Modeling protein complexes with BiGGER

3 AUTHORS:

Ludwig Krippahl
New University of Lisbon
38 PUBLICATIONS **589** CITATIONS

José J G Moura
New University of Lisbon
572 PUBLICATIONS **11,876** CITATIONS

P. Nuno Palma
BIAL
45 PUBLICATIONS **953** CITATIONS

Modeling Protein Complexes with BiGGER

Ludwig Krippahl[1], José J. Moura[1], P. Nuno Palma[1,2]

1- REQUIMTE, Departamento de Química, Centro de Química Fina e Biotecnologia, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa.

2- Current address: Department of Research and Development, BIAL, 4785 S. Mamede do Coronado, Portugal.

Prefered contact:

Ludwig Krippahl: ludi@dq.fct.unl.pt

Aditional email contact

Ludwig Krippahl: ludik@netcabo.pt

Correspondence adress

José J. Moura,

Departamento de Química, Centro de Química Fina e Biotecnologia,

Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa,

2825-114 Monte de Caparica, Portugal

Abstract

This paper describes the method and results of our participation in the Critical Assessment of PRediction of Interactions (CAPRI) experiment, using the protein docking program BiGGER (Bimolecular complex Generation with Global Evaluation and Ranking) [1]. Of five target complexes (CAPRI targets 2, 4, 5, 6, and 7), only one was successfully predicted (target 6), but BiGGER generated reasonable models for targets 4, 5 and 7, which could have been identified if additional biochemical information had been available.

Keywords: docking, modelling, complex prediction, protein interaction.

Introduction

All models we submitted to CAPRI were generated by BiGGER [1]. A detailed explanation of the algorithm is outside the scope of this paper, but we'll give a short description of its function and purpose, as a background for our results and conclusions.

The most basic data needed to model a protein complex are the structures of the proteins involved, but more information is usually available. Our goal for BiGGER is a system that can be used when only the structures of the partners are known, but also use whatever information is available to help pinpoint an accurate model.

To this end, BiGGER is designed as a series of filters that exclude possibilities that don't agree with the experimental data or theoretical assumptions. Candidate models are generated by rotating and translating one molecule (the probe) relative to its partner (the target). Rotations are in steps of 15º (6384 orientations, after removing degenerates), and for each orientation translations are searched in 1Å steps. As an illustration, for a cubic search 150Å wide (approximately the search space for the CAPRI complexes), this amounts to 20 billion possible configurations. Fortunately, this search space can be pruned efficiently by checking for excessive overlap and ignoring regions with too little surface contact.

Thus the first filter is geometric complementarity — the geometric fit between the partners is an important feature in most complex modeling algorithms because of its efficiency and reliability. The surface of each partner is represented as a binary (0,1) grid at 1Å resolution, and surface contact is estimated by counting the superposition of surface cells (those marked with 1 in each grid). The core region of each partner is represented as another binary matrix, and superposition of these core regions indicates excessive overlap and is forbidden.

A candidate model is rejected if it doesn't have a higher surface contact score than the lowest ranking of the models retained at that point. If it has a higher score it replaces the lowest ranking model if it passes the second filter, which is a neural network that evaluates the amino acid contacts between the partners. This network was trained, with a large number of examples, to distinguish between correct and incorrect complexes. Five thousand candidate models are kept by default.

After the search and filtering stage, BiGGER estimates the likelihood of each candidate model using a larger neural network and an evaluation of hydration factors, electrostatics and probabilities of side chain contacts, in addition to the amino acid contacts mentioned above. Both neural networks were trained on a set of 87 known complexes, and tested on 20 other known complexes. In approximately half the cases an accurate model can be found among the 20 models with the highest likelihood estimate, and in most cases it can be found within the first few hundred.

But additional information can help narrow the search. Chemera, the application used to display and manipulate BiGGER results, can score the models by evaluating contacts or distances between atoms or residues. The parameters are flexible enough to model biochemical information from Nuclear Magnetic Resonance (NMR), mutagenesis cross-linking data, or even more general considerations such as conserved regions, surface charge distributions or active sites.

For CAPRI we were to submit five models for each target[1]. Since we work almost exclusively with electron transfer complexes, where BiGGER has demonstrated its usefulness [2, 3, 4, 5, 6, 7], we had neither readily available data nor familiarity with the target proteins that could help us select the most likely candidates. Under these circumstances, our experience indicated that BiGGER often failed to pinpoint the correct models in such a small group as we would submit. Nevertheless, it was important to see how BiGGER would fare in a blind test under such unfavorable conditions, and also how it would compare with other prediction methods.

Though we only succeeded in modeling one of the five CAPRI targets we tried to predict, these results seem to be more due to the difficulty of the task than to shortcomings in our method. We intend to show that BiGGER could suggest useful models in a setting where some experimental data was available.

Method

All models submitted to CAPRI round 2 were generated using the algorithms implemented in Chemera 2.0 (http://www.cqfb.fct.unl.pt/bioin/chemera/)

Targets 4 through 6 were modelled using the default parameters; target 7 was modelled with the hard docking option. Chemera 2.0, the version available at the time of writing this paper, does not allow the user to set the hard docking option, but we plan to make Chemera 3.0 available soon, and this version will allow the user to choose hard or soft docking. Otherwise, researchers interested in replicating our target 7 models can contact the authors to obtain the correct version.

In addition to using hard docking, the T cell receptor protein in target 7 was truncated: all residues up to D118 were used; those from L119 onward were discarded during docking. This was done because the V shape of this structure was favoring contact at the inner portion of the hinge, in detriment to the alternatives. The models submitted to CAPRI were obtained by fitting the complete structure with the fragment used in docking. This was the only case where additional information was used to generate the models, since we assumed that the region from L119 onward was not important for docking.

Round 1 of CAPRI closed while we were preparing Chemera 2.0 for release, so we were only able to submit models for target 2, and these results would be hard to replicate because we did not keep all intermediate versions of the software prior to release. However, a rerun with the current version of BiGGER shows only slight differences.

The procedure was to run BiGGER with the two docking partners, choosing the largest partner as the target, the other as the probe. After the docking run, models were

---

[1] A note on nomenclature: CAPRI documentation uses "target" to refer to the complexes to model. Unfortunately, in the documentation on BiGGER and Chemera you will find "target" referring to the larger partner in the docking complex. In this paper we will be consistent with both usages; unless the meaning is clear through context, when "target" is followed by a number (*e.g.* target 7) it refers to the CAPRI complexes, otherwise it means the largest of the docking partners.

clustered using a 3Å cutoff. The clustering algorithm calculates the distance between the corresponding probe atoms in two models, considering the target fixed. If the root of the means of the squares of these distances (rmsd) is below the cutoff value (3Å in this case) the two models will belong to the same cluster.

Though many clusters contain only one model, in some cases several similar solutions are grouped together, which helps the analysis. From this point onward, when referring to our models we mean a representative model of a 3Å cluster that may contain more than one solution.

Models were sorted according to the global evaluation score that BiGGER calculates. The highest ranking five were selected, except when, upon visual inspection, a candidate was found too similar to another on the list of five candidates (such as having a similar placement or a slightly different placement but similar interface) or was an isolated structure very different from any other high ranking models (this is often an indication of a spurious model). The objective was to submit a representative sample of the highest-ranking models, but with a limit of five models from a choice of, potentially, several hundred, we could not be confident we were submitting the right models.

After the CAPRI target structures were made available, we calculated the rmsd from each model generated by BiGGER to the X-Ray structure of the complex. Note that this differs from the rmsd values used in the clustering calculations by fitting the whole complex instead of considering one partner to be fixed. This allowed us to determine if there were acceptable solutions in the set retained by BiGGER, and how high they ranked in the evaluation score.

Finally, we randomly chose 4 amino acids from the interface region, 3 from the target and 1 from the probe, to estimate if additional information would allow us to improve our results. These amino acids were selected by visual inspection of the X-Ray structure, and may not always correspond to the amino acids listed on the contacts files supplied by CAPRI. Table 1 shows the amino acids selected for each CAPRI target and identifies the partners used as target and probe.

Two scores were then calculated for each model:

1. The number of atoms, in the 3 amino acids chosen on the target, that were within 5Å of any atom on the probe.

2. The number of atoms, in the 3 amino acids chosen on the target, that were within 5Å of any atom on the probe plus the number of atoms, in the amino acid chosen on the probe, that were within 5Å of any atom on the target.

These scores are intended to simulate the effects of experimental data on the quality of the models. This is the sort of scoring that could be used if some amino acids were known to be important for complex formation (*e.g.* by site directed mutagenesis), or present at the interface (*e.g.* by NMR data).

[Table 1]

Results

Table 2 summarizes the results. The first two columns after the target identification show the ranking and rmsd to the X-Ray structure of the highest-ranking acceptable model. By acceptable we mean with an rmsd below 2.5Å or the lowest rmsd if no model meets this criterion.

[Table 2]

The next two columns indicate the rank and rmsd of the model with the lowest rmsd value within the top 5 models, when sorted according to the contact counts for the 3 amino acids selected on the target. The last two columns show the same information, but using all 4 amino acid residues (3 on the target, one on the probe).

In some cases more than five models were at the top ranking, all with the same number of contacts; in these, the number of models having the highest score is indicated in parentheses right after the rank for the best model.

Target 2

As Table 2 shows, the closest model to the X-Ray structure for target 2 has an rmsd of 10Å. This is not an acceptable model at all, and, in this case, additional information could not improve these results because there was no good model in the set of 5000 structures kept by BiGGER.

Target 4

Of three clusters of models with less than 2.5Å of rmsd from the X-Ray structure (ten individual structures in total), the one with the highest global score was in position 1150. The highest scoring models all had rmsd values over 10Å, so in this case it would not be possible to identify an accurate model without additional information.

An acceptable model, with an rmsd of 1.2Å, was found among the 22 models tied for the first place with 25 contacts, when scoring the contacts with the 3 amino acids chosen on the target (see Table 1). When all four amino acids were considered, there were 15 models with 49 contacts each, of which the closest to the X-Ray structure had an rmsd of 3.3Å (see Table 2).

The positions of the probe in these three models are shown on Figure 1, superimposed on the structure of CAPRI target 4 (in thicker lines).

[Figure1]

Target 5

None of the 5000 individual structures generated by BiGGER had an rmsd of less than 2.5Å from the X-Ray structure, the closest being at 3.4Å, ranked in position 1374 by the global score. The ranking of this model improved to 49 when sorting according to the contact scores for the 3 amino acids chosen on the target, and then to 43 when considering all 4 amino acids (see Table 1).

Even when sorting according to the contact scores, the high ranking models had large rmsd values: 6.6Å for the first position (18 models tied with 25 contacts) using 3 amino acids, 5.4 in the second place (the best of the first five, with 40 contacts) when using all 4 amino acids. These three models are shown on Figure 2, superimposed on the structure of CAPRI target 5.

[Figure 2]

Target 6

A model with 2.1Å rmsd from the X-Ray structure was ranked in eighth place by the global scoring function. This was included in our submission to CAPRI as the fifth model, because we rejected 3 higher ranking models due to their similarity to others in the submitted set.

Good results are the hardest to improve, and adding the contact information for the 3 amino acids on the target (see Table 1) promoted a worse model to the top rank (3.3Å rmsd), and among a total of 63 models with the same score of 41 contacts. Adding the contact score for R51, on the probe, reduced this to 6 models tied for the first rank, with the 3.3Å rmsd model still the best among them, but with a good model (1.0Å msd) at position 15. The three models reported on Table 2 are depicted in Figure 3, superimposed on the structure of CAPRI target 6.

[Figure 3]

Target 7

A good model (1.9Å rmsd) was ranked on position 21 just using the score calculated by BiGGER, but this wasn't good enough to place it in the 5 candidates submitted.

Using the contact scores for the 3 amino acids on the target promoted another good model (1.8Å rmsd) to third position, and adding the contact score for the probe amino acid brought it up to second position. These models are shown on Figure 4, superimposed on the structure of CAPRI target 7. Note that the chain of the probe (T Cell Receptor protein) is shown only up to D118, because only this fragment was used in this docking.

[Figure 4]

Discussion

We propose that, with some additional information on the residues or regions critical for docking, our method should have provided reasonable models for the four round 2 targets.

We cannot claim this from our results alone; picking 4 amino acids from the interface region is not a good substitute for experimental data. One alternative would be to do a large number of combinations and a more extensive analysis, but this would be inappropriate because we already know the correct structures, and the more data we interpret, the greater the risk of bias in the interpretation.

Our confidence in this claim comes from the agreement between these results and those obtained using NMR data [4, 5, 6]; all these experiments show that BiGGER can use this information to pinpoint the correct structures.

The exception is in cases like target 2, where the initial filtering stage eliminates all good models. When this happens, it's impossible to recover them. However, it's possible to identify these situations, not only when no model is consistent with experimental data, but sometimes from the overall pattern formed by the models retained when no such information is available.

These considerations led us to model target 7 with hard docking and using only part of the T Cell receptor protein as the probe – even without data on the complex, we could tell there were problems with the first docking runs on this target. It's not guaranteed that such problems can be identified and corrected by changing the parameters on the docking run, but, in our experience, this is often the case.

Overall, we were pleased with our performance on CAPRI. Target 2 was a complete miss when we consider the whole complex, like we did in our analysis on this paper by using rmsd values. But our scores on the interface region of the FAB protein were quite good, with up to 25 of the 27 interface amino acids correctly predicted.

The models submitted for targets 4, 5, and 7 were far from the real structure, but the problem here was in selecting the right models from the ones BiGGER retained, and we think we showed that this could be done with even a small amount of information.

Target 6 was as successful as we can expect using our method alone, without subsequent refinement by energy minimization or molecular dynamics. Due to the way BiGGER accounts for side chain rearrangement by ignoring some clashes, the models we submitted tended to have more close contacts than those of most other groups. Also, since the search is done in 15º steps for rotation, only chance would give us a very low theta angle score.

Our pattern of successes and failures closely matched that of the majority of the participants. It seems that it's the problem itself that is hard – trying to model a complex without biochemical data – and not our method that is at fault. This also favors the idea of combining data with prediction as closely as possible. To this end we are currently developing Chemera 3.0, which will include a constrained docking version of BiGGER

that will be able to use contact information to restrict the search for models, while searching through combinations of the constraints given to account for uncertainty in the data.

We would like to thank the organizers of CAPRI for this opportunity, and for their remarkable work and rigor.

References

1. Palma, P.N., Krippahl, L., Wampler, J.E., Moura, J.J. (2000). BiGGER: a new (soft) docking algorithm for predicting protein interactions. Proteins 39:372-384.

2. Pettigrew, G.W., Gilmour, R., Goodhew, C.F., Hunter, D.J., Devreese, B., Van Beeumen, J., Costa, C., Prazeres, S., Krippahl, L., Palma, P.N., Moura, I., Moura, J.J. (1998). The surface-charge asymmetry and dimerisation of cytochrome c550 from Paracoccus denitrificans-implications for the interaction with cytochrome c peroxidase. Eur J Biochem. 258:559-566.

3. Pettigrew, G.W., Prazeres, S., Costa, C., Palma, P.N., Krippahl, L., Moura, I., Moura, J.J. (1999). The structure of an electron transfer complex containing a cytochrome c and a peroxidase. J Biol Chem. 274:11383-11389.

4. Morelli X., Dolla, A., Czjzek, M., Palma, P.N., Blasco, F., Krippahl, L., Moura, J.J., Guerlesquin, F. (2000), Heteronuclear NMR and soft docking: an experimental approach for a structural model of the cytochrome c553-ferredoxin complex. Biochemistry 39:2530-2537.

5. Morelli, X., Czjzek, M., Hatchikian, C.E., Bornet, O., Fontecilla-Camps, J. C., Palma, N. P., Moura, J. J., and Guerlesquin, F. (2000) Structural model of the Fe-hydrogenase/cytochrome c553 complex combining transverse relaxation-optimized spectroscopy experiments and soft docking calculations. Journal of Biological Chemistry 275, 23204-23210.

6. Morelli, X.J., Palma, P.N., Guerlesquin, F., Rigby, A.C. (2001), A novel approach for assessing macromolecular complexes combining soft-docking calculations with NMR data. Protein Sci. 10:2131-2137.

7. Crowley, P.B., Rabe, K.S., Worrall, J.A.R., Canters, G.W. and Ubbink, M. (2002), Complex of Cytochrome f and Cytochrome c: Identification of a Second Binding Site and Competition for Plastocyanin Binding. ChemBioChem 3:526-533

8. Mathieu, M., Petitpas, I., Navaza, J., Lepault, J., Kohli, E., Pothier, P., Prasad, B. V. V., Cohen, J., Rey, F. A. Atomic structure of the major capsid protein of rotavirus: implications for the architecture of the virion. EMBO J. 2001 Apr 2;20(7):1485-97.

9. Desmyter A, Spinelli S, Payan F, Lauwereys M, Wyns L, Muyldermans S, Cambillau C. Three camelid VHH domains in complex with porcine pancreatic a-amylase. J. Biol. Chem. 2002 277:23645-50

10. Sundberg EJ, Li H, Llera AS, McCormick JK, Tormo J, Schlievert PM, Karjalainen K, Mariuzza RA. Streptococcal superantigens bound to TCR_ chains reveal diversity in the architecture of T cell signaling complexes. Structure 2002, 10:687-699.

Table 1

ID	Target	Residues selected			Probe	Residue
2	Viral Capsid VP6 [8]	—	—	—	FAB (K + IG)	—
4	Pig Amylase [9]	S243	S245	G249	Camel Amyd10 VHH [9]	F47
5	Pig Amylase [9]	S270	G271	G285	Camel Amy07 VHH [9]	F52
6	Pig Amylase [9]	N53	S145	V349	Camel Amy09 VHH [9]	R51
7	Strp. Pyrog. Exotoxin A1 [10]	N20	N54	Y84	14.3.D T Cell A. R. [10]	G53

Table 1 Shows the proteins selected as target and probe for docking. For this paper, some amino acids residues were chosen to simulate experimental data. Models were scored according to the contact between the residues indicated and any residue on the docking partner, to simulate knowledge of the presence of these residues at the interface.

Table 2

| CAPRI Target | Highest acceptable | | Best in 5 highest | | | |
| | | | With 3 AAs | | With 3+1 AAs | |
	Rank	rmsd	Rank	rmsd	Rank	rmsd
2	136	10.8	-	-	-	-
4	1150	2.3	1 (22)	1.2	1 (14)	3.3
5	1374	3.4	1 (18)	6.6	2	5.4
6	8	2.1	1 (63)	3.3	1 (6)	3.3
7	21	1.9	3	1.8	2	1.8

Table 2 The first two columns show the rank and RMSD value for the highest ranking acceptable solution, or the best if no acceptable solution was kept. The last four columns show the same results using simulated information on the interface region. Numbers in parentheses are the total number of models tied for first place when sorted according to the number of contacts counted when simulating information about the interface region, both when considering only three residues from the target and those three residues plus one from the probe. Table 1 indicates which residues were chosen in each case.
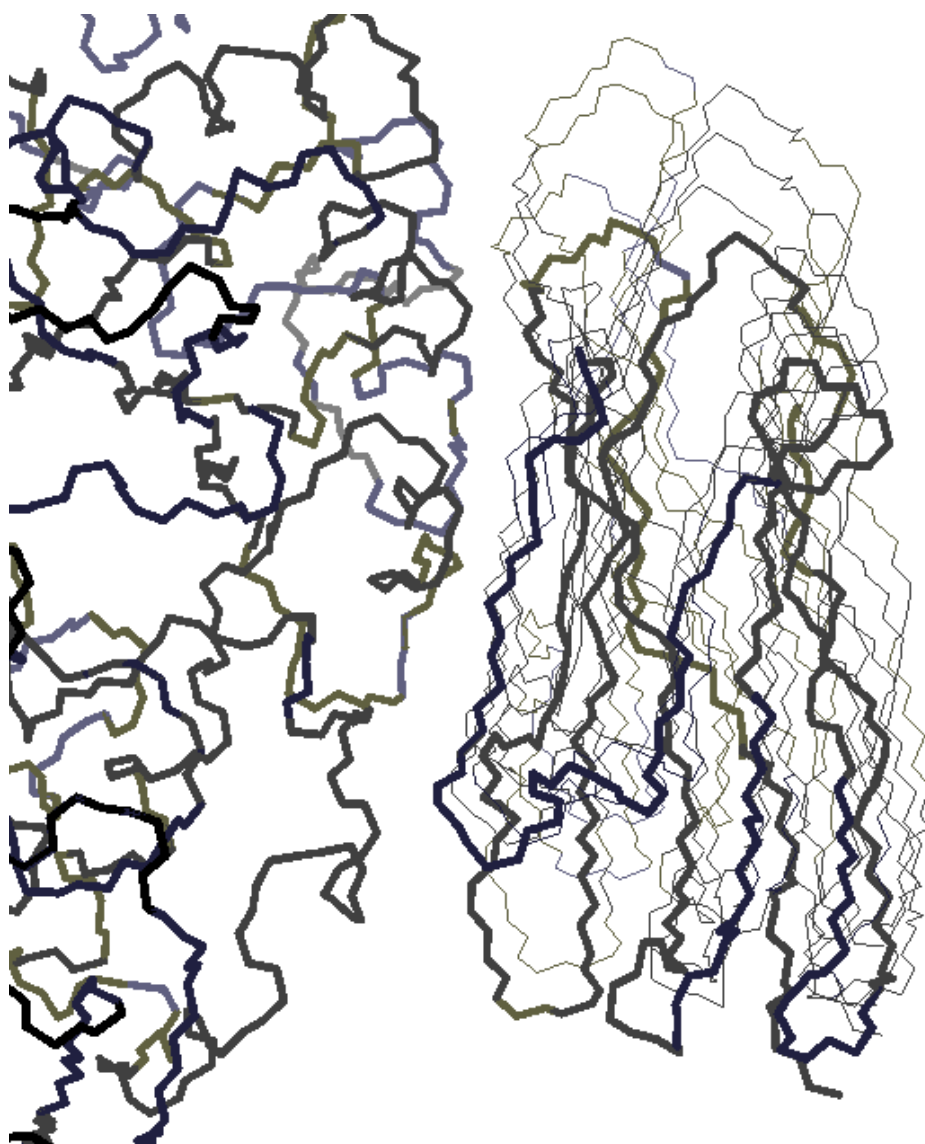
Figure 1

Figure 1 shows the three models for Target 4 in Table 2 (thin lines) superimposed with the X-Ray structure of the complex.
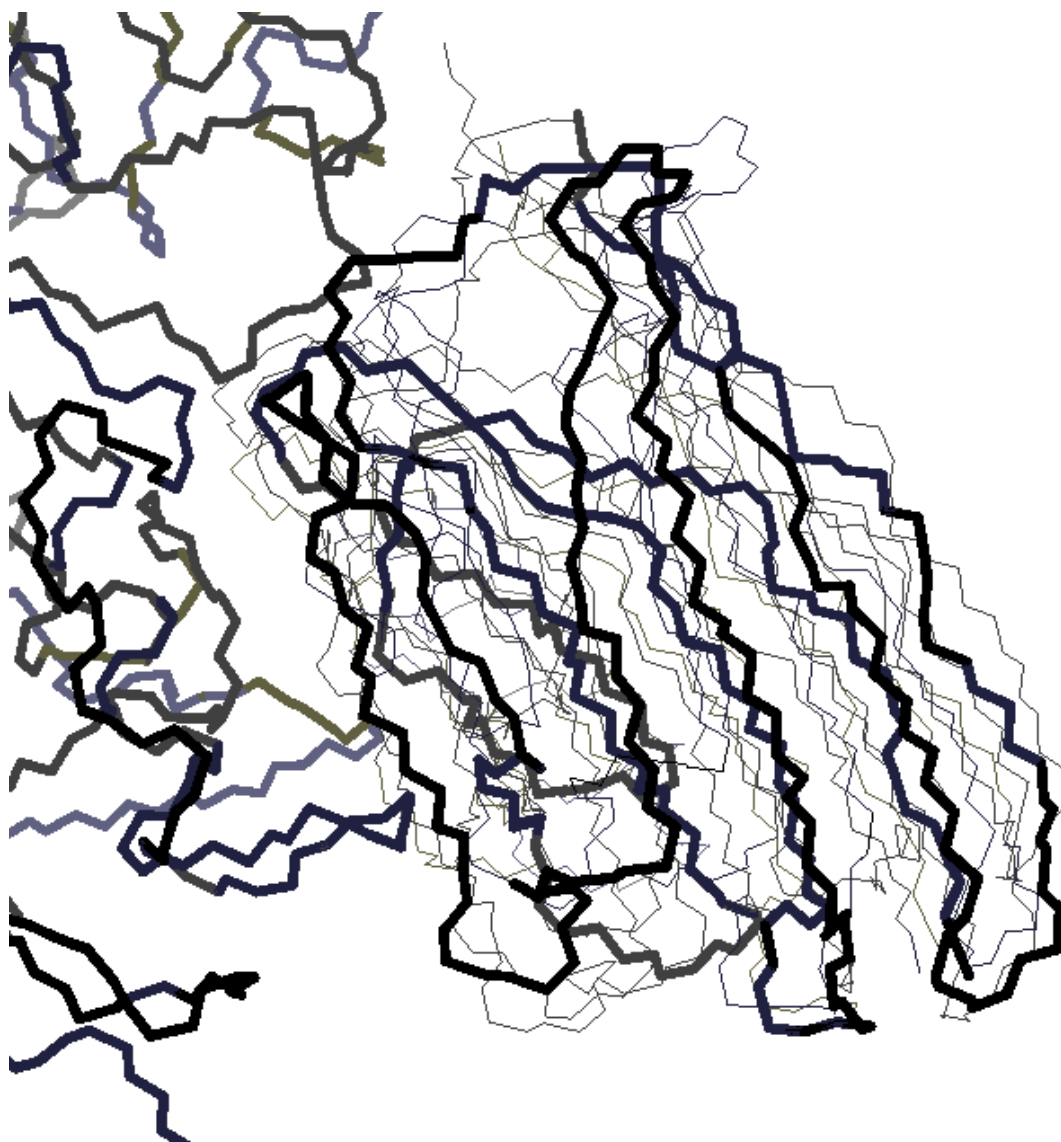
Figure 2

Figure 2 shows the three models for Target 5 in Table 2 (thin lines) superimposed with the X-Ray structure of the complex.

Figure 3

Figure 3 shows the two models for Target 6 in Table 2 (thin lines) superimposed with the X-Ray structure of the complex.
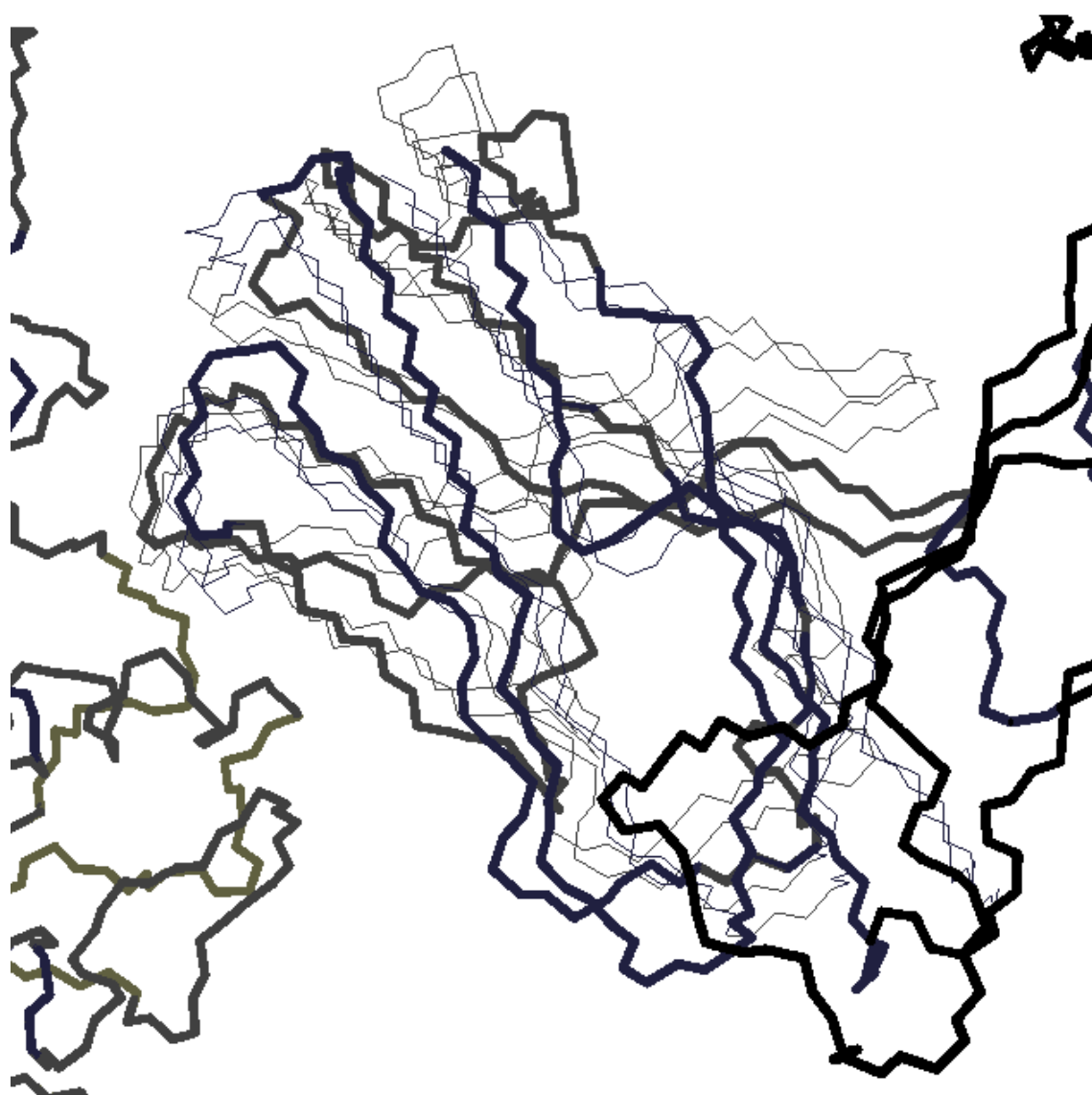
Figure 4

Figure 4 shows the two models for Target 7 in Table 2 (thin lines) superimposed with the X-Ray structure of the complex. Only the fragment up to D118 is shown for our models.