Proc. Natl. Acad. Sci. USA
Vol. 93, pp. 8350–8355, August 1996
Biophysics

Clusters of charged residues in protein three-dimensional structures

ZHAN-YANG ZHU AND SAMUEL KARLIN†

Department of Mathematics, Stanford University, Stanford, CA 94305-2125

Contributed by Samuel Karlin, April 5, 1996

ABSTRACT **Statistically significant charge clusters (basic, acidic, or of mixed charge) in tertiary protein structures are identified by new methods from a large representative collection of protein structures. About 10% of protein structures show at least one charge cluster, mostly of mixed type involving about equally anionic and cationic residues. Positive charge clusters are very rare. Negative (or histidine-acidic) charge clusters often coordinate calcium, or magnesium or zinc ions [e.g., thermolysin (PDB code: 3tln), mannose-binding protein (2msb), aminopeptidase (1amp)]. Mixed-charge clusters are prominent at interchain contacts where they stabilize quaternary protein formation [e.g., glutathione *S*-transferase (2gst), catalase (8act), and fructose-1,6-bisphosphate aldolase (1fba)]. They are also involved in protein–protein interaction and in substrate binding. For example, the mixed-charge cluster of aspartate carbamoyltransferase (8atc) envelops the aspartate carbonyl substrate in a flexible manner (alternating tense and relaxed states) where charge associations can vary from weak to strong. Other proteins with charge clusters include the P450 cytochrome family (BM-3, Terp, Cam), several flavocytochromes, neuraminidase, hemagglutinin, the photosynthetic reaction center, and annexin. In each case in Table 2 we discuss the possible role of the charge clusters with respect to protein structure and function.**

We apply the methods described in our companion paper (1) with the objective to identify statistically significant charge clusters in protein three-dimensional (3D) structures and to suggest correlations with protein structure and function. A linear sequence or 3D-charge cluster signifies an anomalous distribution of the charged residues in a protein constituting one or more regions with excessive concentrations of charge relative to the overall charge composition of the protein.

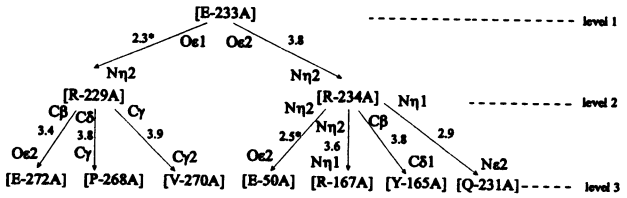
Distinctive charge configurations in a protein may contribute to function and structure in diverse ways. For example, runs of positively or negatively charged amino acids in a protein sequence may be stretched out and exposed structurally or may be arranged to coordinate metal ions, e.g., acidic residues binding to Ca^{2+}. Charge patterns of period two in a β-strand may present a straight line of charge on one side of the strand and those of period three or four in an α-helix present an almost linear path of charge. Electrostatic attractions between charge clusters of unlike or mixed sign may contribute to the formation of multidomain complexes, whereas charge clusters of like sign may help maintain separation between certain protein assemblages. Charge clusters can increase the solubility of the protein in aqueous media. A general function of clusters of charge may be to establish and stabilize protein conformation. Multiple charge clusters within one protein may facilitate intramolecular folding or protein–protein or protein–nucleic acid interactions. Charge clusters and runs appear to be important with respect to protein transport, localization, and regulatory function (2–6). In eukaryotes, charge clusters in linear sequences are associated with transcriptional activation, developmental control, and regulation of membrane receptor activity, and are generally lacking in cytoplasmic enzymes and housekeeping proteins (6).

All 3D-charge clusters at the significance level $P^* \leq 0.01$ from a large nonredundant representative Protein Data Bank (PDB) structure set were determined by Methods 1–3 (M1–M3) described in (1) and M4 described below. Our methods are fundamentally different from those using charge potential calculations (7) as a means to discern charge clusters. For example, mixed-charge clusters would generally appear neutral in the charge potential plot. Moreover, a highly biased charged protein (acidic or basic) would likely not distinguish clusters. The large majority of the charge clusters reported in Table 2 appear to be unknown and could not be easily identified by visual inspection. In this paper, we report the statistics on the 3D-charge clusters in Table 1 and compare it with corresponding statistics of bulk linear protein sequences. We comment on the examples of the charge clusters of Table 2 and venture some interpretations on their possible functions.

METHODS

M1–M3 (see ref. 1) were used with both d_m- and D_m-distance measures. Another method (M4), based on representing the protein structure by a collection of trees, goes as follows. For each residue a_k, a tree S_k is produced with the residue a_k as the root (the first level). The jth level ($j > 1$) consists of the residues within a prescribed distance, d_m or D_m, of all residues at the $(j-1)$th level, which do not occur at any of the preceding levels. Generally, we use distance of 4.0 Å.

An example of three levels of the tree (M4–d_m: M4 and d_m-distance) for the structure of aspartate transcarbamylase (PDB code: 8atc) is presented next. Here we indicated the d_m-distances (in angstroms) and atoms at which the residue distances are attained. The inter sidechain distances between atoms of unlike sign marked with an asterisk portends a salt bridge contact. [E-233A] refers to glutamate at position 233 of chain A.



In the analysis of significance by M1–M3, we use the theory of high-scoring segments (1). For M4, we use a binomial counting model (5). Let the desired charge type have frequency f in the whole protein structure. Observing the count c of the desired residue charge type in a set of w residues, a significant cluster occurs if $t = (c - wf)/\sqrt{wf(1-f)} > 4$. The significance test is checked on all residue sets consisting of the residues from the first level to any other level of tree S_k having $w \geq 20$. The lower bound 20 helps in assuring the Gaussian distribution approximation in the random model. The foregoing statistical procedure is applied to every tree generated from each residue. With respect to positive and negative charge clusters, the count c is conservatively taken to be the net count of positive minus negative or negative minus positive charges, respectively. Non-amino acid entities of the structure appearing in the trees are not counted in the statistical evaluations.

Two significant clusters of the same type are considered to be the same if in the smaller cluster at least 50% of the residues are contained in the larger cluster. The cluster with the greater statistical significance evaluation is retained.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: 3D, three-dimensional; PDB, Protein Data Bank; GST, glutathione *S*-transferase.

†To whom reprint requests should be addressed.

RESULTS, EXAMPLES, AND DISCUSSION

Number of Charge Clusters in Protein 3D Structures. For comparative purposes, it is useful to review briefly the statistics on occurrences of charge clusters in linear sequences. Among protein primary sequences (currently more than 40,000 nonredundant protein sequences available), the percentage of proteins with at least one significant linear charge cluster is about 20%–25% in most eukaryotic species (e.g., human, mouse, yeast), about 35% in *Drosophila*, and about 6%–8% in *Escherichia coli* (and most prokaryotes) (Table 1). The higher percentage in *Drosophila* sequences is probably due to the bias towards developmental regulatory proteins in the *Drosophila* sequence collection. Proteins with multiple charge clusters in primary sequences are uncommon, about 3.5% among human sequences. Principal families of proteins with multiple linear charge clusters feature developmental regulatory proteins, voltage-gated ion channel proteins, and major regulatory proteins of large eukaryotic DNA viruses (6).

With respect to 3D-charge clusters of statistical significance level $P^* \leq 0.01$, there is a scarcity of positive charged clusters. There is only one positive charge cluster ($P^* \leq 0.005$) found in the photosynthetic reaction center structure (PDB code: 1prc), whereas there are only three acidic 3D clusters among all proteins of the representative structure set (Table 1). These markedly contrast with the fractions 4.5% of positively charge clusters and 6% of negatively charged clusters among linear protein sequences of humans. Notably the percent of protein structures showing multiple 3D charge clusters (3.8%) is approximately the same as the percent of multiple charge clusters in general linear protein eukaryotic sequences (about 3.5%). In aggregate, about 10.2% of protein structures show at least one 3D-charge cluster (Table 1). The percent of charge clusters among linear sequences of 3D structures is only about 2.7% (8). Most 3D-charge clusters are of mixed type involving about equally anionic and cationic residues.

The primary sequences of proteins whose 3D structures are in the protein PDB database manifest a near absence of unusual charge distributional properties (Table 1). Moreover, the few occurrences of these properties in the structure set are much less extreme than the most dramatic examples compiled from the Swiss–Prot protein database (see ref. 8). The current PDB set consists predominantly of small soluble globular proteins that lack, both in frequency and extent, statistically significant sequence features that characterize many types of proteins from the Swiss–Prot sequence collection. The paucity of linear charge clusters in the PDB set is probably associated with the scarcity of exposed flexible domains, linker regions, and free termini in these protein structures. Proteins with such features are generally difficult to crystallize (9).

Some Structural Features of 3D-Charge Clusters. The average side-chain solvent accessibility of residues (as assessed by the methods in ref. 10) exceeds 30% in most of the charge clusters. This indicates that a charge cluster is mostly on the surface of a protein structure. By contrast, charge clusters at the interface of two chains tend to be inaccessible to solvent, e.g., glutathione *S*-transferase (GST) (see below). Charge clusters in loops are frequent but not predominant. This suggests that some charge clusters play a role in packing of secondary structures and/or in linking separate modules. The residues in a charge cluster usually derive from separate regions of the primary sequence (segments of the primary sequence separated by at least 10 residues are considered separate) or from 2 or more chains.

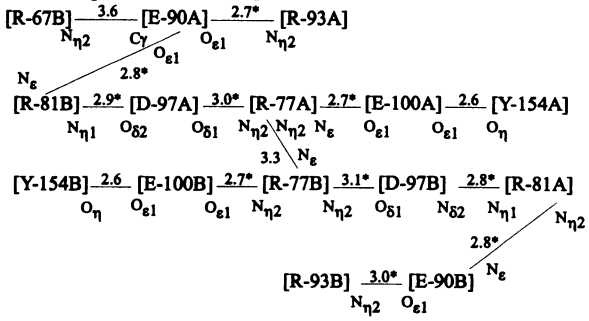
Table 1. The percentage of protein 3D structures with charge clusters

Type of clusters	Protein 3D structures*	Primary sequence†		
		PDB	General database	
			Human	*E. coli*
Single positive	0.0% (0.0%)	0.4%	4.5%	1.0%
Single negative	0.5% (0.0%)	0.4%	6.0%	1.0%
Single mixed	5.9% (2.2%)	1.9%	10.5%	5.0%
Multiple	3.8% (2.1%)	0.0%	3.5%	0.5%
Total	10.2% (4.3%)	2.7%	24.5%	7.5%

*The protein structures examined in the analysis (total of 186) have the sequence identity less than 30%. For a complete list contact Z.-Y.Z. Data in parentheses indicate the percentage of protein 3D structures with clusters involving two or more chains.
†Data are obtained from ref. 8.

Mixed-Charge Clusters and Quarternary Structures. Protein–protein interfaces are generally made up of a mixture of hydrophobic and hydrophilic residues. In many cases, charge interactions over relatively long ranges could orient and position the appropriate protein surfaces between which hydrophobic forces form the preponderance of interface bonds (11, 12). Examples of mixed-charge clusters prominent in interchain contacts, which may stabilize quaternary protein formation or in protein complexes putatively facilitated by electrostatic interactions, include GST (see below), mosaic virus coat protein (4sbv), catalase (8cat), and fructose-1,6-bisphosphate aldolase (1fba) (see Table 2).

The GST are a group of enzymes that play a critical role in detoxification processes of mutagens, carcinogens, and other noxious chemical substances. GST (rat liver) (2gst) occurs as a homodimer that contains one active site in each monomer. The structure exhibits a mixed-charge cluster that apparently mediates dimer formation (of chains A and B) (Fig. 1). The mixed-charge cluster ascertained by the M1–d_m protocol, significance level $P^* = 0.007$, corresponds to the sequence now displayed.



This cluster combines the charged residues 2E, 1D, and 3R from the A chain with 2E, 1D, and 4R from the B chain. The noncharged residues in this cluster are Y-154A and Y-154B, connected by hydrogen bonds with E-100A and E-100B, respectively. The pairings [R-81A] with [E-90B], [R-81A] with [D-97B], [R-81B] with [E-90A], and [R-81B] with [D-97A] establish interchain salt bridge connections. The cluster also shows three salt bridge ties strictly within the A chain ([R-93A] with [E-90A], [D-97A] with [R-77A], [E-100A] with [R-77A]), and correspondingly within the B chain.

No statistically significant charge cluster is discerned in the analysis of the 3D chains A and B separately attesting to the

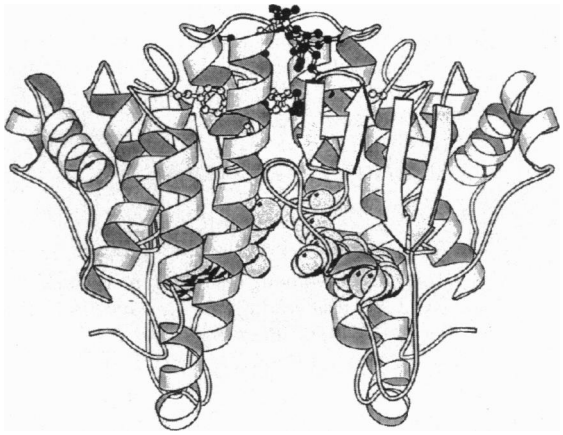


FIG. 1. In the GST structure (2gst), a mixed-charge cluster is identified at its dimer interface. Residues in ball and stick form correspond to a mixed-charge cluster; basic residues are solid, acidic residues open, and the others shaded. Several residues of the charge cluster are buried. Atoms in large spheres indicate 9-(*S*-glutathionyl)-10-hydroxy-9,10-dihydrophenanthrene groups.

Table 2. Charge clusters in protein 3D structures

Example	
1	(2msb) mannose-binding protein A-lectin domain (rat) complex with calcium and glycoprotein; see ref. 1.
2	(1 atp) cAMP-dependent protein kinase (catalytic subunit) (mouse)—336 residues in chain E and 20 residues in chain 1—the peptide inhibitor; frequency of D, E, K, R: 27.0%.
	mixed/M2–d_m; P^* = 0.007*, 32† — Chain E: (***R***-45, ***K***-47, *t*-48, ***R***-56); (**E**-127, **f**-129, **R**-133); (**E**-230); (D-328, D-329, y-330, E-332, E-333, E-334, R-336); Chain I: (R-18); seed R-45E.
3	(2hpd) cytochrome P450-BM-3, hemoprotein domain (*E. coli*)—457 residues in chain A; 457 residues in chain B; frequency of D, E, K, R: 27.4%.
	mixed/M2–d_m; P^* = 0.001; 34.3† — (K-3, m-5); (E-38, *i*-39); (***R***-50, **R**-56); (***D***-68, l-71, **q**-73, **K**-76, **f**-77, **R**-79, **D**-80, D-84, s-89, w-90, t-91, h-92, E-93, **K**-94, **n**-95, **w**-96, **K**-97, **K**-98, **n**-101); (**R**-179, **E**-183, **K**-187, R-190, **v**-198, **E**-200, **n**-201, **R**-203, **q**-204, **f**-205, **E**-207, **D**-208); (D-242, E-244, t-245, E-247); (*y*-334, K-336, E-337, D-338, E-344, *y*-345, E-348, K-349, D-351); (R-398); (all in chain A); seed E-38.
4	(lala) annexin V (chicken)—316 residues in a single chain; frequency of D, E, K, R: 29.5%.
	mixed/M2–D_m; P^* = 0.003; 32.4† — (R-6, D-16, **R**-18, **a**-19, **D**-20, **E**-22, **R**-25, **K**-26, **m**-28, K-29); (R-45, **R**-50, **K**-58, D-64, **v**-66, **D**-67, **D**-68, **K**-70, **s**-71, **E**-72, **K**-76, **l**-80, **l**-84); (**D**-266, R-276, D-280, **l**-281, **D**-283, **R**-285); (**y**-308, **c**-316, **l**-312); seed D-283.
5	(lipd) 3-isopropylmalate dehydrogenase (*Thermus thermophilus*)–345 residues in a single chain frequency of D, E, K, R: 24.9%.
	mixed/M2–D_m; P^* = 0.008; 34.2† — (**K**-159, **p**-160, **E**-163, **R**-164, **R**-167, **f**-170, **E**-171, **a**-173, **R**-174, R-176, K-178); (**E**-201, R-204, D-208); (D-231); seed R-204.
6	(2trx) thioredoxin (*E. coli*)—108 residues in chain A and 108 residues in chain B; frequency of D, E, K, R: 25%.
	mixed/M2–D_m; P^* = 0.005; 32.4† — (D-2, K-3); (**D**-43, **E**-44, **D**-47, **E**-48); (**K**-96, **K**-100, **D**-104); (all in chain A); seed D-2A.
7	(2gst) glutathione *S*-transferase (rat liver); see text.
8	(2por) integral membrane protein porin (*Rhodobacter capsulatus*)—301 residues in a single chain; frequency of D, E, K, R: 22.6%
	mixed/M1–d_m; P^* = 0.009; 34.8† — (***D***-7, ***R***-9); (***R***-24, ***R***-26); ***K***-46, **h**-48, **E**-49, D-58); (D-74); (E-109); seed D-7.
9	(4xis) xylose isomerase (*Streptomyces rubiginosus*) complex with xylose and $MnCl_2$—387 residues in a single chain; frequency of D, E, K, R: 28.7%.
	mixed/M2–D_m; P^* = 0.006; 62.9† — (D-35, **E**-38, **R**-41, **R**-42); (**D**-65, **D**-69, **K**-73, **q**-77, **D**-81); (**R**-109, **R**-113, **R**-117, **R**-121); seed R-109.
10	(8atc) aspartate carbamoyltransferase (*E. coli*—310 residues in chain A and C, 146 residues in chains B and D, respectively; frequency of D, E, K, R: 24.7%.
	mixed/M4–d_m; t = 4.17; 18.4† — (**E**-50, **R**-54); (***R***-105); (**g**-128, h-134); (D-162, K-164, y-165, **R**-167); (**l**-192); (R-229, q-231, E-233, R-234, l-235, **E**-239); (h-265, p-268, R-269, v-270, D-271, E-272,); (all in chain A); seed E-233A.
11	(8cat) catalase (beef liver)—498 residues in chain A and 498 residues in chain B; frequency of D, E, K, R: 23.4%.
	mixed/M2–d_m; P^* = 0.004; 14.5† — Chain A: (**K**-76); (E-118, g-120, D-123); (R-169); (D-206, **R**-209, **h**-210); (K-242); (D-258, **y**-259, **g**-260, **D**-263); Chain B: (K-76); (E-118, g-120, D-123); (R-209); (D-258, **g**-260, **D**-263); seed R-209B.
12	(lfcb) flavocytochrome b_2 (yeast)—494 residues in chains A and B, respectively; frequency of D, E, K, R: 27.8%.
	mixed/M2–d_m; P^* = 0.009; 51.3† — Chain A: (D-173, R-175); Chain B: (D-258, R-259, K-260, **D**-263, **D**-264); (D-327, s-329, t-331, **K**-333, **D**-334, **E**-337, **K**-340, K-341); seed D-258B.
13	(lnn2) neuraminidase (influenza virus)–388 residues in a single chain; frequency of D, E, K, R: 17.1%.
	mixed/M4–d_m; P^* = 4.60; 12.5† — (***R***-118, ***E***-119); (***E***-276, ***E***-277); (***R***-292, n-294, s-298, n-299, R-300); (l-321, v-322, t-325, R-327, D-329, D-330); (R-344, g-348, v-349, K-350); (***m***-362, ***R***-364, t-365, K-368, D-369, R-371, ***E***-375, *f*-377); (s-404, y-406); (***E***-425, *i*-427); seed T-365.
14	(lhge) hemagglutinin (influenza virus)—328 residues in chains A, C, and E and 175 residues in chains B, D, and F, respectively; frequency of D, E, K, R: 22.4%.
	mixed/M1–d_m (M1–D_m, M2–D_m); P^* = 0.004; 17.3† — Chain B: (**R**-127, **E**-128, ***E***-131, y-141); (**R**-163, **R**-170, i-173, K-174); Chain D: (**R**-127, **E**-128, ***E***-131); (**R**-163, **D**-164, **R**-170); Chain F: (**R**-127, **E**-128, ***E***-131, y-141); (**R**-163, **R**-170); seed D-164D.
	mixed/M4–d_m; t = 4.50; 23.9† — Chain A: (***E***-89); (D-104, **a**-106, **s**-107, **R**-109); (***I***-163, ***R***-269); Chain B: (E-67, f-74, g-75, **R**-76, **i**-77, **n**-78, **E**-81, **K**-82); Chain C: (***E***-89); (D-104, **a**-106, **s**-107, **R**-109); (*i*-163, ***R***-269); Chain D: (E-67, f-74, **i**-77, **n**-78, **E**-81, **K**-82); and Chain F: (g-75, **R**-76, **i**-77); seed I-77F.
15	(1prc) photosynthetic reaction center (*Rhodopseudomonas viridis*)—323 residues in chains C and M; 273 residues in chain L and 258 residues in chain H, respectively; frequency of R, K: 8.3%; frequency of D, E, K, R: 16.7%.
	Positive/M1–D_m; P^* = 0.009; 49.6† — Chain C: (K-198, R-199); (R-275, R-272); Chain M: (K-323); seed K-198C.
	mixed/M4–d_m (M3–d_m) t = 6.95; 17.0† — Chain H: (g-40); (R-82, E-84, R-86, l-88); (a-111, s-116, y-117, E-119, R-120, a-121, E-122, v-123, D-125, a-132, K-133, *i*-134, *v*-135, *p*-136, l-137, R-138, v-139, a-140); (R-153); (*t*-169, *D*-170, *w*-172, *D*-174, R-175, s-176, E-177, *y*-179, ***R***-181, *y*-182, *E*-184, *t*-193, *l*-195); (l-224, q-225, R-227, D-228, i-230, t-231, **l**-232, **R**-233, **E**-234, **E**-235, **D**-236. **K**-237); Chain M: (D-230, R-231, **E**-234, D-238, R-239, **E**-244, **l**-248,) Chain L: (l-3, f-5, **K**-8, **y**-9); seed t-231H.

(*Table 2 continues on the opposite page.*)

Table 2. (*Continued*) Charge clusters in protein 3D structures

Example		
16	(1fba) fructose-1,6-bisphosphate aldolase (*Drosophila melanogaster*)—360 residues in chains A, B, C, and D, respectively; frequency of D, E, K, R: 23.3%.	
	mixed/M2–d_m; P^* = 0.001; 37.1†	Chain B: (y-6, s-8, **K**-9, **E**-10, **I**-11, **D**-13, **E**-14, **R**-16); (**K**-138, *D*-143); (R-181); (y-222, **E**-224); Chain C: (D-197, **D**-199, **R**-200, **K**-203, **E**-206); (**R**-258); seed D-143B.
	mixed/M2–d_m; P^* = 0.006; 37.9†	Chain A: (y-6, s-8, **K**-9, **E**-10, **I**-11, **D**-13, **E**-14, **R**-16, **E**-17, **K**-21); (*i*-103); (**K**-138, *D*-143); (R-181); (y-222, **E**-224); Chain D: (**D**-199, **K**-203, *E*-206); (*R*-258); seed D-199D.
17	(3chy) signal transduction protein CheY (*E. coli*)—128 residues in a single chain; frequency of D, E: 14.9%, D, E, R, K: 26.6%.	
	negative/M2–D_m; P^* = 0.007; 34.0†	(D-3, E-5, *f*-8, *v*-10, D-12, D-13); (n-32, *E*-34, *a*-36, D-38, E-37, **D**-41); seed D-3.
	mixed/M2–D_m; P^* = 0.006; 29.2†	(D-12, D-13, R-18, **R**-22, **K**-26, *E*-35, E-37, D-38, **D**-41); seed K-26.
18	(8acn) aconitase complex with nitroisocitrate (bovine heart)—753 residues in a single chain; see ref. 1.	
19	(3tln) thermolysin (*Bacillus thermoproteolyticus*)–316 residues in a single chain; see ref. 1.	
20	(1min) nitrogenase molybdenum–iron protein; frequency of D, E, R, K: 25.9%; see ref. 1 for the other cluster.	
	mixed/M4–d_m (M3–d_m); P^* = 5.77; 14.6†	Chain A: t-104; Chain B: (**y**-12); (*f*-441, **K**-449, **q**-452, **R**-453, **D**-454, **I**-456, **K**-460, **E**-463, *l*-466, *R*-468, D-475, R-476); (**R**-504, **D**-506, **E**-507, **E**-508, R-510); Chain D: (**q**-452, **E**-453, **I**-456, **K**-460, E-463); (**E**-503, **R**-504, **D**-506, **E**-507, **E**-508, t-509, R-510, m-512, **D**-516, h-519, R-523); seed E-507D.

This table presents all charge clusters identified from a collection of 186 nonredundant representative structures (Table 1). P^* is the significance probability for the given charge cluster, i.e., the probability of observing such a cluster in a corresponding random protein structure of the same amino acid composition (1); see *Methods* for the definition of *t*. †The average side chain accessibility of the residues occurring in the charge cluster (10). The residues are arranged in groups determined by their proximity in the primary sequence. Groups are defined provided all mutual residues of the two groups are separated by at least 10 positions in the linear sequence. The average number of groups per charge cluster is 3–4. Residues are formatted as follows: bold underlined letter (i.e., **X**) refers to a charge residue in an α-helix; bold italic letter (i.e., ***X***) in a β-strand; ordinary cap letter (i.e., X) in a loop; lowercase letter for corresponding noncharge residues. The seed gives the residue from which a sequence that contains the charge cluster can be generated by the appropriate M1, M2, M3, or M4 method.

importance of charge interactions across the interface of the two chains.

Another example is the coat viral protein (southern bean mosaic virus) (4sbv). The virus particle has a shell of 180 protein subunits arranged with 3-fold icosohedral symmetry around a core of RNA. There are three rather symmetrically situated Ca^{2+} ions, each at an interface between a pair of subunits coordinated by two aspartate residues from one unit and an asparagine residue from a separate unit. The coat protein composed of chain A (199 residues), chain B (199), and chain C (222) features a mixed charge cluster at the interface of the three chains A, B, and C (M2–d_m, $P^* < 10^{-6}$).

[K-109C] —3.0— [E-252C] —2.6— [D-188B] —2.6— [Y-137B] —3.0— [K-154B] —3.0— [D-65C] —4.1—
[M-64C] —5.3— [S-63C] —3.4— [R-61C] —4.0— [I-68C] —3.5— [R-249C] —2.6*— [E-191B] —2.7*—
[R-196C] —3.6— [K-195C] —3.1— [K-195B] —3.9— [R-196B] —2.6*— [E-191A] —2.9*— [R-249B] —3.7—
[V-66B] —3.9— [D-65B] —3.0— [K-154A] —3.5— [Y-137A] —2.5— [D-188A] —3.1— [E-252B] —2.8—
[I-251B] —3.2— [K-109B] —3.4— [S-193A] —3.6— [Y-197A] —3.3— [E-194A] —2.5— [E-194B] —2.4— [E-194C].

Here, again the individual 3D chains do not contain statistically significant charge clusters of any kind.

Functions of 3D-Charge Clusters. Example 1 of Table 2 (2msb) *mannose-binding protein* is discussed in ref. 1.

Example 2: cAMP-*dependent protein kinase* (mouse) E chain complexed with a peptide inhibitor PKI (1atp). This subunit contains two bound Mn^{2+} ions (labeled Mn-2 and Mn-3) and ATP with its two phosphate groups interacting with T-197E and S-338E. The mixed-charge cluster carries several salt bridges from the major chain: E-333 with R-56 (at about 2.3 Å distance between the sidechain atoms $O_{\varepsilon 1}$ and N_ζ, respectively), R-18I (position 18 of the inhibitor peptide) with E-127 (at about 2.8 Å distance between the sidechain atoms N_ε and $O_{\varepsilon 2}$, respectively), and R-133 with E-230 (at about 2.3 Å between the atoms $N_{\zeta 2}$ and $O_{\varepsilon 2}$, respectively). Both Mn^{2+} ions bind to ATP. The major ligands to Mn-3 are N-171 (2.09 Å) and D-168 (2.14 Å). Two lysine residues, K-168 (at d_m-distance 2.44 Å) and K-72 (2.52 Å) and glutamate E-127 (2.65 Å) of the charge cluster bind ATP. Thus, the mixed-charge cluster is paramount at the catalytic site.

Example 3: Cytochrome P450-BM-3 (fatty acid monooxygenase) (2hpd). The protein binds O_2 and subsequently splits off one oxygen, which is transferred to a nonspecific substrate (e.g., fatty acid). The axial ligand of the heme of each unit is a cysteine thiolate (13). Other close residues to the heme are R-398 (2.84 Å) and W-96 (2.85 Å distance), which are part of the charge cluster. Most of the mixed-charge cluster are generally far from the heme in an exposed location (average solvent accessibility value 34.3%), which suggests a role in protein–protein interactions—e.g., possibly docking to the required reductase essential for electron transfer to the heme center. The three structures of the P450 family (BM-3 in Table 2, Terp and Cam not shown) all carry an exposed mixed-charge cluster distant from the heme site. It is known that P450-Terp possesses an iron–sulfur partner, the ferrodoxin (2Fe—2S) of *Pseudomonas putida* that participates in the redox activity (14). We speculate that the charge cluster modulates the requisite interactions of the two proteins. Similarly, P450-Cam possesses the iron–sulfur partner putidaredoxin.

Example 4: Annexin V (calcium/phospholipid-binding protein) (1ala). Although its active quaternary structure is uncertain, annexin (1ala) seems to have many functions, including movement of vesicles about the cell, secretary regulation, and formation of calcium channels (15, 16). It is known that annexin XII performs as a hexamer (two trimers on top of each other) with multiple Ca^{2+} ions at the interface of the trimers (15). Under conditions of low calcium concentrations, annexin V is principally cytosylic and occurs as a monomer, whereas in the presence of enough calcium ions it binds to phospholipids at the membrane (15). Each unit generally carries three Ca^{2+} ions each coordinated by a single acidic residue and several carbonyl backbone atoms. The mixed-charge cluster intersects two domains of the monomer. The exposed nature of the charge cluster is suggestive of some kind of protein–protein interaction.

Example 5: 3-Isopropylmalate dehydrogenase (IPMDH) (1ipd). The enzyme of *IPMDH* is bifunctional and different from the mono-functional dehydrogenases [e.g., 6ldh (17)] as characterized by x-ray analysis. IPMDH catalyzes substrate decarboxylation simultaneously with dehydrogenation. The two enzymes differ substantially in both their primary sequences and 3D structures (18, 19). A mixed-charge cluster (ascertained by M2–D_m) consists mainly of residues from the solvent exposed side of helix 158–175 (helix e as labeled in ref. 18), the C terminus of helix 190–204 (helix f), and from residues in neighboring loops. The cluster is substantially exposed as assessed by the average side-chain accessibility value 33.2%. Hurley and Dean (20) determined nicotinamide adenine dinucleotide-binding sites in the structure of IPMDH. None of the residues in the mixed-charge cluster is among them. However, the charge cluster may be related to the thermostability and activity of the enzyme at extremely high

temperature. The residues in the cluster show five salt bridge contacts, two side-chain hydrogen bondings, and two hydrogen bonds between the side-chains of residues in the cluster and other residues. The side-chains of the residues in the cluster also form eight hydrogen bonds with water molecules. These hydrophilic interactions of the charge cluster coupled to hydrophobic interactions in its dimer interface may be important factors underlying the high thermostability of the enzyme (21–23).

Example 6: Thioredoxin (2trx). These proteins are ubiquitious and are involved in a variety of cellular redox functions involving reversible disulfide bonding (oxidation of its active center dithiol to a disulfide). Thioredoxins also activate the bacteriophage T7 polymerase by an unknown mechanism (24, 25). It is thought to provide some particular structural environment without which T7 polymerase is inactive (24). Is the mixed-charge cluster vital to this environment?

Example 7: GST (2gst). This is discussed above.

Example 8: Porin (2por). This protein forms a channel surrounded by a 16-strand β-barrel structure that allows the diffusion of small polar noncharged metabolites in and out of the *E. coli* cell. Ions tend to be inhibited in this flow. The side chains of the mixed-charge cluster feature anionic residues on one side and cationic residues on the other side, all facing into the interior of the channel. These presumably help to manipulate metabolites in and out. The charge cluster D-7, R-26, R-24, R-9, K-46, H-48, E-49, D-58, D-74, E-109 clearly displays spatially (apart from seed D-7) a set of positively charged residues followed by a set of negatively charged residues.

Example 9. Xylose isomerase (4xis). Involved in D-xylose catabolism, Mg dependent. The particular crystal structure available contains two Mn^{2+} ions. The charge cluster is largely exposed, e.g., note the amphipathic helix with D-65, E-69, K-73, Q-77, D-81 on the solvent exposed side and a positive charge amphipathic helix featuring R-109, R-113, R-117, R-121, in both cases showing the period 4-linear spacings.

Example 10: Aspartate carbamoyltransferase (ATCase) (8atc). ATCase catalyzes the formation of carbamyl aspartate from carbamyl phosphate and aspartate. Carbamyl aspartate occurs only in the pyrimidine biosynthetic pathway. In this context, ATCase synthesis is suppressed by extra CTP and UTP products and activated by a shortage of ATP presumably to balance pyrimidine and purine content (ref. 25, pp. 448–451). The complete enzyme is composed of two catalytic trimers and three regulatory dimers. A zinc metal ion coordinated by four cysteines is capable of bridging the catalytic and regulatory units. This is interpreted as signifying a structural role for the zinc ions (25). The active site is situated between two domains of adjacent polypeptide chains in the trimers. The charge cluster envelops the aspartate substrate domain and the carbamyl phosphate domain bound to the active site (25). The inherently variant charge interactions (charge associations can vary from weak to strong) putatively allow local conformational changes of the catalytic site (alternating tense and relaxed states) resulting from allosteric attachment and release of effecters of ATP and CTP.

Example 11: Catalase (8cat) (protective metallo-enzyme) A and B chain (498 residues each). This enzyme is ubiquitous in aerobically respiratory organisms and serves to protect cells from the toxic effects of H_2O_2 by affecting its dismutation to water and O_2. The active structure of catalase is a homotetramer. The crystal structure available is that of a dimer. Each unit contains a heme group (with axial tyrosine ligand) and an nicotinamide-adenine dinucleotide phosphate cofactor. The mixed-charge cluster involves contributions from both chains that form both intra- and interchain salt bridges and hydrogen bonds. It appears that the charge cluster mediates the formation of the interface between A and B chains and/or stabilizes quaternary structure. The heme regions are distant from the charge cluster.

Example 12: Flavocytochrome b_2 (1fcb). Cellular location: mitochondrial intermembrane space. This protein is a homotetramer that binds a flavin mononucleotide prosthetic group (26). The cytochrome b_2 core contains the heme-binding region. Chain A contains a heme and flavin mononucleotide, whereas chain B contains a flavin mononucleotide and pyruvic acid. The mixed-charge cluster is totally exposed to solvent (average side-chain accessibility 51.3%). It involves 13 residues of chain B and two residues of chain A. It includes two salt bridges D-263B connecting to K-341B at 2.67 Å and D-334B connecting to R-259B at 2.42 Å. It also includes the H-bonding of D-327B with S-329B at 2.32 Å. The *flavocytochrome* (1ltd) (not shown) of two chains also contains a mixed-charge cluster composed of residues from both chains distant from the heme site and the flavin mononucleotide prosthetic group. There is no 5% significant charge cluster in each chain separately.

Examples 13 and 14: Neuraminidase (1nn2) and *hemagglutinin* (1hge). Neuraminidase and hemagglutinin are the two integral membrane glycoproteins in the influenza viral surface. Neuraminidase cleaves sialic acid from these glycoconjugates, thereby liberating the influenza virus to allow it to spread the infection to new host cells (for review, see ref. 27). A mixed-charge cluster is identified in the structure. Nine of 31 residues in the cluster (including 6 charge residues) are associated with the active site of the protein where sialic acid binds. Sixteen residues (12 charged) in the cluster are among the 54 invariant residues in the amino acid sequences of the neuraminidase N1 N2, N8, N9 subtypes of influenza A viruses and an influenza B (28). One-half of the residues in the cluster are totally buried. These residues form an extensive network of hydrogen bonds and salt bridges. Thus, they make a very stable base that is necessary for the saccharide–protein interactions (29, 30). It is observed in infected populations and in genetic studies that almost any single, uncompensated mutation will disrupt the active site and inactivate the enzyme (31).

The 3D structure of hemagglutinin shows two mixed charge clusters. The same first cluster resulted from all methods: M1–d_m, M1–D_m, and M2–D_m. Table 2 displays the cluster obtained by the M1–d_m protocol. It appears that the major function of this charge cluster is to establish and stabilize the tetramer. The second mixed-charge cluster in the protein 3D structure overlaps the second active site of the protein. This cluster contains primarily conserved residues of hemagglutinin. It has a structural environment and corresponding charge interactions similar to the charge cluster found at the active site of neuraminidase suggesting that both structures bind to sialic acid in a similar way although the primary binding sites are different (32, 33).

The influenza virus glycoproteins both show a mixed-charge cluster that cleaves and binds sialic acid from appropriate membrane receptors. Is it possible that neutralizing the mixed charge cluster could curtail interactions with sialic acid and consequently prevent viral infections?

Example 15: Photosynthetic reaction center (1prc). The photosynthetic reaction center (integral membrane protein of purple bacteria) is a large complex assembly composed of four chains H, M, and L that traverse in an α-helical conformation the lipid bilayers and a tightly bound outside cytochrome chain C. The structure contains many cofactors including bacteriochlorophylls and bacteriopheophytins. The reaction center mediates the initial photochemical event in the electron transfer process of photosynthesis (24, 25). This structure shows the only statistically significant positive charge cluster among all proteins analyzed for Table 2 with residues joined mostly by backbone hydrogen bonds. It is located exterior to the membrane region. The mixed-charge cluster involves all three chains in the cytoplasmic part of the structure. The primary sequence shows in chain L the positive sequence charge run: [R-7L], [K-8L], [Y-9L], [R-10L], [V-11L], [R-12L] and in chain H the intriguing linear charge pattern: [R-33], [R-34], [E-35], [D-36], [R-37], [R-38], [E-39]. Both are on the exterior of the lipid bilayer.

Example 16: Fructose-1,6-bisphosphate aldolase (1fba). This aldolase structure from *Drosophila melanogaster* is a homotetramer (34) where each subunit is an eight-stranded α/β-barrel. The cluster analysis (M1–dm) identifies a mixed-charge cluster for each subunit but at the reduced significance level $P^* \approx 0.05$. When the structure is studied as a tetramer these clusters extend to be highly significant ($P^* \leq 0.001$). The two charge clusters provide the electrostatic interactions to maintain the N-terminal helix of chain A and chain B, respectively, to lie at the open end of the α/β-barrel. Moreover, the two clusters stabilize the chains A and D and chains B and C, respectively, with a salt bridge and a hydrogen bond ([E-224A]-[R-258D], [Y-6A]-[E-206D] and [E-2248]-[R-258C], [Y-6B]-[E-206C]), respectively. The interface of chains A and D and the interface of chains B and C are each stabilized by 15 hydrogen bonds and salt bridges, and not by hydrophobic contacts. This is at

variance with the nature of the interactions seen in a similar structure of the rabbit skeletal muscle aldolase (35).

Example 17: Signal transduction protein CheY (3chy). This protein contains an acidic charge cluster and a mixed-charge cluster overlapping in several acidic residues. In both clusters the closest contacting atoms usually involve backbone atoms. The mixed-charge cluster emphasizes positively charged residues first and subsequently merges with the acidic cluster. The roles of the charge clusters are unknown. A homologous structure (2che) contains Mg^{2+}, which is coordinated by some of the equivalent residues in the cluster.

Examples: 18 Aconitase (8acn), 19 *thermolysin* (3tln), and 20 *nitrogenase molybdenum–iron protein* (1min). These examples are discussed in ref. 1.

Some Structural and Biological Implications of Charge Clusters. Charge clusters can contribute to local structural stability where salt bridges and hydrogen bonds in mixed-charge clusters certainly enhance conformational stability. Highly charged regions in protein structures often form novel coils, sometimes stabilized by metal ions, as in the repetitive calcium-binding β-supercoil structures found in the alkaline tail of *Pseudomonas aeruginosa* (36). Charge clusters can be involved in protein–protein interactions that mediate electron transport or in facilitating diffusion of polar solutes. Many reactions require acid–base groups in order to occur. Electrostatic interactions facilitate important processes such as protein sorting, translocation, docking, localization, orientation, oligomerization, and binding to DNA and other protein molecules. In this context, charge effects are relatively long range, rapid, localized, and flexible.

Acidic clusters often coordinate certain metal ions (1). Charge clusters can occur at the active site and contribute to its stability and enzymatic activity [e.g., substrate binding–aspartate transcarbomylase (8atc)]. Charge clusters may provide interactions more specific than those contributed by hydrophobic residues. Such specificity may be essential in substrate binding. Exposed charge clusters can mediate protein–protein (and protein–nucleic acid) interactions—e.g., in docking. Hemaglutanin of influenza virus infects the human host cells by binding receptors that contain sialic acid. Putatively, the mixed-charge cluster of this structure enhances this binding activity. Neuraminidase at its mixed-charge cluster cleaves sialic acid to free the influenza virus particle. The mixed-charge cluster in the integral membrane protein porin is a run of cation residues followed by a run of anion residues with side chains facing into the pore. The induced charge gradient may assist polar solutes in traversing the outer membrane of *E. coli*.

Three multimeric proteins with central heme porphyrin groups are listed in Table 2; cytochrome P450-BM3, flavocytochrome b_2 (1fcb), and catalase (8cat). Each contains a single mixed-charge cluster distant from the heme region. Does the charge cluster possibly help in recruiting the heme or contribute in interactions with other proteins or substrates? Mutagenesis from this perspective may be very informative.

Charge clusters afford a range of interactions, weak to strong, depending on the pH milieu and the influences of neighboring molecules. Most charge clusters are situated on the surface of the protein structure, in a crevice, or at the interfaces in muitimeric proteins. Specific examples featuring interface charge associations include GST (homodimer and heterodimer forms), catalase, monomer–monomer polymerization of eubacterial RecA (37), and polymerization in intra- or extracellular skeletal determinations. Other examples of mixed-charge clusters at interchain contacts, which may stabilize quaternary protein formation and/or protein complexes, include mosaic virus coat protein (see text), nitrogenase molybdenum–iron protein, and fructose-1,6-bisphosphate aldolase. A mechanism in this respect proposes long-range electrostatic attractions that help orient the surfaces (11). Hydrophobic interactions and hydrogen bonding subsequently act to strengthen the interface contacts. In this vein, it was recently observed that the most frequent interchain specific residue contacts involve charged residues of opposite sign and secondarily nonspecific hydrophobic interactions (38). This result is consistent with many observations on antibody–antigen pairs, hormone–receptor contacts, and large protein complexes where multichain stabilization is facilitated by electrostatic interactions; compare with refs. 39 and 40 on binding of the human growth hormone and the human growth hormone receptor.

The analysis for clusters is done relative to the total quaternary structure as well as for the component chains. Thus, the homodimer GST structure features a mixed-charge cluster at the interface involving both chains, whereas each separate chain shows no significant charge cluster. In such cases, the charge cluster is probably an essential ingredient contributing to the dimer formation. On the other hand, where only a single chain of the structure is available, a charge cluster on its surface may signify that the active structure requires a higher order complex putatively established by means of electrostatic interactions.

Note Added in Proof. We updated the results of Table 1 with respect to an expanded 419 nonredundant protein 3D structure set (PDB release of January, 1996). The protein, prothrombin fragment 1 (2pf1), shows a single positive-charge cluster. About 1.9% carry a single acidic cluster. New examples in this category include inositol polyphosphate 1-phosphatase (1inp), sonic hedgehog (1vhh), α-parvalbumin (1rtp), and black beetle virus capsid protein (2bbv). The number with a single mixed-charge cluster increased to 7.6% and those containing multiple-charge clusters is 3.1%. The total percent structures from among the 419 with at least one charge cluster is about 13.0%. Details and interpretations will be presented elsewhere.

We appreciate useful comments on the manuscript from Drs. B. E. Blaisdell, L. Brocchieri, F. Cohen, J. Griffin, H. Leucke, T. Poulos, and W. I. Weis. We are especially grateful to Dr. K. D. Karlin (John Hopkins University) for many discussions.

1. Karlin, S. & Zhu, Z.-Y. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 8344–8349.
2. Hope, I. A., Mahadevan, S. & Struhl, K. (1988) *Nature (London)* **333,** 635–640.
3. Perutz, M. (1988) *Nature (London)* **336,** 202–203.
4. Richardson, W. D., Roberts, B. L. & Smith, A. E. (1986) *Cell* **44,** 77–85.
5. Karlin, S., Blaisdell, B. E. & Brendel, V. (1990) *Methods Enzymol.* **183,** 338–402.
6. Karlin, S. (1990) in *Structure and Methods: DNA Protein Complexes and Proteins,* eds. Sarma, R. H. & Sarma, M. H. (Adenine, Albany, NY), Vol. 2, pp. 171–180.
7. Honig, B. & Nicholls, A. (1995) *Science* **268,** 1144–1149.
8. Karlin, S. (1995) *Curr. Opin. Struct. Biol.* **5,** 360–371.
9. Patthy, L. & Blundell, T. L. (1995) *Curr. Opin. Struct. Biol.* **5,** 357–359.
10. Richmond, T. J. & Richards, F. M. (1978) *J. Mol. Biol.* **119,** 537–555.
11. Gray, H. B. & Ellis, W. R., Jr. (1994) in *Bioinorganic Chemistry,* eds. Bertini, I., Gray, H. B., Lippard, S. J. & Valentine, J. S. (University Science Books, Mill Valley, CA) pp. 315–364.
12. Northrup, S. H., Reynolds, J. C. L., Miller, C. M., Forrest, K. J. & Boles, J. O. (1986) *J. Am. Chem. Soc.* **108,** 8162–8170.
13. Ravichandran, K. G., Boddupalli, S. S., Hasemann, C. A., Peterson, J. A. & Deisenhofer, J. (1993) *Science* **261,** 731–736.
14. Hasemann, C. A., Ravichandran, K. G., Peterson, J. A. & Deisenhofer, J. (1994) *J. Mol. Biol.* **236,** 1169–1185.
15. Luecke, H., Chang, B. T., Mailliard, W. S., Schlaepfer, D. D. & Haigler, H. T. (1995) *Nature (London)* **378,** 512–515.
16. Moss, S. E. (1995) *Nature (London)* **378,** 446–447.
17. Rossmann, M. G. (1975) *Structure and Conformation of Nucleic Acids and Protein–Nucleic Acid Interactions,* eds. Sundaralingam, M. & Rao, T. (University Park Press, Baltimore), pp. 357–374.
18. Hurley, J. H., Dean, A. M., Sohl, J. L., Koshland, D. E., Jr. & Stroud, R. M. (1990) *Science* **249,** 1012–1016.
19. Imada, K., Sato, M., Tanaka, N., Katsube, Y., Matsuura, Y. & Oshima, T. (1991) *J. Mol. Biol.* **222,** 725–738.
20. Hurley, J. H. & Dean, A. M. (1994) *Structure* **2,** 1007–1016.
21. Perutz, M. F. (1978) *Science* **201,** 1187–1191.
22. Yutani, K., Ogasawara, K., Tsujita, T. & Sugino, Y. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 4441–4444.
23. Kirino, H., Aoki, M., Aoshima, M., Hayashi, Y., Ohba, M., Yamagishi, A., Wakagi, T. & Oshima, T. (1994) *Eur. J. Biochem.* **220,** 275–281.
24. Branden, C. & Tooze, J. (1991) *Introduction to Protein Structure* (Garland, New York).
25. Creighton, T. E. (1993) *Proteins: Structures and Molecular Properties* (Freeman, New York).
26. Xia, Z. & Mathews, F. S. (1990) *J. Mol. Biol.* **212,** 837–863.
27. Colman, P. M. (1994) *Protein Sci.* **3,** 1687–1696.
28. Colman, P. M. (1989) in *The Influenza Viruses,* ed. Krug, R. M. (Plenum, New York), pp. 175–218.
29. Quiocho, F. (1986) *Annu. Rev. Biochem.* **55,** 287–315.
30. Varghese, J. N., McKimm-Breschkin, J. L., Caldwell, J. B., Kortt, A. A. & Colman, P. M. (1992) *Proteins: Struct., Funct., Genet.* **14,** 327–332.
31. Burmeister, W. P., Ruigrok, R. W. & Cusack, S. (1992) *EMBO J.* **11,** 49–56.
32. Sauter, N. K., Hanson, J. E., Glick, G. D., Brown, J. H., Crowther, R. L., Park, S. J., Skehel, J. J. & Wiley, D. C. (1992) *Biochemistry* **31,** 9609–9621.
33. Sauter, N. K., Glick, G. D., Crowther, R. L., Park, S. J., Eisen, M. B., Skehel, J. J., Knowles, J. R. & Wiley, D. C. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 324–328.
34. Hester, G., Brenner-Holzach, O., Rossi, F. A., Struck-Donatz, M., Winterhalter, K. H., Smit, J. D. & Piontek, K. (1991) *FEBS Lett.* **292,** 237–242.
35. Sygusch, J., Beaudry, D. & Allaire, M. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 7846–7850.
36. Baumann, U., Wu, S., Flaherty, K. M. & McKay, D. B. (1993) *EMBO J.* **12,** 3357–3364.
37. Karlin, S. & Brocchieri, L. (1996) *J. Bacteriol.* **178,** 1881–1894.
38. Brocchieri, L. & Karlin, S. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 12136–12140.
39. Wells, J. A. (1991) *Methods Enzymol.* **202,** 390–411.
40. Clackson, T. & Wells, J. A. (1995) *Science* **267,** 383–386.