PROTEINS: Structure, Function, and Genetics 51:360–377 (2003)

An Empirical Model for Electrostatic Interactions in Proteins Incorporating Multiple Geometry-Dependent Dielectric Constants

Michael S. Wisz and Homme W. Hellinga*
Department of Biochemistry, Box 3711, Duke University, Durham, North Carolina 27710

ABSTRACT **Here we introduce an electrostatic model that treats the complexity of electrostatic interactions in a heterogeneous protein environment by using multiple parameters that take into account variations in protein geometry, local structure, and the type of interacting residues. The optimal values for these parameters were obtained by fitting the model to a large dataset of 260 experimentally determined pK_a values distributed over 41 proteins. We obtain fits between the calculated and observed values that are significantly better than the null model. The model performs well on the groups that exhibit large pK_a shifts from solution values in response to the protein environment and compares favorably with other, successful continuum models. The empirically determined values of the parameters correlate well with experimentally observed contributions of hydrogen bonds and ion pairs as well as theoretically predicted magnitudes of charge-charge and charge-polar interactions. The magnitudes of the dielectric constants assigned to different regions of the protein rank according to the strength of the relaxation effects expected for the core, boundary, and surface. The electrostatic interactions in this model are pairwise decomposable and can be calculated rapidly. This model is therefore well suited for the large computations required for simulating protein properties and especially for prediction of mutations for protein design. Proteins 2003;51:360–377.**

Key words: protein design; pK_a values; prediction; environment-dependent; self-energy; charge

INTRODUCTION

Electrostatic interactions in proteins are ubiquitous, affecting protein structure, protein stability, ligand binding, protein-protein interactions, electron transfer, and enzyme catalysis.[1–5] Computational modeling of electrostatic interactions is therefore important for simulating protein properties from first principles,[6,7] and for protein design.[8] At the most basic level, Coulomb's Law for a pair of charges describes the energy of all electrostatic interactions. For this description to be applicable in a protein, an all-atom microscopic model of both the protein and solvent is required.[9,10] Frequently it is not possible to include all these interactions in such detail, however, and models need to be developed that represent several different classes of effects as an average or continuum. In the solvent these effects include the reorientation of water molecules in response to a charge distribution (dipolar solvent relaxation[1,11]), redistribution of salt ions near the surface of the protein (Debye-Hückel screening effects[12]), and the pH of the solution (ionization effects[13]). In the interior of the protein these include dipolar relaxation through structural changes,[14–16] modulation of the charges themselves by induced dipoles,[10,17] and by hydrogen bonds[18,19] or ion pairs.[1,20]

Continuum models typically group electrostatic interactions into two broad categories that are treated separately[6,13,16,21–36]: (a) interactions between the protein interior and the surrounding solvent and (b) interactions within the protein interior. In the first category, solvent dipolar relaxation and ionic screening effects are usually distinguished. In the second category, most models use a uniform dielectric constant to account for all relaxation effects in the protein interior. It has been pointed out that in a highly heterogeneous, anisotropic environment, uniform dielectric constants are an oversimplification.[20] Here we model solvent and interior effects using multiple dielectric constants that are assigned to individual charges and pairwise interactions. The interactions with the solvent are considered individually for each charge and are a function of the location of the charge within the protein relative to the solvent as well as its local environment (static solvent accessibility). Dielectric constants are assigned separately for each pairwise interaction and are determined by location of the two charges relative to the solvent, by the local environment (hydrogen bonding, ion pairs, and solvent accessibility), by the type of interaction between relevant amino acid side chains (charge-charge,

Grant sponsor: National Insititutes of Health and National Science Foundation (to H.W.H.).

M. S. Wisz's present address is Departments of Microbiology & Immunology and Biomedical Engineering, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599.

*Correspondence to: Homme W. Hellinga, Department of Biochemistry, Box 3711, Duke University, Durham, North Carolina 27710. E-mail: hwh@biochem.duke.edu.

Received 20 July 2002; Accepted 25 October 2002

charge-polar), and by the distance between the interacting charges.

The behavior of ionizable groups provides a direct link between effects that can be modeled theoretically and observed experimentally[21–28,37] and can be used both to parameterize and evaluate models. Electrostatic interactions in proteins perturb the intrinsic pK_a values of ionizable groups relative to the values determined in solution. pK_a values of ionizable groups can be accurately determined by titration using NMR. Theoretical titration curves can be calculated by constructing a pH-dependent, Boltzmann-distributed ensemble of all the combinations of charged and neutral forms of the ionizable groups in the protein, using an electrostatic potential function to determine the internal energy of each ensemble member.[38,39] In the study presented here we fit our model to an extensive collection of 260 experimental pK_a values distributed over 41 proteins of known structure.

Models are evaluated by calculating the root-mean square difference (RMSD) between calculated and observed pK_a values. Caution has to be exercised, however, because the RMSD value is dominated by surface groups, the pK_a values of which are similar to those observed in water. The pK_a values of ionizable groups that exhibit large shifts are arguably the most critical to predict correctly, because these groups are often found in functionally important regions of the protein and reflect strong influences by the dielectric nature of the surrounding protein matrix.[37] It is therefore important to classify the observed pK_a values into "small" ($\Delta pK_a < 1$) and "large" ($\Delta pK_a \geq 1$) shifts and to evaluate the model with respect to the latter.[20] Most continuum models still fail to accurately predict large pK_a shifts in proteins. Both by this criterion and overall RMSD value, the model presented here performs well.

METHODS

Collection of pK_a Data

Experimental pK_a data, temperature, and ionic strength conditions were taken without modification from literature (Table I). Only pK_a values with definite values were used, excluding cases where upper or lower bound estimates were reported as well as the special case of reduced human (1ert) and *E. coli* (3trx) thioredoxin, where the close proximity of three interacting negatively charged groups (Cys32, Cys35, Asp26) prevented the individual pK_a values of these side chains from being accurately determined.[40] The temperature and ionic strength used in all calculations were those used experimentally for each protein.

Preprocessing of Protein Structures

The protein structures (Table I) were obtained from the RCSB protein data bank (http://www.rcsb.org/pdb). For NMR data with multiple solution structures, the best reported representative conformer was used. For multiple side-chain conformations, the "A" conformer was selected. The biologically relevant number of subunits was used in all calculations. Hetero-atoms were used if they were present in the experimental pK_a determination, with the exception of water molecules, in which case only the buried water molecules (220, 223 in the Pdb file) of staphylococcal nuclease V66E (phsv66e) were included, treated as polar groups with hydrogen-bonding capability. Protein structures were protonated using the Reduce program,[41] leaving all acidic groups unprotonated, and all basic groups protonated, and histidines either singly or doubly protonated.

Treatment of Charged Groups

The ionizable charged groups and their associated water pK_a values (p$K_a^{solution}$) considered in the electrostatics calculations are Asp (p$K_a^{solution}$ = 4.0), Glu (p$K_a^{solution}$ = 4.4), Tyr (p$K_a^{solution}$ = 9.6), Lys (p$K_a^{solution}$ = 10.4), Arg (p$K_a^{solution}$ = 12.0), reduced Cys (p$K_a^{solution}$ = 8.3), N-terminus (p$K_a^{solution}$ = 7.5), C-terminus (p$K_a^{solution}$ = 3.8), and heme propionate (p$K_a^{solution}$ = 4.4). For all of the amino acids except histidine, these values were taken from data on the titration behavior of denatured proteins.[42] Histidine residues were treated as two independent ionizable groups with two water pK_a values (p$K_a^{solution}$ = 6.5 for the N_δ tautomer, p$K_a^{solution}$ = 6.9 for the N_ϵ tautomer).[43] The nontitrating, partially charged groups include all of the neutral amino acids, histidines involved in heme binding, oxidized cysteines, water, heme, acetylated N-terminus, and amidated C-terminus. Partial charges for the nontitrating amino acids were taken from the PARSE parameter set,[44] and the partial charges for the nontitrating, nonamino acid groups were obtained from CHARMM22.[45] The partial charges of titratable groups in their neutral state were adapted from the PARSE parameter set. For acidic groups with multiple protonation sites (Asp, Glu, C-terminus), the partial charges for the neutral form of these side chains are derived by distributing a unit charge equally over the two carboxylate oxygens, while leaving the group unprotonated in the calculations. For basic groups with more than one possible dissociating proton (Lys, Arg, N-terminus), the partial charges for the neutral forms of these side chains are derived by distributing the subtraction of a unit charge equally over the three hydrogens on each of these groups, while leaving the site fully protonated in the calculations. The charges of all ionizable groups except histidine are localized to one or two atoms of the side chain, obtained by subtracting or adding one unit of charge from the neutral form of an acidic or basic residue, respectively. This charge adjustment is localized to a single formal charge-bearing atom for Lys (Nζ), Tyr (OH), and Cys (Sγ) and is distributed between two atoms for Asp ($O_{\delta 1}$, $O_{\delta 2}$), Glu ($O_{\epsilon 1}$, $O_{\epsilon 2}$), Arg (NH1, NH2), and heme propionate (O1A, O2A, or O1D, O2D). For histidine residues, the tautomerism of the side chain is taken into account by treating the residue as two independent charged groups (charge-bearing atoms N_δ or N_ϵ), while making the appropriate corrections in the energy equations so that the negative ion does not appear and other residues interact with a single histidine only.[22] The ΔG_{solv} term is calculated by considering each ionizable group as a single ± 1 charge of radius R_{Born}.

TABLE I. pK_a Values and Protein Structures

Protein	No. of measured pK_a values	PDB code	Charged group[a]	Region[b]	pK_{exp}	pK_{calc}	$\lvert pK_{exp} - pK_{calc} \rvert$
Hen egg-white lysozyme[69,70]	21	2lzt	N-term1	S	7.90	7.34	0.56
			Lys1	S	10.80	10.36	0.44
			Glu7	S	2.85	3.30	0.45
			Lys13	S	10.50	10.57	0.07
			His15	B	5.40	6.14	0.74
			Asp18	S	2.70	3.83	1.13
			Tyr20	B	10.30	10.18	0.12
			Tyr23	B	9.80	10.08	0.28
			Lys33	S	10.60	10.44	0.16
			Glu35	C	6.20	5.22	0.98
			Asp48	B	1.60	3.05	1.45
			Asp52	B	3.70	4.84	1.14
			Tyr53*	C	12.10	13.76	1.66
			Asp66	C	0.90	2.13	1.23
			Asp87	S	2.10	2.89	0.79
			Lys96	S	10.80	10.63	0.17
			Lys97	S	10.30	10.82	0.52
			Asp101	B	4.10	4.23	0.13
			Lys116	S	10.40	10.45	0.05
			Asp119	S	3.20	3.75	0.55
			C-term129	S	2.80	2.97	0.17
RNAse H1[71,72]	21	2rn2	Glu6	B	4.50	4.06	0.44
			Glu32	B	3.60	3.21	0.39
			Glu48	C	4.40	4.61	0.21
			Glu57*	B	3.20	5.00	1.80
			Glu61	B	3.90	3.50	0.40
			His62	S	7.00	6.87	0.13
			Glu64	S	4.40	3.73	0.67
			His83	S	5.50	6.52	1.02
			Asp94	S	3.20	2.70	0.50
			Asp108	B	3.20	3.45	0.25
			His114	B	5.00	6.10	1.10
			Glu119	B	4.10	3.39	0.71
			His124	S	7.10	6.59	0.51
			His127	S	7.90	7.38	0.52
			Glu129	B	3.60	3.43	0.17
			Glu131	S	4.30	3.88	0.42
			Asp134	S	4.10	3.22	0.88
			Glu135	S	4.30	4.04	0.26
			Glu147	S	4.20	4.23	0.03
			Glu154	S	4.40	4.54	0.14
			C-term155	B	3.40	3.94	0.54
RNAse α-sarcin[73]	21	1de3	Asp9	B	3.90	4.30	0.40
			Glu19	S	4.60	3.58	1.02
			Glu31	S	4.60	3.61	0.99
			His35	S	6.30	6.62	0.32
			His36	B	6.80	6.60	0.20
			His50	S	7.70	6.69	1.01
			Asp57	S	4.30	3.53	0.77
			Asp59	S	4.00	4.77	0.77
			Asp75	B	3.80	3.41	0.39
			His82	C	7.30	7.49	0.19
			Asp85	S	3.80	3.54	0.26
			His92	B	6.90	6.00	0.90
			Glu96	C	5.20	4.25	0.95
			His104	B	6.50	5.90	0.60
			Asp109	S	3.70	3.56	0.14
			Glu115	S	4.80	3.71	1.09

TABLE I. (Continued)

Protein	No. of measured pK_a values	PDB code	Charged group[a]	Region[b]	pK_{exp}	pK_{calc}	$\lvert pK_{exp} - pK_{calc} \rvert$
			His137	C	5.80	4.99	0.81
			Glu140	S	4.20	3.59	0.61
			Glu144	S	4.30	3.97	0.33
			His150	S	7.50	6.19	1.31
			C-term150	S	3.50	3.65	0.15
RNAse A[21]	16	3rn3	N-term1	S	7.60	7.37	0.23
			Glu2	B	2.80	3.88	1.08
			Glu9	S	4.00	3.75	0.25
			His12	B	6.20	6.04	0.16
			Asp14	C	2.00	2.42	0.42
			Asp38	S	3.10	3.35	0.25
			His48	C	6.00	6.37	0.37
			Glu49	B	4.70	4.23	0.47
			Asp53	S	3.90	3.50	0.40
			Asp83	B	3.50	3.46	0.04
			Glu86	B	4.10	4.28	0.18
			His105	S	6.70	6.68	0.02
			Glu111	S	3.50	3.97	0.47
			His119	S	6.10	7.36	1.26
			Asp121	C	3.10	2.21	0.89
			C-term124	S	2.40	2.63	0.23
Protein G B1 domain[74]	13	1pga	Lys10	S	11.00	10.82	0.18
			Glu15	S	4.40	3.78	0.62
			Glu19	S	3.70	3.81	0.11
			Asp22	S	2.90	4.20	1.30
			Glu27	S	4.50	3.73	0.77
			Lys28	S	10.90	11.15	0.25
			Tyr33	B	11.00	10.55	0.45
			Asp36	S	3.80	3.88	0.08
			Asp40	S	4.00	3.93	0.07
			Glu42	S	4.40	4.39	0.01
			Asp46	S	3.60	4.24	0.64
			Asp47	S	3.40	3.88	0.48
			Glu56	B	4.00	3.72	0.28
Human deoxyhemoglobin[75,76]	13	4hhb	N-term1A	S	7.83	7.26	0.57
			His20A	B	6.95	6.62	0.33
			His45A	S	7.18	6.65	0.53
			His50A	S	7.20	6.62	0.58
			His72A	B	6.60	6.88	0.28
			His89A	S	7.18	6.53	0.65
			His112A*	B	7.60	9.52	1.92
			N-term1B	S	6.91	7.43	0.52
			His2B	S	6.38	6.51	0.13
			His77B	B	6.75	6.78	0.03
			His117B	B	8.20	6.68	1.52
			His143B	B	6.25	6.26	0.01
			His146B	S	8.10	6.83	1.27
Human oxyhemoglobin[75,76]	12	1hho	N-term1A	S	7.16	7.66	0.50
			His20A	B	6.70	6.62	0.08
			His45A	S	7.00	6.73	0.27
			His50A	S	7.50	6.78	0.72
			His72A	B	6.00	6.69	0.69
			His89A	S	7.20	6.79	0.41
			His112A	B	7.50	5.95	1.55
			Nterm1B	S	7.00	7.19	0.19
			His2B	S	6.51	6.51	0.00
			His77B	B	6.60	6.61	0.01

TABLE I. (Continued)

Protein	No. of measured pK_a values	PDB code	Charged group[a]	Region[b]	pK_{exp}	pK_{calc}	$\lvert pK_{exp} - pK_{calc} \rvert$
			His117B*	B	8.00	6.07	1.93
			His146B*	S	7.90	5.20	2.70
Barnase[66]	9	1a2p	Asp8	S	3.00	2.55	0.45
			Asp12	S	3.65	3.32	0.33
			Asp22	S	3.30	2.51	0.79
			Glu29	B	3.75	3.42	0.33
			Asp44	S	3.35	3.14	0.21
			Glu60	B	3.40	3.06	0.34
			Asp75	C	3.10	2.37	0.73
			Asp86	B	4.20	3.67	0.53
			C-term110	B	3.30	1.90	1.40
BPTI[77]	10	4pti	N-term1	S	7.94	7.63	0.31
			Asp3	S	3.57	3.57	0.00
			Glu7	B	3.89	4.08	0.19
			Lys15	S	10.43	10.48	0.05
			Lys26	S	10.10	10.44	0.34
			Lys41	S	10.60	10.63	0.03
			Lys46	S	9.87	10.33	0.46
			Glu49	S	4.00	3.76	0.24
			Asp50	B	3.18	3.84	0.66
			C-term58	S	3.05	3.62	0.57
Carbon monoxy sperm whale myoglobin[22]	10	2mb5	His12	S	6.30	6.23	0.07
			His36*	B	8.00	6.23	1.77
			His48	S	5.30	6.57	1.27
			His81	S	6.60	6.59	0.01
			His97	S	5.60	6.72	1.12
			Tyr103	B	10.30	10.19	0.11
			His113	B	5.40	6.35	0.95
			His116	S	6.50	6.45	0.05
			His119	B	6.10	5.30	0.80
			Tyr151	B	10.50	10.47	0.03
Designed leucine zipper[78]	10	1fmh	Glu1A	S	4.14	3.29	0.85
			Glu6A	S	4.82	3.42	1.40
			Glu8A	B	4.52	4.22	0.30
			Glu13A	B	4.37	3.41	0.96
			Glu15A	S	4.11	2.60	1.51
			Glu20A	B	4.41	4.44	0.03
			Glu22A	S	4.82	4.21	0.61
			Glu27A	B	4.65	3.98	0.67
			Glu29A	S	4.63	3.77	0.86
			Glu1B	S	4.22	4.23	0.01
Calbindin D_{9K} (apo)[79]	9	2bca	Lys1	S	10.59	11.65	1.06
			Lys7	S	11.36	11.25	0.11
			Lys12	S	11.06	11.28	0.22
			Lys16	B	11.06	11.48	0.42
			Lys29	S	11.28	11.21	0.07
			Lys41	S	10.93	10.98	0.05
			Lys55	S	12.12	11.43	0.69
			Lys71	S	10.72	11.06	0.34
			Lys72	B	11.33	11.53	0.20
Xylanase[80]	9	1xnb	Asp4	S	3.00	3.72	0.72
			Asp11	B	2.50	3.17	0.67
			Glu78	B	4.60	4.26	0.34
			Asp106	B	2.70	3.47	0.77
			Asp119	S	3.20	4.24	1.04
			Asp121	S	3.60	3.68	0.08
			His156	B	6.50	6.59	0.09

TABLE I. (Continued)

Protein	No. of measured pK_a values	PDB code	Charged group[a]	Region[b]	pK_{exp}	pK_{calc}	$\lvert pK_{exp} - pK_{calc} \rvert$
			Glu172*	B	6.70	4.74	1.96
			C-term185	B	2.70	2.95	0.25
Xylanase Q127A[81]	2	1hv1	Glu78	S	4.10	4.73	0.63
			Glu172*	S	7.30	4.56	2.74
Xylanase Y80F[81]	2	1hv0	Glu78	B	4.80	4.33	0.47
			Glu172*	B	7.70	4.81	2.89
Xylanase E172C[81]	1	1bcx	Glu78	B	4.00	4.58	0.58
Mouse epidermal growth factor[82]	8	1epi	N-term1	S	7.70	7.55	0.15
			Asp11	S	3.90	3.89	0.01
			His22	B	6.80	6.54	0.26
			Glu24	S	4.10	3.92	0.18
			Asp27	S	4.00	3.36	0.64
			Asp40	S	3.60	3.33	0.27
			Asp46	S	3.80	3.69	0.11
			C-term53	S	3.50	3.52	0.02
Reduced human thioredoxin[83]	6	1ert	Glu13	S	4.80	4.19	0.61
			Asp16	S	3.70	3.73	0.03
			Asp20	S	3.60	4.22	0.62
			His43	S	5.50	6.71	1.12
			Glu103	S	4.90	4.28	0.62
			C-term105	S	3.20	3.26	0.06
RNAse T1[84,85]	6	1rga	His27	B	7.30	7.61	0.31
			Glu28*	S	5.90	4.33	1.57
			His40	B	7.90	6.42	1.48
			Glu58	B	4.30	3.73	0.57
			Asp76*	C	0.50	2.78	2.28
			His92	B	7.80	6.72	1.08
Cytochrome c[86]	5	1hrc	His18*	B	2.40	5.61	3.21
			His26*	B	2.90	5.67	2.77
			His33	B	6.40	6.44	0.36
			Tyr67	B	11.00	11.85	0.85
			Lys79	S	9.00	9.93	0.93
Ribosomal protein L9[87]	6	1div	Asp8	S	2.99	3.43	0.44
			Glu17	B	3.57	3.88	0.31
			Asp23	S	3.05	3.31	0.26
			Glu38	S	4.04	3.61	0.43
			Glu48	B	4.21	4.34	0.13
			Glu54	B	4.21	4.10	0.11
Asp(B9) human insulin[88]	6	1mhi	Glu4A	B	2.62	4.15	1.53
			C-term21A*	B	3.17	4.94	1.77
			Asp9B	S	2.60	3.77	1.17
			Glu13B*	B	2.20	4.31	2.11
			Glu21B	B	3.71	4.67	0.96
			C-term30B	S	2.38	3.28	0.90
Turkey ovomucoid inhibitor[89,90]	4	1ppf	Glu10	S	4.10	3.76	0.34
			Glu19	B	3.20	4.20	1.00
			Asp27	B	2.30	3.39	1.09
			Glu43	S	4.80	4.30	0.50
Neurotoxin III[91]	4	1ans	Tyr7	S	9.80	9.53	0.27
			Tyr18	B	10.10	10.77	0.67
			Glu20	S	5.40	4.27	1.13
			C-term27	B	3.30	4.63	1.33

TABLE I. (Continued)

Protein	No. of measured pK_a values	PDB code	Charged group[a]	Region[b]	pK_{exp}	pK_{calc}	$\lvert pK_{exp} - pK_{calc} \rvert$
Staphylococcal nuclease[92]	4	1ey0	His8	S	6.82	6.74	0.08
			His46*	B	5.80	7.60	1.80
			His121*	B	5.49	7.52	2.03
			His124	S	5.99	6.71	0.72
Staph. Nuclease V66K[37]	1	2snm	Lys66*	C	5.50	7.86	2.36
Staph. Nuclease V66E[93]	1	phsv66e[c]	Glu66*	C	8.70	5.47	3.23
Proteinase inhibitor IIA[94]	4	2bus	Asp6	S	4.00	3.11	0.89
			Glu9	S	4.30	3.62	0.68
			Asp12	S	3.60	3.71	0.11
			Glu20	S	4.10	4.30	0.20
Fibronectin Type III[95]	4	1fna	Glu38	S	3.79	3.72	0.07
			Glu47	B	3.94	5.10	1.16
			Asp67	B	4.18	4.78	0.60
			Asp80	S	3.40	3.54	0.14
Semi-synthetic RNAse A D21N[96]	3	3srn	His12	B	6.02	5.37	0.65
			His105	B	6.54	5.51	1.03
			His119	S	6.18	6.59	0.41
Oxidized *E. coli* thioredoxin[40]	3	2trx	N-term1	S	7.34	7.74	0.40
			His6	S	6.04	6.81	0.77
			Asp26*	C	7.50	5.93	1.57
Glutathione transferase[97]	3	6gst	His83	B	5.20	6.45	1.25
			His84	S	7.10	6.51	0.59
			His167*	B	7.80	6.20	1.60
Tyrosine phosphatase[98]	2	1dg9	His66	B	8.29	7.44	0.85
			His72*	B	9.19	7.01	2.18
T4 lysozyme[55]	2	2lzm	His31*	B	9.10	7.24	1.86
			Asp70	B	0.50	1.41	0.91
Reduced *E. coli* thioredoxin[40]	1	3trx	N-term1	S	7.35	7.58	0.23
T4 lysozyme Q105E[99]	1	1l98	Glu105	C	6.00	4.98	1.02
T4 lysozyme M102K[100]	1	1l54	Lys102*	C	6.50	8.36	1.86
Subtilisin inhibitor[101]	2	3ssi	His43*	C	3.25	5.68	2.43
			His106	B	6.00	6.71	0.71
c-Myc-Max heterodimer[102]	2	1a93	His29A	S	6.85	6.77	0.08
			His11B	B	7.19	6.46	0.73
Erabutoxin[103]	1	3ebx	His6*	B	2.75	5.99	3.24
Subtilisin[104]	1	1dui	His64	B	7.17	7.54	0.37

[a]Outlier groups (see Fig. 3) are marked with asterisks.
[b]C = core, B = boundary, S = surface. A single region classification is shown for each histidine residue, determined by using the C_β-surface distance and the average solvent-accessibility of the imidazole nitrogens.
[c]Staphylococcal nuclease V66E X-ray coordinates kindly provided by Bertrand Garcia-Moreno.

Region-dependent Classification

All residues are classified either as core, boundary, or surface according to criteria involving depth below the surface and solvent-accessible surface area (Fig. 1). Atomic radii parameters were obtained from CHARMM22, a 1.4-Å radius probe was used for solvent-accessible surface area calculations, and a 7-Å probe was used for generating the pseudo-solvent accessible surface used to determine side-chain depth.[46] The N_δ and N_ϵ nitrogens in the histidine imidazole ring can experience distinct environments, requiring separate classifications of these atoms.

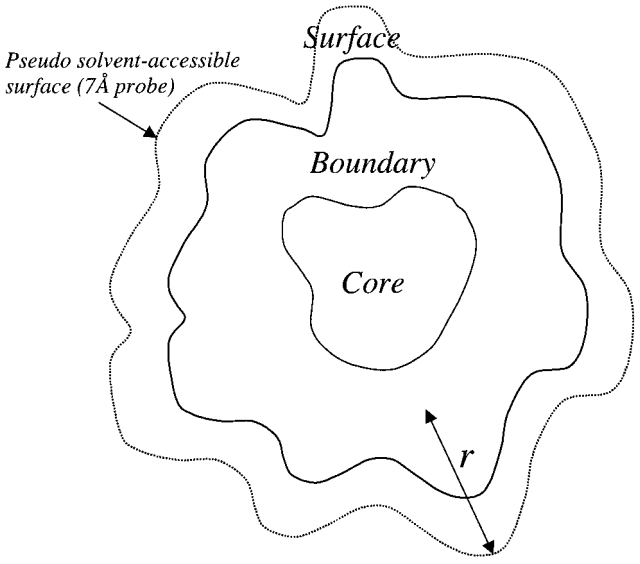

Fig. 1. Definition of the core, boundary, and surface regions. All residues are classified according to distance of their C_β atom to a pseudo solvent-accessible surface (r), calculated only in the presence of backbone and C_β atoms, and the fractional solvent-accessible surface area (A_f) of atoms involved in the ionization process (see Methods). Core: $r \geq 7.0$ Å and $A_f < 0.30$, or 5.7 Å $< r < 7.0$ Å and $A_f \leq 0.10$; boundary: $r \geq 7.0$ Å and $A_f \geq 0.30$, or 5.7 Å $< r < 7.0$ Å and $0.10 < A_f < 0.65$, or $r \leq 5.7$ Å and $A_f \leq 0.30$; surface: 5.7 Å $< r < 7.0$ Å and $A_f \geq 0.65$, or $r \leq 5.7$ Å and $A_f \geq 0.30$.

Local Interactions

Local interactions include ion pairs and hydrogen bonds and are determined by the type of groups involved and the geometry of the interaction. An ion pair is defined as two oppositely charged side chains within 4 Å of each other that are not involved in hydrogen bonds, independent of the angle of interaction. A proton donor and acceptor are considered involved in a hydrogen bond if the energy of interaction of the charge-charge or charge-polar pair is below a certain threshold (−1 kcal/mol) as determined by a distance- and angle-dependent potential function.[47]

Determination of pK_a Values

The fractional protonation state of titratable groups at a given pH is obtained by applying a Metropolis-Monte Carlo scheme[39] consisting of 10 equilibration steps per ionized group, followed by 10,000 sampling steps per group, at pH increments of 1.0. The pH value at half protonation (pK_{calc}) is obtained by linear interpolation of the pH titration curve. Calculation time was reduced by not recalculating all pairwise interactions after a single group changes protonation state, but rather by only calculating interactions with the changed group. To further decrease compute time, while retaining precision, only charged groups within 15 Å of any of the side chains whose pK_a values were being compared with experimental data were included in calculations. Using this methodology, the 260 pK_a values from the 41 proteins were calculated in approximately 20 min on a 1.7-GHz processor.

THEORY

Here we summarize the computational methods for calculating pK_a values of ionizable groups (based on published methods)[38,39] and present the construction of a new function that describes electrostatic interactions using multiple dielectric constants, and the method used to parameterize this function based on a large set of experimentally observed pK_a values.

Calculation of pK_a Values

Theoretical pK_a values can be calculated for all ionizable groups in a protein by constructing a pH-dependent, Boltzmann-distributed ensemble of all the combinations of charged and neutral forms of these groups in the protein, using the electrostatic potential function described below to determine the internal energy of each ensemble member.

The energy for a particular protonation state $\mathbf{x}_m$ of all N ionizable groups in a protein at a given pH is given by[38,39]

$$\Delta G(\mathbf{x}_m) = \sum_{i=1}^{N} x_i 2.3RT(pH - pK_{a,i}^{intr}) + \frac{1}{2}\sum_{\substack{i,j \\ i \neq j}}^{N} \Delta G_{QQ,ij}(q_i^0 + x_i, q_j^0 + x_j), \quad (1)$$

where $\mathbf{x}_m$ is a vector describing the protonation state x_i (0 or 1; i: $1 \rightarrow N$), $\Delta G_{QQ,ij}(q_i^0 + x_i, q_j^0 + x_j)$ is the pairwise electrostatic interaction energy between ionizable groups i and j, that have charges q_i^0 and q_j^0 in their deprotonated state (q^0 is −1 or 0). $pK_{a,i}^{intr}$ is given by

$$pK_{a,i}^{intr} = pK_{a,i}^{solution} - \frac{q}{2.3RT}\Delta G_{self,i}^{w \rightarrow p}, \quad (2)$$

where $pK_{a,i}^{solution}$ is the solution pK_a value of group i in water, $\Delta G_{self,i}^{w \rightarrow p}$ the change in self-energy of transferring the group from water to a neutral protein (i.e., a reference state in which all ionizable groups have a charge of 0), q the charge of the group in its ionized form (± 1), R the Boltzmann constant, and T the temperature. The$\Delta G_{QQ,ij}$ and $\Delta G_{self,i}^{w \rightarrow p}$ terms form the electrostatic potential and are described in the next section.

The fractional protonation state $\langle x_i \rangle$ of each group at a given pH is the Boltzmann-weighted average of the ionization state of that group within the ensemble of all protonation states of the protein:

$$\langle x_i \rangle = \frac{1}{Z}\sum_m x_{i,m} e^{-\Delta G(\mathbf{x}_m)/RT} \quad (3)$$

where $\Delta G(\mathbf{x}_m)$ the electrostatic energy of the mth member of the ensemble, $x_{i,m}$ is the protonation state of group i in the mth ensemble member, and Z the partition function:

$$Z = \sum_m e^{-\Delta G(\mathbf{x}_m)/RT} \quad (4)$$

Enumeration of Z increases as 2^N, where N is the number of ionizable groups, and is intractable for $N > 25$.

This upper limit can be extended to ~35 groups by not considering changes in protonation states when the group is >95% protonated or deprotonated at the pH value of the calculation.[38] In the proteins included in our calculations, we had to consider up to 131 ionizable groups in the case of hemoglobin. We therefore used a Metropolis-Monte Carlo procedure to construct the ensemble of protonation states.[39] Here members of the ensemble are constructed by randomly assigning protonation states x_i in a member of the ensemble $\{\mathbf{x}_m\}$, and constructing a Boltzmann-distributed Markov chain.[39] Fractional protonation states are then obtained as a simple arithmetic average of all values of x_i in $\{\mathbf{x}_m\}$:

$$\langle x_i \rangle = \frac{1}{M} \sum_m x_{i,m} \tag{5}$$

where M is the size of the ensemble (we find that M =10,000 is sufficient to obtain stable values). pK_a values are calculated by obtaining $\langle x_i \rangle$ at different pH values and determining by interpolation the pH value at which $\langle x_i \rangle$ = 0.5.

The Electrostatic Energy Function

The total electrostatic free energy, ΔG, of a system of point charges and permanent dipoles, relative to the charges at infinite separation *in vacuo*, can be written as follows[1]:

$$\Delta G = \Delta G_{QQ} + \Delta G_{self} \tag{6}$$

Here, ΔG_{QQ} is the Coulomb term for charge-charge interactions, and the self-energy ΔG_{self} is the interaction between the charges and the polarizable environment. This equation holds for any collection of charges in a polarizable medium and is therefore generalizable to the description of electrostatic interactions in both solvent and proteins. Relative to a collection of charges at infinite separation in solvent, the total free energy of charges in a solvated protein now becomes

$$\Delta G = \Delta G_{QQ} + \Delta G_{self}^{w \rightarrow p} \tag{7}$$

where ΔG_{QQ} is the Coulomb charge-charge interaction term (0 for infinitely separated charges in solvent), and $\Delta G_{self}^{w \rightarrow p}$ is the change in self-energy of the charges upon transfer from solvent to neutral protein.

Self-energy

The change in self-energy of a charge upon transfer from solvent to neutral protein, $\Delta G_{self}^{w \rightarrow p}$, consists of several terms, representing desolvation of the charge (ΔG_{self}^{solve}), the dipolar relaxation of the solvent in and around the protein (ΔG_{self}^{qw}), interactions of the charge with permanent partial charges in the protein ($\Delta G_{self}^{q\mu}$), and induced dipoles within the protein ($\Delta G_{self}^{q\alpha}$)[10,20]:

$$\Delta G_{self}^{w \rightarrow p} = \Delta G_{self}^{q\mu} + \Delta G_{self}^{q\alpha} + \Delta G_{self}^{qw} - \Delta G_{self}^{solv} \tag{8}$$

The terms that represent solvent interactions within the protein can be combined:

$$\Delta G_{solv} = \Delta G_{self}^{qw} - \Delta G_{self}^{solv} \tag{9}$$

Similarly, the interactions within the protein are combined as

$$\Delta G_{QP} = \Delta G_{self}^{q\mu} + \Delta G_{self}^{q\alpha} \tag{10}$$

yielding,

$$\Delta G_{self}^{w \rightarrow p} = \Delta G_{solv} + \Delta G_{QP} \tag{11}$$

We divide the protein into three, objectively defined regions: core, boundary, and surface (Fig. 1) and use a modified Born equation[26] to calculate a region-dependent solvation energy, $\Delta G_{solv,i}(p_i)$ (in kcal/mol) for a charged group i:

$$\Delta G_{solv,i}(p_i) = \frac{166 q_i^2}{R_{Born}} \left(\frac{1}{{}^{solv}\varepsilon(p_i)} - \frac{1}{\varepsilon_w e^{\kappa R_{Born}}} \right) (1 - A_{f,i}) \tag{12}$$

where q_i is the charge of the group in electron units, R_{Born} is the Born radius of the charge (in Å), ${}^{solv}\varepsilon(p_i)$ is a region-dependent protein desolvation dielectric, a constant that describes the dipolar solvent relaxation effects for a particular region p, ε_w is the dielectric of water, $\varepsilon^{\kappa R_{Born}}$ is a salt screening term $\kappa = 50.29 \times \sqrt{\frac{I}{\varepsilon_w T}}$, where I is the ionic strength of the buffer and T is the temperature),[12] and $A_{f,i}$ the fractional solvent-accessible surface area of the ionizable atoms in the charged group in the folded protein structure determined relative to the group in an extended tripeptide.[48]

In Equation 12, the value of ${}^{solv}\varepsilon(p_i)$ is the dominant determinant in the description of the dipolar solvent relaxation effect.[1,11] In most continuum models, the value of ${}^{solv}\epsilon$ is set to a single value, usually 4 or 20.[21,49,50] In our model we allow ${}^{solv}\varepsilon(p_i)$ to vary in a region-dependent manner to capture the dependence of dipolar solvent relaxation effects on the depth of the buried charge. ${}^{solv}\varepsilon(p_i)$ takes on two separate values (Table II), depending on the region that the ionizable amino acid is located in the following: one for the core (${}^{solv}\varepsilon(p_i) = {}^{solv}\varepsilon_C$), and one for the boundary region (${}^{solv}\varepsilon(p_i) = {}^{solv}\varepsilon_B$). In the surface region, groups are exposed to solvent and the desolvation penalty is negligible; $\Delta G_{solv,i}(p_i)$ is therefore set to 0 in this region. A single Born radius is used for all ionizing groups. The $(1 - A_{f,i})$ term in Equation 12 corrects for any local structural details in the core or boundary region that allow solvent access, diminishing the magnitude of $\Delta G_{solv,i}(p_i)$ accordingly. The values for ${}^{solv}\varepsilon_C$, ${}^{solv}\varepsilon_B$, and R_{Born} are determined by empirical fit of calculated pK_a values to experimentally observed values (see below).

The pairwise interactions between charged and polar groups described by the second term of the self-energy, $\Delta G_{QP,ij}(p_{ij}, r_{ij})$, depend on region (core, boundary, surface) as well, but also on local environment (hydrogen bonding, nonhydrogen bonding), and distance between the charges:

TABLE II. Model Parameters

Parameter	Optimal value[a]	Number of contributing residues[b]	Meaning of parameter
$^{solv}\varepsilon_C$	19	18	Solvation parameters
$^{solv}\varepsilon_B$	32	100	
R_{Born}	2.0	118	
$^{QQ}\alpha_C$	110	42	Pairwise charge-charge (*QQ*) and charge-polar (*QP*) interaction parameters
$^{QQ}\alpha_B$	110	107	
$^{QQ}\alpha_S$	150	144	
$^{QQ}\lambda_C$	0.2	42	
$^{QQ}\lambda_B$	0.4	107	
$^{QQ}\lambda_S$	1.4	144	
$^{QP}\alpha_C$	90	42	
$^{QP}\alpha_B$	110	107	
$^{QP}\alpha_S$	150	144	
$^{QP}\lambda_C$	0.2	42	
$^{QP}\lambda_B$	0.5	107	
$^{QP}\lambda_S$	0.9	144	
$\Delta^{HbQP}G_C$	−1.0	15	Charge-polar (*HbQP*) and charge-charge (*HbQQ*) hydrogen bonding parameters
$\Delta^{HbQP}G_B$	−0.25	31	
$\Delta^{HbQP}G_S$	1.0	14	
$\Delta^{HbQQ}G_C$	−5.25	4	
$\Delta^{HbQQ}G_B$	−2.25	10	
$\Delta^{HbQQ}G_S$	3.75	10	
$^{IP}\varepsilon_C$	16	9	Ion pairing parameters
$^{IP}\varepsilon_B$	135	23	
$^{IP}\varepsilon_S$	165	18	

[a]Determined by fit of model to experimental data.
[b]The number of residues used to determine parameter minima for the core and boundary solvation dielectrics ($^{solv}\varepsilon_C$, $^{solv}\varepsilon_B$) differs from those used for the core and boundary charge-charge and charge-polar pairwise interaction parameters ($^{QQ}\alpha_C$, $^{QQ}\alpha_B$, $^{QQ}\alpha_S$, $^{QQ}\lambda_C$, $^{QQ}\lambda_B$, $^{QQ}\lambda_S$, $^{QP}\alpha_C$, $^{QP}\alpha_B$, $^{QP}\alpha_S$, $^{QP}\lambda_C$, $^{QP}\lambda_B$, $^{QP}\lambda_S$), because the special treatment of histidine side chains as two independent ionizable groups results in cases of a single histidine residue belonging to two different regions (see Fig. 2 and Methods).

$$\Delta G_{QP,ij}(p_{ij}, r_{ij}) = \begin{cases} 332\,\dfrac{q_i q_j}{^{QP}\varepsilon(p_{ij}, r_{ij}) r_{ij} e^{\kappa r_{ij}\overline{A_{f,ij}}}} & i,j \text{ are not hydrogen bonded} \\ \Delta^{HbQP}G(p_{ij}) & i,j \text{ form a hydrogen bond,} \end{cases} \quad (13)$$

where q_i and q_j represent the charge (electron units) of ionizable group i in its charged state and partial charge of atom j, respectively, r_{ij} is the distance between the groups (in Å), $^{QP}\varepsilon(p_{ij}, r_{ij})$ is the distance- and region-dependent dielectric assigned to this interacting pair (see below), $e^{\kappa r_{ij}\overline{A_{f,ij}}}$ is a solvent exposure-dependent salt screening correction term,[12] where $\overline{A_{f,ij}}$ is the average solvent-accessible surface area of interacting groups i and j. In these nonbonded interactions, the dielectric constant describes relaxation effects not explicitly accounted for, such as induced dipoles[10] and small vibrations,[15,16] which are dependent on local environment and distance.[10,51] We therefore again assign dielectric constants separately for the core, boundary, and surface regions ($^{QP}\varepsilon_C$, $^{QP}\varepsilon_B$, $^{QP}\varepsilon_S$). We use a distance-dependent dielectric of an exponential form,[1,35]

$$^{QP}\varepsilon(p_{ij}, r_{ij}) = 1.0 + {}^{QP}\alpha(p_{ij})(1.0 - e^{-^{QP}\lambda(p_{ij})r_{ij}}) \quad (14)$$

where $^{QP}\alpha(p_{ij})$ is the dielectric constant at large separation distance and $^{QP}\lambda(p_{ij})$ is a damping constant, both of which are region dependent. For interactions between pairs located in different regions, we use a combination rule in which the arithmetic average of the parameters assigned to the two regions is used.

The hydrogen-bonding term, $\Delta^{HbQP}(p_{ij})$, is used for all charge-polar pairs that form a hydrogen bond and is also assigned separately for each of the core, boundary, and surface regions ($\Delta^{HbQP}G_C$, $\Delta^{HbQP}G_B$, $\Delta^{HbQP}G_S$). The arithmetic average is used to combine inter-region interactions.

Charging Energy

Pairwise interaction energies between two ionizable groups are also calculated in a region-, environment-, and distance-dependent manner:

$$\Delta G_{QQ,ij}(p_{ij}, r_{ij}) = \begin{cases} 332\,\dfrac{q_i q_j}{^{QQ}\varepsilon(p_{ij}, r_{ij}) r_{ij} e^{\kappa r_{ij}\overline{A_{f,ij}}}} & i,j \text{ are not hydrogen bonded} \\ \Delta^{HbQQ}G(p_{ij}) & i,j \text{ form a hydrogen bond.} \end{cases} \quad (15)$$

In this equation, q_i and q_j are the pH-dependent charges of the interacting ionizable groups (Eq. 1), and $^{QQ}\varepsilon(p_{ij}, r_{ij})$ is a region- and distance-dependent dielectric constant. We distinguish between three classes of interactions: charge-charge hydrogen bonds, non-hydrogen bonding ion pairs, and simple Coulombic. Charge-charge hydrogen bonds between two ionizable groups are modeled using a region-dependent energy term, $\Delta^{HbQQ}G(p_{ij})$, analogous to $\Delta^{HbQP}G(p_{ij})$, described above. Ion pairs are defined as oppositely charged ionizable groups that lie within 4 Å of each other[52] but that do not conform to a canonical hydrogen bond geometry,[47] and are treated using a position-, but not distance-dependent dielectric, $^{IP}\varepsilon(p_{ij})$. For the remaining, simple Coulombic interactions we use a region- and distance-dependent dielectric[1,27,35]:

$$^{QQ}\varepsilon(p_{ij}, r_{ij}) = 1.0 + {}^{QQ}\alpha(p_{ij})(1.0 - e^{-^{QQ}\lambda(p_{ij})r_{ij}}) \quad (16)$$

analogous to Equation 14. The $^{QQ}\varepsilon(p_{ij}, r_{ij})$ parameters account for dipolar relaxation effects within the protein matrix in response to presence of charged groups,[14] which are also dependent on local environment and distance.[14,20] As before, the arithmetic average of the parameters is used to combine inter-region interactions.

Parameterization

There are 109 partial atomic charges that describe the permanent charge distribution of the 20 amino acids in their neutral states. The values for these were taken from the PARSE parameter set.[44] Equations 12–16 constitute the model for simulating the electrostatic behavior of charged groups and contain 24 independent parameters

TABLE III. Summary of the Minimization Procedure

	Iteration[a]					
Parameter	1	2	3	4	5	6
$^{solv}\varepsilon_C$	[4, 80]/4	[12, 28]/1	19	19	19	19
$^{solv}\varepsilon_B$	[4, 80]/4	[16, 80]/4	32	32	32	32
R_{Born}	[1.5, 2.5]/0.1	2.0	2.0	2.0	2.0	2.0
$^{QQ}\alpha_C$	[40, 150]/10	[40, 150]/10	[40, 150]/10	[40, 150]/10	[70, 150]/10	110
$^{QQ}\alpha_B$	[40, 150]/10	[40, 150]/10	[80, 150]/10	[80, 150]/10	[80, 150]/10	110
$^{QQ}\alpha_S$	[40, 150]/10	[40, 150]/10	[80, 150]/10	[100, 150]/10	[140, 150]/10	150
$^{QQ}\lambda_C$	[0.1, 1.5]/0.1	[0.1, 1.5]/0.1	[0.1, 1.5]/0.1	[0.1, 1.5]/0.1	[0.1, 1.5]/0.1	0.2
$^{QQ}\lambda_B$	[0.1, 1.5]/0.1	[0.1, 1.5]/0.1	[0.1, 1.5]/0.1	[0.1, 1.5]/0.1	[0.2, 1.5]/0.1	0.4
$^{QQ}\lambda_S$	[0.1, 1.5]/0.1	[0.1, 1.5]/0.1	[0.2, 1.5]/0.1	[0.2, 1.5]/0.1	[0.3, 1.5]/0.1	1.4
$^{QP}\alpha_C$	[40, 150]/10	[40, 150]/10	[40, 150]/10	[40, 150]/10	[40, 150]/10	90
$^{QP}\alpha_B$	[40, 150]/10	[40, 150]/10	[40, 150]/10	[40, 150]/10	[60, 150]/10	110
$^{QP}\alpha_S$	[40, 150]/10	[40, 150]/10	[70, 150]/10	[70, 150]/10	[100, 150]/10	150
$^{QP}\lambda_C$	[0.1, 1.5]/0.1	[0.1, 1.5]/0.1	[0.1, 1.5]/0.1	[0.1, 1.5]/0.1	[0.2, 1.5]/0.1	0.2
$^{QP}\lambda_B$	[0.1, 1.5]/0.1	[0.1, 1.5]/0.1	[0.2, 1.5]/0.1	[0.2, 1.5]/0.1	[0.4, 1.5]/0.1	0.5
$^{QP}\lambda_S$	[0.1, 1.5]/0.1	[0.1, 1.5]/0.1	[0.3, 1.5]/0.1	[0.3, 1.5]/0.1	[0.7, 1.5]/0.1	0.9
$\Delta^{HbQP}G_C$	[−3.0, 3.0]/0.5	[−1.5, 1.5]/0.25	[−2.0, 1.0]/0.5	−1.0	−1.0	−1.0
$\Delta^{HbQP}G_B$	[−3.0, 3.0]/0.5	[−1.5, 1.5]/0.25	[−1.5, 1.5]/0.5	−0.25	−0.25	−0.25
$\Delta^{HbQP}G_S$	[−3.0, 3.0]/0.5	[1.0, 4.0]/0.25	[0.0, 3.0]/0.5	1.0	1.0	1.0
$\Delta^{HbQQ}G_C$	[−6.0, 6.0]/0.5	[−6.0, 0.0]/0.5	[−6.0, −3.5]/0.25	[−6.0, −3.5]/0.25	−5.25	−5.25
$\Delta^{HbQQ}G_B$	[−6.0, 6.0]/0.5	[−6.0, 0.0]/0.5	[−3.0, −1.0]/0.25	[−3.0, −1.0]/0.25	−2.25	−2.25
$\Delta^{HbQQ}G_S$	[−6.0, 6.0]/0.5	[0.0, 6.0]/0.5	[2.5, 4.5]/0.25	[2.5, 4.5]/0.25	3.75	3.75
$^{IP}\varepsilon_C$	[4, 204]/10	[4, 204]/10	[8, 24]/2	[12, 22]/1	16	16
$^{IP}\varepsilon_B$	[4, 204]/10	[4, 204]/10	[34, 204]/10	[100, 200]/10	[130, 160]/5	135
$^{IP}\varepsilon_S$	[4, 204]/10	[54, 204]/10	[84, 204]/10	[130, 200]/10	[160, 200]/5	165

[a]The numbers in brackets are the minimum and maximum values of the parameter for the iteration; the number following the slash corresponds to the step size. Entries with single values represent the use of a single parameter value for the iteration (i.e., converged parameters).

describing the multiple, geometry-dependent dielectric constants (Table II). The values of these were obtained by fitting the model to the set of 260 experimentally observed pK_a values, minimizing the RMSD between experimental and calculated values.

The minimization (Table III) was performed using a stochastic search method in which random values of the 24 parameters were generated. For each randomly generated parameter set, the corresponding 260 pK_a values were calculated by generating titration curves at 1 pH unit intervals (from 0 to 14) using the Monte Carlo method, as described above, allowing an RMSD value to be determined that correlate the calculated pK_a values associated with the parameter set with the experimental observation. This parameter space is extremely large. An iterative minimization scheme was therefore used in which Monte Carlo sampling progressively focuses in on the minima of the individual parameters (Fig. 2).

RESULTS

The model was first parameterized by a fit to the empirically observed pK_a values of 260 ionizable groups (Table I). Convergence for all 24 parameters was achieved in five iterations of the Monte Carlo search method described above (see Supplementary Material for details). To investigate the stability of the parameters obtained in this manner, we carried out additional rounds of Monte Carlo explorations in which parameters that had converged in early stages of the search were allowed to vary while keeping late-converged parameters constant. We found little change in the initially established values, nor was the overall RMSD of the parameter set affected (not shown). The values of the parameters therefore appear to have converged stably. Table II shows the final values for all the parameters.

The overall RMSD between the experimental and calculated values of the 260 ionizable groups included in this study is 0.95 pH units. For the subset of 68 groups that exhibit large experimentally observed pK_a shifts (≥1.0 pH units), the RMSD is 1.56 pH units; for the other 192 groups it is 0.56 pH units. In all cases, the method performs better than the null model (predicted pK_a values correspond simply to the solution pK_a values). Scatter plots show that deviations are dominated by a relatively small number of outliers (Fig. 3). If 10% of the groups, representing the worst outliers, are removed, the RMSD errors are improved to 0.65, 0.53, and 0.98 pH units respectively for the overall fit, for small shifts, and for large shifts in experimentally observed pK_a values.

DISCUSSION

The model presented here attempts to capture the major components of electrostatic interactions in proteins using a set of 24 empirically fit parameters (Table II). These parameters can be divided into three classes, representing contributions from dipolar solvent relaxation effects (R_{Born}, $^{solv}\varepsilon$), relaxation effects in the protein interior (α, λ), and noncovalent interactions ($^{IP}\varepsilon$, $\Delta^{Hb}G$). Within each class,

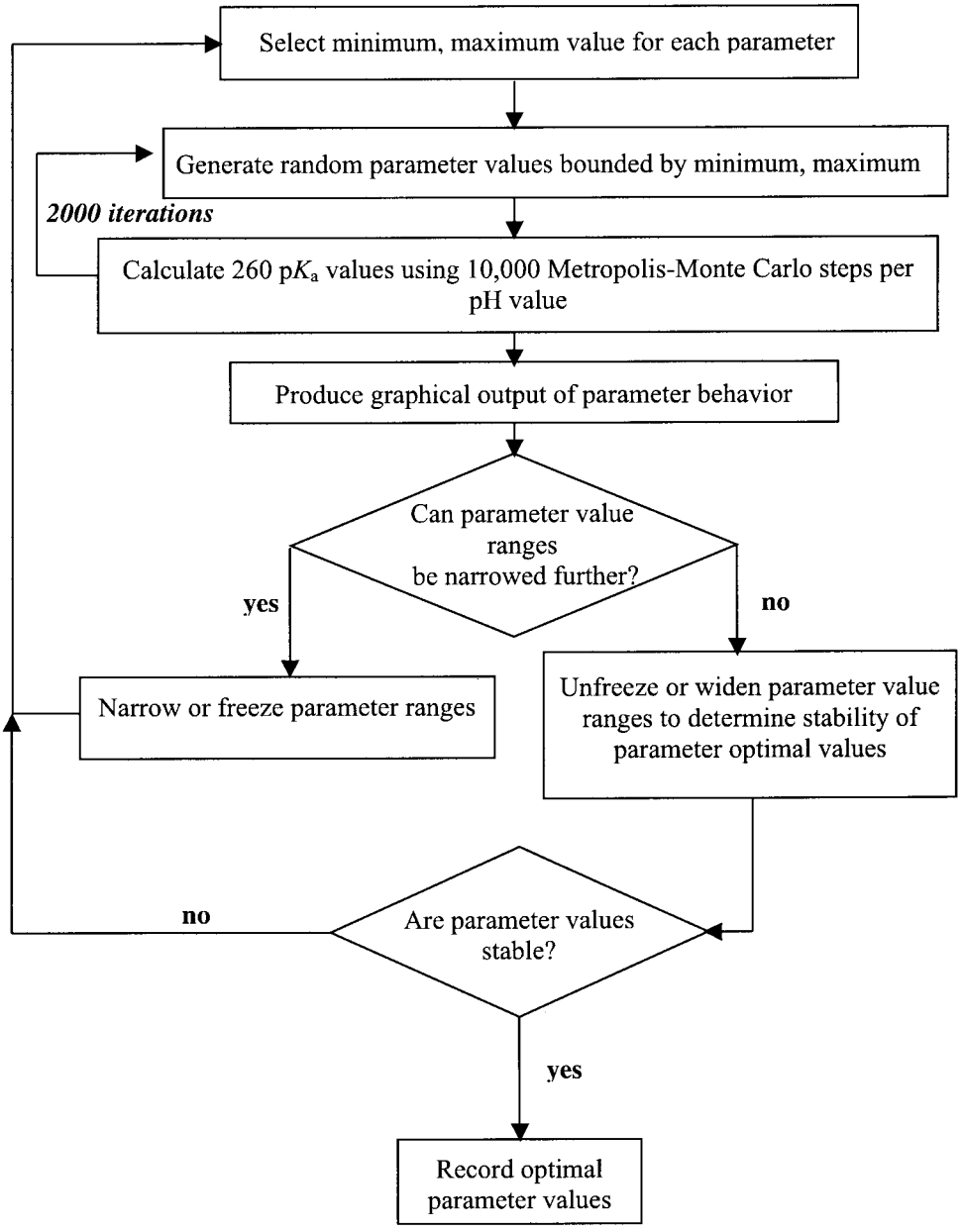

Fig. 2. Procedure to find parameter minima. For each iteration step, 2000 random combinations of the parameters were generated using the stochastic method described in the text, sampling each parameter at fixed intervals within a specified range. At the completion of each iteration, trends in the variation of the RMSD fit were visualized graphically for each parameter. To generate plots recording the RMSD as a function of an individual parameter, the RMSD was calculated over a subset containing only those ionizable groups that are most affected by that individual parameter (e.g., selecting ion pairing groups in the core for ${}^{IP}\varepsilon_C$, core charge-polar hydrogen bonding groups for $\Delta^{HbQP}G_C$, and so on for each region-specific and interaction-specific parameter). These trends were examined to identify emerging minima. This information was used in the next iteration to change the sampling limits and densities for individual parameters. At convergence, all 24 parameters are limited to single values. The parameters determining the distance dependence of charge-charge and charge-polar pairwise interactions (Eqs. 14 and 16), and the dielectric constant for ion-pair interactions in the boundary and surface (${}^{IP}\varepsilon_B$, ${}^{IP}\varepsilon_S$), do not show visually identifiable minima, whereas the minimum for ${}^{IP}\varepsilon_C$ is clear (Supplemental Figure 4). Nevertheless, for each of these parameters, an optimal value was established by numerically identifying the minimum.

the parameters are further subdivided depending on the region in which the interacting groups are located (C, core; B, boundary; S, surface), and interaction type (QQ, charge-charge; QP, charge-polar). This scheme provides an objective and physically intuitive description of the different types of residue microenvironments. Core and surface regions of proteins clearly differ in their proximity to the solvent and the rigidity of the residues located in these

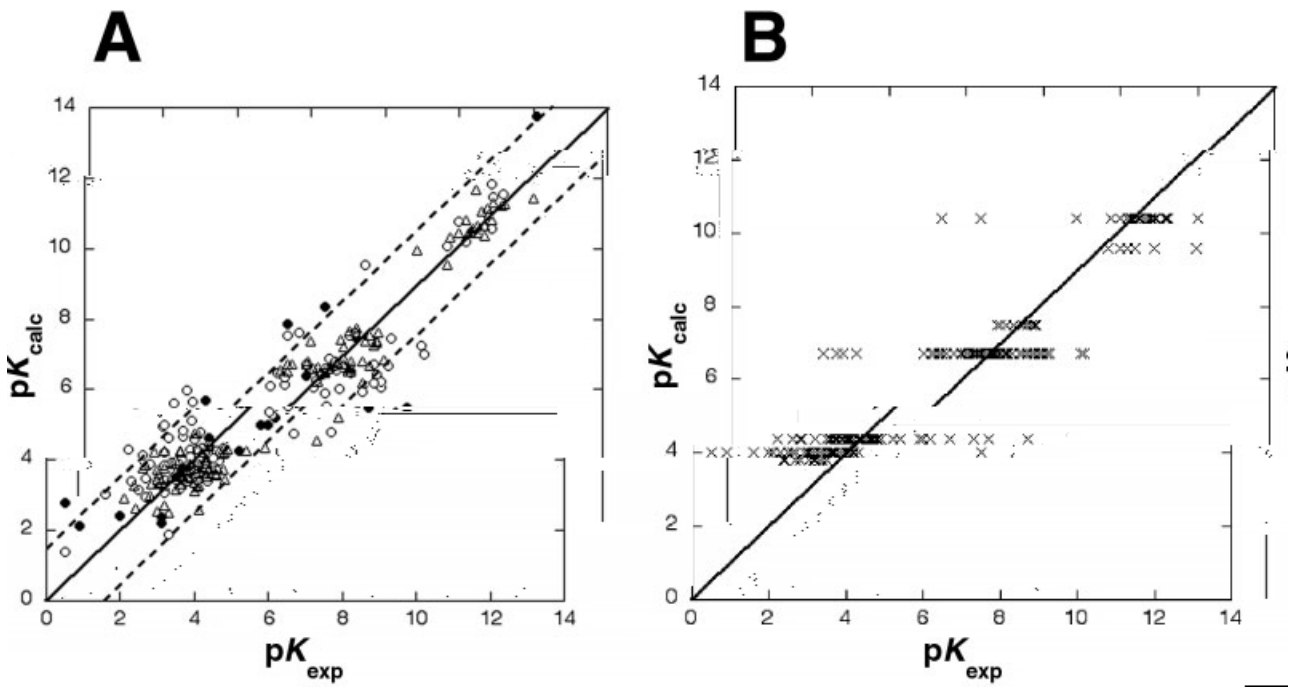

Fig. 3. Scatter plots showing the correlation between pK_{calc} and pK_{exp} for our model (A) and the null model (B). Closed circles represent the core residues, open circles represent boundary residues, triangles represent surface residues, and crosses represent all residues. The solid diagonal line represents perfect agreement with experiment ($pK_{exp} = pK_{calc}$), and the dashed diagonal lines show the worst 10% outlier boundary (inside these lines, $|pK_{exp} - pK_{calc}| < 1.55$ pH units).

regions. Distinct treatment of these regions was used to capture differences in solvent relaxation effects (proximity to the solvent) and interior relaxation effects (rigidity of the protein matrix). The boundary classification was introduced for those residues that cannot be placed clearly in either region. In all cases, we find that there is strong region dependence of the parameters that are assigned in this manner (Table II). Furthermore, we find that the magnitudes of these parameters rank the interactions in a physically intuitive order. This correlation between the model and microscopic physical mechanisms provides a strong validation for the approach presented here, because the parameter values are obtained objectively from an empirical fit of the pK_a data.

Dipolar solvent relaxation effects are captured using the Born radius to describe the cavity size of the charge and separate desolvation dielectric constants for the core and boundary regions of the protein (there is no surface desolvation dielectric, because in this region groups are exposed to solvent and the desolvation penalty is negligible). The optimal value of the Born radius is similar to that used in other studies.[21,24–26] The dielectric constant determining the desolvation penalty is smaller in the core (${}^{solv}\varepsilon_C = 19$) than in the boundary (${}^{solv}\varepsilon_B = 32$), corresponding to a larger penalty for placing charges in the core than in the boundary, consistent with experimental observations on buried charges[37] and theoretical studies finding that use of a dielectric constant of 20 results in better prediction of pK_a values.[21,53]

Relaxation effects in the protein interior are the result of induced dipoles,[10] small vibrations,[15,16] and dipolar relaxation effects,[14] which in this model are captured using distance- and environment-dependent dielectric constants. We use an exponential function to model the distance dependence (a linear function, or constant did not produce adequate fits; not shown), where the parameters α and λ describe the dielectric at infinite at distance and the dampening factor, respectively. High values of these two parameters imply that the electrostatic interactions are either weak (α) or attenuated over short distances (λ) in the protein matrix. These two parameters are further divided into core, boundary, and surface regions to account for differences in the average rigidity of these regions.

Furthermore, separate interactions are assigned for charge-polar (*QP*) and charge-charge (*QQ*) interactions to model induced dipoles or small vibrations and dipolar relaxation effects, respectively. The distance dependence of pairwise charge-charge and charge-polar dielectric constants is region dependent. In all cases, the dielectric constants are smallest in the core (${}^{QQ}\alpha_C \approx {}^{QQ}\alpha_B < {}^{QQ}\alpha_S$; ${}^{QP}\alpha_C < {}^{QP}\alpha_B < {}^{QP}\alpha_S$) and are also much less dampened in the core (${}^{QQ}\lambda_C < {}^{QQ}\lambda_B < {}^{QQ}\lambda_S$; ${}^{QP}\lambda_C < {}^{QP}\lambda_B < {}^{QP}\lambda_S$), consistent with increased interaction strength in the core relative to the other two regions, presumably as a consequence of increased rigidity. Furthermore, the relaxation effects are weaker for charge-polar than charge-charge interactions, as expected from theory.[14]

Contributions from noncovalent interactions are divided into hydrogen bonding and ion pairing. Hydrogen bonds are modeled as partially covalent interactions, identified both by distance and angular geometry,[47] and assigned an interaction energy, whereas ion pairs are modeled as

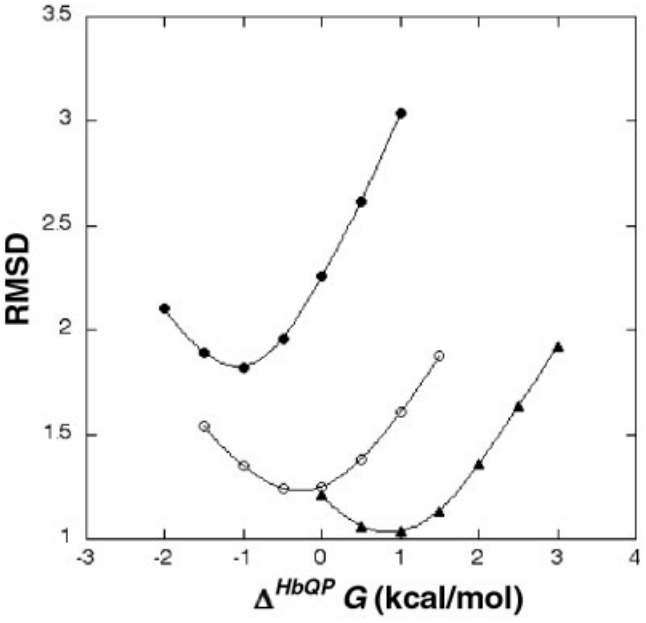

Fig. 4. Dependence of RMSD on the value of $\Delta^{HbQP}G_C$ (filled circles), $\Delta^{HbQP}G_B$ (open circles), and $\Delta^{HbQP}G_S$ (triangles) for the final iteration of the Monte Carlo search for these parameters (iteration 3).

purely Coulombic, distance-dependent interactions. The hydrogen bond energies are classified both by region and by charge type of the interacting partners. Ion pairs are classified by region only. The empirically obtained rank orderings for the free energy of formation of hydrogen bonds in the three regions indicate that hydrogen bonds are favored in core and boundary regions, but penalized on the surface (Fig. 4). These differences in the contributions made by hydrogen bonds is consistent with the interactions on the protein surface between side chains and solvent being favored relative to those with other protein groups.

This effect is further emphasized by the observation that the penalty for forming surface hydrogen bonds is much greater for charge-charge than for charge-polar interactions ($\Delta^{HbQQ}G_S \gg \Delta^{HbQP}G_S$). The dielectric constants for ion pairs rank order in a similar manner ($^{IP}\varepsilon_C < {}^{IP}\varepsilon_B < {}^{IP}\varepsilon_S$), consistent with ion pairs exerting the strongest effect in the hydrophobic core, as shown by experiment,[54,55] whereas ion pairs on the surface have little effect.[56–59] Furthermore, the experimentally observed contributions both of core ion pairs (3–5 kcal/mol)[54,55] and charge-polar hydrogen bonds (1–2 kcal/mol)[60] are similar to the values obtained here by empirical fit ($\Delta^{HbQQ}G_C = -5.25$ kcal/mol, and $\Delta^{HbQP}G_C = -1.0$ kcal/mol).

We have obtained relatively good fits for a large number of experimental observations, and the parameters exhibit physically meaningful behavior. Although these results suggest that a number of physical effects are captured appropriately, there is room for improvement. The significant decrease in RMSD values obtained by omitting 10% of the worst outliers (26 groups) suggests that there remain effects that are not captured correctly for a relatively small number of cases. Deviations between predicted and observed values are correlated with region: the largest deviations are observed in the core, the smallest on the surface [Fig. 5(A)]. This in turn correlates with the magnitude in pK_a shifts [Fig. 5(B)], the largest effects being inside the protein, whereas at the surface a high-dielectric environment dampens out any shifts from solution pK_a values, as has been noted before.[34,35] There are potentially four major sources for inaccuracies in the model: errors in the empirical fit of the parameters, resulting from incorrect minimization; omission of relevant physical effects; molecular modeling deficiencies due to incorrect placement of charges on the amino acids, insufficient sampling of the protein conformations in the pK_a calculations, and incorrect modeling of the unfolded state; and experimental errors involving deficiencies in structure and pK_a determinations.

It is unlikely that the minimization has led to the identification of false minima, because the final values were shown to be stable in Monte Carlo searches that allowed subsets to vary while holding the remaining parameters constant. Despite the use of 260 pK_a values, some parameters are determined by a relatively small numbers of experimental observations (Table II). In general, ~10% of the groups contribute to the core, ~40% to the boundary, and ~50% to the surface. Consequently, the most important parameters in this model, namely those associated with the core, are still somewhat underdetermined. This problem is exacerbated for the environment-dependent parameters. In particular, pairwise interactions involving ion pairs and hydrogen bonds are severely underdetermined in the core, where there are 28 such interactions (compared with 64 in the boundary, and 42 on the surface).

The model appears to capture correctly a significant number of physical effects, because parameters such as the dielectric constants and hydrogen bonding strengths are correlated with experimentally verifiable values or microscopic models. However, inaccuracies in the dominant terms are a possible source of error. Approximately 70% of the total electrostatic energy arises from the self-energy term ($\Delta G_{self}^{w\rightarrow p}$) that describes the contributions of the movement of a charge from solvent to protein interior. Correct modeling of the effects that contribute to this factor is therefore of paramount importance (Eqs. 11–14). As has been noted before, parameterization of $\Delta G_{self}^{w\rightarrow p}$ is a general problem for continuum models. The use of multiple, position-, and environment-dependent $\Delta G_{self}^{w\rightarrow p}$ dielectric constants is an improvement over models that use a single dielectric constant[21,22,53] and models that ignore the self-energy term.[27,34] $\Delta G_{self}^{w\rightarrow p}$ is modeled by 11 independent parameters, which capture dependence on position ($^{solv}\varepsilon_C$, $^{solv}\varepsilon_B$), environment ($\Delta^{HbQP}G_C$, $\Delta^{HbQP}G_B$, $\Delta^{HbQP}G_S$), and distance ($^{QP}\alpha_C$, $^{QP}\alpha_B$, $^{QP}\alpha_S$, $^{QP}\lambda_C$, $^{QP}\lambda_B$, $^{QP}\lambda_S$). It is unlikely that the addition of more parameters will improve the model, without introducing arbitrary parameters that have no clear physical basis.

Molecular modeling deficiencies, such as the oversimplification of charge distribution, the conformation of ionizable side chains, and the assumption of an absence of

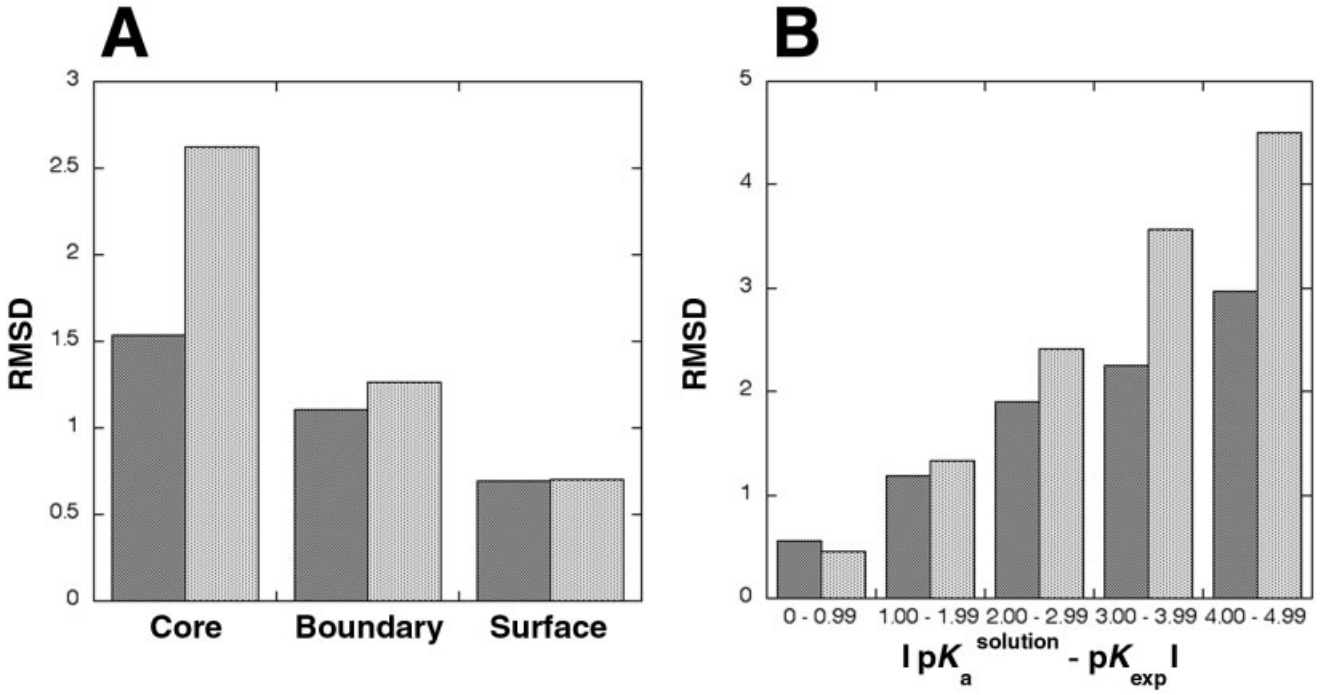

Fig. 5. Comparison of our model (dark bars) and null model (light bars), sorted by core, boundary, and surface regions (A) and the absolute value of $pK_a^{solution} - pK_{exp}$ (B).

electrostatic interactions in the unfolded state, is a likely source of error in our model. The ionizable charges are modeled by subtracting or adding a unit charge to one or two atoms of the side chain (see Methods). A more realistic model requires distribution of the electron density across more side-chain atoms.[26,61,62] The use of static X-ray or NMR structures for all calculations is another shortcoming in our model. Small changes in side-chain conformation affect the magnitude of interactions between groups. This effect is especially strong for the self-energy terms. Incorporation of side-chain flexibility is therefore likely to improve our model, as has been shown by others.[28,49,63–65] Neglecting the possibility of unfolded state structure is another cause of inaccuracy. Though our model assumes that the pK_a values of charged groups in the unfolded state are equivalent to $pK_a^{solution}$, this may not always be the case.[66]

Finally, discrepancies between experimental conditions (pH, ionic strength, temperature) used to measure pK_a values and to determine protein structures will inevitably result in model error. Any type of structural change related to these parameters, including small side-chain movements as well as large conformational changes leading to local or global unfolding,[67] will not be captured by the model. Experimental uncertainty in pK_a determination is usually small (0.1–0.2 pH units) and therefore does not contribute significantly to model error.

When the RMSD error between calculated and experimentally measured pK_a values for all groups and for "large shifter" groups ($|pK_a^{solution} - pK_{exp}| \geq 1.0$) are used as benchmarks, our model compares favorably with existing models based on solution of the Poisson-Boltzmann equation,[21,24] and Coulombic models with sigmoidal[25,26] distance-dependent dielectric constant (Table IV), which in these other studies had been tested on a much smaller set of proteins. However, because the real strength of our model lies in its improved ability to predict the pK_a values of buried charged groups, it must be evaluated using this more appropriate criterion. As shown in Table IV, our model predicts the pK_a values of core and boundary groups (RMSD = 0.81 pH units) better than all other models shown, with the exception of the microenvironment-dependent sigmoidal distance-dependent dielectric model[25] (RMSD = 0.72 pH units). Moreover, the pK_a prediction error of the worst predicted groups ($|pK_{exp} - pK_{calc}| \geq 1.0$) for our model (RMSD = 1.29 pH units) is the lowest of the models shown. The success of our model in these buried regions is probably due to the multiple environment- and region-dependent parameters used to determine the magnitude of the self-energy term, an improvement over the other methods shown, which ignore either charged group local environment[21] or region.[21,24–26]

CONCLUSIONS

Here we present a new continuum electrostatic model that replaces the isotropic dielectric constant with multiple, geometry-dependent dielectric constants that model relaxation effects as several independent cases. The model has been parameterized empirically by a fit to 260 experimentally determined pK_a values. We obtain fits between the calculated and observed values that are significantly better than the null model. The model performs well on the groups that exhibit large pK_a shifts due to the protein environment and compares favorably with other, success-

TABLE IV. Comparison to Other Models

Model[a]	Protein	RMSD[b,c]	RMSDL[b,c]	RMSD core, boundary[b,c]	RMSD deviants[b,c]
Finite difference Poisson-Boltzmann (FDPB) ($\varepsilon_p = 20$)[21]	Lysozyme	0.77 (19)	0.86 (9)		
	RNase A	0.75 (16)	0.19 (3)		
	BPTI	0.47 (10)	—		
	Overall	0.71 (45)	0.74 (12)	0.97 (19)	1.55 (7)
FDPB (Environment-dependent ε_p)[24]	Lysozyme	0.79 (21)	1.01 (9)		
	RNase A	0.61 (16)	0.21 (3)		
	BPTI	0.36 (10)	—		
	Overall	0.65 (47)	0.96 (12)	0.96 (19)	1.49 (7)
Coulombic (sigmoidal $\varepsilon_p(r)$)[26]	Lysozyme	0.88 (18)	1.16 (8)		
	RNase A	0.47 (16)	0.61 (3)		
	BPTI	0.28 (10)	—		
	Overall	0.64 (44)	1.03 (11)	0.90 (16)	1.50 (5)
Coulombic (Environment-dependent sigmoidal $\varepsilon_p(r)$)[25]	Lysozyme	0.70 (21)	0.96 (9)		
	RNase A	0.55 (16)	0.69 (3)		
	BPTI	0.34 (10)	—		
	Overall	0.59 (47)	0.90 (12)	0.72 (19)	1.36 (4)
Our model	Lysozyme	0.77 (21)	1.05 (9)		
	RNase A	0.54 (16)	0.68 (3)		
	BPTI	0.36 (10)	—		
	Overall	0.63 (47)	0.97 (12)	0.81 (19)	1.29 (7)
Null model	Lysozyme	1.32 (21)	1.97 (9)		
	RNase A	0.86 (16)	1.69 (3)		
	BPTI	0.49 (10)	—		
	Overall	1.04 (47)	1.89 (12)	1.39 (19)	1.91 (12)

[a] ε_p denotes a constant protein dielectric; $\varepsilon_p(r)$ denotes a distance-dependent protein dielectric.
[b] The number of residues used to calculate the RMSD, RMSDL (RMSD of "large shifter" residues: $|pK_a^{solution} - pK_{exp}| \geq 1.0$), RMSD core, boundary (RMSD of core and boundary residues), and RMSD deviants (RMSD of the worst predicted groups: $|pK_{calc} - pK_{exp}| \geq 1.0$) are in parentheses next to each value.
[c] Values given may differ from those originally published, because we use the most recent experimental data for lysozyme,[69,70] RNase A,[21] and BPTI.[77]

ful continuum models. The terms in this model involve only a single charge or pairs of charges. Pairwise decomposability is a strict requirement for the dead-end elimination algorithms needed to solve the vast combinatorial problems that arise in computational protein design.[47,68] The model presented here is therefore well suited for incorporation of electrostatic interactions in protein design, an element that has been largely absent in these calculations.

ACKNOWLEDGMENTS

We thank J.J. Dwyer and L.L. Looger for helpful discussions and M.G. Prisant for construction of the Beowulf computer cluster.

REFERENCES

1. Warshel A, Russell ST. Calculations of electrostatic interactions in biological systems and in solutions. Q Rev Biophys 1984;17:283–422.
2. Warshel A, Papazyan A. Electrostatic effects in macromolecules: fundamental concepts and practical modeling. Curr Opin Struct Biol 1998;8:211–217.
3. Nakamura H. Roles of electrostatic interaction in proteins. Q Rev Biophys 1996;29:1–90.
4. Sharp KA, Honig B. Electrostatic interactions in macromolecules—theory and applications. Annu Rev Biophys Biophys Chem 1990;19:301–332.
5. Matthew JB. Electrostatic effects in proteins. Annu Rev Biophys Biophys Chem 1985;14:387–417.
6. Daggett V, Kollman PA, Kuntz ID. Molecular dynamics simulations of small peptides—dependence on dielectric model and pH. Biopolymers 1991;31:285–304.
7. Okamoto Y. Dependence on the dielectric model and pH in a synthetic helical peptide studied by Monte Carlo simulated annealing. Biopolymers 1994;34:529–539.
8. Marshall SA, Morgan CS, Mayo SL. Electrostatics significantly affects the stability of designed homeodomain variants. J Mol Biol 2001;316:189–199.
9. Lee FS, Chu ZT, Warshel A. Microscopic and semimicroscopic calculations of electrostatic energies in proteins by the POLARIS and ENZYMIX programs. J Comput Chem 1993;14:161–185.
10. Sham YY, Chu ZT, Warshel A. Consistent calculations of pK_a of ionizable residues in proteins: semi-microscopic and microscopic approaches. J Phys Chem B 1997;101:4458–4472.
11. Warshel A, Levitt M. Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. J Mol Biol 1976;103:227–249.
12. Garcia-Moreno EB. Estimating binding constants for site-specific interactions between monovalent ions and proteins. Methods Enzymol 1994;240:645–667.
13. Matthew JB, Gurd FRN, Garciamoreno EB, Flanagan MA, March KL, Shire SJ. pH-dependent processes in proteins. CRC Crit Rev Biochem 1985;18:91–197.
14. Sham YY, Muegge I, Warshel A. The effect of protein relaxation

on charge-charge interactions and dielectric constants of proteins. Biophys J 1998;74:1744–1753.
15. Simonson T, Perahia D, Bricogne G. Intramolecular dielectric screening in proteins. J Mol Biol 1991;218:859–886.
16. Nakamura H, Sakamoto T, Wada A. A theoretical study of the dielectric constant of protein. Protein Eng 1988;2:177–183.
17. Warshel A. Molecular biophysics—what about protein polarity. Nature 1987;330:15–16.
18. Warshel A, Aqvist J. Electrostatic energy and macromolecular function. Annu Rev Biophys Biophys Chem 1991;20:267–298.
19. Hassan SA, Guarnieri F, Mehler EL. Characterization of hydrogen bonding in a continuum solvent model. J Phys Chem B 2000; 104:6490–6498.
20. Schutz CN, Warshel A. What are the dielectric "constants" of proteins and how to validate electrostatic models? Proteins 2001;44:400–417.
21. Antosiewicz J, McCammon JA, Gilson MK. Prediction of pH-dependent properties of proteins. J Mol Biol 1994;238:415–436.
22. Bashford D, Case DA, Dalvit C, Tennant L, Wright PE. Electrostatic calculations of side-chain pK_a values in myoglobin and comparison with NMR data for histidines. Biochemistry 1993;32: 8045–8056.
23. Honig B, Sharp K, Sampogna R, Gunner MR, Yang AS. On the calculation of pK_a values in proteins. Proteins 1993;15:252–265.
24. Demchuk E, Wade RC. Improving the continuum dielectric approach to calculating pK_a's of ionizable groups in proteins. J Phys Chem 1996;100:17373–17387.
25. Mehler EL, Guarnieri F. A self-consistent, microenvironment modulated screened Coulomb potential approximation to calculate pH-dependent electrostatic effects in proteins. Biophys J 1999;77:3–22.
26. Mehler EL. Self-consistent, free energy based approximation to calculate pH dependent electrostatic effects in proteins. J Phys Chem 1996;100:16006–16018.
27. Sandberg L, Edholm O. A fast and simple method to calculate protonation states in proteins. Proteins 1999;36:474–483.
28. Rabenstein B, Knapp EW. Calculated pH-dependent population and protonation of carbon monoxy myoglobin conformers. Biophys J 2001;80:1141–1150.
29. Onufriev A, Bashford D, Case DA. Modification of the generalized Born model suitable for macromolecules. J Phys Chem B 2000;104:3712–3720.
30. Still CS, Tempczyrk A, Hawley RC, Hendrickson T. Semianalytical treatment of salvation for molecular mechanics and dynamics. J Am Chem Soc 1990;112:6127–6129.
31. Edinger SR, Cortis C, Friesner RA. Solvation free energies of peptides: comparison of approximate continuum salvation models with accurate solution of the Poisson-Boltzmann equation. J Phys Chem 1997;101:1190–1197.
32. Voges D, Karshikoff A. A model of a local dielectric constant in proteins. J Chem Phys 1998;108:2219–2227.
33. Tanford C, Kirkwood JG. Theory of protein titration curves. I. General equations for impenetrable spheres. J Am Chem Soc 1957;79:5333–5339.
34. Tanford C, Roxby R. Interpretation of protein titration curves. Application to lysozyme. Biochemistry 1972;11:2192–2198.
35. Warshel A, Russell ST, Churg AK. Macroscopic models for studies of electrostatic interactions in proteins—limitations and applicability. Proc Natl Acad Sci USA 1984;81:4785–4789.
36. Rocchia W, Alexov E, Honig B. Extending the applicability of the nonlinear Poisson-Boltzmann equation: multiple dielectric constants and multivalent ions. J Phys Chem B 2001;105:6507–6514.
37. Garcia-Moreno EB, Dwyer JJ, Gittis AG, Lattman EE, Spencer DS, Stites WE. Experimental measurement of the effective dielectric in the hydrophobic core of a protein. Biophys Chem 1997;64:211–224.
38. Bashford D, Karplus M. Multiple-site titration curves of proteins—an analysis of exact and approximate methods for their calculation. J Phys Chem 1991;95:9556–9561.
39. Beroza P, Fredkin DR, Okamura MY, Feher G. Protonation of interacting residues in a protein by a Monte Carlo method: application to lysozyme and the photosynthetic reaction center of *Rhodobacter sphaeroides.* Proc Natl Acad Sci USA 1991;88:5804–5808.
40. Chivers PT, Prehoda KE, Volkman BF, Kim BM, Markley JL, Raines RT. Microscopic pK_a values of *Escherichia coli* thioredoxin. Biochemistry 1997;36:14985–14991.
41. Word JM, Lovell SC, Richardson JS, Richardson DC. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. J Mol Biol 1999;285:1735–1747.
42. Nozaki Y, Tanford C. Examination of titration behavior. Methods Enzymol 1967;11:715–734.
43. Tanokura M. H1-NMR study on the tautomerism of the imidazole ring of histidine residues. 1. Microscopic pK values and molar ratios of tautomers in histidine containing peptides. Biochim Biophys Acta 1983;742:576–585.
44. Sitkoff D, Sharp KA, Honig B. Accurate calculation of hydration free-energies using macroscopic solvent models. J Phys Chem 1994;98:1978–1988.
45. MacKerell AD, Bashford D, Bellott M, Dunbrack RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher WE, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, WiorkiewiczKuczera J, Yin D, Karplus M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B 1998;102:3586–3616.
46. Lee B, Richards FM. The interpretation of protein structures: estimation of static accessibility. J Mol Biol 1971;55:379–400.
47. Dahiyat BI, Mayo SL. *De novo* protein design: fully automated sequence selection. Science 1997;278:82–87.
48. Shrake A, Rupley JA. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. J Mol Biol 1973;79:351–371.
49. Bashford D, Gerwert K. Electrostatic calculations of the pK_a values of ionizable groups in bacteriorhodopsin. J Mol Biol 1992;224:473–486.
50. Gilson MK, Rashin A, Fine R, Honig B. On the calculation of electrostatic interactions in proteins. J Mol Biol 1985;184:503–516.
51. Van Belle D, Couplet I, Prevost M, Wodak SJ. Calculations of electrostatic properties in proteins—analysis of contributions from induced protein dipoles. J Mol Biol 1987;198:721–735.
52. Barlow DJ, Thornton JM. Ion pairs in proteins. J Mol Biol 1983;168:867–885.
53. Antosiewicz J, McCammon JA, Gilson MK. The determinants of pK_a's in proteins. Biochemistry 1996;35:7819–7833.
54. Fersht AR. Conformational equilibria in ct- and 8-chymotrypsin. The energetics and importance of the salt bridge. J Mol Biol 1972;64:497–509.
55. Anderson DE, Becktel WJ, Dahlquist FW. pH-Induced denaturation of proteins—a single salt bridge contributes 3–5 kcal/mol to the free energy of folding of T4 lysozyme. Biochemistry 1990;29: 2403–2408.
56. Horovitz A, Serrano L, Avron B, Bycroft B, Fersht AR. Strength and cooperativity of contributions of surface salt bridges to protein stability. J Mol Biol 1990;216:1031–1044.
57. Erwin CR, Barnett BL, Oliver JD, Sullivan JF. Effects of salt bridges on the stability of subtilisin BPN′. Protein Eng 1990;4:87–97.
58. Dao-pin S, Sauer U, Nicholson H, Matthews BW. Contributions of engineered salt bridges to the stability of T4 lysozyme determined by directed mutagenesis. Biochemistry 1991;30:7142–7153.
59. Sali D, Bycroft M, Fersht AR. Surface electrostatic interactions contribute little to stability of barnase. J Mol Biol 1991;220:779–788.
60. Myers JK, Pace CN. Hydrogen bonding stabilizes globular proteins. Biophys J 1996;71:2033–2039.
61. Antosiewicz J, Briggs JM, Elcock AH, Gilson MK, McCammon JA. Computing ionization states of proteins with a detailed charge model. J Comput Chem 1996;17:1633–1644.
62. Yang AS, Gunner MR, Sampogna R, Sharp K, Honig B. On the calculation of pK_a*'s* in Proteins. Proteins 1993;15:252–265.
63. You TJ, Bashford D. Conformation and hydrogen ion titration of proteins: a continuum electrostatic model with conformational flexibility. Biophys J 1995;69:1721–1733.
64. Beroza P, Case DA. Including side chain flexibility in continuum electrostatic calculations of protein titration. J Phys Chem 1996;100:20156–20163.
65. Alexov EG, Gunner MR. Incorporating protein conformational

flexibility into the calculation of pH-dependent protein properties. Biophys J 1997;72:2075–2093.
66. Oliveberg M, Arcus VL, Fersht AR. pK_a values of carboxyl groups in the native and denatured states of barnase—the pK_a values of the denatured state are on average 0.4 units lower than those of model compounds. Biochemistry 1995;34:9424–9433.
67. Imoto T, Johnson LN, North ACT, Phillips DC, Rupley JA. In: Boyers PD, editor. The Enzymes. New York: Academic; 1972. p 665–868.
68. Looger LL, Hellinga HW. Generalized dead-end elimination algorithms make large-scale protein side-chain structure prediction tractable: implications for protein design and structural genomics. J Mol Biol 2001;307:429–445.
69. Bartik K, Redfield C, Dobson CM. Measurement of the individual pK_a values of acidic residues of hen and turkey lysozymes by 2-dimensional H1-NMR. Biophys J 1994;66:1180–1184.
70. Kuramitsu S, Hamaguchi K. Analysis of the acid-base titration curve of hen lysozyme. J Biochem 1980;87:1215–1219.
71. Oda Y, Yamazaki T, Nagayama K, Kanaya S, Kuroda Y, Nakamura H. Individual ionization constants of all the carboxyl groups in ribonuclease HI from *Escherichia coli* determined by NMR. Biochemistry 1994;33:5275–5284.
72. Oda Y, Yoshida M, Kanaya S. Role of histidine 124 in the catalytic function of ribonuclease HI from *Escherichia coli.* J Biol Chem 1993;268:88–92.
73. Perez-Canadillas JM, Campos-Olivas R, Lacadena J, del Pozo AM, Gavilanes JG, Santoro J, Rico M, Bruix M. Characterization of pK_a values and titration shifts in the cytotoxic ribonuclease alpha-sarcin by NMR. Relationship between electrostatic interactions, structure, and catalytic function. Biochemistry 1998;37: 15865–15876.
74. Khare D, Alexander P, Antosiewicz J, Bryan P, Gilson M, Orban J. pK_a measurements from nuclear magnetic resonance for the B1 and B2 immunoglobulin G-binding domains of protein G: comparison with calculated values for nuclear magnetic resonance and X-ray structures. Biochemistry 1997;36:3580–3589.
75. Russu IM, Ho NT, Ho C. A proton nuclear magnetic resonance investigation of histidyl residues in human normal adult hemoglobin. Biochemistry 1982;21:5031–5043.
76. Garner MH, Bogardt RA, Gurd FRN. Determination of the pK values for the α-amino groups of human hemoglobin. J Biol Chem 1974;250:4398–4404.
77. March KL, Maskalick DG, England RD, Friend SH, Gurd FRN. Analysis of electrostatic interactions and their relationship to conformation and stability of bovine pancreatic trypsin inhibitor. Biochemistry 1982;21:5241–5251.
78. Marti DN, Jelesarov I, Bosshard HR. Interhelical ion pairing in coiled coils: solution structure of a heterodimeric leucine zipper and determination of pK_a values of Glu side chains. Biochemistry 2000;39:12804–12818.
79. Kesvatera T, Jonsson B, Thulin E, Linse S. Measurement and modelling of sequence-specific pK_a values of lysine residues in calbindin D_{9k}. J Mol Biol 1996;259:828–839.
80. Joshi MD, Hedberg A, McIntosh LP. Complete measurement of the pK_a values of the carboxyl and imidazole groups in *Bacillus circulars* xylanase. Protein Sci 1997;6:2667–2670.
81. Joshi MD, Sidhu G, Nielsen JE, Brayer GD, Withers SG, McIntosh LP. Dissecting the electrostatic interactions and pH-dependent activity of a family 11 glycosidase. Biochemistry 2001;40:10115–10139.
82. Kohda D, Sawada T, Inagaki F. Characterization of pH titration shifts for all the nonlabile proton resonances in a protein by 2-dimensional NMR—the case of mouse epidermal growth factor. Biochemistry 1991;30:4896–4900.
83. Forman-Kay JD, Clore GM, Gronenborn AM. Relationship between electrostatics and redox function in human thioredoxin—characterization of pH titration shifts using 2-dimensional homonuclear and heteronuclear NMR. Biochemistry 1992;31:3442–3452.
84. Giletto A, Pace CN. Buried, charged, non-ion-paired aspartic acid 76 contributes favorably to the conformational stability of ribonuclease T1. Biochemistry 1999;38:13379–13384.
85. Inagaki F, Kawano Y, Shimada I, Takahashi K, Miyazawa T. Nuclear magnetic resonance study on the micro-environments of histidine residues of ribonuclease T1 and carboxymethylated ribonuclease T1. J Biochem 198 1;89:1185–1195.
86. Shaw RW, Hartzell CR. Hydrogen ion titration of horse heart ferricytochrome c. Biochemistry 1976;15:1909–1914.
87. Kuhlman B, Luisi DL, Young P, Raleigh DP. pK_a values and the pH dependent stability of the N-terminal domain of L9 as probes of electrostatic interactions in the denatured state. Differentiation between local and nonlocal interactions. Biochemistry 1999; 38:4896–4903.
88. Sorensen MD, Led JJ. Structural details of Asp(B9) human insulin at low pH from 2-dimensional NMR titration studies. Biochemistry 1994;33:13727–13733.
89. Swint-Kruse L, Robertson AD. Hydrogen bonds and the pH-dependence of ovomucoid 3rd domain stability. Biochemistry 1995;34:4724–4732.
80. Schaller W, Robertson AD. pH, ionic strength, and temperature dependences of ionization equilibria for the carboxyl groups in turkey ovomucoid 3′ domain. Biochemistry 1995;34:4714–4723.
81. Norton RS, Cross K, Braachmaksvytis V, Wachter E. HL-NMR study of the solution properties and secondary structure of neurotoxin III from the sea anemone *Anemonia sulcata.* Biochem J 1993;293:545–551.
92. Alexandrescu AT, Mills DA, Ulrich EL, Chinami M, Markley JL. NMR assignments of the 4 histidines of staphylococcal nuclease in native and denatured states. Biochemistry 1988;27:2158–2165.
93. Dwyer JJ, Gittis AG, Karp DA, Lattman EE, Spencer DS, Stites WE, Garcia-Moreno EB. High apparent dielectric constants in the interior of a protein reflect water penetration. Biophys J 2000;79:1610–1620.
94. Ebina S, Wuthrich K. Amide proton titration shifts in bull seminal inhibitor IIa by two-dimensional correlated H1 nuclear magnetic resonance (COSY)—manifestation of conformational equilibria involving carboxylate groups. J Mol Biol 1984;179:283–288.
95. Koide A, Jordan MR, Horner SR, Batori V, Koide S. Stabilization of a fibronectin type III domain by the removal of unfavorable electrostatic interactions on the protein surface. Biochemistry 2001;40:10326–10333.
96. Cederholm MT, Stuckey JA, Doscher MS, Lee L. Histidine pK_a shifts accompanying the inactivating Asp 121-Asn substitution in a semisynthetic bovine pancreatic ribonuclease. Proc Natl Acad Sci USA 1991;88:8116–8120.
97. Zhang P, Graminski GF, Armstrong RN. Are the histidine-residues of glutathione-S-transferase important in catalysis—an assessment by C13 NMR spectroscopy and site-specific mutagenesis. J Biol Chem 1991;266:19475–19479.
98. Tishmack PA, Bashford D, Harms E, VanEtten RL. Use of H1 NMR spectroscopy and computer simulations to analyze histidine pK_a changes in a protein tyrosine phosphatase: Experimental and theoretical determination of electrostatic properties in a small protein. Biochemistry 1997;36:11984–11994.
99. Pjura P, McIntosh LP, Wozniak JA, Matthews BW. Perturbation of Trp-138 in T4 lysozyme by mutations at Gln-105 used to correlate changes in structure, stability, solvation, and spectroscopic properties. Proteins 1993;15:401–412.
100. Dao-pin S, Anderson DE, Baase WA, Dahlquist FWI, Matthews BW. Structural and thermodynamic consequences of burying a charged residue within the hydrophobic core of T4 lysozyme. Biochemistry 1991;30:11521–11529.
101. Fujii S, Akasaka K, Hatano H. Acid denaturation steps of Streptomyces subtilisin inhibitor—a proton magnetic resonance study of individual histidine environment. J Biochem 1980;88: 789–796.
102. Lavigne P, Kondejewski LH, Houston ME, Sonnichsen FD, Lix B, Sykes BD, Hodges RS, Kay CM. Preferential heterodimeric parallel coiled-coil formation by synthetic max and c-myc leucine zippers—a description of putative electrostatic interactions responsible for the specificity of heterodimerization. J Mol Biol 1995;254:505–520.
103. Inagaki F, Miyazawa T, Hori H, Tamiya N. Conformation of erabutoxin a and erabutoxin b in aqueous solution as studied by nuclear magnetic resonance and circular dichroism. Eur J Biochem 1978;89:433–442.
104. Russell AJ, Thomas PG, Fersht AR. Electrostatic effects on modification of charged groups in the active site cleft of subtilisin by protein engineering. J Mol Biol 1987;193:803–813.