Are predicted structures good enough to preserve functional sites?

Liping Wei[1], Enoch S Huang[2] and Russ B Altman[1]*

Background: A principal goal of structure prediction is the elucidation of function. We have studied the ability of computed models to preserve the microenvironments of functional sites. In particular, 653 model structures of a calcium-binding protein (generated using an *ab initio* folding protocol) were analyzed, and the degree to which calcium-binding sites were recognizable was assessed.

Results: While some model structures preserve the calcium-binding microenvironments, many others, including some with low root mean square deviations (rmsds) from the crystal structure of the native protein, do not. There is a very weak correlation between the overall rmsd of a structure and the preservation of calcium-binding sites. Only when the quality of the model structure is high (rmsd less than 2 Å for atoms in the 7 Å local neighborhood around calcium) does the modeling of the binding sites become reliable.

Conclusions: Protein structure prediction methods need to be assessed in terms of their preservation of functional sites. High-resolution structures are necessary for identifying binding sites such as calcium-binding sites.

Addresses: [1]Stanford Medical Informatics, 251 Campus Drive, MSOB x215, Stanford University School of Medicine, Stanford, CA 94305-5479, USA and [2]Department of Biochemistry and Molecular Biophysics, Box 8231, Washington University School of Medicine, St Louis, MO 63110, USA.

*Corresponding author.
E-mail: altman@smi.stanford.edu

Key words: calcium-binding site, function analysis, sidechain modeling, site recognition, structure prediction

Received: 23 November 1998
Revisions requested: 27 January 1999
Revisions received: 2 March 1999
Accepted: 9 March 1999

Published: 28 May 1999

Structure June 1999, 7:643–650
http://biomednet.com/elecref/0969212600700643

 ISSN 0969-2126

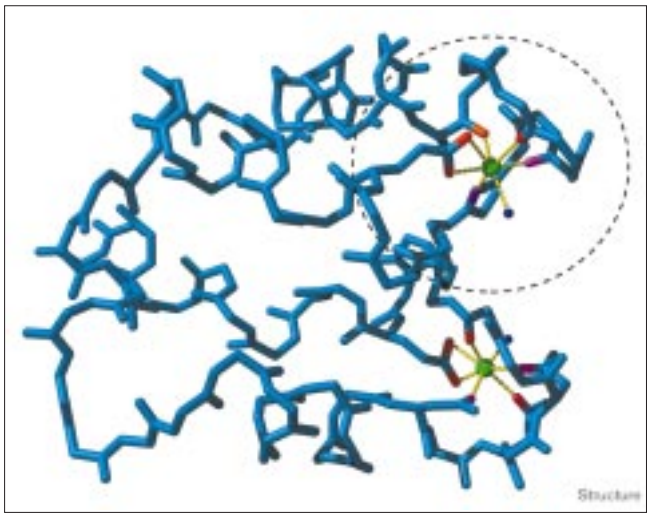
Introduction

In recent years, numerous computational techniques have emerged for predicting protein three-dimensional (3D) structures from their primary amino acid sequences. These techniques include comparative modeling, threading and *ab initio* folding [1,2], and are crucial to bridge the gap between the number of known protein sequences and number of known protein 3D structures [3,4]. The quality of the predicted structures has improved, as measured by blinded tests in the Critical Assessment of Methods of Protein Structure Prediction (CASP) series of meetings [2,5]. These tests have gauged success primarily with numerical metrics that judge the quality of alignments and root mean square deviation (rmsd) relative to the native structure [6–11]. Ultimately, however, the quality and usefulness of a predicted structure will also be gauged by its ability to provide information on molecular function. It is therefore important to assess the accuracy required for a predicted structure to be useful in functional analysis.

An important step towards the elucidation of function from a 3D structure is to determine the occurrence and location of active sites and binding sites within the structure. A 'well-predicted' structure should preserve the geometric, biochemical and physical features in the local microenvironments of these sites. Recent studies by Fetrow and Skolnick [12,13] have found that predicted structures with low to moderate resolution are sufficient to identify enzymatic active sites that have specific residue geometry. They have developed 'fuzzy functional forms' for active sites based mainly on the pairwise 3D distances between the Cα atoms of the conserved residues involved in the active sites. However, if a functional site lacks strongly conserved amino acid residues or conserved residue geometry, or if the site can be recognized accurately only by considering atomic coordinates (and not just overall residue locations), their studies and conclusions are not directly applicable. One such class of sites is the calcium-binding site. Recognizing the occurrence and location of calcium-binding sites is important because of the critical roles calcium has in intercellular and intracellular communication [14]. Unfortunately, because of the variation in the coordinating residues and the heterogeneity of coordination geometry, calcium-binding sites are difficult to recognize accurately in structural analysis at the residue level [14,15]. These sites are most reliably recognized at the atomic level, with consideration of not only geometric, but also biochemical and physical features [15–17]. The complexity of different functional and structural sites varies greatly, and the quality of a computed structure required for functional analysis would be expected to vary accordingly with the nature of the sites as well as with the level of detail of the functional analysis.

In this paper, we analyze the preservation of calcium-binding sites in model structures of a vitamin D dependent calcium-binding protein from bovine intestine. The native structure of the vitamin D dependent calcium-binding

Figure 1



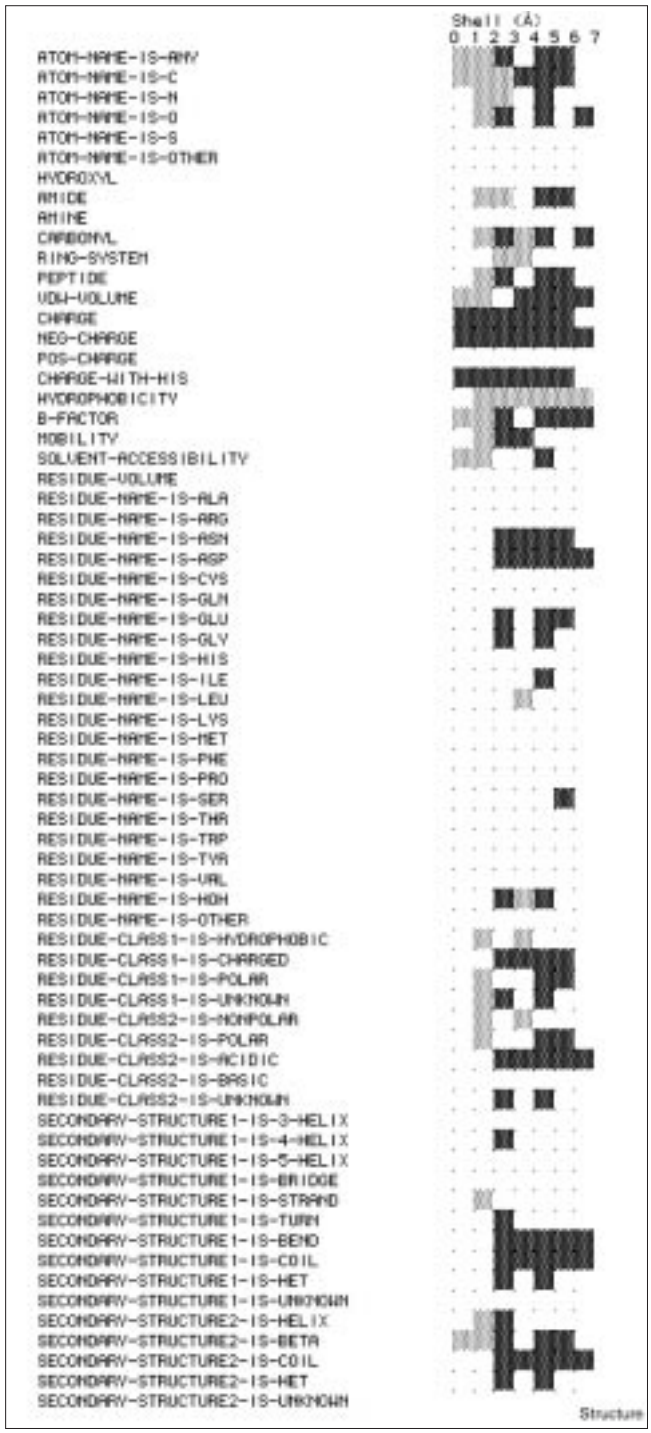
The native structure of vitamin D dependent calcium-binding protein from bovine intestine. The native protein binds two calcium ions, shown as green spheres, one near the N terminus (upper-right) and the other near the C terminus (lower-right). Both calcium-binding sites occur at loop regions near the surface of the protein. In the figure, the calcium ions are shown connected to their liganding oxygen atoms with yellow dashed lines. The black dashed-line circle around the N-terminal site illustrates the 7 Å spherical microenvironment for the calcium-binding site. (The figure was generated using MidasPlus [26,27].)

protein has been determined crystallographically at a resolution of 2.3 Å [18,19] (Figure 1; Protein Data Bank accession code 3ICB). The protein is small, 75 residues long, with two calcium-binding sites, one near the N terminus and the other near the C terminus. We have systematically studied the level of accuracy required to reliably recognize calcium-binding capability in models of this structure.

Park and coworkers [20,21] have constructed several sets of decoy protein structures to study the ability of different energy functions in selecting the correct folds. One of the sets they created comprised 653 decoy structures for the vitamin D dependent calcium-binding protein. These model structures are used in this study. The rmsd between the Cα atoms in the decoy structures and the corresponding Cα atoms in the native structure ranges from 0.95 Å to 9.39 Å. The structures with an rmsd less than or equal to 4 Å are considered 'near native' models because they maintain the general fold of the native structure [20,21]. There are 150 low rmsd near native structures among the 653 decoys.

In this study, three questions were addressed. Do any of the 653 decoy structures preserve the correct microenvironments for the two calcium-binding sites? For the 150 low rmsd structures, how does the level of preservation of the calcium-binding microenvironments relate to the quality of the structures? And when the calcium-binding sites are not

Figure 2



Conserved features in the microenvironments of calcium-binding sites compared to control nonsite regions that do not bind calcium. The column on the left shows the list of all biochemical and physical properties. The row on the top specifies the distance (in Å) of concentric radial shells from the calcium ion. A dark box means that that property in that shell is significantly more abundant for calcium-binding sites than for nonsites. A gray box indicates that the property in that shell is significantly more abundant for nonsites. A white box means that there is no statistically significant difference between sites and nonsites.

Table 1

Percentage of decoy structures recognizable as calcium-binding molecules.

Range of rmsd	Number of decoy structures	Number/percentage recognizable as calcium-binding molecules
≤ 4	150	123 (82%)
> 4	503	380 (75%)
All	653	503 (77%)

preserved in a low rmsd structure, what is wrong with the structure? To answer these questions, all of the decoy structures were searched for the occurrence and location of calcium-binding sites. We used a site-recognition method that we have previously developed that recognizes sites based on their microenvironments, represented by spatial distributions of multilevel chemical and physical properties [17,22]. The method compares a query region with the 3D motif of a calcium-binding site (Figure 2). This motif was constructed by statistical comparison of 59 known calcium-binding sites and 140 control nonsites (regions within protein structures that do not bind calcium ions). Our method reports the locations of potential calcium-binding sites — if any — as well as a score for each site indicating how likely the site is. The method has been shown to be highly accurate in recognizing calcium-binding sites. It has a sensitivity of 86% and a specificity near 100% in cross-validation analyses on known calcium-binding sites and control nonsites [17].

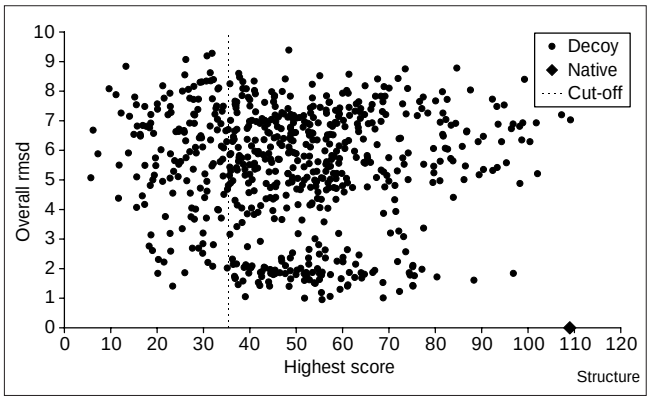
We have found that even though some structures successfully maintain the calcium-binding site microenvironment, many others, including some low rmsd structures, have incorrectly modeled one or both of the binding sites. We show two examples of the types of errors in the model structures. The quality of the calcium-binding microenvironments has only a very weak correlation with the overall rmsd. Only when the structures are of high quality (rmsd less than 2 Å for atoms in the 7 Å local neighborhood around calcium) are the microenvironments of calcium-binding sites consistently well modeled.

Results

Recognition of model structures as calcium-binding molecules

As a positive control, we first searched the native structure for calcium-binding sites using our site-recognition method. The N-terminal site is predicted at 0.6 Å from the true location of the calcium ion in the crystal structure, and the C-terminal site is predicted at 1.1 Å from the true location. Both sites are predicted with high scores — 103.5 and 109.0 for the N- and C-terminal sites, respectively. These scores are about six standard deviations higher than the mean of the scores for the 140 random regions used as control nonsites. There are no false positives.

Figure 3



Plot of the overall root mean square deviation (rmsd) of decoy structures from the crystal structure versus the highest calcium-binding site score in those structures. Each black circle denotes one decoy structure; the black diamond denotes the native structure. The dotted line shows the score cut-off, structures to the right of this line are recognized as calcium-binding molecules.

The 653 decoy structures were then searched for calcium-binding sites. First, we looked for the occurrence of any high-scoring regions, indicating a high-confidence prediction of calcium-binding sites. Table 1 shows the number of structures that have at least one high-scoring region. Figure 3 shows the plot of the overall rmsd against the highest score in a structure. The score cut-off is chosen to be four standard deviations higher than the mean of the scores for regions that do not bind calcium; the cross-validation sensitivity and specificity using this score cut-off are 86% and 100% for the training data including 59 known calcium-binding sites and 140 control nonsites. Overall, 77% of the decoy structures show a strong calcium-binding signal.

Relationship between rmsd values and preservation of sites

In addition to analyzing whether the overall calcium-binding function can be identified in these structures, it is also important to evaluate whether the local microenvironments of both native calcium-binding sites are conserved. Unfortunately, more than two-thirds of the 503 structures recognized as 'calcium-binding molecules' contain only one high-scoring region, even though the native structure has two calcium-binding sites. Table 2 shows how well the 150 low rmsd structures preserve the native calcium-binding sites. Only 25% of these structures identify both calcium sites. Figure 4 shows a plot of the overall backbone rmsd relative to the native protein against the scores for each of the two native calcium-binding sites. In both

Table 2

Percentage of low rmsd structures that preserve the two calcium-binding sites.

Total number of low rmsd structures	150
Number with both sites preserved	37 (25%)
Number with only N-terminal site preserved	17 (11%)
Number with only C-terminal site preserved	65 (43%)
Number with putative sites at wrong locations (i.e., more than 7 Å away from both sites)	4 (3%)

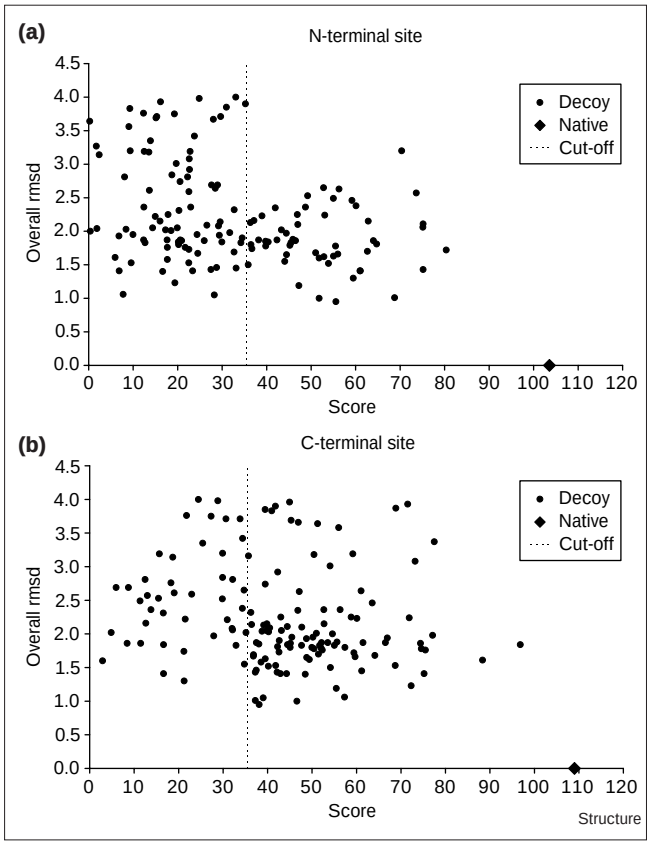
cases, the native calcium-binding site (rmsd = 0) has the highest score. The backbone rmsd correlates only very weakly with the score (and therefore with the strength of the calcium-binding site signal) for both sites, with a correlation coefficient of –0.31 and –0.16 for the N- and C-terminal sites, respectively.

Given that the overall backbone rmsd does not correlate well with the score of the site, we computed the rmsd of the model to the native structure only for atoms found within 7 Å of the calcium-binding sites in the native structure. Figure 5 shows the plot of this 'local rmsd' versus the score of the binding site. The correlation is significant (with a significance F of 9×10^{-6} and 1.8×10^{-5} for the two sites, respectively, in linear regression analysis) and stronger (with a correlation coefficient of –0.37 and –0.35, respectively) than the correlation between overall rmsd and score, but still weak.

Preservation of sites in high-resolution models

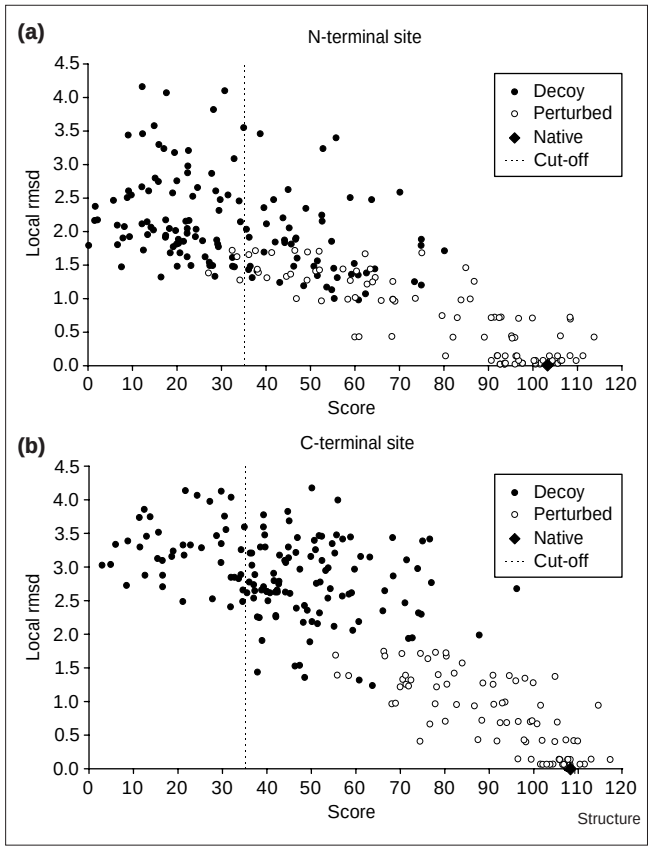
In order to investigate whether structures more accurate than the 653 decoy structures can more routinely preserve calcium-binding sites, we perturbed the position of atoms within the native structure of the protein to generate 100 decoys with local rmsd values in the range 0 to 1.7 Å. Figure 5 shows local rmsd values plotted against the score for the two calcium-binding sites in these structures. When the accuracy is high, the binding sites are reliably modeled and begin to have similar features to those in the

Figure 4



Plot of the overall root mean square deviation (rmsd) of the 150 low rmsd decoy structures from the crystal structure versus (a) the score for the N-terminal site and (b) the score for the C-terminal site. There is very weak correlation between the overall rmsd and the score for the binding sites.

Figure 5



Plot of the local all-atom root mean square deviation (rmsd) of 150 low rmsd decoy structures (shown as black circles) and the higher resolution perturbed structures (shown as white circles) versus (a) the score for the N-terminal site and (b) the score for the C-terminal site. Only when the resolution is high is there a strong correlation between the rmsd value and the score for the binding sites.

native structure. The rmsd measure of these structures also correlates much more strongly with the quality of the binding sites — the correlation coefficient between the local rmsd and the score is –0.85 and –0.80, respectively, for the two calcium-binding sites.

Examination of missed sites

The 150 low rmsd structures that miss one or both functional sites were examined in order to establish what features of the binding sites are missing. Two types of errors were found. First, some models have incorrect local backbone conformations and do not allow the sidechains to pack properly. Second, some models have the correct backbone conformation but incorrect sidechain orientations, so a recognizable binding site cannot be created. Figure 6 graphically illustrates examples of these two types of errors. A computer-generated analysis further explains which biochemical and physical features are missing in the model shown in Figure 6b that make it unrecognizable as a calcium-binding site (Table 3).

Discussion

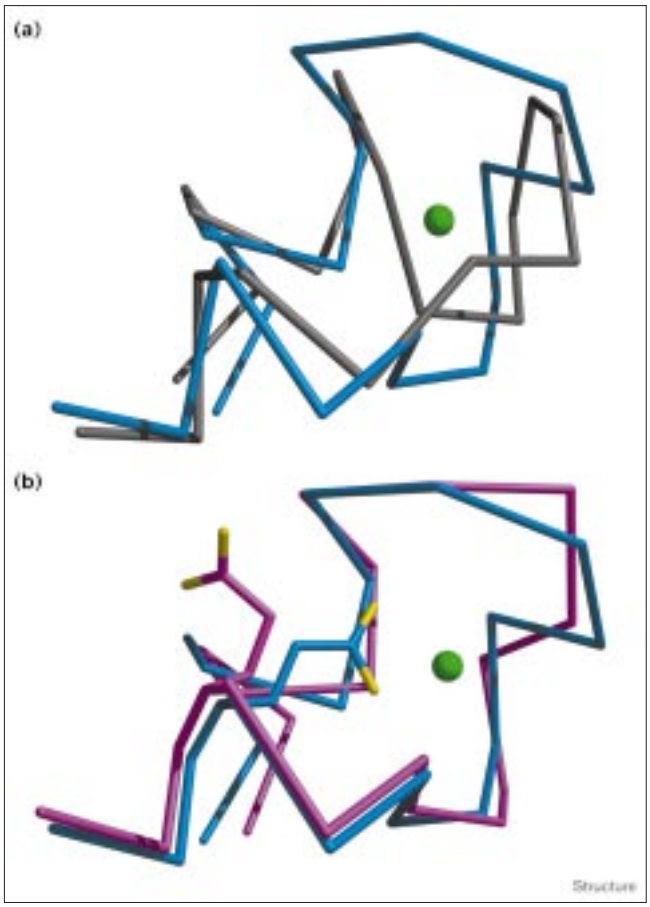
Accuracy of the site-recognition method

An important prerequisite for this study is that the site-recognition method be highly accurate. We have previously shown that the site-recognition method used here has high sensitivity and specificity [17]. The score cut-off employed is very strict, with 86% sensitivity and nearly 100% specificity in cross-validation analysis. The trade-off between sensitivity and specificity can be made by moving the score cut-off required for predicting sites. In this study, a strict score cut-off gives us confidence in structures recognized as calcium-binding structures.

Functional analysis at two levels

Functional analysis can be performed at different levels. At a coarse level, we test whether the overall structure can be recognized as a calcium-binding structure. The analysis shows that the general calcium-binding capabilities of this protein can be recognized fairly frequently even for high rmsd structures. However, this is likely to be a consequence of the high content of aspartate and glutamate residues in this 75-residue protein — four aspartate and 13 glutamate residues in total, representing nearly one-fourth of the whole protein. As shown in Figure 2, the abundance of aspartate and glutamate residues is important for the formation of calcium-binding sites; these residues provide the liganding oxygen atoms as well as the correct hydrophobicity contrast in the microenvironment. As the sequence has a high content of acidic sidechains, it is possible for wrong conformations to bring sidechains together to create credible calcium-binding sites by chance. More than 70% of the structures with rmsd values higher than 4 Å can still be recognized as calcium-binding structures. Such sequence bias may be less important in larger structures where the chances of having amino acids randomly associate to produce candidate sites will be lower.

Figure 6



Examples of two types of errors that result in the loss of the N-terminal calcium-binding site in two decoy structures. (a) Bad backbone conformation. (b) Bad sidechain orientation. In (b), Glu27, a crucial residue in the binding of the calcium ion, is incorrectly modeled in the decoy structure. Two crucial oxygen atoms in Glu27 normally form part of the coordination octahedra in the native structure, but point in the wrong direction in the decoy. In both figures only part of the structure – the local region around the N-terminal site – is shown. The decoy structures are shown in grey (a) and magenta (b); the backbone of the native structure is shown in light blue for comparison. The position of the calcium ion in the native structure is shown with a green sphere. (The figures were generated using MidasPlus [26,27].)

At the finer level, we test whether the biochemical and physical microenvironment of each of the binding sites is correctly modeled in the structures. In order to understand how the microenvironment of each of these two sites is maintained or disrupted (literally beyond recognition), we focused on the near-native decoy structures (rmsd less than 4 Å), which allows us to define the locations in the 3D structure where the calcium sites should have been recognized. Table 2 shows that only 25% of the high-resolution structures preserve both binding sites. An additional 54%

Table 3

Computer-generated analysis to demonstrate that a decoy structure does not have the correct calcium-binding microenvironment.

ATOM-NAME-IS-ANY in shell 2 is low
ATOM-NAME-IS-O in shell 2 is low
CARBONYL in shell 2 is low
PEPTIDE in shell 2 is low
CHARGE in shell 1, 2, 3 is low
NEG-CHARGE in shell 1, 2, 3 is low
CHARGE-WITH-HIS in shell 1, 2, 3 is low
HYDROPHOBICITY in shell 1, 2 is high
MOBILITY in shell 2 is low
RESIDUE-NAME-IS-ASP in shell 2, 3, 5 is low
RESIDUE-CLASS1-IS-CHARGED in shell 2 is low
RESIDUE-CLASS2-IS-ACIDIC in shell 2 is low
SECONDARY-STRUCTURE1-IS-BEND in shell 5 is low

The explanation function of the site-recognition method was used to analyze the region in the decoy structure shown in Figure 6b that corresponds to the N-terminal calcium-binding site in the native structure. Listed in the table are features – properties and shells in which the properties occur – that make the region in the decoy NOT likely to be a calcium-binding site.

preserve one binding site, but not both. Thus, even for structures that would be considered excellent fold predictions, the microenvironments of the functional sites are not modeled consistently.

The relationship between rmsd values and preservation of functional sites

The overall backbone rmsd is clearly not a good indicator of the quality of the microenvironments of functional sites (Figure 4). This is not surprising because functional sites tend to be local entities rather than global ones. However, our finding that models with an overall backbone rmsd as low as 1.5 Å do not necessarily preserve the microenvironment of the binding sites is quite sobering. The accuracy of the structures predicted by current prediction techniques ranges widely, but models with an accuracy of 1.5 Å are usually taken as excellent results. As shown by our study, however, such models can still fail to give strong functional site signals if the local backbone or sidechains are not accurately modeled. Good calcium-binding microenvironments require more than close positioning of the correct residues. The microenvironments require the right liganding atoms (i.e., the oxygen atoms) at the right distance (between 1.4 Å and 3.5 Å), and the right hydrophobicity contrast around the center of the site (a shell of hydrophilic atomic groups embedded within a larger shell of hydrophobic atomic groups) [15,16]. In addition, in the native structure of the vitamin D dependent calcium-binding protein, both of the calcium-binding sites occur at loop regions on the surface of the protein, which are known to have higher error rates in protein modeling. Thus, functional sites such as calcium-binding sites can be difficult to model correctly. Better modeling of loops and better placement of sidechains in predicted models will both improve the likelihood of recognition of such functional sites.

A better measure of the quality of local regions in predicted structures is the local all-atom rmsd value. Figure 5 shows that there is a stronger correlation between the local all-atom rmsd and the quality of the microenvironments of the two binding sites. These plots also show that there is a significant gap between the best model structure and the native structure. If we can increase the accuracy of the predicted structure further, the quality of the modeled calcium-binding sites approaches that in the native structure, as also shown in Figure 5 with the model structures generated by perturbation. Some of the model structures generated by perturbation with very low rmsd values even have calcium-binding sites that have a slightly higher score than those in the native structure.

Implications for protein structure prediction

Modeling functional sites correctly is critical for detailed analysis of protein functions and for the design of novel proteins. Our case study shows that the preservation of functional sites requires a high accuracy of structure prediction and details in the modeled structures matter. Thus, further progress in the modeling of loop regions and sidechains, with the goal of increased resolution, is crucial. The errors shown in Figure 6 provide examples of ostensibly good predictions that would need to be improved to allow for site recognition.

If the function of a protein is known, then the occurrence of an expected site may be used as an indication of a good structure prediction. Our results show that using site recognition to separate correct topologies from incorrect ones is dangerous. For some functional sites, such as calcium-binding sites, the quality of the functional sites modeled may be related to the overall quality of the fold only very weakly, because of the localized and detailed characteristics of the sites. Our site-recognition algorithm uses the detailed positions of atoms to recognize sites, and is appropriately sensitive to changes in these positions. Using more coarse representations of sites, such as using only the presence of appropriate residue types within a range of distances, may reduce the sensitivity of site recognition to change and be more appropriate as a filter for predicted topologies [13]. Unfortunately, such a strategy is likely to also increase the false-positive rate (i.e., decrease specificity) of site recognition.

Our results suggest, however, that filtering with functional site-recognition methods may be helpful in the selection

of the best structures among structures constructed by homology modeling, which are sufficiently close to the native structure. When the local rmsd becomes low, the functional sites begin to be routinely recognizable and can be used as filters on homology models. Our results underscore the observation that rmsd related measures are not sufficient to gauge the quality of models. It is important to include the preservation of functional sites as an additional evaluation measure of the quality of the structures, and such measures may be a valuable addition to the CASP experiments.

Other binding sites

The site-recognition method used in this study is general purpose. We have used the method elsewhere to characterize ATP-binding sites and serine protease active sites [22,23]. These sites have a much more complex set of requirements for binding, and so it will be important to evaluate the sensitivity of these sites to computed models. The calcium-binding model is dominated by the need for relatively symmetric coordination chemistry around the calcium ion. For small ligand binding sites, the binding site microenvironment may be even more sensitive to changes in the positions of backbone and sidechain elements.

Biological implications

Modeling functional sites correctly is important for novel protein design and structure-based drug discovery. Recently, there has been interest in assessing the ability of model structures to be used for functional prediction. Here, we have studied 653 model structures and assessed their ability to reliably preserve recognizable calcium-binding sites. We find that predicted structures may be good enough for high-level functional analysis. However, for detailed analysis such as the precise localization of functional sites, which is important in protein design, high-resolution structures at the accuracy of 1–2 Å are required. Unfortunately, this level of accuracy is, at present, beyond the capabilities of typical *ab initio* methods [11]. Because many binding sites occur at loop regions and may involve atomic-level interactions that cannot be recognized with a residue-level site model, the correct modeling of loop regions and sidechain atoms becomes critical.

Materials and methods

Construction of decoy structures

The 653 models of vitamin D dependent calcium-binding protein were generated by the method of Park and Levitt [20], which exhaustively constructs backbone traces using four discrete points in Ramachandran space. The native helical boundaries were enforced during the fold generation procedure. Sidechain atoms were built using the software SegMod [24]. The all-atom models were subjected to 2000 steps steepest descent minimization against the ENCAD forcefield [25]. These particular structures were selected in order to represent the full range of structures that might result from an *ab initio* folding experiment. This decoy set may be downloaded at http://dd.stanford.edu/.

Structures that were higher in accuracy than the decoy set were also generated by perturbation of the native structure. To generate a perturbed structure, each atomic coordinate in the native structure is perturbed by a small random number. We used a series of perturbations ranging from 0.01 to 1.2 Å and generated ten structures for each level of perturbation. The resulted structures had a local all-atom rmsd from the native structure of 0.01 to 1.7 Å.

The site-recognition method

To search for potential calcium-binding sites within a 3D structure, we define and search a grid that covers the whole structure, with 1.652 Å spacing between the closest neighboring grid points. At each search grid point, we determine whether the local region around the search point is likely to be a calcium-binding site. A site-recognition method is used, which is briefly described below, that recognizes sites by their microenvironments [17,22]. In 3D space, microenvironments are spherical regions with 7 Å radius around the center point of interest, divided into concentric shells of 1 Å thickness. We have defined a list of geometric, biochemical and physical properties that span multiple levels of details, from atomic to residue to secondary structural. For each property, we calculate its value for all atoms and sum the property values up in each shell. Thus the microenvironment for a site is described by a series of triples (property-volume-value) in which the value of a property within the radial shell volume is maintained. The microenvironment around a search point in a test structure is compared to the microenvironments in the training set consisting of known calcium-binding sites and nonsites. A Bayesian scoring function determines a score that indicates the likelihood that the region around the search point is a calcium-binding site. The version of the site-recognition method used in this paper is an improved, slightly different version compared to the one described previously [17]. We now have a larger training set, 59 known calcium-binding sites and 140 control nonsites, chosen from proteins with pairwise sequence identity of less than 25%. For higher specificity, we used a higher score cut-off of 35.431, which is four standard deviations higher than the mean of the scores for the 140 known nonsites. If the score for a search point exceeds the cut-off, the search point is labeled a potential calcium-binding location. Cross-validation sensitivity and specificity using this cut-off are 86% and 100%, respectively, for the training set. After all search points in a structure are scored, the potential calcium-binding locations are clustered. First, the highest scoring point is found and labeled a calcium-binding site; then all potential calcium-binding locations that are within 7 Å distance of the highest scoring point are removed, and the clustering procedure is repeated for the remaining potential locations.

We scanned each of the 653 decoy structures and 100 perturbed structures, applied the score cut-off to all search points within each structure, clustered the resulting high-scoring points, and reported the clusters as putative sites. For the low rmsd decoy structures, if one and only one putative site lies within 7 Å of the location of a calcium ion in the native structure, that calcium-binding site is said to be ‘preserved’ or ‘identified’ in the decoy.

More information about the site-recognition method can be found at http://www.smi.stanford.edu/projects/helix/features.html. The time required to scan a structure on a medium-range workstation ranges from a few seconds to a few minutes, depending on the size of the structure. The development of an online server for automated annotation of functional and structural sites is under consideration.

Acknowledgements

This work is supported by the National Institutes of Health (NIH) grants LM-05652 and LM-06422, National Science Foundation (NSF) grant DBI-9600637, NSF agreement ACI-9619020, a Faculty Scholar grant from IBM and a gift from Sun Microsystems Inc. ESH is a Jane Coffin Childs Fellow. The authors wish to thank Michael Levitt and his laboratory for the use of the SegMod software and technical assistance. We also thank Teri Klein for her help with generating the color figures using the MidasPlus software.

References

1. Eisenhaber, F., Persson, B. & Argos, P. (1995). Protein structure prediction: recognition of primary, secondary, and tertiary structural features from amino acid sequence. *Crit. Rev. Biochem. Mol. Biol.* 30, 1-94.
2. Moult, J., Hubbard, T., Bryant, S.H., Fidelis, K. & Pedersen, J.T. (1997). Critical assessment of methods of protein structure prediction (CASP): round II. *Proteins* (Supplement 1), 2-6.
3. Bairoch, A. & Apweiler, R. (1997). The SWISS-PROT protein sequence data bank and its supplement TrEMBL. *Nucleic Acids Res.* 25, 31-36.
4. Bernstein, F.C., *et al.*, & Tasumi, M. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* 112, 535-542.
5. Moult, J., Pedersen, J.T., Judson, R. & Fidelis, K. (1995). A large-scale experiment to assess protein structure prediction methods. *Proteins* 23, ii-v.
6. Venclovas, C., Zemla, A., Fidelis, K. & Moult, J. (1997). Criteria for evaluating protein structures derived from comparative modeling. *Proteins* (Supplement 1), 7-13.
7. Martin, A.C., MacArthur, M.W. & Thornton, J.M. (1997). Assessment of comparative modeling in CASP2. *Proteins* (Supplement 1), 14-28.
8. Marchler-Bauer, A. & Bryant, S.H. (1997). Measures of threading specificity and accuracy. *Proteins* (Supplement 1), 74-82.
9. Marchler-Bauer, A., Levitt, M. & Bryant, S.H. (1997). A retrospective analysis of CASP2 threading predictions. *Proteins* (Supplement 1), 83-91.
10. Zemla, A., Venclovas, C., Reinhardt, A., Fidelis, K. & Hubbard, T.J. (1997). Numerical criteria for the evaluation of *ab initio* predictions of protein structure. *Proteins* (Supplement 1), 140-150.
11. Lesk, A.M. (1997). CASP2: report on *ab initio* predictions. *Proteins* (Supplement 1), 151-166.
12. Fetrow, J.S., Godzik, A. & Skolnick, J. (1998). Functional analysis of the *Escherichia coli* genome using the sequence-to-structure-to-function paradigm: identification of proteins exhibiting the glutaredoxin/thioredoxin disulfide oxidoreductase activity. *J. Mol. Biol.* 282, 703-711.
13. Fetrow, J.S. & Skolnick, J. (1998). Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. *J. Mol. Biol.* 281, 949-968.
14. McPhalen, C.A., Strynadka, N.C. & James, M.N. (1991). Calcium-binding sites in proteins: a structural perspective. *Adv. Protein Chem.* 42, 77-144.
15. Nayal, M. & Di Cera, E. (1994). Predicting Ca^{2+} binding sites in proteins. *Proc. Natl Acad. Sci. USA* 91, 817-821.
16. Yamashita, M.M., Wesson, L., Eisenman, G. & Eisenberg, D. (1990). Where metal ions bind in proteins. *Proc. Natl Acad. Sci. USA* 87, 5648-5652.
17. Wei, L. & Altman, R.B. (1998). In *Pacific Symposium on Biocomputing*. (Altman, R.B., Dunker, A.K., Hunter, L. & Klein, T.E., eds), pp. 497-508, World Scientific, Maui, Hawaii.
18. Szebenyi, D.M. & Moffat, K. (1986). The refined structure of vitamin D-dependent calcium-binding protein from bovine intestine. Molecular details, ion binding, and implications for the structure of other calcium-binding proteins. *J. Biol. Chem.* 261, 8761-8777.
19. Szebenyi, D.M., Obendorf, S.K. & Moffat, K. (1981). Structure of vitamin D-dependent calcium-binding protein from bovine intestine. *Nature* 294, 327-332.
20. Park, B. & Levitt, M. (1996). Energy functions that discriminate X-ray and near native folds from well-constructed decoys. *J. Mol. Biol.* 258, 367-392.
21. Park, B.H., Huang, E.S. & Levitt, M. (1997). Factors affecting the ability of energy functions to discriminate correct from incorrect folds. *J. Mol. Biol.* 266, 831-846.
22. Wei, L., Chang, J.T. & Altman, R.B. (1998). Statistical analysis of protein structures: using environmental features for multiple purposes. In *Computational Methods in Molecular Biology*. (Salzberg, S., Searls, D. & Kasif, S., eds), pp. 207-226, Elsevier Science, France.
23. Bagley, S.C. & Altman, R.B. (1996). Conserved features in the active site of nonhomologous serine proteases. *Fold. Des.* 1, 371-379.
24. Levitt, M. (1992). Accurate modeling of protein conformation by automatic segment matching. *J. Mol. Biol.* 226, 507-533.
25. Levitt, M., Hirshberg, M., Sharon, R. & Daggett, V. (1995). Potential energy function and parameters for simulations of the molecular dynamics of proteins and nucleic acids in solution. *Comp. Phys. Commun.* 91, 215-231.
26. Ferrin, T.E., Huang, C.C., Jarvis, L.E. & Langridge, R. (1988). The MIDAS display system. *J. Mol. Graph.* 6, 13-27.
27. Huang, C.C., Pettersen, E.F., Klein, T.E., Ferrin, T.E. & Langridge, R. (1991). Conic: a fast renderer for space-filling molecules with shadows. *J. Mol. Graph.* 9, 230-236.

Because *Structure with Folding & Design* operates a 'Continuous Publication System' for Research Papers, this paper has been published on the internet before being printed (accessed from http://biomednet.com/cbiology/str). For further information, see the explanation on the contents page.