PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Towards predicting Ca^{2+}-binding sites with different coordination numbers in proteins with atomic resolution

Xue Wang,[1] Michael Kirberger,[2] Fasheng Qiu,[1] Guantao Chen,[1,3]* and Jenny J. Yang[2]*

[1] Department of Computer Science, Georgia State University, Atlanta, Georgia 30303
[2] Department of Chemistry, Center for Drug Design and Biotechnology, Georgia State University, Atlanta, Georgia 30303
[3] Department of Mathematics and Statistics, Georgia State University, Atlanta, Georgia 30303

ABSTRACT

Ca^{2+}-binding sites in proteins exhibit a wide range of polygonal geometries that directly relate to an equally-diverse set of biological functions. Although the highly-conserved EF-Hand motif has been studied extensively, non-EF-Hand sites exhibit much more structural diversity which has inhibited efforts to determine the precise location of Ca^{2+}-binding sites, especially for sites with few coordinating ligands. Previously, we established an algorithm capable of predicting Ca^{2+}-binding sites using graph theory to identify oxygen clusters comprised of four atoms lying on a sphere of specified radius, the center of which was the predicted calcium position. Here we describe a new algorithm, MUG (MUltiple Geometries), which predicts Ca^{2+}-binding sites in proteins with atomic resolution. After first identifying all the possible oxygen clusters by finding maximal cliques, a calcium center (CC) for each cluster, corresponding to the potential Ca^{2+} position, is located to maximally regularize the structure of the (cluster, CC) pair. The structure is then inspected by geometric filters. An unqualified (cluster, CC) pair is further handled by recursively removing oxygen atoms and relocating the CC until its structure is either qualified or contains fewer than four ligand atoms. Ligand coordination is then determined for qualified structures. This algorithm, which predicts both Ca^{2+} positions and ligand groups, has been shown to successfully predict over 90% of the documented Ca^{2+}-binding sites in three datasets of highly-diversified protein structures with 0.22 to 0.49 Å accuracy. All multiple-binding sites (i.e. sites with a single ligand atom associated with multiple calcium ions) were predicted, as were half of the low-coordination sites (i.e. sites with less than four protein ligand atoms) and 14/16 cofactor-coordinating sites. Additionally, this algorithm has the flexibility to incorporate surface water molecules and protein cofactors to further improve the prediction for low-coordination and cofactor-coordinating Ca^{2+}-binding sites.

Proteins 2009; 75:787–798.
© 2008 Wiley-Liss, Inc.

Key words: prediction; ligand; graph; clique; water; penalty; statistics; coordination; clag.

INTRODUCTION

The biological role of calcium is mainly fulfilled by interacting with different classes of Ca^{2+}-binding proteins (CaBPs).[1–3] Depending on the roles and cellular locations of CaBPs, their affinities may vary by as much as 10^6 fold.[4] Intracellularly, calcium binding to trigger proteins such as calmodulin and troponin C with helix-loop-helix motifs results in Ca^{2+}-induced conformational change which in turn mediates the activity of different cellular processes. Buffer proteins such as calbindin D9K and parvalbumin are essential to maintain proper calcium homeostasis. Conversely, many extracellular CaBPs, such as metabotropic glutamate receptors (mGluRs) and calcium sensing receptors, have weak Ca^{2+}-binding affinities (~m*M*) yet are essential for extracellular calcium signaling and cell–cell communication.[5] Membrane proteins and free peptides, such as Ca^{2+} channels and STIM1,[6] which also play major roles in Ca^{2+} homeostasis, are known to have Ca^{2+} affinities in the millimolar range.

One of the main barriers to understanding the role of calcium in biological systems is identification of the Ca^{2+}-binding sites in proteins, especially for proteins with weak Ca^{2+}-binding affinity (~m*M*). Although X-ray crystallography is a major method for determining the structure of CaBPs with more than 1000 structures determined, the identification of weak affinity (0.05–2 m*M*) binding sites in certain proteins (e.g. mGluR[7,8]) can be problematic with this method, due to a lack of binding during the crystallization process. Further, few structures of transmembrane

Additional Supporting Information may be found in the online version of this article.
Grant sponsor: NSF; Grant number: DMS0070059; Grant sponsor: NIH; Grant numbers: GM62999, GM081749; Grant sponsors: GSU RPE Bioinformatics, GSU MBD fellowship.
*Correspondence to: Guantao Chen, Georgia State University, Department of Mathematics & Statistics, 787 COE, Atlanta, GA 30303. E-mail: gchen@gsu.edu or Jenny J. Yang, Georgia State University, Department of Chemistry, 50 Decatur Street, 550 NSC, Atlanta, GA 30303. E-mail: chejjy@langate.gsu.edu.
Received 4 July 2008; Revised 5 September 2008; Accepted 12 September 2008
Published online 30 September 2008 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.22285

CaBPs are available due to difficulties associated with purification and crystallization of these proteins. In addition, NMR techniques cannot reveal the coordination of all Ca^{2+}-binding sites, although they contribute greatly to our understanding of Ca^{2+}-dependent conformational change and the Ca^{2+}-binding properties of different classes of CaBPs.[9,10] Therefore, there is an urgent need to develop computational methods to predict Ca^{2+}-binding sites in proteins with high accuracy, which will greatly facilitate improved understanding of the role of calcium in biological systems, as well as the capability to design CaBPs with different functional properties.

Ca^{2+}-binding sites are often classified as continuous or discontinuous types.[11,12] Discontinuous Ca^{2+}-binding sites are formed by residues that are spatially proximate in the folded structure but distant in the primary sequence, whereas continuous sites are comprised of amino acids adjacent in the sequence, usually in flexible loop regions such as the helix-loop-helix EF-hand motif. Ca^{2+}-binding sites in proteins are largely formed by ligand oxygen atoms from the side chains of charged carboxyl residues (i.e., Asp and Glu), main chain carbonyl, and solvent water molecules [Fig. 1(A)]. In addition, prosthetic groups such as phosphate and carbohydrates are often found in Ca^{2+}-binding sites in proteins. Further, there are situations where a single ligand atom binds to more than one calcium ion [Fig. 1(C)], forming multiple-binding sites.[13] Consistent with previous extensive analyses,[11,12,14–24] our recent statistical analysis[25] of all Ca^{2+}-binding sites in proteins reported that the non-EF-hand proteins utilize fewer protein ligands (4 ± 2 vs. 6 ± 1) and more water ligands (2 ± 2 vs. 1 ± 0) than EF-hand sites.[25] The variation of ligand distances and bond angles [Fig. 1(B)] is much greater for non-EF-hand than for EF-hand proteins.[25] These significant variations in calcium coordination properties correlate with the diversified biological functions of CaBPs with different binding affinities and Ca^{2+}-induced conformational change. Therefore, based on observations of the extreme heterogeneity of the coordination geometry and structural architecture of Ca^{2+}-binding sites,[26,27] it was concluded that prediction of Ca^{2+}-binding sites in proteins based on uniform structural patterns is unlikely to succeed.[28]

Nevertheless, significant progress has been made in developing algorithms to predict Ca^{2+}-binding sites in proteins.[26,27,29–36] Yamashita *et al.*[31] reported that metals in the binding sites were chelated by a shell of hydrophilic atomic groups, which was further embedded within a larger hydrophobic shell. To predict Ca^{2+}-binding sites, they embedded the whole protein structure into a 3-dimensional grid and then measured qualitatively the hydrophobicity contrast at each grid point. Nayal and Di Cera[26] applied a similar approach by embedding the protein structure into grids and searching for grid points with high values of a valence function whose value was contributed from the bond-strength of each ligand oxygen atom. Wei and Altman[29] developed the program FEATURE based on a Bayesian statistical method to select and measure a variety of physical and chemical features in Ca^{2+}-binding sites and non-sites. For a given protein embedded in grids, they computed a score for each grid based on identified features to indicate the likelihood of Ca^{2+}-binding. In a recent study, Deng *et al.*[35] discovered a strong correlation between Ca^{2+}-binding sites and oxygen clusters containing exactly four atoms lying on a sphere of specified radii. Consequently their approach first identified oxygen clusters, and the calcium location was determined at an equidistant center within each cluster if the distance ranged from 1.8 to 3.0 Å. However, limited studies were reported in determining the precise location of a Ca^{2+}-binding site and in prediction algorithms which include structural water molecules. Recently, Schymkowitz *et al.*[27] reported the capability of predicting the precise locations of Ca^{2+}-binding sites with high coordination numbers based on the Fold-X empirical force field.[27] However, their results also demonstrated a lower success rate and more significant error for Ca^{2+}-binding sites with fewer protein coordination ligand atoms. In addition, this algorithm generally fails to predict multiple-binding sites and sites with multiple water molecules, as well as binding sites that involve prosthetic groups and organic compounds due to lack of defined Fold-X force field parameters.

In this article, we establish an algorithm that utilizes multiple geometrical characteristics of Ca^{2+}-binding sites and predicts the precise location of Ca^{2+}-binding sites including the local coordination for irregular sites. This algorithm is able to predict regular Ca^{2+}-binding sites, multiple-binding sites, sites with multiple structural water molecules, and sites that include prosthetic groups and organic compounds. Using three datasets previously tested by three different algorithms, MUG achieves similar or better performance in terms of sensitivity and selectivity. The mean deviations from predicted and documented sites range from 0.22 to 0.49 Å. Further, 100% sensitivity is achieved using an additional dataset consisting of proteins with multiple-binding sites. This algorithm predicts not only calcium locations but also ligand groups. The precise position of Ca^{2+} ion and its complete-ligand-atom-group (CLAG) can be determined for high coordination Ca^{2+}-binding sites (i.e. at least four ligand atoms) without introducing the structural water molecules. Improved prediction for the low-coordination Ca^{2+}-binding sites (less than four ligands) can be achieved when structural water is included, as demonstrated using a dataset containing 27 sites of coordination number ≤3, where 24 sites are predicted. Similarly, the improved ability to predict the Ca^{2+}-binding sites involving protein cofactors can be achieved by including cofactor atoms, and our results demonstrate that 15 out

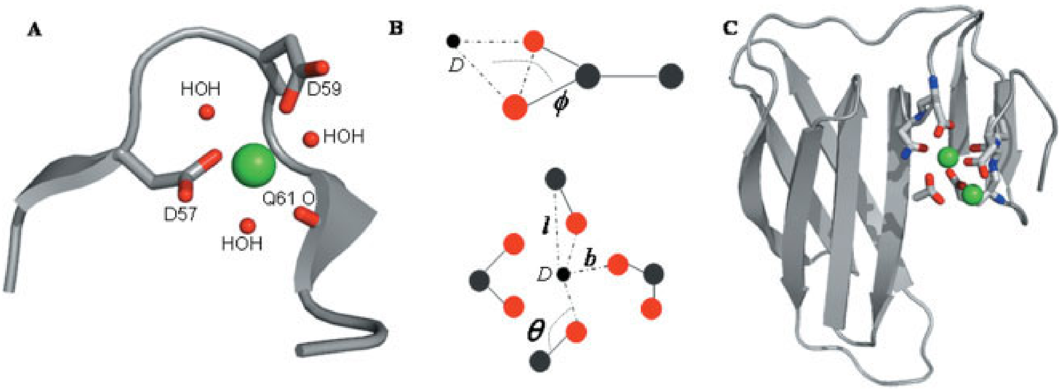

Figure 1

(**A**) Calcium ion (sequence ID 2463, green ball) of thermolysin (1FJ3.pdb) from Protein Data Bank (PDB). Binding residues and water oxygen atoms are labeled in the figure. (**B**) *D* is a spatial point; *b* is *D*-oxygen (red spheres) distance; *l* is *D*-carbon (dark grey spheres) distance; θ is *D*-carbon-oxygen angle; ϕ is the dihedral angle between the plane formed by the side chain carboxyl group (—COO) and the plane formed by the two carboxyl oxygen atoms (bidentate pair) and *D*. (**C**) Calcium ions (sequence ID 8251 and 8252) of serum amyloid P component (1SAC.pdb), its binding residues and cofactor acetic acid.

of 16 cofactor-coordinating sites are predicted with this algorithm.

METHODS

Algorithm description

The process of MUG can be divided into three major steps. In Step 1, for a given protein of known holo (bound) structure, MUG finds oxygen clusters, that is, groups of oxygen atoms that are near to each other in the 3-dimensional structure which will be treated as a potential ligand group. In Step 2, for each oxygen cluster MUG identifies a point CC (calcium center) for the placement of a calcium ion using a grid algorithm. In Step 3, filters consisting of various restrictions are sequentially applied to the structure of each (cluster, CC) pair. If it passes all filters, the cluster is a predicted ligand group and CC is the predicted calcium position; otherwise, MUG removes the oxygen atom with the longest distance to the CC from that cluster and recursively calls Step 2 and Step 3 until either a "pass" is obtained or the number of oxygen atoms remaining in the cluster is less than four. The schematic diagram and flow chart of MUG are illustrated in Figures 2 and 3, respectively.

Finding oxygen clusters

MUG finds oxygen clusters by identifying maximal cliques (defined below) in a graph constructed from a given protein. For a given protein structure, MUG extracts the coordinates of all oxygen atoms, either from amino acids cofactors, or water, depending on the input parameters. Nitrogen atoms are not considered as ligands, as their role in metal-binding is controversial.[37] The distances between each pair of oxygen atoms are calculated and a graph *G*(*V*,*E*) is constructed, where *V* is the vertex set and *E* is the edge set of *G*. Each vertex represents an oxygen atom. An edge is assigned between two vertices if and only if the distance between these two vertices does

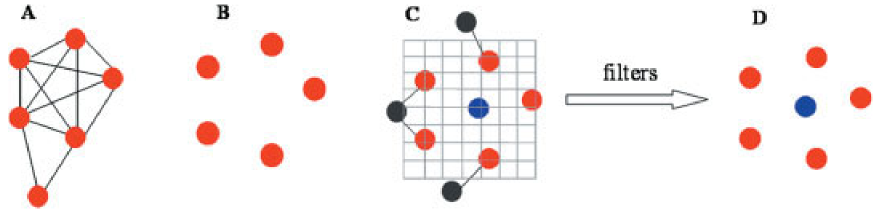

Figure 2

Schematic diagram of MUG. (**A**) The extracted oxygen atoms (red spheres) and the constructed graph *G*. (**B**) A maximal clique of size five in graph *G*. (**C**) The calcium center (blue sphere) of the cluster. The dark gray spheres represent carbon atoms covalently connected to a ligand oxygen atom. (**D**) Predicted ligand group and Ca^{2+} position.

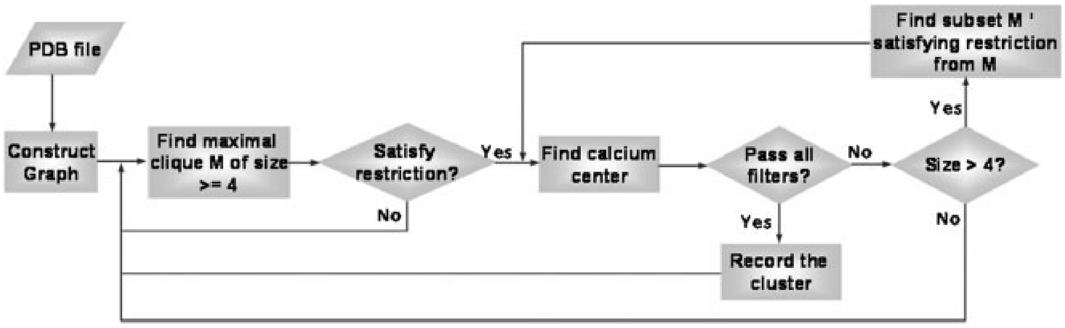

Figure 3

Flow chart of MUG. Restriction: a cluster must have at least 4 oxygen atoms of specified types where at least n (n = 2, 3, 4) oxygen atoms are from amino acids. Both the type and n are the input parameters.

not exceed a preset O-O cutoff of 6.0 Å. Figure 2(A) illustrates a graph containing six vertices where each vertex represents an oxygen atom.

A clique Q is a subset of V such that every two vertices in Q are adjacent. A maximal clique M is a clique that is not a proper subset of any other clique. The size of a maximal clique is the number of vertices it contains. Each maximal clique corresponds to an oxygen cluster. Intuitively, oxygen atoms that belong to a maximal clique are spatially close to each other and thus form a cluster. Figure 2(B) illustrates a maximal clique of size five.

Enumerating all maximal cliques of a general graph is NP-hard by computational complexity theory, requiring more than polynomial computational time to process.[38,39] Fortunately, in the context of the graph generated above, the size of any maximal clique is no more than twelve, because oxygen atoms maintain some distance from each other due to charge repulsion. This convenient property allows us to apply a well-established algorithm and produce all the maximal cliques efficiently (within $O(n)$ time, where n is then number of vertices in the graph). As a result, MUG implements the algorithm of Bron and Kerbosch[40] to find all the maximal cliques of size at least four in graph G.

Identifying CC

MUG identifies the CC within each oxygen cluster by finding the point that reaches the least penalty according to a penalty function defined in Eq. (2). Suppose $o_1,o_2,...,o_n$ are n oxygen atoms in an oxygen cluster and the coordinates of o_i are (x_i,y_i,z_i), $1 \leq i \leq n$. Moreover, we assume the first m oxygen atoms $o_1,...,o_m$, $1 \leq m \leq n$ are all nonwater oxygen atoms among $o_1,o_2,...,o_n$. Initially, let o be the center of $o_1,o_2,...,o_n$, that is, its coordinates (x,y,z) are:

$$x = \frac{\sum_{i=1}^{n} x_i}{n}, \quad y = \frac{\sum_{i=1}^{n} y_i}{n}, \quad z = \frac{\sum_{i=1}^{n} z_i}{n} \tag{1}$$

MUG divides the space within 1.5 Å from o into fine grids where the distance between two adjacent grid points is 0.1 Å. For each grid point D of coordinates (x,y,z) the penalty $P(x,y,z)$ is calculated with Eq. (2), and the point of the least penalty is the CC of that cluster.

$$\begin{aligned} P(x,y,z) = {} & \sum_{i=1}^{n} \lambda_{b,i}(b_i - b_i^*)^2 + \sum_{i=1}^{n} \lambda_{l,i}(l_i - l_i^*)^2 \\ & + \sum_{i=1}^{n} \lambda_{\theta,i}(\theta_i - \theta^*)^2 I(\theta_i < \theta^*) + \sum_{i=1}^{n} \lambda_{\phi}(\phi_i - \phi^*)^2 I(\phi_i < \phi^*) \\ & + \begin{cases} \lambda_s(s_m - s^*) + m^*\lambda_t(b'm - b'_1)^2 & \text{No irregular bidentate pair;} \\ \lambda_s(s_{m-1} - s^*) + (m-1) * \lambda_t(b'_{m-1} - b'_1) & \text{Otherwise} \end{cases} \end{aligned} \tag{2}$$

The detailed description of the variables and constants in Eq. (2) is given in the supporting information. A brief explanation and illustration [Fig. 1(B)] are presented as follows:

n is the total number of oxygen atoms in an oxygen cluster; m is the number of nonwater oxygen atoms; i indexes oxygen atom i. b^*,l^*,θ^*,ϕ^* are reference values of calcium-oxygen distance, calcium-carbon distance, calcium-oxygen-carbon angle, and dihedral angle, respectively, obtained through our previous statistical analysis of documented Ca^{2+}-binding sites (Supporting Table S9). b, l, θ, ϕ, s_m, s_{m-1}, b'_m, b'_{m-1}, b'_1 are calculated values with respect to grid point, $D(x,y,z)$(D-oxygen (D-O) distance, D-carbon (D-C) distance, D-oxygen-carbon (D-O-C) angle, etc.). λ is the weight assigned to each term; $I(\cdot)$ is an indication function where $I(\cdot) = 1$ if the inequality inside the parentheses is satisfied and $I(\cdot) = 0$ otherwise.

In Eq. (2), the first term accounts for a penalty incurred by the deviation of D-O distance from a reference value; the second term for D-C distance penalty; the third term for the D-O-C angle penalty; the fourth term for dihedral angle penalty; whereas the fifth and sixth terms apply less penalty to grid points of small average D-O distance and symmetric distances to oxygen atoms, respectively.

The dynamic penalty function is inspired by various energy functions[41] which have been used to empirically approximate free energy or potential energy of a protein structure. However, there are three major differences between our penalty function and energy functions: (1) The value of the penalty function does not correspond to energy but to a pure numerical value to reflect a statistically more suitable position of the calcium ion. (2) The terms in this penalty function are purely structural, accounting for structural preference. Unlike energy functions, there are no electrostatic and solvation terms. (3) This penalty function assumes a different formula and assigns different weights and reference values with respect to different grid points, which is further discussed in the supporting information.

Filters

Filters are used to inspect the structure of each (cluster, CC) pair. If a structure passes all the filters, the oxygen cluster is predicted to be a ligand group and the CC is the predicted calcium location. Two sets of filters are currently applied in MUG: distance-filters and angle-filters (Supporting Table S10). Distance-filters examine whether the distances between the CC and oxygen atoms, the distances between the CC and the carbon atoms connected to the ligand oxygen atoms, and the ratio of these two distances are within a certain range. The angle-filters examine whether certain angles are within a certain range, including angles among the CC, oxygen atom, and carbon atom connected to the oxygen atom; and the dihedral angles between the plane formed by the side chain carboxyl group (—COO) and the plane formed by the two carboxyl oxygen atoms (bidentate pair) and CC.

Implementing and rendering tools

The MUG program was implemented in Java (http://java.sun.com/). PyMOL (http://pymol.sourceforge.net/) was used for molecular visualization.

Datasets

All X-ray crystallography resolved protein structures used in this study were obtained from PDB.[42] The training dataset reproduced from Nayal and Di Cera[26] contains 32 proteins with 62 Ca^{2+}-binding sites. Testing dataset I, which contains 19 proteins with 48 Ca^{2+}-binding sites, is reproduced from Schymkowitz *et al.*[27] Testing dataset II contains 92 Ca^{2+}-binding sites from 44 proteins of Pidcock and Moore[12] obtained from 515 fully-normalized crystal structures of CaBPs deposited in PDB from 1994 to 1999 with a resolution between the range 1.0 to 2.5 Å. Testing dataset III reproduced from Liang *et al.*[43] contains 40 CaBPs with 91 Ca^{2+}-binding sites and 14 non-CaBPs. A combined dataset is obtained by incorporating the CaBPs from testing sets I, II, and III, where duplicate references are removed. This combined dataset contains 109 CaBPs with 215 sites. A low-coordination dataset is formed by selecting all proteins containing at least one site of coordination number three or less from the combined dataset. The coordination number of a Ca^{2+}-binding site in this study is defined to be the number of oxygen atoms from amino acids within 3.5 Å of the calcium ion. 3.5 Å is normally considered to be the maximum possible distance between ligand oxygen atoms and the calcium ion.[26,35,44] The low-coordination set contains 23 proteins with 27 sites of coordination number less than or equal to three and a total of 44 Ca^{2+}-binding sites. A cofactor-coordinating dataset is constructed by selecting from the combined dataset of all proteins containing at least one site with cofactor oxygen atom(s). The cofactor-coordinating set contains seven proteins with 16 sites of least one cofactor oxygen atom and a total of 22 Ca^{2+}-binding sites. A multiple-binding dataset is constructed by selecting from the combined dataset all proteins containing at least one multiple-binding site. The multiple-binding set contains five proteins with 18 multiple-binding sites and a total of 26 Ca^{2+}-binding sites.

Measurements

MUG predicts both calcium positions and ligand clusters. If the structure of a (cluster, CC) pair passes all the filters, the cluster is a qualified cluster. A qualified cluster is a predicted ligand group and its CC is the predicted calcium position. A correct prediction (CP) is a qualified cluster where the CC falls within a cutoff distance (3.5 Å by default) to a true site (TS), that is, a documented site. A truly predicted site (TPS) is a documented Ca^{2+}-binding site where there is at least one CP for it.

Sensitivity (SEN), selectivity (SEL), and the smallest deviation between predicted Ca^{2+} and documented Ca^{2+} position are the three measurement criteria of the performance of MUG. SEN represents the percentage of TPS in the total documented sites and SEL represents the per-

Table I
Summary Results of Testing Datasets I, II, and III

Dataset	Total protein	Total sites	Total predicted sites	Total predictions	Total correct predictions	SEN%	SEL%	R[a]	CLAG
I	19	48	45	243	230	94	95	0.22	43
II	44	92	83	457	317	90	69	0.38	66
III	54	91	83	468	344	91	74	0.49	63

[a] *R*: Mean deviation (in Å) between the documented and predicted sites.

centage of CP in total predictions (TP), respectively, that is

$$\mathrm{SEN} = \mathrm{TPS}/\mathrm{TS}$$

$$\mathrm{SEL} = \mathrm{CP}/\mathrm{TP}$$

RESULTS

Analysis of Ca^{2+}-binding sites in datasets

As shown in Table I and Supporting Tables S2, S3, and S4, there are a total of 109 (107 with resolution ≤ 2.5 Å) CaBPs with 215 sites in the three testing datasets. Two proteins (1ALA.pdb and 1B9O.pdb) are shared by testing datasets I and II, two (1SNC.pdb and 2PRK.pdb) are shared by testing datasets I and III, and four (1CEL.pdb, 1ESL.pdb, 1KIT.pdb, and 1SRA.pdb) are shared by testing datasets II and III. Three proteins (3CLN.pdb, 1SNC.pdb, and 3EST.pdb) are in both the training dataset and testing dataset I, four (1OVA.pdb, 1SNC.pdb, 2POR.pdb and 4SBV.pdb) are in both the training dataset and testing dataset III, whereas no protein is duplicated between the training dataset and testing dataset II. According to the classifications obtained from Structural Classification of Proteins (SCOP[45]), all three testing datasets contain all alpha proteins, alpha and beta proteins, and all beta proteins (Supporting Table S5). All three testing datasets contain continuous (sites in 1AUI.pdb, site in 1B9O.pdb, sites in 1TCO.pdb), semicontinuous (second site in 2PRK.pdb, first site in 1OIL.pdb, second site in 1AXN.pdb), and discontinuous sites (first site in 1THM.pdb, site in 1AI4. pdb, site in 1PNK.pdb).

Ca^{2+}-binding sites in the testing sets have coordination numbers ranging from 1 to 8. 45 of 48 sites in testing dataset I have at least four ligand oxygen atoms from amino acids, two sites have three and one site has two ligand oxygen atoms from amino acids. In testing dataset II, 13 of 92 sites have fewer than four coordinating oxygen atoms from amino acids. In testing dataset III, 14 of 91 sites have fewer than four coordinating oxygen atoms from amino acids. For the combined dataset, Figure 4(A,B) present the distribution of the number of different Ca^{2+}-binding sites.

Input and output

Input data for the MUG program are taken from PDB structural data files. Outputs for MUG are the predicted

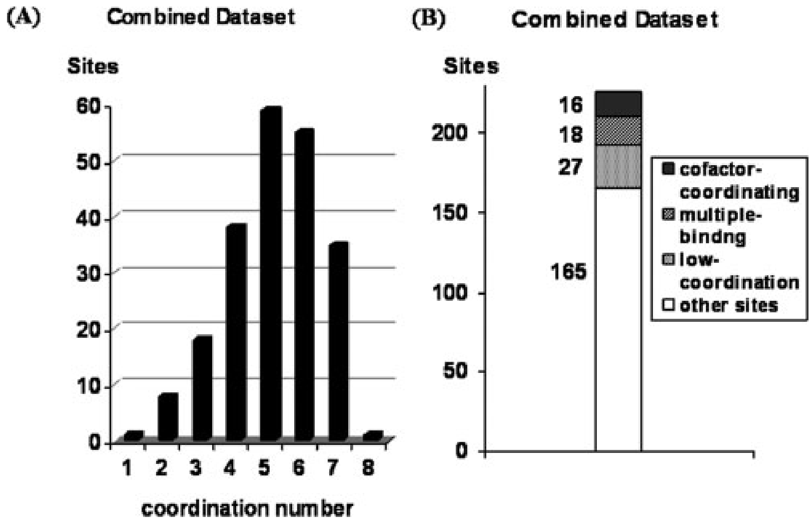

Figure 4
(**A**) Distribution of the number of Ca^{2+}-binding sites with respect to coordination number of the combined dataset. (**B**) Population distribution of multiple-binding sites, cofactor-coordinating sites and low-coordination sites in the combined dataset.

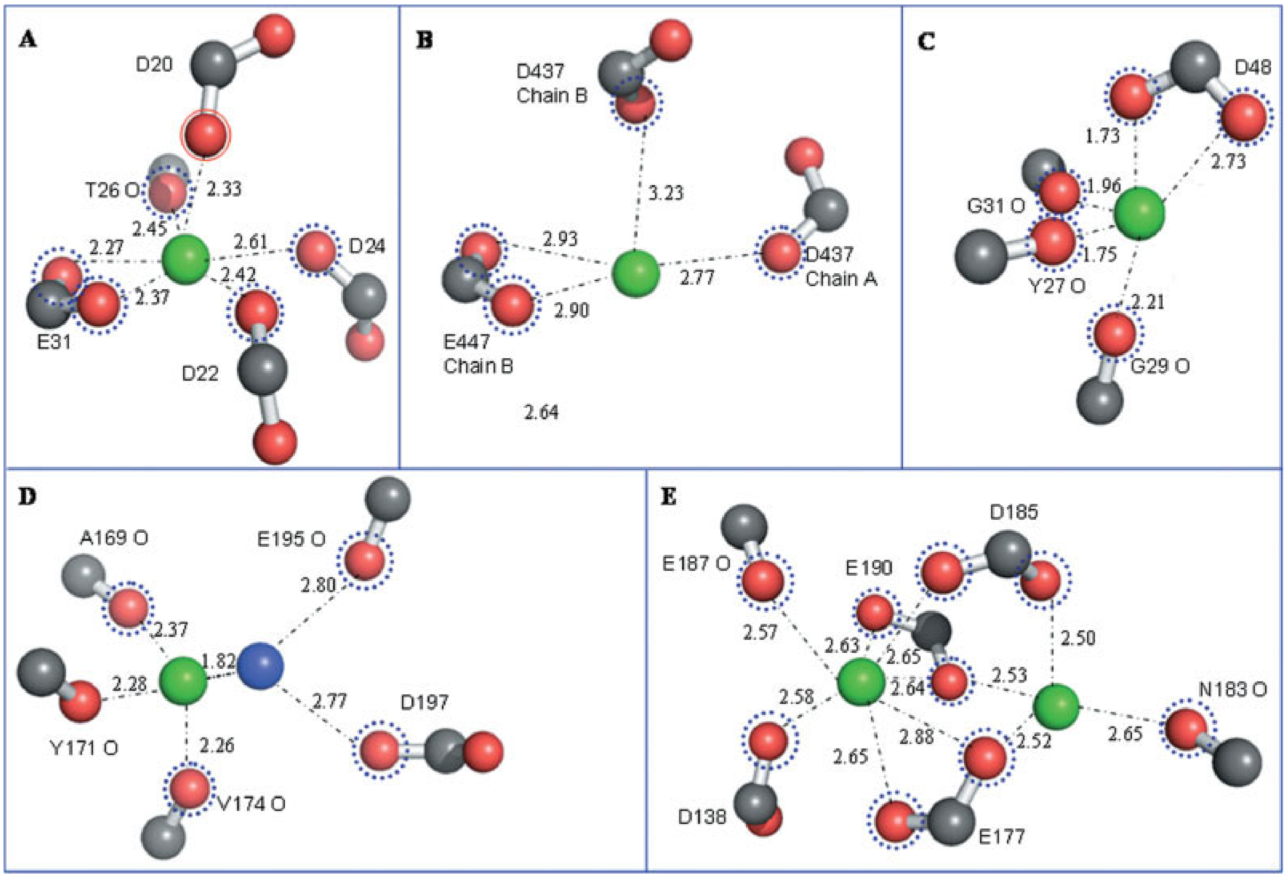

Figure 5

Calcium ions and their surrounding atoms for (**A**) calmodulin (sequence ID 1128, 3CLN.pdb), (**B**) hcv helicase (sequence ID 6434, 1HEI.pdb), (**C**) phospholipase A2 (sequence ID 2789, 1PSH.pdb), (**D**) subtilisin (sequence ID 1943, 1SBH.pdb), and (**E**) themolysin (sequence ID 2461 and 2462, 1FJ3.pdb). [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

ligand group and calcium location with respect to that ligand group. By default, the results presented in this study were obtained under the same set of input parameters without utilizing cofactor and water oxygen atoms.

Although a single documented calcium site could have more than one different set of predicted (ligand group, CC) pairs, we did not combine the multiple predictions for the following several reasons. First, prediction results (Table I) show that the number of predictions per documented site is small (around five). Second, for multiple-binding sites, the combination process may result in missing sites. Third, different ligand groups could provide useful information for the design of calcium pockets and CaBPs. Fourth, a protocol (described below) has been developed to select one group from multiple ligand groups. Related to the protocol, two definitions are provided: predicted ligand group could be either discrete, meaning no other group shares any atom with it or non-discrete with overlapping ligands between the predicted groups. When multiple predictions are made, we identify the following predicted groups as the most favorable ligand group: (1) discrete groups and (2) the ligand group containing the most atoms in a set of overlapping groups, and if two overlapping groups have the same number of atoms, the one with the most appropriate average distance between its predicted calcium location and its ligand atoms. "Appropriate" here means close to 2.5 Å because the mean calcium-oxygen distances are observed to range from 2.4 to 2.6 Å, based on statistical analysis (Supporting Table S9).[25]

Prediction of Ca^{2+}-binding sites of diverse structure architectures

Figure 5 illustrates documented Ca^{2+}-binding sites of different structural architectures. The circled oxygen atoms (red spheres) are the identified ligand atoms. The ones marked with dashed lines connected to documented Ca^{2+} (green spheres) are the true ligand atoms. In A, a regular pentagonal-bipyramid geometry with five oxygen atoms in the same plane and an additional oxygen above and below the center of the plane is seen in calmodulin (3CLN.pdb) which represents an important (EF-hand like) class of CaBPs. The distances from ligand atoms to Ca^{2+} is typically close to 2.4 Å, and the O-Ca-O angle is close to 72°. Because of this regularity, the predicted Ca^{2+} binding site is close to the documented site with a deviation of 0.21 Å. In B, a longer average distance (2.96 Å) between ligand atoms and Ca^{2+} is seen in hcv helicase (1HEI.pdb). The calcium ion is chelated between two different chains with one ligand atom from chain A and three from chain B. The deviation of predicted and documented position is 0.52 Å.

Figure 5(C) illustrates calcium ion (sequence ID 2789) of phospholipase A2 (1PSH.pdb) and its surrounding atoms. Three of the five ligand oxygen atoms are from carbonyl groups. The distance between the documented calcium ion and carbonyl oxygen of Y27 is 1.75 Å, which is lower than the Van der Waals radius between calcium and oxygen (1.76 Å). MUG still correctly identified the binding ligand atoms. However, because of the abnormally short distance between ligand atoms and calcium ion, the predicted position deviates 0.75 Å from the documented one.

Figure 5(D) illustrates calcium ion (sequence ID 1943) of subtilisin (1SBH.pdb) and its surrounding atoms. As the documented calcium ion has only three oxygen atoms (from A169, Y171, and V174) within 3.5 Å, this Ca^{2+}-binding site was expected to be difficult to predict. However, MUG still correctly identified it by identifying ligand oxygen atoms from A169, V174 and two other oxygen atoms from E195 and D197. The deviation between the predicted and documented ions is 1.82 Å.

Figure 5(E) illustrates the multiple sites of thermolysin (1FJ3.pdb) where the two calcium ions are close (3.80 Å) to each other. There are only three ligand residues for both sites, including two oxygen atoms that are ligands for both ions. MUG successfully identified both CLAGs for the two calcium ions, and the deviations between the predicted and documented locations were 0.19 Å and 0.25 Å, respectively.

General performance

Table I presents the summary results of testing datasets I, II, and III. For testing dataset I, the total of all predictions made by MUG was 243 (Column 5) and the total number of correct predictions was 230 (Column 6), which means the ratio of the number of predictions per documented site was 243/48 = 5.1, whereas the number of correct predictions per documented site was 230/48 = 4.8. 45 of 48 documented sites were predicted. Thus the sensitivity is 45/48 = 94% (Column 7), whereas the selectivity was 230/243 = 95% (Column 8). For documented sites that are correctly predicted, the mean deviation between the documented and predicted sites was 0.22 Å (Column 9). With the protocol described in section 3.2, 43 CLAGs of 45 Ca^{2+}-binding site of coordination number $\geq$4 are identified.

The sensitivities of all three testing sets are similar (~90%). However, the selectivities varied from 95% to 69%, which might result from different degrees of structural diversity in the three testing sets. The deviations of datasets II and III are larger than that of dataset I due to the fact that datasets II and III contain larger populations of sites having fewer than four coordinating oxygen atoms.

For the combined dataset, as shown in Figure 6(A), MUG was able to predict 186/188 Ca^{2+}-binding sites having four or more coordinating oxygen atoms and 11/18 sites coordinated by three oxygen ligands. However, it only predicted 1/8 sites having two coordinating protein oxygen atoms, and it fails to identify the site reported to have only one coordinating atom.

As shown in Figure 6(B), there is an inverse correlation between the coordination number of sites and mean deviation between the documented and predicted Ca^{2+} locations. For sites chelated by four or more oxygen atoms, the predicted positions of calcium ions are accurate (mean deviation less than 0.30 Å) because its ligand groups are almost always correctly identified. For documented sites having fewer than four coordinating oxygen atoms that are correctly predicted, there are typically larger deviations (around 2.0 Å) from documented positions. This is because some nonligand oxygen atom(s) are included in the qualified oxygen cluster. The inverse pattern is consistent with the phenomena that sites of larger coordination number usually have more geometrically-predictable structural architecture, and constraints are apparent for the position of the calcium ion.

Prediction on low-coordination dataset with water oxygen atoms

Besides protein oxygen atoms, calcium ions are frequently bound by structural water molecules. For Ca^{2+}-binding sites of low coordination number, water molecules play a role in forming a regular Ca^{2+}-binding pocket. For instance, in the N-terminal domain of e-selectin (1ESL.pdb) (see Fig. 7), the calcium ion (sequence ID 1269) is chelated by two protein oxygen atoms and five water oxygen atoms. In the absence of water molecules, the calcium ion is supported by two carbonyl oxygen atoms forming a low-structured triad. Conversely, with the inclusion of water molecules, the calcium ion is located in the center of a seven-oxygen pocket with distances to these six oxygen atoms ranging from 2.41 Å to 2.82 Å. In this section, we utilize oxygen atoms from both protein and structural water to predict Ca^{2+}-binding sites. The positions of the water molecule oxygen atoms are obtained directly from the PDB file. Algorithmically, the coordinates of both protein and

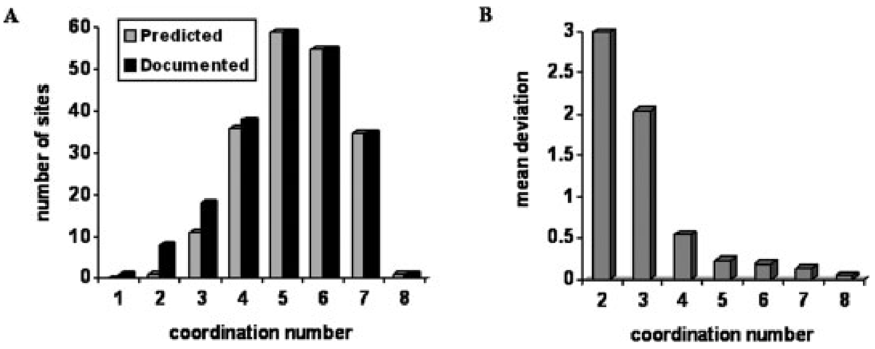

Figure 6

(**A**) Distribution of the number of documented sites (black bar) and true predicted sites (gray bar) with respect to coordination number i ($1 \leq i \leq 8$) (**B**) Mean deviations (in Å) between the documented and predicted sites with respect to coordination number i ($1 \leq i \leq 8$).

water oxygen atoms are used to identify the CC within each oxygen cluster, and it is required that in a ligand group there are at least four oxygen atoms, at least two of which are from amino acids. We then test the modified program on the low-coordination dataset.

For black bars in Figure 7, a cutoff at 0.5 Å of the deviation between the documented and predicted sites is observed, where a majority of the deviations are observed below the line at 0.5 Å. A further inspection reveals that each of these small deviations corresponds to the CC of the ligand group that is a subset of the CLAG. Each of these large deviations corresponds to the CC of the ligand group that contains some nonligand atom(s), which probably results from the lack of sufficient oxygen atoms ($\leq$3), even when both water oxygen and oxygen from amino acid are considered. It is also clearly seen that deviations when water is excluded are larger than those observed when water is included. 40 of 44 documented sites are predicted, and 24 (12 when excluding water) of 27 documented sites of coordination number $\leq$3 are predicted. A total of 2756 predictions are made, of which 320 are within 3.5 Å of a documented site. Thus 63 predictions are made, on average, per documented site, leading to 12% selectivity. Although low selectivity suggests that water oxygen atoms are not as good an indicator of Ca^{2+}-binding sites as amino acid oxygen atoms, water molecules participating in Ca^{2+}-binding sites regularize and enable the binding pocket, thus playing an essential role in identification of half of the sites of coordination number $\leq$3.

Prediction on multiple-binding and cofactor-coordinating datasets

Occasionally, the same oxygen atom binds more than one calcium ion, forming multiple-binding sites [Fig. 5(E)]. In the multiple-binding dataset, there are nine pairs of calcium ions that share at least one oxygen atom. For the multiple-binding set, all of the multiple-binding sites are predicted. The number of average predictions per documented calcium ion is 4.3, with 3.8 within 3.5 Å of the documented Ca^{2+}.

Other than amino acids, ligand atoms may be provided by cofactors [Fig. 8(A)]. To account for this, we utilized oxygen atoms from both the protein and its cofactors to predict Ca^{2+}-binding sites. Algorithmically, the coordinates of both protein and cofactor oxygen atoms are used to identify the CC within each oxygen cluster. We then test the modified program on the cofactor-coordinating dataset.

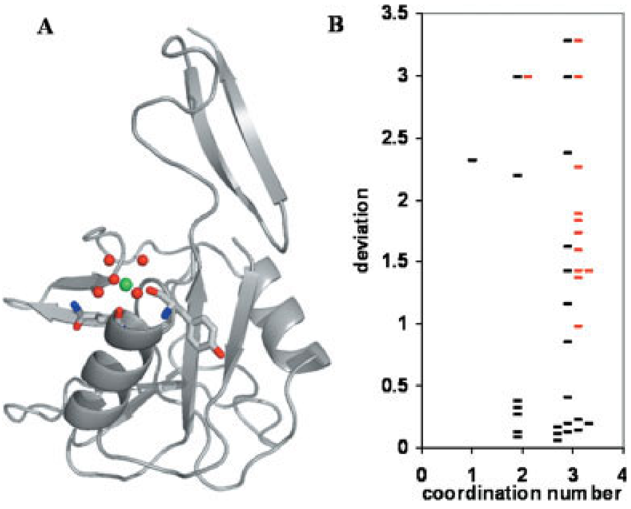

Figure 7

(**A**) E-selectin (1ESL.pdb) and calcium ion (sequence ID 1269). (**B**) Distribution of deviations (in Å) (black bars for inclusion of water, red bars for exclusion of water) between the documented and predicted sites with respect to coordination number i ($1 \leq i \leq 3$) for sites that are predicted. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

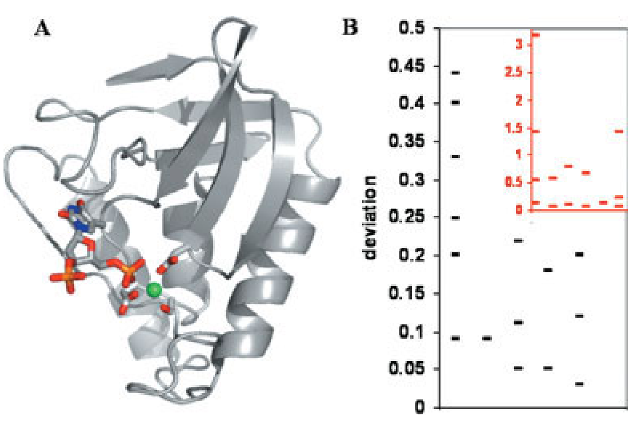

Figure 8

(**A**) Staphylococcus nuclease (1SNC.pdb), its cofactor tetrahydropyranyl and calcium ion (sequence ID 1084). (**B**) Distribution of deviations (in Å) (black bars for inclusion of cofactor atoms, red bars for exclusion of cofactor atoms) between the documented and predicted sites for sites that are predicted. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

On the cofactor-coordinating dataset, the average number of predictions per documented site is 5.2, where 4.0 are correct predictions. 15 of 16 cofactor-coordinating sites are predicted [Fig. 8(B)], where the missing site is found in ovalbumin (1OVA.pdb), which has four ligand oxygen atoms, two of the four from phosphite. The two ligands from amino acids are 2.56 Å and 1.85 Å distant from the documented Ca^{2+}, whereas the two phosphite ligand atoms are 3.30 Å and 1.61 Å relative to the documented Ca^{2+}. This site is irregular because the distance variation is large. For these 15 predicted sites, the deviations between predicted and documented Ca^{2+} locations were less than 0.5 Å, significantly smaller than those excluding cofactor atoms [Fig. 8(B)].

DISCUSSION

Design of algorithm

It is largely reported in the literature that the coordination number of Ca^{2+}-binding sites varies and Ca^{2+}-binding sites exhibit large diversification of both bond angle and bond length. To ensure the prediction of different classes of Ca^{2+}-binding sites in proteins with irregular geometric parameters, the MUG algorithm was developed based on three major considerations—finding maximal cliques instead of finding clusters of fixed size, addition of a penalty function, and selection of input parameters based on structural statistics of Ca^{2+}-binding sites.[25]

A maximal clique was used in this study to represent a potential ligand group (cluster). By its definition, no restrictions are imposed on the number of atoms in the cluster, as long as all atoms are close to each other, which fits the true Ca^{2+}-binding sites where a coordination number between four and seven is most frequently observed.[25] Furthermore, as shown in Supporting Table S1, about 89% of Ca^{2+}-binding sites in the training dataset consist of at least four ligands. And according to the analysis of a large dataset, more than 87% of non-EF-hand binding sites have at least four ligands, and all EF hand and pseudo-EF Ca^{2+}-binding sites have a coordination number larger than four.[25] Consequently, MUG is designed to find maximal cliques of size at least four.

Because a maximal clique may be a true ligand group, the primary goal of the dynamic penalty function is to locate the CC appropriately so that the structure of each (cluster, CC) pair is consistent with that of the true Ca^{2+}-binding sites. Also, because there are situations when some nonligand atoms are so close to a ligand atom that they appear in the maximal clique, another goal of the dynamic penalty function is to locate the CC close to the ligand atoms yet distant from the nonligand atoms. A novel feature of MUG is the recursive removal of an atom from an originally unqualified oxygen cluster. To ensure that the removed atom, which is the atom most distant from the CC, is a nonligand atom, the penalty function was devised to be dynamic, which assumes different formulae and different reference values and weights with respect to different grid points. The supporting material includes a further detailed description of the penalty function.

The cutoff of 6.0 Å in the graph construction is selected according to empirical results from our previous study[35] which indicated 6.0 Å was the maximum distance between vertices to ensure capture of all relevant binding ligands and exclude those that would more appropriately be classified as having second shell occupancy. For determination of the search area to find the least-penalty point, an analysis shows that documented calcium ions are almost always within 1.5 Å from the initial point calculated with Eq. (1), provided that the oxygen cluster contains all oxygen atoms within 3.5 Å of the calcium ion. The analysis of the training dataset is listed in Supporting Table S1. MUG therefore divides the space within 1.5 Å from the initial point into fine grids where the distance between two adjacent grid points is 0.1 Å. For the selection of reference values in the penalty function, a thorough statistical analysis was conducted[25] and key results are listed in Supporting Table S9. The mean value of each characteristic is taken as the corresponding reference value. Weights are assigned to obtain a satisfactory performance of the training dataset. Supporting Table S9 and Figure S1 present statistical analysis results and reference values, respectively. The input parameters of the geometrical filters are set to be lower or higher than the sample mean plus or minus two standard deviations. These sample means and standard deviations were obtained from prior analysis.[25] The restrictions set in filters are given in Supporting Table S10 while the statistical analysis results are presented in Supporting Table S9.

Comparison with previous algorithms

Testing dataset III is reproduced from the dataset of Liang *et al.*[43] Their FEATURE program determined if a query site is a Ca^{2+}-binding site by scoring a variety of biochemical properties within a predefined radius including secondary structure, polarity, charge, acidity etc. By adjusting the cutoff score, FEATURE achieves different (sensitivity, selectivity) pairs. On testing dataset III, FEATURE has 37% selectivity to achieve 90% sensitivity, and has 68% sensitivity in order to achieve 74% selectivity. Its standard of correct prediction of Ca^{2+}-binding sites is within 6.0 Å of the documented calcium ion, in contrast to MUG's 3.5 Å. Even so, MUG has 74% selectivity and 90% sensitivity on the same dataset. Moreover, because this dataset of 91 documented sites contains 14 sites of coordination number ≤ 3, FEATURE's 68% sensitivity indicates that a significant number of sites (≥ 15) of coordination number ≥ 4 were not identified. On the other hand, MUG is able to predict 99% of the sites of coordination number ≥ 4, and half of the sites of coordination number ≤ 3. Our results show the characteristic roles of appropriately combined structural parameters in the Ca^{2+}-binding sites.

Testing dataset II is reproduced from the dataset that was originally constructed for statistical analysis purposes by Pidcock and Moore.[12] We previously reported GG (Graph and Geometry) algorithm to detect Ca^{2+}-binding sites and tested it using this dataset.[35] GG was based on oxygen clusters of size exactly four and the CC was determined at an equidistant center within each cluster. It achieved 87% sensitivity and 74% selectivity, if a prediction within 3.5 Å distance to the documented position is assumed as a correct prediction. The number of predictions per documented site was more than 10. The deviation between predicted and documented location was around 0.73 Å. As a comparison, MUG made five predictions per documented sites and predicted Ca^{2+} location within 0.38 Å to documented positions with 90% sensitivity and 69% selectivity. Furthermore, for binding sites of coordination number ≤ 3, GG reported a 45% prediction rate, whereas MUG achieves an 86% prediction rate when structural water molecules are utilized.

Testing dataset I is reproduced from Schymkowitz *et al.*[27] Their algorithm measures the contribution of atoms from 2.3 Å to 5.7 Å from binding sites containing 4–6 ligand atoms. If we regard a predicted calcium ion as a correct prediction when it is within 3.5 Å from a documented one, the algorithm predicted 43 of 48 documented Ca^{2+}-binding sites. However, it failed to identify five documented sites including three sites of coordination number ≤ 3 and two sites of coordination number ≥ 4, one of which is a multiple-binding site, leading to 90% sensitivity (selectivity was not reported). Comparatively, we have shown that MUG identifies all sites of coordination number ≥ 4, leading to 94% sensitivity and 95% selectivity. When structural water molecules are utilized, two of three sites of coordination number ≤ 3 are identified. If we regard a predicted calcium ion as a correct prediction if it is within 1.0 Å from a documented one, their algorithm predicted Ca^{2+} location within 0.25 Å to documented positions on average. 40 of 48 sites were predicted, leading to 83% sensitivity (selectivity was not reported). Comparatively, MUG predicted 44 of 48 sites, predicted Ca^{2+} location within 0.19 Å to documented positions on average, with 92% sensitivity and 77% selectivity. Moreover, while the prediction of multiple-binding sites and prosthetic groups are problematic by Schymkowitz's algorithm, the MUG algorithm is able to predict all multiple-binding sites in the selected datasets. Prosthetic groups or protein cofactors that are involved in the Ca^{2+}-binding sites can be unambiguously predicted.

In summary, by exploring the multiple geometric characteristics of Ca^{2+}-binding sites and utilizing water molecules and protein cofactors, we have developed the MUG algorithm which is able to predict Ca^{2+}-binding sites of different classes in proteins with high resolution. This represents a major achievement towards understanding the role of calcium in biological systems by facilitating accurate prediction of Ca^{2+}-binding sites in proteins. Our results largely extend our capability to predict diversified Ca^{2+}-binding sites in proteins, especially for Ca^{2+}-binding sites with low coordination numbers from proteins such as extracellular proteins. Proteins with different prosthetic groups such as carbohydrates and phosphate groups involved in Ca^{2+}-binding can also be predicted. A website has been developed to allow scientists in the field to apply this developed algorithm to various proteins through a user-friendly interface. In addition, this algorithm can be further developed to enhance the design of CaBPs with tailored biological functions, such as sensors and biosensors.

ACKNOWLEDGMENTS

We thank Susan Fu and Hai Deng for their initial work.

REFERENCES

1. Herzberg O, Moult J, James MN. A model for the Ca2+-induced conformational transition of troponin C: a trigger for muscle contraction. J Biol Chem 1986;261:2638–2644.
2. Holmes KC, Popp D, Gebhard W, Kabsch W. Atomic model of the actin filament. Nature 1990;347:44–49.
3. Mann KG, Nesheim ME, Church WR, Haley P, Krishnaswamy S. Surface-dependent reactions of the vitamin K-dependent enzyme complexes. Blood 1990;76:1–16.

4. Linse S, Forsen S. Determinants that govern high-affinity calcium binding. Adv Second Messenger Phosphoprotein Res 1995;30:89–151.
5. Huang Y, Zhou Y, Yang W, Butters R, Lee HW, Li SY, Castiblanco A, Brown EM, Yang JJ. Identification and dissection of Ca^{2+}-binding sites in the extracellular domain of Ca^{2+}-sensing receptor. J Biol Chem 2007;282:19000–19010.
6. Stathopulos PB, Li G-Y, Plevin MJ, Ames JB, Ikura M. Stored Ca2+ depletion-induced oligomerization of stromal interaction molecule 1 (STIM1) via the EF-SAM region: an initial mechanism for capacitive Ca2+ entry. J Biol Chem 2006;281:35855–35862.
7. Hwang J-I, Kim HS, Lee JR, Kim E, Ryu SH, Suh P-G. The interaction of phospholipase C-beta3 with shank2 regulates mGluR-mediated calcium signal. J Biol Chem 2005;280:12467–12473.
8. Hu J, Spiegel A. Structure and function of the human calcium-sensing receptor: insights from natural and engineered mutations and allosteric modulators. J Cell Mol Med 2007;11:908–922.
9. Banci L, Bertini I, Mangani S. Integration of XAS and NMR techniques for the structure determination of metalloproteins: examples from the study of copper transport proteins. J Synchrotron Radiat 2005;12:94–97.
10. Feng Y, Liu D, Yao H, Wang J. Solution structure and mapping of a very weak calcium-binding site of human translationally controlled tumor protein by NMR. Arch Biochem Biophys 2007;467:48–57.
11. Yang W, Lee HW, Hellinga H, Yang JJ. Structural analysis, identification and design of calcium-binding sites in proteins. Proteins 2002;47:344–356.
12. Pidcock E, Moore GR. Structural characteristics of protein binding sites for calcium and lanthanide ions. J Biol Inorg Chem 2001;6:479–489.
13. Monzingo AF, Matthews BW. Binding of *N*-carboxymethyl dipeptide inhibitors to thermolysin determined by X-ray crystallography: a novel class of transition-state analogues for zinc peptidases. Biochemistry 1984;23:5724–5729.
14. Chakrabarti P. Geometry of interaction of metal ions with sulfur-containing ligands in protein structures. Biochemistry 1989;28:6081–6085.
15. Marsden BJ, Shaw GS, Sykes BD. Calcium binding proteins. Elucidating the contributions to calcium affinity from an analysis of species variants and peptide fragments. Biochem Cell Biol 1990;68:587–601.
16. Glusker JP. Structural aspects of metal liganding to functional groups in proteins. Adv Protein Chem 1991;42:1–76.
17. Lippard SJ, Berg JM. Principles of bioinorganic chemistry. Mill Valley, Calif.: University Science Books; 1994; pp 411.
18. Bagley SC, Altman RB. Characterizing the microenvironment surrounding protein sites. Protein Sci 1995;4:622–635.
19. Bagley SC, Altman RB. Conserved features in the active site of nonhomologous serine proteases. Fold Des 1996;1:371–379.
20. Nelson MR, Chazin WJ. Structures of EF-hand Ca(2+)-binding proteins: diversity in the organization, packing and response to Ca2+ binding. Biometals 1998;11:297–318.
21. Harding MM. The geometry of metal-ligand interactions relevant to proteins. Acta Crystallogr D Biol Crystallogr 1999;55(Pt 8):1432–1443.
22. Nelson MR, Thulin E, Fagan PA, Forsen S, Chazin WJ. The EF-hand domain: a globally cooperative structural unit. Protein Sci 2002;11:198–205.
23. Babu CS, Dudev T, Casareno R, Cowan JA, Lim C. A combined experimental and theoretical study of divalent metal ion selectivity and function in proteins: application to E. coli ribonuclease H1. J Am Chem Soc 2003;125:9318–9328.
24. Dudev T, Lim C. Principles governing Mg. Ca, and Zn binding and selectivity in proteins. Chem Rev 2003;103:773–787.
25. Kirberger M, Wang X, Deng H, Yang W, Chen G, Yang JJ. Statistical analysis of structural characteristics of protein Ca^{2+}-binding sites. JBIC 2008;13:1169–1181.
26. Nayal M, Di Cera E. Predicting Ca^{2+}-binding sites in proteins. Proc Natl Acad Sci USA 1994;91:817–821.
27. Schymkowitz JW, Rousseau F, Martins IC, Ferkinghoff-Borg J, Stricher F, Serrano L. Prediction of water and metal binding sites and their affinities by using the Fold-X force field. Proc Natl Acad Sci USA 2005;102:10147–10152.
28. McPhalen CA, Strynadka NC, James MN. Calcium-binding sites in proteins: a structural perspective. Adv Protein Chem 1991;42:77–144.
29. Wei L, Altman RB. Recognizing protein binding sites using statistical descriptions of their 3D environments. Pac Symp Biocomput World Scientific; 1998;497–508.
30. Barker JA, Thornton JM. An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis. Bioinformatics 2003;19:1644–1649.
31. Yamashita MM, Wesson L, Eisenman G, Eisenberg D. Where metal ions bind in proteins. Proc Natl Acad Sci USA 1990;87:5648–5652.
32. Lin C-T, Lin K-L, Yang C-H, Chung I-F, Huang C-D, Yang Y-S. Protein metal binding residue prediction based on neural networks. Int J Neural Syst 2005;15:71–84.
33. Lin H, Han L, Zhang H, Zheng C, Xie B, Cao Z, Chen Y. Prediction of the functional class of metal-binding proteins from sequence derived physicochemical properties by support vector machine approach. BMC Bioinformatics 2006;7(Suppl 5):S13.
34. Jambon M, Imberty A, Deleage G, Geourjon C. A new bioinformatic approach to detect common 3D sites in protein structures. Proteins 2003;52:137–145.
35. Deng H, Chen G, Yang W, Yang JJ. Predicting calcium-binding sites in proteins—a graph theory and geometry approach. Proteins 2006;64:34–42.
36. Zhou Y, Yang W, Kirberger M, Lee HW, Ayalasomayajula G, Yang JJ. Prediction of EF-hand calcium-binding proteins and analysis of bacterial EF-hand proteins. Proteins 2006;65:643–655.
37. Vallee BL, Auld DS. Zinc coordination, function, and structure of zinc enzymes and other proteins. Biochemistry 1990;29:5647–5659.
38. Lawler EL, Lenstra JK, Rinnooy Kan AHG. Generating all maximal independent sets: NP-hardness and polynomial-time algorithms. SIAM J Comput 1980;9:558–565.
39. Tomita E, Tanaka A, Takahashi H. The worst-case time complexity for generating all maximal cliques. Volume 3106. LNCS. Heidelberg: Springer Berlin; 2004. pp 161–170.
40. Bron C, Kerbosch J. Algorithm 457: finding all cliques of an undirected graph. Commun ACM 1973;16:575–579.
41. Zhang X, Minear RA, Barrett SE. Characterization of high molecular weight disinfection byproducts from chlorination of humic substances with/without coagulation pretreatment using UF-SEC-ESI-MS/MS. Environ Sci Technol 2005;39:963–972.
42. Bernstein FC, Koetzle TF, Williams GJB, Meyer EFJ, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. The Protein Data Bank: A computer-based archival file for macromolecular structures. J Mol Biol 1977;112:535–542.
43. Liang MP, Brutlag DL, Altman RB. Automated construction of structural motifs for predicting functional sites on protein structures. Pac Symp Biocomput 2003:204–215.
44. Dudev T, Lin YL, Dudev M, Lim C. First-second shell interactions in metal binding sites in proteins: a PDB survey and DFT/CDM calculations. J Am Chem Soc 2003;125:3168–3180.
45. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.