshaw G, Thoene M, Bott R. The 0.78 Å structure of a serine protease: *Bacillus lentus* Subtilisin. Biochemistry 1998;37(39):13446–13452.
68. Mauguen Y, Hartley RW, Dodson EJ, Dodson GG, Bricogne G, Chothia C, Jack A. Molecular structure of a new family of ribonucleases. Nature 1982;297(5862):162–164.
69. Tesmer JJ, Nbsp G, Sunahara RK, Johnson RA, Gosselin G, Gilman AG, Sprang SR. Two-metal-ion catalysis in adenylyl cyclase. Science 1999;285(5428):756–760.
70. Tews I, Findeisen F, Sinning I, Schultz A, Schultz JE, Linder JU. The structure of a pH-sensing mycobacterial adenylyl cyclase holoenzyme. Science 2005;308(5724):1020–1023.
71. Drum CL, Yan S-Z, Bard J, Shen Y-Q, Lu D, Soelaiman S, Grabarek Z, Bohm A, Tang W-J. Structural basis for the activation of anthrax adenylyl cyclase exotoxin by calmodulin. Nature 2002;415:396–402.
72. Doublie S, Tabor S, Long AM, Richardson CC, Ellenberger T. Crystal structure of a bacteriophage T7 DNA replication complex at 2.2[thinsp][angst] resolution. Nature 1998;391(6664):251–258.
73. Jeanmougin F, Thompson JD, Gouy M, Higgins DG, Gibson TJ. Multiple sequence alignment with Clustal X. *Trends Biochem Sci* 1998;23(10):403–405.
74. Barbato G, Cicero DO, Cordier F, Narjes F, Gerlach B, Sambucini S, Grzesiek S, Matassa VG, Francesco RD, Bazzo R. Inhibitor binding induces active site stabilization of the HCV NS3 protein serine protease domain. EMBO J 2000;19:1195–1206.
75. Katz BA, Luong C, Ho JD, Somoza JR, Gjerstad E, Tang J, Williams SR, Verner E, Mackman RL, Young WB. Dissecting and designing inhibitor selectivity determinants at the S1 site using an artificial Ala190 protease (Ala190 uPA). J Mol Biol 2004;344(2):527–547.
76. Murthy HMK, Clum S, Padmanabhan R. Dengue virus NS3 serine protease. Crystal structure and insights into interaction of the active site with substrates by molecular modeling and structural analysis of mutational effects. J Biol Chem 1999;274(9):5573–5580.
77. Leahy DJ, Axel R, Hendrickson WA. Crystal structure of a soluble form of the human T cell coreceptor CD8 at 2.6 A resolution. Cell 1992;68(6):1145–1162.
78. Linder JU, Schultz JE. The class III adenylyl cyclases: multi-purpose signalling modules. Cell Signal 2003;15(12):1081–1089.
79. Katz BA, Elrod K, Verner E, Mackman RL, Luong C, Shrader WD, Sendzik M, Spencer JR, Sprengeler PA, Kolesnikov A, Tai VWF, Hui HC, Guy Breitenbucher J, Allen D, Janc JW. Elaborate manifold of short hydrogen bond arrays mediating binding of active site-directed serine protease inhibitors. J Mol Biol 2003;329(1):93–120.
80. Housset D, Mazza G, Gregoire C, Piras C, Malissen B, Fontecilla-Camps JC. The three-dimensional structure of a T-cell antigen receptor $V\alpha V^{\beta}$ heterodimer reveals a novel arrangement of the V^{β} domain. EMBO J 1997;16:4205–4216.
81. Shapiro L, Doyle JP, Hensley P, Colman DR, Hendrickson WA. Crystal structure of the extracellular domain from P0, the major structural protein of peripheral nerve myelin. Neuron 1996;17(3):435–449.
82. Brenner SE, Chothia C, Hubbard TJ. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. Proc Natl Acad Sci USA 1998;95(11):6073–6078.
83. Mirny LA, Shakhnovich EI. Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. J Mol Biol 1999;291(1):177–196.
84. Okada T, Fujiyoshi Y, Silow M, Navarro J, Landau EM, Shichida Y. Functional role of internal water molecules in rhodopsin revealed by X-ray crystallography. PNAS 2002;99:5982–5987.
85. Li J, Edwards PC, Burghammer B, Villa C, Schertler GFX. Structure of bovine rhodopsin in a trigonal crystal form. J Mol Biol 2004;343:1409–1438.
86. Okada T, Sugihara M, Bondar AN, Elstner M, Entel P, Buss V. The retinal conformation and its environment in rhodopsin in light of a new 2.2 A crystal structure. J Mol Biol 2004;342:571–583.
87. Palczewski K, Kumasaka T, Hori T, Behnke CA, Motoshima H, Fox BA, LeTrong I, Teller DC, Okada T, Stenkamp RE, Yamamoto M, Miyano M. crystal structure of rhodopsin: A G-protein coupled receptor. Science 2000;289:739–745.
88. Floresca CZ, Chen SW, Kortagere S, Schetz JA Reciprocal mutations in TM2/TM3 in a D2 dopamine receptor background confirms the importance of this microdomain as a selective determinant of para-halogenated 1,4-disubstituted aromatic piperazines. Arch Pharm 2005;338(5/6):268–227.
89. Kortagere S, Gmeiner P, Weinstein H, Schetz JA. Certain 1,4-disubstituted aromatic piperidines and piperazines with extreme selectivity for the dopamine D4 receptor interact with a common receptor microdomain. Mol Pharmacol 2004;66(6):1491–1499.
90. Visiers I, Ballesteros JA, Weinstein H. Three dimensional representations of GPCR structures and mechanisms. In: Iyengar I, Hildebrandt J, editors. Methods in Enzymology, Vol. 343. New York: Academic Press; 2002. pp. 329–371.
91. Ballesteros JA, Shi L, Javitch JA. Structural mimicry in g protein-coupled receptors: implications of the high-resolution structure of rhodopsin for structure-function analysis of rhodopsin-like receptors. Mol Pharmacol 2001;60(1):1–19.
92. Gingrich JA, Caron MG. Recent advances in the molecular biology of dopamine receptors. Ann Rev Neurosc 1993;16(1):299–321.
93. Ballesteros JA, Weinstein H. Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein-coupled receptors. Meth Neurosci 1995;25:366–428.
94. Mooney SD, Liang, MH-P, DeConde R, Altman RB. Structural characterization of proteins using residue environments. Proteinsorm 2005;61:741–747.
95. Eisenberg D, Wilcox W, McLachlan AD. Hydrophobicity and amphiphilicity in protein structure. J Cell Biochem 1986;31(1):11–17.