Bioinformatics, 31(13), 2015, 2098–2105
doi: 10.1093/bioinformatics/btv092
Advance Access Publication Date: 19 February 2015
Original Paper

OXFORD

Structural bioinformatics

AIDA: *ab initio* domain assembly for automated multi-domain protein structure prediction and domain–domain interaction prediction

Dong Xu[1], Lukasz Jaroszewski[1,2], Zhanwen Li[1] and Adam Godzik[1,2,3,*]

[1]Bioinformatics and Systems Biology Program, Sanford-Burnham Medical Research Institute, 10901 North Torrey Pines Road, La Jolla, CA 92037, USA, [2]Center for Research in Biological Systems, University of California, San Diego, 9500 Gilman Dr. La Jolla, CA 92093-0446, USA and [3]Center of Excellence in Genomic Medicine Research (CEGMR), King Fahad Medical Research Center, King Abdulaziz University, Jeddah, Kingdom of Saudi Arabia

*To whom correspondence should be addressed.
Associate Editor: Anna Tramontano

Received on September 22, 2014; revised on January 16, 2015; accepted on February 10, 2015

Abstract

Motivation: Most proteins consist of multiple domains, independent structural and evolutionary units that are often reshuffled in genomic rearrangements to form new protein architectures. Template-based modeling methods can often detect homologous templates for individual domains, but templates that could be used to model the entire query protein are often not available.
Results: We have developed a fast docking algorithm *ab initio* domain assembly (AIDA) for assembling multi-domain protein structures, guided by the *ab initio* folding potential. This approach can be extended to discontinuous domains (i.e. domains with 'inserted' domains). When tested on experimentally solved structures of multi-domain proteins, the relative domain positions were accurately found among top 5000 models in 86% of cases. AIDA server can use domain assignments provided by the user or predict them from the provided sequence. The latter approach is particularly useful for automated protein structure prediction servers. The blind test consisting of 95 CASP10 targets shows that domain boundaries could be successfully determined for 97% of targets.
Availability and implementation: The AIDA package as well as the benchmark sets used here are available for download at http://ffas.burnham.org/AIDA/.
Contact: adam@sanfordburnham.org
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Domains are evolutionarily, structurally and functionally independent units in proteins. Proteins, especially large, eukaryotic ones are often composed of multiple domains, mostly due to the duplications and recombinations of the coding regions during evolution (Bjorklund *et al.*, 2006; Chothia *et al.*, 2003). Average eukaryotic and prokaryotic proteins have 2.1 and 1.5 domains, respectively (Apic *et al.*, 2003; Brocchieri and Karlin, 2005; Ekman *et al.*, 2005; Zmasek and Godzik, 2012)—a low estimate, as only known Pfam domains were used in the count, and new domains are continuously being discovered. Domains usually have compact 3D shapes and specific functions that are, at least to some extent, conserved between homologous domains in different proteins. The same domains are often found in proteins with different domain architectures, including standalone, single-domain ones. Moreover, structures of many domains are solved only as single-domain constructs. As a result, domain assembly is a necessary step for structure predictions of full-length proteins. Correctly arranged domain structures are often crucial for full understanding of the function of multi-domain proteins (Ben-Zeev *et al.*, 2005).

Residues within each domain usually have strong interactions with each other, stabilizing the conserved domain structure, while only a fraction of all residues are involved in the interactions with other domains (Han *et al.*, 2007). This increases the difficulty of accurate determination of the relative domain positions and orientations. The interactions involved in domain–domain interactions and those between different protein chains are very similar, therefore the domain–domain and protein–protein docking pose exactly the same challenges (Kanaan *et al.*, 2009) and could be approached with the same tools (Cheng *et al.*, 2008; Inbar *et al.*, 2005; Lise *et al.*, 2006). However, due to the restriction of polypeptide chain connectivity, the search space for domain position and orientation is much smaller than that for protein–protein docking. The problem of domain assembly could also be regarded as a special case of the folding process, in which only the conformation of the linker region is changing and domain structures are rigid. Hence, potentials and algorithms for single-domain protein structure prediction, such as Rosetta (Wollacott *et al.*, 2007), could be applied to this problem.

In this article, we describe a fast energy minimization method for domain assembly guided by the *ab initio* folding potential (Xu and Zhang, 2012). This method is implemented in the first publicly available server in the field, AIDA (*ab initio* domain assembly), available at http://ffas.burnham.org/AIDA/ (Xu *et al.*, 2014a,b). This server also provides access to a recursive protocol, which combines template-based modeling with domain assembly in an iterative method suitable for automated domain assignment, modeling and assembly for a one-stop structure prediction of multi-domain proteins.

2 Methods

2.1 Statistics of protein domain structures

We first split all the individual chains from the set of non-redundant protein structures in the Protein Data Bank (PDB) (Berman *et al.*, 2000) into domains, using the automatic tool DomainParser (Xu *et al.*, 2000). The statistics of the number of domains in the non-redundant set of protein chains are shown in Figure 1a. Most of the protein chains in the PDB contain only one domain, while only 32.7% contain multiple domains. Very few chains have more than five domains and the maximum number of domains is 20 (in *Human Complement Factor H protein*, PDB ID 3gaw, chain A). This is clearly a bias introduced by constraints of experimental structure determination, because the distribution of domains in known proteomes is strongly biased toward multi-domain proteins.

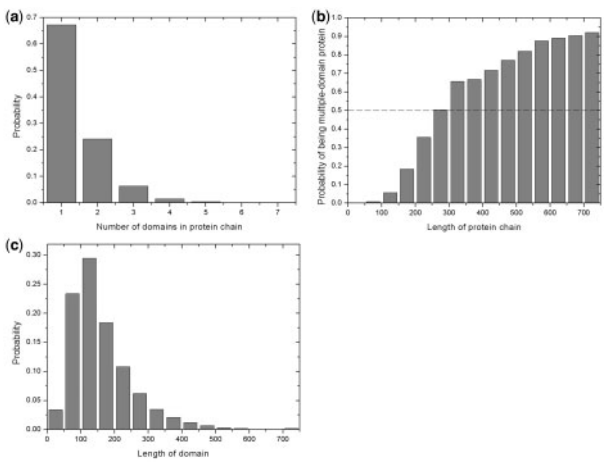

Fig. 1. (**a**) Distribution of domains in protein chains in PDB. (**b**) Probability of finding multiple domains in protein chains of different lengths. (**c**) Length distribution of all protein domains

As expected, the probability of finding multiple domains in a chain increases with its length. As shown in Figure 1b, more than half of the chains longer than 275 amino acids contain at least two domains. We also analyzed the length distribution of the domains (Fig. 1c). Majority of them (71.3%) have lengths around 150 residues and very few contain more than 600 amino acids. The maximum domain length is 1317 [the third domain of *fungal fatty acid synthase* (PDB ID 2uv9, chain D)]. Definitions of domain boundaries are often somewhat fuzzy and may differ between different domain parsing algorithms and even between manual assignments by different experts. Here, we consistently use DomainParser definitions.

2.2 Potential describing domain–domains interactions

AIDA represents protein structures by a reduced model, where each residue contains four backbone atoms and a single point representing the side-chain center. In the reduced model used by AIDA, the positions of side-chain centers are always estimated based on the backbone geometry. Because we only change the conformations of the linker regions, keeping domains' conformations unchanged, intra-domain energies remain constant. Therefore, energy for the assembled multi-domain protein structure E_{tot} (Equation 1), that needs to be minimized, is the sum of inter-domain interaction energy E_{int} and the conformational energy of the linker E_{link}. The potential for domain–domain and domain–linker interactions contains five terms, which were already tested in *ab initio* folding calculations. E_{prm} and E_{prs} describe the pairwise short-range interactions between main-chain atoms and side-chain center, which originally were introduced in Distance-scaled, Finite Ideal-gas REference (DFIRE) (Zhou and Zhou, 2002) knowledge-based potential. Excluded-volume term E_{ev} prevents atomic clashes between different domains. E_{hb} is the backbone hydrogen-bond potential, especially important in describing interaction between two strands belonging to different domains. When domains come close to each other, solvent accessibilities of the interface residues become smaller, as compared with those in separated domains. E_{sa} compares the solvent accessibility in the assembled structure with that predicted by a neural network. Even if solvent accessibility prediction contains some degree of error, this term generally could guide the simulation to find the correct minimum of the domain–domain interaction energy.

Because the linker region is often lacking any regular secondary structure and we need its conformation to be thoroughly sampled, we only use two basic terms to control the quality of its structure. $E_{C\alpha}$ is the penalty when the distance between every two consecutive $C\alpha$ atoms exceeds the standard distance while E_{dh} is the dihedral-angle potential, which prevents unfavorable torsion angles in the Ramachandran plot (Ramachandran and Sasisekharan, 1968).

$$\begin{aligned} E_{\text{tot}} &= E_{\text{int}} + E_{\text{link}} \\ E_{\text{int}} &= E_{\text{prm}} + E_{\text{prs}} + E_{\text{ev}} + E_{\text{hb}} + E_{\text{sa}} \\ E_{\text{link}} &= E_{C\alpha} + E_{\text{dh}} \end{aligned} \quad (1)$$

By default, four residues around the domain boundary are treated as a linker. AIDA users may truncate terminal residues in the domain structures and they will be included in the linker region as well.

2.3 Energy minimization simulation

Due to the chain connectivity constraints and fixed domain structures, the conformational sampling space is significantly reduced.

In the AIDA program, we only perform a single-trajectory energy minimization to obtain one assembled structure at a time. During each energy minimization, the new conformation resulting from a random movement in the linker region is accepted only if it has lower total energy than the previous one. The simulation stops when the total number of attempts exceeds $1000L$ (where L is the length of the entire sequence) or the assembled structure reaches the local or global minimum state (i.e. conformation with lower energy cannot be found after 200 consecutive random movements). At the end of the simulation, all side-chain atoms are added to the reduced model using SCWRL4 (Krivov *et al.*, 2009).

We use nine types of local movements adopted from QUARK program (Xu and Zhang, 2012) to change the conformation of the linker region. Those movements include the change of bond lengths, bond angles and torsion angles of one residue in the torsion-angle coordinate system and perturbation of coordinates of backbone atoms in a short segment in the Cartesian coordinate system.

2.4 Assembly of continuous domains with non-continuous domains

If all the domains are continuous at the sequence level, such as the 2-domain protein in Figure 2a, then the position and orientation of the second domain is completely flexible and dependent only on the conformation of the linker (colored in red). However, if there is a domain that contains two discontinuous parts in the sequence, such as in the 2-domain protein shown in Figure 2b, then we treat the two parts of the discontinuous domain as the first and third domains and the whole region in the middle as the second domain. We recalculate the conformation of the two linkers and the position of the middle domain and keep the positions of the first and third domains fixed. In the beginning of the simulation, the second linker and the third domain may be disconnected, but the penalty potential $E_{C\alpha}$ gradually pulls them together. Sometimes, the distance between the two ends of the predicted middle domain is significantly different from the distance between the two break points of the discontinuous domain. In this situation, it is impossible to generate a model without breaking the chain if we only change the conformation of a small number of residues in the two linkers. To address such situations, the program gradually extends the length of the linker region until it fulfills conformational constraints.

2.5 Recursive domain splitting, modeling and assembly

The first step of template-based structure prediction is usually threading, where one tries to identify the best template and generate the alignment between the query sequence and the template. We use local–local alignment programs such as FFAS (Fold and Function Assignment System) (Jaroszewski *et al.*, 2005; Xu *et al.* 2014a,b) for threading, since in the local–local alignment, template selection is not affected by the length difference between the query and the template. Based on the initial threading alignment, at most three domains can be defined (N-terminal unaligned region, aligned region in the middle, C-terminal unaligned region), as shown in Figure 3. The middle region, which is aligned to the template, may contain multiple domains. It is noteworthy that Phyre2 server (Kelley and Sternberg, 2009) also enables basic prediction of domain architecture by presenting threading alignments in a graphical form.

The region matched to a template is modeled by the Modeller program (Sali and Blundell, 1993) based on the threading alignment from FFAS-3D. However, the alignment of this part may contain a large gap in the middle, which is treated as an additional domain and modeled separately. Hence, the original modeling result of this gap region by Modeller is deleted and the other two parts, which now form a discontinuous domain, will be assembled together with the middle domain (see the modeling procedure of Dom2 in Fig. 3).

For the two unaligned terminal regions outside the central domain, two separate threading procedures are performed. In the example shown in Figure 3, most of the N-terminal region is aligned with a template. We stop splitting of the terminal regions if the number of unaligned residues is smaller than 20 (which in many cases is a signal peptide) or if <30% of that region is predicted to be in alpha-helical or beta-sheet structures. In the example shown in Figure 3, few unaligned residues are built by Modeller as a random coil. The C-terminal region is split into two domains, one of which will be modeled directly since it is aligned with the template. The procedure of splitting and modeling is continued until remaining regions cannot be further divided into smaller domains. If models of all the sub-regions have been generated, then AIDA will assemble them together in a hierarchal fashion to build a model covering the entire protein (Fig. 3). We want to stress here that, while many other servers can also generate full-length models for multi-domain proteins, they can do it only if a multi-domain template is available (otherwise the relative orientations of individual domains become arbitrary).

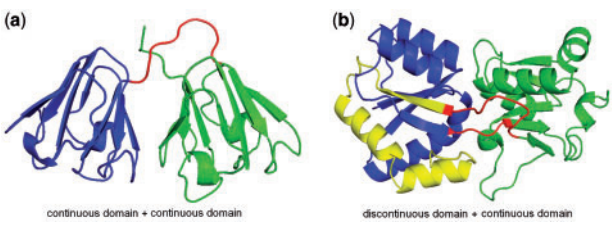

Fig. 2. Continuous and discontinuous domains. (**a**) *GammaB-crystallin* (PDB chain: 1ammA) contains two continuous domains (in blue and green, respectively). (**b**) *Azotobacter vinelandii periplasmic molybdate-binding protein* (PDB chain: 1atgA) contains one discontinuous domain (in blue and yellow) and one continuous (in green), inserted in the middle of the first one. Linker regions are colored in red

3 Results

3.1 Construction of the benchmark set

We first obtained a non-redundant set of protein chains from the PISCES (*Protein* Sequence Culling Server) (Wang and Dunbrack, 2003) server. We selected structures solved by X-ray crystallography

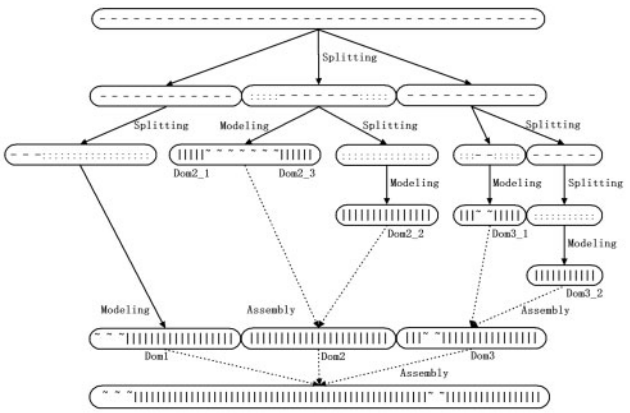

Fig. 3. The schema of the domain splitting, modeling and assembly procedure. '–' and ':' denote gap and aligned region after the threading. '~' and '|' stand for coil and reliable region after template-based modeling

with resolutions better than 2.5 Å, clustered at sequence identity cut-off of 40%. We then used DomainParser to determine the number of domains in each chain. Single-domain proteins were discarded and multi-domain proteins whose domains have no contacts (i.e. there is no pair of atoms in two domains with a distance below 5 Å) were also not considered. For multi-domain proteins, whose domains interact with other molecules in heterocomplexes, domain arrangements usually cannot be correctly predicted if no additional information is provided. We manually selected 13 proteins with more than three domains. Based on the distribution of the numbers of domains in Figure 1a, we then selected 136 2-domain proteins and 36 3-domain proteins *pro rata*. All the domains in these 185 proteins are continuous. The 76 2-domain proteins used in the earlier work (Wollacott *et al.*, 2007) are also included in this set for the purpose of comparison. We also picked up 20 2-domain proteins containing discontinuous domains and inserted domains. This set was used for testing the assembly of discontinuous domains.

For an additional, more realistic test, we selected multi-domain protein structures where individual domains have also been solved independently. Because structures of separate domains are usually slightly different than the corresponding domain structures in multi-domain proteins, this test is more realistic and more demanding than the test described in previous paragraph. In the first step, we selected 4222 non-redundant 2-domain structures (as assigned by DomainParser) from PDB. Then for each domain in a 2-domain structure, we searched for single-domain proteins by running PSI-BLAST (Altschul *et al.*, 1997) against the entire PDB database. From PSI-BLAST output, we selected PDB coordinate sets that were aligned with each individual domain. We selected examples where independently solved single-domain structures were highly similar to corresponding domains in multi-domain proteins (sequence identity of at least 80% and alignment covering at least 75% of a single domain). The above procedure yielded 122 examples where both domains in a 2-domain protein have corresponding single-domain structures. We then removed cases where two domains in the 2-domain protein had no strong interactions or the linker structure was not determined. The final set (called 2unbound) includes 24 examples of 2-domain proteins where both domains have independently solved structures.

The five categories of multi-domain proteins (229 in total) included in the benchmark set are shown in Table 1. Because for the first four categories, individual domains are directly taken from the original multi-domain proteins, each domain was randomly rotated and translated in a $20 \times 20 \times 20$ box before running the tests.

Table 1. Domain numbers and assembly result for different categories of examples in the benchmark set

Number of domains	Number of proteins	Number of proteins with correct predicted assembly	Number of proteins with correct assembly in five lowest-energy models	Success rate of energy-based selection (%)
2	76 + 60	88	73	83.0
3	36	21	14	66.7
>=4	13	3	1	33.3
2dis	20	19	14	73.7
2unbound	24	13	7	53.8

2dis: 2-domain proteins with one discontinuous domain. 2unbound: 2-domain proteins with each domain solved independently.

Domain boundaries were determined manually. Lengths of linkers in the benchmark examples range from 4 to 31 residues.

3.2 Assembly of native domain structures

In this test, the structures of individual domains were adapted directly from native protein structures. The initial conformation of the linker regions was constructed based on the secondary structure types predicted by PSIPRED (Jones, 1999). Solvent accessibility of each residue (used in energy calculation) was predicted by a 2-layer neural network (Xu and Zhang, 2012).

We generated 5000 assembled structures for each of the 76 2-domain proteins. We assumed that domain assembly procedure failed when none of the assembled structures in that set had Root Mean Square Deviation of C-alpha atoms (RMSD) < 3.0 Å to the native structure. In total, AIDA failed for 11 targets, which is slightly lower than 13 cases where Rosetta (Wollacott *et al.*, 2007) failed by the same criterion on the same set, as shown in Table 2. Five unsuccessful predictions are common for both methods. We can reason that these failures are due to the relatively small interface between the two domains and the relatively long linker regions (~20 residues), which significantly expand the search space. There are also eight targets, which were not correctly predicted by Rosetta but were successfully assembled by AIDA, three of which are shown in Figure 4. The predicted model in Figure 4a has RMSD = 0.5 Å to the native structure (PDB ID 1nkr, chain A). Domain–domain interaction is weak there since the two domains are separated. However, three residues in the linker region form a parallel beta-sheet with three residues in the second domain. AIDA correctly predicted the domain orientations as well as the conformation of the linker region.

In Figure 4b, the four residues in the linker region of the *N-terminal domain of N-ethylmaleimide-sensitive factor (NSF)* (PDB ID 1qcs, chain A) form a coil structure, which has no interaction with the two adjacent domains. However, the two domains have strong side-chain atomic interactions with each other. The best AIDA model has RMSD = 0.5 Å to the native structure, which correctly captured those interactions. The success in this case is mainly due to the synergistic effect of pairwise side-chain potential and solvent accessibility term. Interestingly, only one residue in the middle of the linker shows a big deviation from its position in the native structure. The realistic energy function is expected to guide the simulation toward the native domain assembly. Hence, we expect higher number of near-native models in the resulting set of models and significant correlation between RMSD and TM-score (Zhang and Skolnick, 2004) (Supplementary Fig. S1a and b in the Supplementary Materials). In contrast, inaccurate potential would not guide the simulations toward native assembly and as a result, simulations generate very few near-native models (Supplementary Fig. S1c and d).

Table 2. Comparison with Rosetta on 76 2-domain proteins

Method	Number of assembly attempts per protein	Number of proteins with assembly having RMSD < 3.0 Å to native	Number of proteins with five lowest-energy models having RMSD < 3.0 Å to native	Success rate of energy-based selection (%)
ROSETTA	5000	63	38	60.3
AIDA	50	51	46	90.2
AIDA	200	58	49	84.5
AIDA	1000	61	52	85.2
AIDA	5000	65	52	80.0

Sulfur-substituted rhodanese (PDB ID 1rhs, chain A) in Figure 4c has a long linker of 19 residues, which wraps around one half of the first domain. The two domains have a large interface, which allowed AIDA to correctly identify their relative positions. The RMSD between the model and the native structure is 1.7 Å. From the figure, we can find that higher RMSD is probably a result of the slight shift of the first domain and the structural error in the long linker.

From six targets which were not correctly predicted by AIDA but were correctly predicted by Rosetta, three have unusually long inter-domain linkers and for the other three, AIDA failed probably because of the insufficient accuracy of the folding potential. One may anticipate that the inadequate sampling in structures with long inter-domain linkers may be another key factor affecting the accuracy of domain assemblies. In order to gain insight into this, we split all the 136 2-domain proteins into three subsets according to their linker lengths. As we can see from Supplementary Table S1, the success rate decreases dramatically for proteins with longer linkers. When the linker is short, domains can only occupy relatively limited space with respect to each other and their interactions usually involve residues that are close to the linker. In contrast to that, when a linker is long, one domain can interact with another domain via many possible interfaces, which dramatically increase the search space for the simulation. This observation is in agreement with the general trend—the accuracy of simulation decreases with the increasing sequence separation between interacting residues (Xu and Zhang, 2012).

From Table 2, we can see that using more models increases the probability of obtaining a correctly assembled structure among generated models. Unfortunately, it is difficult to identify the best model without knowing the native structure. We selected five models with the lowest total energies as the representative output of each protein. The total energy is calculated based on the reduced model using Equation (1). We list the cases that contain low-RMSD (<3.0 Å) model in the five selected models in Table 2 for different numbers of generated models. If we only generate 50 models per target, 51 of the 76 targets contain low-RMSD models and we successfully select good models for 46 of the 51 targets. It is much higher than 38 as reported for Rosetta, where five models were selected from 5000 models based on the Rosetta's high-resolution energy. Each of these low-resolution models was refined via energy minimization and subject to side-chain repacking to generate a full-atom model. The comparison between domain assembly results obtained with AIDA and Rosetta reveals that the model selection by AIDA's coarse-grained energy not only increases the accuracy, but also reduces computational cost (Table 2).

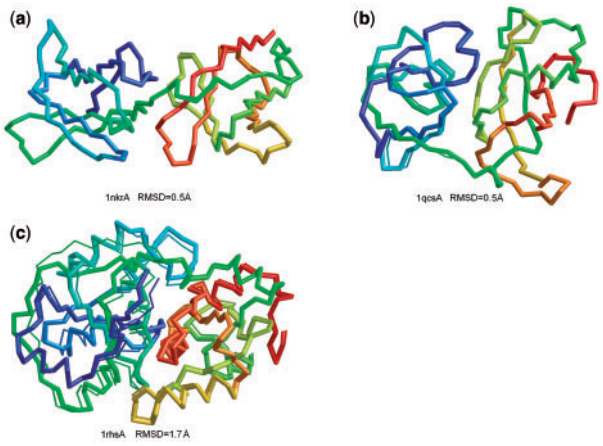

Fig. 4. Three examples of 2-domain assembly. Predicted assembly is represented by thick backbone while native structure is by thin backbone. (**a**) *The inhibitory receptor for human natural killer cells* 1nkrA, linker length = 4, RMSD = 0.5 Å. (**b**) *The N-terminal domain of NSF* 1qcsA, linker length = 4, RMSD = 0.5 Å. (**c**) *Sulfur-substituted rhodanese* 1rhsA, linker length = 19, RMSD = 1.7 Å

With the increasing number of generated models, the number of targets where correct models were selected is also increasing, confirming the accuracy of the energy terms used in AIDA. However, the success rate, which is defined as the ratio of the number of targets with successful selection to the number of targets containing good models, is generally decreasing. Obviously, the cost of generating 5000 models is 100 times higher than for 50 models. Therefore, in our tests, we generated 50 models for each target in the benchmark set.

Because RMSD is sequence-length dependent, the assembled multi-domain structure may have RMSD > 3.0 Å if the sequence is long, even when the arrangement of the domains is correct. Hence, we used TM-score, which takes the length dependence into account to evaluate the difference between each model and the native structure. The highest TM-score $TM^{\max}$ of all the individual domains is normally from the longest domain. The assembled structure should have TM-score larger than $TM^{\max}$ even when the domain arrangement is completely wrong. We consider the assembly result is correct if the remaining domains (other than the one which individually yields the highest TM-score) contributed at least half of the total TM-score value (Xu and Zhang, 2010). Hence, the TM-score cutoff value for correct domain arrangement is defined by Equation (2). Here, TM-score of the ith domain structure is denoted as TM^i and their sum should be close to 1 if each domain was directly extracted from a multi-domain protein.

$$TM^{\text{cutoff}} = TM^{\max} + 0.5 \times \left(\sum_{i=1}^{m} TM^i - TM^{\max}\right)$$
$$= 0.5 \times \left(\sum_{i=1}^{m} TM^i + TM^{\max}\right) \quad (2)$$

We then check the 50 models for each benchmark target to see if any good model was generated and if it was selected by the AIDA potential. The detailed results for proteins with different number of domains are shown in Table 1. From the last column of the table, we can find that the probability of finding a good model among 50 models becomes lower for a larger number of domains in a protein. It seems that the success rate of energy-based selection is also dependent on the number of domains. This is caused by the larger search space and higher chance of selecting incorrect domain assembly for proteins with a larger number of domains.

One successful domain assembly result for 5-domain beta-galactosidase from *Arthrobacter sp. C2-2 protein* (PDB ID 1yq2, chain A) is shown in Figure 5a. All the five domains in the model are in the correct position as compared with the native structure. The most challenging part was the arrangement of the fifth domain since the linker between the fourth and fifth domains is 12-residue long. Even if this linker is not predicted accurately, the fifth domain contacts with the third domain via the correct interface. For some of the failed cases, we still can assemble some of the domains correctly when domain number is larger than 2. One failed case, *human cytosolic X-prolyl aminopeptidase* (PDB ID 3ctz, chain A) in Figure 5b contains three domains, whose TM-score 0.719 is lower than the cutoff value 0.742. The third domain in the native structure and the model match with each other, but at most it can lead to the

TM-score gain of 0.480. Part of the second domain in the model has the correct position and orientation, which contributes to the final TM-score. The first domain is in the correct position, but the orientation is completely wrong.

For the set of proteins with discontinuous domains, 19 out of 20 targets are successfully assembled. Because there are two linker regions between the discontinuous domain and the 'inserted' domain, the position of the 'inserted' domain is more restricted, which makes the assembly of those 2-domain proteins much easier than for the proteins with two continuous domains. All the 19 targets have models with TM-score > 0.9. The only case, where assembly failed is *ATP-Phosphoribosyltransferase(hisG) from Thermus thermophilus HB8* (PDB ID 1ve4, chain A) with TM-score $= 0.747$ which is close to the cutoff value 0.796.

3.3 Assembly of independently determined domain structures

Unlike the reassembly of domains directly extracted from the native proteins, assembly of independently solved domains reflects a scenario that is likely to be encountered in a real research problem. Our dataset contains 24 such examples and similarities between structures of individual domains and corresponding domains in 2-domain proteins vary from 0.5 to almost 1.0, as evaluated with TM-score. For example, the second domain in *FADD (MORT1)* protein (PDB ID 2gf5, chain A) and the first domain in *SMT fusion Peptidyl-prolyl cis-trans isomerase* (PDB ID 3uf8, chain A) only share the same folds with single-domain proteins—*human FADD death domain* (PDB ID 1e41, chain A) and *ubiquitin-like protein* Smt3 (PDB ID 1l2n, chain A) separately and the details of their backbone and side-chain conformations are quite different.

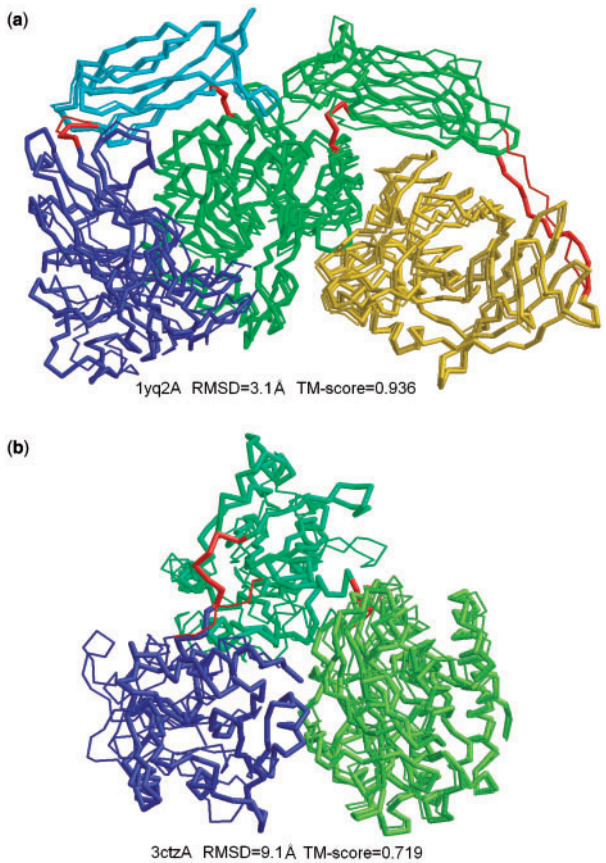

Fig. 5. Two examples of multi-domain assembly. Domains are colored differently and linkers are shown in red. Predicted assembly is represented by thick backbone while native structure is by thin backbone. (**a**) Successful assembly of 5-domain protein *beta-galactosidase from Arthrobacter sp.* (PDB ID 1yq2, chain A), RMSD = 3.1 Å, TM-score = 0.936. (**b**) Partially successful assembly of 3-domain protein *human cytosolic X-prolyl aminopeptidase* (PDB ID 3ctz, chain A), RMSD = 9.1 Å, TM-score = 0.719

We performed domain assembly experiment for those proteins by generating 50 assembled models for each of them using independently solved domain structures and selected top five models based on energy. The overall result is shown in the last row of Table 1. As compared with the benchmark result on the 136 pairs of domains extracted from multi-domain proteins, the success rate of generating good models decreases from 65 to 54%. It indicates that structural difference in independently solved domain structures indeed affects the accuracy of the method. However, there are still several successful examples of domain assembly using unbounded domain structures. As shown in Supplementary Figure S2a, the result of the assembly of independently solved domains structures is generally correct for 2gf5A. In particular, the single-domain protein 1e41A matches well with the second domain of 2gf5A after the superposition even if the TM- score between them is only 0.57. We also didn't find that the linker length is the major factor affecting the success rate. Two out of the four proteins, which have linker lengths > 20 are correctly assembled. For example, the linker between the two domains in *scFv-IL-1B complex* (PDB ID 2kh2, chain B) consists of 21 residues and, despite the fact that the predicted linker conformation significantly differs from the real one (shown in Supplementary Fig. S2b), the two domains were correctly assembled by AIDA. This is probably due to the strong interactions between them (two pairs of beta-sheets are formed between the two domains).

The success rate of selecting correct domain configuration by energy is significantly lower for assembly of independently determined domains than for domains extracted from the structure of the same protein (53.8 and 83%, respectively). The lower success rate indicates that small changes in the independent domain structures (including side-chain positions) cause some incorrect domain configuration results to have low energies. More importantly, the correct (near-native) assembly result may contain significant clashes when built from independently solved structures. For instance, for the *human ATP-dependent splicing and export factor UAP56* (PDB ID 1xtk, chain A), AIDA generated relatively accurate domain

Table 3. Domain splitting, modeling and assembly result on 95 CASP10 targets

Number of domains	Number of proteins	Number of successfully split proteins	Number of proteins with all domains correctly folded	Number of successfully assembled proteins
1	71	71	63	63
2in1	9	9	4	0
2	11	9	3	1
3	1	1	1	0
6in3	1	1	0	0
dis	2	1	1	1
Total	95	92	72	65

2in1: two domains are modeled together on one two-domain template. 6in3: six domains are modeled as three parts, with four domains modeled together on one four-domain template. dis: targets with one discontinuous domain.

configurations but they don't have low energy values. In Supplementary Figure S3, we show the superposition of 1xtkA with two incorrect domain configurations with the lowest energies. In the native structure, only a few atoms in one helix of the second domain interact with the first domain (the structure shown in red). However, AIDA generated domain configurations with stronger inter-domain interactions which have comparable or lower energies (models shown in green and blue). Model selection based on clustering may be a better solution in such cases.

3.4 Automated prediction of CASP10 targets

The assembly of predicted domain structures whose domain boundaries are determined automatically is obviously more challenging than predicting arrangement of domains derived from native structures. We conducted a blind test of the iterative AIDA protocol on the 95 CASP10 (The 10th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction) targets. In the automated protocol, FFAS-3D is used to generate sequence-template alignments, which are then used to determine domain boundaries.

In this protocol, the accuracy of the prediction of domain boundaries depends on the accuracy of the FFAS-3D alignments. As summarized in Table 3, only two 2-domain proteins T0671, T0724 and one 3-domain protein T0726, which includes discontinuous domain, have incorrectly predicted domain boundaries. For targets T0724 and T0726, the best template is correctly found for one domain, but the alignment also covers half of the next domain. Because most of the CASP10 targets are either single-domain proteins or 2-domain proteins with corresponding 2-domain PDB templates (rows 2–3 in Table 3), domain assembly is not applicable to them. The detailed analysis of these examples is given in the Supplementary Materials. Below we only analyze the small subset of 15 targets, where non-trivial domain assembly was needed (rows 4–7 in Table 3). It is noteworthy that domain assembly using homology-based models is a more demanding test for AIDA than domains extracted from multi-domain proteins and independently solved domain structures.

There are total 15 multi-domain targets for which AIDA was used to assembly the models of individual domains. For the nine 2-domain targets, whose domains are correctly identified but built from different templates, only three have both domains' structures correctly predicted, while each of the six remaining targets contains at least one free-modeling domain. Supplementary Figure S4a shows one successful assembly result for target T0674, whose domains have TM-score = 0.600 and 0.845 to the corresponding native structures. In particular, one beta-hairpin is incorrectly predicted for the first domain. However, the final assembled model still has acceptable TM-score of 0.539. As it is shown in the Supplementary Figure S4a, majority of the first domain and part of the second domain align well with the native structure.

The 3-domain protein T0651 is correctly split into domains and modeled, but the real linker region between the first and second domains is more than 15 residues long, which probably led to the wrong assembly prediction. Target T0719 contains six domains, but the protocol split it into two parts, the first of which contains four domains and the second contains two. Five of the domains were modeled correctly while the last one is a free-modeling domain. Unfortunately, the arrangement of the four domains in the model, which is directly copied from the template, was not correct.

For the discontinuous 2-domain protein T0755, we split it into two continuous domains. The modeling results of the two domains are still mostly correct, with TM-scores of 0.696 and 0.606. The assembly result for this target is shown in Supplementary Figure S4b. Most of the first domain in the model matches the native structure well and part of the second domain is placed in the correct position, which resulted in the TM-score = 0.538. The C-terminal part of the whole structure is modeled as part of the second domain, while in fact it belongs to the first domain.

FFAS-3D outputs a significance score (Z-score) for each alignment. If this score falls below accepted cutoff (−34), then the resulting model is expected to have correct fold. Thus, AIDA assembly procedure should only be applied when all individual domains can be modeled based on alignments with significant Z-scores.

4 Conclusions

Proteins, especially eukaryotic ones, often contain multiple domains. Fold prediction programs can often identify boundaries of individual domains even in the absence of the template containing all the domains present in the target but are typically not able to correctly predict the full-length structure because of the lack of appropriate template. We developed the AIDA server for modeling and assembly of domains in such cases. AIDA supports the assembly of any number of continuous domains and can be used for discontinuous domains. The method was tested using a benchmark set of 229 proteins, which includes multi-domain proteins with different number of domains, 2-domain proteins with one domain inserted in the other, and 2-domain proteins whose individual domains' structures were solved independently. AIDA could generate accurately assembled models (RMSD < 3.0 Å to native structure) within 5000 models for 65 out of 76 cases, which is slightly better than the current best result reported in literature. AIDA also has a high success rate in selecting the most accurate model based on the final energy value. Due to the presence of two linkers, which significantly reduce the search space, discontinuous domain assembly is easier than the assembly of continuous domains. Results of the assembly of independent domain structures are less accurate than results for domain structures extracted from the original multi-domain proteins. Apparently small difference in domain conformations interferes with the generation of near-native models and affects the energy-based model selection.

AIDA also provides an iterative protocol for automatic domain splitting, modeling and assembly, which is useful for automated protein structure prediction. The protocol was tested on 95 CASP10 targets, whose sequences were split and aligned with templates using FFAS-3D. Then models were built for each domain by Modeller. As many as 97% of the targets were correctly split into domains by the local–local alignment program FFAS-3D. From the blind test on CASP10 targets, we also observed that the arrangements of domain structures in different proteins are not necessarily the same even if structures of individual domains are similar. However, only for 5 out of the 15 multi-domain targets, all individual domains were correctly modeled. The low prediction accuracy of domain models made it difficult to assemble them correctly. Nevertheless, AIDA still could predict two out of five domain assemblies correctly (in cases when structures of individual domains were correctly predicted).

Acknowledgements

Part of energy terms and movements were developed in Prof. Yang Zhang's lab.

Funding

This work was supported by the National Institute of Health [grant number GM101457 to A.G.] and the National Natural Science Foundation of China [grant number F020504 to D.X.].

Conflict of Interest: none declared.

References

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Apic,G. *et al.* (2003) Multi-domain protein families and domain pairs: comparison with known structures and a random model of domain recombination. *J. Struct. Funct. Genomics*, **4**, 67–78.

Ben-Zeev,E. *et al.* (2005) Docking to single-domain and multiple-domain proteins: old and new challenges. *Proteins*, **60**, 195–201.

Berman,H.M. *et al.* (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.

Bjorklund,A.K. *et al.* (2006) Expansion of protein domain repeats. *PLoS Comput. Biol.*, **2**, e114.

Brocchieri,L. and Karlin,S. (2005) Protein length in eukaryotic and prokaryotic proteomes. *Nucleic Acids Res.*, **33**, 3390–3400.

Cheng,T.M. *et al.* (2008) Structural assembly of two-domain proteins by rigid-body docking. *BMC Bioinformatics*, **9**, 441.

Chothia,C. *et al.* (2003) Evolution of the protein repertoire. *Science*, **300**, 1701–1703.

Ekman,D. *et al.* (2005) Multi-domain proteins in the three kingdoms of life: orphan domains and other unassigned regions. *J. Mol. Biol.*, **348**, 231–243.

Han,J.H. *et al.* (2007) The folding and evolution of multidomain proteins. *Nat. Rev. Mol. Cell Biol.*, **8**, 319–330.

Inbar,Y. *et al.* (2005) Combinatorial docking approach for structure prediction of large proteins and multi-molecular assemblies. *Phys. Biol.*, **2**, S156–S165.

Jaroszewski,L. *et al.* (2005) FFAS03: a server for profile–profile sequence alignments. *Nucleic Acids Res.*, **33**, W284–W288.

Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.

Kanaan,S.P. *et al.* (2009) Inferring protein-protein interactions from multiple protein domain combinations. *Methods Mol. Biol.*, **541**, 43–59.

Kelley,L.A. and Sternberg,M.J. (2009) Protein structure prediction on the web: a case study using the Phyre server. *Nat. Protoc.*, **4**, 363–371.

Krivov,G.G. *et al.* (2009) Improved prediction of protein side-chain conformations with SCWRL4. *Proteins*, **77**, 778–795.

Lise,S. *et al.* (2006) Docking protein domains in contact space. *BMC Bioinformatics*, **7**, 310.

Ramachandran,G.N. and Sasisekharan,V. (1968) Conformation of polypeptides and proteins. *Adv. Protein Chem.*, **23**, 283–438.

Sali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815.

Wang,G. and Dunbrack,R.L. Jr. (2003) PISCES: a protein sequence culling server. *Bioinformatics*, **19**, 1589–1591.

Wollacott,A.M. *et al.* (2007) Prediction of structures of multidomain proteins from structures of the individual domains. *Protein Sci.*, **16**, 165–175.

Xu,D. *et al.* (2014a) AIDA: ab initio domain assembly server. *Nucleic Acids Res.*, **42**, W308–W313.

Xu,D. *et al.* (2014b) FFAS-3D: improving fold recognition by including optimized structural features and template re-ranking. *Bioinformatics*, **30**, 660–667.

Xu,D. and Zhang,Y. (2012) Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins*, **80**, 1715–1735.

Xu,J. and Zhang,Y. (2010) How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics*, **26**, 889–895.

Xu,Y. *et al.* (2000) Protein domain decomposition using a graph-theoretic approach. *Bioinformatics*, **16**, 1091–1104.

Zhang,Y. and Skolnick,J. (2004) Scoring function for automated assessment of protein structure template quality. *Proteins*, **57**, 702–710.

Zhou,H. and Zhou,Y. (2002) Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci.*, **11**, 2714–2726.

Zmasek,C.M. and Godzik,A. (2012) This *Déjà Vu* feeling—analysis of multi-domain protein evolution in eukaryotic genomes. *PLoS Comput. Biol.*, **8**, e1002701.