Journal of
Proteome Science &
Computational Biology

HOAJ
Herbert Open Access Journals

Research Article Open Access

Blind prediction of quaternary structures of homo-oligomeric proteins from amino acid sequences based on templates

Mizuki Morita[1], Masanori Kakuta[2], Kentaro Shimizu[2] & Shugo Nakamura[2*]

Abstract

Background: Prediction of protein tertiary and quaternary structures helps us to understand protein functionality. While tertiary structure prediction techniques have been much improved over the last two decades, quaternary structure (homo-oligomer) prediction has not been paid much attention to.

Results: We show the results of the assessment of our simple auto server prediction and manual prediction of protein quaternary structure from its amino acid sequence based on templates. They were tested in the 9th Critical Assessment of Protein Structure Prediction (CASP9) experiment. CASP experiments are the only true blind test for protein tertiary and quatenary structure prediction from amino acid sequence alone and therefore they are the most severe tests in the field of protein stucture prediction. Our simple auto server prediction could generate successful models for 14 out of 58 targets. Human experts could generate successful models for 11 out of 16 targets and most of them were better than those by our auto server.

Conclusions: The results show the efficiency of our template-based protein quaternary structure prediction approaches and provide useful information for improvement of the accuracy of template-based quaternary structure prediction.

Background

Proteins are capable of self-assembly and form homo-oligomers or homo-multimers. It has been estimated that about half of the known proteins are homo-oligomers [1,2]. Although the practical biological reasons that induce self-association of proteins are not yet completely understood, various examples of relationships between formation of homo-oligomers and their functions have been identified [3]. Catalytic sites of enzymes and ligand-binding sites of receptors are frequently located at the interface of subunits in order to form large binding pockets (Morita et al., unpublished results). Multiple binding sites formed in a given homo-oligomer can increase binding affinity between the protein and its ligand in a phenomenon known as the "multivalent effect" [4-6]. Certain membrane proteins control signal transduction by transient dimerization [7,8]. Consequently, defects in oligomeric states can cause diseases, and control of oligomeric states may provide a means for treating human diseases [9].

Thus, prediction of protein quaternary structures as well as tertiary structures is expected to shed light on biological issues. Over the last two decades, protein tertiary structure prediction techniques from their amino acid sequences have been significantly improved based on the development of novel prediction techniques and enlargement of protein structure databases. A number of websites and tools for protein tertiary structure prediction have been developed. On the other hand, few methods for predicting protein quaternary structures from their amino acid sequences have been reported so far. Baker and co-workers have extended their tertiary structure prediction framework ROSETTA to predict quaternary structures of symmetrical oligomeric proteins. This method is applicable when the symmetry type of a protein is known *a priori* [10]. Chen and Skolnick developed M-TASSER to predict dimeric structures from amino acid sequences [11]. Protein-protein docking tools can also be applied for the prediction of a quaternary structure of a homo-oligomeric protein with or without extensions [12]. These methods are aimed at *de novo* prediction (prediction without templates).

On another front, there are currently many templates for protein structure prediction available from the Protein Data Bank (PDB) [13], and such templates are constantly increasing in number. For tertiary structure prediction, model structures by using template-based modeling are so accurate that they can be applied for molecular replacement in many cases. Recently, Krissinel and Henrick developed PISA (Protein Interactions, Surface, and Assembly) to detect the most probable homo-oligomeric states of proteins solved by X-ray diffraction [14]. The success rate of PISA is 80–90% according to the authors' benchmark results. These advances have made the concept of template-based quaternary structure modeling a practical reality.

In this paper, we describe our template-based protein quaternary (homo-oligomer) structure prediction methods from amino acid sequences, which were tested in the 9th Critical Assessment of Protein Structure Prediction (CASP9) experiment [15]. CASP experiments are the only true blind test for protein tertiary and quaternary structure prediction from amino acid sequences alone, in that various groups predict structures of proteins whose coordinates are not yet publically available. Usually, the performance of a prediction technique is estimated by the prediction results for a set of proteins whose coordinates are publically available. Though information about their structures are of course concealed during the test, there is room for overfitting of prediction parameters, which CASP experiments do not

*correspondence: shugo@bi.a.u-tokyo.ac.jp
[2]Department of Biotechnology, The University of Tokyo, 1-1-1 Yayoi, Bunkyo-ku, Tokyo, Japan.
Full list of Author's information is available at the end of the article

doi: 10.7243/2050-2273-1-1

have in principle. Actually, we often experience that performance of our developed prediction tools for CASP experiments are worse than we expect. So CASP experiments are the most severe tests in the field of protein structure prediction. In CASP9, quaternary structure prediction was assessed for more than 20 targets for prediction by auto-servers and more than 50 targets for manual prediction. Models from 7 auto servers and 16 human groups were submitted. Although there were a few targets for hetero-dimer predictions using sequence information in Critical Assessment of Prediction of Interactions (CAPRI) experiment [16], this is the first systematic assessment of the blind prediction of protein homo-oligomer structures for a number of targets. This requires to predict not only quaternary structures but also oligomeric states.

Our methods based on templates were ranked first among all quaternary structure predictors in both the auto server category and the human prediction category in CASP9 [15]. Our results demonstrate the effectiveness of the concept of template-based protein quaternary structure modeling.

Methods

We have tried 2 approaches for predicting protein quaternary structures: prediction by an auto server and prediction by a manual method utilizing the results of our another server predicting tertiary structures. Protein structures to be predicted (called "targets") in CASP9 were grouped into 2 categories: targets designated only for server prediction (denoted as "server-only" targets) and targets for both human and server prediction (denoted as "human/server" targets). Auto servers were supposed to make predictions for all targets, while human predictors only made predictions for selected human/server targets. Auto servers and human predictors were required to submit their predictions from amino acid sequences within 72 hours and 3 weeks, respectively. For human predictors, the number of chains in the native oligomer was occasionally available (Table S1).

Table S1. All targets in CASP9

Target	Type	Length	Method	PDB	Information about the number of chainsa
T0515	Human/Server	365	X-RAY	3mt1	2
T0516	Server only	229	X-RAY	3no6	4
T0517	Human/Server	159	X-RAY	3pnx	3 or 6
T0518	Server only	288	X-RAY	3nmb	1
T0519	Server only	180	X-RAY	canceled	-
T0520	Human/Server	189	X-RAY	3mr7	-
T0521	Server only	179	X-RAY	3mse	-
T0522	Server only	134	X-RAY	3nrd	2 or 4
T0523	Human/Server	120	X-RAY	3mqo	-
T0524	Server only	325	X-RAY	3mwx	1
T0525	Server only	215	X-RAY	3mqz	-
T0526	Human/Server	290	X-RAY	3nre	1
T0527	Server only	142	X-RAY	3mr0	-
T0528	Server only	388	X-RAY	3n0x	1
T0529	Human/Server	569	X-RAY	3mwt	-
T0530	Server only	115	X-RAY	3npp	2
T0531	Human/Server	65	NMR	2kjx	-
T0532	Server only	506	X-RAY	3mx3	1
T0533	Server only	313	X-RAY	canceled	-
T0534	Human/Server	384	X-RAY	3n8u	-
T0535	Server only	294	X-RAY	canceled	-
T0536	Server only	152	X-RAY	canceled	-
T0537	Human/Server	381	X-RAY	3n6z	4
T0538	Server only	54	NMR	2l09	1
T0539	Server only	81	NMR	2l0b	-
T0540	Human/Server	90	X-RAY	N.A.	-
T0541	Server only	106	NMR	2l0d	-
T0542	Server only	590	X-RAY	3n05	-
T0543	Human/Server	887	X-RAY	2xrg	-
T0544	Human/Server	135	NMR	2l3w	-
T0545	Server only	158	NMR	2l3f	-
T0546	Server only	134	NMR	canceled	-
T0547	Human/Server	611	X-RAY	3nzp	-
T0548	Server only	106	X-RAY	3nnq	-
T0549	Server only	84	NMR	canceled	-
T0550	Human/Server	339	X-RAY	3nqk	-
T0551	Server only	74	NMR	3obh	3 or 2
T0552	Server only	122	NMR	2l3b	-
T0553	Human/Server	141	NMR	2ky4	-
T0554	Server only	135	NMR	canceled	-
T0555	Server only	148	NMR	2l06	-
T0556	Human/Server	73	NMR	canceled	-
T0557	Server only	145	NMR	2kyy	-
T0558	Human/Server	294	X-RAY	3no2	-
T0559	Server only	69	NMR	2l01	1
T0560	Server only	74	NMR	2l02	-
T0561	Human/Server	161	X-RAY	2xse	-
T0562	Human/Server	123	NMR	2kzx	-
T0563	Server only	279	X-RAY	3on7	-
T0564	Human/Server	89	NMR	2l0c	-
T0565	Server only	326	X-RAY	3npf	-
T0566	Human/Server	156	X-RAY	3n72	-
T0567	Server only	145	X-RAY	3n70	2
T0568	Human/Server	158	X-RAY	3n6y	8
T0569	Human/Server	79	NMR	2kyw	-
T0570	Server only	258	X-RAY	3no3	-
T0571	Human/Server	344	X-RAY	3n91	-
T0572	Server only	93	NMR	2kxy	-
T0573	Server only	311	X-RAY	3oox	-
T0574	Human/Server	126	X-RAY	3nrf	-
T0575	Server only	216	X-RAY	3nrg	4
T0576	Human/Server	172	X-RAY	3na2	2
T0577	Server only	116	NMR	canceled	-

doi: 10.7243/2050-2273-1-1

Target	Type	Length	Method	PDB	Information about the number of chainsa
T0578	Human/Server	164	X-RAY	3nat	-
T0579	Human/Server	124	NMR	2ky9	-
T0580	Human/Server	105	X-RAY	3nbm	-
T0581	Human/Server	136	X-RAY	3npd	1
T0582	Human/Server	222	X-RAY	3o14	1
T0583	Server only	152	NMR	canceled	-
T0584	Human/Server	352	X-RAY	3nf2	-
T0585	Server only	234	X-RAY	3ne8	-
T0586	Human/Server	125	X-RAY	3neu	-
T0587	Server only	373	X-RAY	canceled	-
T0588	Human/Server	400	X-RAY	3nfv	-
T0589	Server only	465	X-RAY	3net	1
T0590	Human/Server	137	NMR	2kzw	-
T0591	Server only	406	X-RAY	3nra	-
T0592	Human/Server	144	X-RAY	3nhv	-
T0593	Server only	208	X-RAY	3ngw	-
T0594	Human/Server	140	X-RAY	3ni8	-
T0595	Server only	123	NMR	canceled	-
T0596	Human/Server	213	X-RAY	3ni7	-
T0597	Server only	429	X-RAY	3nie	-
T0598	Human/Server	161	X-RAY	3njc	-
T0599	Server only	399	X-RAY	3os6	-
T0600	Server only	125	X-RAY	3nja	-
T0601	Server only	449	X-RAY	3qtd	-
T0602	Human/Server	123	X-RAY	3nkz	-
T0603	Server only	305	X-RAY	3nkd	-
T0604	Human/Server	549	X-RAY	3nlc	-
T0605	Human/Server	72	X-RAY	3nmd	-
T0606	Human/Server	169	X-RAY	3noh	-
T0607	Server only	471	X-RAY	3pfe	-
T0608	Human/Server	279	X-RAY	3nyy	-
T0609	Server only	340	X-RAY	3os7	-
T0610	Human/Server	186	X-RAY	3ot2	-
T0611	Server only	227	X-RAY	3nnr	-
T0612	Server only	129	X-RAY	3o0l	-
T0613	Server only	287	X-RAY	3obi	-
T0614	Human/Server	135	X-RAY	3nqw	-
T0616	Human/Server	103	X-RAY	3nrt	-
T0617	Server only	148	X-RAY	3nrv	-
T0618	Human/Server	182	X-RAY	3nrh	-
T0619	Human/Server	111	X-RAY	3nrw	-
T0620	Server only	312	X-RAY	3nr8	-
T0621	Human/Server	172	X-RAY	3nkg	-
T0622	Human/Server	138	X-RAY	3nkl	-
T0623	Server only	220	X-RAY	3nkh	-
T0624	Human/Server	81	X-RAY	3nrl	-
T0625	Human/Server	233	X-RAY	3oru	-
T0626	Server only	283	X-RAY	3o1l	-
T0627	Human/Server	261	X-RAY	3oql	-
T0628	Human/Server	295	X-RAY	3nuw	-
T0629	Human/Server	216	X-RAY	2xgf	-
T0630	Human/Server	132	X-RAY	2kyt	-
T0631	Human/Server	303	X-RAY	canceled	-
T0632	Server only	168	X-RAY	3nwz	-
T0633	Human/Server	462	X-RAY	canceled	-
T0634	Server only	140	X-RAY	3n53	-
T0635	Server only	191	X-RAY	3n1u	-
T0636	Server only	336	X-RAY	3p1t	-
T0637	Server only	146	X-RAY	2x3o	-
T0638	Server only	269	X-RAY	3nxh	-
T0639	Server only	128	X-RAY	3nym	-
T0640	Server only	250	X-RAY	3nyw	-
T0641	Server only	296	X-RAY	3nyi	-
T0642	Server only	387	X-RAY	canceled	-
T0643	Human/Server	83	X-RAY	3nzl	-

Auto server:

The tertiary structure (monomer) prediction of a target protein was performed using a combination of template-based modeling and a consensus-based model quality assessment. Template structures were detected by widely used template detection tools: PDB-BLAST [17], FUGUE [18], and HHsearch [19]. Sequence-structure alignments were generated by T-COFFEE [20] and our pairwise alignment tool REALIZE, which employs an environment dependent position-specific gap penalty. About 300~3000 models per target were generated from these alignments using MODELLER [21]. One model was selected according to a consensus-based model quality assessment approach. For this quality assessment, distances between all-by-all model structures were calculated using an S-score [22] per residue and the model with minimum averaged distances was selected based on the heuristics defining the most probable model as being the closest to the center of the generated models [23].

The quaternary structure of the selected monomer model was deduced from those of the template proteins used in tertiary structure modeling. The quaternary structures of the template proteins were obtained from the PISA server [http://pdbe.org/pisa/] [14]. The quaternary structure of the selected model was decided upon by the majority of the quaternary structures of those templates (i.e. voting). Here, only templates with TM-scores [24], a measure of structural similarity (from 0 to 1 and 1 means perfect match), greater than 0.3 with the selected monomer model were used. Finally, the quaternary structure of the target protein was obtained by superimposing the monomer structure of the selected model on the quaternary structure of the template protein with the highest TM-score with respect to the selected model. Structural refinements to avoid collisions of atoms around interfaces and application of distance constraints between subunits were not performed because of the prediction time limitation and our computational resources. Although refinement to avoid collisions is important for precise quaternary structure prediction, we wanted to observe the ability of this rather simple automated prediction to detect interfaces between chains in oligomeric states.

doi: 10.7243/2050-2273-1-1

Manual method

For human predictions, human experts made modifications to our server prediction. Selections of template proteins and quaternary structures were reexamined by human experts with particular attention to select the most probable number of chains in the quaternary structure. When a proper template could not be obtained for building reasonable template-based models, quaternary structure modeling was not tried. For the other cases, templates which were suitable for quaternary structure prediction rather than for tertiary structure prediction were selected here, e.g., templates which have good alignments in the interface regions between different chains were used. The sequence alignments between the target protein and the selected templates were also manually corrected according mainly to position-specific sequence profile, secondary structure, and environment dependent position-specific gap penalty. Sequence alignments in not only hydrophobic core region but also inter-chain interface region were taken more care of. Model building with MODELLER using oligomeric state templates was also performed manually considering to avoid collisions around the interface regions.

Performance evaluation

The performance of our predictions was assessed according to the quaternary structures given in REMARK 350 records of the PDB files. An author-determined biological unit was used when it was provided. Otherwise, a software-determined quaternary structure, which appeared first in the PDB header, was used. We evaluated prediction accuracy according to the consistency of the number of subunits, "contact agreement score (S_{agree})," which was used in the official assessment of CASP9, and overall TM-score calculated by MM-align [25], which we have called "MM-score" in this paper to avoid confusion.

The S_{agree} score represents the structural similarity at interfaces between chains. S_{agree} is defined as fraction of correctly predicted contacts between subunits and is calculated by the following equations:

$$S_{agree} = \frac{\sum_{i,j} f(x_{ij}, y_{ij})}{\sum_{i,j} g(x_{ij}, y_{ij})}$$

$$f(x_{ij}, y_{ij}) = \begin{cases} 1 & if\ x_{ij} = y_{ij} \\ 0 & otherwise \end{cases}$$

$$g(x_{ij}, y_{ij}) = \begin{cases} 1 & if\ x_{ij} = 1\ or\ y_{ij} = 1 \\ 0 & otherwise \end{cases}$$

where residue *i* of 1 subunit and residue *j* of another subunit are regarded as being "in contact" when the inter-subunit distance between their Cβ-atoms is less than 12 Å. x_{ij} and y_{ij} are 1 if residue *i* and *j* are in contact in protein *x* (model) and *y* (native), respectively, and 0 otherwise. $\Sigma f(x_{ij}, y_{ij})$ represents the number of correctly predicted contacts, and $\Sigma g(x_{ij}, y_{ij})$ represents the number of unions of residue pairs in protein *x* and *y*.

The MM-score is a multiple-chain extension of TM-score [24] and calculated according to the following equation:

$$\mathrm{MM-score} = \max\left[\frac{1}{L}\sum_{i=1}^{L_{ali}} \frac{1}{1 + d_{ij}^2 / d_0^2(L)}\right]$$

where *L* is the total length of all chains in the target and L_{ali} is the number of aligned residue pairs between the native and the model structures. d_{ij} is the distance between the Ca atoms of the aligned residues *i* and *j* after superposition of the native structure and the model. $d_0(L)$ is a parameter to correct the effect of chain length:

$$d_0(L) = 1.24\sqrt[3]{L - 15} - 1.8$$

Thus the MM-score represents the overall structural similarity when multiple chains are considered.

The range of the similarity scores of S_{agree} and MM-score is from 0 to 1 where 1 indicates perfect similarity.

Results

Targets for performance evaluation

Our methods were tested in the CASP9 experiment. For all 129 targets, 60 were human/server targets and 69 were server-only targets. From these, we excluded the following targets for our analyses: (1) structures with coordinates not released on April 7, 2011; (2) PDB structures with insertions in locations other than the N- and C- termini compared to the corresponding CASP target sequence; (3) structures with coordinates obtained by methods other than X-ray diffraction analysis; (4) targets determined as monomers from PDB headers of the native structure (described below); and (5) the models that were submitted as monomer by our group. Finally, we have 16 targets in human/server category and 58 targets in both human/server and server-only categories for our analyses.

It is notable that we submitted our models as monomer in cases when we predicted the target as monomer; we could not build reliable single-chain model for the target; CASP organizers gave us information that the structure of the target was solved by methods other than X-ray diffraction analysis; or we could not obtain enough information about the quaternary structure of the target from our template search.

Summary of the prediction results

Although the predictors in CASP9 could submit up to 5 models with ranking for each target, only the first model (i.e. the model the predictors believed to be the most probable) was considered in the official assessment. We mention only the first model in this paper.

Table 1 shows the summary of our predictions provided by the auto server and by the manual method. Figure 1 shows examples of our models and the native structures. Note that human/server targets were more difficult than server-only targets because CASP organizers tended to avoid assigning easy targets for human prediction.

doi: 10.7243/2050-2273-1-1

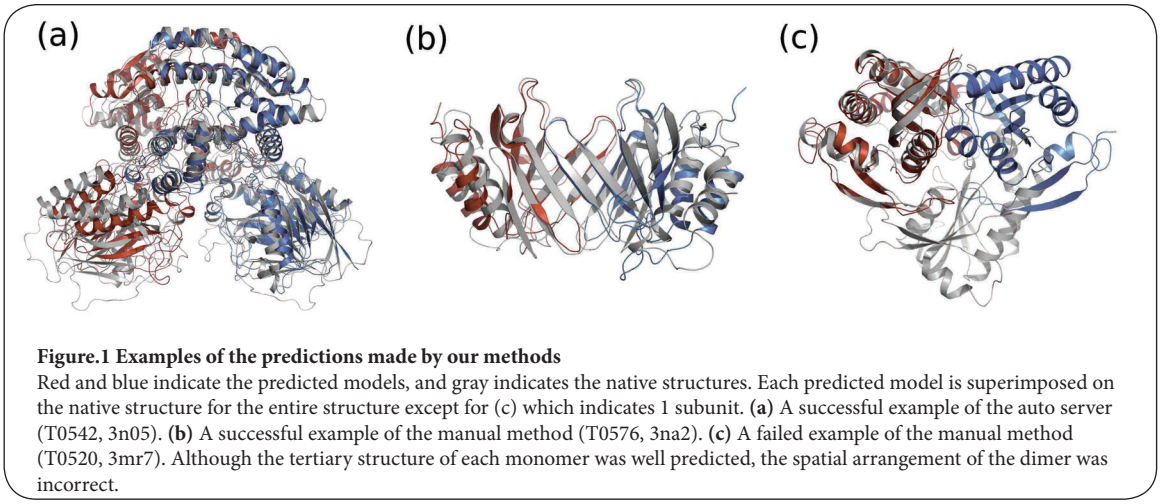

Figure.1 Examples of the predictions made by our methods
Red and blue indicate the predicted models, and gray indicates the native structures. Each predicted model is superimposed on the native structure for the entire structure except for (c) which indicates 1 subunit. **(a)** A successful example of the auto server (T0542, 3n05). **(b)** A successful example of the manual method (T0576, 3na2). **(c)** A failed example of the manual method (T0520, 3mr7). Although the tertiary structure of each monomer was well predicted, the spatial arrangement of the dimer was incorrect.

Table.1 Prediction performance summary

Target type	Prediction method	Targets	Number of chain matches	Successful predictions[a]	S_{agree} Avg (Stdev)	MM-score Avg (Stdev)
Server-only	Auto server	58	46	14	0.145 (0.203)	0.468 (0.234)
Human/ Server	Auto server	25	19	4	0.106 (0.175)	0.406 (0.216)
Human/ Server	Manual method	16	15	11	0.358 (0.256)	0.635 (0.179)

a $S_{agree} \geq 0.25$ and MM-score ≥ 0.5

With the auto server, we submitted our first models as oligomers for 58 targets in human/server and server-only categories. Out of them, 46 models had the correct number of chains and 14 models had correct quaternary structures. Here, we determined that the quaternary structure of a model is "correct" when $S_{agree} \geq 0.25$ and MM-score ≥ 0.5 according to the visual inspection, as described below. For prediction of human/server targets, we submitted our first models as oligomers for 25 targets. Of these, 19 models had the correct number of chains and 4 models had correct quaternary structures. The best server-only model (T0522) satisfies $S_{agree} \geq 0.7$ and MM-scores ≥ 0.8, showing that our auto server could generate rather accurate quaternary structure models under this completely blind test.

With the manual method, we submitted our first models as oligomers for 16 targets in human/server category. Out of them, 15 models had the correct number of chains and 11 models had correct quaternary structures. Among them, 2 models had $S_{agree} \geq 0.7$ and MM-score ≥ 0.8.

There were 15 multimer targets for which both our auto server and manual prediction submitted first models as multimers. To assess the effect of voting to determine the number of chains in our auto server predictions, the predicted number of chains for these 15 targets are listed in Table 2. Fourth column of Table 2 shows the number of chains predicted by the number of top hit templates of HHsearch (quaternary structures were determined by PISA) for comparison. Among these 15 targets, 8 were predicted correctly. On the other hand, the numbers of correctly predicted targets by our auto server and manual method were 11 and 14, respectively. These results show the effectiveness of our approaches. For these targets, the predicted number of chains, S_{agree}, and MM-score of the models from auto server prediction and manual prediction are also listed in Table 3. Among the 15 targets, models from our manual prediction outperform those from the auto server for 11 targets with respect to S_{agree} and 12 targets with respect to the MM-score. Figure 2 shows the plot for comparison of MM-score for the auto server and manual prediction per target.

Examples of the prediction models

Figure 1 shows examples of the submitted models.

T0542 is a server-only target (Figure 1(a)). The PDB ID of the native structure is 3n05. The protein is a 2-mer with 590 amino acids per chain. The top hit of the HHsearch fold recognition tool in our server was 3dla (8-mer). However, our server discarded this hit because the result of the prediction of the number of chains by voting

doi: 10.7243/2050-2273-1-1

Table.2 Comparison of the predicted number of chains

Target	Answer		Predicted with HH-search's top hit	Predicted by our methods	
	PISA	PDBa		Auto server	Manual method
T0515	2	**2**	2	2	2
T0517	6	**3**	12	2	3
T0520	2	**2**	4	2	2
T0523	4	**2**	12	2	2
T0547	2	**2**	4	2	2
T0576	2	**2**	1	2	2
T0584	2	**2**	2	2	2
T0586	2	**2**	10	2	2
T0592	3	**3**	3	2	3
T0596	2	**2**	2	2	2
T0602	4	**4**[b]	4	2	2
T0605	2	**2**	3	2	2
T0625	2	**2**	2	2	2
T0627	4	**4**	4	4	4
T0629	3	**3**	3	2	3
Correct			8	11	14

[a]The quaternary structures from PDB entries were used for assessment in this study (see Methods section).

[b]Authors of PDB entry (3nkz) wrote about the probable quaternary structure in the REMARK 300 record: "Experimentally unknown. The chains A and B, C, and D likely form dimers, respectively." When we performed the annotation with dimer, we obtained the scores as follows. S_{agree}: 0.006 and MM-score: 0.305 for the auto server, and S_{agree}: 0.150 and MM-score: 0.429 for the manual method.

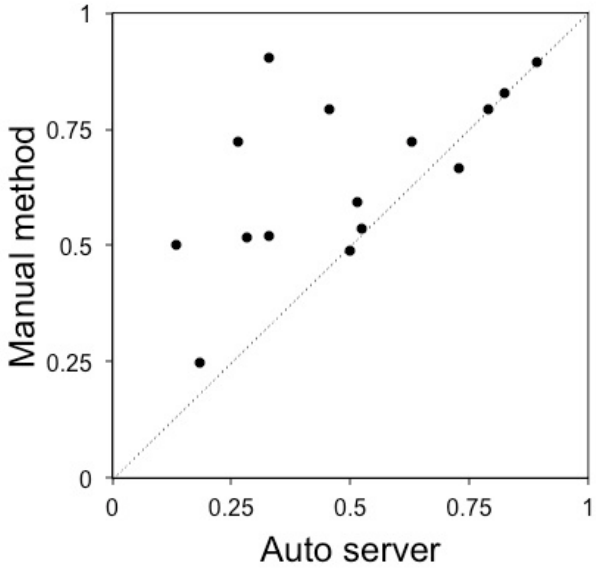

Figure.2 Auto server vs. manual method
Comparison of MM-score of the models by manual method and auto server prediction.

indicated that the most probable oligomeric state for this target was a 2-mer, and selected 3ilv (2-mer) as the template instead. This led to the correct prediction of both the monomer and the quaternary structure of this target, with an S_{agree} of 0.298 and an MM-score of 0.856. This is an example that shows the framework of our server prediction works well.

T0576 is a human/server target (Figure 1(b)). The PDB ID of the native structure is 3na2. The protein is a 2-mer with 172 amino acids per chain. The two chains interact with each other at their beta sheets and if quaternary structure is not formed, hydrophobic residues of these regions will be exposed. So quaternary structure is essential for this protein. The original sequence of this protein included a short tag sequence at the N-terminus and the effects of this tag led to false hits in sequence based template search methods, e.g. 2grg or 1z1s by HHsearch. Prediction by our auto server, which chose 1u83 as the template, also failed because of the same reason (S_{agree} and the MM-score were 0.004 and 0.265, respectively). On the other hand, in our manual prediction, we eliminated the tag region from query of HHsearch. The top hit provided by this improved HHsearch was 3fm2 (2-mer) and the second hit was 2hqv (4-mer). We found that the corresponding residues of the interface of 2hqv were probably missing in this target according to the alignment of 2hqv and the target, so we selected 3fm2 to generate the quaternary structure of this target. S_{agree} and the MM-score of our model have rather high values of 0.826 and 0.724, respectively. This shows an example of the effectiveness of manual inspection for template-based quaternary structure prediction.

T0520 is an example of a failure of manual prediction. T0520 is a human/server target (Figure 1(c)). The PDB ID of the native structure is 3mr7. The protein is a 2-mer with 189 amino acids per chain. There are several templates, including 1wc3, 1ybt, and 2w01, which have structures similar to the first model of our auto server prediction, which we originally considered as the most probable monomer structure. This prediction was accurate. Among these templates, the spacial arrangement of chains 2w01 and 1wc3 are similar, so we modeled our quaternary structure based on these 2 templates. The native structure, however, has a chain orientation that is different from the chain orientation in all templates in spite of their similarity as monomers. The TM-score between the native structure and our model provided by human prediction was 0.810, but the S_{agree} was 0.002 and the MM-score was 0.490.

Discussion

Server prediction versus manual prediction

The ratios of the correctly predicted number of chains for human/server targets were 19/25 (0.76) and 15/16 (0.94) for auto server prediction and manual prediction, respectively (Table 3). The average of S_{agree} for the auto server and manual prediction were 0.167 and 0.378, and that of the MM-scores were 0.492 and 0.648, respectively. In our predictions, manual prediction outperforms auto server prediction for all of these measures. This indicates that human intervention, i.e.

doi: 10.7243/2050-2273-1-1

Table.3 - Comparison of prediction performance between auto server and manual method

	Auto server		Manual method	
Target	S_{agree}	MM-score	S_{agree}	MM-score
T0515	0.558	0.893	0.533	0.893
T0517	0.000	0.330	0.296	0.521
T0520	0.000	0.499	0.002	0.490
T0523	0.096	0.514	0.347	0.592
T0547	0.112	0.729	0.490	0.665
T0576	0.004	0.265	0.826	0.724
T0584	0.523	0.825	0.525	0.828
T0586	0.380	0.456	0.444	0.794
T0592	0.009	0.282	0.364	0.516
T0596	0.060	0.523	0.055	0.536
T0602	0.003	0.182	0.082	0.248
T0605	0.000	0.330	0.906	0.904
T0625	0.349	0.791	0.310	0.792
T0627	0.394	0.630	0.355	0.724
T0629	0.010	0.133	0.137	0.500
Avg	0.167	0.492	0.378	0.648

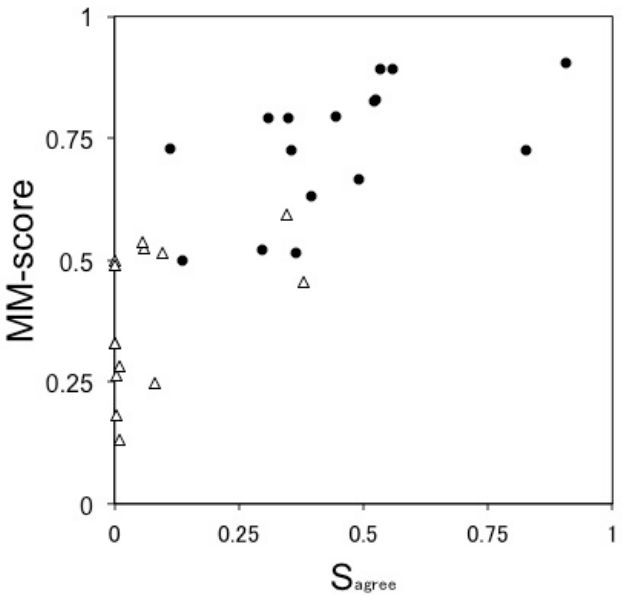

Figure.3 Scatter plot between Sagree and MM-score for the prediction targets in Table 3.
The closed circles are targets that we regarded as correct by visual inspection, and the open triangles are others. The correlation coefficient for all points was 0.787.

more accurate chain number predictions, re-selection of templates, re-alignment between sequences of targets and templates, and avoiding collisions could considerably improve the prediction performance in template-based quaternary structure prediction, as described in the example of T0576 above.

S_{agree} and MM-score

In this paper, we used 2 scores (S_{agree} and MM-score; S_{agree} was used in the official assessment of CASP9) to provide quality assessments of the predicted models. Since an assessment method for quaternary structure prediction has not been established, we discuss these 2 scoring measures.

Figure 3 is a scatter plot between S_{agree} and MM-score for the data in Table 3. These were moderately correlated with each other, and the correlation coefficient was 0.787. The closed circles in Figure 3 are the targets judged by visual inspection to have correctly predicted quaternary structures. According to this judgment, despite some exceptions, predicted models with an S_{agree} value over 0.25 and an MM-score over 0.5 could be considered as good models. Note that a score called IS-score [26] that can consider significance of similarity between target and model structures compared to random pairs was developed recently by the same group as MM-score. They reported that all near native docking models in CAPRI have an IS-score above 0.17. According to this threshold, two models (T0517 and T0592) among our "good" models in Table 3 have IS-scores below 0.17 and two models (T0602, T0629) among our "not good" models in Table 3 have IS-scores above 0.17. However, classification of "good" and "not good" for other targets was the same, and we found that IS-score is correlated with S_{agree} and also MM-score (correlation coefficients are 0.89 between S_{agree} and IS-score and 0.73 between MM-score and IS-score; data not shown). This also supports that our criteria of "good models" are appropriate.

There are 3 types of prediction failures:

(1) incorrect number of chains,

(2) incorrect arrangement of subunits, and

(3) low quality of each monomer of the model.

(1) When the number of chains is incorrect, S_{agree} is extremely low while the MM-score appears to be reasonably low. This can be seen by comparing predictions between the auto server and the manual method for T0517 and T0592.

(2) When the spatial arrangement of subunits is incorrect, S_{agree} is extremely low but the MM-score is not significantly low, as we can see for T0520 described above and also for T0596. To improve the prediction for these targets, a detailed assessment of the interfaces in a model is needed.

(3) When a portion of each monomer is not well modeled and the portion is located near the subunit interface, both S_{agree} and MM-score tend to be low. Predicted models of T0517, T0602, and T0629 by the manual method are examples of this scenario.

Conclusions

We have discussed the overall performance of our local tools, successful examples as well as unsuccessful examples, and current challenges. Our quaternary structure predictions in CASP9 shows

the effectiveness of the template-based method in this field. As far as we know, this is the first systematic performance assessment of completely blind protein quaternary structure prediction from its amino acid sequence. We have demonstrated that simple auto server prediction could generate successful models for 14 out of 58 targets. Human experts could generate better models than our auto server through more accurate chain number predictions, re-selection of templates, re-alignment between sequences of targets and templates, and avoiding collisions. We have also shown in several cases that good monomer templates could provide different chain orientations, leading to poor quality of predicted quaternary structures based on templates. Our template-based quaternary structure prediction methods are simple and general and therefore can be expanded further with various options or combinations with other methods. Our results should be useful for improvement of the accuracy of template-based quaternary structure prediction.

Authors' contributions

MM, MK, and SN have developed the protein tertiary and quaternary structure prediction server. MM and MK performed manual prediction of protein quaternary structures, and SN carried out result analysis. KS helped to draft the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank the CASP organizers and assessors for their contributions and hard work, and the experimental groups for providing their protein structures.

Competing Interests

The Authors declare that they have no competing interests.

Author information

[1]Department of Fundamental Research, National Institute of Biomedical Innovation, 7-6-8 Saito-asagi, Ibaraki-shi, Osaka, Japan.

Article history

Received: 17-Jan-2012 Revised: 09-Feb-2012
Accepted: 14-Mar-2012 Published: 20-Apr-2012

References

1. Kuhner S, van Noort V, Betts MJ, Leo-Macias A, Batisse C, et al: **Proteome organization in a genome-reduced bacterium.** *Science* 2009, **326:**1235-1240. | Article | Pubmed
2. Levy ED, Boeri Erba E, Robinson CV, Teichmann SA: **Assembly reflects evolution of protein complexes.** *Nature* 2008, **453:**1262-1265. | Article | Pubmed Abstract | Pubmed Full text
3. Marianayagam NJ, Sunde M, Matthews JM: **The power of two: protein dimerization in biology.** *Trends Biochem Sci* 2004, **29:**618-625. | Article | Pubmed
4. Lee RT, Lee YC: **Affinity enhancement by multivalent lectin-carbohydrate interaction.** *Glycoconj J* 2000, **17:**543-551. | Article | Pubmed
5. Pluckthun A, Pack P: **New protein engineering approaches to multivalent and bispecific antibody fragments.** *Immunotechnology* 1997, **3:**83-105. | Pubmed
6. Weis WI, Drickamer K: **Structural basis of lectin-carbohydrate recognition.** *Annu Rev Biochem* 1996, **65:**441-473. | Article | Pubmed
7. Heldin CH: **Dimerization of cell surface receptors in signal transduction.** *Cell* 1995, **80:**213-223. | Article | Pubmed
8. Stock J: **Receptor signaling: dimerization and beyond.** *Curr Biol* 1996, **6:**825-827. | Article | Pubmed
9. Elrod KC, Moore WR, Abraham WM, Tanaka RD: **Lactoferrin, a potent tryptase inhibitor, abolishes late-phase airway responses in allergic sheep.** *Am J Respir Crit Care Med* 1997, **156:**375-381. | Article | Pubmed
10. Andre I, Bradley P, Wang C, Baker D: **Prediction of the structure of symmetrical protein assemblies.** *Proc Natl Acad Sci U S A* 2007, **104:**17656-17661. | Article | Pubmed Abstract | Pubmed Full text
11. Chen H, Skolnick J: **M-TASSER: an algorithm for protein quaternary structure prediction.** *Biophys J* 2008, **94:**918-928. | Article | Pubmed Abstract | Pubmed Full text
12. Diemand AV, Lupas AN: **Modeling AAA+ ring complexes from monomeric structures.** *J Struct Biol* 2006, **156:**230-243. | Article | Pubmed
13. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al: **The Protein Data Bank.** *Nucleic Acids Res* 2000, **28:**235-242. | Article | Pubmed Abstract | Pubmed Full text
14. Krissinel E, Henrick K: **Inference of macromolecular assemblies from crystalline state.** *J Mol Biol* 2007, **372:**774-797. | Article | Pubmed
15. Mariani V, Kiefer F, Schmidt T, Haas J, Schwede T: **Assessment of template based protein structure predictions in CASP9.** *Proteins* 2011, **79 Suppl 10:**37-58. | Article | Pubmed
16. **CAPRI: Critical Assessment of PRediction of Interactions** | Online |
17. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402. | Article | Pubmed Abstract | Pubmed Full text
18. Shi J, Blundell TL, Mizuguchi K: **FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties.** *J Mol Biol* 2001, **310:**243-257. | Article | Pubmed
19. Soding J: **Protein homology detection by HMM-HMM comparison.** *Bioinformatics* 2005, **21:**951-960. | Article | Pubmed
20. Notredame C, Higgins DG, Heringa J: **T-Coffee: A novel method for fast and accurate multiple sequence alignment.** *J Mol Biol* 2000, **302:**205-217. | Article | Pubmed
21. Sali A, Blundell TL: **Comparative protein modelling by satisfaction of spatial restraints.** *J Mol Biol* 1993, **234:**779-815. | Article | Pubmed
22. Fasnacht M, Zhu J, Honig B: **Local quality assessment in homology models using statistical potentials and support vector machines.** *Protein Sci* 2007, **16:**1557-1568. | Article | Pubmed Abstract | Pubmed Full text
23. Cozzetto D, Kryshtafovych A, Tramontano A: **Evaluation of CASP8 model quality predictions.** *Proteins* 2009, **77 Suppl 9:**157-166. | Article | Pubmed
24. Zhang Y, Skolnick J: **Scoring function for automated assessment of protein structure template quality.** *Proteins* 2004, **57:**702-710. | Article | Pubmed
25. Mukherjee S, Zhang Y: **MM-align: a quick algorithm for aligning multiple-chain protein complex structures using iterative dynamic programming.** *Nucleic Acids Res* 2009, **37:**e83. | Article | Pubmed Abstract | Pubmed Full text
26. Gao M, Skolnick J: **New benchmark metrics for protein-protein docking methods.** *Proteins* 2011, **79:**1623-1634. | Article | Pubmed