Protein Science (1994), *3*:1556–1569. Cambridge University Press. Printed in the USA.

Optimization of the electrostatic interactions in proteins of different functional and folding type

VELIN Z. SPASSOV,[1,2] ANDREJ D. KARSHIKOFF,[1] AND RUDOLF LADENSTEIN[1]

[1] Centre for Structural Biochemistry, Karolinska Institute, NOVUM, S-14157 Huddinge, Stockholm, Sweden
[2] Central Laboratory of Biophysics, Bulgarian Academy of Sciences, 1113 Sofia, Bulgaria

(RECEIVED March 11, 1994; ACCEPTED June 21, 1994)

Abstract

The 3-dimensional optimization of the electrostatic interactions between the charged amino acid residues was studied by Monte Carlo simulations on an extended representative set of 141 protein structures with known atomic coordinates. The proteins were classified by different functional and structural criteria, and the optimization of the electrostatic interactions was analyzed. The optimization parameters were obtained by comparison of the contribution of charge–charge interactions to the free energy of the native protein structures and for a large number of randomly distributed charge constellations obtained by the Monte Carlo technique. On the basis of the results obtained, one can conclude that the charge–charge interactions are better optimized in the enzymes than in the proteins without enzymatic functions. Proteins that belong to the mixed $\alpha\beta$ folding type are electrostatically better optimized than pure α-helical or β-strand structures. Proteins that are stabilized by disulfide bonds show a lower degree of electrostatic optimization. The electrostatic interactions in a native protein are effectively optimized by rejection of the conformers that lead to repulsive charge–charge interactions. Particularly, the rejection of the repulsive contacts seems to be a major goal in the protein folding process. The dependence of the optimization parameters on the choice of the potential function was tested. The majority of the potential functions gave practically identical results.

Keywords: energy calculations; ion pairs; Monte Carlo simulations; potential functions; protein electrostatics; protein folding

The nature and spatial distribution of charged residues in a folded protein can be considered as an evolutionary solution of 2 different tasks: first, the stabilization of the native structure by the contribution of the electrostatic interactions to the free energy and, second, the display of a functional role by creating a specific electrostatic field, necessary for the enhancement of the enzymatic reactions, intermolecular recognition, and assembly. In principle, the solution of the second task could be opposite to the stabilization effect. The magnitude of the stabilization effect of the other forces governing protein folding could counteract the necessity of significant electrostatic interactions. A variety of experimental and theoretical observations provide evidence that the interactions between the charged groups contribute to protein stability: the pH dependence of protein stability (Perutz, 1978), the energies estimated for ion pair contacts (Fersht, 1972; Perutz & Raidt, 1975; Anderson et al., 1990; Meiering et al., 1992), and the number of salt bridges observed in proteins with known X-ray structure (Barlow & Thornton, 1983). Concerning charge–charge interactions, however, the viewpoint has changed considerably over the years, from the assignment of a dominant contribution to a nondominant force in the folding process (Dill, 1990). There are strong arguments (Dill, 1990; Ponnuswamy, 1993) that the hydrophobic interactions represent the major folding force, but there is also clear experimental evidence (Anderson et al., 1990) that removing of only 1 salt bridge can significantly influence the stability of a protein molecule. Therefore, the question of the importance of electrostatic interactions for protein stability and folding, in our opinion, is not generally solved.

One of the possible ways to estimate the structural significance of the electrostatic interactions is to study the 3-dimensional distribution of the charged groups in folded proteins with known atomic coordinates. Recently, a model approach has been proposed (Spassov & Atanasov, 1994) to estimate and compare on a common scale the degree of spatial optimization of the interaction between ionized groups in proteins with different structure and properties. This approach was based on a comparison of the electrostatic term of the free energy calculated for the charge constellation of a native protein with the corresponding energies calculated for a large set of charge distributions generated by a Monte Carlo technique. The probability of the oc-

Reprint requests to: Andrej D. Karshikoff, Centre for Structural Biochemistry, Karolinska Institute, NOVUM, S-14157 Huddinge, Stockholm, Sweden; e-mail: aka@csb.ki.se.

currence of random charge constellations with an energy lower than the energy of the native structure has been suggested as a criterion for spatial optimization of the electrostatic interactions. The results of the calculations show clearly that the electrostatic interactions in the native structures are better optimized than in the average random charge constellations for almost all the 44 tested proteins. Significant differences were obtained for different proteins, from a very low degree of optimization (near to the expected value for the random charge distribution) to structures characterized by a very high optimization. These theoretical observations indirectly support the viewpoint that the role of interactions between the ionized groups in the folding process may be strongly individual for the different types of protein structures.

In the present study, we continue the analysis of the problems discussed above on the basis of an extended set of 141 nonhomologous structures—very near to the maximum number of the nonequivalent high-resolution X-ray structures available in a recent collection (April 1993) of the Protein Data Bank (PDB) (Bernstein et al., 1977). The protein structures selected were divided into representative groups using the following criteria: (1) functional type (grouped as the enzymes and proteins without enzymatic function); (2) secondary-structure folding type (all-α, α/β, and all-β structures); (3) type of covalent structure (proteins with or without disulfide bonds). The latter criterion has been applied previously to a set of 44 protein structures (Spassov & Atanasov, 1994). Here we analyze the frequency of occurrence of proteins with different ion pair and disulfide bridge saturation. The results confirm the observation that the charge interactions in the native protein structures are characterized by an effective rejection of the repulsive contacts.

In principle, the results may depend on the electrostatic model used and its parametrization. Here we tested this possibility, repeating the calculations with several potential functions that describe the charge–charge interactions. It was found that, apart from the Coulomb potential, the results are essentially independent of the type of potential function.

Results and discussion

The entry list (RS1) is represented in Table 1 in descending order with respect to the S_{opt} values, i.e., from high to low degree of optimization. The optimization parameters S_{opt} were calculated by Equation 8 (Methods section) as a function of the energy terms for the native structure, $\Delta G_{ei,ntv}$ (Equations 3, 3′, 4), and the mean energies, $\langle \Delta G_{ei,rnd} \rangle$, of the generated 1,000 random charge distributions for each entry (Table 1). The results represent only the electrostatic interactions between the charged amino acid residues in the given crystallographic structure: all structures are taken in their "apo" form; the charges of metal ions, substrates, inhibitors, etc., are not included in the calculations. The binding of charged ligands is governed not only by the electrostatic interactions. Factors such as hydrophobic effect, proper coordination of the metal ions, etc., play a determining role here. Because these factors are a result of an already folded structure with specific clefts, hydrophobic patches, etc., ligand charges cannot participate in the randomization as performed in this study. It is clear that the hypothetical "apo" structures may differ from the real ones. In the cases where this difference influences the electrostatic interaction between ionizable groups, or is mainly due to the strong electrostatic interactions between the ligand charge and the protein charge constellation, the entries were excluded from the data set (see Methods). For the highly optimized structures (top regions of Table 1), the low S_{opt} values are an indirect indication of the minor influence of the ligand-binding effects or intermolecular interactions on integral electrostatic properties of the proteins.

In almost all structures selected in RS1, the S_{opt} values show a negative sign, i.e., the energy terms calculated for the native structures, $\Delta G_{ei,ntv}$, have lower values than the averaged energies of the random charge distributions (see Table 1). This regularity is not trivial. The establishment of the native 3-dimensional structure in the folding process is a result of the simultaneous work of different factors and, a priori, it is not absolutely necessary that the electrostatic term is optimized. The S_{opt} values can be considered as a measure of the importance of the electrostatic interactions in the folding process of each individual protein structure. More clearly, S_{opt} serves as a relative measure of the gain in energy that results from the minimization of the electrostatic term alone in reaching a low energy state that, in general, characterizes the native structure. It has to be noted that S_{opt} is not directly related to the electrostatic stabilization of the proteins. Thus, for example, some small proteins (Table 1, entries 1FXB, 2MLT, 1RDG, 5RXN) with $S_{opt} < 0$ are characterized by a positive value of $\Delta G_{ei,ntv}$, i.e., the electrostatic interactions between titratable groups destabilize the native structure.

Correlation between the electrostatic optimization and the presence of disulfide bridges

A certain tendency of protein structures to compensate for a decrease in the energetic optimization of the electrostatic interactions by the appearance of structure-stabilizing factors, i.e., disulfide bridges, has been observed by Spassov and Atanasov (1994). The analysis of this important feature of protein structures is extended in the present work. The number of disulfide bridges, shown in Table 1, was obtained from the header records in the PDB files, as well as using a distance criterion. The cross-chain connections Fe–Cys 4 in rubredoxin were taken as disulfide bridges as well. For each of the representative sets (RS1 and RS2), the frequencies of occurrence of disulfide bond-containing proteins are calculated in S_{opt} intervals (RS1, Fig. 1A; RS2, Fig. 1B). It is seen that the distributions obtained for proteins without disulfide bridges have S_{opt} values shifted to more negative values (i.e., good electrostatic optimization) and the distributions of the proteins stabilized by disulfide bridges to less negative values. It is immediately obvious that the low level of spatial optimization of the electrostatic interactions is compensated by the statistic appearance of covalent crosslinks in the protein structures and vice versa. In terms of protein stability, this means that the insufficiency of the electrostatic stabilization of the native structure is compensated by introduction of chemical crosslinks, such as disulfide bridges. Thus, for example, 3 entries at the bottom of Table 1 are characterized by $S_{opt} > 0$. One of them, 9WGA, has 16 disulfide bridges, whereas the other 2 proteins are cytochromes, where 2 cysteine residues are covalently bound to the heme. The strong noncovalent interactions of the prosthetic group with the protein moiety, as well as the 2 thioether bridges, appear to compensate for the low electrostatic optimization. Another clear tendency is that smaller proteins (<100 residues) are characterized by a lower electrostatic optimization (see Table 1; $S_{opt} > -2$). This is in accord

Table 1. *Selected protein structures (set RS1) listed in descending order with respect to* S_{opt}[a]

Code	Protein (source)	S_{opt}	ΔG_{ei}	$\langle \Delta G_{ei} \rangle$[b]		N	SSF
6ACN *r*	Aconitase (pig)	−6.23	−184.0	−2.5	754	0	E
2CPP *r*	Cytochrome P450 Cam (*Pseudomonas putida*)	−4.32	−80.8	−6.6	414	0	E
8CAT *r*	Catalase (beef)	−4.09	−81.6	−4.6	506	0	E
*8ADH *r*	Alcohol dehydrogenase (horse)	−3.88	−61.5	−4.1	374	0	E
*RUBA	Rubisco (*Rhodospirillum rubrum*)	−3.83	−75.1	−6.7	466	0	E
3GRS *r*	Glutathione reductase (human)	−3.72	−60.2	−4.0	478	1	E
1GOX *r*	Glycolate oxidase (spinach)	−3.58	−59.3	−5.1	369	0	E
1WSY *r*	Trypt. synthase (*Salmonella typhimurium*), β-chain	−3.28	−58.6	−6.8	397	0	E
*1CTS *r*	Citrate synthase (pig)	−3.15	−48.1	−3.9	437	0	E
*1GPD	D-Gyceraldehyde-3-P dehydrogenase	−3.13	−44.5	−5.0	334	0	E
*1MBD *r*	Myoglobin (sperm whale)	−3.06	−26.9	3.6	153	0	N
9PAP *r*	Papain (papaya)	−3.05	−27.2	2.1	212	3	E
1SGT *r*	Trypsin (*Streptomyces griseus*)	−3.03	−30.3	−3.7	223	3	E
2I1B *r*	Interleukin-1 beta (human)	−2.99	−30.3	−3.4	153	0	N
2HLA *r*	Histocompatibility antigen, AW 68.1, α-chain	−2.99	−43.2	−5.2	270	2	N
2ALP	Alpha-lytic protease (*Lysobacter enz.*)	−2.98	−15.2	4.6	198	3	E
2FB4 *r*	Igg1 Fab (human) L-chain	−2.92	−22.7	−2.6	216	2	N
*2HHB *r*	Hemoglobin (human) α-chain	−2.90	−16.9	0.2	141	0	N
1WSY	Trypt. synthase (*S. typhimurium*), α-chain	−2.90	−30.4	−4.0	268	0	E
2GBP *r*	D-Galactose binding protein (*Escherichia coli*)	−2.88	−51.3	−4.8	309	0	N
GTRA	Glutathione transferase (pig lung)	−2.88	−35.4	−2.9	198	0	E
4XIA *r*	D-Xylose isomerase (*Arthrobacter*)	−2.88	−46.9	2.7	393	0	E
2FB4	Igg1 Fab (human) H-chain	−2.87	−21.2	−1.7	229	3	N
2GD1 *r*	D-Glyceraldehyde dehydrogenase (*Bacillus stearothermophilus*)	−2.86	−48.4	−5.7	334	0	E
1PHH *r*	*P*-Hydroxybenzoate hydroxylase (*P. fluor.*)	−2.82	−56.1	−6.1	394	0	E
*1GCR *r*	γ-II Crystallin (calf)	−2.79	−33.5	−3.5	174	0	N
*3LZM *r*	Lysozyme T4 (*E. coli*)	−2.79	−27.9	0.2	164	0	N
*5CPA *r*	Carboxypeptidase A alpha (bovine)	−2.76	−36.6	−3.9	307	1	E
1TIM *r*	Triose phosphate isomerase (chicken)	−2.73	−37.0	−3.3	247	0	E
*1PPD	Papain D (papaya)	−2.71	−24.7	2.2	212	3	E
*1SBC *r*	Subtilisin Carlsberg (*Bacillus subtilis*)	−2.70	−23.5	−2.8	275	0	E
*2PRK	Proteinase K (fungus)	−2.64	−32.5	−3.6	279	2	E
2CI2 *r*	Chymotrypsin inhibitor CI-2 (barley)	−2.61	−20.5	−2.9	83	0	N
3RP2	Rat mast cell protease II (rat)	−2.59	−24.4	0.0	224	3	E
*2AZA *r*	Azurin (*Alcaligenes denitrificans*)	−2.59	−27.7	−3.9	129	1	N
4MDH	Cytoplasmic malate dehydrogenase (porcine)	−2.57	−41.1	−5.5	334	0	E
*1RHD *r*	Rhodanese (bovine)	−2.55	−36.6	−3.9	293	0	E
1HOE *r*	α-Amylase inhibitor Hoe-467A	−2.54	−14.1	−1.4	74	2	N
4FD1 *r*	Ferredoxin (*Azotobacter vinelandii*)	−2.51	−3.6	17.2	106	0	N
1PMB	Myoglobin (porcine)	−2.50	−24.5	−1.3	153	0	N
2STV	Satellite tobacco necrosis virus	−2.49	−17.9	−0.5	195	0	N
1GP1 *r*	Glutathione peroxidase (bovine)	−2.48	−31.5	−3.6	198	0	E
1TON	Tonin (rat)	−2.44	−26.0	−3.2	235	5	E
*1ABP	L-Arabinose-binding protein (*E. coli*)	−2.42	−39.5	−4.4	306	0	N
*1CA2	Carbonic anhydrase form C (human)	−2.41	−34.1	−2.9	259	0	E
*2SNS *r*	Staphylococcal nuclease	−2.40	−27.4	0.4	149	0	E
2MHR *r*	Myohemerythrin (sipunculan worm)	−2.39	−28.8	−3.8	118	0	N
3DFR	Dihydrofolate reductase (*Lactobacillus casei*)	−2.37	−20.5	−3.5	162	0	E
256B *r*	Cytochrome b562 (*E. coli*)	−2.35	−22.1	−3.9	106	0	N
1CLA *r*	Chloramphenicol acetyltransferase (*E. coli*)	−2.34	−18.9	−1.8	213	0	E
1REI	Bence-Jones immunoglobulin (human)	−2.33	−10.3	−1.4	107	1	N
2CDV *r*	Cytochrome c_3 (*Desulfovibrio vulgaris*)	−2.33	−12.5	4.1	107	0	N
3PGM *r*	Phosphoglycerate mutase (yeast)	−2.28	−29.0	−1.0	241	0	E
*3GAP *r*	Catabolite gene activator protein (*E. coli*)	−2.28	−21.2	−1.6	209	0	N
*2LIV	Leu-Ile-Val-binding protein (*E. coli*)	−2.25	−32.2	−1.1	344	1	N
2CAB *r*	Carbonic anhydrase form B (human)	−2.22	−26.8	−2.3	261	0	E
*6LDH *r*	Lactate dehydrogenase (dogfish)	−2.22	−30.9	−1.8	330	0	E
2PAZ *r*	Pseudoazurin (*Alcaligenes faecalis*)	−2.15	−17.1	−2.9	123	0	N
4MBA	Myoglobin (sea hare)	−2.09	−15.8	−2.4	147	0	N
*1UTG *r*	Uteroglobin oxidase (rabbit)	−2.09	−11.9	−2.4	70	0	N
3BLM	β-Lactamase (*Staphylococcus aureus*)	−2.08	−31.9	1.5	257	0	E

(*continued*)

Table 1. *Continued*

Code	Protein (source)	S_{opt}	ΔG_{ei}	$\langle \Delta G_{ei} \rangle^b$		N	SSF
8ATC	Aspartate carbamoyltransferase (*E. coli*), β-chain	−2.05	−17.1	−2.8	153	0	E
*2ACT	Actinidin (Chinese gooseberry)	−2.03	−11.2	9.9	220	3	E
3B5C *r*	Cytochrome b_5 (bovine)	−2.02	−14.7	−1.6	93	0	N
2HHB	Hemoglobin (human), β-chain	−2.02	−12.8	−1.0	146	0	N
1FXB *r*	Ferredoxin (*Bacillus thermoproteolyticus*)	−1.97	4.4	15.8	81	0	N
*1CRN	Crambin (Abyssinian cabbage)	−1.94	−6.2	−1.5	46	3	N
*8DFR *r*	Dihydrofolate reductase (chicken)	−1.91	−20.6	−0.8	189	0	E
2LBP *r*	Leucine-binding protein (*E. coli*)	−1.90	−25.1	1.4	346	1	N
3HVP *r*	HIV-1 protease (synthetic)	−1.90	−9.4	−1.8	99	0	E
1TNF *r*	Tumor necrosis factor (human)	−1.89	−13.3	−2.5	157	1	N
4INS *r*	Insulin (pig)	−1.86	−5.6	−1.2	21	3	N
8ATC *r*	Aspartate carbamoyltransferase (*E. coli*), α-chain	−1.86	−32.0	−5.6	310	0	E
2PKA	Kallikrein A (porcine)	−1.85	−14.0	2.8	232	5	E
1HMQ	Hemerythrin (*Themiste dyscritum*)	−1.82	−20.7	−4.0	113	0	N
2CCY *r*	Cytochrome *c'* (*Rhodomonas molichianum*)	−1.80	−13.4	−2.4	128	0	N
2PAB *r*	Prealbumin (human)	−1.75	−13.4	−2.9	127	0	N
*3ADK *r*	Adenylate kinase (porcine)	−1.75	−22.8	−3.7	195	0	E
*3EST	Elastase (porcine)	−1.74	−10.8	0.4	240	4	E
2AAT *r*	Aspartate aminotransferase (*E. coli*)	−1.71	−28.7	−6.2	396	0	E
*2SGA *r*	Proteinase A (*Streptomyces griseus*)	−1.69	−9.6	−1.7	181	2	E
2CRO	434 Cro protein (phage 434)	−1.68	−5.4	2.9	71	0	N
2UTG	Uteroglobin	−1.68	−22.7	−4.5	140	2	N
1CYC	Ferrocytochrome *c* (bonito)	−1.65	−3.3	6.4	103	0	N
2RHE	Bence-Jones protein (human)	−1.65	−8.5	0.5	114	1	N
1LZ1 *r*	Lysozyme (human)	−1.63	−9.3	4.1	130	4	E
2LH4 *r*	Leghemoglobin (lupin)	−1.63	−15.7	−3.6	153	0	N
*5CYT	Cytochrome *c* (albacore tuna)	−1.59	−1.9	7.7	104	0	N
*1UBQ *r*	Ubiquitin (human)	−1.59	−11.8	−2.2	76	0	N
1ECA *r*	Erythrocruorin (*Chironomus thummi thummi*)	−1.56	−12.5	−2.5	136	0	N
2MLT	Melittin (honeybee)	−1.53	1.4	3.4	27	0	N
1PP2	Phospholipase A2 (rattlesnake)	−1.52	−10.4	1.5	122	7	E
*1NXB *r*	Neurotoxin B (sea snake)	−1.52	−9.1	−1.3	62	4	N
1CTF *r*	50S Ribosomal protein (*E. coli*)	−1.48	−12.8	−2.7	74	0	N
3FXC *r*	Ferredoxin (*Spirulina platensis*)	−1.46	12.5	20.3	98	0	N
RVSA	Riboflavin synthetase	−1.45	−10.3	−1.4	153	0	E
1HNE	Neutrophil elastase (human)	−1.42	−6.4	3.6	218	4	E
*4FXN *r*	Flavodoxin (*Clostridium* MP)	−1.42	−3.0	10.3	138	0	N
2RSP *r*	Rous sarcoma virus protease	−1.40	−11.7	−2.8	124	0	E
1LDB	Lactate dehydrogenase (*B. stearothermophilus*)	−1.36	−18.3	−1.5	317	0	E
2HLA	Histocompatibility antigen AW 68.1, β-chain (human)	−1.34	−11.0	−2.7	99	1	N
1HIP *r*	High potential iron protein (*Chrysomonas vinosum*)	−1.34	−8.7	−2.7	85	0	N
2PTN	Trypsin (bovine)	−1.27	−8.3	1.0	223	6	E
1CCR *r*	Cytochrome *c* (rice)	−1.23	−7.8	0.1	112	0	N
2LHB	Hemoglobin (sea lamprey)	−1.21	−13.8	−3.8	149	0	N
2LZ2	Lysozyme (turkey)	−1.17	−2.9	5.9	129	4	E
351C *r*	Cytochrome c_{551} (*Pseudomomas aeruginosa*)	−1.16	−9.4	−2.8	82	0	N
1FX1	Flavodoxin (*Desulfovibrio vulgaris*)	−1.16	−0.4	8.6	148	0	N
*1BP2 *r*	Phospholipase A2 (bovine)	−1.15	−12.0	−3.4	123	7	E
5EBX	Erabutoxin A (sea snake)	−1.14	−6.8	−1.5	62	4	N
1R69 *r*	434 Repressor (phage 434)	−1.13	−4.4	1.5	69	0	N
5DFR	Dihydrofolate reductase (*E. coli*)	−1.08	−13.2	−4.3	159	0	E
*1RNT *r*	Ribonuclease T1 (*Aspergillus oryzae*)	−1.06	−2.1	3.5	104	2	E
THIA	Glutaredoxin (bacteriophage T4)	−1.04	−8.3	−2.5	87	1	N
3PGK *r*	Phosphoglycerate kinase (baker's yeast)	−1.04	−22.3	−6.9	416	0	E
*1RN3 *r*	Ribonuclease A (bovine)	−1.01	−9.1	−2.3	124	4	E
*2PCY *r*	Plastocyanin (poplar)	−1.01	−4.6	1.3	99	0	N
*1ACX *r*	Actinoxanthin (*Actinomyces globisporus*)	−1.00	−2.1	−0.1	108	2	N
*2SOD *r*	Superoxide dismutase (bovine)	−0.97	−11.4	−3.4	152	1	E
*1LZT	Lysozyme (hen egg)	−0.97	−3.8	3.5	129	4	E
2CGA	Chymotrypsinogen A (bovine)	−0.94	−11.1	−4.5	245	5	E
2WRP *r*	Trp repressor (*E. coli*)	−0.91	−6.9	−1.7	107	0	N

(*continued*)

Table 1. *Continued*

Code	Protein (source)	S_{opt}	ΔG_{ei}	$\langle \Delta G_{ei} \rangle$[b]		N	SSF
1RBB	Ribonuclease B (bovine)	−0.89	−6.2	−0.6	124	4	E
2OVO *r*	Ovomucoid third domain (silver pheasant)	−0.87	−3.9	−1.4	56	3	N
1SN3 *r*	Scorpion neurotoxin	−0.80	−6.1	−2.3	65	4	N
1CC5	Cytochrome c_5 (*A. vinelandii*)	−0.79	−7.0	−2.4	83	1	N
*1RDG	Rubredoxin (*Desulfovibrio gigas*)	−0.78	0.7	4.4	53	4	N
1RNS	Ribonuclease-S (bovine)	−0.75	−6.6	−1.9	104	4	E
1CTX *r*	α-Cobratoxin (cobra)	−0.74	−3.5	0.3	71	5	N
*5PTI *r*	Trypsin inhibitor (bovine)	−0.61	−0.4	1.6	58	3	N
2GCH	γ-Chymotrypsin A (bovine)	−0.58	−7.7	−3.7	245	5	E
2ABX	α-Bungarotoxin (braided krait)	−0.55	−3.1	−0.2	74	5	N
2SSI *r*	*Streptomyces* subtilisin inhibitor	−0.54	−2.5	−0.5	113	2	N
5RXN *r*	Rubredoxin (*Clostridium pasteurianum*)	−0.50	9.0	11.7	54	4	N
5CHA *r*	α-Chymotrypsin A (cow)	−0.46	−6.7	−3.6	245	5	E
2GN5 *r*	Gene 5 DNA binding protein	−0.43	−5.1	−2.6	87	0	N
*1ALC	α-Lactalbumin (baboon)	−0.28	−2.2	−0.2	123	4	N
*1PPT	Avian pancreatic polypeptide (turkey)	−0.19	−0.8	−0.4	36	0	N
155C	Cytochrome c_{550}	0.07	−1.8	−2.4	135	0	N
3C2C	Cytochrome c_2 (*Rhodospirillum rubrum*)	0.18	−1.7	−2.6	112	0	N
9WGA *r*	Wheat germ agglutinin	0.63	1.8	−0.6	171	16	N

[a] The entries of the unbiased representative set RS2 are designated by ***r***. The entries used in Spassov and Atanasov (1994) are designated by an asterisk. Column designations: ΔG_{ei} and $\langle \Delta G_{ei} \rangle$, the electrostatic term of free energy for the "native" and for the random charge distributions; N, number of the amino acid side chains; SS, number of disulfide bridges; F, functional type (E, enzymes; N, proteins without enzymatic functions).
[b] Some $\langle \Delta G_{ei} \rangle$ values presented here differ by about 0.1–0.2 kcal/mol from those obtained by Spassov and Atanasov (1994). These differences are due to the different random generators used in both works.

with the fact that the disulfide-bridge density is higher for small proteins. This result not only correlates with the well-known role of disulfide bonds as structure-stabilizing factors but provides evidence that, at least in the cases of the electrostatically well-optimized structures (i.e., $S_{opt} \ll 0$), the role of the charged groups seems to be important for the optimization of the total energy of a protein structure during the folding process.

Enzymes and proteins without enzymatic functions

We have analyzed the structures within the representative data set on the basis of 2 functional classes: enzymes and proteins without enzymatic functions (designated as E and N, respectively, in Table 1). The estimated frequencies of the occurrence of group E and group N members in S_{opt} intervals are shown in Figure 2. The interactions between the ionized groups among the enzyme molecules show much better optimization than among the enzymatically inactive proteins. The estimated mean values of the optimization parameters are: $\langle S_{opt}(\mathrm{E}) \rangle = 2.25$ and $\langle S_{opt}(\mathrm{N}) \rangle = 1.63$ (for the unbiased set RS2 they are: $\langle S_{opt}(\mathrm{E}) \rangle = 2.43$ and $\langle S_{opt}(\mathrm{N}) \rangle = 1.69$). In terms of the optimization criterion P_{opt} (Equation 7), the probabilities of occurrence of random charge constellations with an energy lower than the average electrostatic energy of the enzymes (group E) are 0.012 for RS1 and 0.007 for RS2, whereas for group N, these values are about 5 times higher—0.05 and 0.045. Thus, the frequency of occurrence of group E in the interval of the highly optimized structures ($S_{opt} < -3$) is much higher than the corresponding frequency of group N ($S_{opt} > -1$). It is known from structural studies that enzymes in general form more compact globular structures than proteins without enzymatic functions. This observation is in line with the easier crystallizability of the enzymes. In terms of the electrostatic energy, it is obvious that an average compact globular structure allows a better optimization, which may result in an increased stability.

Secondary structure classes

The percentages of the α-helical and β-strand regions in the structures of the extended set RS1 are presented in Table 2. The entries were assigned to 3 types of dominant secondary structure according to the rules given by Equation 10: α-helix (α), β-strand (β), and αβ. The corresponding assignments for the entries of the unbiased set RS2 (Boberg et al., 1992) obtained with the DSSP package and the rules of Equation 9 are shown also in Table 3. The application of the rules of Boberg et al. (1992) gives almost the same assignment as is proposed for the unbiased set. This result shows that our algorithm for the calculation of the percentages of secondary structure is sufficiently accurate for the purpose of structural classification. The frequencies of occurrence of α, β, and αβ proteins in S_{opt} intervals obtained on the basis of the sets RS1 and RS2 are shown in Figure 3A and B, respectively. Classes α+β and α/β are taken as αβ for the set RS2. The distributions show a very similar shape in spite of the fact that the folding class and secondary-structure assignments were performed by different classification criteria and different algorithms. The maxima of the frequencies for the αβ class (mixed structures) are shifted by about 1σ unit to more negative S_{opt} values as compared to pure α or β class structures. The distributions are not symmetric: the class αβ proteins occur in the region of highly optimized structures ($S_{opt} < -3$), whereas class α and class β are grouped in a region ($S_{opt} \geq -1$) that re-

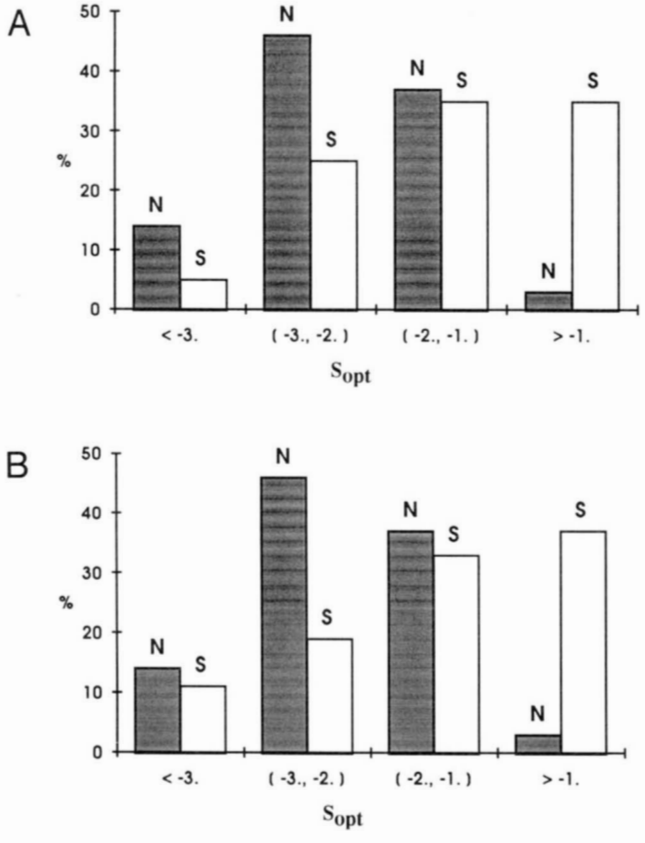

Fig. 1. Frequencies of occurrence of protein structures without disulfide bridges (N) and containing disulfide bonds (S) in intervals of the computed S_{opt}. **A:** Results obtained on basis of the extended set RS1. **B:** Results obtained on basis of the representative set RS2.

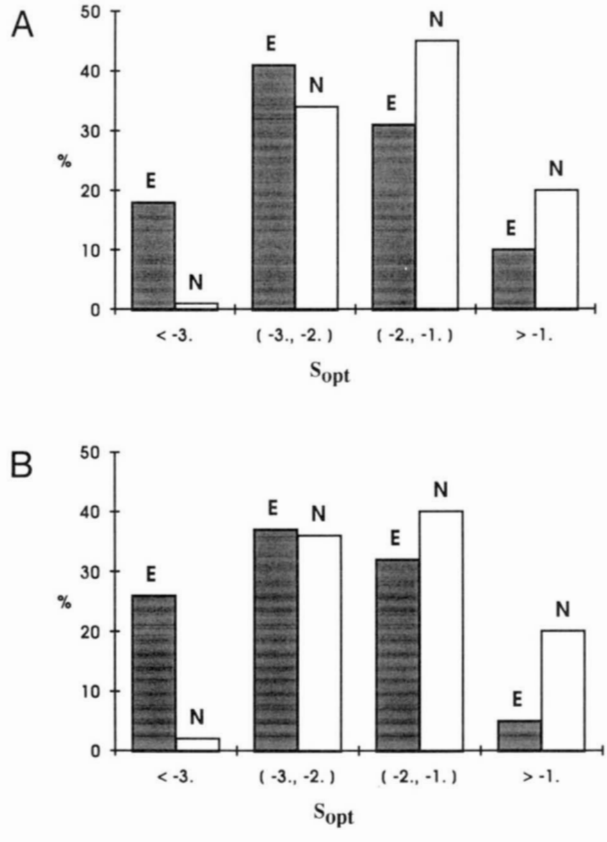

Fig. 2. Frequencies of occurrence of proteins of different functional type. E, Enzymes; N, proteins without enzymatic functions. **A:** Results obtained on basis of the extended set RS1. **B:** Results obtained on basis of the representative set RS2.

flects inferior optimization of the electrostatic interactions. This finding shows that proteins built from either α or β secondary-structure elements are characterized by essentially lower optimization. The increased variability and adaptability of the structural patterns, which certainly exist in the mixed $\alpha\beta$-type structures, seems to result in a better optimized electrostatic energy term. In the case of α and β folding types, the effect of the network of main-chain hydrogen bonds in the usually longer secondary-structure domains appears to be an alternative stabilizing factor. It follows that, for these types of proteins, the optimization of the electrostatic interactions is not necessarily the major stabilizing factor.

Attractive and repulsive electrostatic interactions

The following definitions are used in this analysis: 2 charges separated less than 5 Å are defined as an ion pair; ion pairs of charges of opposite sign are defined as salt bridges. The number of salt bridges realized in the native structures ($N_{slt,ntv}$) does not differ significantly from the number of salt bridges generated by a random process ($\langle N_{slt,rnd} \rangle$), and has a shape very similar to a Gaussian distribution (Fig. 4A). It is notable, however, that the repulsive contacts are rejected very effectively in most of the native structures, i.e., $N_{rep,ntv} \ll \langle N_{rep,rnd} \rangle$ (Fig. 4B). This shows that the number of attractive electrostatic short-range contacts in a folded protein is near the number of ion pairs statistically expected (on average 4–5 ion pairs/100 residues) for randomly distributed charged groups on the protein surface. However, the shape of the frequencies of repulsive contacts estimated for the native structures is essentially different: most of the proteins show from zero to no more than 2 ion pairs with equal sign per 100 residues. The electrostatic term of the free energy in a folded protein is therefore effectively minimized by rejecting the conformers, leading to repulsive electrostatic contacts, but remains at this stage without further improving the network of attractive interactions. This result confirms the previous observation (Spassov & Atanasov, 1994) based on a reduced set of protein structures. Our analysis shows that salt bridges seem to occur in folded proteins with a relatively constant ratio of about 4 per 100 amino acid residues. In principle, the electrostatic energy term could be optimized by increasing the number of salt bridges (attractive interactions) or by decreasing the number of repulsive interactions. It appears to be a general rule for proteins that a gain in electrostatic stabilization is achieved rather by minimizing the number of repulsive contacts. Particularly, the rejection of repulsive contacts seems to be 1 major goal in the protein folding process to achieve an energetically optimized distribution of charged amino acid side chains. This may also be reflected in the functional properties of proteins. Thus, for example, it was shown for the case of thrombin (Bode et al., 1992; Bode & Karshikov, 1993) that ion-

Table 2. *Computed percentage of α-helical and β-strand regions and secondary structure folding classes*

Code	S_{opt}	α-Helix %	β-Strand %	Folding class[a] C1	C2
6ACN	−6.23	33.5	33.2	αβ	α/β
2CPP	−4.32	49.8	21.3	α	α+β
8CAT	−4.09	32.8	28.2	αβ	α+β
8ADH	−3.88	29.0	34.0	αβ	α+β
RUBA	−3.83	24.3	26.2	αβ	–
3GRS	−3.72	32.0	35.7	αβ	α+β
1GOX	−3.58	44.5	23.3	αβ	α/β
BWSY	−3.28	41.9	27.6	αβ	α/β
1CTS	−3.15	63.3	7.3	α	α
1GPD	−3.13	28.6	26.2	αβ	–
1MBD	−3.06	83.6	0.0	α	α
9PAP	−3.05	23.7	29.9	αβ	α+β
1SGT	−3.03	8.1	43.7	β	β
2I1B	−2.99	0.0	67.8	β	β
AHLA	−2.99	28.3	41.3	αβ	α+β
2ALP	−2.98	3.6	46.2	β	–
LFB4	−2.92	7.4	64.7	β	β
2HHB	−2.90	75.0	4.3	α	α
1WSY	−2.90	55.1	23.1	α	α/β
2GBP	−2.88	41.5	24.8	αβ	α/β
GTRA	−2.88	60.2	12.1	α	–
4XIA	−2.88	49.5	16.6	α	α/β
2FB4	−2.87	0.0	62.3	β	β
2GD1	−2.86	30.9	34.5	αβ	α/β
1PHH	−2.82	35.6	31.6	αβ	α+β
3LZM	−2.79	65.6	6.7	α	α
1GCR	−2.79	5.2	48.0	β	β
5CPA	−2.76	37.3	22.2	αβ	α+β
1TIM	2.73	47.6	19.1	α	α/β
1PPD	−2.71	23.7	28.0	αβ	–
1SBC	−2.70	30.4	24.2	αβ	α/β
2PRK	−2.64	24.8	26.3	αβ	–
2CI2	−2.61	17.2	45.3	β	α+β
3RP2	−2.59	8.1	52.5	β	–
2AZA	−2.59	14.1	45.3	β	α+β
4MDH	−2.57	42.5	25.9	αβ	–
1RHD	−2.55	32.2	32.5	αβ	α/β
1HOE	−2.54	0.0	61.6	β	β
4FD1	−2.51	33.3	18.1	αβ	α+β
1PMB	−2.50	80.3	0.0	α	–
2STV	−2.49	12.6	54.1	β	–
1GP1	−2.48	25.7	28.4	αβ	α/β
1TON	−2.44	8.0	49.6	β	–
1ABP	−2.42	44.9	17.0	α	–
1CA2	−2.41	12.2	48.2	β	–
2SNS	−2.40	27.9	30.0	αβ	α+β
2MHR	−2.39	70.9	13.7	α	α
3DFR	−2.37	21.7	47.2	β	–
256B	−2.35	79.0	0.0	α	α
1CLA	−2.34	29.2	42.0	αβ	α+β
1REI	−2.33	0.0	64.2	β	–
2CDV	−2.33	25.5	30.2	αβ	α
3PGM	−2.28	35.4	15.3	α	α
3GAP	−2.28	33.8	27.5	αβ	α+β
2LIV	−2.25	44.6	25.7	αβ	–
2CAB	−2.22	17.3	45.5	β	α+β
6LDH	−2.22	43.9	25.6	αβ	α/β
2PAZ	−2.15	16.4	54.1	β	α+β
4MBA	−2.09	80.7	3.4	α	–
1UTG	−2.09	78.3	0.0	α	α
3BLM	−2.08	41.8	25.8	αβ	–
8ATC	−2.05	22.1	39.3	αβ	α+β
2ACT	−2.03	25.7	26.1	αβ	–
3B5C	−2.02	48.8	20.2	α	α/β
2HHB	−2.02	77.2	0.0	α	α
1FXB	−1.97	18.8	16.3	αβ	α
1CRN	−1.94	46.7	26.7	αβ	–
8DFR	−1.91	23.2	49.7	β	α/β
2LBP	−1.90	41.2	25.2	αβ	α/β
3HVP	−1.90	7.1	61.2	β	β
1TNF	−1.89	0.0	60.9	β	β
4INS	−1.86	54.0	24.0	α	α
8ATC	−1.86	39.5	20.7	αβ	α+β
2PKA	−1.85	7.4	45.5	β	–
1HMQ	−1.82	70.5	9.8	α	–
2CCY	−1.80	77.0	6.3	α	α
2PAB	−1.75	12.4	55.8	β	β
3ADK	−1.75	62.2	20.2	α	α/β
3EST	−1.74	7.1	49.0	β	–
2AAT	−1.71	41.0	12.9	α	α
2SGA	−1.69	7.8	43.3	β	β
2CRO	−1.68	62.5	0.0	α	–
2UTG	−1.68	79.9	2.9	α	–
1CYC	−1.65	41.2	8.8	α	–
2RHE	−1.65	5.3	64.6	β	–
1LZ1	−1.63	38.0	10.9	α	α
2LH4	−1.63	77.6	2.6	α	α
5CYT	−1.59	46.1	13.7	α	–
1UBQ	−1.59	16.0	45.3	β	α+β
1ECA	−1.56	78.5	3.0	α	α
2MLT	−1.53	96.0	0.0	α	–
1PP2	−1.52	42.1	11.6	α	–
1NXB	−1.52	0.0	57.4	β	β
1CTF	−1.48	52.2	23.9	α	α/β
3FXC	−1.46	8.2	27.8	β	α+β
RVSA	−1.45	47.4	28.9	αβ	–
1HNE	−1.42	6.3	41.2	β	–
4FXN	−1.42	38.7	24.8	αβ	α/β
2RSP	−1.40	6.1	75.4	β	β
1LDB	−1.36	41.6	22.9	αβ	–
2HLA	−1.34	0.0	70.4	β	–
1HIP	−1.34	12.0	15.7	*i*	α+β
2PTN	−1.27	9.0	42.3	β	–
1CCR	−1.23	45.5	17.3	α	α
2LHB	−1.21	75.7	7.4	α	–
2LZ2	−1.17	35.9	10.2	α	–
351C	−1.16	51.9	11.1	α	α
1FX1	−1.16	28.8	28.1	αβ	–
1BP2	−1.15	50.0	14.8	α	α
5EBX	−1.14	0.0	75.4	β	–
1R69	−1.13	69.4	0.0	α	α
5DFR	−1.08	22.9	46.4	β	–
1RNT	−1.06	16.5	34.0	β	α+β
THIA	−1.04	36.0	23.3	α	–
3PGK	−1.04	38.9	16.2	α	α/β
1RN3	−1.01	19.5	42.3	β	α+β
2PCY	−1.01	4.1	57.1	β	β
1ACX	−1.00	0.0	40.2	β	β

(*continued*)

Table 2. *Continued*

Code	S_{opt}	α-Helix %	β-Strand %	Folding class[a] C1	C2	Code	S_{opt}	α-Helix %	β-Strand %	Folding class[a] C1	C2
2SOD	−0.97	0.0	44.0	β	β	2GCH	−0.58	9.4	47.2	β	–
1LZT	−0.97	41.4	10.2	α	–	2ABX	−0.55	0.0	5.5	*i*	–
2CGA	−0.94	14.3	45.9	β	–	2SSI	−0.54	16.0	37.7	β	α+β
2WRP	−0.91	78.8	3.8	α	α	5RXN	−0.50	0.0	37.7	β	β
1RBB	−0.89	22.0	48.0	β	–	5CHA	−0.46	9.4	42.6	β	α+β
2OVO	−0.87	18.2	36.4	β	β	2GN5	−0.43	0.0	38.4	β	*i*
1SN3	−0.80	12.5	39.1	β	α+β	1ALC	−0.28	32.2	14.0	α	–
1CC5	−0.79	48.8	4.9	α	–	1PPT	−0.19	51.4	17.1	α	–
1RDG	−0.78	0.0	35.3	β	–	155C	0.07	33.8	9.0	α	–
1RNS	−0.75	21.7	46.7	β	–	3C2C	0.18	48.6	16.2	α	–
1CTX	−0.74	0.0	31.4	β	β	9WGA	0.63	11.8	24.1	β	α
5PTI	−0.61	24.6	38.6	αβ	α+β						

[a] Classification C1 is obtained according to Equation 10; classification C2 is taken from Boberg et al. (1992).

izable groups that do not take part in salt bridges are responsible for the establishment of the electrostatic field in and around the protein molecule, a feature that is essential for substrate and inhibitor binding.

Dependence of the optimization parameters on the electrostatic model

In this study, we tested the sensitivity of the computed optimization parameter, S_{opt}, by repeating the calculations with a number of different potential functions describing the electrostatic interactions (Table 3; see also Methods). In terms of the approach used, the S_{opt} values are a measure of the difference between the charge distribution in a folded structure and the random state. This is a result of the minimization of ΔG_{ei} in the native structure. If we consider ΔG_{ei} as a function of the algebraic form and parameterization of the electrostatic potential function $W_{mod}(r_{ij})$, S_{opt} could also be used as a quality criterion of the different models. Incorrect models are expected to give positive S_{opt} values, whereas the more adequate models will give more negative S_{opt} values. From Equations 2 and 8, it can be seen that S_{opt} is independent of the calibration of the potential function: $S_{opt}[w(r_{ij})] = S_{opt}[a \cdot w(r_{ij})]$ for any $a > 0$. This allows us to use normalized functions for the further analysis. The shapes of the normalized functions $|W_{mod}(r_{ij})| =$

Table 3. *Averaged results of the test of different electrostatic models*[a]

No.	Potential function $W_{mod}(r_{ij})$	$\langle S_{opt}[W'_{mod}]\rangle$	$\langle S_{opt}[W_{mod}]\rangle$
1	$W_{se} = a1/r + a2/r^2 + a3/r^3$	−1.923	−1.720
2	$W_{DD} = B/D \cdot r$, $D = r$	−1.935	−1.732
3	$W_{CL} = B/D \cdot r$, D = constant	−1.746	−1.626
4	$W_{TK}(r, R_a, d, D_i)$, $d = 1.0$ Å	−1.911	−1.718
5	$W_{TK}(r, R_a, d, D_i)$, $d = 0.4$ Å	−1.899	−1.709
6	$W_{TK}(r, R_a, d, D_i)$, $d = 0.0$ Å	−1.805	−1.668
7	W_{DH}, $\kappa = 3.5$ Å	−1.973	−1.737
8	W_{DH}, $\kappa = 22.0$ Å	−1.813	−1.673

[a] $\langle S_{opt}\rangle$ represents the mean values of the optimization parameter, S_{opt}, obtained on the full set of 141 protein structures. The calculations performed with the corrections for the solvent accessibility are designated as $W'_{mod} = (1 - SA_{ij}) \cdot W_{mod}$.

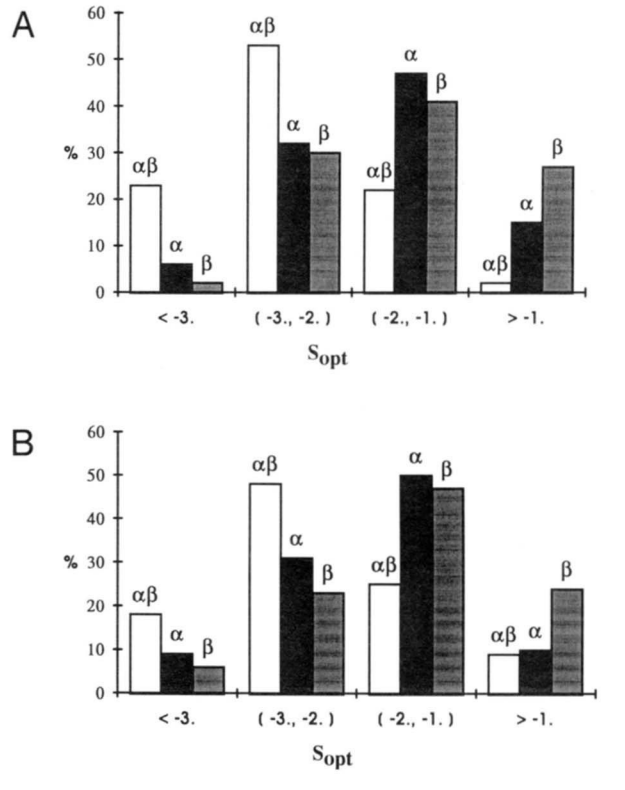

Fig. 3. Frequencies of occurrence of proteins of different secondary structure folding type. **A:** Set RS1 and classification C1 (see Table 2 and text). **B:** Set RS2 and classification C2 (the classes α/β and $\alpha+\beta$ are taken as $\alpha\beta$).

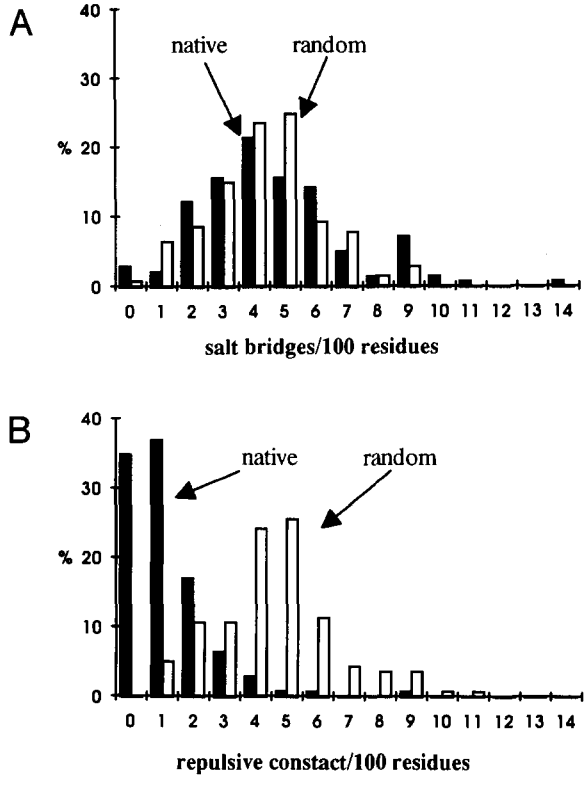

Fig. 4. Frequencies of occurrence of proteins with different number of ion pairs per 100 residues, obtained on the extended set of 141 structures (RS1). On the abscissa: **(A)** ($N_{slt,ntv}$/100 residues), number of salt bridges in the native, and ($\langle N_{slt,rnd}\rangle$/100 residues), number of the expected salt bridges in random charge constellations; **(B)** ($N_{rep,ntv}$/100 residues), number of repulsive contacts in the native structures, and ($\langle N_{rep,rnd}\rangle$/100 residues), number of the expected repulsive contacts in random charge constellations.

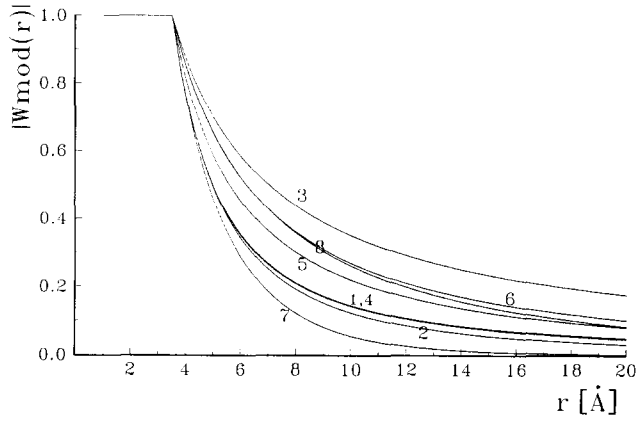

Fig. 5. The model potential functions $|W_{mod}(r)| = W_{mod}(r)/W_{mod}(3.5)$ normalized at $r_{ij} = 3.5$ Å and used to test different electrostatic models. The enumeration of the different potential functions corresponds to that in Table 3. The Tanford–Kirkwood potential function (curves 4, 5, and 6) corresponds to an internal dielectric constant $D_i = 4$ and a molecular radius $R_a = 20$ Å. However, the calculations are performed with values of R_a estimated for each particular structure. Therefore, the observed identity of models 1 and 4 is valid only for proteins with R_a close to 20 Å.

$W_{mod}(r_{ij})/W_{mod}(3.5)$ are shown in Figure 5. These functions, together with the mean values $\langle S_{opt}\rangle$ of the computed optimization parameters, averaged over the set RS1, are listed in Table 3. The values $\langle S_{opt}[W_{mod}\cdot(1 - SA_{ij})]\rangle$ and $\langle S_{opt}[W_{mod}]\rangle$ represent the results of the application of the electrostatic models with and without the correction by the mean static solvent accessibilities (SA_{ij}) of the interacting groups, as discussed in the Methods section. The computed S_{opt} values for the individual proteins are compared for the different potential functions $W_{mod}(r_{ij})$ in Figure 6.

On the basis of the averaged results (Table 3) and the individual differences in S_{opt} (Figs. 6, 7), the following conclusions can be made: (1) The classification of protein structures by the degree of spatial optimization (S_{opt}) of the electrostatic interactions is independent of the choice of the potential function. As shown in Figure 6, the majority of the potential functions W_{mod} give practically identical results. The more significant differences are observed in the case of the Coulomb potential (model 3) and in the case of models 6 and 8, which are characterized by a shape close to $1/r$. However, these differences do not essentially influence the arrangement of the individual structures with respect to the optimization of the electrostatic interactions. (2) The averaged values $\langle S_{opt}\rangle$ (see Table 3) strongly correlate with the slope of the potential functions used (Fig. 5). The models, where the normalized function $W_{mod}(r)$ decreases more rapidly with distance, show more negative values of $\langle S_{opt}\rangle$. This regularity can also be seen in the individual S_{opt} values (see Fig. 6) for a large number of proteins, and it is stronger in "well"-optimized structures. Although the model functions are tested only as a formal algebraic expressions (some of them have different parameterization), the results indicate that the short-range interactions seem to be the major part, forming the free energy term ΔG_{ei}. (3) The inclusion of the correction term $(1 - SA_{ij})$ in the electrostatic models gives systematically better results for most of the structures investigated (Fig. 7). The solvent accessibility correction seems to be a reasonable tool to account for the differences in the local dielectric properties at the protein-water interface in the simplified electrostatic models, based on mean force field potentials.

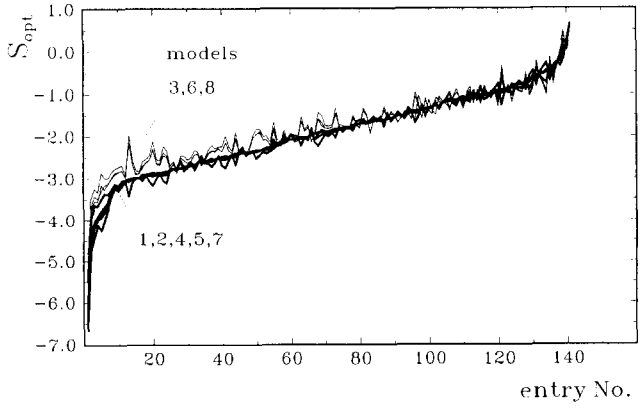

Fig. 6. S_{opt} calculated for the set RS1 by means of the different model potential functions. The enumeration of the models corresponds to that in Table 3. The enumeration of the protein items corresponds to that in Table 1.

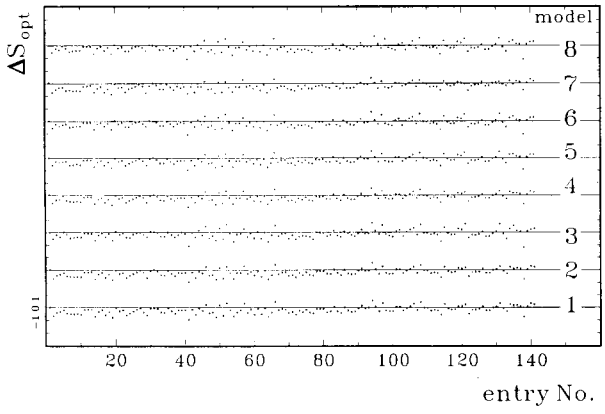

Fig. 7. The effect of the correction term $(1 - SA_{ij})$ on the values of S_{opt}: $\Delta S_{opt} = S_{opt}[W_{mod} \cdot (1 - SA_{ij})] - S_{opt}(W_{mod})$. The enumeration of the models corresponds to that in Table 3. The enumeration of the protein items corresponds to that in Table 1. The negative values of ΔS_{opt} correspond to an improvement of the optimization parameter.

Conclusions

The analysis of the extended set of 141 protein structures shows that different proteins exhibit different degrees of optimization of the charge–charge interactions. The majority of native structures, however, is characterized by an optimized charge distribution compared to the corresponding random states. Obviously, the reason for these differences is hidden in the structure of the folded state, which, in any individual case, corresponds to a concrete architecture related to a concrete function. Therefore, the question of the importance of the electrostatic interactions for stability and folding of globular proteins seems to have no universal answer—the charge constellation in each individual structure is optimized individually with different efficiency. Nevertheless, the approach used gives the opportunity not only to classify the structures by the degree of optimization of the electrostatic interactions on a common scale but to search for correlation between the estimated optimization parameters and the most common structural and functional characteristics of proteins.

Most of the known theoretical models for analysis of the electrostatics in proteins are mainly focused on molecular properties such as structural stability, dissociation characteristics, electrostatic field, and potential. The method used in the present work gives the opportunity to estimate how optimal the geometry of the charge constellation is for any given protein structure and provides a different measure for the role of electrostatic interactions as a factor in the folding process.

One of the most substantial criticisms against our results could arise from questioning the use of a simplified macroscopic electrostatic model based on a mean force field (in view of the existing numerical methods for the solution of the Poisson–Boltzmann equation). We do not claim that the obtained classification is absolute, but the test of a variety of electrostatic potentials with very different shapes shows that the results seem to be independent of the choice of the electrostatic model. An indirect indication for this is the fact that the observed correlation between the computed optimization parameters and the functional and structural classes can be interpreted in terms of observable properties of the proteins.

Methods

Energy calculations

The contribution of the acidic and basic side chains to the free energy, ΔG_{el}, can be represented by:

$$\Delta G_{el}(R, \text{pH}, \text{I}, \dots) = \Delta G_{Born}(R, \text{pH}, \text{I}, \dots) + \Delta G_{pc}(R, \text{pH}, \text{I}, \dots) + \Delta G_{ei}(R, \text{pH}, \text{I}, \dots), \quad (1)$$

where R is the vector of the charge coordinates, and the other arguments represent the physicochemical properties of the protein molecule and the surrounding solution (pH, ionic strength, internal and external dielectric constants, etc.). ΔG_{Born} is the Born energy term of the individual charged group and ΔG_{pc} is the contribution of the interaction energy between charged groups and other permanently charged atoms in the protein. In this work, the attention is focused on the term ΔG_{ei}—the contribution of the electrostatic interaction between the charged groups. This term is a direct function of the specific 3-dimensional distribution of the charged groups, whereas ΔG_{Born} and ΔG_{pc} are sums of the individual interaction energies of the residual charges with the environment. The development of numerical techniques for the solution of the Poisson–Boltzmann equation for macromolecules (Warwicker & Watson, 1982; Gilson et al., 1987; You & Harvey, 1993), and the use of Boltzmann statistics to describe the site protonation/deprotonation equilibrium (Bashford & Karplus, 1990; Yang et al., 1993), as well as the microscopic description of the electrostatic interactions (Warshel & Russell, 1984; King et al., 1991; Lee et al., 1993) demonstrate the recent success in the extremely complicated problem of modeling electrostatic properties and titration behavior of protein molecules. However, these techniques are inappropriate for the purposes of this study because the computational effort increases dramatically in a Monte Carlo experiment on a large set of structures. Therefore, it is convenient to use a more simple physical approach to the electrostatic problem. The interaction energy between the charged groups was taken as:

$$\Delta G = \tfrac{1}{2} \sum_i \sum_j Q_i Q_j w(r_{ij}) \quad (i, j = 1, \dots, NG), \quad (2)$$

where $w(r_{ij})$ is the potential function of electrostatic interaction between the charges Q_i and Q_j separated by a distance r_{ij}, and NG is the number of the charged groups. The calculations were repeated using several types of the uniform potential function $w(r_{ij})$. Each of them was tested with and without the correction term corresponding to the static solvent accessibility, SA_{ij}, first introduced by Shire et al. (1974):

$$w(r_{ij}) = W_{mod}(r_{ij}) \cdot (1 - SA_{ij}) \quad (3)$$

or

$$w(r_{ij}) = W_{mod}(r_{ij}), \quad (3')$$

where SA_{ij} represent the averages of the normalized static solvent accessibilities of the ionized groups (see Matthew et al.,

1979). $W_{mod}(r_{ij})$ represents the concrete potential function. Two classes of uniform potential function are usually used for description of the electrostatic interactions in proteins. The first class represents potentials, used in models for analysis of pH-dependent properties of proteins, and are parameterized and tested on the basis of experimental titration data (Tanford & Roxby, 1972; Matthew et al., 1979; Spassov et al., 1989). The second class represents the electrostatic potential functions, defined and parameterized as part of the force field in different programs based on molecular mechanics and molecular dynamics methods (AMBER, CHARM, TRIPOS). Initially, the calculations described in this work were performed using the semi-empirical potential function $W_{se}(r_{ij})$, proposed by Spassov et al. (1989):

$$W_{mod}(r_{ij}) = W_{se}(r_{ij}) = a_1/r_{ij} + a_2/r_{ij}^2 + a_3/r_{ij}^3, \quad (4)$$

where the values of the empirical parameters ($a_1 = 2.9$ [Å·kcal/mol], $a_2 = 40.6$ [Å^2·kcal/mol], and $a_3 = 40.8$ [Å^3·kcal/mol]) were obtained by fitting protein titration experiments. In addition, the following potential functions were applied in order to test the dependence of the results on the type of the potential function:

1. Kirkwood-Tanford (KT) models based on the analytical solution of the Poisson-Boltzmann equation (Kirkwood, 1934; Tanford & Kirkwood, 1957). In this case, the corresponding potential, $W_{KT}(r_{ij})$, may be represented as 3-parameter function:

$$W_{mod}(r_{ij}) = W_{KT}(r_{ij}, R_a, d, D_i), \quad (5)$$

where R_a is the radius of the protein molecule, the parameter d represents the postulated depth of the charges under the protein surface, and D_i (usually equal to 4) is the assumed internal dielectric constant. The explicit expression for $W_{KT}(r_{ij})$ is given in Tanford and Kirkwood (1957). Three values of the parameter d were used that give a different ratio of short-range to long-range interactions: $d = 1$ Å (Tanford, 1957); $d = 0.4$ Å (Tanford & Roxby, 1972); $d = 0$ Å (Karshikov et al., 1989).

2. Coulomb potential functions, most frequently used as part of the force field in molecular mechanics:

$$W_{mod}(r_{ij}) = K/D(r_{ij})/r_{ij},$$

where $K = 332$ is chosen so that W_{mod} is in kcal/mol, the charge values in proton units, and the distance r_{ij} in Å. The functions chosen from this class are the Coulomb potential, $W_C(r_{ij}) = K/D \cdot 1/r_{ij}$ (D = const), and $W_{dd} = K/r_{ij}^2$, a potential with distant-dependent dielectric constant ($D(r_{ij}) = r_{ij}$).

3. Debye-Hückel potential:

$$W_{mod}(r_{ij}) = W_{DH}(r_{ij}) = K\frac{e^{-r_{ij}/\kappa}}{Dr_{ij}}, \quad (6)$$

where κ is the Debye length. The Debye-Hückel potential has recently been proposed as the function that gives the best fit in the statistical analysis of the distance distribution of the charged residues in protein molecules, $\kappa = 22$ Å (Bryant & Lawrence, 1991), and in the statistical "projection" method, $\kappa = 3.5$ Å (G. Casary & A. Beyer, pers. comm.). A number of other potential functions are described in the literature (see, for example, Warshel et al., 1984; Hingerty et al., 1985); however, for the purpose of this study, the selected potential functions, $W_{mod}(r_{ij})$, cover a sufficiently large diversity of potential curve shapes.

Monte Carlo simulations

The computational scheme for estimating the degree of spatial optimization of charge-charge interactions in different protein structures in general follows the algorithm suggested and described in Spassov and Atanasov (1994). For each of the investigated proteins, the electrostatic energy term of the native structure, $\Delta G_{ei,nat}$, is calculated by Equation 2 and Equation 3 or 3′. The charges are defined as point charges with coordinates of the Nζ atoms for lysines, and the average coordinates of the carboxyl oxygens, Nδ1 and Nϵ2, Nη1 and Nη2 atoms for aspartic and glutamic acids, histidines, and arginines, respectively. The standard set of charges, Q_i, corresponds to neutral pH: values of -1 for the carboxyl groups, $+1$ for lysines, arginines, and the terminal amino group, and $+1/2$ for histidines. The validity of this charge assignment was tested for the native structure of each protein by the comparison of the calculated $\Delta G_{ei,nat}$ with the value, ΔG_{ei}(pH = 7), estimated for pH 7. The titration of the individual groups and the charge values at pH 7 were calculated by means of the method proposed by Spassov et al. (1989). A small number of proteins show significant differences between ΔG_{ei} and ΔG_{ei}(pH = 7). These cases will be discussed below in this section.

The second step consists of the generation of a large number of random charge constellations with the same type and number of charged groups. Each of them is randomly distributed on the protein surface. The *SA* calculations show no or only a small percentage of charged groups characterized by a zero accessibility to the solvent for all 141 native structures investigated. In accordance with this result, the set of possible coordinates of the charges, R^s, is defined by the centers of the protein atoms characterized by a non-zero solvent accessibility calculated by the method of Lee and Richards (1971). The peptide backbone atoms, as well as the Cβ atoms, were excluded from this set. For each of the investigated proteins, 1,000 nonequivalent random distributions, ($R_{rnd,i}$, $i = 1, \ldots, NR$), of the coordinates of NG charges are extracted from the set of surface atoms (R^s) using an integer-number Monte Carlo algorithm (Spassov & Atanasov, 1994). A similar approach was used by Barlow and Thornton (1986) for analysis of charge asymmetry. The frequencies, $f(\Delta G_{ei,rnd})$, of a random constellation in the intervals of ΔG_{ei} are very similar to the normal (Gaussian) distribution, $p(\Delta G_{el})$, indicating that $R_{rnd,i}$ is sufficiently large for a statistical investigation. The electrostatic energy term for the individual random charge distributions, $\Delta G_{ei,rnd}$, was calculated in the same way as $G_{ei,ntv}$, keeping the number, types, and SA_i values of the charged groups equal to those in the native protein, but replacing the coordinates with $R_{rnd,i}$.

It is convenient to introduce a dimensionless statistical criterion for the spatial optimization of the electrostatic interactions that can provide an opportunity to compare the electrostatic properties of proteins with different numbers and types of charged

residues, different shape, surface area, etc., on a common scale. This criterion can be the probability, P_{opt}, of generating a random charge constellation with an energy lower than the energy of the native structure:

$$P_{opt} = P(\Delta G_{ei,rnd} < \Delta G_{ei,ntv}) = \int_{-\infty}^{\Delta G_{ei,ntv}} p(\Delta G_{ei})\, d\Delta G_{ei}, \quad (7)$$

where

$$p(\Delta G_{ei}) = \left(\frac{1}{\sigma}\sqrt{2\pi}\right) e^{\frac{-(\Delta G_{ei} - \langle \Delta G_{ei,rnd} \rangle)}{2\sigma^2}}, \quad (7')$$

where ΔG_{ei} is the electrostatic energy of the native structure, $\langle \Delta G_{ei,rnd} \rangle = \sum G^i_{ei,rnd}/NR$ is the mean of the corresponding energies of NR charge distributions, and σ is the standard deviation. The statistical criterion P_{opt} can be substituted by the alternative optimization parameter S_{opt},

$$S_{opt} = (\Delta G_{ei,ntv} - \langle \Delta G_{ei,rnd} \rangle)/\sigma, \quad (8)$$

used in this study. A similar expression is commonly used in statistics to calculate probabilities in the case of Gaussian distributions and allows a more compact record of the results. The dimensionless parameter S_{opt} represents the degree of deviation of the electrostatic energy of the native structure from that expected if the electrostatic interactions do not take part in the stabilization of the folded structure (i.e., if the coordinates of the charged groups are not related to the electrostatic interactions between them). For a given structure, the more negative values of S_{opt} (or $P_{opt} \to 0$) indicate the better optimized charge–charge interactions. Thus, for example, $S_{opt} = 0$ corresponds to $P_{opt} = 0.5$, $S_{opt} = -1$ to $P_{opt} = 0.16$, $S_{opt} = -2$ to $P_{opt} = 0.23$, $S_{opt} = -3$ to $P_{opt} = 0.01$, etc.

Structural input data and selection of the representative protein set

The protein structures used in this study were extracted from the PDB. The collection of crystallographic and NMR structures in the PDB (April 1993) contains about 800 files of 3-dimensional atomic coordinates of protein molecules. A preliminary selection was made by excluding the structures with a resolution of ≥ 3 Å and the entries containing only C_α atoms or with significant incompleteness in the atomic coordinates, as well as the model and mutant structures. For the entries of protein structures with equal sequence but obtained at different conditions by different authors, or at different resolution, the representative structures were selected by the criteria of the most complete atomic coordinates and the superior resolution.

For a small number of proteins, the use of the standard charge values at neutral pH was found to be not valid. The values $\Delta G_{ei,ntv}$, computed with standard Q_i, and $\Delta G_{ei,ntv}$(pH), estimated by the method given in Spassov et al. (1989), differ essentially (more than 0.5 kcal/mol) for these proteins. This is an indication that some titratable residues are characterized by pK_a values shifted from the usual values, so, in terms of our study, these groups cannot be treated as standard charges. Therefore, some of the aspartic proteinases (endothiapepsin, penicillopepsin) and some proteins whose structure seems to be strongly dependent on binding of Ca^{2+} ions (calmodulin, troponin C, parvalbumin, and thermolysin) are excluded from the data set. These cases need more detailed analysis, which is in progress and will be described elsewhere. For the multisubunit proteins, the subunits with different primary structure are presented as separate entries. On this basis, 139 items were selected. In addition, the inactivated form of rubisco (RUBA in Table 1) and of T4 thioredoxin were included. The coordinates were kindly provided by the authors, G. Schneider and H. Eklund, BMC, Uppsala, Sweden. This collection of 141 structures, listed in Table 1, is defined as extended set RS1. This set contains representatives of almost all known protein structures that have nonidentical primary structure, however, with some degree of sequence homology. The structures with known sequence similarity included in RS1 (e.g., the family of serine proteinases, globins, etc.) show, however, essential differences in their charge constellation (nonequal number of charged groups, different abundance of ionizable groups of a given type, different positions, etc.). Thus, it is a serious problem to construct a representative set of proteins with minimal sequence homology. Boberg et al. (1992) have proposed an unbiased representative set of 103 proteins obtained by sequence alignment of the PDB structures with the GCG program GAP (Devereux et al., 1984), statistical estimation of the significance of sequence similarity (Lipman et al., 1985), and an original clustering algorithm. Eighty-four proteins of 103 from the unbiased set proposed by Boberg et al. (1992) coincide with those selected in RS1; the other 19 entries, characterized by a resolution of ≥ 3 Å or by incomplete atomic coordinates, were not appropriate for this study. Here we define these 84 structures as the representative set RS2. RS2 is a subset of RS1 and the corresponding entries are marked by ***r*** in Table 1. Most of the results given below are represented independently for both sets RS1 and RS2, given in Table 1. Some minor differences in the primary structure data (the header sequence records in PDB) are possible because of the poorly defined electron density maps. The residues with undefined side-chain atomic coordinates were considered as mobile and strongly exposed to the solvent, i.e., their averaged effect on the electrostatic free energy terms is negligible.

Classification of the protein structures

For the purpose of the search for possible relationships between the electrostatic optimization parameters and some common characteristics of the investigated proteins such as functional type or folding class, the representative sets were partitioned according to different criteria. It would certainly be interesting to compare the structures by more concrete functional criteria. However, the recent PDB version does not contain functional classes with a sufficiently large number of different structures. Thus, the dimensions of RS1 and RS2 (141 and 84 entries) do not allow us to obtain good statistics by dividing the proteins into more than 2 or 3 groups of similarity. A most simple functional classification, satisfying the condition for a sufficient statistical weight, is the division of the proteins into enzymes and proteins without enzymatic function (in Table 1, marked E or N, respectively). The second simple classification used in this work was made by considering proteins with and without disul-

fide bonds. The reason for this is to see how the existence of covalent cross bridges as a structure-stabilizing factor relates to the efficiency of the electrostatic interactions expressed by the parameter S_{opt}. This type of relationship has already been observed on a smaller set of 44 proteins (Spassov & Atanasov, 1994). Here we check the validity of this observation by estimating the frequencies of occurrence of proteins of both types in intervals of S_{opt} for both the enlarged (RS1) and the unbiased (RS2) sets.

A technically more complicated problem is to classify the protein structures on the basis of their secondary or tertiary structure, i.e., by folding classes. According to the most frequently used classification (Levitt & Chothia, 1976), the proteins can be related to 3 secondary-structure classes: α-helical, β-sheet, and mixed $\alpha\beta$ type. The topological and 3-dimensional organization of the secondary-structural elements can be represented by different types of β sheets, e.g., $\alpha+\beta$ or α/β folding types (see Lesk, 1991). Boberg et al. (1992) have proposed a rule for the classification into folding types based on the percentages of α and β secondary-structural elements (see also Nakashima et al., 1986):

$$\text{Folding type} = \begin{cases} \boldsymbol{\alpha}: & (\alpha \geq 10\%,\ \beta < 10\%) \text{ and } \alpha \geq 2\beta \\ \boldsymbol{\beta}: & (\beta \geq 10\%,\ \alpha < 10\%) \text{ and } \beta \leq 2\alpha \\ \boldsymbol{\alpha+\beta},\ \boldsymbol{\alpha/\beta}: & (\alpha \text{ and } \beta \geq 10\%) \\ & \text{or } (\alpha < 10\%,\ \beta < 10\%, \text{ and } \alpha < 2\beta) \\ & \text{or } (\beta \geq 10\%,\ \alpha < 10\%, \text{ and } \beta < 2\alpha). \end{cases} \quad (9)$$

Although this rule is quite suitable and accurate for the objective estimation of the type of protein architecture, we have defined a similar but more simple rule:

$$\text{Folding type} = \begin{cases} \boldsymbol{\alpha}: & \alpha > 10\% \text{ and } \alpha/\beta > 2 \\ \boldsymbol{\beta}: & \beta > 10\% \text{ and } \beta/\alpha > 2 \\ \boldsymbol{\alpha\beta}: & \alpha+\beta > 20\% \text{ and } 0.5 < \alpha/\beta < 2. \end{cases} \quad (10)$$

The latter criteria allow us to distinguish between the structure classes of "dominant" α or β type in a more flexible way. For some of the $\alpha\beta$ structures assigned by Equation 9, Equation 10 gives α or β if the ratio of the corresponding percentages is more then 2, even if both α and β are greater than the 10% limit. A computer program was developed for assignment of regions of specific secondary structure and for the calculation of the corresponding α or β percentages. The input data for this program are the peptide backbone dihedral angles. Residues with $(-180° \leq \phi \leq 0°,\ 0° \leq \psi \leq 180°)$ are assigned to type β and $(-180° \leq \phi \leq 0°,\ -180° \leq \psi \leq 0°)$ to type α. The residues with (ϕ, ψ) values corresponding to the other 2 quadrants of the Ramachandran plot are assumed to participate in segments of irregular structure. We use the following definition for secondary-structure regions: regular α-helix or β-strand, if 4 or more subsequent residues are in type α or β; if otherwise, an irregular structure is assumed. The percentages of both α and β structures computed by our algorithm and Equation 9 are systematically higher in comparison to those obtained by Kabsch and Sander (1983), where the criteria were based on the H-bonding network and chirality of the peptide chain. However, we obtain almost identical secondary-structure assignments as obtained by Boberg et al. (1992) and shown for the set RS2 in Table 2.

References

Anderson DE, Becktel WJ, Dahlquist FW. 1990. pH-induced denaturation of proteins: A single salt bridge contributes 3-5 kcal/mol to the free energy of folding of T4 lysozyme. *Biochemistry 29*:2403-2408.

Barlow DJ, Thornton JM. 1983. Ion-pairs in proteins. *Mol Biol 168*:867-885.

Barlow DJ, Thornton JM. 1986. The distribution of charged groups in proteins. *Biopolymers 25*:1717-1733.

Bashford D, Karplus M. 1990. p*K*'s of ionizable groups in proteins: Atomic detail from a continuum electrostatic model. *Biochemistry 29*:10219-10225.

Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rogers JR, Kennard O, Shimanouchi T, Tasumi M. 1977. The Protein Data Bank: A computer based archival file for macromolecular structures. *J Mol Biol 112*:535-542.

Boberg J, Salakoski T, Vihinen M. 1992. Selection of a representative set of structures from Brookhaven Protein Data Bank. *Proteins Struct Funct Genet 14*:265-276.

Bode W, Karshikov A. 1993. The electrostatic properties of thrombin: Importance for structural stabilization and ligand binding. *Semin Thromb Hemostasis 19*:334-343.

Bode W, Turk D, Karshikov A. 1992. The refined 1.9 Å X-ray crystal structure of D-PheProArg chloromethylketone-inhibited human α-thrombin. Structure analysis, overall structure, electrostatic properties, detailed active-site geometry, structure-function relationship. *Protein Sci 1*:426-471.

Bryant SH, Lawrence CE. 1991. The frequency of ion-pair substructures in proteins is quantitatively related to electrostatic potential: A statistical model for nonbonded interactions. *Proteins Struct Funct Genet 9*:108-119.

Devereux J, Haeberli P, Smithies O. 1984. A comprehensive set of sequence analysis programs for VAX. *Nucleic Acids Res 12*:387-395.

Dill KA. 1990. Dominant forces in protein folding. *Biochemistry 29*:7133-7155.

Fersht AR. 1972. Conformational equilibria in α- and δ-chymotrypsin. The energetics and importance of the salt bridge. *J Mol Biol 64*:497-509.

Gilson MK, Sharp KA, Honig BH. 1987. Calculating the electrostatic potential of molecules in solution: Method and error assessment. *J Comput Chem 9*:327-335.

Hingerty BE, Ritchie RH, Ferrel TL, Turner JE. 1985. Dielectric effects in biopolymers: The theory of ionic saturation revisited. *Biopolymers 24*:427-439.

Kabsch W, Sander C. 1983. Dictionary of protein secondary structure: Pattern recognition of hydrogen bonded and geometrical features. *Biopolymers 22*:2577-2637.

Karshikov A, Engh R, Bode W, Atanasov B. 1989. Electrostatic interactions in proteins: Calculation of the electrostatic term of free energy and the electrostatic potential field. *Eur Biophys J 17*:287-297.

King G, Lee SL, Warshel A. 1991. Microscopic simulations of macroscopic dielectric constants of solvated proteins. *J Phys Chem 95*:4366-4377.

Kirkwood JG. 1934. Theory of solution of molecules containing widely separated charges with special application to zwitterions. *J Chem Phys 2*:351-361.

Lee B, Richards FM. 1971. The interpretation of protein structures: Estimation of static accessibility. *J Mol Biol 55*:379-400.

Lee F, Chu ZT, Warshel A. 1993. Microscopic and semimicroscopic calculations of electrostatic energies in proteins by POLARIS and ENZYMIX programs. *J Comput Chem 14*:161-185.

Lesk AM. 1991. *Protein architecture. A practical approach*. Oxford, UK: Oxford University Press.

Levitt M, Chothia C. 1976. Structural patterns in globular proteins. *Nature 261*:552-558.

Lipman DJ, Pearson WR. 1985. Rapid and sensitive protein similarity searches. *Science 227*:1435-1441.

Matthew JB, Hanania GIH, Gurd FRN. 1979. Electrostatic effects in hemoglobin: Hydrogen ion equilibria in human deoxy- and oxyhemoglobin A. *Biochemistry 18*:1919-1928.

Meiering EM, Serrano L, Fersht AR. 1992. Effect of active site residues in barnase on activity and stability. *J Mol Biol 225*:585-589.

Nakashima H, Nishikava K, Ooi T. 1986. The folding type of a protein is relevant to the amino acid composition. *J Biochem 39*:153-162.

Perutz MF. 1978. Electrostatic effects in proteins. *Science 201*:1187-1191.

Perutz MF, Raidt H. 1975. Stereochemical basis of heat stability in bacterial ferredoxin and in haemoglobin A2. *Nature 255*:256–259.

Ponnuswamy PK. 1993. Hydrophobic characteristics of folded proteins. *Prog Biophys Mol Biol 59*:57–103.

Shire SJ, Hanania GIH, Gurd FRN. 1974. Electrostatic effect in myoglobin: H ion equilibria in sperm whale ferrimyoglobin. *Biochemistry 13*:2967–2974.

Spassov VZ, Atanasov BP. 1994. Spatial optimisation of electrostatic, interactions between the ionised groups in globular proteins. *Proteins Struct Funct Genet*. Forthcoming.

Spassov VZ, Karshikov AD, Atanasov BP. 1989. Electrostatic interactions in proteins. A theoretical analysis of lysozyme ionization. *Biochim Biophys Acta 999*:1–6.

Tanford CH. 1957. The location of electrostatic charges in Kirkwood's model of ionoorganic ions. *J Am Chem Soc 79*:5348–5352.

Tanford CH, Kirkwood JG. 1957. Theory of protein titration curves. General equations for impenetrable spheres. *J Am Chem Soc 79*:5333–5339.

Tanford CH, Roxby R. 1972. Interpretation of protein titration curves. Application to lysozyme. *Biochemistry 11*:2192–2198.

Warshel A, Russell ST. 1984. Calculation of electrostatic interactions in biological systems and in solution. *Q Rev Biophys 17*:283–422.

Warshel A, Russell ST, Churg AK. 1984. Macroscopic models for studies of electrostatic interactions in proteins: Limitations and applicability. *Proc Natl Acad Sci USA 81*:4785–4789.

Warwicker J, Watson HC. 1982. Calculation of electric potential in the active site cleft due to α-helix dipoles. *J Mol Biol 157*:671–679.

Yang A, Gunner MR, Sampogna R, Sharp K, Honig B. 1993. On the calculation of pK_a's in proteins. *Proteins Struct Funct Genet 15*:252–265.

You TJ, Harvey SC. 1993. Finite elements approach to the electrostatics of macromolecules with arbitrary geometry. *J Comput Chem 14*:484–501.