Published in final edited form as:

J Mol Biol. 2009 January 30; 385(4): 1314–1329. doi:10.1016/j.jmb.2008.10.093.

OPUS-Dom: Applying the Folding-Based Method VECFOLD to Determine Protein Domain Boundaries

Yinghao Wu^{1,†}, Athanasios D. Dousis^{1,†}, Mingzhi Chen^{2,†}, Jialin Li², and Jianpeng Ma^{1,2,3,*}

¹Department of Bioengineering, Rice University, Houston, TX 77005, USA

²Graduate Program of Structural and Computational Biology and Molecular Biophysics, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA

³Verna and Marrs McLean, Department of Biochemistry and Molecular Biology, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA

Abstract

In this article, we present a *de novo* method for predicting protein domain boundaries, called OPUS-Dom. The core of the method is a novel coarse-grained folding method, VECFOLD, which constructs low-resolution structural models from a target sequence by folding a chain of vectors representing the predicted secondary-structure elements. OPUS-Dom generates a large ensemble of folded structure decoys by VECFOLD and labels the domain boundaries of each decoy by a domain parsing algorithm. Consensus domain boundaries are then derived from the statistical distribution of the putative boundaries and three empirical sequence-based domain profiles. OPUS-Dom generally outperformed several state-of-the-art domain prediction algorithms over various benchmark protein sets. Even though each VECFOLD-generated structure contains large errors, collectively these structures provide a more robust delineation of domain boundaries. The success of OPUS-Dom suggests that the arrangement of protein domains is more a consequence of limited coordination patterns per domain arising from tertiary packing of secondary-structure segments, rather than sequence-specific constraints.

Keywords

protein folding; structure prediction; chain skeleton; statistical scores; domain boundary

Introduction

Protein structures are usually formed hierarchically,^{1,2} and the smallest folding and functional units are called structural domains.³ Experimentally, structure determination by techniques such as X-ray crystallography or NMR spectroscopy is much more successful when solving

*Corresponding author. One Baylor Plaza, BCM-125, Baylor College of Medicine, Houston, TX 77030, USA. jpma@bcm.tmc.edu.

[†]Y.W., A.D.D., and M.C. made equal contributions to this work.

Edited by M. Levitt

Supplementary Data Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2008.10.093

Publisher's Disclaimer: This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues. Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited. In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit: <http://www.elsevier.com/copyright>

smaller domain units.⁴ Computationally, structure prediction methods are essentially restricted to small single-domain proteins.⁵ Thus, the identification of protein domain boundaries from an amino acid sequence is of practical interest for both experimental and computational structure determination.

For a known protein structure, it is often nontrivial to define the domains,⁶ and it is even more difficult to determine domain boundaries if only the amino acid sequence is available. Several computational methods have been developed over the last few years to tackle this problem. They include algorithms based on domain size distribution,⁷ sequence alignments,⁸⁻¹⁰ artificial neural networks and machine learning,¹¹⁻¹⁷ information theory,^{18,19} and statistical potentials.^{20,21} Other methods such as SnapDRAGON²² and RosettaDOM²³ assign domains by analyzing an ensemble of structures predicted by *ab initio* methods such as DRAGON²⁴ and Rosetta.²⁵ Finally, meta-domain prediction methods choose domain boundaries by combining different domain prediction approaches.²⁶⁻²⁸

In this study, we developed a procedure called OPUS-Dom for predicting the domain boundaries of a protein from its amino acid sequence. At the heart of the method is VECFOLD, a novel coarse-grained folding method that models a protein structure as a chain of three-dimensional vectors representing the predicted secondary-structure elements (SSEs). VECFOLD generates a large number of folded structures, which are then analyzed individually by a protein domain parsing method, and the results are accumulated into a Z-score profile. The domain boundaries are then determined based on the Z-score profile and three empirical filters.

For a systematic evaluation of the predictive power of OPUS-Dom, three multidomain protein databases were used as benchmark sets. Additionally, targets from the sixth and seventh Critical Assessment of Techniques for Protein Structure Prediction experiments (CASP6 and CASP7) were also tested. We found that OPUS-Dom generally performs better than all previous methods. For instance, we obtained an overall sensitivity of 55% for CASP6 multidomain targets, which is substantially better than other published results,²³ and 51% for CASP7 targets.

Based on our benchmark results, the consistent and robust behavior of our method suggests that it can locate potential domain boundaries with high confidence. Furthermore, the success of OPUS-Dom provides new insight into the fundamental principles of domain formation.

Results

Overall procedure for VECFOLD-based folding and domain boundary determination

Figure 1 is a flowchart of the overall procedure for OPUS-Dom. Given a query sequence, the SSEs are first predicted by PSI-PRED.²⁹ The SSEs are then grouped into three-SSE windows, each representing a super secondary structure motif (SSSM). The sequence of each target SSSM is then aligned with SSSMs in a structure template library to extract possible structures of the target SSSM. For each SSSM window on the query sequence, the vector coordinates and sequence profiles of the 10 most likely structure candidates are saved after sequence-based alignment.

Guided by a geometric scoring function that describes the packing preference of SSEs, the vector model is folded from an initially extended chain into a compact tertiary structure by simulated annealing Monte Carlo (MC) simulation. In each MC move, an SSSM conformation is replaced by one of the 10 saved structure candidates extracted from the template library. At the end of VECFOLD, a C α trace is constructed from the vector model in order to carry out further domain boundary analysis.

For each query sequence, 10^4 compact structural models are generated using VECFOLD, and then the domain boundaries of each model are analyzed using a structure-based domain parsing algorithm, e.g., DOMID³⁰ or the algorithm of Taylor.³¹ Along the sequence, a frequency profile is constructed from the residue-specific domain boundaries identified by the domain parsing algorithm. This profile is then smoothed and converted into a Z-score profile. The residues corresponding to Z-scores above 1.0 are selected as potential candidates for the domain boundaries. Three additional sequence-based profiles are then applied to filter the domain boundary candidates and improve the specificity of OPUS-Dom.

Examples of VECFOLD results

To illustrate the results of VECFOLD, several target proteins are chosen as examples and final models for these proteins are selected from 10^4 independently converged trajectories. The structures of these models together with their corresponding native structures are shown in Fig. 2.

Figure 2a is the C-MYB DNA-binding domain [Protein Data Bank (PDB) code 1MSE]. A cartoon of the native structure is shown on the far right and consists of two domains according to its Molecular Modeling Database (MMDB) definition.³² The N-terminus domain is in black, while the C-terminus domain is in white. The leftmost two structure models are generated by VECFOLD using the same color index as for the native domain assignment. Despite having different tertiary topologies, the VECFOLD models share similar domain features. Such models are generated frequently and repeatedly by VECFOLD.

The copper chaperone for superoxide dismutase (PDB code 1QUP) and human heart short-chain α -3-hydroxyacyl CoA dehydrogenase (PDB code 2HDH) are two examples with relatively larger sizes than 1MSE. Their native structures and VECFOLD models with similar domain arrangements are shown in Fig. 2b and c. All these examples (1MSE, 1QUP, 2HDH) indicate that VECFOLD is capable of obtaining conformations that are similar to the native protein at the domain level. Such a finding inspired us to apply statistical analysis to VECFOLD-based folding results to identify domain boundary regions.

Examples of domain boundary prediction

We follow the criteria of Dumontier *et al.*²⁰ to assess the accuracy of domain boundary prediction. A predicted domain boundary is regarded as correct when the predicted boundary fully or partially overlaps the true (official) boundary within a margin of tolerance of ± 20 residues. In addition, the 20 residues nearest the N and C termini are excluded from domain boundary prediction. The true boundaries are assigned from the MMDB³² or the official CASP definitions.³³ In order to generate results comparable to those of previous methods for CASP6 and CASP7 test proteins,²³ the ± 20 residue margin of tolerance is replaced by a ± 10 residue margin for CASP test sets.

Several examples help to illustrate the domain boundary prediction results. The first case is 1QUP, whose VECFOLD results are discussed in the previous section and whose native structure is given in Fig. 2b. The 1QUP prediction results are shown in Fig. 3a. The black line-dot profile in the lower part of the figure is the Z-score profile of the statistical distribution of structure-based domain assignments by DOMID on models generated by VECFOLD. The continuous, dashed, and dotted lines in the upper part of the figure represent the three additional filters. The horizontal continuous line at Z-score=1 and the horizontal dashed-dotted line at Z-score=-2 are two cutoffs used in this study. Combining the results of DOMID and the additional filters yielded a final predicted domain boundary at residues 73–74 (marked by the continuous arrow). The shaded bar below the residue index gives the MMDB domain

definition. Comparing the MMDB definition to our results shows that OPUS-Dom successfully predicted the domain boundary for this protein.

The second example is the Ap-2 clathrin adaptor alpha-appendage (PDB code 1QTS), for which the domain boundary prediction results are shown in Fig. 3b. Our method predicted two domain boundaries for this protein. The predicted boundary region of residues 141–155 (peak at residue 150, marked by a continuous arrow) falls within 20 residues of the true domain boundary spanning residues 130 to 131. The other predicted boundary region at residues 72–84 (peak at residue 79, marked by dashed arrow) is outside the margin of tolerance of the true boundary.

The next example, bacterial chitobiase (PDB code 1QBA), is a much larger protein with a very complicated domain assignment (Fig. 3c). This protein consists of 858 residues distributed over four domains, with the second domain straddling a long discontinuous region as indicated by the shaded bar. For comparison with previous works, we adopted the domain boundary definition provided in Miyazaki *et al.*,¹² where the linker regions span residues 173–222, 335–339, and 780–787. Our method predicted five domain boundaries, three of which match the predefined domain linker regions. The successful identification of all three predefined linker regions suggests that our method is a sensitive and powerful tool for helping experimentalists determine possible domain regions of large proteins.

The fourth example is the CASP7 prediction target T0356 (PDB code 2IDB), one of the largest and most difficult targets in the competition. This target consists of 505 residues divided into three domains as illustrated by the shaded bar at the bottom of Fig. 3d. The first and third domains of this protein belong to the CASP “free-modeling” category, which is composed of structures with no homology information. Although the second domain is a template-based modeling target, few groups in the competition found the correct structure analogs. By combining the results from DOMID and the three filters, our method predicted two boundaries (marked with continuous arrows) that are within 20 residues of the official boundaries. The official domain assignment of T0356 is indicated by different shades of gray on the native structure in Fig. 4a, and our predicted domains are shaded with the same grayscale index on the native structure in Fig. 4b. A comparison of these two figures yields only slight differences near the domain boundaries.

The fifth example is *Enterococci faecium* D-alanyl-D-lactate ligase (PDB code 1E4E), for which the domain boundary prediction results are shown in Fig. 3e. In this case, OPUS-Dom fails to recognize any domain boundaries. This error is partly due to the two key interdependent weaknesses of OPUS-Dom: (i) the coarseness of the loop-building method in VECFOLD, and (ii) the accuracy of secondary-structure prediction from sequence, e.g., by PSI-PRED. In this example, PSI-PRED predicts a large loop region between residues 52 and 74 (with a small helix turn at residues 56–58). As a result, many of the structure decoys generated by VECFOLD exhibit a dumbbell shape with residues 52–74 serving as the linker. The domain parser interprets this gap as a domain boundary, and the resulting domain labels at residues 52–74 dominate the structure-based Z-score profile, bringing the other peaks below the Z-score cutoff of 1.0. Note that the domain boundary profile library (DBPL) filter finds the boundaries at residues 119, 146, and 210, and thus, elevating the structure-based Z-score profile would have resulted in two correct boundary predictions at residues 153 and 208.

The sixth and final example is the 903-residue, six-domain DNA polymerase of bacteriophage RB69 (PDB code 1IH7), for which the Z-score profiles are plotted in Fig. 3f. Here, the specificity is 100%, but the sensitivity is only 33%. This example illustrates the sensitivity–specificity trade-off arising from our choice of Z-score cutoffs; slightly less strict cutoffs would enhance sensitivity at the cost of decreased specificity.

Benchmark evaluations

For the purpose of systematically evaluating the predictive power of our method, three multidomain protein databases and two sets of CASP domain targets are used as benchmark sets. The first three testing sets were generated for performance evaluation in previous works and could therefore be used for comparative study, and the latter two were created for the CASP6 and CASP7 competitions. We calculated the sensitivity and specificity of our method for all the benchmark testing sets as described in Methods. All specific examples mentioned in the previous sections are drawn from these benchmark testing sets.

For the first three data sets, we compared OPUS-Dom to five other methods: (i) the domain linker propensity index (DLI) of Dumontier *et al.*,²⁰ (ii) the residue entropy index (REI) method by Galzitskaya and Melnik,³⁴ (iii) the residue index method by George and Heringa²² derived from amino acid propensity of all linkers (GHL); (iv) the hydropathy index method by Kyte and Doolittle³⁵ derived from the propensity of residues to be in the hydrophobic core (KDH); and (v) the DLI+REI consensus domain predictor Armadillo (ARM) by Dumontier *et al.*²⁰ In addition to measuring the performance of OPUS-Dom for the entire set, we also analyzed sensitivity and specificity *versus* protein chain length (Fig. 5).

The first data set, called MMDB, is composed of 211 proteins, each with two contiguous domains and one linker and an average length of 283 residues.²⁰ The native domain assignments for these proteins follow the MMDB definitions. For this benchmark, our method achieved 50% sensitivity and 39% specificity, both of which are mostly better than previously reported results as shown in Table 1. The performance of OPUS-Dom is uncorrelated with protein chain length, with linear regression *R*-squared values of 0.063 and 0.127 for sensitivity and specificity, respectively (Fig. 5a).

The second benchmark set used in our study is the Galzitskaya and Melnik (GM) representative data set comprising 29 structures from 44 Structural Classification of Proteins (SCOP) superfamilies.³⁴ The GM data set is one of the most challenging benchmarks because it includes some very large proteins and it also contains multidomain proteins with multiple domain linkers. Our method achieved 42% sensitivity and 59% specificity. As before, the sensitivity and specificity are uncorrelated with chain length (Fig. 5b). The PDB codes and individual predictions for all 29 proteins are given in Table 2.

The third benchmark set, originally presented by Miyazaki *et al.*,¹² consists of 74 protein structures with 99 SCOP-derived domain linkers. Compared with previous methods, our method achieved comparable specificity and sensitivity as shown in Table 1. As before, the sensitivity and specificity are uncorrelated with chain length (Fig. 5c).

We also used OPUS-Dom to predict the domain boundaries for CASP6 and CASP7 protein targets. To provide comparable results with previous methods, a 10-residue margin of tolerance was used instead of the typical 20-residue margin. For all multidomain protein targets in CASP6 whose domain boundaries were officially defined by the CASP assessor³³ (the set will be called CASP6-Multi), we achieved almost twice the sensitivity of one of the most successful domain prediction methods, Rosetta-DOM²³ (the results are shown in Table 3).

In addition, OPUS-Dom was applied to a set of 95 CASP7 targets with domain definitions, split into single-domain (68 targets), two-domain (25 targets), and three-domain (2 targets) subsets. For single-domain proteins, we counted the predictions as correct when OPUS-Dom did not identify any domain boundaries. As shown in Table 4, OPUS-Dom achieved sensitivities of 54%, 47%, and 40% on the single-, double-, and triple-domain sets, respectively. Overall sensitivity and specificity were 51% and 50%, respectively, for the full 95-target set. We also tested OPUS-Dom on a subset of multidomain proteins that lack any significant

sequence homology information. This subset, which we call CASP7-Multi, is a useful benchmark for *de novo* methods such as OPUS-Dom, and it consists of 14 proteins with more than 150 residues and for which PSI-BLAST³⁶ and the fold recognition method FFAS03³⁷ cannot find any proper templates. For these 14 targets, OPUS-Dom achieved 44% sensitivity and 47% specificity.

In summary, the performance of OPUS-Dom on several different benchmark testing sets suggests that the method can locate potential domain boundaries with relatively high confidence.

Other domain prediction methods not included in benchmark

Several other domain prediction methods are discussed below, but due to differences in testing sets or metrics, we did not make a direct performance comparison with OPUS-Dom. SnapDRAGON²² is a folding-based prediction method like OPUS-Dom and Rosetta-DOM. SnapDRAGON uses DRAGON to generate a large number of atomic structure models, and then it assigns domain boundaries using Taylor's method. SnapDRAGON achieved 42% sensitivity and 40% specificity on a custom multidomain testing set. DomNet¹⁶ is a very recent machine-learning algorithm that uses a new compact domain profile and outperforms many other machine-learning methods on the Benchmark_2⁶ and CASP7 data sets. CHOPnet¹⁴ is a neural-network-based method that takes into account several parameters such as amino acid composition and flexibility, predicted secondary structure, and solvent accessibility. CHOPnet achieved 51% sensitivity on a custom two-domain testing set.

Comparing domain prediction results using Taylor's method with DOMID

There is no standard domain definition for a given protein structure, and different structure-based domain parsing algorithms typically give different results. We used DOMID to assign domain boundaries for the VECFOLD-generated structure models in our initial study. For comparison, we also predicted the domain boundaries using the structure-based domain parsing method developed by Taylor.³⁸ In most cases (e.g., Fig. 6a and b), the Z-score profiles from DOMID and Taylor's method had similar shapes and were highly correlated, with only slight differences in peak shift and relative peak height (Fig. 6c is an example of less correlated results). Furthermore, the specificities and sensitivities on all benchmark testing sets (Table 5) are comparable, with Taylor's method performing better on the GM, Miyazaki, and MMDB benchmark sets, and DOMID performing better on the CASP6 and CASP7 sets. The success of the OPUS-Dom method using DOMID or Taylor's algorithm suggests that OPUS-Dom is mostly insensitive to the choice of domain parser, and it underscores the power of VECFOLD to generate structures that are topologically correct. Even so, a future direction for our work is to combine different structure-based domain parsing methods to give better results by consensus.

Using sequence-based filters to enhance structure-based domain prediction

Three sequence-based filters were added to OPUS-Dom to improve the specificity of the structure-based domain prediction algorithm. Table 6 shows that the specificity improves by as much as 18 percentage points for the five benchmark test sets (MMDB, GM, Miyazaki, CASP6-Multi, and CASP7-Multi) when the sequence-based filters are applied, which comes at the cost of decreased sensitivity. This specificity-sensitivity trade-off must be considered when fine-tuning OPUS-Dom for different applications, i.e., the structure-based method alone may be sufficient for cases in which sensitivity is most important.

Discussion

Here, we attempt to address both the challenges of domain prediction and assessment. Based on our encouraging benchmark results on several different testing sets, we believe that our method is a potentially useful tool to help experimentalists determine possible domain regions for large proteins. Our method generally outperformed the other methods in our benchmark comparison (DLI, REI, GHL, KDH, Armadillo, Rosetta-DOM) using the MMDB, GM, Miyazaki, and CASP6 sets, as indicated in Tables 1 and 3. Furthermore, OPUS-Dom performed especially well relative to the state of the art in predicting domain boundaries of structures with no available sequence homologs. In the CASP7 competition, OPUS-Dom predicted two of three boundaries of very difficult prediction target T0356 (Figs. 3d and 4), for which few teams found the correct structure analogs. In addition, OPUS-Dom obtained the highest score in the contest for the comparably difficult CASP7 target T0301, which was highlighted in Fig. 7d of Tress *et al.*³⁹ (our group ID was DP229).

OPUS-Dom is several times faster than other folding-based prediction methods such as Rosetta-DOM because of our unique coarse-grained, vector-based sampling method VECFOLD. VECFOLD can easily handle targets with more than 600 residues and achieve thousands of trajectories in a short time on a single processor. For example, in our study, VECFOLD generated 10^4 trajectories in roughly 0.016 h per residue (on average for the Miyazaki¹² test set) on a single 900-MHz Intel Itanium2 processor, i.e., VECFOLD can generate 10^4 compact tertiary structure models of a 250-residue protein in roughly 4 h. Because VECFOLD is very efficient in searching protein conformational space, it might sample the protein native topology with far fewer steps than other residue-based or atom-based folding algorithms. Furthermore, as evidenced by the benchmark results, VECFOLD seems to generate structures with global features that are roughly consistent with the native structure.

A distinct advantage of folding-based domain boundary prediction methods is that they rely less on sequence homology information than purely sequence-based methods. In our case, evolutionary information is used only to establish the library of structure fragment candidates, whereas the overall tertiary structure is determined by VECFOLD. Although each folded structure carries relatively large errors, the statistical consensus derived from an ensemble of such structures reveals a much more robust delineation of domain boundaries. This suggests that the domain arrangement in protein structures is less a consequence of sequence-specific constraints. Instead, tertiary structure packing of secondary-structure segments yields a limited set of coordination patterns per domain. An interesting analogy is the short-range ordering in simple liquids.⁴⁰ Although the instantaneous configuration of liquid molecules may be constantly changing, the position of the peak of the first coordination shell, as manifested by the radial distribution function, is very stable for a given set of macroscopic environmental parameters. Such a static pattern of thickness in coordination shells resembles the domain volume arrangement in proteins, i.e., if the vectorial arrangement of secondary-structure segments are given, tertiary packing of a single domain would result in a relatively stable domain volume distribution due to topological constraints.

It is worth mentioning that decomposing protein structures into domains is challenging even for human experts⁴¹ because of the diverse nature of domains. Furthermore, structure-based domain parsing methods such as Domain-Parser,⁴² NCBI,⁴³ PDP,⁴⁴ and PUU⁴⁵ cannot yet match the level of accuracy of human experts.⁶ Thus, the inconsistency of domain assignments makes the unbiased assessment of prediction results even more difficult. For example, in the CASP domain prediction competition, assessors use human inspection and several automatic domain assignment methods to define the domains, often assigning multiple domain definitions to a protein target.³³

Finally, OPUS-Dom successfully defines the domain boundaries, but it is less reliable in specifying the domain identity of a polypeptide segment in between two contiguous domain boundaries. For example, consider a two-domain protein in which the sequence of domain A is continuous, and domain B is formed by two separate sequence fragments (B1 and B2), yielding a sandwich-like B1–A–B2 sequence motif. In this case, it is difficult to specify which fragment belongs to which domain, even though the domain boundaries are determined via statistical analysis. This is a direction for future improvement.

Methods

As summarized in Fig. 1, OPUS-Dom (i) folds 10^4 candidate structures from a query sequence using VECFOLD, (ii) generates a structure-based Z-score profile from the distribution of domain boundaries determined for each candidate structure by a domain parsing method, e.g. DOMID, and (iii) predicts domain boundaries using a consensus of the structure-based profile and three sequence-based domain profiles. In this section, we first describe the vector geometry of VECFOLD, and we introduce the nonredundant structure database used for the geometric scoring function and the SSSM (super secondary structure motif) fragment library described subsequently. Then, we explain the simulated annealing MC scheme used to fold vector models into compact tertiary structures. Next, we summarize two domain parsing methods and the process for generating a structure-based domain boundary profile from the distribution of boundaries assigned by either of these methods. The three sequence-based profiles used to improve the domain boundary prediction are conveyed last.

VECFOLD Method

Vector representation of protein conformation—In the VECFOLD algorithm, a protein is modeled as a highly coarse-grained chain of vectors representing SSEs,⁴⁶ such that no atomic coordinates are needed during folding. For an α -helix, the vector is defined on the primary axis of the helix. The direction is calculated from the center of the first four alpha carbon (C^α) atoms in the helix to the center of the last four C^α atoms in the helix. The length of the helix vector is $(N_\alpha - 1) \times 1.5 \text{ \AA}$, where N_α is the number of residues in the helix. The midpoint of the vector resides on the helix center of mass. Each β -strand is represented by two vectors, an axial vector that connects the two terminal C^α atoms of the strand, and a normal vector describing the orientation of the strand. For a loop region, the vector connects the end points of the two adjacent SSE vectors. With this vector representation, illustrated in Fig. 7, a protein conformation is described by a set of local internal coordinates: vector lengths, packing angles between pairs of adjacent vectors, and dihedral angles for sets of three adjacent vectors.

Nonredundant structure database—The template library and geometric scoring function described below are constructed from a nonredundant structure database, which is generated by PISCES⁴⁷ and consists of 2701 nonhomologous proteins with homologies less than 20% and resolutions better than 1.8 $\text{\AA}$.

Geometric packing scoring function—In order to evaluate the tertiary-structure packing quality of VECFOLD vector models, a scoring function based on the statistical distribution of different types of long-range SSE packing is derived from the nonredundant structural database. The geometric scoring function E_{Geometry} consists of four terms:

$$E_{\text{Geometry}} = E_{\text{HH}} + E_{\text{SS}} + E_{\text{SH}} + E_{R_g} \quad (1)$$

where E_{HH} , E_{SS} , and E_{SH} are helix–helix, strand–strand, and strand–helix packing terms, respectively, and E_{R_g} is a radius of gyration term.

For the helix–helix packing term E_{HH} , two helices are considered in contact if there is at least one pair of side-chain atoms, one atom on each helix, within the 5 Å cutoff distance. E_{HH} consists of three terms representing the packing distance d_{ij} , the packing angle φ_{ij} , and the coupling effect of helix lengths L_i and L_j :

$$E_{HH}(d_{ij}, \varphi_{ij}, L_i, L_j) = -RT \ln \frac{N_{bin}^d N_{obs}(d_{ij})}{\sum_{d_{kl}} N_{obs}(d_{kl})} - RT \ln \frac{N_{bin}^\varphi N_{obs}(\varphi_{ij})}{\sum_{\varphi_{kl}} N_{obs}(\varphi_{kl})} - RT \ln \frac{N_{bin}^{d,\varphi,L} N_{obs}(d_{ij}, \varphi_{ij}, L_i, L_j)}{\sum_{d_{kl}, \varphi_{kl}, L_k, L_l} N_{obs}(d_{kl}, \varphi_{kl}, L_k, L_l)} \quad (2)$$

Here, N_{obs} is the number of observed occurrences of a specific coarse-grained parameter (e.g., d_{ij}) in the nonredundant structural database, and $N_{bin}^{(d,\varphi,L)}$ is the number of bins for a set or subset of parameters $\{d, \varphi, L\}$ (the bin number and ranges for each parameter are defined in Supplemental Materials). Figure 8a illustrates the definitions of d_{ij} , φ_{ij} , L_i , and L_j . For helices H_i and H_j defined by helix vectors $\mathbf{v}_i$ and $\mathbf{v}_j$, d_{ij} equals the transversal distance (i.e., the minimal distance between the two vectors),⁴⁸ φ_{ij} is the dihedral angle between $\mathbf{v}_i$ and $\mathbf{v}_j$, and $L_i = \|\mathbf{v}_i\|$ and $L_j = \|\mathbf{v}_j\|$ ($\|\cdot\|$ is the Euclidian distance or L2-norm) are the lengths of the helices. Figure 9a and b show the packing score as a function of helix–helix packing angle or distance, and the minima in the packing angle curve (Fig. 9a) correspond to the orthogonal and up–down packing preferences.

For the strand–strand packing term E_{SS} in Eq. (1), two strands are considered in contact if there is at least one hydrogen bond between the backbones of these two strands. Like E_{HH} , E_{SS} includes three terms representing the packing (transversal) distance d_{ij} , packing angle φ_{ij} , and the coupling effect of normal angles β_i and β_j :

$$E_{SS}(d_{ij}, \varphi_{ij}, \beta_i, \beta_j) = -RT \ln \frac{N_{bin}^d N_{obs}(d_{ij})}{\sum_{d_{kl}} N_{obs}(d_{kl})} - RT \ln \frac{N_{bin}^\varphi N_{obs}(\varphi_{ij})}{\sum_{\varphi_{kl}} N_{obs}(\varphi_{kl})} - RT \ln \frac{N_{bin}^{d,\varphi,\beta} N_{obs}(d_{ij}, \varphi_{ij}, \beta_i, \beta_j)}{\sum_{d_{kl}, \varphi_{kl}, \beta_k, \beta_l} N_{obs}(d_{kl}, \varphi_{kl}, \beta_k, \beta_l)} \quad (3)$$

Figure 8b illustrates the definitions of d_{ij} , φ_{ij} , β_i , and β_j . In addition to the direction vectors $\mathbf{v}_i$ and $\mathbf{v}_j$, strands S_i and S_j require normal vectors $\mathbf{n}_i$ and $\mathbf{n}_j$ (perpendicular to the sheet plane) to define their orientations. Figure 9c and d show the packing score as a function of strand–strand packing angle or distance, and the minima in the packing angle curve (Fig. 9c) correspond to the parallel and antiparallel strand–strand packing preferences.

For the strand–helix packing term E_{SH} , the helix and strand are in contact if there is at least one pair of side-chain atoms, one atom on the helix and the other on the sheet region, within the 5 Å distance cutoff. Strand–helix packing is illustrated in Fig. 8c, where $d_{ij} = \|\mathbf{c}_i - \mathbf{p}_j\|^2$ is the distance between the center of mass $\mathbf{c}_i$ of helix H_i and its projection $\mathbf{p}_j$ onto the plane of sheet S_j , θ_{ij} is the angle between the helix vector $\mathbf{v}_i$ and the vector represented by d_{ij} (connecting $\mathbf{c}_i$ and $\mathbf{p}_j$), $c_{ij} = \|\mathbf{c}_j - \mathbf{p}_j\|^2$ is the distance between the center of mass $\mathbf{c}_j$ of the sheet plane and $\mathbf{p}_j$, and φ_{ij} is the dihedral angle between consensus sheet vector $\mathbf{v}_j$ and $\mathbf{v}_i$ (i.e., the angle between $\mathbf{v}_j$ and the vector projection of $\mathbf{v}_i$ in the sheet plane). The strand–helix packing energy also includes three terms corresponding to θ_{ij} , φ_{ij} , and the coupling effect of distances c_{ij} and d_{ij} :

$$E_{SH}(\theta_{ij}, \varphi_{ij}, c_{ij}, d_{ij}) = -RT \ln \frac{N_{bin}^{\theta} N_{obs}(\theta_{ij})}{\sum_{\theta_{kl}} N_{obs}(\theta_{kl})} - RT \ln \frac{N_{bin}^{\varphi} N_{obs}(\varphi_{ij})}{\sum_{\varphi_{kl}} N_{obs}(\varphi_{kl})} - RT \ln \frac{N_{bin}^{c,d} N_{obs}(c_{ij}, d_{ij})}{\sum_{c_{kl}, d_{kl}} N_{obs}(c_{kl}, d_{kl})} \quad (4)$$

Figure 9e and f show the packing score as a function of angle φ_{ij} or θ_{ij} .

In our study, we use a harmonic scoring function based on the radius of gyration R_g to promote compactness in the global shape of the VECFOLD models:

$$E_{R_g} = \frac{\varepsilon}{2} \left(\frac{R_{g,Calculated} - R_{g,Empirical}}{\sigma} \right)^2 \quad (5)$$

where weight parameter ε and standard deviation σ are set to 10.0 and 5.0 Å, respectively, based on the statistical data, $R_{g,Calculated}$ is the calculated radius of gyration, and $R_{g,Empirical}$ is the radius of gyration of an N -residue protein that obeys the empirical expression by Bahar and Jernigan:⁴⁹

$$\log(R_{g,Empirical})^2 = \frac{2}{3} \log N + 0.92. \quad (6)$$

For a series of consecutive vectors packed in space, the radius of gyration may be calculated as:

$$R_{g,Calculated} = \sqrt{\frac{\sum_i n_i \left(\frac{v_i^2}{12 + \|\mathbf{R}_{MC} - \mathbf{r}_{m,i}\|^2} \right)}{N}} \quad (7)$$

where N is the number of residues in the query protein, i is the index over all vectors $\{\mathbf{v}_i\}$, n_i is the number of residues in $\mathbf{v}_i$, $\mathbf{R}_{MC}$ is the protein center of mass, and $\mathbf{r}_{m,i}$ is the midpoint of $\mathbf{v}_i$. We choose 100 test proteins to validate the accuracy of vector-based $R_{g,Calculated}$ (the list of our test set is in Supplemental Materials). These test proteins are well packed and lack long disordered regions. We calculate the radius of gyration for these proteins using both the vector approach and the traditional atomic method, and we plot these values in Fig. 10. The strong correlation between these two methods (correlation coefficient 0.936) suggests that our vector-based method approximates the traditional method well.

Generating the SSSM move set by aligning the query sequence to the fragment library

An SSSM is a mesoscale protein fragment consisting of three consecutive SSEs and their two connecting loops. VECFOLD folds protein tertiary structures by drawing fragments as MC moves from an ensemble of SSSM candidates.

Prior to executing VECFOLD, an SSSM template library is constructed from the nonredundant structure database. Each structure is divided into overlapping SSSM windows based on the SSEs identified in the RCSB Protein Data Bank records, and an SSE vector model is built to match the structure. In parallel, a position-specific sequence profile is generated from the query

sequence by three iterations of PSI-BLAST with an E value below 0.001 (the profile is normalized to unity at each residue position). The sequence and secondary-structure profiles and SSSM vector coordinates (vector lengths, packing angles, and dihedrals) are recorded in the template library for each SSSM.

To generate an ensemble of SSSM candidates, VECFOLD first generates a secondary-structure profile from the query sequence using PSI-PRED,²⁹ and then it separates the resulting P SSEs into $P - 2$ SSSM windows. Next, as for the template library, a PSSM-based sequence profile is generated from the query sequence by three iterations of PSI-BLAST with an E value below 0.001 (the profile is again normalized to unity at each residue position). The structural candidates for each query SSSM window are generated by aligning the sequence profile and secondary-structure information to all SSSMs in the template library. We use Smith–Waterman local affine-gap dynamic programming to align the profiles.⁵⁰ The gap opening and gap extension parameters are -4.5 and -0.5 , respectively.

For query sequence position i and template position j , the alignment score consists of two terms, the first related to sequence profiles and the second related to secondary structures:

$$S(i, j) = -w_{\text{seq}} \sum_{k=1}^{20} F_{\text{query}}(i, k) M_{\text{temp}}(j, k) w_{2\text{nd}} \delta(S_{\text{query}}(i), S_{\text{temp}}(j)), \quad (8)$$

where F_{query} is the $N \times 20$ frequency profile of the query sequence, M_{temp} is the $N \times 20$ log-odds profile of the template sequence, and $s_{\text{query}}(i), s_{\text{temp}}(j)$ are the SSE types for query position i and template position j , respectively. The value of $\delta(\cdot)$ is 1 if the SSE types for positions i and j match, i.e., $s_{\text{query}}(i) = s_{\text{temp}}(j)$, and -1 otherwise. Both weights w_{seq} and $w_{2\text{nd}}$ are set to unity. Note that when the sequence homology is low, the secondary-structure profile dominates the alignment score.

For each query SSSM window, the 10 template SSSMs with highest alignment scores are selected as structure candidates. Note that among these top 10 candidates, the SSE content may vary, as the secondary-structure information is but one of three terms in the alignment. Moreover, the number of residues per SSE may vary, as Smith–Waterman includes the processes of residue insertion and deletion. Therefore, the residue lengths of the target SSSM window and the chosen template SSSM may differ. However, the number of SSEs per SSSM is always three.

Generating tertiary structures by simulated annealing

Armed with our geometric packing scoring function and the ensembles of SSSM candidates, we apply MC simulated annealing to fold the query protein into a compact tertiary structure in vector form. We begin with the vector chain in its fully extended conformation, such that each vector packing angle is 120° and each dihedral is 180° . Then the chain conformation is adjusted using SSSM-based fragment assembly. In each simulation step, an SSSM region in the query sequence is randomly chosen, and the conformation of that query SSSM is replaced by the conformation of a template structure candidate selected randomly from the top 10 candidates for that SSSM. The geometric packing score is evaluated at each step and the SSSM substitution is accepted or rejected by the Metropolis criterion. The effective temperature factor T is adjusted linearly from 20 to 0.5 over the course of 5000P steps, where P is the number of SSEs in the query sequence.

At the end of each individual run, a C^α trace is reconstructed from the vector chain. First, ideal C^α traces of helices and strands are superimposed onto the target vectors as rigid bodies, and

then the C $^{\alpha}$ atoms of the loop regions are built in a crude but very fast way. The C $^{\alpha}$ atoms in a loop region are assigned to an isosceles triangle on the plane defined by two vectors, (i) the vector representing the loop connecting the previous and next SSE, and (ii) the resultant of the normalized vectors representing the previous and next SSE. The loop C $^{\alpha}$ atoms are spaced 3.8 Å apart (the pseudo-bond length) and bridged by 3.8 Å to the C $^{\alpha}$ atoms in adjoining SSEs. Higher resolution loop modeling may be realized using more accurate local move sets and high-resolution energy functions, but this is not the purpose of our current study, as we are interested in *rapidly* constructing topologically correct structure models for domain parsing.

Domain boundary prediction

Structure-based domain assignment by DOMID and Taylor's method—For each structure model generated by VECFOLD, we assign domain boundaries by the structure-based method DOMID. For this study, we applied, in parallel, the domain parsing method of Taylor³⁸ to test the sensitivity of our overall VECFOLD-based domain prediction method to different domain assignment algorithms.

In DOMID, scores for the likelihood of rigid-body movement for all the possible domain divisions are derived from the interaction energy of pairs of residues in the target structure. The hypothesis is that the contrast of intradomain and interdomain interaction energies is maximized for correct domain partitions, i.e., pairwise residue interactions are strong within the domain and weak between domains. DOMID then assigns domain boundaries to the structural subsets that have considerable tendency for domain movement.

In Taylor's algorithm, each residue in a protein chain is assigned a numerical label (simply, its sequential residue number), and this label changes according to a consensus value of the labels of the residue's physical neighbors. Thus, the labels of nearby residues tend to consolidate, and label boundaries begin to form in regions where the residue density is low. Taylor's method assigns domains to residues that share one of the reduced set of numerical labels.

Domain boundary Z-score profile—Once the domain boundaries for all 10⁴ VECFOLD folded models are obtained by DOMID or Taylor's method, they are accumulated to give a distribution of domain boundaries along the query sequence. This distribution is smoothed using a triangular sliding window²² in order to combine nearby peaks. The window slides over the distribution one residue at a time and sums the weighted values within the window, such that the values at more central positions in the window are weighted more than values toward either of the ends. The specific expression for the smoothed score is the convolution sum:

$$S(k) = \sum_{i=1}^n \bar{S}\left(k - \frac{n+1}{2} + i\right) \times W_n(n-i+1) \quad (9)$$

where $S(k)$ is the smoothed score at position k , $\bar{S}$ is the raw score before smoothing, and $W_n(i)$ is the window weight function:

$$W_n(i) = \frac{(p - |p - i|)}{p^2}; \quad p = \frac{n+1}{2}. \quad (10)$$

Here, n is the size of the sliding window, which is set to 5% of the whole sequence length and rounded up to the nearest odd integer. This definition ensures that the sum of the weights is normalized to unity and the weights are biased to the center of the window.

The smoothed domain boundary distribution is then converted into a Z-score profile. The residue positions of all peaks above a Z-score of 1.0 are regarded as potential domain boundary regions. This Z-score threshold is chosen based on empirical data (not shown).

Sequence-based filters—For the potential domain boundaries identified by DOMID or Taylor's method based on VECFOLD models, three sequence-based filters are applied to improve the results: (i) a residue entropy index filter,²⁰ which is based on the hypothesis that the domain boundary is dominated by residues with low side-chain entropy and hence small side-chain size; (ii) a domain linker index filter,³⁴ which is derived from a structure database and is based on the relative frequency of residues found in linker regions compared to compact domains; and (iii) a domain boundary profile library (DBPL) that we developed to extract domain boundary profile information from a nonredundant structure database. All the filtered profiles are in the form of Z-score profiles, and minima are preferred since the filters are energy-like.

We construct the DBPL by recording the sequence profile information near domain linkage regions of all multidomain proteins in the nonredundant structure database used to construct the SSSM template library. The profile includes the PSSM and secondary-structure information. The DBPL filter score for position i in the query sequence is generated by using an 11-residue window and aligning the sequence profile of the window to all the domain linkage profiles in the DBPL. The filtering profile for the whole query sequence is then generated after recording the lowest alignment score for position i and sliding the window through the whole sequence. The alignment score profile along the sequence is converted into a Z-score profile.

We choose the domain boundary candidates for which at least one filtered Z-score profile falls below -2.0 within ± 5 residues of the candidate boundary position (as determined by the structure-based domain parser, e.g. DOMID). Note that domain boundaries are not assigned to residues that are within 30 residues of the N- or C-termini; furthermore, if two boundaries are less than 30 residues apart, we merge them into a single linkage region. In addition, for cases in which no domain boundary candidate remains after filtering, we consider the target to be a single-domain protein, i.e., having no domain boundaries.

Assessment of domain boundary prediction—For a systematic test of our prediction algorithm, we use the assessment method of Dumontier *et al.*²⁰ First, our predicted boundaries consist of whole linker regions rather than single residues in accordance with the literature.²⁰ As stated earlier in Methods, these linker regions fall within five residues of the sequence-based filter predictions and are composed of all residues for which the structure-based Z-score is above the 1.0 cutoff.

We calculate two measures of success, prediction sensitivity and specificity, with respect to the benchmark tests. Prediction sensitivity is the percentage of accurately predicted boundaries out of all official boundaries, $\text{Sens} = \text{TP} / (\text{TP} + \text{FN})$, where TP is the number of true positive boundary predictions and FN is the number of false negatives, i.e., real boundaries that our method failed to predict. To be consistent with the domain prediction literature, we define specificity per Kim *et al.*²³ and Dumontier *et al.*²⁰ as the percentage of accurate domain boundary predictions out of all predictions, $\text{Spec} = \text{TP} / (\text{TP} + \text{FP})$, where FP is the number of false positives, i.e., predictions that fail to fall near a real boundary. If two or more predictions are within the same range of an official boundary, we keep the prediction with the better Z-score in order to prevent biasing the results toward an improved sensitivity and specificity.

When presented with a single-domain target, we increment the true positive count if OPUS-Dom does not predict a domain boundary, and otherwise we increment the false negative count. In other words, we can define $\text{TP}' = \text{TP} + \text{TP}_0$ and $\text{FN}' = \text{FN} + \text{FN}_0$, where TP_0 and FN_0 are the

true positive and false negative counts for all single-domain protein targets, respectively, such that

$$\begin{aligned}\text{Sens} &= \text{TP}' / (\text{TP}' + \text{FN}') \\ \text{Spec} &= \text{TP}' / (\text{TP}' + \text{FP})\end{aligned}\tag{11}$$

Future Improvements

The predictive power of OPUS-Dom is dependent on the accuracy of secondary-structure predictions. Thus, consensus secondary-structure predictions will enable VECFOLD to generate more accurate structure models. Another future direction for our work is to combine different structure-based domain parsing methods such as DOMID and Taylor's method to generate a more sensitive Z-score profile of the domain boundary distribution. A third direction is to improve the accuracy of the C^α trace by implementing a more sophisticated loop-building method.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

J.M. acknowledges support from the National Institutes of Health (R01-GM067801), the National Science Foundation (MCB-0818353), and the Welch Foundation (Q-1512). Y.W. is partially supported by a grant from the Doer Foundation. M.C. is partially supported by a predoctoral fellowship from the Keck Center for Interdisciplinary Bioscience Training of the Gulf Coast Consortia. A.D.D. is supported by a predoctoral fellowship from the Keck Center through the National Library of Medicine Computational Biology and Medicine Training Program (NLM Grant No. 5T15LM07093). Computer resources were provided by the Rice Terascale Cluster funded by the NSF (Grant No. EIA-0216467), Intel, and HP.

References

1. Rose GD. Hierarchic organization of domains in globular proteins. *J Mol Biol* 1979;134:447–470. [PubMed: 537072]
2. Panchenko AR, Luthey-Schulten Z, Wolynes PG. Foldons, protein structural modules, and exons. *Proc Natl Acad Sci USA* 1996;93:2008–2013. [PubMed: 8700876]
3. Lesk AM, Rose GD. Folding units in globular proteins. *Proc Natl Acad Sci USA* 1981;78:4304–4308. [PubMed: 6945585]
4. Radhakrishnan I, Perez-Alvarado GC, Parker D, Dyson HJ, Montminy MR, Wright PE. Structural analyses of CREB–CBP transcriptional activator–coactivator complexes by NMR spectroscopy: implications for mapping the boundaries of structural domains. *J Mol Biol* 1999;287:859–865. [PubMed: 10222196]
5. Ginalski K, Grishin NV, Godzik A, Rychlewski L. Practical lessons from protein structure prediction. *Nucleic Acids Res* 2005;33:1874–1891. [PubMed: 15805122]
6. Holland TA, Veretnik S, Shindyalov IN, Bourne PE. Partitioning protein structures into domains: why is it so difficult? *J Mol Biol* 2006;361:562–590. [PubMed: 16863650]
7. Wheelan SJ, Marchler-Bauer A, Bryant SH. Domain size distributions can predict domain boundaries. *Bioinformatics* 2000;16:613–618. [PubMed: 11038331]
8. Suyama M, Ohara O. DomCut: prediction of inter-domain linker regions in amino acid sequences. *Bioinformatics* 2003;19:673–674. [PubMed: 12651735]
9. Marsden RL, McGuffin LJ, Jones DT. Rapid protein domain assignment from amino acid sequence using predicted secondary structure. *Protein Sci* 2002;11:2814–2824. [PubMed: 12441380]
10. Gewehr JE, Zimmer R. SSEP-Domain: protein domain prediction by alignment of secondary structure elements and profiles. *Bioinformatics* 2006;22:181–187. [PubMed: 16267083]

11. Sikder AR, Zomaya AY. Improving the performance of DomainDiscovery of protein domain boundary assignment using inter-domain linker index. *BMC Bioinformatics* 2006;7(Suppl 5):S6. [PubMed: 17254311]
12. Miyazaki S, Kuroda Y, Yokoyama S. Characterization and prediction of linker sequences of multi-domain proteins by a neural network. *J Struct Funct Genomics* 2002;2:37–51. [PubMed: 12836673]
13. Cheng J, Randall AZ, Sweredoski MJ, Baldi P. SCRATCH: a protein structure and structural feature prediction server. *Nucleic Acids Res* 2005;33:W72–W76. [PubMed: 15980571]
14. Liu J, Rost B. Sequence-based prediction of protein domains. *Nucleic Acids Res* 2004;32:3522–3530. [PubMed: 15240828]
15. Sim J, Kim SY, Lee J. PPRODO: prediction of protein domain boundaries using neural networks. *Proteins* 2005;59:627–632. [PubMed: 15789433]
16. Yoo PD, Sikder AR, Taheri J, Zhou BB, Zomaya AY. DomNet: protein domain boundary prediction using enhanced general regression network and new profiles. *IEEE Trans Nanobiosci* 2008;7:172–181.
17. Chen L, Wang W, Ling S, Jia C, Wang F. KemaDom: a Web server for domain prediction using kernel machine with local context. *Nucleic Acids Res* 2006;34:W158–W163. [PubMed: 16844982]
18. Nagarajan N, Yona G. Automatic prediction of protein domains from sequence information using a hybrid learning system. *Bioinformatics* 2004;20:1335–1360. [PubMed: 14962932]
19. Cheng J, Sweredoski MJ, Baldi P. DOMpro: protein domain prediction using profiles, secondary structure, relative solvent accessibility, and recursive neural networks. *Data Min Knowl Discov* 2006;13:1–10.
20. Dumontier M, Yao R, Feldman HJ, Hogue CW. Armadillo: domain boundary prediction by amino acid composition. *J Mol Biol* 2005;350:1061–1073. [PubMed: 15978619]
21. Dovidchenko NV, Yu Lobanov M, Galzitskaya OV. Prediction of number and position of domain boundaries in multi-domain proteins by use of amino acid sequence alone. *Curr Protein Pept Sci* 2007;8:189–195. [PubMed: 17430200]
22. George RA, Heringa J. SnapDRAGON: a method to delineate protein structural domains from sequence data. *J Mol Biol* 2002;316:839–851. [PubMed: 11866536]
23. Kim DE, Chivian D, Malmstrom L, Baker D. Automated prediction of domain boundaries in CASP6 targets using GinzU and RosettaDOM. *Proteins* 2005;61(Suppl 7):193–200. [PubMed: 16187362]
24. Aszodi A, Gradwell MJ, Taylor WR. Global fold determination from a small number of distance restraints. *J Mol Biol* 1995;251:308–326. [PubMed: 7643405]
25. Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C, Baker D. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins* 1999;34:82–95. [PubMed: 10336385]
26. Cheng J. DOMAC: an accurate, hybrid protein domain prediction server. *Nucleic Acids Res* 2007;35:W354–W356. [PubMed: 17553833]
27. Saini HK, Fischer D. Meta-DP: domain prediction meta-server. *Bioinformatics* 2005;21:2917–2920. [PubMed: 15840708]
28. Bryson K, Cozzetto D, Jones DT. Computer-assisted protein domain boundary prediction using the Dom-Pred server. *Curr Protein Pept Sci* 2007;8:181–188. [PubMed: 17430199]
29. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* 1999;292:195–202. [PubMed: 10493868]
30. Guoguang, L. DOMID. Version 1.0 Mar-15-1999. Karolinska Institute. Version 2.5 Nov-9-1999. Lund Univ; 1999.
31. Taylor WR. Protein structural domain identification. *Protein Eng* 1999;12:203–216. [PubMed: 10235621]
32. Chen J, Anderson JB, DeWeese-Scott C, Fedorova ND, Geer LY, He S, et al. MMDB: Entrez's 3D-structure database. *Nucleic Acids Res* 2003;31:474–477. [PubMed: 12520055]
33. Tress M, Tai CH, Wang G, Ezkurdia I, Lopez G, Valencia A, Lee B, Dunbrack RL Jr. Domain definition and target classification for CASP6. *Proteins* 2005;61(Suppl 7):8–18. [PubMed: 16187342]

34. Galzitskaya OV, Melnik BS. Prediction of protein domain boundaries from sequence alone. *Protein Sci* 2003;12:696–701. [PubMed: 12649427]
35. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. *J Mol Biol* 1982;157:105–132. [PubMed: 7108955]
36. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;25:3389–3402. [PubMed: 9254694]
37. Jaroszewski L, Rychlewski L, Li Z, Li W, Godzik A. FFAS03: a server for profile–profile sequence alignments. *Nucleic Acids Res* 2005;33:W284–W288. [PubMed: 15980471]
38. Taylor WR. Protein structure comparison using iterated double dynamic programming. *Protein Sci* 1999;8:654–665. [PubMed: 10091668]
39. Tress M, Cheng J, Baldi P, Joo K, Lee J, Seo JH, et al. Assessment of predictions submitted for the CASP7 domain prediction category. *Proteins* 2007;69(Suppl 8):137–151. [PubMed: 17680686]
40. Shoemaker BA, Wang J, Wolynes PG. Structural correlations in protein folding funnels. *Proc Natl Acad Sci USA* 1997;94:777–782. [PubMed: 9023333]
41. Veretnik S, Bourne PE, Alexandrov NN, Shindyalov IN. Toward consistent assignment of structural domains in proteins. *J Mol Biol* 2004;339:647–678. [PubMed: 15147847]
42. Xu Y, Xu D, Gabow HN. Protein domain decomposition using a graph-theoretic approach. *Bioinformatics* 2000;16:1091–1104. [PubMed: 11159328]
43. Madej T, Gibrat JF, Bryant SH. Threading a database of protein cores. *Proteins* 1995;23:356–369. [PubMed: 8710828]
44. Alexandrov N, Shindyalov I. PDP: protein domain parser. *Bioinformatics* 2003;19:429–430. [PubMed: 12584135]
45. Holm L, Sander C. Parser for protein folding units. *Proteins* 1994;19:256–268. [PubMed: 7937738]
46. Wu Y, Chen M, Lu M, Wang Q, Ma J. Determining protein topology from skeletons of secondary structures. *J Mol Biol* 2005;350:571–586. [PubMed: 15961102]
47. Wang G, Dunbrack RL Jr. PISCES: a protein sequence culling server. *Bioinformatics* 2003;19:1589–1591. [PubMed: 12912846]
48. Allen, MP.; Evans, GT.; Frenkel, D.; Mulder, BM. Hard convex body fluids. In: Prigogine, I.; Rice, SA., editors. *Advances in Chemical Physics*. Vol. 86. John Wiley Sons, Inc.; New York, NY: 2007. p. 1-166.
49. Bahar I, Jernigan RL. Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation. *J Mol Biol* 1997;266:195–214. [PubMed: 9054980]
50. Smith TF, Waterman MS. Identification of common molecular subsequences. *J Mol Biol* 1981;147:195–197. [PubMed: 7265238]

Abbreviations used

SSE	secondary-structure element
SSSM	super secondary structure motif
MC	Monte Carlo
DLI	domain linker index
REI	residue entropy index
GHL	George and Heringa linker index

KDH	Kyte and Doolittle hydropathy index
ARM	DLI+REI consensus domain predictor Armadillo
GM	Galzitskaya and Melnik
SCOP	Structural Classification of Proteins
DBPL	domain boundary profile library

Fig. 1.

Flowchart of VECFOLD-based domain boundary prediction method OPUS-Dom. Sequence profiles are generated from the query sequence by PSI-PRED and PSI-BLAST and fed into the VECFOLD method, which folds the predicted SSEs into a compact tertiary structure by template-based MC guided by a geometric scoring function. Both the scoring function and template library are derived from a nonredundant structure database. DOMID then labels the domains on each of 10^4 candidate structures generated by VECFOLD. The domain boundaries are counted by residue and a Z-score profile is generated from the smoothed distribution. This structure-based domain boundary profile is then combined with sequence-based profiles generated by REI, DLI, and DBPL from the query sequence. REI, DLI, and DBPL are

sequence-based domain boundary predictors that serve to enhance the specificity of the structure-based predictor.

Fig. 2.

(a) The C-MYB DNA-binding domain (PDB code 1MSE). (b) The copper chaperone for superoxide dismutase (PDB code 1QUP). (c) The human heart short-chain α -3-hydroxyacyl CoA dehydrogenase (PDB code 2HDH). A cartoon of the native structure is shown on the far right and consists of two domains specified by the MMDB. The N-terminus domain is shaded black and the C-terminus domain is in white. The two left structure models were generated by VECFOLD, and the color index is the same as for the native domain assignment.

Fig. 3. Domain boundary prediction examples for (a) protein 1QUP, (b) protein 1QTS, (c) protein 1QBA, (d) protein 2IDB, (e) protein 1E4E, and (f) protein 1IH7. The black line-dot profile in the lower part of the figure is the Z-score of the statistical distribution of structure-based domain assignments by DOMID on models generated by VECFOLD. The continuous, dashed, and dotted lines in the upper part of the figure are Z-scores from the three sequence-based filters. The horizontal continuous and dashed-dotted lines are the Z-score cutoffs. The true boundaries are indicated by the shaded bar below each set of Z-score profiles. The continuous arrows indicate correctly predicted boundaries (true positives), whereas the dashed arrows indicate incorrect boundary predictions (false positives).

Fig. 4. (a) The official domain assignment for protein 2IDB indicated by shades of gray on a cartoon of the native structure and (b) the domains predicted by OPUS-Dom, shaded with the same grayscale index as on the native structure. Note that the domain boundary regions of (a) and (b) differ only slightly. This figure was generated by PyMOL (DeLano Scientific, LLC).

Fig. 5. Sensitivity and specificity versus protein size for the (a) MMDB, (b) GM, and (c) Miyazaki benchmark sets. The correlation coefficients indicate that the sensitivities and specificities are uncorrelated with chain length. Note that the MMDB benchmark set is composed of two-domain proteins with only a single linker, such that the sensitivity is bimodal (0% or 100%).

Fig. 6.

Comparison of the domain boundary predictions from DOMID and Taylor's method. (a and b) Examples of globally similar results with highly correlated Z-score profiles. Here we found slight differences in peak shift and relative peak height. (c) An example of a structure with less correlated Z-score profiles.

Fig. 7.

Geometry of two consecutive secondary structures connected by a loop. An α -helix is represented by a vector along the cylindrical axis of the helix directed from the N terminus to the C terminus. For a loop or β -strand, the vector runs from the first C^α atom to the last C^α atom of the loop or strand. For any three vectors, we define the packing angle θ_1 with respect to vectors V_1 and V_2 , the packing angle θ_2 with respect to vectors V_2 and V_3 , and the dihedral angle ϕ with respect to all three vectors V_1 , V_2 , and V_3 .

Fig. 8. Schematic illustration of secondary-structure packing geometry and geometric parameters used to describe the packing. (a) helix–helix packing; (b) strand–strand packing; (c) helix–sheet packing.

Fig. 9.

Statistical behavior of the packing scoring function: for helix-helix packing with respect to the packing angle (a) and packing distance (b); for strand-strand packing with respect to the packing angle (c) and packing distance (d); for helix-sheet packing with respect to the packing angle ϕ (e) and packing angle θ (f). The range and division of bins is described in detail in Supplemental Materials.

Fig. 10. Correlation between protein radii of gyration calculated by the vector-based and traditional atom-based methods. The correlation coefficient is 0.936.

Table 1
Comparisons of OPUS-Dom domain prediction sensitivity and specificity to previous methods (DLI, REI, GHL, KDH, ARM) using three benchmark sets

	DLI ^a	REI ^a	GHL ^a	KDH ^a	ARM ^a	OPUS-Dom
MIMDB						
Sensitivity (%)	35	26	30	27	56	50
Specificity (%)	31	23	29	26	32	39
GM						
Sensitivity (%)	29	35	29	25	44	42
Specificity (%)	35	35	42	34	40	59
Miyazaki						
Sensitivity (%)	40	41	39	23	45	47
Specificity (%)	37	38	42	27	46	41

^aData in column from Dumontier *et al.*²⁰

Table 2
Number of actual and correctly predicted linkers for 29 protein structures from the GM data set

PDB (Chain ID)	Linker	TP	FN	FP	TP ₀	FN ₀
1CD9 (B)	1	1	0	0	0	0
1D9X (A)	5	2	3	0	0	0
1DR9 (A)	1	0	1	2	0	0
1DXH (A)	2	1	1	1	0	0
1E43 (A)	3	2	1	0	0	0
1E4E (A)	3	0	3	0	0	0
1EK4 (A)	0	0	0	1	0	1
1EP3 (B)	1	1	0	0	0	0
1F5Q (B)	1	1	0	0	0	0
1FNG (B)	1	1	0	0	0	0
1GE8 (A)	0	0	0	1	0	1
1HBN (A)	2	1	1	0	0	0
1HR6 (A)	0	0	0	3	0	1
1HWX (A)	3	1	2	1	0	0
1HYH (A)	0	0	0	2	0	1
1I4G (A)	1	1	0	0	0	0
1IH7 (A)	6	2	4	0	0	0
1ISA (A)	1	0	1	1	0	0
1JCF (A)	3	1	2	0	0	0
1JR3 (D)	2	1	1	1	0	0
1KMM (A)	2	2	0	0	0	0
1MPY (A)	1	0	1	1	0	0
1PDZ (-)	3	3	0	0	0	0
1PMT (-)	0	0	0	1	0	1
1QH4 (A)	1	0	1	0	0	0
1QM9 (A)	1	0	1	1	0	0
1X01 (A)	1	0	1	1	0	0
2NAP (A)	6	2	4	1	0	0
2PGD (-)	1	1	0	1	0	0
Total	52	24	28	19	0	5

PDB (Chain ID)	Linker	TP	FN	FP	TP ₀	FN ₀
TP, true positive; FN, false negative; FP, false positive; TP ₀ , true positive for zero domain linkers; FN ₀ , false negative for zero domain linkers.						

Table 3

Comparisons of OPUS-Dom domain prediction sensitivity and specificity to that of Rosetta-DOM for the CASP6 multidomain benchmark set

Method	Rosetta-DOM ^a	OPUS-Dom
Sensitivity (%)	29	55
Specificity (%)	55	56

^aData in column from Kim *et al.*²³

Table 4 OPUS-Dom domain prediction sensitivity and specificity for the CASP7 benchmark set

OPUS-Dom	One domain ^a	Two domains ^b	Three domains ^c	Multi domain ^d	All
Sensitivity (%)	54	47	40	44	51
Specificity (%)	52	41	100	47	50

^aConsists of 68 targets.
^bConsists of 25 targets.
^cConsists of 2 targets.
^dConsists of 14 targets with >150 residues and for which PSI-BLAST³⁵ and the fold recognition method FFAS03³⁶ cannot find any proper templates.

Table 5

Sensitivity and specificity of domain prediction using DOMID and Taylor's method

	DOMID	Taylor
MMDB		
Sensitivity (%)	50	59
Specificity (%)	39	43
GM		
Sensitivity (%)	42	49
Specificity (%)	59	68
Miyazaki		
Sensitivity (%)	47	58
Specificity (%)	41	45
CASP6-Multi		
Sensitivity (%)	55	42
Specificity (%)	56	46
CASP7-Multi		
Sensitivity (%)	44	39
Specificity (%)	47	39

Table 6

Sensitivity and specificity of domain prediction with and without sequence-based filters

	OPUS-Dom	OPUS-Dom without filters
MMDB		
Sensitivity (%)	50	74
Specificity (%)	39	32
GM		
Sensitivity (%)	42	60
Specificity (%)	59	41
Miyazaki		
Sensitivity (%)	47	69
Specificity (%)	41	33
CASP6-Multi		
Sensitivity (%)	55	58
Specificity (%)	56	41
CASP7-Multi		
Sensitivity (%)	44	50
Specificity (%)	47	26