Proc. Natl. Acad. Sci. USA
Vol. 91, pp. 817–821, January 1994
Biophysics

Predicting Ca^{2+}-binding sites in proteins

(ion binding/molecular recognition/computational biochemistry)

MURAD NAYAL AND ENRICO DI CERA*

Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine, Box 8231, St. Louis, MO 63110

Communicated by Carl Frieden, October 27, 1993 (received for review August 9, 1993)

ABSTRACT **The coordination shell of Ca^{2+} ions in proteins contains almost exclusively oxygen atoms supported by an outer shell of carbon atoms. The bond-strength contribution of each ligating oxygen in the inner shell can be evaluated by using an empirical expression successfully applied in the analysis of crystals of metal oxides. The sum of such contributions closely approximates the valence of the bound cation. When a protein is embedded in a very fine grid of points and an algorithm is used to calculate the valence of each point representing a potential Ca^{2+}-binding site, a typical distribution of valence values peaked around 0.4 is obtained. In 32 documented Ca^{2+}-binding proteins, containing a total of 62 Ca^{2+}-binding sites, a very small fraction of points in the distribution has a valence close to that of Ca^{2+}. Only 0.06% of the points have a valence ≥ 1.4. These points share the remarkable tendency to cluster around documented Ca^{2+} ions. A high enough value of the valence is both necessary (58 out of 62 Ca^{2+}-binding sites have a valence ≥ 1.4) and sufficient (87% of the grid points with a valence ≥ 1.4 are within 1.0 Å from a documented Ca^{2+} ion) to predict the location of bound Ca^{2+} ions. The algorithm can also be used for the analysis of other cations and predicts the location of Mg^{2+}- and Na^{+}-binding sites in a number of proteins. The valence is, therefore, a tool of pinpoint accuracy for locating cation-binding sites, which can also be exploited in engineering high-affinity binding sites and characterizing the linkage between structural components and functional energetics for molecular recognition of metal ions by proteins.**

Calcium ions are involved in a variety of important biological functions, many of which are accomplished through interaction with proteins. Events originating at the level of the cell membrane trigger a cascade of processes leading to the mobilization of Ca^{2+} for interaction with intracellular target proteins (1). Major conformational transitions induced by Ca^{2+} binding allow troponin C and actin to accomplish their key role in muscle contraction (2, 3) and lead to the activation of a number of target enzymes by calmodulin (4, 5). Calcium ions also interact with a number of extracellular proteins to confer thermal stability, as in the case of thermolysin, protection against autolysis, as in the case of trypsin, and are required for the activity of a number of enzymes—e.g., phospholipase A_2 (6) and some of the vitamin K-dependent factors of the coagulation cascade (7).

Although involvement of Ca^{2+} has long been documented in many key regulatory activities and the mechanism of action of this cation is sufficiently well understood, prediction of the structural architecture necessary to coordinate Ca^{2+} in proteins remains elusive. Understanding molecular recognition of Ca^{2+} by proteins in structural terms is *conditio sine qua non* for eventually relating structural features to functional energetics and is also beneficial to the design of new Ca^{2+}-specific biological carriers that can be exploited in pharmacological applications. The difficulty of predicting Ca^{2+}-binding sites in proteins stems from the extreme heterogeneity of the coordination geometry and structural architecture (8). Ca^{2+}-binding sites in proteins have a coordination number anywhere between 3 (e.g., taka amylase A) and 8 (e.g., thermolysin). Distances between Ca^{2+} and its ligands in the coordination shell vary from 1.6 Å (e.g., ovalbumin) to 3.3 Å (e.g., DNase I–actin). McPhalen *et al.* (8) have also noted that, in general, Ca^{2+} does not lie in the plane of the ligand group. The average rms deviation of the ligands in the coordination shell from an ideal polygon structure in a sample of 26 Ca^{2+}-binding sites was found to be 0.43 Å, a value significantly larger than the 0.1- to 0.2-Å error in the determination of the crystal structures (8). Due to all these factors, it has been concluded that prediction of Ca^{2+}-binding sites in proteins based on purely structural analysis is unlikely to succeed (8).

In 1990 Eisenberg *et al.* (9) have pointed out that the environments of metal ions in proteins seemingly share a remarkable feature, "regardless of the metal and its precise pattern of ligation to the protein." The metal is coordinated by an inner sphere of hydrophilic groups, embedded in an outer sphere of hydrophobic groups, giving rise to a center of substantial hydrophobicity contrast. An algorithm based on calculation of the hydrophobicity contrast for the protein effectively locates ion-binding sites in a number of cases (9). The importance of Eisenberg's observation is in the "local" nature of the origin of ion specificity because the hydrophobicity contrast function is determined by groups located within 7 Å from the metal. Other components that are naturally long-range, such as the electrostatic properties of the protein, do not seem to set the rules for recognition and fail to provide a simple algorithm for the prediction of metal-binding sites (9). In this paper we introduce an alternative algorithm that, much in the spirit of Eisenberg's approach, focuses on local properties of the environment of Ca^{2+} in proteins but provides a prediction of Ca^{2+}-binding sites with greater accuracy.

The Protein Data Bank (PDB) contains a total of 32 different protein structures with a total of 62 documented, distinct Ca^{2+}-binding sites. The analysis of this complete set of protein structures allows one to construct a plot of the radial distribution of atoms found around documented Ca^{2+} ions; the results are shown in Fig. 1. Practically all Ca^{2+}-binding sites contain an inner shell of oxygens clustering around 2.1–2.7 Å from the Ca^{2+} ion. A second shell, 3.2–3.8 Å away from the Ca^{2+} ion, is mostly populated by carbons and presumably supports the coordinating oxygens of the inner shell. A third shell, 4–4.8 Å away from the Ca^{2+} ion, contains nitrogens in significant amounts. At a distance >5 Å, no preferential clustering of particular atom groups is observed. The molecular architecture for recognition of Ca^{2+} by proteins involves the organization of atom groups within 5 Å from the ion in a fashion consistent with the observation by Eisenberg *et al.* (9). In addition, the rules for ligation seem

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.

Abbreviation: PDB, Protein Data Bank.

*To whom reprint requests should be addressed.

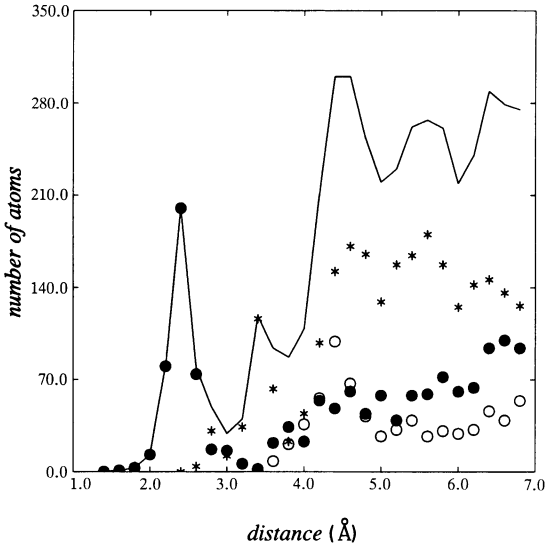

FIG. 1. Distribution of atoms as a function of the distance from a Ca^{2+} ion, as derived from the analysis of 32 Ca^{2+}-binding proteins in the PDB (see also Table 1). The continuous line depicts the sum of all atoms. ●, Oxygens; *, carbons; ○, nitrogens. Note how the inner coordination shell (peak, ≈2.4 Å) is composed almost exclusively by oxygen atoms. A supporting outer shell of carbon atoms shows a peak at ≈3.3 Å. The nitrogen-atom peak occurs at ≈4.2 Å.

to be set exclusively by the interaction of Ca^{2+} with oxygen groups because out of 376 atoms composing the inner coordination shell within 2.7 Å in 62 cases examined, 371 are oxygen atoms (198 carboxyl, 94 carbonyl, 6 hydroxyl, and 73 water oxygens). This observation has prompted us to look at Ca^{2+}-binding sites in terms of the "electrostatic valence principle" formulated by Pauling some 60 yr ago (10). Assuming local neutralization of charges, the bond strength, or bond order, contributed by each oxygen ligand to the ligated cation is the charge of the cation divided by the number of ligands, or the coordination number. When ligands are asymmetrically disposed around the ligated cation, different bonds are expected to have different strength. Here the bond-length correlation to bond order, which is also seen in covalent bonding, can be used to estimate the strength of different bonds in structures determined by x-ray crystallography. Several empirical expressions relating bond length to bond order have been proposed (11, 12). The most thorough characterization has been provided by Brown from the analysis of the environments of 884 cations in 417 different structures of small inorganic molecules (13–16). The bond strength or valence v, contributed by a given ligating group to the cation, is given by

$$v = (R/R_1)^{-N}, \quad [1]$$

where R is the distance of the atom from the cation, R_1 is the value of R giving $v = 1$, and N is an empirical exponent. The overall valence of a given disposition of ligating oxygens around a cation is calculated from Eq. 1 by summing over all atoms. Values of R_1 and N have been derived for a number of cation–oxygen pairs from analysis of crystals of metal oxides, and the empirical expression in Eq. 1 approximates extraordinarily well the valence of the cation being ligated (13). We have therefore decided to exploit this approach in the analysis of Ca^{2+}-binding proteins to assess whether Ca^{2+}-binding sites lined up by oxygen atoms possess values of the valence significantly different than those found anywhere else in the protein.

METHODS

Algorithm. The protein structure is embedded in a very fine three-dimensional grid of points with a spacing of 0.1 Å. For each potential Ca^{2+}-binding site j, defined as a grid point not within the van der Waals radius of a protein atom increased by the ionic radius of Ca^{2+} and surrounded by at least three oxygen atoms within a probe radius of 3.4 Å, the value of the valence v_j is computed by summing the contribution of all oxygens according to Eq. 1. The values of R_1 and N for the Ca^{2+}—O pair are R_1 = 1.909 Å and N = 5.4 (14). In constructing the algorithm, we have found that looping over grid points would be very inefficient, an aspect already noted by Eisenberg and coworkers (9). In practice, the algorithm loops over all atoms and ignores water molecules (water oxygen atoms were found to add considerable noise and very little information). For each oxygen atom, the allowable grid points within the probe radius are determined, and the contribution of the oxygen atom to the valence of each point is computed according to Eq. 1. All potential Ca^{2+}-binding sites as defined above are stored as "pseudo-atoms" in a PDB-formatted output file. The file can be used to retrieve grid points to be superimposed to the protein structure for molecular graphics. The valences are then sorted for quantitative analysis. The executable code is available upon request (to enrico@caesar.wustl.edu).

The Protein Sample. All protein structures currently published in the PDB were considered for analysis, except for the following: The Ca^{2+} sites in ferritin, satellite tobacco necrosis virus, dihydrofolate reductase, and Ca135 in oncomodulin were excluded because these Ca^{2+}-binding sites contain atoms from symmetry-related molecules in the crystal not listed in the PDB file. Also excluded were the Ca^{2+} ions bound to acid proteinase, ribonuclease, and macromomycin because they are ligated by only one or two protein ligands. All other Ca^{2+}-binding sites in the PDB were included in the analysis. For multiple structures of the same protein only what seemed to be the best determination was used. The final sample processed for analysis contained 32 different proteins (see Table 1), with a total of 62 distinct Ca^{2+}-binding sites.

RESULTS

The distribution of valence values for all proteins analyzed in this study is given in Fig. 2. The distribution is peaked around a value of the valence of 0.4, quite different from the expected value of 2 for the Ca^{2+} ion; <0.06% of the points show a valence of 1.4 or higher. This result suggests that among all possible potential Ca^{2+}-binding sites of a protein, defined as cavities lined up by at least three oxygen atoms, only a very small fraction possesses a high enough value of the valence to accommodate a Ca^{2+} ion. If the portion of the valence distribution for $v \geq 1.4$ in Fig. 2 is exploded by 1000 times (discontinuous-dotted line), a small peak is observed around v = 1.9, a value of the valence very close to the expected value for the Ca^{2+} ion. This very small fraction of high-valence grid points correlates strongly with the position of Ca^{2+} ions. In fact, of 4321 grid points with a valence ≥1.4 in the complete protein sample analyzed, 3740 (87%) are found within 1 Å from a documented Ca^{2+} ion. Of 1873 grid points with $v \geq 1.6$, 1814 (97%) are within 1 Å, and 1868 (99.7%) are within 3.5 Å from a Ca^{2+} ion. High-valence values may reflect a structural organization peculiar of cation-binding sites (17–19). Fig. 3 shows the valence as a function of the distance from a documented Ca^{2+} ion; the results are averaged over the entire protein sample. The average value of the valence changes from 1.6 at the Ca^{2+} ion to ≈1.0 at 0.5 Å away from the Ca^{2+} ion and approaches the background level of the protein around 3.0 Å. Because only a very small fraction of points has a valence $v \geq 1.4$ (Fig. 2) and only points very close to the Ca^{2+} ion have

Table 1. Results of calculations of valence values for documented Ca^{2+}-binding proteins

Protein	PDB*	N_{tot}†	$N_{1.4}$‡	$N_{1.4}$§	$N_{1.4}$¶	Ca‖	Rank**	R††	v‡‡
Ca^{2+}-binding protein	3icb	80,479	221	221	217	1	1 (1)	0.2	2.19
						2	1 (14)	0.1	2.01
Calmodulin	3cln	119,267	348	348	313	2	1 (1)	0.3	1.99
						4	1 (11)	0.4	1.85
						3	1 (16)	0.8	1.84
						1	1 (18)	0.2	1.82
Con A	3cna	206,415	79	1	0	2	67 (67)	2.0	1.41
DNase I–actin	1atn	460,083	245	245	200	2	1 (1)	2.9	1.79
						3	1 (2)	0.5	1.77
						4	1 (13)	0.5	1.67
						5	1 (54)	2.3	1.58
Elastase	3est	136,353	129	129	129	280	1 (1)	0.6	1.72
Enolase	5enl	219,153	0	0	0	438	4 (4)	0.9	1.35
Galactose binding protein	3gbp	142,199	104	104	104	309	1 (1)	0.3	2.05
α-Lactalbumin	1alc	102,639	56	56	12	200	1 (1)	0.1	1.88
Lysozyme	3lhm	67,271	164	162	122	131	1 (1)	0.7	1.65
Mannose-binding protein A	2msb	186,878	650	650	621	D2	1 (1)	0.5	2.37
						D1	1 (117)	0.4	1.98
						C1	1 (118)	0.4	1.96
						C2	1 (119)	0.5	1.95
						D3	1 (193)	2.7	1.77
						C3	1 (405)	2.2	1.58
Mesenterico peptidase	1mee	184,590	152	150	141	400	1 (1)	0.2	2.18
						401	3 (149)	0.5	1.45
Neuraminidase N9	2nn9	376,401	6	1	1	94	4 (4)	0.8	1.42
Neutral protease	1npc	238,585	263	263	263	319	1 (1)	0.5	1.90
						320	1 (2)	3.3	1.88
						322	1 (8)	0.3	1.82
						321	1 (59)	0.3	1.69
Oncomodulin	1omd	112,655	204	204	203	109	1 (1)	0.2	1.93
						110	1 (3)	0.2	1.92
Ovalbumin	1ova	996,242	31	0	0	500	44 (44)	1.9	1.39
Parvalbumin B	5cpv	77,928	161	161	157	110	1 (1)	0.2	2.16
						109	1 (34)	0.3	1.91
Pea lectin	2ltn	349,252	160	95	95	A	1 (1)	0.3	1.77
						C	1 (3)	0.2	1.72
Porin	2por	321,846	116	116	115	303	1 (1)	0.2	2.12
						304	1 (2)	0.3	2.00
						302	56 (361)	0.3	1.19
Prophospholipase A_2	4bp2	88,120	41	41	35	201	1 (1)	0.5	1.64
Protease K	3prk	152,691	2	1	1	285	2 (2)	0.5	1.42
Serine protease B	4sgb	161,279	95	94	92	8	1 (1)	0.4	1.91
Soybean mosaic virus	4sbv	479,525	68	62	50	A1	1 (1)	0.3	1.79
						C1	1 (4)	0.6	1.65
						B1	1 (14)	1.4	1.55
Staphylococcus nuclease	1snc	109,425	0	0	0	142	1 (1)	0.7	1.35
Subtilisin BPN	2st1	144,687	59	56	46	276	1 (1)	0.3	2.33
						277	4 (58)	1.3	1.41
Subtilisin Carlsberg	1cse	165,106	130	128	123	430	1 (1)	0.3	2.15
						401	1 (126)	2.8	1.44
Taka amylase A	2taa	448,794	62	1	0	A1	61 (61)	3.2	1.40
Thermitase	2tec	167,911	149	149	149	343	1 (1)	0.6	2.12
						344	1 (130)	0.3	1.69
Thermolysin	1tmn	224,995	260	260	260	1	1 (1)	0.4	1.91
						4	1 (19)	0.4	1.74
						3	1 (190)	0.1	1.46
						2	1 (260)	0.1	1.40
Tobacco mosaic virus	2tmv	147,944	77	34	20	1	3 (3)	1.0	1.70
Troponin C	5tnc	163,324	226	225	213	163	1 (1)	0.3	2.14
						164	1 (12)	0.3	1.91
β-Trypsin	4ptp	128,884	31	31	29	247	1 (1)	0.4	1.75
Trypsinogen	2tgt	123,442	34	30	29	480	1 (1)	0.1	1.74

*PDB code.
†Total number of points examined.
‡Number of points with valence ≥1.4.
§Number of points with valence ≥1.4 and within 3.5 Å from a documented Ca^{2+} ion.
¶Number of points with valence ≥1.4 and within 1.0 Å from a documented Ca^{2+} ion.
‖Residue number or code for the Ca^{2+} site in the PDB file.
**Relative rank of point; the absolute rank is given in parentheses.
††Distance of point in Å from documented Ca^{2+} ion.
‡‡Valence of point.

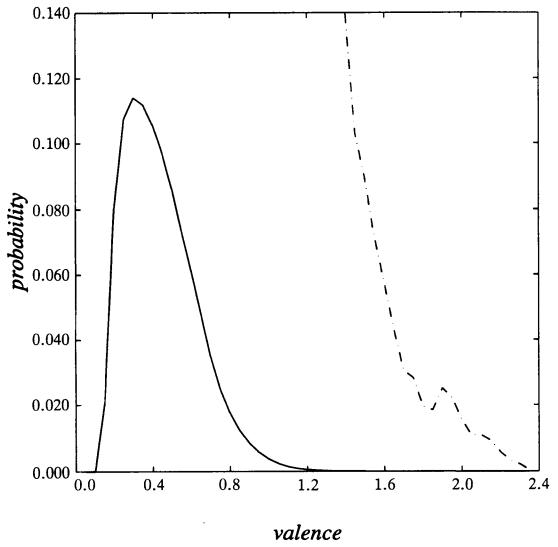

FIG. 2. Probability distribution of valence values for 32 Ca^{2+}-binding proteins in the PDB (see also Table 1), as determined with the algorithm based on Eq. 1 in the text (continuous line). The discontinuous-dotted line is the portion of the distribution for $v \geq 1.4$ times 1000. Note how the distribution for all proteins sharply peaks around $v = 0.4$, a value of the valence significantly smaller than that expected for Ca^{2+}; only ≈0.06% of the points populated valence values ≥1.4. The expanded, high-valence tail of the distribution contains all the information to locate Ca^{2+}-binding sites. This tail is characterized by a second peak, around $v = 1.9$, indicative of a significant structural organization of regions of the protein involved in Ca^{2+} binding. The distribution shown is typically observed for Ca^{2+}-binding proteins analyzed separately.

a high valence (Fig. 3), it is expected that the algorithm based on Eq. 1 will provide a most accurate prediction of where Ca^{2+} binds in proteins.

The predictive value of the algorithm can be assessed in each case by computing the distribution of valence values for the protein under investigation and ranking the grid points according to their valence. All information about the location of Ca^{2+}-binding sites is contained in the high-valence tail of the distribution. The results of our analysis are summarized in Table 1. Points with a valence of $v \geq 1.4$ typically represent 0.1% of the total, and on the average predict the location of Ca^{2+}-binding sites within 0.7 Å or, at most, within 3.2 Å. In 25 proteins out of 32, the point with the highest valence in the distribution, which ranks first, is on the average within 0.5 Å from a Ca^{2+} ion and has a valence of 1.94. Absolute ranking of the points becomes misleading for proteins containing multiple Ca^{2+}-binding sites and is replaced by relative ranking. The relative rank of a point is computed as follows. Starting from the highest valences of the distribution, the first point within 3.5 Å from a Ca^{2+} ion is determined. To rank points predicting additional Ca^{2+}-binding sites, all points within 3.5 Å from the first Ca^{2+} ion are deleted from the distribution, and the search is repeated on the new distribution. The relative rank of points predicting the second Ca^{2+}-binding site is the same as the absolute rank of those points in a new distribution, where all points within 3.5 Å from the first Ca^{2+}-binding site have been deleted. The procedure is repeated for all Ca^{2+}-binding sites. For example, in thermitase the first point of the distribution is within 0.6 Å from the Ca^{2+} ion denoted by 343 in the PDB file, but the highest-ranking point predicting the second Ca^{2+} ion denoted by 344 in the PDB file has an absolute rank of 130 in the distribution and is within 0.3 Å from the Ca^{2+}. However, this result is because all 129 higher-ranking points are within 1.0 Å from the first Ca^{2+}-binding site. Out of 62 sites, 52 are predicted by a point in the distribution with a relative rank of 1, which, on average, is within 0.6 Å from the Ca^{2+} ion and has a valence of 1.85. Although the cutoff of $v \geq 1.4$ proves most adequate for predicting many of the Ca^{2+}-binding sites, it should be used with some flexibility. In the case of staphylococcal nuclease none of the points in the distribution has a valence >1.4. Yet, the highest-ranking point ($v = 1.35$) predicts the position of the bound Ca^{2+} within 0.7 Å. Only in the case of Con A, Ca302 of porin, ovalbumin, and taka amilase A, the algorithm fails to predict the position of the Ca^{2+} ion as the region with the highest valence. A closer look at Con A and ovalbumin reveals that the Ca^{2+}-binding cavity in these proteins is entirely occupied by the van der Waals radii of ligating oxygen atoms increased by the ionic radius of Ca^{2+}, so that no grid point can be found *a priori* in the cavity. These sites have unusually short Ca^{2+}—O distances (1.96, 2, and 2.1 Å in Con A; 1.6 and 1.9 Å in ovalbumin) compared with the average Ca^{2+}—O ligating distance of 2.4 Å. In the case of

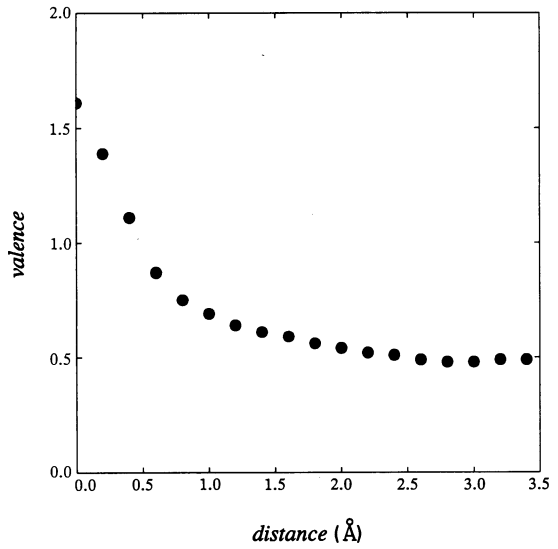

FIG. 3. Valence as a function of the distance from documented Ca^{2+} ions, as determined by averaging the results obtained for 32 Ca^{2+}-binding proteins in the PDB (see also Table 1). The valence assumes an average value of 1.6 at the documented Ca^{2+} ion and decreases rapidly with increasing distance to reach the background valence level of the protein.

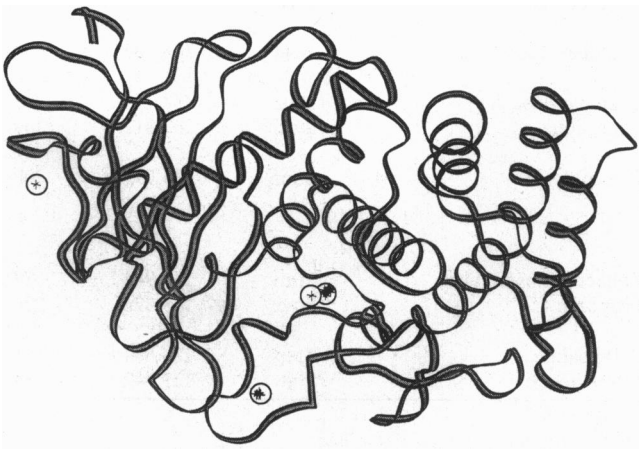

FIG. 4. Ribbon plot of thermolysin obtained with the program INSIGHTII (Biosym Technologies, San Diego) using the PDB file listed in Table 1. Bound Ca^{2+} ions are depicted by circles. Crosses depict all points of the valence distribution with a valence $v \geq 1.4$. These points represent potential Ca^{2+}-binding sites and cluster with remarkable accuracy at documented Ca^{2+} ions.

taka amylase A, Ca^{2+} is ligated by three oxygen atoms; two of them are located unusually far away from the Ca^{2+} ion (3 and 3.1 Å), so that the calculated valence at the site is low. A similar situation is seen for Ca302 of porin.

The pinpoint accuracy of the algorithm is best illustrated by superimposing all points with a valence ≥ 1.4 in the distribution on the crystal structure. These high-valence points tend to cluster at the bound Ca^{2+} ions with remarkable precision. The results obtained for thermolysin are shown in Fig. 4. The algorithm predicts Ca^{2+}-binding sites as dense clusters of high-valence grid points, which clearly stand out from the background noise level of the ensemble of grid points of the protein. This feature of the algorithm may be of particular interest in practical applications and especially in the analysis of *de novo* structures. Clusters of high-valence grid points may serve to direct the analysis on well-defined, localized regions of the structure that may contain all the information on Ca^{2+}-binding sites.

DISCUSSION

We have demonstrated that it is possible to predict the location of Ca^{2+}-binding sites in proteins by taking into account local properties of the binding site. The valence used to rank potential Ca^{2+}-binding sites is based on an empirical expression used extensively in the analysis of structures of minerals containing Ca^{2+} ions. A high enough value of the valence (≥ 1.4) seems both necessary (58 Ca^{2+}-binding sites out of 62 documented in the PDB have a valence ≥ 1.4) and sufficient (95% of the grid points in 32 Ca^{2+}-binding proteins with a valence ≥ 1.4 are closely associated with a Ca^{2+} ion) to predict the location of bound Ca^{2+} ions. This result draws attention to the role of the inner coordination shell composed by hydrophilic oxygen groups. Because oxygen atoms are only supported by carbon atoms in proteins, it is clear that the inner hydrophilic core must be surrounded by a hydrophobic shell of supporting carbon atoms. The high hydrophobicity contrast observed at Ca^{2+} and other metal-binding sites in proteins (9) may well be a consequence of the need of supporting the inner shell of ligating oxygen atoms.

The potential in Eq. 1 and the algorithm based on it are cation-specific because they depend on the values of R_1 and N for a specific metal–ligand pair and the radius of a given cation. This is an element of considerable significance in the computational approach to cation binding to proteins because, in principle, it allows an understanding of the origin of ion-binding specificity. Practically, a protein can be screened with R_1 and N values specific of any cation to assess what cations can be bound and where. The distribution of valences obtained for Ca^{2+} may differ significantly, in the same protein, from that obtained for Mg^{2+}, Na^+, or K^+. We have analyzed a structure of parvalbumin (4pal) that contains both Ca^{2+} and Mg^{2+} ions bound. When the algorithm was run with R_1 and N values for Ca^{2+}, all points with $v \geq 1.4$ in the distribution were found to cluster within 0.2 Å from the bound Ca^{2+} ion; the average valence was 2.1. When the algorithm was run with R_1 and N values for Mg^{2+}, the 50 highest-ranking points in the distribution were found to cluster within 0.3 Å from the bound Mg^{2+} ion; the average valence was 1.7. Analysis of the Mg^{2+}-binding proteins H-Ras p21 protein (1q21), D-xylose isomerase (2xis), enolase (7enl), isocitrate dehydrogenase (8icd), and ribulose 1,5-bisphosphate carboxylase/oxygenase (8rub) shows that the highest-ranking point of the valence distribution is, on the average, within 1.8 Å from the bound Mg^{2+} and has an average valence of 1.70. Analysis of the Na^+-binding proteins proteinase A (2sga) and insulin (9ins) shows that the highest-ranking point of the valence distribution is within 0.9 (2sga) or 1.1 Å (9ins) from the bound Na^+ ion and has a valence of 1.17 (2sga) or 0.93 (9ins). The algorithm also clearly indicates the presence of a potential Na^+-binding site in thrombin, as predicted by functional studies (20). The results relative to the prediction of Mg^{2+}, Na^+, and other cation-binding sites in proteins will be dealt with in detail elsewhere.

Perhaps the most interesting aspect of the algorithm introduced here stems from the possibility of formulating a connection between structural components and functional energetics for cation binding to proteins. Our approach can be exploited and optimized along several different lines of investigation. For example, one could try to correlate the empirical values of the valences at the Ca^{2+} sites with measured binding affinities and test whether high-affinity binding sites are characterized by high valence values. If such a correlation exists, then Ca^{2+}-binding sites could be engineered in proteins with the desired binding affinity. Specificity could be added by determining coordination geometries that favor binding of a particular cation, based on the prediction of the valence term. As the number of documented Ca^{2+}-binding sites increases, it will be possible to determine R_1 and N from analysis of all protein crystals and test whether they differ from those derived from the analysis of crystals of metal oxides and lead to improved predictive algorithms. The approach introduced here also provides a convenient method of screening new protein structures for potential Ca^{2+}-binding sites, which may be of considerable practical utility to crystallographers. Finally, the algorithm can be used to search for additional Ca^{2+}-binding sites in those cases where the presence of multiple classes of sites is suggested from functional studies.

This work was done during the tenure of an Established Investigator Award from the American Heart Association and Genentech to E.D.C. and was supported by National Science Foundation Grant DMB91-04963 and a Grant from the Lucille P. Markey Charitable Fund.

1. Rasmussen, H. (1990) *J. Biol. Chem.* **371**, 191–206.
2. Holmes, K. C., Popp, D., Gebhard, W. & Kabsch, W. (1990) *Nature (London)* **347**, 44–49.
3. Kabsch, W. & Vandekerckhove, J. (1992) *Annu. Rev. Biophys. Biomol. Struct.* **21**, 49–76.
4. Klee, C. B. & Vanaman, T. C. (1982) *Adv. Protein Chem.* **35**, 213–321.
5. Cox, J. A., Comte, M., Malnoe, A., Burger, D. & Stein, E. A. (1984) in *Metal Ions in Biological Systems*, ed. Sigel, H. (Decker, Basel, Switzerland), Vol. 17, pp. 215–273.
6. Einspahr, H. & Bugg, C. E. (1984) in *Metal Ions in Biological Systems*, ed. Sigel, H. (Decker, Basel, Switzerland), Vol. 17, pp. 51–97.
7. Mann, K. G., Nesheim, M. E., Church, W. R., Haley, P. & Krishnaswamy, S. (1990) *Blood* **76**, 1–16.
8. McPhalen, C. A., Strynadka, N. C. J. & James, M. N. G. (1991) *Adv. Protein Chem.* **42**, 77–144.
9. Yamashita, M. M., Wesson, L., Eisenman, G. & Eisenberg, D. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 5648–5652.
10. Pauling, L. (1929) *J. Am. Chem. Soc.* **51**, 1010–1026.
11. Pauling, L. (1947) *J. Am. Chem. Soc.* **69**, 542–553.
12. Donnay, G. & Allman, R. (1970) *Am. Miner.* **55**, 1003–1015.
13. Brown, I. D. & Shannon, R. D. (1973) *Acta Crystogr. A* **29**, 266–282.
14. Brown, I. D. & Wu, K. K. (1976) *Acta Crystogr. B* **32**, 1957–1959.
15. Altermatt, D. & Brown, I. D. (1985) *Acta Crystogr. B* **41**, 240–244.
16. Brown, I. D. & Altermatt, D. (1985) *Acta Crystogr. B* **41**, 245–247.
17. Diebler, H., Eigen, M., Ilgenfritz, G., Maass, G. & Winkler, R. (1969) *Pure Appl. Chem.* **20**, 93–115.
18. Suelter, C. H. (1970) *Science* **168**, 789–795.
19. Eisenman, G. & Dani, J. A. (1987) *Annu. Rev. Biophys. Biophys. Chem.* **16**, 205–226.
20. Wells, C. M. & Di Cera, E. (1992) *Biochemistry* **31**, 11721–11730.