PROTEINS: Structure, Function, and Genetics 37:379–387 (1999)

Classification of Protein Sequences by Homology Modeling and Quantitative Analysis of Electrostatic Similarity

Niklas Blomberg, Razif R. Gabdoulline, Michael Nilges,* and Rebecca C. Wade
European Molecular Biology Laboratory, Heidelberg, Germany

ABSTRACT **Protein electrostatics plays a key role in ligand binding and protein–protein interactions. Therefore, similarities or dissimilarities in electrostatic potentials can be used as indicators of similarities or dissimilarities in protein function. We here describe a method to compare the electrostatic properties within protein families objectively and quantitatively. Three-dimensional structures are built from database sequences by comparative modeling. Molecular potentials are then computed for these with a continuum solvation model by finite difference solution of the Poisson-Boltzmann equation or analytically as a multipole expansion that permits rapid comparison of very large datasets. This approach is applied to 104 members of the Pleckstrin homology (PH) domain family. The deviation of the potentials of the homology models from those of the corresponding experimental structures is comparable to the variation of the potential in an ensemble of structures from nuclear magnetic resonance data or between snapshots from a molecular dynamics simulation. For this dataset, the results for analysis of the full electrostatic potential and the analysis using only monopole and dipole terms are very similar. The electrostatic properties of the PH domains are generally conserved despite the extreme sequence divergence in this family. Notable exceptions from this conservation are seen for PH domains linked to a Dbl homology (DH) domain and in proteins with internal PH domain repeats. Proteins 1999;37:379–387.** © 1999 Wiley-Liss, Inc.

Key words: PH domains; signal transduction; electrostatics; similarity index; homology modeling; phospholipid binding

INTRODUCTION

The function of many proteins is governed by their electrostatic properties. In particular, ligand–protein binding and protein–protein interactions are often determined by electrostatic forces.[1] As two molecules approach, the electrostatic potentials determine their relative orientation, whereas on binding, charge–charge and hydrogen bonding interactions contribute to binding specificity and affinity. Similar electrostatic potentials from a set of proteins may indicate similar behavior and function. For example, electrostatic similarity led to the suggestion of a common cell-recognition function for cholinesterases and structurally similar but catalytic inactive cell-adhesion proteins.[2] Another example is the identification of functionally equivalent sites on the small copper proteins cytochrome c_6 and plastocyanin[3] by superposition based on electrostatic potentials. The analysis of regions with similar electrostatic potentials from triose phosphate isomerases,[4] Cu,Zn-superoxide dismutases, and class A β-lactamases[5] from several organisms allowed identification of residues responsible for diffusion-controlled catalysis. The presence or absence of the symmetry in electrostatic potentials of cytokines was analyzed to explain and predict their homo- or hetero-dimeric receptor binding properties.[6]

Our recent systematic study of Pleckstrin homology (PH) domains[7] showed striking electrostatic differences between the domains of different proteins, indicating that the function of the domain has diverged during evolution. In that study, we analyzed a single component of the electrostatic potential, the dipole moment, which depends simply on the average separation of charges. Models of the PH domain sequences could be clustered by analyzing the direction of the dipole moment in relation to the phospholipid-binding sites reported for spectrin and phospholipase C-δ (PLC-δ) PH domains. In about a quarter of the domains, the dipole moment pointed away from the phospholipid-binding site. Visual inspection of the electrostatic potentials from this cluster of PH domains indicated that binding of phospholipids to the binding site described for the spectrin and PLC-δ PH domains is unlikely, because this site is completely surrounded by negative electrostatic potential.

This simplified description of the molecular electrostatic potential seemed appropriate for the PH domains because the electrostatic potentials of domains with experimentally determined structures show strong polarization with simple electrostatic potential distributions. This is not the case for proteins or protein domains in general. In order to apply the same type of analysis to other proteins and protein domains with potentially more complicated electrostatic properties in an automated way, it would clearly be desirable to use a more detailed comparison of the potentials. Sequence databases grow rapidly and will provide us with a growing number of protein sequences with unknown function. It is therefore necessary to develop techniques that allow rapid and automated analysis of protein properties, to draw conclusions about possible functions, and to suggest directions for experimental studies.

*Correspondence to: Michael Nilges, European Molecular Biology Laboratory, Meyerhofstr. 1, D-69117 Heidelberg, FRG.
Received 22 March 1999; Accepted 26 June 1999

In this study, we investigate the use of similarity indices,[8] and analytical approximations of the electrostatic potentials that go beyond the dipole moment description. We investigate the applicability of these methods to the comparison of homology models derived from structural templates with low sequence identity to the target sequence. These quantitative methods are tested using our database of PH domain models, and the results are compared to those of our earlier study.

THEORY

Similarity Indices for Molecular Electrostatic Potentials

Molecular electrostatic potentials can be compared quantitatively by calculating similarity indices.[8,9] Similarity indices have been developed for, and are usually applied to, the comparison of small molecules.[10–12] They are, however, also useful for the analysis of macromolecules, allowing classification as well as prediction of the functional properties and active sites of macromolecules.[3–6,13]

The Hodgkin index[8] is commonly used to measure the similarity of two molecular potentials. It detects differences in sign, magnitude, and spatial behavior in the potentials and is given by

$$SI_{12} = \frac{2(\mathbf{p}_1, \mathbf{p}_2)}{(\mathbf{p}_1, \mathbf{p}_1) + (\mathbf{p}_2, \mathbf{p}_2)} \tag{1}$$

where $(\mathbf{p}_1, \mathbf{p}_2)$, $(\mathbf{p}_1, \mathbf{p}_1)$, and $(\mathbf{p}_2, \mathbf{p}_2)$ are the scalar products of the electrostatic potentials over the region where the potentials are compared. Thus, $SI_{12} = 1$ if the two potentials are identical; $SI_{12} = 0$ if they are fully uncorrelated; $SI_{12} = -1$ if they are anti-correlated.

If the electrostatic potentials (ϕ) are computed on a three-dimensional grid over the two superimposed molecules, the scalar product is

$$(\mathbf{p}_1, \mathbf{p}_2) = \sum_{i,j,k} \phi_1(i, j, k)\phi_2(i, j, k) \tag{2}$$

where the summation is carried out over the grid points (i, j, k) that are within the region of interest. The regions in this study were chosen to represent a "skin" around the molecules having thickness δ. The "skin" of each molecule is defined as the volume remaining after excluding the region inside the surface accessible to the center of a probe with radius σ, and the region outside the surface accessible to a probe of radius $\sigma + \delta$. The "region of interest" is the intersection of the "skins" of the two molecules. Throughout the study we used $\sigma = 2$ Å and $\delta = 1$ Å to represent the relevant interaction properties of proteins.

Analytical Expression for Electrostatic Potential Similarity Indices

In this article, we propose a simple analytical expression for comparison of two molecular electrostatic potentials. If the molecular dipole alone is expected to define the electrostatic potential, then formula 1 may be used for SI_{12} with $\mathbf{p}$ being $\mathbf{d}$—the molecular dipole moment. For proteins having net charges, the monopole term may be comparable to or dominate the dipole term. Therefore, we introduce a similarity measure taking into account not only the dipole moments of the molecules but also their net charges. We assume the electrostatic potential of each molecule to be given by its charge, q, and dipole moment, d, i.e. [omitting the $(1/4\pi\epsilon_0\epsilon)$ scaling factor]:

$$\phi(r) = \frac{q}{r} + \frac{(\mathbf{d}, \mathbf{r})}{r^3} \tag{3}$$

The scalar product of two such potentials over a spherical layer of thickness δ is then

$$\int_R^{R+\delta} \int_\Omega \phi_1(r)\phi_2(r)r^2 \, dr \, d\omega = q_1q_2 \int_R^{R+\delta} \int_\Omega dr \, d\omega + \frac{(\mathbf{d}_1, \mathbf{d}_2)}{3} \int_R^{R+\delta} \int_\Omega \frac{1}{r^2} dr \, d\omega \tag{4}$$

where dr and $d\omega$ denote integration over radial and angular variables, respectively. In contrast to the comparison of Poisson-Boltzmann potentials, r is integrated over a thin spherical layer close to the chosen radius. Integration of $d\omega$ is carried out over the full range of the celestial angle Ω.

Introducing the parameter

$$\alpha = \frac{1}{3} \int_R^{R+\delta} \int_\Omega \frac{1}{r^2} dr \, d\omega \Big/ \int_R^{R+\delta} \int_\Omega dr \, d\omega \tag{5}$$

dependent on the spherical layer in which the potentials are compared, the similarity index (1) taking into account both the monopole and dipole moments of the molecules is

$$SI_{12} = \frac{2(q_1q_2 + \alpha(\mathbf{d}_1, \mathbf{d}_2))}{q_1q_1 + q_2q_2 + \alpha((\mathbf{d}_1, \mathbf{d}_1) + (\mathbf{d}_2, \mathbf{d}_2))}. \tag{6}$$

If the layer in which the potentials are compared is thin and has a radius of R, then the parameter α is simply $\alpha = 1/3R^2$. The optimal value of this parameter was obtained by fitting of the similarity indices derived from electrostatic potential grids to those derived from equation 6. The value giving the best correlation for PH domains is $\alpha = (17\text{Å})^{-2} = \frac{1}{3}(9.815\text{Å})^{-2}$, and thus R is comparable to the radii of gyration of these proteins, 12–13 Å. A general expression for comparing the molecular electrostatic potentials in a specified region of space, e.g., an enzyme active site, is described in Appendix A.

METHODS

Protein Structures

The homology models of the PH domain family were taken from our previous study[7] and can be accessed via Internet (http://www.nmr.embl-heidelberg/~blomberg/PHdomains) together with validation data. Details of the solvated molecular dynamics simulation of the PH domain from mouse β-spectrin are published in reference 14.

Electrostatic Calculations

The electrostatic potentials were calculated using the UHBD package.[15] Prior to electrostatic calculations, polar hydrogen atoms were added to the models using the XPLOR program,[16] and the models were minimized, keeping the positions of the non-hydrogen atoms fixed. The OPLS non-bonded parameter set[17] was used to assign partial atomic charges and radii. The relative dielectric constant of the solvent was set at 78.5. The relative dielectric constant of the solute was set to 2.0. When the ionic strength of the solvent was set to 150 mM, the solute was surrounded by a 2-Å-thick Stern layer. The temperature was set to 300 K. The coordinates of the proteins were superimposed based on the sequence alignment used for homology modeling, and the electrostatic potentials of the proteins were computed by finite-difference solution of the linearized Poisson-Boltzmann equation on a 110^3 grid with spacing of 1 Å centered at the same point for all proteins.

SI Calculations

The similarity indices were calculated from equations 1 and 2 using in-house programs written for the purpose. For each pair of proteins, the similarity index was calculated using the precomputed potentials in the regions defined by probe radius σ and "skin" thickness δ.

RESULTS

Sensitivity of the Potentials and Similarity Indices to Structural Variations

Electrostatic potentials were calculated with the UHBD package[15] for 104 homology-modeled PH domain structures.[7] These models have low "resolution" owing to the sequence divergence of the PH domain family. The electrostatic potentials were calculated at two different nominal ionic strengths, 0 mM and 150 mM, to investigate the influence of ionic strength on the comparisons. The accuracy of the molecular electrostatic potentials for the models was evaluated by comparison to potentials calculated from the eight experimentally determined PH domain structures. The similarity indices of the electrostatic potentials of the models versus the electrostatic potentials of the experimental structures range from 0.71 to 0.95 for potentials at 0 mM ionic strength and from 0.64 to 0.84 for potentials at 150 mM ionic strength (Fig. 2, diamonds). We attribute the higher accuracy of the modeled potentials at the low ionic strength to the different relative contributions of the potential terms at the two ionic strengths. Higher-order multipole terms can be expected to contribute more to the potential at 150 mM ionic strength and therefore to be more sensitive to conformation than the mono- and dipole terms.

The influence of structural variability on similarity indices and the significance of the similarities between the models and the experimental structures was assessed by comparison of electrostatic self-similarity within the structure ensembles from PH domain structures solved by nuclear magnetic resonance (NMR) spectroscopy,[18–23] and by comparison to an 800-ps molecular dynamics (MD) simulation of the solvated PH domain from mouse β-spectrin.[14] The distribution of the self-similarities of the electrostatic potentials at 0 mM ionic strength of the 20 lowest-energy structures from the NMR ensemble of the PH domains peaks around 0.9, but there is a considerable spread of the similarity indices. For the potentials at 150 mM ionic strength, the self-similarity of the structures in the NMR ensembles is lower, confirming the observation that the potentials calculated with higher ionic strength are more sensitive to the molecular conformation. The distribution of self-similarities in the MD trajectory is similar to the data from the NMR structure ensemble (Fig. 1).

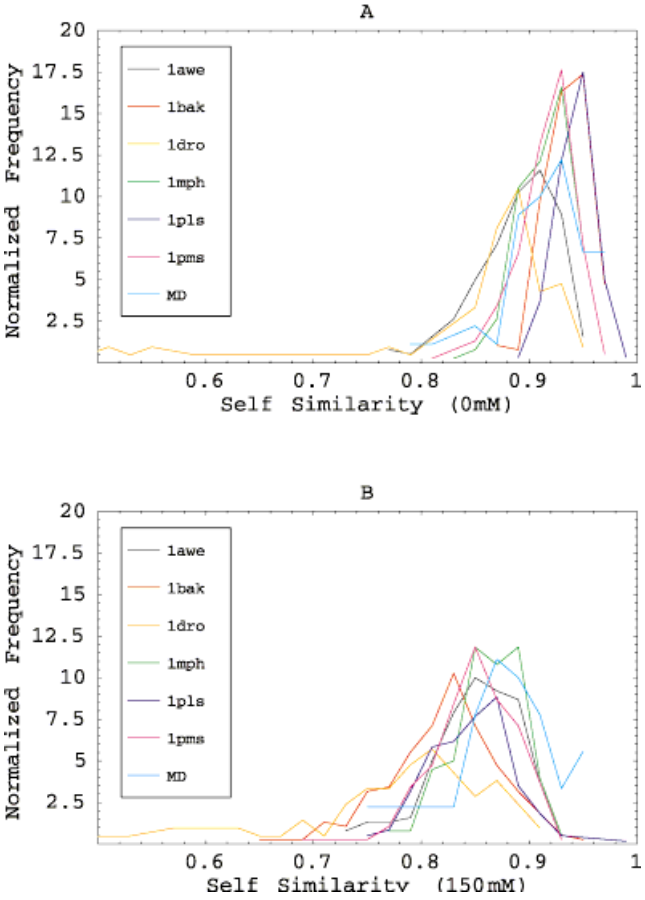

Fig. 1. Distribution of self-similarity indices of the electrostatic potentials for an ensemble of NMR structures and an MD trajectory. The distributions were calculated from the all-versus-all comparison matrix within each ensemble. **A:** SI distribution of potentials calculated at 0 mM ionic strength. **B:** SI distribution of potentials calculated at 150 mM ionic strength.

Reproduction of Electrostatic Potential Similarities By Homology Models

For our purpose, it is more important to assess how well the potentials from the homology models can reproduce the potential similarity as opposed to simply reproducing the electrostatic potentials. For this, the pairwise similarity indices from the experimentally determined PH do-

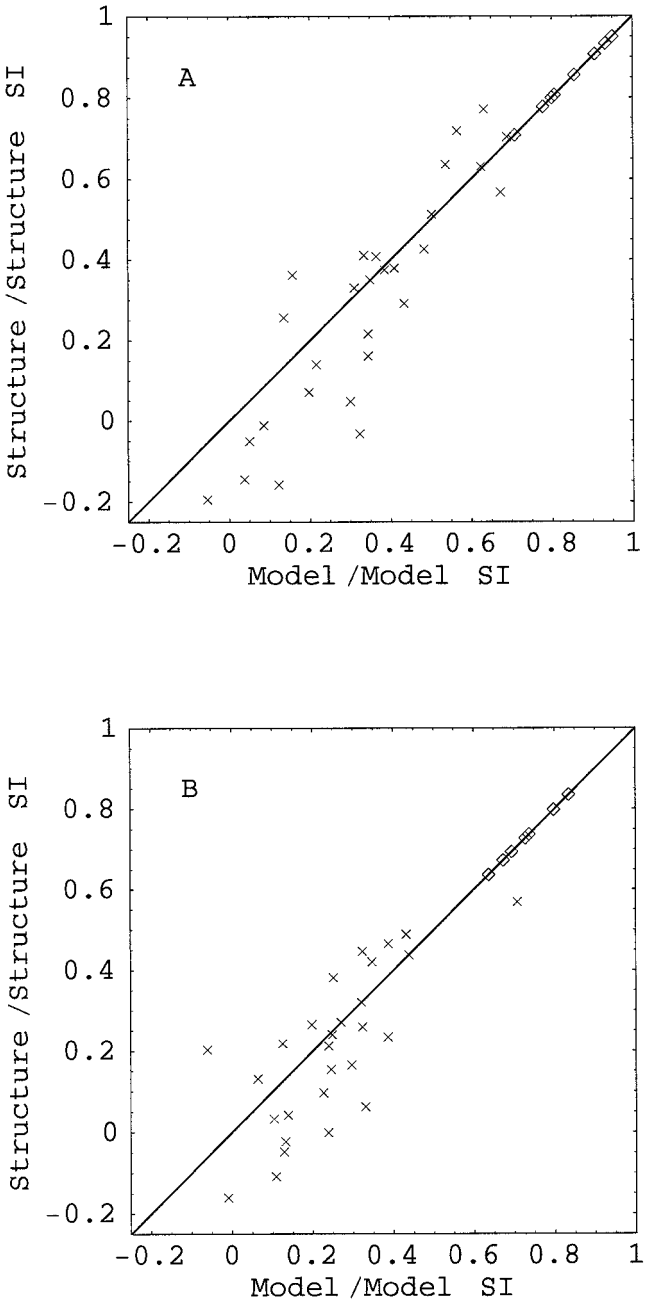

Fig. 2. Evaluation of the accuracy of the electrostatic potentials and the similarity indices. Diamonds represent the SI of a model to the experimental structure of the same PH domain. Crosses are SIs between two experimental PH domain structures versus SIs between the corresponding models. **A:** Potentials calculated at 0 mM ionic strength. **B:** Potentials calculated at 150 mM ionic strength.

main structures were compared to those for the models (Fig. 2). For the potentials at 0 mM ionic strength, the correlation between the similarity indices from the models and those from the experimental structures is good ($r^2 = 0.88$, slope 0.95), whereas the reproduction of similarity indices from the potentials at 150 mM ionic strength is slightly worse ($r^2 = 0.80$, slope 0.85). However, both these correlations are significant ($P < 0.01$, data not shown). Interestingly, for the zero ionic strength potentials, the fit of the model–model similarity indices to the experimental structure similarity indices is better for pairs having high similarity, whereas the spread is larger for the more dissimilar potentials (Fig. 2).

The similarity indices calculated for all models (Fig. 3) from the potentials at the two different ionic strengths also show differences (Fig. 3A). Overall, the similarity indices calculated for the higher ionic strength have a narrower range compared to the zero ionic strength potentials. The similarity indices calculated from the 0 mM potentials range from −0.66 to 0.75, whereas, for the 150 mM potentials, the similarity indices range from −0.45 to 0.72.

A comparison of electrostatic similarity to the sequence identity for all pairs of PH domain sequences (Fig. 4) shows that although the PH domains are in general very divergent in sequence, this is not correlated with electrostatic similarity. There is strong conservation of electrostatic potential despite the very low sequence conservation (15% average pairwise identity) in the PH domain family.

Analysis of the Electrostatic Similarity

The matrix of electrostatic similarities can be conveniently analyzed using principal component analysis (PCA). PCA is a standard technique to analyze correlations between data.[24,25] The aim is to represent the original high-dimensional problem as well as possible by a few variables. PCA is usually applied to the correlation matrix, but can also be used with a similarity matrix such as the matrix of pairwise Hodgkin similarity indices. The result of PCA is an eigenvector expansion of the original matrix, where the eigenvectors with the largest eigenvalues explain most of the variability in the original matrix. The decay of the eigenvalues is then a measure of the quality of the lower-dimensional representation. The application of PCA to the 104×104 Hodgkin similarity matrix allows the representation of the electrostatic potentials in a few dimensions. Potentials with a high degree of similarity will have similar locations along the principal components, and sequences can thus be classified according to their modeled electrostatic potential.

A projection into a subspace defined by the first two eigenvectors is a convenient way of visualizing the electrostatic similarity of the protein family members (Figs. 5 and 6; Table I). For all three potential models (analytical, 0 mM, and 150 mM ionic strength), the eigenvalues decay rapidly (Fig. 5D), and consequently the full matrix can be approximated by the first few components. The most important features are conserved in the projections of molecular electrostatic potentials for all PH domain sequences for all three potential descriptions used in this study. The majority of the domains are located in one cluster (Fig. 5A–C), indicating that the electrostatic potential is generally conserved among the PH domains. When the first two eigenvectors are considered, there is a particularly good agreement between the analytical description of the potential and the potential computed from the Poisson-Boltzmann equation at 0 mM ionic strength, whereas the

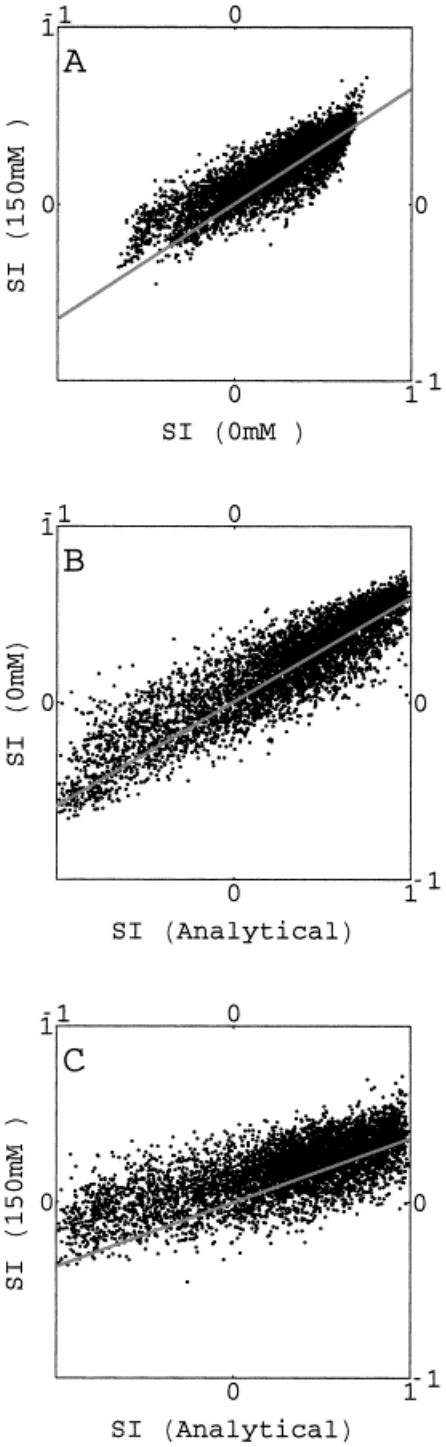

Fig. 3. Comparison of similarity indices for different descriptions of the electrostatic potentials (modeled structures only). **A:** Poisson-Boltzmann (PB) potential at 0 mM ionic strength versus PB potential at 150 mM. **B:** Analytical potential versus PB potential at 0 mM ionic strength. **C:** Analytical potential versus PB potential at 150 mM ionic strength.

separation along the second eigenvector differs for the potential at 150 mM. However, there is only a very slow decay of the eigenvalues after the first, showing that the

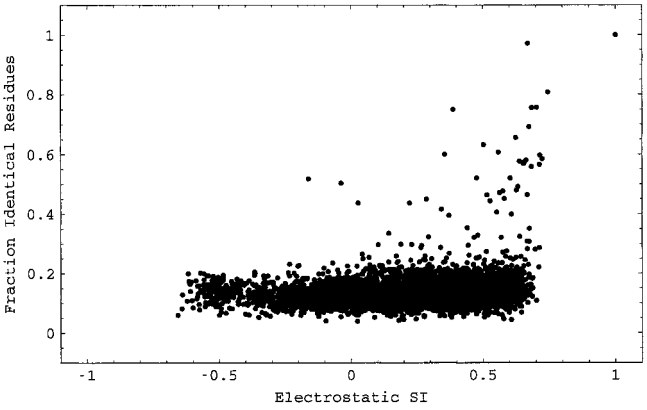

Fig. 4. Comparison for all pairs of PH domain models of the electrostatic SI from the electrostatic potentials at 0 mM with the fraction of identical residues in each pair of sequences.

order of these components is not clearly determined. The separation in the second eigenvector, and differences therein are therefore less significant.

Whereas the majority of PH domains have a conserved positive electrostatic potential, roughly 15% of the PH domains display a different overall pattern. In our previous study, we discerned two groups of PH domains with different potentials, a set of Dbl homology (DH) domain–linked PH domains and PH domains from proteins with internal PH domain repeats (Fig. 6 and Fig. 5A–C, red numbers: N-terminal PH domain, blue numbers: C-terminal domain, green: DH domain linked). There are six DH–PH domains found outside the main cluster in the analysis of all three potential models—models 16, 40, 53, 62, 83, and 91 (UNC89, LBS, RSC453, ECT2, SCD1, and CDC24)—and the magnitude of these differences is easily seen from a display of the potential (Fig. 6). For the internal repeat PH domains, the more N-terminally located PH domains (Fig. 5A–C, red numbers) tend to localize in the main cluster, whereas there is a larger number of C-terminally located PH domains among the outliers. A closer inspection of the general pattern reveals that the N-terminal domains outside the main cluster in projections of all potentials (red, models 70 and 85, rabbit and *Torpedo C.* synthrophins) come from proteins whose C-terminal PH domain is also found among the outliers (Fig. 5A–C and Fig. 6, models 71 and 86). Other PH domains consistently located outside the main cluster (models 6, 11, and 95: unidentified genomic sequences; models 23, 46, 98 *Drosophila* GPK-1, human PLC-γ_2, and yeast STE5) are difficult to link to any function, but all have a clearly non-standard potential.

DISCUSSION

We have compared the electrostatic properties of protein models using several different representations of the electrostatics. Similarity indices of electrostatic potentials can be used to compare a large number of structures and to infer functional properties. Our results indicate that a simplified analytical expression may be sufficient for the

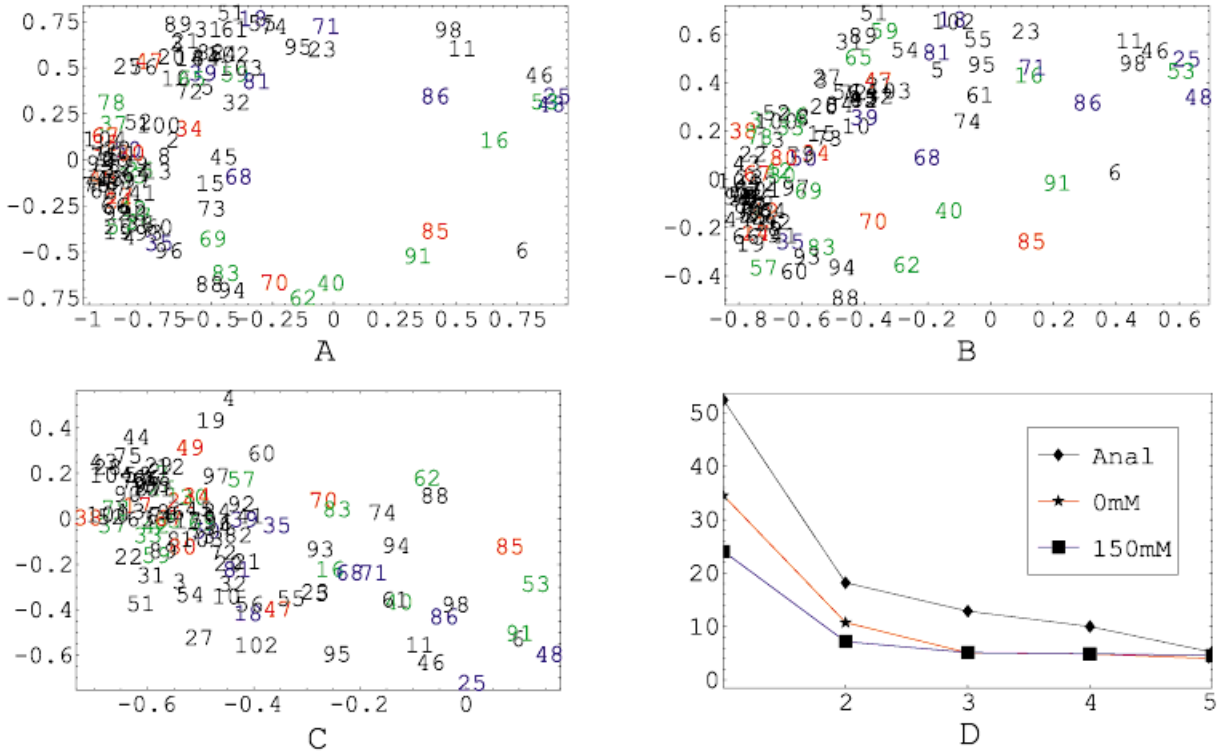

Figure 5.

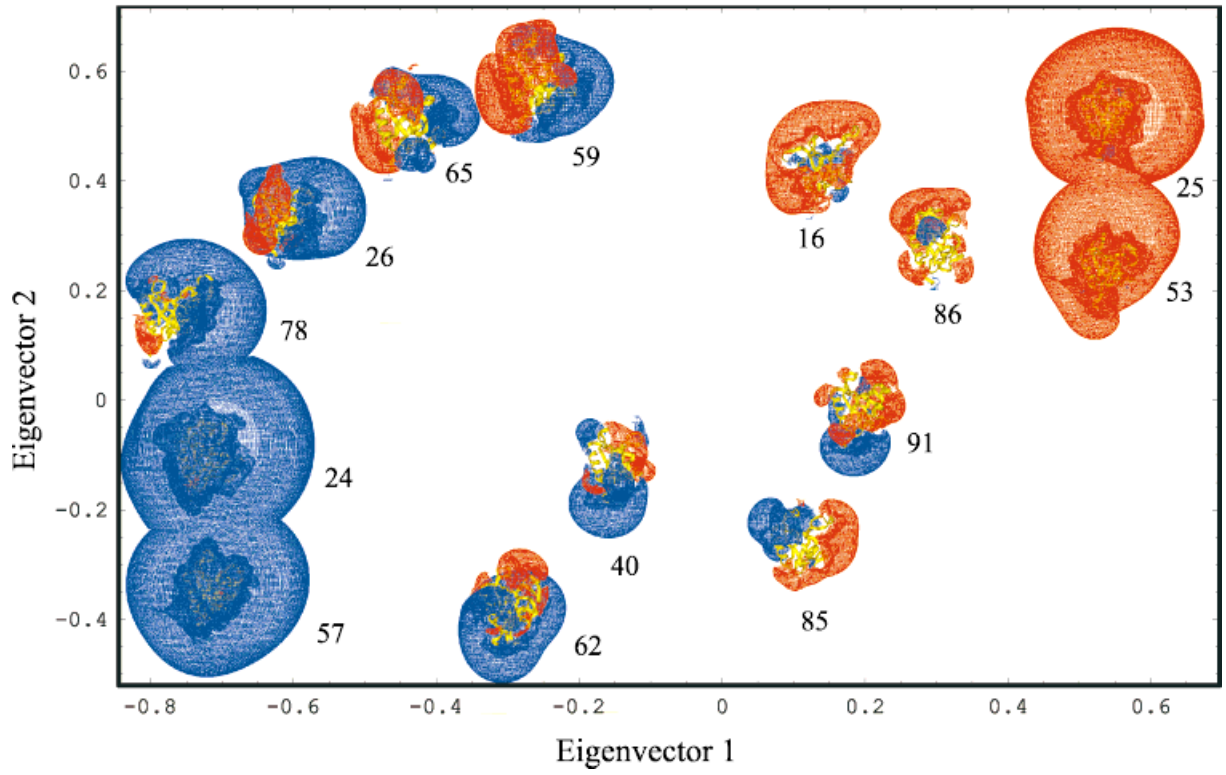

Figure 6.

calculation of the electrostatic potential: a simple monopole–dipole based model reproduces potential similarities calculated from the linearized Poisson-Boltzmann equation at 0 mM ionic strength. The correlation of the similarities is worse for potentials calculated at 150 mM ionic strength, showing that at higher ionic strength, the higher-order multipole terms become more prominent. This can also be appreciated from the decay of the eigenvalues for the more complex molecular electrostatic potentials at higher ionic strength (Fig. 5D).

The comparison of the electrostatic variability within NMR structure ensembles and during molecular dynamics simulations indicate that the electrostatics of proteins are modulated by protein motion. The errors in the homology models cause deviations in the molecular electrostatic potential comparable to the changes in potentials due to protein motions, making low resolution homology modeling a viable route to large-scale studies of electrostatic properties.

Similarity indices calculated with the analytical model reproduce the ones calculated from the Poisson-Boltzmann potentials at 0 mM well (Fig. 3, Table II), whereas the simple description is not sufficient to model potentials at high ionic strength. The range of similarity indices from the analytical potentials is wider than for the other potentials. This can be understood by considering electrostatic screening effects which are expected to increase the contribution of more complicated multipole terms of the electrostatic potential at higher ionic strengths. The similarity indices calculated from the dipole moment alone show a wider spread than the similarity indices from the other potentials (Table II), and, in contrast to the analytical model, the magnitude of the similarity indices is overestimated compared to the Poisson-Boltzmann electrostatic potentials. This also shows that although the general trend is the same using the dipole moments as a simplified description, one can achieve a much better agreement with the Poisson-Boltzmann potentials for proteins with an overall charge by also including the monopole term in the potential expression.

An objective of this study was to compare the observations from our previous study based on the direction of the dipole moments with a quantitative method: principal component analysis of the similarity index matrix to display the similarity between the full potentials. The two main results from our previous study[7] were a striking electrostatic complementarity of PH domains from proteins carrying internal PH domain repeats and the presence of a group of DH domain–linked PH domains having electrostatic properties incompatible with phospholipid binding. The conclusions from our previous analysis hold in this quantitative study. Thus, we find a complementarity of the tandem PH domains, e.g., the two domains from the two PLC-γ isoforms; models 24, 25, and 47, 48 (Figs. 5A–C, 6). The DH-linked PH domain outliers (Figs. 6 and 5A–C, green) do not form a cluster in the manner seen in our previous study. However, all of these domains are also now found among the outliers and are clearly seen to have different potentials (Fig. 5), supporting our previous observation of two subgroups among the DH-linked PH domains.

The main incentive for the present study is the rapid growth of the biological sequence and structure databases. With the advent of what has been coined "structural genomics," large-scale efforts to determine the structure of all proteins within an organism,[26] we can soon hope to have structural homologues for the majority of globular proteins. Many protein sequences will thus be accessible for homology modeling. Modeling the structures for most sequences in the databanks with a structural homologue is already a feasible task. The SWISS-MODEL initiative[27] is an effort to routinely integrate models of proteins in the SWISS-PROT database for experimental design and analysis.

The availability of structural information on this scale will require large-scale analysis of structural features and studies of structure–function relationships for whole protein families with statistical techniques. In this study, we have demonstrated the use of similarity indices for large-scale comparisons of protein electrostatics. Similarity indices were found to be useful for objective and quantitative comparison of electrostatic potentials calculated from homology models. An attractive feature of similarity indices is that the resulting quantity is a scalar, thus simplifying the analysis and allowing a straightforward application of multivariate statistical technique.

The analytical approximation is particularly appropriate for very large datasets or for datasets for which it can be expected that the monopole and dipole terms will dominate the molecular electrostatic potential. However, electrostatic potentials calculated from the Poisson-Boltzmann equation offer a much more accurate description of the molecular electrostatic potentials. The computational requirements of such an analysis depend on the size of the grids used for the finite difference solution of the Poisson-Boltzmann equations. The grid size determines the time required not only to compute the potentials but also to compute the similarity matrix. For example, on a SGI Origin 2000 (250 MHz R10000, IRIX 6.5), comparison of 110^3 grids requires about 10 minutes per 100 SIs, and a total of 16 hours for the 104×104 PH domain similarity matrix; however, by using 65^3 grids, the computational time is reduced to 4 minutes per 100 SIs. In conclusion, for small to medium-sized datasets (up to around 100 proteins), the comparison of Poisson-Boltzmann potentials offers greater accuracy and flexibility (variation of ionic

Fig. 5. The first two eigenvectors of the electrostatic similarity matrix (abscissa is the first principal component). The key to the numbering is found in Table I. The colors symbolize the following: green, DH-linked PH domains; red, N-terminal PH domain from proteins with PH domain repeats; blue, C-terminal PH domains from same proteins. N- and C-terminal PH domains from the same protein have consecutive numbers (e.g., 85 and 86 come from syntrophin). **A:** SI from analytical potential. **B:** SI from 0 mM potential. **C:** SI from 150 mM potential. **D:** Decay of eigenvalues from the principle coordinate analysis.

Fig. 6. Display of the electrostatic potential calculated at 0 mM ionic strength for DH-linked PH domains and internal repeat PH domains on the first two principal components of the similarity matrix (cf. Fig. 5B). The key to the numbering is in Table I.

TABLE I. Key to Figure Numbering[†]

1	AN_APSA	2	B_PLC-D2
3	CE10H12	4	CE21D1
5	CEC04D8.1	6	CEC07B5.4
7	CEC35B8.2	8	CEC38D4.5
9	CEF10E9.6	10	CEK06H7.4
11	CEK10B2.5	12	CEZK1248.10
13	CEZK632.12	14	CE_DYN1
15	CE_UNC104	16	CE_UNC89
17	C_AFAP_1	18	C_AFAP_2
19	DD_CRAC	20	DM_AKT
21	DM_DYNAMIN	22	DM_GAP
23	DM_GPRK-1	24	DM_PLCGD_1
25	DM_PLCGD_2	26	DM_SOS
27	DM_SPCB	28	EM_A09787
29	HSORFV	30	H_ABR
31	H_AKT2	32	H_B-ARK-1
33	H_BCR	34	H_BSYN2_1
35	H_BSYN2_2	36	H_BTK
37	H_DBL	38	H_FGD1_1
39	H_FGD1_2	40	H_LBC
41	H_MIG2	42	H_NEP1
43	H_ORFA	44	H_OSBP
45	H_PKC_MU	46	H_PLCG2_2
47	H_PLCG_1	48	H_PLCG_2
49	H_PLEC_1	50	H_PLEC_2
51	H_RACA	52	H_RASA-GAP
53	H_RSC453	54	H_SEC7
55	H_SPCB	56	H_SPCBR
57	H_TIM	58	H_TSK
59	MM_SOS-2	60	M_3BP2
61	M_BSPCB	62	M_ECT2
63	M_GRB10	64	M_GRB7
65	M_SOS-1	66	M_TEC
67	M_TIAM1_1	8	M_TIAM1_2
69	M_VAV	70	OC_59DAP_1
71	OC_59DAP_2	72	R_B-ARK-2
73	R_CAPS	74	R_DYNAMIN
75	R_GAP1M	76	R_IRS-1
77	R_LL5	78	R_OST
79	R_PLC-III	80	R_RASGRF_1
81	R_RASGRF_2	82	R_TES-DYN
83	SP_SCD1	84	TB_NRKA
85	TC_SYN_1	86	TC_SYN_2
87	Y_BEB1	88	Y_BEM2
89	Y_BEM3	90	Y_BOB1
91	Y_CDC24	92	Y_CLA4
93	Y_L9470.23	94	Y_L9470.4
95	Y_L9576.5	96	Y_NUM1
97	Y_SIP3	98	Y_STE5
99	Y_SWH1	100	Y_YBR1004
101	Y_YHR073W	102	Y_YHR131C
103	Y_YHR155W	104	Y_YIL105C

[†]Key to the protein numbering used in Figures 3 and 4. The model codes are those used in reference, 7 and the corresponding coordinate files and alignments can be found at http://www.nmr.embl-heidelberg.de/blomberg/PHdomains together with evaluation data for the models.

strength conditions, charge models) but is currently unsuitable for large datasets. In contrast, the analytical expression is well suited for rapid comparison of even very large datasets, due to the small computational requirements.

TABLE II. Linear Regression of SIs From the Three Electrostatic Potentials[†]

	r^2	Slope	Standard error
Analytical/0 mM	0.85	0.59	0.0033
Analytical/150 mM	0.62	0.36	0.0038
Analytical/Dipole	0.55	0.71	0.0089
0 mM/Dipole	0.33	0.88	0.0170
150 mM/Dipole	0.41	1.35	0.0223
0 mM/150 mM	0.81	0.62	0.0042

[†]Correlation of the similarity indices from different potentials. These were calculated from comparisons of the matrix of 104×104 pairwise similarity indices for the PH domain models.

We have demonstrated the applicability of large-scale comparisons of electrostatic potentials by the use of similarity indices using the PH domain family as an example and show that similarity indices in combination with principal coordinate analysis allow rapid comparison of electrostatic properties within a large protein family. Combining the information available in sequence databases directly with predicted structural features provides an important tool for the analysis and prediction of biological function. As we have previously shown,[7] the use of structural data can significantly enhance the information available from sequence alignments and allow predictions on function and ligand binding.

Supplementary Material

Scripts for performing the comparisons and for calculating the analytical SIs can be found on http://www.embl-heidelberg.de/ExternalInfo/wade/pub/soft/pipsa.html

ACKNOWLEDGMENTS

Roger Abseher is gratefully acknowledged for providing the molecular dynamics trajectory and Michael Habeck for stimulating discussions.

REFERENCES

1. Honig B, Nicholls A. Classical electrostatics in biology and chemistry. Science 1995;268:1144–1149.
2. Botti S, Felder C, Sussman J, Silman I. Electrotactins: a class of adhesion proteins with conserved electrostatic and structural motifs. Protein Eng 1998;11:415–420.
3. Ullmann GM, Hauswald M, Jensen A, Kostic N, Knapp E-W. Comparison of the physiologically equivalent proteins cytochrome c_6 and plastocyanin on the basis of their electrostatic potentials. Tryptophan 63 in cytochrome c_6 may be isofunctional with tyrosine 83 in plastocyanin. Biochemistry 1997;36:16187–16196.
4. Wade RC, Gabdoulline RR, Luty BA. Species dependence of enzyme-substrate encounter rates for triose phosphate isomerases. Proteins 1998;31:406–416.
5. Wade RC, Gabdoulline RR, Lüdemann SK, Lounnas V. Electrostatic steering and ionic tethering in enzyme-ligand binding: insights from simulations. Proc Natl Acad Sci USA 1998;95:5942–5949.
6. Demchuk E, Müller T, Oschkinat H, Sebald W, Wade RC. Receptor binding properties of four-helix-bundle growth factors deduced from electrostatic analysis. Protein Sci 1994;20:203–215.
7. Blomberg N, Nilges M. Functional diversity of PH domains: an exhaustive modelling study. Folding Design 1997;2(6):343–355.
8. Hodgkin EE, Richards WG. Molecular similarity based on electrostatic potential and electric field. Int J Quant Chem Quant Biol Symp 1987;14:105–110.

9. Good AC. 3D Molecular similarity indices and their application in QSAR studies. Molecular Similarity in drug design. Dean PM, editor. London: Blackie Academics, 1995. p 24-53.
10. Carbo R, Domingo L. LCAO-MO similarity measures and taxonomy. Int J Quant Chem 1987;17:517–545.
11. Burt C, Richards WG. The application of molecular similarity calculations. J Comput Chem 1990;11:1139–11.
12. Richard AM. Quantitative comparison of molecular electrostatic potentials for structure-activity studies. J Comp Chem 1991;12: 959–969.
13. Tomic S, Gabdoulline RR, Kojic-Prodic B, Wade RC. Classification of auxin plant hormones by interaction property similarity indices. J Comp Aid Mol Design 1998;12:1–17.
14. Abseher R, Horstink L, Hilbers CW, Nilges M. Essential spaces defined by NMR structure ensembles and molecular dynamics simulation show significant overlap. Proteins 1998;31:370–382.
15. Madura JD, Briggs JM, Wade RC, Davis ME, Luty BA, Ilin A, Antosiewicz J, Gilson MK, Bagheri B, Scott LR, McCammon JA. Electrostatics and diffusion of molecules in solution: simulations with the University of Houston Browinian Dynamics program. Comp Phys Comm 1995;91:57–95.
16. Brünger AT. X-PLOR. A system for X-ray crystallography and NMR. New Haven: Yale University Press; 1992. 382 p.
17. Jorgensen WL, Tirado-Rives J. The OPLS potential function for proteins. Energy minimization for crystals of cyclic peptides and crambin. J Am Chem Soc 1988;110:1657–1666.
18. Macias MJ, Musacchio A, Ponstingl H, Nilges M, Saraste M, Oschkinat H. Structure of the pleckstrin homology domain from beta-spectrin. Nature 1994;369:675–677.
19. Yoon HS, Hajduk PJ, Petros AM, Olejniczak ET, Meadows RP, Fesik SW. Solution structure of a pleckstrin-homology domain. Nature 1994;369:672–675.
20. Zhang P, Talluri S, Deng H, Branton D, Wagner G. Solution structure of the pleckstrin homology domain of drosophila β-spectrin. Structure 1995;3(2):1185–1195.
21. Fushman D, Najmabadi-Haske T, Cahill S, Zheng J, LeVine III H, Cowburn D. The solution structure and dynamics of the pleckstrin homology domain of G-protein coupled receptor kinase 2 (β-ARK1): a binding partner of Gβγ subunit. J Biol Chem 1998;273: 2835–2843.
22. Koshiba S, Kigawa T, Kim J-H, Shirouzu M, Bowtell D, Yokoyama S. The solution structure of the pleckstrin homology domain of mouse son-of-sevenless 1 (mSos1). J Mol Biol 1997;269:579–591.
23. Zheng J, Chen R-H, Corblan-Garcia S, Cahill SM, Bar-Sagi D, Cowburn D. The solution structure of the pleckstrin homology domain of human Sos1: a possible structural role for the sequential association of dbl homology and pleckstrin homology domains. J Biol Chem 1997;272(48):30340–44.
24. Gower JC. Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 1966;53(3 and 4):325–38.
25. Mardia KV, Kent JT, Bibby JM. Multivariate analysis, 6th ed. Academic Press: London; 1997. 521 p.
26. Rost B. Marrying structure and genomics. Structure 1998;6:259–263.
27. Guex N, Peitsch MC. Swiss-model and the swiss-pdbviewer: an environment for comparative protein modeling. Electrophoresis 1997;18(15):2714–2723.

APPENDIX A
Extension of the Analytical Expression to Comparison of the Electrostatic Potentials in a Defined Region, e.g., the Active Site

In the theory section of the report, we introduced an analytical expression to compute the similarity of electrostatic potentials based on their monopole and dipole terms. This expression measures the similarity over the whole "skin" of both molecules. However, this similarity may not be a good measure of similarity of the proteins in terms of their function. Usually, only a part of the protein is responsible for its function; thus, only a part of its electrostatic potential may represent the functional properties of the protein. For example, only the similarity at the active site of two enzymes would indicate their similar enzymatic function; similarity at other sites is not required. Therefore, the measure of global similarity of the electrostatic potentials (6) may not be sufficient to establish similar functionality of proteins.

However, the similarity index (6) can be redefined so that it measures the similarity of the potentials in some specified region of the three-dimensional space. Let this be the angular region around some vector e with angular extent θ_0, i.e., the angle θ between the coordinate vector r and e is required to be less than θ_0. Here the coordinate system is centered on the superimposed center of the proteins. Then integration of the left-hand side of formula 4 in the theory section gives

$$\int_R^{R+\delta}\int_0^{\theta_0}\int_0^{2\pi} \phi_1(r)\phi_2(r)r^2\,dr\,\sin\theta\,d\theta\,d\psi$$
$$= 4\pi q_1 q_2 \sin^2(\theta_0/2)\int_R^{R+\delta} dr +$$
$$4\pi\left(\sin^2(\theta_0/2) - \frac{1}{4}\cos\theta_0\sin^2\theta_0\right)(\mathbf{d}_1, \mathbf{d}_2)\int_R^{R+\delta}\frac{1}{3r^2}\,dr$$
$$+ 4\pi\cos\theta_0\sin^2\theta_0(\mathbf{d}_1, \mathbf{e})(\mathbf{d}_2, \mathbf{e})\int_R^{R+\delta}\frac{1}{4r^2}\,dr \quad (7)$$

where dr denotes integration over the radial variable, and $(\sin\theta\,d\theta\,d\psi)$ denotes integration over the region $(r, e) < |r||e|\cos\theta_0$.

Then formula 6 is replaced by

$$SI_{12} = \frac{2(q_1q_2f_1 + \alpha(\mathbf{d}_1, \mathbf{d}_2)f_2 + \alpha(\mathbf{d}_1, \mathbf{e})(\mathbf{d}_2, \mathbf{e})f_3)}{(q_1^2 + q_2^2)f_1 + \alpha(|\mathbf{d}_1|^2 + |\mathbf{d}_2|^2)f_2 + \alpha((\mathbf{d}_1, \mathbf{e})^2 + (\mathbf{d}_2, \mathbf{e})^2)f_3} \quad (8)$$

where the first two factors, f_1 and f_2, are functions of θ_0, changing from 0 to 1 as θ_0 changes from 0 to π, and f_3 is an oscillatory function:

$$f_1 = \sin^2(\theta_0/2) \quad (9)$$

$$f_2 = \sin^2(\theta_0/2) - \tfrac{1}{4}\cos\theta_0\sin^2\theta_0 \quad (10)$$

$$f_3 = \tfrac{3}{4}\cos\theta_0\sin^2\theta_0 \quad (11)$$

If the layer in which the potentials are compared is thin and has a radius of R, the parameter $\alpha = 1/3R^2$. The previously derived value of α of $(17\text{Å})^{-2} = \frac{1}{3}(9.815\text{Å})^{-2}$ may be used.

Use of this new measure is as computationally efficient as use of formula 6. The value of θ_0 gives the angular extent of the functional region of the proteins to be compared (as seen from the protein center). Appropriate values range from 10° (e.g., enzymes with small active sites) to 90° (for example, membrane-binding proteins). The vector e points to the location of the functional site.