Notes for Ca^{2+}-binding sites in proteins

Shushan He, 2019

Outline

1. What to do?
2. How to do?
   - Statistical analysis of structural characteristics of protein Ca^{2+}-binding sites
   - Prediction of EF-Hand Calcium-Binding Proteins and Analysis of Bacterial EF-Hand Proteins
   - Integration of Diverse Research Methods to Analyze and Engineer Ca2+-Binding Proteins: From Prediction to Production
   - GG: A Graph Theory and Geometry Approach
   - MUG (MUltiple Geometries)
   - MUG^C

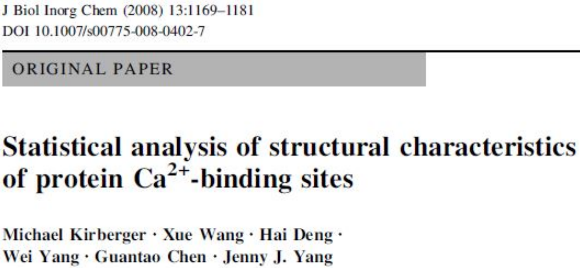
What to do?

- Identifying Ca^{2+}-binding sites in proteins is the first step towards understanding the molecular basis of diseases related to Ca^{2+}-binding proteins.
- **Challenge:** these sites are identified in structures either through X-ray crystallography or NMR analysis. However, Ca^{2+}-binding sites are not always visible in X-ray structures due to flexibility in the binding region or low occupancy in a Ca^{2+}-binding site. Similarly, both Ca^{2+} and its ligand oxygens are not directly observed in NMR structures.
- **Goal:** To improve our ability to predict Ca^{2+}-binding sites in both X-ray and NMR structures.

Outline

1. What to do?

2. How to do?
   - Statistical analysis of structural characteristics of protein Ca^{2+}-binding sites
   - Prediction of EF-Hand Calcium-Binding Proteins and Analysis of Bacterial EF-Hand Proteins
   - Integration of Diverse Research Methods to Analyze and Engineer Ca2+-Binding Proteins: From Prediction to Production
   - GG: A Graph Theory and Geometry Approach
   - MUG (MUltiple Geometries)
   - MUG^C



J Biol Inorg Chem (2008) 13:1169–1181
DOI 10.1007/s00775-008-0402-7

ORIGINAL PAPER

Statistical analysis of structural characteristics of protein Ca^{2+}-binding sites

Michael Kirberger · Xue Wang · Hai Deng ·
Wei Yang · Guantao Chen · Jenny J. Yang

Summary

- a comprehensive statistical analysis of calcium-binding proteins from the Protein Data Bank to identify structural parameters associated with EF-hand and non-EF-hand Ca2+-binding sites.
- Comparatively, non-EF-hand sites utilize lower coordination numbers (6 ± 2 vs. 7 ± 1), fewer protein ligands (4 ± 2 vs. 6 ± 1), and more water ligands (2 ± 2 vs. 1 ± 0) than EF-hand sites.
- The orders of ligand preference for non-EF-hand and EF-hand sites, respectively: $H_2O(33.1\%) > side - chainAsp(24.5\%) > main - chaincarbonyl(23.9\%) > side - chainGlu(10.4\%)$,
  side-chain Asp (29.7%) > side-chain Glu (26.6%) > main-chain carbonyl (21.4%) > $H_2O(13.3\%)$.
- Less formal negative charge was observed in the non-EF-hand than in the EF-hand binding sites ($1 \pm 1 vs. 3 \pm 1$).

Summary

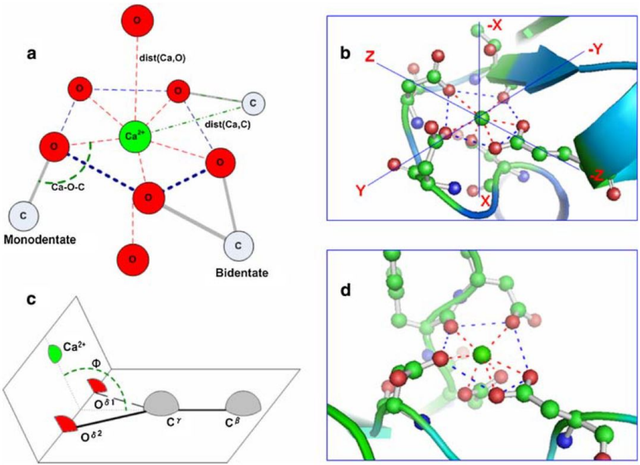
- Additionally, over 20% of non-EF-hand sites had formal charge values of zero due to increased utilization of water and carbonyl oxygen ligands.
- Moreover, the EF-hand sites presented a narrower range of ligand distances and bond angles than non-EF-hand sites, possibly owing to the highly conserved helix–loop–helix motif.
- Significant differences between ligand types (carbonyl, side chain, bidentate) demonstrated that angles associated with each type must be classified separately, and the EF-hand sidechain Ca-O-C angles exhibited an unusual bimodal quality consistent with an Asp distribution that differed from the Gaussian model observed for non-EF-hand proteins.
- The results of this survey more accurately describe differences between EF-hand and non-EF-hand proteins and provide new parameters for the prediction and design of different classes of Ca2+-binding proteins.

Ca2+-binding sites

- According to the nature of liganding residues, Ca2+-binding sites are often classified either as continuous or discontinuous.
    - continuous: where ligands originate in a continuous short sequence (Fig. 1b)
    - discontinuous: where multiple ligands are distant from each other within the sequence (Fig. 1d)

Ca2+-binding sites

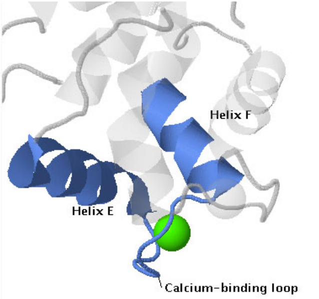
Fig. 1 **a** Illustration of key structural characteristics. The physical relationships between the Ca^{2+} ion, the ligand oxygen, and the ligand oxygen atoms covalently bound to carbon are defined by the angle Ca–O–C and distances dist(Ca,C) and dist(Ca,O). **b** Pentagonal-bipyramidal geometry of EF-hand binding site surrounding Ca^{2+} ion (Protein Data Bank, PDB, sequence ID 1168) from calmodulin (3cln.pdb). The water molecule at $-X$ is not shown. **c** Dihedral angle of bidentate ligands. **d** Pseudo-EF-hand binding site surrounding Ca^{2+} ion (PDB sequence ID 1420) from synaptotagmin I C2B domain (1uow.pdb). Pentagonal-bipyramidal geometry is formed by ligands distant from one another in the sequence

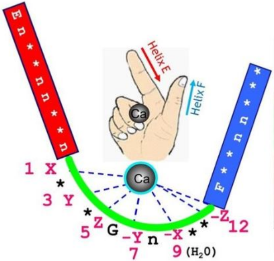


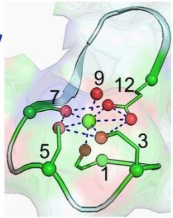
the EF-hand motif:

- first described by Kretsinger and Nockolds in 1973
- a continuous binding site that has been studied extensively.
- It is the most common motif associated with Ca2+ binding, characterized by a helix-loop-helix structure composed of approximately 30 amino acids
- can be subdivided into the following two classes found in the S100 protein N-termini [44-47].
    - canonical EF-hand (e.g., calmodulin):
      (1) calcium ions bind in a 12-residue central loop, utilizing side-chain oxygen ligands from loop positions 1, 3, 5, 9, and 12, as well as a main-chain carbonyl oxygen from position 7.
      (2) Ligands associated with EF-hands are typically Asp at position 1, Asp or Asn at position 3, Asp, Ser, or Asn at position 5, a water molecule at position 9, and a bidentate Glu at position 12 [30].
    - pseudo EF-hands: coordinate the Ca2+ ion predominantly with main-chain carbonyl oxygen atoms in a 14-residue loop.

EF-hand

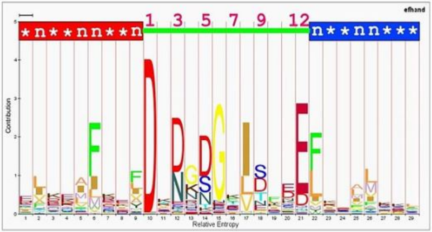






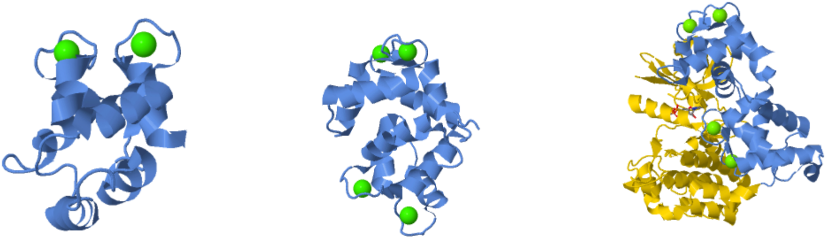
Residue: D-x-[DNS]-{ILVFYW}-[DENSTG]-[DNQGHRK]-{GP}-[LIVMC]-[DENQSTAGC]- x(2) -[DE]

Position: 1-2- 3 - 4 - 5 - 6 - 7 - 8 - 9 -10,11 - 12



where x indicates any residue; any residue in square brackets [] is possible at that position; none of the residues in curly brackets {} are possible: and $x(2)$ indicates a series of two x's.

EF-hand



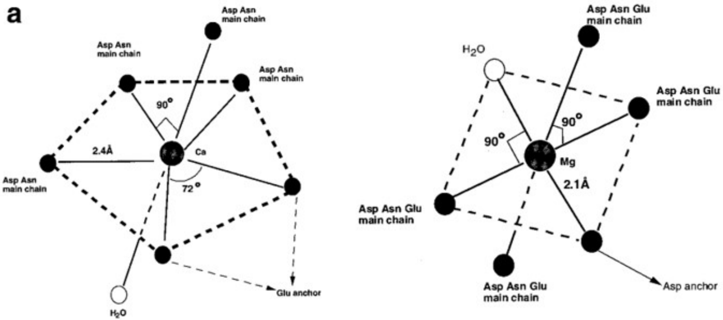
[1]

- Parvalbumin is a monomeric protein that has a pair of EF hands,
- calmodulin is a monomer with two pairs,
- Calcium-dependent protein kinase is a monomer with a protein kinase catalytic domain and its calcium-binding domain has two pairs of EF hands.

As shown in these examples EF hands often occur in interacting pairs, which enables the cooperative binding of calcium ions.

Ca2+: in green space fill. The calcium binding domain of CDPK is blue and the kinase catalytic domain is in gold and has an ATP analog (sticks in CPK colors) bound in its active site.

[1] https://proteopedia.org/wiki/index.php/EF_hand



b

Classic EF-hand

1	2	3	4	5	6	7	8	9	10	11	12
s		s		s		m-c		H_2O			b-s
D (99%)		D (76%) N (23%)		D (52%) S (23%) N (21%)		T (23%) F (16%) K (12%)					E (92%)

Pseudo EF-hand

1	2	3	4	5	6	7	8	9	10	11	12	13	14
m-c			m-c		m-c			m-c					b-s E

Outline

1 What to do?

2 How to do?

- Statistical analysis of structural characteristics of protein Ca^{2+}-binding sites
- Prediction of EF-Hand Calcium-Binding Proteins and Analysis of Bacterial EF-Hand Proteins
- Integration of Diverse Research Methods to Analyze and Engineer Ca2+-Binding Proteins: From Prediction to Production
- GG: A Graph Theory and Geometry Approach
- MUG (MUltiple Geometries)
- MUG^{C}

WILEY InterScience

DISCOVER SOMETHING GREAT

PROTEINS: Structure, Function, and Bioinformatics 65:643–655 (2006)

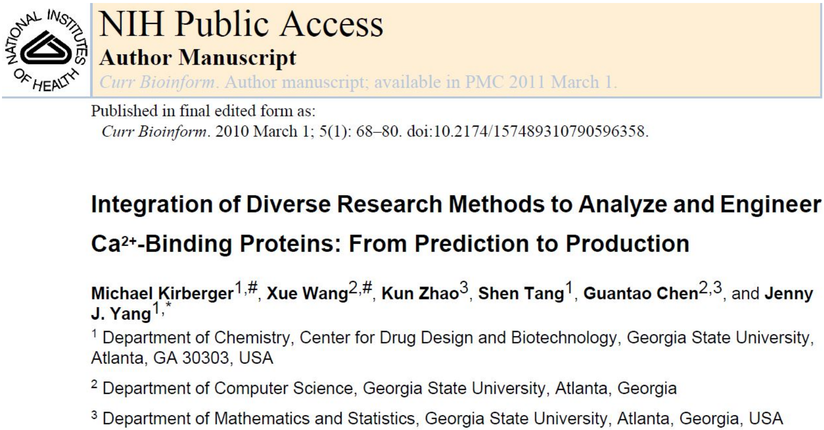
Prediction of EF-Hand Calcium-Binding Proteins and Analysis of Bacterial EF-Hand Proteins

Yubin Zhou, Wei Yang, Michael Kirberger, Hsiau-Wei Lee, Gayatri Ayalasomayajula, and Jenny J. Yang*
Department of Chemistry, Georgia State University, Atlanta, Georgia 30303

Outline

1. What to do?

2. How to do?
   - Statistical analysis of structural characteristics of protein Ca^{2+}-binding sites
   - Prediction of EF-Hand Calcium-Binding Proteins and Analysis of Bacterial EF-Hand Proteins
   - Integration of Diverse Research Methods to Analyze and Engineer Ca2+-Binding Proteins: From Prediction to Production
   - GG: A Graph Theory and Geometry Approach
   - MUG (MUltiple Geometries)
   - MUG^C



Published in final edited form as:

Curr Bioinform. 2010 March 1; 5(1): 68–80. doi:10.2174/157489310790596358.

Integration of Diverse Research Methods to Analyze and Engineer Ca^{2+}-Binding Proteins: From Prediction to Production

Michael Kirberger[1,#], **Xue Wang**[2,#], **Kun Zhao**[3], **Shen Tang**[1], **Guantao Chen**[2,3], and **Jenny J. Yang**[1,*]

[1] Department of Chemistry, Center for Drug Design and Biotechnology, Georgia State University, Atlanta, GA 30303, USA

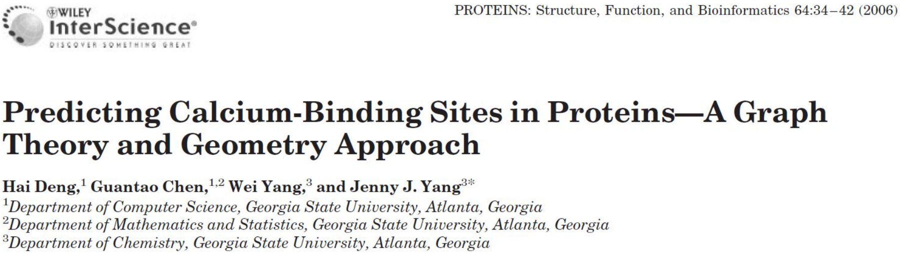
[2] Department of Computer Science, Georgia State University, Atlanta, Georgia

[3] Department of Mathematics and Statistics, Georgia State University, Atlanta, Georgia, USA

Outline

1. What to do?

2. How to do?
   - Statistical analysis of structural characteristics of protein Ca^{2+}-binding sites
   - Prediction of EF-Hand Calcium-Binding Proteins and Analysis of Bacterial EF-Hand Proteins
   - Integration of Diverse Research Methods to Analyze and Engineer Ca2+-Binding Proteins: From Prediction to Production
   - GG: A Graph Theory and Geometry Approach
   - MUG (MUltiple Geometries)
   - MUG^C



WILEY InterScience® DISCOVER SOMETHING GREAT

PROTEINS: Structure, Function, and Bioinformatics 64:34–42 (2006)

Predicting Calcium-Binding Sites in Proteins—A Graph Theory and Geometry Approach

Hai Deng,[1] Guantao Chen,[1,2] Wei Yang,[3] and Jenny J. Yang[3*]

[1] *Department of Computer Science, Georgia State University, Atlanta, Georgia*

[2] *Department of Mathematics and Statistics, Georgia State University, Atlanta, Georgia*

[3] *Department of Chemistry, Georgia State University, Atlanta, Georgia*

Datasets

Three datasets were used.

- Nayal-Di Cera's dataset (Dataset I) containing 32 proteins with 62 calcium-binding sites was used for parameter adjustment.
- Liang's dataset (Dataset II) contains 54 proteins with 91 calcium-binding sites and 14 noncalcium-binding proteins.
- Dataset III from Pidcock and Moore contains 94 sites in 44 proteins, representing all classes and folds of calcium-binding proteins.
- All structures are obtained from X-ray crystallography, and 114 of these 123 structures have resolution $<$ 2.4 A.
- Except for those from water, all of the oxygen atoms including those from proteins, carbohydrates, lipids, and other cofactors, are included in the calculation.

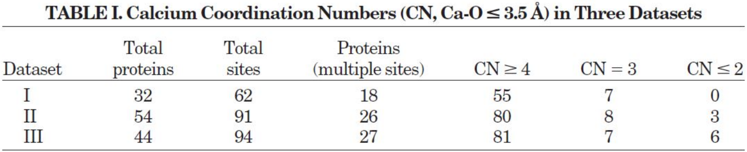


TABLE I. Calcium Coordination Numbers (CN, Ca-O $\leq$ 3.5 Å) in Three Datasets

Dataset	Total proteins	Total sites	Proteins (multiple sites)	CN $\geq$ 4	CN = 3	CN $\leq$ 2
I	32	62	18	55	7	0
II	54	91	26	80	8	3
III	44	94	27	81	7	6

Algorithm

Definitions

- Graph $G(V, E)$:
  each vertex in V represents an oxygen atom
  each edge in E represents a pair of oxygen atoms apart within an $O - O$ cutoff distance.
- Clique Q: represents oxygen clusters, is a subset of V such that every two vertices of Q are adjacent.
- the size of a clique: the number of vertices in the clique.

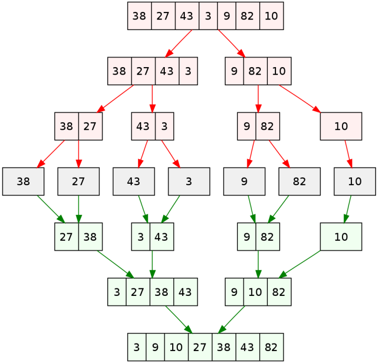
  Note that: Cliques may overlap representing multiple sites with shared ligand oxygens. They use a backtracking algorithm to find cliques of certain sizes (four in most of the studies reported here) in G (Scheme 1).
- the circumcenter (CC): the point which has the same distance to the four vertices of a clique (denoted as psdCa-O).

  A unique CC exists as long as the four oxygens are not in one plane. To eliminate false positives, a clique is considered as a putative calciumbinding site only if psdCa-O falls into the range $(R1, R2)$, where $R1$ and $R2$ are the lower and upper limits, respectively.

Algorithm

Removal of the Redundant Predictions

To remove redundant predictions in one location, a merging algorithm was adopted. All putative binding sites in a protein are input in a vector and sorted by psdCa-O. The one with the shortest psdCa-O is determined and the putative sites within 3.5 A (Center-Center distance) from it are deleted. The procedure is repeated until the vector is empty.



Algorithm

Performance Measurement

- A qualified clique is a true prediction (TP): if its CC falls into the cutoff distance (3.5 or 1.0 Å in this study) from a documented calcium ion in a crystal structure.
- A documented calcium-binding site is a true predicated site (TPS): if there is any prediction within the cutoff distance from this site.
- The performance of the method is evaluated by Site Sensitivity (SEN), Site Selectivity (SEL), and Redundancy (RE)
  Site Sensitivity (SEN): the percentage of TPS in the total sites
  Site Selectivity (SEL): the percentage of TP in the total predictions (hits)
  Redundancy (RE): the true predictions per site.
- The value of RE is not less than 1.

Algorithm

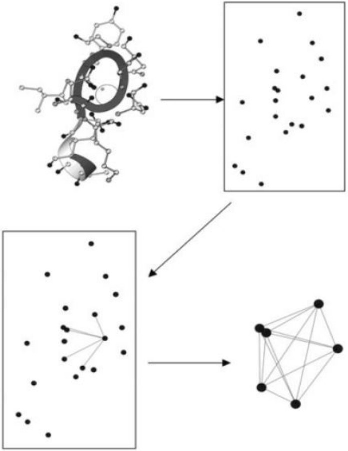


Fig. 1. The schematic model of GG program is shown. The positions of oxygen atoms (dark dots) are extracted from the protein structure while the other atoms (light dots) are excluded. The distance between two oxygen atoms are calculated and an edge is assigned if the distance is below a cutoff distance (O–O cutoff). A potential calcium-binding position is a clique (right bottom), which is a group in which every oxygen atom is linked to all other members of the group by edges.

```
Loop through all vertices a in V
  Loop through all vertices b in the adjacecy list of a
  output $V:=V-\{a\}$ if no such b exists
    Loop through all vertices c adjacent to both a and b
    output $V:=V-\{a\}$ if no such c exists
      Loop through all vertices d adjacent to both a, b, and c
        output clique {a,b,c,d}
      output $V:=V-\{a\}$ if no such d exists
```

Scheme 1. The pseudo code of the algorithm for finding all cliques with size four.

Results and Discussions

Analysis of Calcium-Binding Sites in Three Datasets

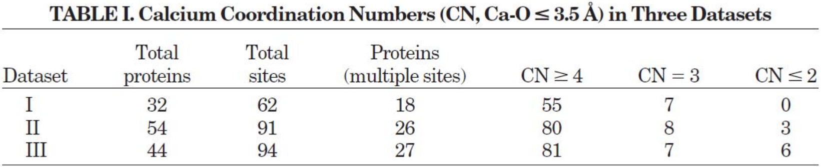


TABLE I. Calcium Coordination Numbers (CN, Ca-O $\leq$ 3.5 Å) in Three Datasets

Dataset	Total proteins	Total sites	Proteins (multiple sites)	CN $\geq$ 4	CN = 3	CN $\leq$ 2
I	32	62	18	55	7	0
II	54	91	26	80	8	3
III	44	94	27	81	7	6

- there are a total of 123 proteins and 231 calcium-binding sites in three datasets. There are three proteins (1OVA, 2POR, and 4SBV) in both Datasets I and II and four proteins (1CEL, 1ESL, 1KIT, and 1SRA) in both Datasets II and III.
- 55(89%) of 62 calcium-binding sites in 32 proteins in Dataset I have four or more oxygen ligands within 3.5 Å of calcium, including 5(16%) EF-hand proteins. This dataset was used for parameter adjustment.
- Dataset II contains 54 proteins with 91 sites, 80(88%) of which have four or more oxygen ligands including two (4%) EF-hand proteins.
- Dataset III from Pidcock and Moore contains 94 sites in 44 proteins representing all classes and folds of calcium-binding proteins, 81 (86%) of which have four or more oxygen ligands.

These three datasets represent different classes of calcium binding sites with different protein fold topologies (Fig. 2). For example,

- continuous sites in calmodulin (3CLN) with four EF-hand motifs are largely helical, and lectin (2TEP) is predominately a β-sheet.
- Semicontinuous sites such as in galactose-binding protein (1GCG) or penicillin acylase (1AI4) and discontinuous sites such as in annexin (1ALA) or cellulase (1CEL) are also included.
- About 55% of proteins in the datasets contain two or more calcium-binding sites.
- Some proteins, such as calmodulin, do not have calcium-binding sites with shared ligand residues, although the calcium binding process of this protein is tightly cooperative.
- In other proteins, the clustered calcium ions share ligand residues and even ligand oxygen atoms. For example, mannose-binding protein A (2MSB), thermolysin (1TMN, 1HYT, and 8TLN), neutral protease (1NPC), serum amyloid P component (1SAC), and carboxypeptidase T (1OBR) in the datasets have calcium ions that share the same oxygen atoms (Fig. 2).
- It is important that the datasets contain proteins representing different protein folds and families, and contain different types of calcium-binding sites. For example, the dataset III from Pidcock and Moore only contains one EF-hand protein (1SRA) to overcome the bias of EF-hand proteins in the protein data bank.

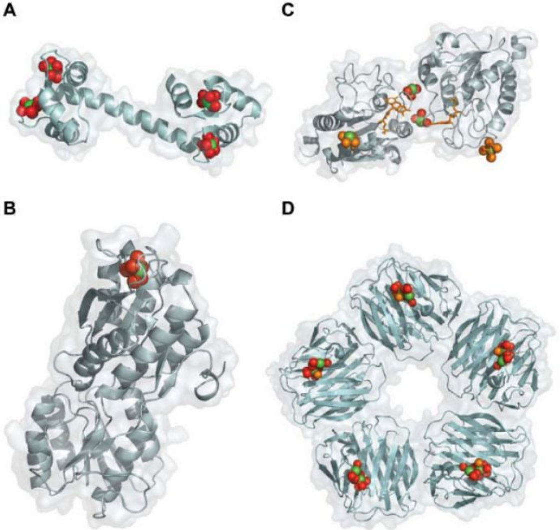


Fig. 2. The calmodulin (**A**, 3CLN), galactose-binding protein (**B**, 1GCG), flavodoxin (**C**, 1AG9), and serum amyloid P component (**D**, 1SAC) represent different classes of calcium-binding sites including different ligand distributions (continuous, semicontinuous, and discontinuous), binding numbers (single site, independent multiple sites, and sites with shared ligands), cofactor conditions, and protein sizes. The green balls are calcium ions. The ligand oxygen atoms are from proteins (red balls) and cofactors (orange balls). The protein frames are in light blue while the cofactors are in orange. The pictures are generated by PyMol (DeLano Scientific).

Parameter Optimization

The oxygen atoms from water are not included, because

(1) the inclusion of water molecules results in tremendous false predictions in the bulk solution, and

(2) the NMR and modeling structures typically do not contain water molecules.

However, the oxygen atoms from the cofactors, such as sugars or lipids, are included in this study. To achieve high accuracy and speed, clique size, O-O cutoff, PsdCa-O range, and D-filter in GG algorithm have been optimized using Dataset I.

Parameter Optimization

Clique Size: 4

- To allow a clique in graphs to represent the calcium binding location in a protein, the vertices and the edges should accurately describe the calcium-binding ligands and the relationships among them.
- To reveal key features of calcium-binding coordination and increase the speed of calculation, the model of calcium binding is simplified by using the minimal required cliques.
- As shown in Table I, more than 85% of the sites contain four or more close oxygen ligands.
- The use of clique size 3 resulted in many false positive cases and the increase of computational complexity (data not shown), although the use of clique size > 4 resulted in many false negative cases in GG. To focus on the main features of calcium-binding proteins, we therefore chose the clique size 4 in this study.
- It is expected that the calcium-binding sites with three or fewer ligands cannot be found. However, we will show that part of these sites can still be identified.
- To better illustrate GG, the datasets were analyzed either including or excluding the sites that have three or fewer ligand oxygen atoms.

Parameter Optimization

O-O Cutoff: 6.0 A

- The upper limit of O-O distance is restricted by the O-O cutoff
- the lower limit is determined by van der Waals radius
- Theoretically, the upper limit of the O-O distance is no more than twice of the maximum Ca-O distance.
- Most of the ligand oxygen atoms are in the distance of 2-3 Å to the calcium.

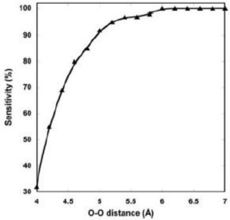


Fig. 3. The site sensitivity (the true predicted sites in the total sites) as a function of O–O distance cutoff.

- Figure 3 shows the SEN of the GG algorithm using a series of O-O cutoffs from 4 to 7 Å for Dataset I.
- At the cutoff of 6.0 Å, the SEN reaches 100% for Dataset I. Therefore, an O-O cutoff of 6.0 Å was used for all sites.

Parameter Optimization

O-O Cutoff: 6.0 A

- When the cutoff is 5.0 Å,
  - 89% (55 of 62) of the sites were identified.
  - Among the 55 sites with four or more ligands, 52 (94%) have been identified within 3.5 Å and 51 (93%) within 1 Å to the documented ions.
  - This result is consistent with the statistical study that shows the average Ca-O distance in calcium-binding sites is about 2.4 Å, suggesting that a cutoff of 5 Å covers most of the O-O distances.
  - The longer O-O distances mainly originate from longer O-Ca distances.
  - Further increase of the O-O cutoff results in the identification of more calcium-binding sites (within 3.5 Å to the real ions).

Parameter Optimization

PsdCa-O range=D-filter=1.8-3.0

- To eliminate false positives, a clique is considered as a putative calcium-binding site only if psdCa-O falls into the range (R1, R2).
- A D-filter is further applied to eliminate any cliques that contain nonoxygen atoms within a short distance (D-filter) from the CC because the space for calcium binding should not be occupied by other atoms.
- A clique is considered to be a potential calcium-binding site only when the psdCa-O is in a given range as long as there are no other atoms within a distance of D-filter to the circumcenter.
- SEN exhibits an upward trend, while SEL shows a downward trend with the increase of the psdCa-O range.
- The best performance was obtained when D-filter equals the psdCa-O of each clique.

PsdCa-O range=D-filter=1.8-3.0

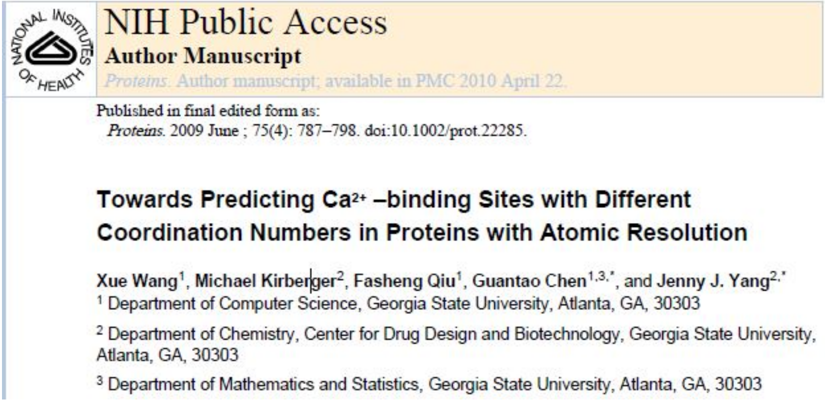
- Corresponding to the average Ca-O distance of 2.4 Å, 48 sites (77% of all sites and 87% of the sites with four or more ligands) in Dataset I are identified within 1 Å to the documented ions.
- If the psdCa-O of 2.3-2.5 Å is used, the corresponding SEL is 96%.
- When the psdCa-O range is enlarged to 2.1-2.7 Å, the SEN increases to 95% and the SEL is 87% for the sites with four or more ligands.
  In this case, three calciumbinding sites in ovalbumin (1OVA),47 staphylococcus nuclease (1SNC),48 and substilisin BPN (2ST1)49 are not identified.
- When the psdCa-O range is 1.8-3.0 Å, the only unidentified site is that in ovalbumin.
  Under such conditions, four of seven sites with three or fewer ligands in the Dataset I have also been identified within 3.5 Å of the documented ions, but none of them is within 1.0 Å.
- Taken together, using an O-O cutoff of 6.0 Å, a psdCa–O range of 1.8-3.0 Å and a D-filter equal to the psdCa-O, the optimal performance for the prediction of calcium-binding sites in Dataset I possesses a SEN of 95% and a SEL of 86% within 3.5 Å or a SEN of 87% and a SEL of 68% within 1.0 Å to the documented ions.

Outline

1. What to do?

2. How to do?
   - Statistical analysis of structural characteristics of protein Ca^{2+}-binding sites
   - Prediction of EF-Hand Calcium-Binding Proteins and Analysis of Bacterial EF-Hand Proteins
   - Integration of Diverse Research Methods to Analyze and Engineer Ca2+-Binding Proteins: From Prediction to Production
   - GG: A Graph Theory and Geometry Approach
   - MUG (MUltiple Geometries)
   - MUG^C

MUG



NIH Public Access

Author Manuscript

Proteins. Author manuscript; available in PMC 2010 April 22.

Published in final edited form as:
Proteins. 2009 June ; 75(4): 787–798. doi:10.1002/prot.22285.

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, GA, 30303

[2] Department of Chemistry, Center for Drug Design and Biotechnology, Georgia State University, Atlanta, GA, 30303

[3] Department of Mathematics and Statistics, Georgia State University, Atlanta, GA, 30303

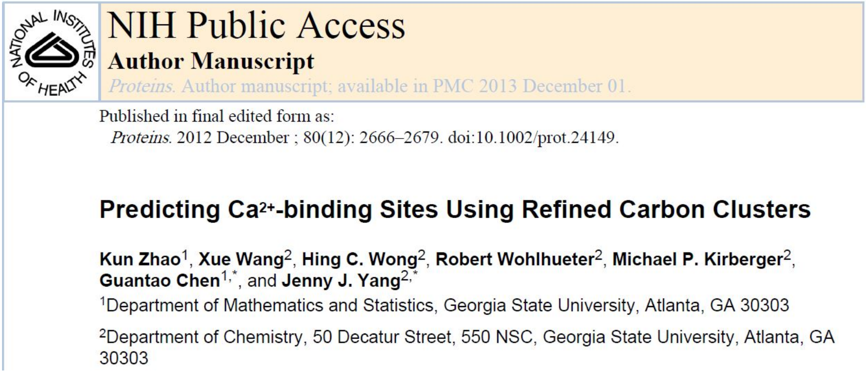
Outline

1 What to do?

2 How to do?

- Statistical analysis of structural characteristics of protein Ca^{2+}-binding sites
- Prediction of EF-Hand Calcium-Binding Proteins and Analysis of Bacterial EF-Hand Proteins
- Integration of Diverse Research Methods to Analyze and Engineer Ca2+-Binding Proteins: From Prediction to Production
- GG: A Graph Theory and Geometry Approach
- MUG (MUltiple Geometries)
- MUG^C

MUG^C



Published in final edited form as:
Proteins. 2012 December ; 80(12): 2666–2679. doi:10.1002/prot.24149.

Predicting Ca^{2+}-binding Sites Using Refined Carbon Clusters

Kun Zhao[1], **Xue Wang**[2], **Hing C. Wong**[2], **Robert Wohlhueter**[2], **Michael P. Kirberger**[2], **Guantao Chen**[1,*], and **Jenny J. Yang**[2,*]

[1]Department of Mathematics and Statistics, Georgia State University, Atlanta, GA 30303

[2]Department of Chemistry, 50 Decatur Street, 550 NSC, Georgia State University, Atlanta, GA 30303

For Further Reading I

A. Author.
Handbook of Everything.
Some Press, 1990.

S. Someone.
On this and that.
Journal of This and That, 2(1):50–100, 2000.