

Prediction of structures of zinc-binding proteins through explicit modeling of metal coordination geometry

Chu Wang,^{1,2} Robert Vernon,¹ Oliver Lange,¹ Michael Tyka,¹
and David Baker^{1,2*}

¹Department of Biochemistry, University of Washington, Seattle, Washington 98195

²Howard Hughes Medical Institute, University of Washington, Seattle, Washington 98195

Received 11 October 2009; Revised 9 December 2009; Accepted 11 December 2009

DOI: 10.1002/pro.327

Published online 6 January 2010 proteinscience.org

Abstract: Metal ions play an essential role in stabilizing protein structures and contributing to protein function. Ions such as zinc have well-defined coordination geometries, but it has not been easy to take advantage of this knowledge in protein structure prediction efforts. Here, we present a computational method to predict structures of zinc-binding proteins given knowledge of the positions of zinc-coordinating residues in the amino acid sequence. The method takes advantage of the “atom-tree” representation of molecular systems and modular architecture of the Rosetta3 software suite to incorporate explicit metal ion coordination geometry into previously developed *de novo* prediction and loop modeling protocols. Zinc cofactors are tethered to their interacting residues based on coordination geometries observed in natural zinc-binding proteins. The incorporation of explicit zinc atoms and their coordination geometry in both *de novo* structure prediction and loop modeling significantly improves sampling near the native conformation. The method can be readily extended to predict protein structures bound to other metal and/or small chemical cofactors with well-defined coordination or ligation geometry.

Keywords: *de novo* structure prediction; loop modeling; metalloproteins; zinc binding

Introduction

Zinc is one of the most abundant and important metal ions in biology, playing an indispensable role in a broad range of cellular processes, such as DNA replication and transcription,¹ cell apoptosis,² and metabolism.³ Catalytically, zinc acts as the critical electrophile in many hydrolases⁴ and structurally, zinc stabilizes many protein domains, for example,

“zinc-finger” proteins.⁵ Genome analysis studies have revealed thousands of potential zinc-binding protein sequences⁶; however, only a small percentage of them have been structurally characterized.⁷ Therefore, it is of substantial interest to develop computational structure prediction methods that are able to generate three-dimensional structural models of zinc-binding proteins from their sequences with accuracy in terms of both overall topology and atomic details around zinc-binding site.

Many previous studies have reviewed and classified the coordination geometry and amino acid preferences in zinc-binding sites in known zinc-binding proteins.^{8–10} Patel *et al.* estimated that a majority (82%) of zinc ions in proteins are tetrahedrally coordinated with the rest pentahedrally or hexahedrally coordinated.¹⁰ For a structural zinc-binding site, cysteine (Cys) and histidine (His) are the preferred coordinating residues and usually there are no water

Abbreviations: NMR, nuclear magnetic resonance; PDB, Protein Data Bank; RMSD, root mean square deviation.

Additional Supporting Information may be found in the online version of this article.

Chu Wang's current address is The Scripps Research Institute, La Jolla, California 92037.

Grant sponsor: Henry Wellcome Fellowship Grant.

*Correspondence to: David Baker, Department of Biochemistry, Box 357350, University of Washington, Seattle, WA 98195. E-mail: dabaker@u.washington.edu

molecules in the primary coordination sphere.⁹ In a recent comprehensive survey of zinc-binding proteins, Grishin and coworkers structurally classified zinc finger domains into eight distinct fold groups with three dominant categories: C2H2-like finger, treble clef finger, and zinc ribbon.¹¹ Torrance *et al.* studied the evolutionary divergence of metal-binding sites in proteins and identified two “archetypal” zinc-binding site structures — Cys-Cys-Cys-Cys and His-Cys-Cys-Cys, each of which appears to have evolved independently multiple times.¹² These sequence and structural patterns have led to development of methodologies to predict potential zinc-binding sites from either protein primary sequences or structural information. Hovmoller and coworkers combined support vector machine and homology-based approaches to predict zinc-coordinating Cys and His from protein sequences with a success rate of 86%.¹³ METSITE developed by Sodhi *et al.* uses neural network classifiers to distinguish metal-binding sites from nonsites in protein structure models with moderate quality with a mean accuracy of 94.5%.¹⁴ The empirical force field Fold-X developed by Serrano and coworkers is able to predict from the high-resolution crystal structures the positions of single-atom ligand including zinc with an overall deviation less than 1.0 Å.¹⁵ Recently, it was shown by the Montelione group that zinc-coordinating cysteine residues can also be identified based on NMR ¹³C_α and ¹³C_β chemical shift data.¹⁶ Despite this progress, information on likely zinc-binding sites has not generally been incorporated into protein tertiary structure prediction methods to generate models for zinc-binding proteins. Previous studies have included zinc in docking metalloprotein–ligand complexes¹⁷ and modeling active site of metalloenzymes by molecular dynamics (MD) simulations^{18,19}; however, modeling in these cases starts from existing protein structures, and only a narrow range of protein conformational space is searched.

The two key components of computational protein structure prediction methods are the procedure for carrying out the conformational search (sampling) and the free energy function used for evaluating possible conformations (scoring).²⁰ Challenges in both areas have hindered modeling metal binding explicitly in protein structure prediction. First, conformational sampling is generally limited to the protein backbone and sidechain torsional degrees of freedom, and it is difficult to simultaneously sample the rigid-body degrees of freedom of the metal ion during folding. Second, to reduce computational complexity, nonbonded physical interactions among multiple atoms are simplified by treating the total energy as the sum of pairwise additive distance-dependent interactions. However, this two-body approximation does not suffice to model metal–protein interactions because metal coordination geometries around the favored coordination sphere have

angular and multibody dependencies. In the example of zinc, the tetrahedral coordination of the four liganding residues requires distances, angles, and dihedrals among multiple atoms to be satisfied simultaneously. New algorithms must be developed to address such challenges to model metal-binding sites explicitly in protein structure prediction.

The *de novo* structure prediction and homology modeling methods in Rosetta software suite use a Monte Carlo strategy to assemble short fragments of known protein structures into compact conformations followed by gradient-based refinement with respect to all backbone and sidechain torsional angles in a detailed all-atom force field.^{21,22} The power of the methods has been demonstrated by the generation of structural models with atomic accuracy for a handful of benchmark and blind prediction protein targets in the last few years.^{23–25} Recently, a “fold-tree” representation²⁶ of the molecular system has been developed in Rosetta that can seamlessly integrate the torsional degrees of freedom and rigid-body degrees of freedom, which has allowed explicit treatment of backbone flexibility in protein–protein docking²⁷ and protein–ligand docking.²⁸ Taking advantage of this new capability, we developed an approach for predicting the structure of proteins with ion-binding sites with known coordination geometries. In this new method, zinc ions are explicitly represented and are tethered to their liganding residues with naturally observed geometries to maintain the integrity of the zinc-binding site and drive the folding of the protein chain. We show that in both *de novo* structure predictions and loop modeling, the explicit incorporation of zinc ions significantly improves sampling toward native protein conformation, and we expect that this method can be readily extended to predict protein structures bound with other metal ions and other ligands/cofactors with known coordination geometries.

Results

Incorporation of zinc into the molecular system for protein structure prediction

It has been well established that the majority of the structural zinc-binding sites are arranged in a tetrahedral coordination, and the most preferred zinc-liganding residues in these sites are cysteines and histidines.^{9,10} To capture this coordination geometry, zinc is represented as a ligand with five atoms forming the center and vertices of a tetrahedron. The actual zinc atom is centered in the tetrahedron and each of the four virtual atoms occupies a vertex. The distance between zinc and a virtual atom, 2.20 Å [Fig. 1(A)], is given by the average bond lengths of Zn-S_γ of Cys and Zn-N_δ/N_ε of His in a set of structural zinc-binding sites. The four virtual atoms defined in the zinc residue serve to (1) set a

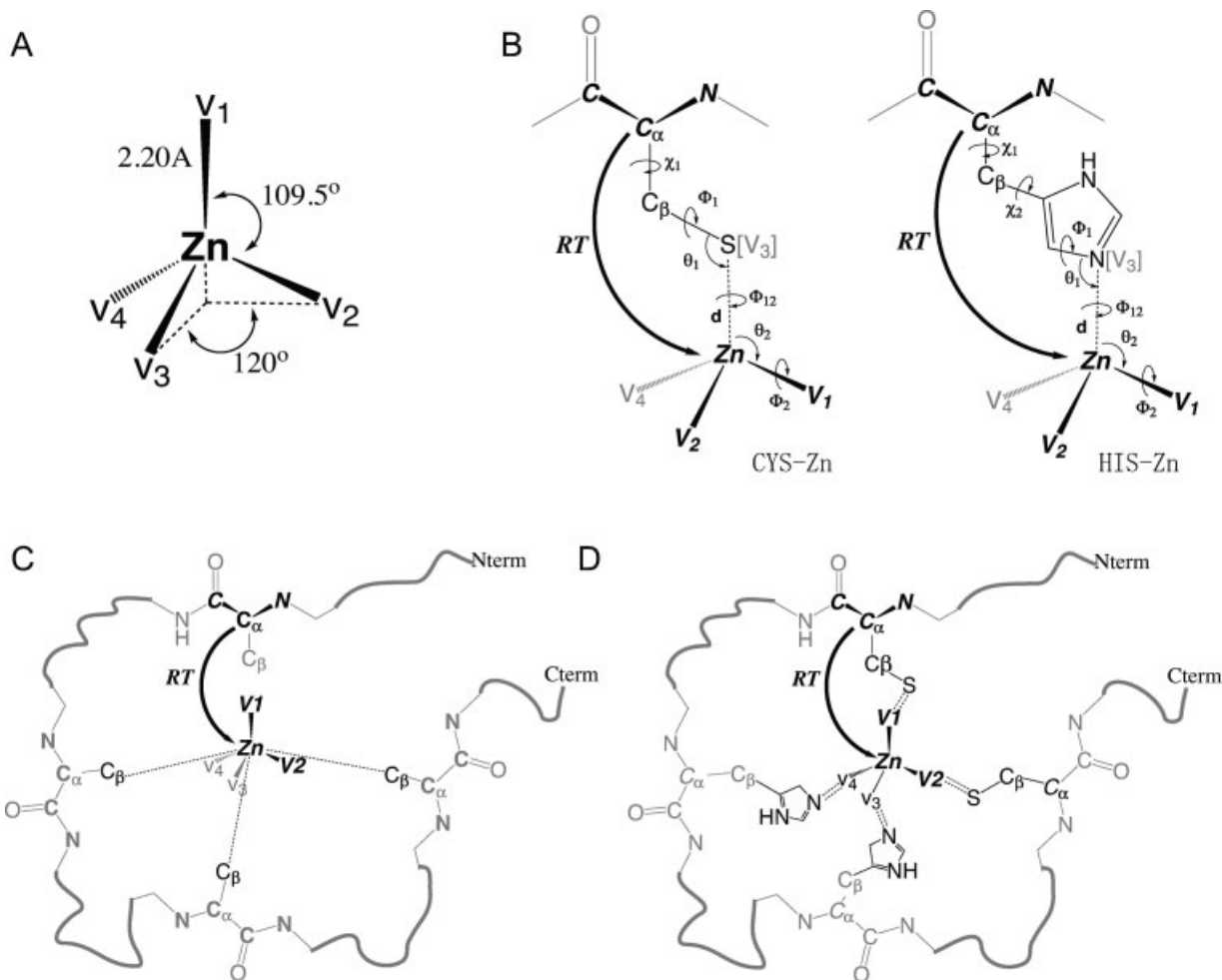


Figure 1. Incorporation of zinc coordination geometry into structure prediction. (A) Zinc is modeled as a tetrahedral ligand with four virtual atoms on the vertices and one zinc atom in the center. The distance between the zinc atom and each of the virtual atoms is 2.20 Å. (B) Rigid-body transformation (RT) from zinc-coordinating residue Cys (left) and His (right) to zinc. Coordinate frames are defined by N, C α , C in protein backbone and Zn, V1, V2 in the zinc. The spatial relationship between the sidechain and zinc can be described by six internal coordinate parameters (Table I), which can have a range of possible values (Table I). Combination of these possibilities with different possible sidechain rotamer conformations results in a discrete set of rigid-body transformations (a “jump” library) that can be used to sample the rigid-body degrees of freedom of the zinc during structure prediction. (C) Low-resolution *de novo* structure prediction. The fold tree is setup so that backbone conformational changes are propagated from protein N-terminal to C-terminal upon changes in the backbone torsion angles by fragment insertion, while a through-space transform (curly arrow) is established from one zinc-coordinating residue to the zinc ligand that samples alternative rigid-body transforms from the “jump” library created in (B). Distance constraints (single dash) tether the C β atoms of the remaining zinc-coordinating residues to the zinc atom. (D) High-resolution refinement. The fold tree is unchanged but with all atoms represented, more precise constraints (double dash, see Materials and Methods) are defined to maintain the integrity of the zinc-coordination site. Virtual atoms are included in the constraints definition to enforce tetrahedral coordination arrangements, which would otherwise be very complicated to realize with the single zinc atom alone.

reference frame for calculating the internal rigid-body transformation from the protein to the zinc, and (2) define constraints consistent with the coordination geometry between the zinc and the coordinating residues. A related “dummy-atom” approach was implemented in a MD study of zinc-bound farnesyltransferase.²⁹

Rosetta’s “fold-tree” representation of the molecular system integrates torsional degrees of freedom with rigid-body degrees of freedom together so they can be optimized simultaneously.^{26,27} In the fold

tree, the zinc ligand is attached to one of its coordinating protein residues via a through-space connection (“jump”) and protein backbone conformational changes propagate through this rigid-body jump to determine the new position of the zinc ligand [Fig. 1(C,D)]. While protein backbone torsion degrees of freedom are sampled through inserting short fragments from known protein structures, the rigid-body degrees of freedom of the jump are also sampled using a precomputed library of sidechain–zinc transforms. To generate this library, the rigid-body

Table I. The Six Internal Coordinate Parameters Defining Local Interactions Between Cys/His Sidechains and Zinc

Residue-Zn	d	θ_1	θ_2	Φ_1	Φ_2	Φ_{12}
Cys-Zn	S $_{\gamma}$ -Zn 2.20 Å	C $_{\beta}$ -S $_{\gamma}$ -Zn 112.0°	S $_{\gamma}$ -Zn-V $_1$ 109.5°	C $_{\alpha}$ -C $_{\beta}$ -S $_{\gamma}$ -Zn -180°:180°:30°	C $_{\beta}$ -S $_{\gamma}$ -Zn-V $_1$ -180°:180°:30°	S $_{\gamma}$ -Zn-V $_1$ -V $_2$ 120.0°
His(D)-Zn	N $_{\delta 1}$ -Zn 2.20 Å	C $_{\gamma}$ -N $_{\delta 1}$ -Zn 120.0°	N $_{\delta 1}$ -Zn-V $_1$ 109.5°	C $_{\beta}$ -C $_{\gamma}$ -N $_{\delta 1}$ -Zn 0.0°	C $_{\gamma}$ -N $_{\delta 1}$ -Zn-V $_1$ -180°:180°:30°	N $_{\delta 1}$ -Zn-V $_1$ -V $_2$ 120.0°
His(E)-Zn	N $_{\epsilon 2}$ -Zn 2.20 Å	C $_{\delta 2}$ -N $_{\epsilon 2}$ -Zn 120.0°	N $_{\epsilon 2}$ -Zn-V $_1$ 109.5°	C $_{\gamma}$ -C $_{\delta 2}$ -N $_{\epsilon 2}$ -Zn 180.0°	C $_{\delta 2}$ -N $_{\epsilon 2}$ -Zn-V $_1$ -180°:180°:30°	N $_{\epsilon 2}$ -Zn-V $_1$ -V $_2$ 120.0°

The parameters consist of one bond length (d), two bond angles (θ_1 and θ_2), and three torsion angles (Φ_1 , Φ_2 , and Φ_{12}), and their defining atoms are listed. For each parameter, the sampling range is indicated, for example, Φ_1 in Cys-Zn is sampled every 30° in a full rotatable manner, but the same torsion in His-Zn is fixed at either 0 or 180° to keep zinc aligned in the plane of the imidazole ring as observed in natural zinc-binding sites. These parameters are used in combination with side-chain rotamers to create a set of rigid-body transformations from residue backbone to the zinc ligand [Fig. 1(B)]. They also serve as a reference for all-atom constraints, which enforce optimal zinc-coordination geometries.

relationships between Cys and His sidechain atoms and zinc were first characterized in structures of proteins with natural zinc-binding sites. Based on this analysis, allowed ranges for each of the six rigid-body degrees of freedom [d , θ_1 , θ_2 , Φ_1 , Φ_2 , and Φ_{12} in Fig. 1(B)] relating the sidechain and the zinc were defined (Table I). Combining these possible sidechain-zinc interactions with all Cys and His sidechain rotamers (χ_1 and χ_2) yields a set of rigid-body transformations from the backbone “N-C $_{\alpha}$ -C” triplet in the zinc-coordinating residue to the “Zn-V $_1$ -V $_2$ ” triplet in the zinc ligand [Fig. 1(B), Table I]. Our library contains about 1300 different transformations (jumps) from the backbone of Cys residues to zinc and 2000 jumps from the backbone of His residues to zinc.

The rigid-body jump described in the previous paragraph anchors the zinc to the protein. Constraints between the zinc and the other three zinc-coordinating residues are included during folding to maintain the integrity of zinc-binding site. In the low-resolution search stage in which protein side-chain atoms are approximated by centroids,²¹ a distance constraint term defined from the zinc atom to the C $_{\beta}$ atom of each of the remaining zinc-coordinating residues favors the formation of a protein topology that can accommodate a zinc-binding site [Fig. 1(C)]. In the subsequent high-resolution refinement stage in which all atoms are represented, the Rosetta all-atom energy function is supplemented with distance, angular, and dihedral constraints derived from structures of zinc-binding sites to ensure that low-energy models are generated that contain a zinc coordination site with correct geometry [Fig. 1(D)]. The distance constraints are defined between a protein zinc-coordinating atom and a virtual atom with a target distance of zero (see Materials and Methods), which enforces the overall tetrahedral coordination geometry around the zinc because in the creation of the zinc ligand, the four virtual atoms occupy the four vertexes of the tetra-

hedron centered at the zinc atom [Fig. 1(A)]. Such treatment allows generation of correct zinc coordination geometries without complicated computation of nonpair-additive interactions between zinc-liganding residues.

With incorporation of zinc into the fold-tree framework of the molecular system by (1) defining a zinc ligand residue with virtual atoms, (2) creating a jump library sampling rigid-body transformations from protein to zinc, and (3) adding constraint energy terms maintaining the geometry of zinc coordination, previously developed Rosetta structure prediction methods can be seamlessly adapted to perform various tasks to generate structure models for zinc-binding proteins. In the next sections, we present results from implementing this new method in *de novo* structure prediction and loop modeling of zinc-binding proteins.

De novo structure prediction

Starting from the amino acid sequence only, Rosetta *de novo* structure prediction and high-resolution refinement have generated structure models with atomic-level accuracy for a handful of benchmark and blind prediction cases.^{23,25} In this study, a benchmark set of nine zinc-binding proteins was constructed to test the performance of the new method with explicit modeling of zinc (Table II). The set represents six of the eight fold groups of zinc fingers as defined by Grishin and coworkers,¹¹ including two targets from each of the three major zinc-finger fold groups — classical C2H2-like zinc finger, treble clef finger, and zinc ribbon. Models were generated for each protein using the Rosetta *de novo* structure prediction method without and with zinc incorporation. Energy versus RMSD plots are shown for low-energy (5%) predictions in Figure 2(A). For six of nine cases (1co4, 1d0q, 1ef4, 1fv5, 1ncs, and 2b9d), improved sampling toward near-native conformations is observed. For four cases (1ef4, 1fv5, 1wjb, and 2b9d), the overall “energy-funnel” character was

Table II. Benchmark Sets of Zinc-Binding Proteins from Various Fold Groups for Testing De Novo Structure Prediction

PDB	Length	Structural classification	Ligands	BL5	
				Control	Zinc
1co4	1–34	Zn2/Cys6-like finger	C11, C14, C23, H25	6.25	6.14
1d0q	2–103	Zinc ribbon	C40, H43, C61, C64	11.99	5.20
1dsv	54–84	Gag knuckle	C58, C61, H66, C71	6.33	7.31
1ef4	1–55	Treble clef	C6, C9, C43, C44	4.06	3.71
1fv5	10–33	C2H2-like finger	C11, C14, H27, C32	3.82	0.70
1irn	1–53	Zinc ribbon	C6, C9, C39, C42	5.90	7.35
1ncs	21–60	C2H2-like finger	C34, C39, H52, H56	8.54	1.59
1wjb	1–46	TAZ2 domain-like	H12, H16, C40, C43	2.89	2.07
2b9d	42–93	Treble clef	C52, C55, C85, C88	6.46	3.04

The lowest RMSD of the lowest energy five models (BL5) is reported for the test without zinc (control) and the test with zinc (zinc).

improved when zinc is explicitly incorporated in the process of structure prediction. In both cases of the classical C2H2-like zinc finger proteins (1fv5 and 1ncs), near-native models (<2 Å backbone RMSD) were identified among the five best energy-ranked

predictions [Fig. 2(A), Table II]. Accurately predicted models of 1fv5, 2b9d, and 1wjb are shown in Figure 2(B); for 1fv5 the prediction has atomic level accuracy with backbone RMSD less than 1.0 Å and all-atom RMSD less than 2.0 Å.

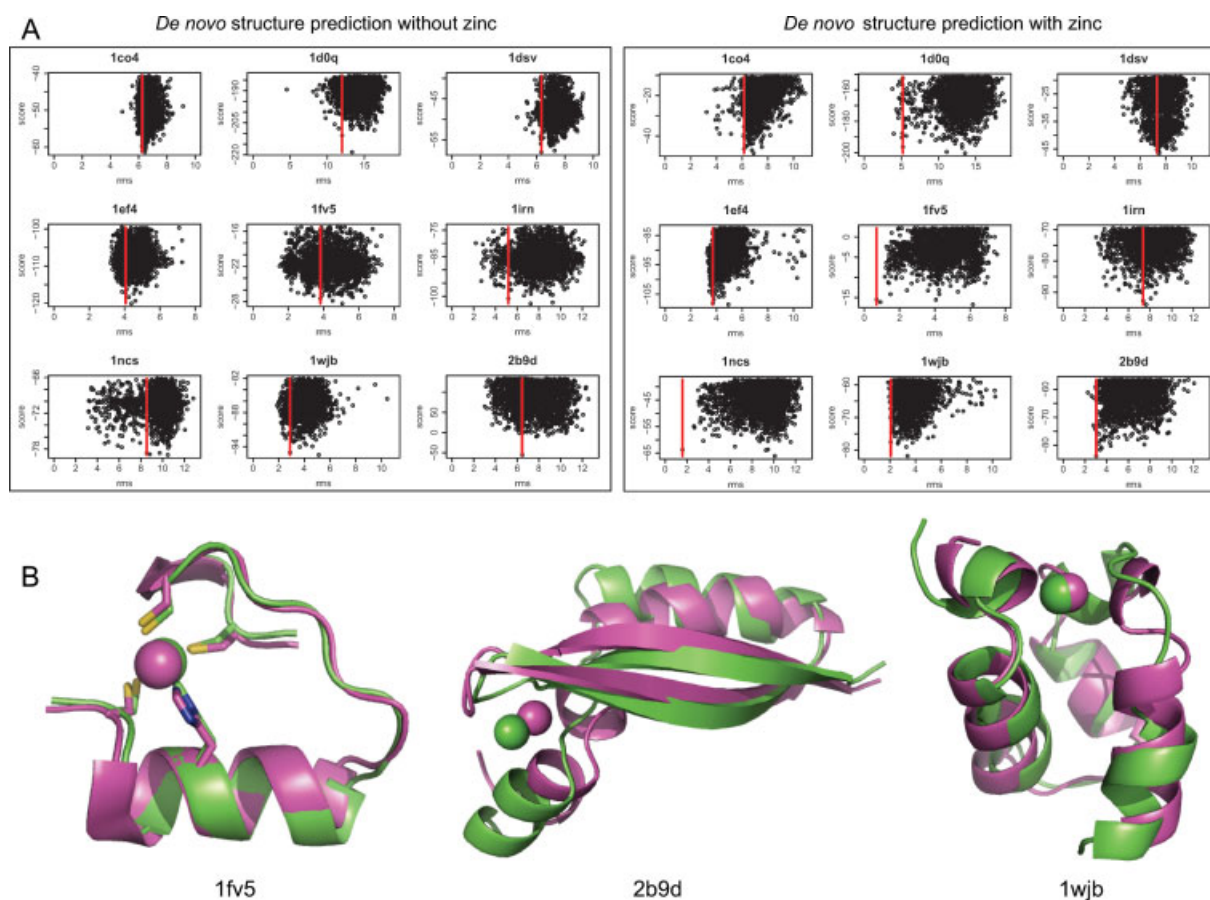


Figure 2. De novo prediction of the structures of zinc-binding proteins. (A) Energy (y -axis) versus RMSD (x -axis) plots of the lowest energy 5% of models generated without (left) and with (right) explicit zinc modeling. The red line in each plot indicates the lowest RMSD value of the lowest energy five models (“BL5” in Table II). (B) Accurate predictions of 1fv5 (left, 0.7 Å RMSD), 2b9d (middle, 3.04 Å RMSD), and 1wjb (right, 2.07 Å RMSD) from low-energy models with zinc incorporated. The predicted model (pink) is superimposed onto the native structure (green). Backbone traces are drawn in cartoon, zinc ions are drawn in spheres, and zinc-coordinating sidechains are drawn in sticks.

De novo structure prediction using NMR chemical shift information

It was demonstrated recently that the robustness of Rosetta *de novo* structure prediction method can be improved by using a fragment library generated with NMR chemical shift data (CS-fragment).^{30,31} For metal-binding proteins, chemical shift information can further provide valuable insights on structure features such as metal ligation. Montelione and coworkers recently showed that overlapped $^{13}\text{C}_\beta$ chemical shift distributions of zinc-liganding and nonmetal-liganding cysteine residues are largely resolved by the inclusion of the corresponding $^{13}\text{C}_\alpha$ chemical shift information.¹⁶ Here, we take the nine proteins in that study whose chemical shift data were used to identify their zinc-liganding cysteines¹⁶ and generate models using Rosetta *de novo* structure prediction and refinement protocols. Four protocols including with/without zinc and with/without CS-fragments were tested, and the results are summarized in Figure 3(A) and Table III. Compared to the control protocol (black curve, without zinc and without CS-fragments), the new protocol (blue curve, with explicit zinc and with CS-fragments) has RMSD distributions shifted toward near-native conformations in seven of nine cases [Fig. 3(A)], and the energy-based ranking of modeled structures is improved for six cases (Table III). Incorporating zinc and using a chemical shift-based fragment library produce different levels of improvement for different protein targets. In 1lv3 and 1r9p, improvements mainly come from incorporating zinc, whereas in 1m3v and 1iyf, CS-fragments play a more important role in creating near-native models. In 1exk, including both zinc and CS-fragments have a synergistic impact. Significantly improved results are obtained for both 1r9p (with one zinc-binding site) and 1m3v (with two zinc-binding sites) with predictions of 2.21 and 2.66 Å backbone RMSD identified in the best five energy-ranked models. As illustrated in Figure 3(B), both overall protein topology and zinc positions are predicted accurately.

Loop modeling

One of the important goals of computational structure biology is to model protein structures accurately from homologues of known structures. A critical step in this process is the modeling of structurally divergent regions using “loop modeling” methods. Several loop modeling methods have been developed in Rosetta^{27,32} and have been applied in CASP blind predictions to create accurate models.^{24,25} In the current test, 16 crystal structures of zinc-binding proteins were selected, which have at least two zinc-coordinating residues residing in one or more loop regions (Table IV). These loop regions were built using a previously published protocol²⁷ coupling

cyclic coordinate descent (CCD) algorithm³³ with Monte Carlo energy minimization.³⁴ For each test case, 6000 models were generated with or without the explicit incorporation of zinc. Distributions of the global loop RMSD values from the 300 lowest energy models are plotted in Figure 4(A), and the best global loop RMSD value from the five lowest energy models (BL5) is reported in Table IV. When loops are modeled in the presence of zinc, 7 of 16 cases show improved results, while the performance for the rest of the cases does not become significantly worse. For 1d0q, 2ayd, 2ioi, and 2orw, the accuracies of modeled loop conformations and the energetic discrimination between near-native and incorrect models are dramatically improved as evidenced by both a substantial RMSD distribution shift toward the native loop conformation [Fig. 4(A)] and the significantly lower RMSD values among the five lowest energy models (Table IV). For 2ayd, two loops containing all four zinc-coordinating residues are modeled simultaneously with a RMSD of 1.34 Å, whereas for 2orw, a 15-residue long loop accommodating two zinc-coordinating residues is predicted with an RMSD of 1.29 Å. In both cases, as illustrated in Figure 4(B), accurate predictions are achieved not only for loop backbone conformations but also for the sidechain conformations of the zinc-coordinating residues as well as the zinc position, which would not be possible without explicit incorporation of the zinc into the modeling process.

Discussion

Metal ions are essential to maintain the function, structure, and stability of proteins and, as the second abundant metal ion found in eukaryotic organisms, zinc plays important roles in many biological processes. About 10% of the structures deposited in the Protein Data Bank have zinc listed as a ligand in their structure records.⁷ Despite the flourishing development of computational tools to generate protein structure models either from sequence alone or from structures of close homologues, few methods take the binding of metal cofactors into account explicitly. The method presented in this article is a step toward overcoming this limitation.

The development of the current method has greatly benefited from both the representation of the molecular system by a “fold tree”²⁶ and the recent effort to reshape Rosetta software with a modular object-oriented design (Leaver-Fay A, Baker D, and Bradley P, unpublished). The fold tree lays out a general kinematic framework for a wide spectrum of structure modeling tasks in which torsional degrees of freedom and rigid-body degrees of freedom can be integrated seamlessly and optimized simultaneously. The power and generality of this framework is illustrated by predictions of the structures of 1m3v (protein folding with two zinc ions) and 2ayd (two loops

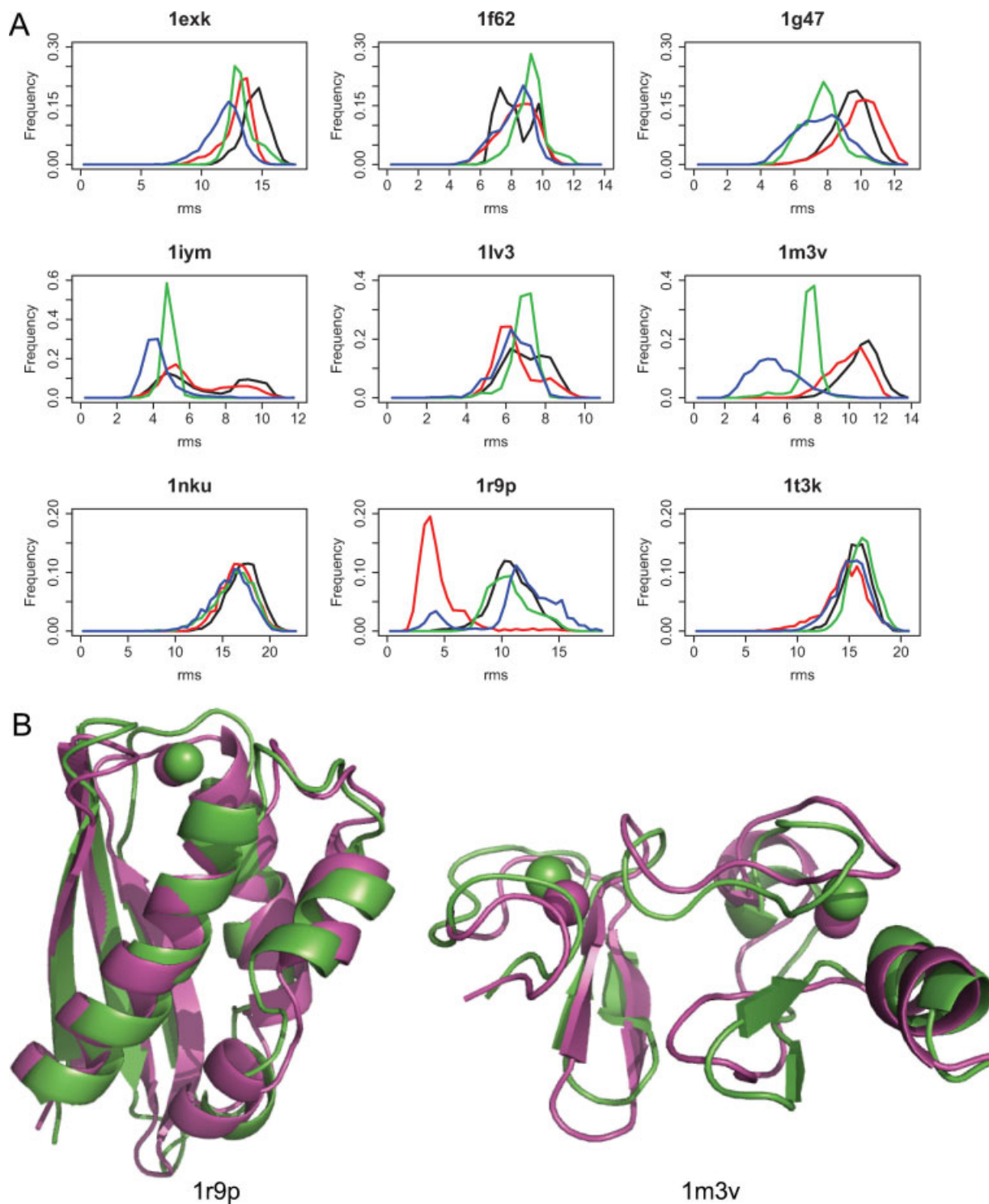


Figure 3. Prediction of the structures of zinc-binding proteins using NMR chemical shift information. (A) Comparison of RMSD distribution of low-energy models generated with and without explicit zinc modeling. Backbone heavy-atom RMSD values of the lowest energy 5% of models for each of nine NMR protein structures were grouped into 0.5 Å bins. Histograms are shown for calculations without zinc using standard fragments (black), without zinc using CS-fragments (green), with zinc using standard fragments (red), and with zinc using CS-fragments (blue). (B) Accurate predictions of 1r9p (left, 2.08 Å RMSD) and 1m3v (right, 2.66 Å RMSD, two zinc-binding sites) from low-energy models predicted with zinc incorporated. The predicted model (pink) is superimposed onto the native structure (green). Backbone traces are drawn in cartoon and zinc ions are drawn in spheres.

Table III. Benchmark Set of Zinc-Binding Proteins for Testing Structure Prediction Using Chemical Shift Information

PDB	Length	Ligands	BL5			
			Control	Control-CS	Zinc	Zinc-CS
1exk	13–75	C14, C17, C67, C70 C31, C34, C53, C56	13.25	11.38	11.66	11.96
1f62	1–48	C3, C6, H26, C29 C18, C21, C44, C47	7.06	7.67	4.07	4.86
1g47	8–66	C10, C13, H32, C35 C38, C41, C59, H61	6.78	5.97	9.53	5.04
liym	133–179	C134, C137, H158, C161 C153, H155, C172, C175	2.85	4.73	3.95	3.71
1lv3	5–40	C9, C12, C28, C32	5.85	6.80	5.42	7.11
1m3v	7–67	C8, C11, H29, C32 C35, C38, C58, D61	8.80	6.88	8.88	2.66
1nku	1–184	C4, H17, H175, C179	14.68	11.45	14.79	11.37
1r9p	26–122	C37, C63, H105, C106	7.04	6.70	1.93	2.08
1t3k	1–132	H39, C120, C122, C127	15.04	15.11	6.33	14.86

The best RMSD value of the lowest energy five models (BL5) is reported for tests without zinc using normal fragments (control), tests without zinc using CS-fragments (control-CS), tests with zinc using normal fragments (zinc), and tests with zinc using CS-fragments (zinc-CS).

built to coordinate the zinc ion) presented here. The new modular architecture allows easy integration into protein structure prediction and design calculations of nonprotein molecules with well-defined coordination geometries, such as metal ions and clusters, water molecules, and other small molecules with well-defined hydrogen bond acceptors and donors. Once the coordination/ligation geometry is specified, the Rosetta3 framework enables fast and efficient development of modeling methodologies with these

compounds based on existing protocols (e.g., *de novo* structure prediction and loop modeling).

In conventional protein structure prediction methods, interactions among protein atoms are often approximated using pair-additive distance-dependent potentials. This approach is problematic for modeling metal–protein interactions because formation of correct metal coordination geometries requires simultaneous satisfaction of distance, angle, and dihedral geometric constraints from multiple protein

Table IV. Benchmark Set of Zinc-Binding Proteins for Testing Loop Modeling

PDB	Length	Ligands	Loops	Nres	BL5	
					Control	Zinc
1d0q	2–103	C40, H43, C61, C64	38–49 61–70	4	3.86	1.56
1ee8	1–266	C238, C241, C258, C261	238–242 257–264	4	1.21	1.12
1kk1	6–198	C60, C62, C72, C75	60–80	4	8.48	8.43
1oqj	90–179	C113, H170, C174, C178	107–120 162–177	3	3.31	3.73
1v33	1–346	C106, H108, C114, C117	97–115	3	5.42	3.27
1vsr	23–156	C66, H71, C73, C117	64–82 115–127	4	4.33	3.79
1zin	1–217	C130, C133, C150, C153	130–134 138–165	4	8.78	6.96
2ayd	293–368	C332, C337, H361, H363	332–340 358–366	4	1.75	1.34
2d5b	1–287	C127, C130, C144, H147	127–152	4	4.35	6.12
2gmw	24–205	C112, H114, C127, C129	112–135	4	7.69	6.83
2ioi	1097–1283	C1173, H1176, C1235, C1239	1233–1248	2	3.12	0.88
2j6a	1–136	C11, C16, C112, C115	8–33 112–116	4	5.30	5.29
2olm	3–135	C29, C32, C49, C52	22–50	3	6.47	6.00
2orw	2–181	C140, C143, C173, C176	135–149	2	7.89	1.29
2pq8	177–305	C210, C213, H226, C230	210–214 229–238	3	2.10	2.49
2znr	270–436	H362, C402, H408, H410	401–431	3	5.79	3.93

The number of zinc-coordinating residues residing in the defined loop regions (Nres) is indicated. The lowest RMSD of the lowest energy five models ranked by energy (BL5) is reported for tests without zinc (control) and tests with zinc (zinc).

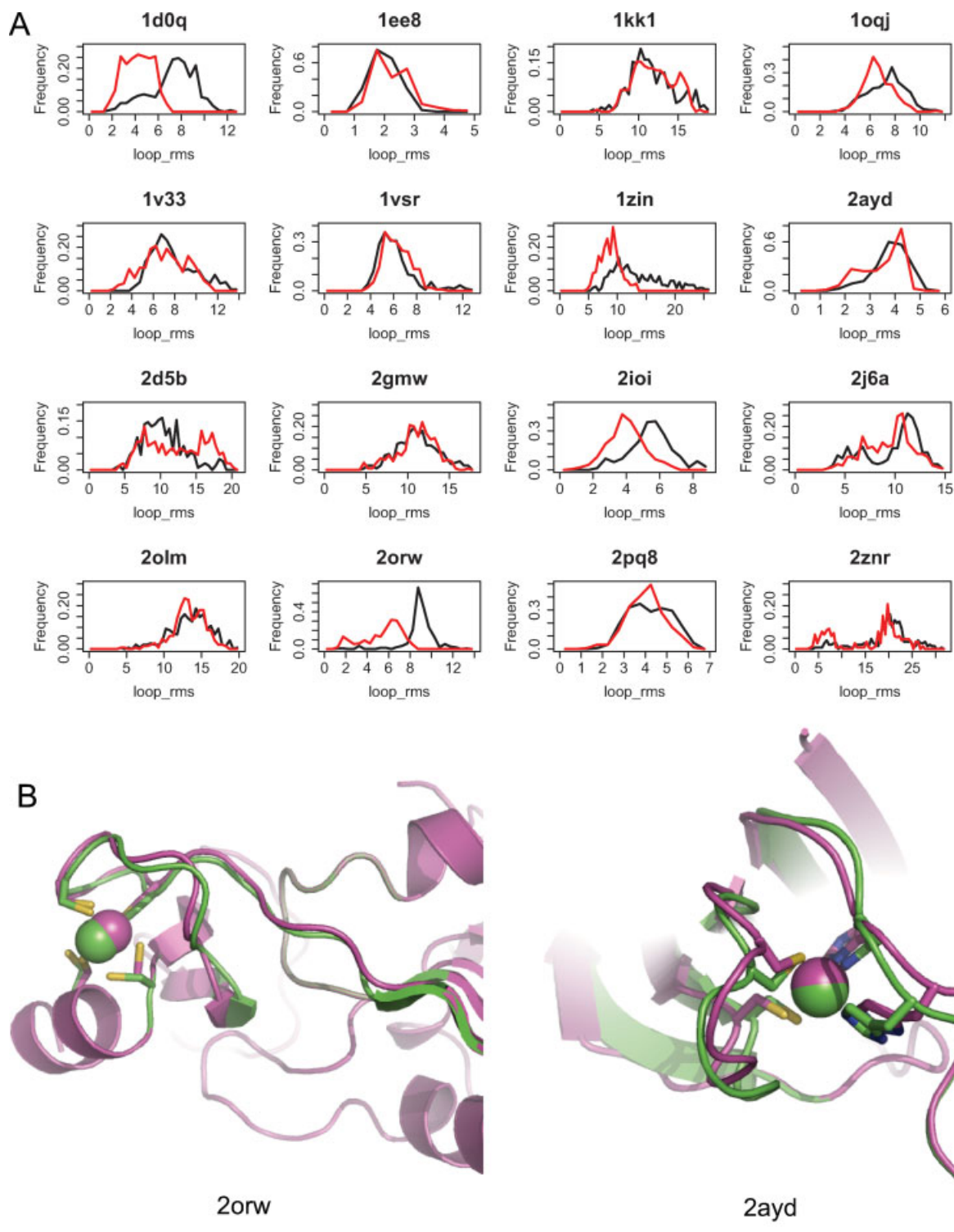


Figure 4. Loop modeling near zinc-binding sites. (A) Comparison of RMSD distribution of low-energy models generated with and without explicit zinc modeling. Histograms of loop backbone heavy-atom RMSD of the lowest energy 5% of models are shown for tests without zinc (black) and with zinc (red). (B) Accurate predictions of 2orw (left, 1.29 Å RMSD) and 2ayd (1.34 Å RMSD) from low-energy models with zinc incorporated. The predicted model (pink) is superimposed onto the native structure (green). Backbone traces are drawn in cartoon, zinc ions are drawn in spheres, and zinc-coordinating sidechains are drawn in sticks.

atoms surrounding the ion. Our method addresses this challenge by representing the zinc ion with a tetrahedron-shaped residue. In this pseudo-residue,

the actual zinc atom is positioned at the center and the four virtual atoms occupy the vertexes of the tetrahedron, mimicking zinc coordination spheres in

native protein structures. By enforcing distance constraints between zinc-coordinating atoms from proteins and these virtual atoms, as well as angular/dihedral constraints, correct zinc-coordinating geometries are favored by the energy function without incurring the complexity of computing multibody interactions.

The approach to zinc incorporation in this article can easily be extended to model protein structures bound with other metals, such as calcium, iron, and magnesium. Metal ion coordination geometry and sidechain preferences have been extensively studied,³⁵ and once such information is encoded as done here for zinc (creation of ligand residue, “jump” library, and coordination constraints), these metals (or more broadly, small chemical ligands) can be readily incorporated into existing methods to predict structures of other metalloproteins and/or dock metalloproteins with protein or ligand partners.

Our method currently relies on knowledge of the locations of the zinc-coordinating residues in the protein primary sequence. Such information may be obtained from analysis of consensus metal-binding sequence patterns in genome sequences,^{6,13} alignments with other known homologues,¹¹ and experimental data such as NMR chemical shifts¹⁶ or mutagenesis around metal-binding sites. With the incorporation of zinc-coordinating constraints in prediction, the conformational space to be searched is certainly reduced; however, we still have several test cases (1dsv, 1t3k, 1nku, 2d5b, etc.) in which near-native conformations are not sampled, highlighting the importance of developing algorithms to better search conformational space. Although the method described in this article mainly focuses on structurally bound zinc metals in proteins, the catalytic role of zinc binding in many metalloenzymes should not be overlooked. To model their structures, energy functions need to be improved to monitor electrostatic interactions among cationic metals, protonated waters, and more acidic residues in the active site such as Asp and Glu.³⁶

Metal binding promotes protein stability and catalytic activity, and attention has been increasingly focused on designing interactions between protein and metal ions.^{37,38} Previous studies have explored the introduction of zinc- and iron-binding sites into static protein scaffolds^{39–43} such as four-helix bundles, however, as suggested by recent work on *de novo* protein structure design⁴⁴ and enzyme design,⁴⁵ successful creation of metalloproteins with novel structure and function will likely require iterative rounds of design and prediction of protein scaffolds with structural and/or catalytic metal-binding sites. The method described in this article can serve to create an initial structure model for sequence optimization and to refine designed sequences and structures.

Materials and Methods

Datasets

Nine protein targets were selected from Krishna *et al.*¹¹ representing six of the eight defined classes of zinc-binding proteins to test the *de novo* structure prediction protocol. Two targets were selected for each of the three major fold groups: C2H2-like finger, treble clef finger, and zinc ribbon. The nine protein targets used to test structure prediction with NMR chemical shift information were selected from the set of Kornhaber *et al.*¹⁶ The first model in the NMR ensemble was used as the native conformation with the flexible terminal residues removed. To test the loop modeling protocol, 16 crystal structures with resolution better than 2.5 Å were selected from the Protein Data Bank,⁷ which contain loop regions with at least two zinc-coordinating residues. Information on the four residues coordinating the zinc was extracted from the structures and used to guide *de novo* structure prediction and loop modeling.

De novo structure prediction

The Rosetta *de novo* structure prediction method has been described in detail.^{21,46} Models are built from fragments starting from an extended chain and then subjected to all-atom refinement. Two sets of torsional fragment libraries were tested, one created with standard procedures based on local sequence similarity⁴⁶ and the other created with additional chemical shift information³¹ retrieved from Biological Magnetic Resonance Data Bank (BMRB, <http://www.bmrb.wisc.edu/published/>). When zinc is incorporated, it is treated as an additional ligand residue and is attached to one coordinating residue (closest to protein N-terminal) via a long-range connection in the fold tree. A library of rigid-body transformations from the backbone of the coordinating residue to the zinc ligand is generated by combining all parameters of freedom as listed in Table I. During the course of fragment assembly, the “jump” fragment from this library can be inserted and selected using a Monte Carlo strategy to sample the rigid-body orientation of zinc with respect to protein backbone. All backbone and sidechain torsional degrees of freedom and zinc rigid-body degrees of freedom are optimized simultaneously in the all-atom refinement stage. In both the low-resolution folding and high-resolution stages, tethering constraints are implemented to favor keeping zinc-coordination geometry (see the section of “energy function”). For each protein, 50,000 models were generated and the first 2500 models (5%) ranked by energy were selected for further analysis. With the incorporation of zinc, the computational cost of Rosetta *de novo* structure prediction generally increases by about 20–50% depending on the size of protein being modeled.

Loop modeling

The loop modeling method used in this article and the fold-tree setup were described in detail by Wang *et al.*²⁷ It couples CCD algorithm³³ with Monte Carlo energy minimization³⁴ to build loops onto protein template structures. The native loop conformations were removed from template structures before modeling. Multiple loops are constructed in the low-resolution stage in a randomly selected order and then optimized simultaneously in the high-resolution refinement stage. The zinc ligand is allowed to be freely moved in space, and its interactions with the four coordinating residues are rewarded by constraint terms defined in energy function. Six thousand models were generated for each test case, and the first 300 models (5%) ranked by energy were selected for further analysis.

Energy function

Standard Rosetta low-resolution and all-atom energy functions were used in generating and ranking models.^{21,22} The virtual atoms in the zinc ligand have no physical interactions with other protein atoms, and they are implemented only for the purpose of defining zinc-coordination constraints. The zinc atom is treated as a backbone C_α atom in the low-resolution stage and in the all-atom stage, force field parameters for zinc ion from CHARMM27⁴⁷ were used to model its interaction with the rest of protein. Additional constraints energies are defined to favor formation of zinc-coordination sites with satisfying geometry. In the low-resolution stage, a distance constraint is defined between the zinc atom and the C_β atom of each zinc-coordinating residue with a penalty function form of

$$E_{\text{cst}} = \begin{cases} \left(\frac{d-u_b}{\Delta}\right)^2 & \text{if } d > u_b \\ 0 & \text{if } l_b \leq d \leq u_b \\ \left(\frac{l_b-d}{\Delta}\right)^2 & \text{if } d < l_b \end{cases},$$

where d is the actual distance between zinc and C_β, Δ is a constant of 0.2 Å. u_b and l_b are 2.8 and 3.8 Å for Cys-zinc coordination and 3.2 and 4.0 Å for His-zinc coordination. In the all-atom refinement stage, the constraint energy for each zinc-residue coordination is composed of three terms:

$$E_{\text{cst}} = E_{\text{dis}} + E_{\text{ang}} + E_{\text{dih}} \\ = \left(\frac{d-d_0}{\Delta_d}\right)^2 + \left(\frac{\theta_1-\theta_0}{\Delta_\theta}\right)^2 + \left(\frac{\Phi_1-\Phi_0}{\Delta_\Phi}\right)^2,$$

where θ_1/θ_0 and Φ_1/Φ_0 are the actual/optimal values of bond angles and dihedral angles for zinc coordination as defined in Table I, respectively, with Δ_θ and Δ_Φ both equal to 20°. d is the distance between the zinc-coordinating atom and one of the virtual atoms,

d_0 is 0.0 Å and Δ_d 0.2 Å. The purpose of defining the distance constraints using virtual atoms instead of the actual zinc atom is to explicitly favor tetrahedral zinc coordination while keeping the coordination distance optimal. As virtual atoms are tethered to four unique zinc-coordinating residues in protein sequence, the zinc atom essentially becomes a chiral center with (A₁/V₁, A₂/V₂, A₃/V₃, and A₄/V₄) and (A₁/V₁, A₂/V₂, A₃/V₄, A₄/V₃) corresponding to two different zinc-coordination sites (V₁, V₂, V₃, and V₄ are four virtual atoms in the zinc ligand and A₁, A₂, A₃, and A₄ are the zinc-coordinating atoms from the four residues ordered from N-terminal to C-terminal). When the modeling process enters all-atom refinement, both “chiral” constraints are provided and one is randomly chosen to proceed to generate a final model. All constraint penalties for each pair of zinc-residue interaction are summed together and added to the total energy of the model with a weight of 0.01 and 0.1 for low-resolution and high-resolution energy functions, respectively.

Evaluation of model accuracy

To evaluate model accuracy in the loop modeling test, the RMSD is calculated over all backbone heavy atoms in the loop region between the model and the native structure after the backbones of non-loop regions of the two proteins are superimposed. To evaluate the accuracy of models generated in *de novo* structure prediction tests, the RMSD is calculated over all backbone heavy atoms in the entire protein chain after the model and native structure are optimally superimposed.

Plots and figures

R (<http://www.r-project.org/>) was used to make energy versus RMSD plots and RMSD distributions, and PYMOL (<http://www.pymol.org>) was used to produce figures for protein models.

BOINC and Rosetta@Home

Rosetta@Home (<http://boinc.bakerlab.org/rosetta/>), a distributed computing project running the Rosetta software on personal computers of volunteers from all over the world using the Berkeley Open Infrastructure for Network Computing (BOINC) technology, was critical to the method development and model production described in this article. This substantial computing resource allowed us to rapidly test and improve the new methodology at a level not possible with only in-house computing resources.

Software availability

The software described in this article is available free for academic use at <http://www.rosettacommons.org/> as part of the Rosetta software suite release 3.1 (SVN#33180) or newer. The command line

options used for this study are provided in the electronic Supporting Information available over the internet as part of the Electronic Edition of Protein Science.

Acknowledgments

The authors thank many scientists who have participated in the development of the suite of computational tools used in the Baker laboratory for computations on the structure of proteins. In particular, Philip Bradley and Andrew Leaver-Fay made key contribution to reshaping Rosetta's software architecture into an advanced modular design, which paves the way for efficient software development. David Kim built and maintained the Rosetta@Home project. Keith Laidig and Darwin Alonso maintained reliable, state-of-the-art computing resources. We thank all the Rosetta@Home users worldwide for generously donating their computer time for their scientific research.

References

1. Wu FY, Wu CW (1987) Zinc in DNA replication and transcription. *Annu Rev Nutr* 7:251–272.
2. Sunderman FW, Jr (1995) The influence of zinc on apoptosis. *Ann Clin Lab Sci* 25:134–142.
3. Murakami M, Hirano T (2008) Intracellular zinc homeostasis and zinc signaling. *Cancer Sci* 99:1515–1522.
4. McCall KA, Huang C, Fierke CA (2000) Function and mechanism of zinc metalloenzymes. *J Nutr* 130:1437S–1446S.
5. Berg JM, Shi Y (1996) The galvanization of biology: a growing appreciation for the roles of zinc. *Science* 271:1081–1085.
6. Andreini C, Banci L, Bertini I, Rosato A (2006) Counting the zinc-proteins encoded in the human genome. *J Proteome Res* 5:196–201.
7. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. *Nucleic acids Res* 28:235–242.
8. Christianson DW (1991) Structural biology of zinc. *Adv Protein Chem* 42:281–355.
9. Auld DS (2001) Zinc coordination sphere in biochemical zinc sites. *Biometals* 14:271–313.
10. Patel K, Kumar A, Durani S (2007) Analysis of the structural consensus of the zinc coordination centers of metalloprotein structures. *Biochim Biophys Acta* 1774:1247–1253.
11. Krishna SS, Majumdar I, Grishin NV (2003) Structural classification of zinc fingers: survey and summary. *Nucleic Acids Res* 31:532–550.
12. Torrance JW, Macarthur MW, Thornton JM (2008) Evolution of binding sites for zinc and calcium ions playing structural roles. *Proteins* 71:813–830.
13. Shu N, Zhou T, Hovmoller S (2008) Prediction of zinc-binding sites in proteins from sequence. *Bioinformatics* 24:775–782.
14. Sodhi JS, Bryson K, McGuffin LJ, Ward JJ, Wernisch L, Jones DT (2004) Predicting metal-binding site residues in low-resolution structural models. *J Mol Biol* 342:307–320.
15. Schymkowitz JW, Rousseau F, Martins IC, Ferkinghoff-Borg J, Stricher F, Serrano L (2005) Prediction of water and metal binding sites and their affinities by using the Fold-X force field. *Proc Natl Acad Sci USA* 102:10147–10152.
16. Kornhaber GJ, Snyder D, Moseley HN, Montelione GT (2006) Identification of zinc-ligated cysteine residues based on $^{13}\text{C}\alpha$ and $^{13}\text{C}\beta$ chemical shift data. *J Biomol NMR* 34:259–269.
17. Seebeck B, Reulecke I, Kamper A, Rarey M (2008) Modeling of metal interaction geometries for protein-ligand docking. *Proteins* 71:1237–1254.
18. Pang YP (2001) Successful molecular dynamics simulation of two zinc complexes bridged by a hydroxide in phosphotriesterase using the cationic dummy atom method. *Proteins* 45:183–189.
19. Calhoun JR, Liu W, Spiegel K, Dal Peraro M, Klein ML, Valentine KG, Wand AJ, DeGrado WF (2008) Solution NMR structure of a designed metalloprotein and complementary molecular dynamics refinement. *Structure* 16:210–215.
20. Baker D, Sali A (2001) Protein structure prediction and structural genomics. *Science* 294:93–96.
21. Rohl CA, Strauss CE, Misura KM, Baker D (2004) Protein structure prediction using Rosetta. *Methods Enzymol* 383:66–93.
22. Das R, Baker D (2008) Macromolecular modeling with rosetta. *Annu Rev Biochem* 77:363–382.
23. Bradley P, Misura KM, Baker D (2005) Toward high-resolution de novo structure prediction for small proteins. *Science* 309:1868–1871.
24. Das R, Qian B, Raman S, Vernon R, Thompson J, Bradley P, Khare S, Tyka MD, Bhat D, Chivian D, Kim DE, Sheffler WH, Malmström L, Wollacott AM, Wang C, Andre I, Baker D (2007) Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home. *Proteins* 69 (Suppl 8):118–128.
25. Qian B, Raman S, Das R, Bradley P, McCoy AJ, Read RJ, Baker D (2007) High-resolution structure prediction and the crystallographic phase problem. *Nature* 450:259–264.
26. Bradley P, Baker D (2006) Improved beta-protein structure prediction by multilevel optimization of nonlocal strand pairings and local backbone conformation. *Proteins* 65:922–929.
27. Wang C, Bradley P, Baker D (2007) Protein-protein docking with backbone flexibility. *J Mol Biol* 373:503–519.
28. Davis IW, Baker D (2009) RosettaLigand docking with full ligand and receptor flexibility. *J Mol Biol* 385:381–392.
29. Pang YP, Xu K, Yazal JE, Prendergas FG (2000) Successful molecular dynamics simulation of the zinc-bound farnesyltransferase using the cationic dummy atom approach. *Protein Sci* 9:1857–1865.
30. Shen Y, Lange O, Delaglio F, Rossi P, Aramini JM, Liu G, Eletsky A, Wu Y, Singarapu KK, Lemak A, Ignatchenko A, Arrowsmith CH, Szyperski T, Montelione GT, Baker D, Bax A (2008) Consistent blind protein structure generation from NMR chemical shift data. *Proc Natl Acad Sci USA* 105:4685–4690.
31. Shen Y, Vernon R, Baker D, Bax A (2009) De novo protein structure generation from incomplete chemical shift assignments. *J Biomol NMR* 43:63–78.
32. Rohl CA, Strauss CE, Chivian D, Baker D (2004) Modeling structurally variable regions in homologous proteins with rosetta. *Proteins* 55:656–677.
33. Canutescu AA, Dunbrack RL, Jr (2003) Cyclic coordinate descent: a robotics algorithm for protein loop closure. *Protein Sci* 12:963–972.
34. Li Z, Scheraga HA (1987) Monte Carlo-minimization approach to the multiple-minima problem in protein folding. *Proc Natl Acad Sci USA* 84:6611–6615.
35. Messerschmidt A (2001) Handbook of metalloproteins. Chichester: Wiley.

36. Vallee BL, Auld DS (1992) Active zinc binding sites of zinc metalloenzymes. *Matrix Suppl* 1:5–19.
37. Kennedy ML, Gibney BR (2001) Metalloprotein and redox protein design. *Curr Opin Struct Biol* 11:485–490.
38. Ghosh D, Pecoraro VL (2005) Probing metal-protein interactions using a de novo design approach. *Curr Opin Chem Biol* 9:97–103.
39. Handel T, DeGrado WF (1990) Denovo design of a Zn²⁺-binding protein. *J Am Chem Soc* 112:6710–6711.
40. Hellinga HW, Caradonna JP, Richards FM (1991) Construction of new ligand binding sites in proteins of known structure. II. Grafting of a buried transition metal binding site into *Escherichia coli* thioredoxin. *J Mol Biol* 222:787–803.
41. Klemba M, Gardner KH, Marino S, Clarke ND, Regan L (1995) Novel metal-binding proteins by design. *Nat Struct Biol* 2:368–373.
42. Wisz MS, Garrett CZ, Hellinga HW (1998) Construction of a family of Cys2His2 zinc binding sites in the hydrophobic core of thioredoxin by structure-based design. *Biochemistry* 37:8269–8277.
43. Bender GM, Lehmann A, Zou H, Cheng H, Fry HC, Engel D, Therien MJ, Blasie JK, Roder H, Saven JG, DeGrado WF (2007) De novo design of a single-chain diphenylporphyrin metalloprotein. *J Am Chem Soc* 129:10732–10740.
44. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D (2003) Design of a novel globular protein fold with atomic-level accuracy. *Science* 302:1364–1368.
45. Murphy PM, Bolduc JM, Gallaher JL, Stoddard BL, Baker D (2009) Alteration of enzyme specificity by computational loop remodeling and design. *Proc Natl Acad Sci USA* 106:9215–9220.
46. Simons KT, Kooperberg C, Huang E, Baker D (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J Mol Biol* 268:209–225.
47. Stote RH, Karplus M (1995) Zinc binding in proteins and solution: a simple but accurate nonbonded representation. *Proteins* 23:12–31.