J Mol Model (2011) 17:3163–3172
DOI 10.1007/s00894-011-0990-y

ORIGINAL PAPER

Study of protein complexes via homology modeling, applied to cysteine proteases and their protein inhibitors

Özlem Tastan Bishop · Matthys Kroon

Received: 5 December 2010 / Accepted: 24 January 2011 / Published online: 2 March 2011
© Springer-Verlag 2011

Abstract This paper develops and evaluates large-scale calculation of 3D structures of protein complexes by homology modeling as a promising new approach for protein docking. The complexes investigated were papain-like cysteine proteases and their protein inhibitors, which play numerous roles in human and parasitic metabolisms. The structural modeling was performed in two parts. For the first part (evaluation set), nine crystal structure complexes were selected, 1325 homology models of known complexes were rebuilt by various templates including hybrids, allowing an analysis of the factors influencing the accuracy of the models. The important considerations for modeling the interface were protease coverage and inhibitor sequence identity. In the second part (study set), the findings of the evaluation set were used to select appropriate templates to model novel cysteine protease-inhibitor complexes from human and malaria parasites *Plasmodium falciparum* and *Plasmodium vivax*. The energy scores, considering the evaluation set, indicate that the models are of high accuracy.

Keywords Cathepsin · Chagasin · Cystatin · Malaria

Introduction

Protein interactions are crucial for many cellular processes. The identification and analysis of complexes provide biological information with applications from drug design to understanding the causes of diseases. Diverse approaches have been used toward understanding which and how proteins interact, yet the answers are still limited. The Protein Data Bank (PDB) [1] contains, relative to single protein structures, only a few complexes due to the difficulty of isolation and solving complex structures. Thus, computational methods, such as protein docking, are important in solving and understanding protein interactions. However, current methods for protein docking are not easy. Calculating protein complexes by homology modeling is a promising new approach for protein docking. Some groups built 3D databases of templates to analyze the accuracy of homology based approaches for protein complexes [2–4]. However, previous work is quite limited in scope. In this paper, we develop the subject further. The general issue is, given two families of proteins with a number of known crystal structures of complexes between them, how accurately can we model complexes for cases with unknown structure? As a novel feature of this work, a specific protein family, cysteine proteases and their interactions with protein inhibitors, is systematically analyzed, both in template selection and model accuracy.

Cysteine proteases are a diverse family of enzymes found in all living organisms. They are subdivided into clans, and then into families. Our interest is Clan CA, papain-like family C1 proteases (cathepsin-L and cathepsin-B like proteases), whose members play numerous roles in human and parasitic metabolisms [5–7]. In humans, they are important in the immune system, the protein renewal process and the resorption of bone and cartilage [7]. An imbalance between these proteins and their natural inhibitors leads to diseases such as cancer, immune system defects, osteoporosis and rheumatoid arthritis [8]. In parasites, they have major roles in physiological processes including immune evasion, digestion, and cellular invasion. Thus, they are crucial for parasitic

Electronic supplementary material The online version of this article (doi:10.1007/s00894-011-0990-y) contains supplementary material, which is available to authorized users.

Ö. Tastan Bishop (✉) · M. Kroon
Rhodes University Bioinformatics (RUBi),
Department of Biochemistry, Microbiology and Biotechnology,
Rhodes University,
Grahamstown 6140, South Africa
e-mail: o.tastanbishop@ru.ac.za

diseases such as malaria [9, 10], Chagas disease [11], leishmaniasis [12, 13]. They usually function differently from homologous enzymes in the host, thus many have been identified as promising drug targets. For example, falcipain-2 of *P. falciparum* is a validated drug target and functions as a hemoglobinase [14].

Cystatins regulate endogenous proteases and also defend against foreign cysteine proteases [15]. Thus, the analysis of cysteine protease-inhibitor structures to understand enzyme regulation is important for medical and agricultural applications. Available crystal structures show how cystatins interact with the enzyme: interactions are formed by residues from the N-terminal segment as well as by two β-hairpin loops, one in the middle and one in the C-terminal segments of the molecule. These segments bind in the active site cleft of the protease and block access of the substrate to the catalytic residues. In some parasites no cystatin homolog could be identified. In *Trypanosoma cruzi* a novel inhibitor protein, chagasin [16], and in *P. falciparum* falstatin [17] have been found. Complexes with chagasin in the PDB show that chagasin interacts with a protease in a similar fashion to cystatin [18]. The major difference is that the N-terminal interaction of cystatin is replaced by another loop interaction.

This conserved interaction between cysteine proteases and inhibitors, and the availability of crystal structures, made it possible to apply homology modeling as a promising new tool to obtain many protein complex structures and analyze their accuracy.

Methodology

The study comprises two parts. In part 1 (evaluation set), homology models of complexes with known crystal structures were built to evaluate modeling accuracy. In part 2 (study set), models of novel homolog complexes were built using templates that, according to the evaluation set results, were appropriate. The model accuracy was estimated using results from the evaluation set. Fig. 1 gives a schematic overview.

Data retrieval

Nine crystal structures were identified that contain a cysteine protease in complex with either chagasin or a cystatin by using BLAST [19] and keyword searches in the PDB. For the study set, proteases and inhibitors that occur in humans and in human malaria parasites, *P. falciparum* and *P. vivax*, were chosen. Sequences belonging to *Carica papaya, Gallus gallus* and *T. cruzi* were also included. The sequences were retrieved from the UniProt database [20] (Table 1).

Sequence alignment

A multiple sequence alignment was constructed for the proteases, cystatins and chagasin-like inhibitors using the Promals3D [21] web server. Sequences from UniProt and PDB files were included. Promals3D pre-groups similar

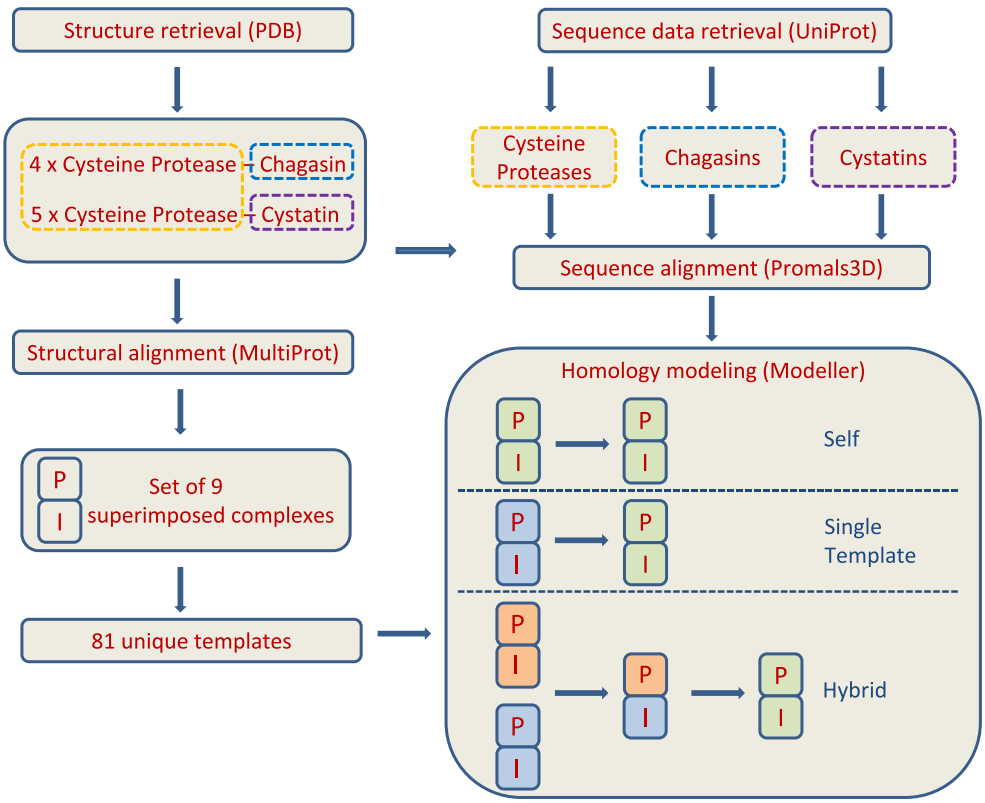

Fig. 1 Schematic overview of the modeling process for both evaluation and study set

Table 1 Dataset of cysteine proteases and their inhibitors used in this study. The PDB ID of the nine complexes, the UniProt protein sequence accession numbers of the individual proteins as well as the name of the organisms to which proteins belong are given. Cystatin refers to egg-white-cystatin in the text

Protease	Organism	Inhibitor	Organism	PDB ID	Accession
Cathepsin-B	*H. sapiens*	Cystatin-A	*H. sapiens*	3K9M	
Cathepsin-B	*H. sapiens*	Chagasin	*T. cruzi*	3CBJ	
Cathepsin-H	*S. Scrofa*	Cystatin-A	*H. sapiens*	1NB5	
Cathepsin-L	*H. sapiens*	Chagasin	*T. cruzi*	2NQD	
Falcipain-2	*P. falciparum*	Cystatin	*G. gallus*	1YVB	
Falcipain-2	*P. falciparum*	Chagasin	*T. cruzi*	2OUL	
Papain	*C. papaya*	Chagasin	*T. cruzi*	3E1Z	
Papain	*C. papaya*	Cystatin-B	*H. sapiens*	1STF	
Papain	*C. papaya*	Tarocystatin	*C. esculenta*	3IMA	
Cathepsin-B	*H. sapiens*				P07858
Cathepsin-H	*H. sapiens*				P09668
Cathepsin-L	*H. sapiens*				P07711
Cathepsin-S	*H. sapiens*				P25774
Falcipain-2	*P. falciparum*				Q9N6S8
Falcipain-2'	*P. falciparum*				Q8I6U5
Falcipain-3	*P. falciparum*				Q9NAW4
Vivapain-2(a)	*P. falciparum*				Q6J131
Vivapain-2(b)	*P. falciparum*				Q5IZD8
Vivapain-3	*P. falciparum*				Q7Z0B2
Papain	*C. papaya*				P00784
		Cystatin-A	*H. sapiens*		P01040
		Cystatin-B	*H. sapiens*		P04080
		Cystatin-C	*H. sapiens*		P01034
		Kininogen-1	*H. sapiens*		P01042
		Kininogen-2	*H. sapiens*		P01042
		Kininogen-3	*H. sapiens*		P01042
		Falstatin	*P. falciparum*		Q2PZB1
		Cystatin	*G. gallus*		P01038
		Chagasin	*T. cruzi*		Q966X9
		Tarocystatin	*C. esculenta*		Q8L5J8

sequences, thus dealing with redundancy. The double inclusion was used because sometimes there are minor differences between UniProt and PDB sequences, and the evaluation set models used the exact sequence as in the crystal structure. Errors in the alignment of cathepsin-B relative to other proteases, due to the extra occluding loop, became apparent later when comparing models to native structures. The errors were left however since they would not have been discovered without the native structures and thus their presence represents part of the unavoidable errors associated with high throughput homology modeling and should form part of the error estimation for the models. This error does not affect any models in the study set as better templates were selected in all cases.

Structural alignment

Non-protein atoms were removed from coordinate files. Where multiple copies of protein complexes were present in one asymmetric unit, the one with the fewest missing residues was chosen and the others deleted. All crystal structures were simultaneously superimposed using MultiProt [22]. MultiProt produces various possible solutions. The solution with the largest number of aligned residues (178 residues) that contained all crystal structures was chosen. From nine superimposed structures, each containing one protease and one inhibitor, 81 complexes were extracted by selecting the protease and inhibitor independently (Fig. 1).

Homology modeling

Homology models were generated using Modeller 9v6 [23] in high throughput fashion. For each target-template pair, the alignment was automatically extracted from the Promals3D alignment. For every template-target combination five models were built using standard 'automodel' routine of the program with very slow refinement option.

Evaluation set

Homology models were built in three ways: Self, single template and hybrid (Fig. 1). In "self", target complexes were modeled based on their own crystal structure to estimate the modeling error. In "single template", target complexes were modeled from a single homolog crystal structure complex. In "hybrid" target complexes were constructed based on templates generated by combining a protease and an inhibitor from different crystal structures. Hybrid templates were constructed by copying coordinates for protease and inhibitor from the superimposed structures.

Every crystal structure complex was modeled using every combination of protease and inhibitor templates in the superimposed complexes (excluding itself). The crystal structures containing cystatin were modeled based on 33 different templates, and those containing chagasin on 25 different templates. Five models were built for each template, making a total of 1325.

Study set

Models were built for every combination of protease and inhibitor sequences. Eleven proteases and 10 inhibitors gave 110 target complexes. The templates were constructed by selecting the protease and inhibitor independently as no systematic differences between the quality of the single template and hybrid models were detected. The proteases were selected on the basis of sequence coverage and the inhibitors on sequence identity. If more than one structure of the same inhibitor was present then the one in the complex with the protease showing the highest sequence identity to the target protease was chosen.

Model evaluation

The quality of models was evaluated by calculating several parameters. Two energy scores, the DOPE Z score and Rosetta energy score were calculated for models of both sets. Additionally, i-RMSD, C_α-RMSD (RMSD), GDT-ha [24] and Tm-score [25] were calculated for the evaluation set models.

Results and discussion

Global accuracy measures in the evaluation set

The evaluation set models were built to investigate parameters that can be used to estimate the accuracy of a model. For this purpose, homology models of complexes were calculated for known crystal structures in a variety of ways as described under Methodology. Later, we applied our conclusions to the study set models.

Percentage sequence identity is an often used measure of evolutionary distance between proteins. The sequence identity values between the proteases and inhibitors in the crystal structures are given in Table 2. There are cases where the template and target sequences are identical; however this is not redundancy since these provide information about conformational changes upon binding due to different spatial conformations. This can give detailed insights about interactions between cysteine proteases and the inhibitor proteins.

Energy versus RMSD

The models were superimposed on corresponding crystal structures and RMSD values were found. Median energy values for DOPE Z score and Rosetta energy score of the five models of every template were plotted against median RMSD values (Fig. 2), and so it is not necessarily the energy and RMSD of the same model that are plotted (see online resource 1 for the complete dataset containing the values for every single model). As expected, the self models had very small energy and RMSD values. The single template and hybrid models were distributed throughout the graph. According to Shen and Sali [26] the DOPE score has higher discriminatory power when the models are close to the native state than when they are distant; while the Rosetta energy retains its discriminatory power better for distant models. The scope of that paper was monomers, but our results suggest it applies to protein complexes too. The figure shows that for a given DOPE Z score or Rosetta energy, there is a wide range of possible RMSD values and vice versa, especially at Rosetta energies above 2000 and DOPE Z scores above -0.5. At lower values of RMSD (<4 Å), the correlation between RMSD and energy functions is better. Thus, energy functions can be used to distinguish near native models from distant models.

Correlation between metrics

The correlation coefficients between various measurements were calculated (Table 3). Since correlation coefficient requires linear relationship, the usage of it here, is merely to get an overall idea. As expected there is very low correlation between independent quantities such as protein identity against inhibitor identity. On the other hand, some interesting observations can be made. There is a high correlation (0.78) between Rosetta energy and DOPE Z score, and good correlation between DOPE Z score and RMSD (0.68) as well as Rosetta energy and RMSD (0.69).

Table 2 Percent sequence identity between proteins in crystal structures. The percentage sequence identities between the proteins in the crystal structures are shown. The proteases are shown in the top-left half while the inhibitors are shown in the bottom-right. Cystatin refers to egg-white-cystatin in the text

		3IMA	3CBJ	2OUL	1YVB	1NB5	3K9M	2NQD	3E1Z	1STF	
		Papain	Cathepsin-B	Falcipain-2	Falcipain-2	Cathepsin-H	Cathepsin-B	Cathepsin-L	Papain	Papain	
1STF	Papain	98.58	29.13	37.02	37.02	40.29	29.13	39.61	98.58	100.00	Cystatin-B
3E1Z	Papain	100.00	29.61	37.02	37.02	41.26	29.61	39.61	100.00	0.00	Chagasin
2NQD	Cathepsin-L	39.61	29.44	35.98	35.98	46.26	29.58	100.00	100.00	0.00	Chagasin
3K9M	Cathepsin-B	29.61	99.21	25.36	25.36	30.09	100.00	NA	NA	53.06	Cystatin-A
1NB5	Cathepsin-H	41.26	30.09	37.80	37.80	100.00	100.00	NA	NA	53.06	Cystatin-A
1YVB	Falcipain-2	37.02	25.12	100.00	100.00	13.95	13.95	NA	NA	19.77	Cystatin
2OUL	Falcipain-2	37.02	25.12	100.00	NA	NA	NA	100.00	100.00	0.00	Chagasin
3CBJ	Cathepsin-B	29.61	100.00	100.00	NA	NA	NA	100.00	100.00	0.00	Chagasin
3IMA	Papain	100.00	NA	NA	31.76	19.51	19.51	NA	NA	19.51	Tarocystatin
		Tarocystatin	Chagasin	Chagasin	Cystatin	Cystatin-A	Cystatin-A	Chagasin	Chagasin	Cystatin-B	

This conclusion can be made at least for the low RMSD models (<4 Å), since data shows linearity (Figs. 2A and B). Energy scores seem to correlate with Tm-score and GDT-ha as well (Fig. 3). There is a clear correlation observed between protease sequence identity and GDT-ha, although the identity range from 50% to 90% is badly sampled. Of particular interest is any indicator that correlates well with RMSD and which can be calculated independently of the structure of the complex. The indicator of this type with the highest correlation coefficient was total coverage. Sequence coverage is defined as the percentage of the target sequence that is aligned to the template. We defined a weighted combination of inhibitor and protease values as total identity and total coverage. We gave the protease double the weight of the inhibitor as it is roughly double the size. The correlation between identity and coverage is higher for the inhibitors than for proteases, which is possibly because of chagasin biasing the dataset. In conclusion, total coverage is the best parameter to estimate the RMSD at high RMSD (>4 Å) and the energy functions can be used at lower RMSD. We suggest using energy functions when the protease coverage is above 92% and the inhibitor coverage over 95% (Online Resource 2).

Models with low RMSDs

Since we know that models with low RMSD values are accurate models, we investigated them in detail. The analysis is in two categories: High accuracy models (RMSD≤1.5 Å), and medium accuracy models (1.5 Å≤ RMSD≤4.0 Å). The whole data set with five models for each complex was used. The high accuracy models consist of all self models and 90 hybrid models. Interestingly, RMSD values ranged between 0.13 (3IMA; papain–tarocystatin) to 0.45 Å (1YVB; falcipain-2–egg-white-cystatin) for the self models illustrating the error range of homology modeling. Most hybrid models were built with very high template sequence identity either to the protease or inhibitor or both. The two lowest sequence identity cases are (a) 1NB5, with cathepsin-L of 2NQD (46.26% sequence identity and 97.27% sequence coverage to cathepsin-H) and cystatin-B of 1STF (53.06% and 100% to cystatin A); and (b) 1STF, with cathepsin-L of 2NQD (39.61% and 97.64% to papain) and cystatin-A of 1NB5 (53.06% and 100% to cystatin-B). The medium accuracy models contained a mixture of 548 single template and hybrid models. One example of a single template model with low sequence identity is model of 3IMA by template 1NB5 (cathepsin-H with 41.26% and 97.17% sequence identity and coverage respectively; cystatin-A with 19.51 and 96.47%) giving 2.12 Å RMSD. In all cases, we observed high sequence coverage demonstrating the correlation between RMSD and coverage.

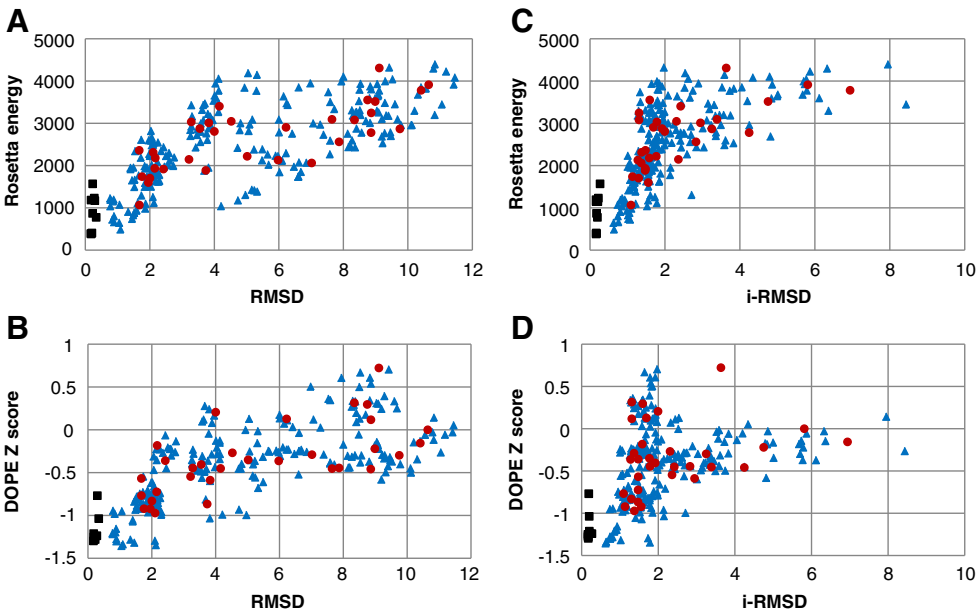

Fig. 2 Energy scores as function of RMSD and i-RMSD. The respective median Rosetta energy (**A**) and (**C**) and the DOPE Z score (**B**) and (**D**) are shown as a function of the median RMSD and i-RMSD values for the five models of each target-template combination. Self-models are shown as black squares, single template models as red circles and hybrid models as blue triangles

Interface accuracy in the evaluation set by i-RMSD

The i-RMSD is the backbone RMSD between the model and native structures using only interface residues. Residues are classified as being in the interface if they have at least one atom within 10 Å of the interaction partner. The models were categorized into four groups with i-RMSD values: (A) 0.62 to 2.13 Å, (B) 1.47 to 2.89 Å, (C) 1.63 to 3.72 Å, and (D) 2.40 to 8.43 Å (Fig. 4). The models in group D are all the models where the target protease was cathepsin-B and the template protease was a cathepsin-L like protease. The large error in this group is due to the incorrect modeling of the occluding loop. Group C contains all models where the target protease is a cathepsin-L like protease with a cathepsin-B template. A slight alignment error at the borders of the occluding loop was identified as the main reason for the increase in i-RMSD value. Group B contains all models excluding group C and D where the target inhibitor was cystatin-A or cystatin-B and the template inhibitor egg-white-cystatin or tarocystatin. The increased i-RMSD is due to a C-terminal extension in the interface in the target inhibitors. All self models and models not falling in the other groups fall in Group A.

Table 3 Pearson correlation coefficients between model parameters and evaluation metrics. The values were calculated using the median value of each metric for the five models of all template-target combinations. Detailed explanation of the parameters is given in the text

	A DOPE Z Score	B Rosetta	C Protease id	D Protease coverage	E Inhibitor id	F Inhibitor Coverage	G Total id	H Total coverage	I RMSD	J Tm-RMSD	K Tm score	L GDT-ha	M i-RMSD
A	1.00	0.78	-0.62	-0.52	-0.45	-0.59	-0.79	-0.71	0.68	0.79	-0.82	-0.83	0.27
B	0.78	1.00	-0.71	-0.56	-0.36	-0.48	-0.80	-0.69	0.69	0.77	-0.82	-0.86	0.63
C	-0.62	-0.71	1.00	0.54	-0.06	0.02	0.76	0.44	-0.47	-0.71	0.59	0.80	-0.41
D	-0.52	-0.56	0.54	1.00	-0.12	0.15	0.35	0.87	-0.88	-0.42	0.77	0.63	-0.59
E	-0.45	-0.36	-0.06	-0.12	1.00	0.72	0.60	0.26	-0.19	-0.48	0.36	0.35	-0.07
F	-0.59	-0.48	0.02	0.15	0.72	1.00	0.49	0.61	-0.46	-0.54	0.56	0.48	-0.08
G	-0.79	-0.80	0.76	0.35	0.60	0.49	1.00	0.52	-0.50	-0.89	0.71	0.87	-0.38
H	-0.71	-0.69	0.44	0.87	0.26	0.61	0.52	1.00	-0.93	-0.61	0.90	0.74	-0.51
I	0.68	0.69	-0.47	-0.88	-0.19	-0.46	-0.50	-0.93	1.00	0.60	-0.94	-0.76	0.65
J	0.79	0.77	-0.71	-0.42	-0.48	-0.54	-0.89	-0.61	0.60	1.00	-0.80	-0.94	0.39
K	-0.82	-0.82	0.59	0.77	0.36	0.56	0.71	0.90	-0.94	-0.80	1.00	0.90	-0.62
L	-0.83	-0.86	0.80	0.63	0.35	0.48	0.87	0.74	-0.76	-0.94	0.90	1.00	-0.55
M	0.27	0.63	-0.41	-0.59	-0.07	-0.08	-0.38	-0.51	0.65	0.39	-0.62	-0.55	1.00

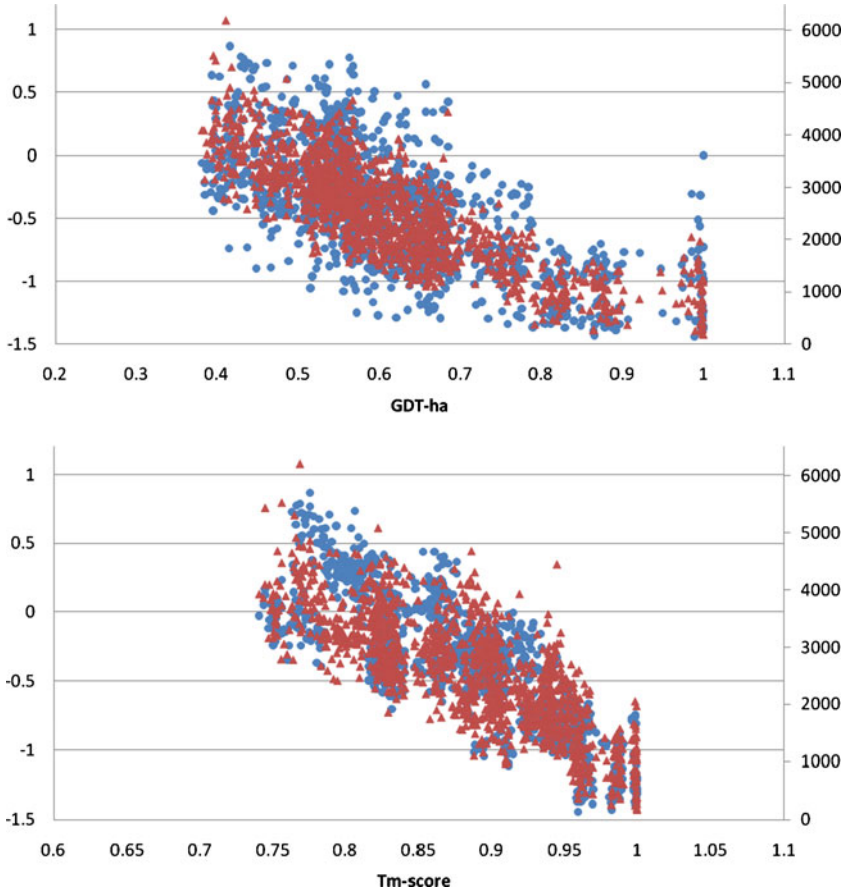

Fig. 3 Energy scores of models at different GDT-ha and Tm-score values for all target complexes in the evaluation set. The Rosetta energy is indicated with red triangles and the values shown on the right axis. The DOPE Z scores are indicated with blue circles and the values shown on the left axis. Values for all models are shown

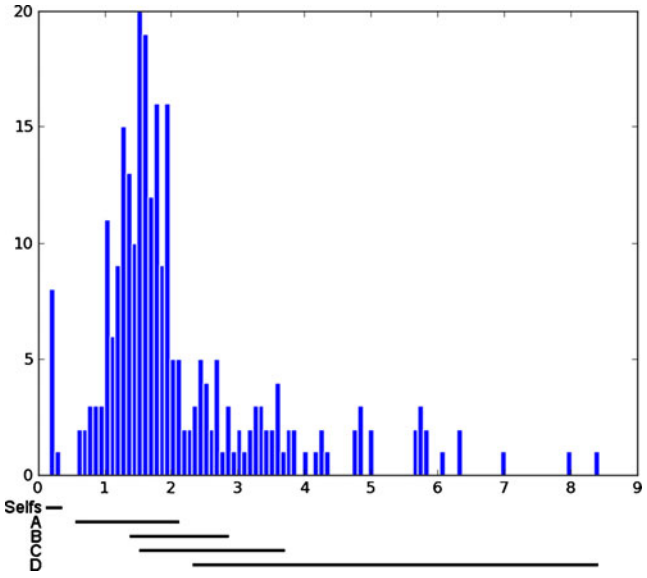

Fig. 4 Histogram showing the distribution of i-RMSD values of models in the evaluation set. The median i-RMSD value of the 5 models for each template for each target complex is shown. Frequency is shown on the Y-axis and i-RMSD on the X-axis

Energy functions

The correlation coefficient between energy functions and i-RMSD was calculated (Table 3). Although a good correlation was observed between Rosetta energy and i-RMSD (0.63), strangely it was not the case for DOPE Z score (0.27). Median energy values of models were plotted against their median i-RMSD values (Figs. 2C & D). Both graphs show a similar pattern. A Rosetta energy below 2000 and DOPE Z score below -0.5 typically correspond to models with i-RMSD below 2.5 Å. According to CAPRI assessment, high accuracy and medium accuracy models can be defined as, among other parameters, models which show i-RMSD within 1.0 Å and 2.5 Å respectively [4]. It seems that Rosetta energy is better at identifying models with low i-RMSD but that the correlation is not fine enough to differentiate between close models.

Sequence identity and coverage

There are some excellent models (i-RMSD <2 Å) with sequence identity as low as 20% for the inhibitor (Online resource 3). There are also models below 5 Å with inhibitor sequence identity below 20% and protease sequence identity below 35%. In general higher sequence identity

corresponds to better models but the relationship is not simple. Sequence identity above 50% almost guarantees a good model for that component of the complex.

There is a clear trend that higher coverage values correspond to lower i-RMSD values but there are some models with low i-RMSD values in spite of weighted total coverage values of only 250. The total coverage provides a good idea of the upper limit that can be expected for the i-RMSD of a model. The results make sense for both identity and coverage since the interface region is the most conserved area. Thus even if the sequence identity or coverage is low, the interface may be similar. Only occasionally, namely in modeling the cathepsin-B occluding loop and in modeling cystatin-A or cystatin-B based on egg-white-cystatin or tarocystatin, does coverage come into play for i-RMSD.

Overall, picking templates with the highest identity and coverage values yields good models. However, this will not guarantee that the best template is used. If there are multiple templates with similar levels of identity and coverage models should be built for all of them and the Rosetta energy calculated.

Accuracy of the study set models

Our interest here is proteases and inhibitors that occur in human, and in human malaria parasites, *P. falciparum* or *P. vivax*. Sequences belonging to *C. papaya, G. gallus* and *T. cruzi* were also included. Eleven proteases and 10 inhibitors of this set gave 110 different target complexes. For each target five models were calculated.

In the study set, protease identity varies from 54% to 100% and inhibitor identity from 12% to 100%. Protease targets below 90% identity are Plasmodial proteases apart from falcipain-2 and falcipain-2' which range from 54% to 66% and cathepsin-S at 57%. The inhibitor with the lowest identity is falstatin at 12%. The other inhibitors modeled below 90% identity are cystatin-C at 46%, kininogens at 26%, 21% and 28%. Protease coverage varies from 98% to 100%. Inhibitor coverage is from 87% to 100%. In the case of falstatin, the coverage is artificially high at 97.25% because large inserts were removed to make modeling possible. These inserts constitute more than half the protein so the part that was modeled is just one domain of a multi-domain protein. Other inhibitors below 95% coverage are cystatin-C at 92.5%, the kininogens at 92%, 88% and 88% and tarocystatin at 93%. The crystal structure containing tarocystatin has six unresolved residues, causing the incomplete coverage.

Overall, the models of the Study Set show good energy function scores. The DOPE Z scores and Rosetta energies of the models are presented in Fig. 5: the measures are reasonably correlated, and most of the models have both DOPE Z score ≤-0.5 and Rosetta energy ≤2000. The mean values of the energy functions for each target complex are summarized in Table 4. For most models, according to the Evaluation Set results, the correlation between energy functions and RMSD and i-RMSD in this range is strong: The number of models, out of a total of 550, in this category is 396 for DOPE Z score ≤-0.5, 380 for Rosetta energy ≤2000.

The lowest scoring model according to Rosetta energy is the complex of cathepsin-L with tarocystatin. It has a score of 322 which is better than many self models. The worst is 3462 for a model of falstatin with vivapain-3, for which case coverage and identity are better indicators of accuracy. In fact, the inhibitor, falstatin, shares only 12.26% sequence identity with chagasin making alignment very difficult, and several big inserts had to be removed. Still, this energy is lower than some models in the evaluation set and when looking at the graph of Rosetta energy versus RMSD, the RMSD is at a minimum four and probably closer to about

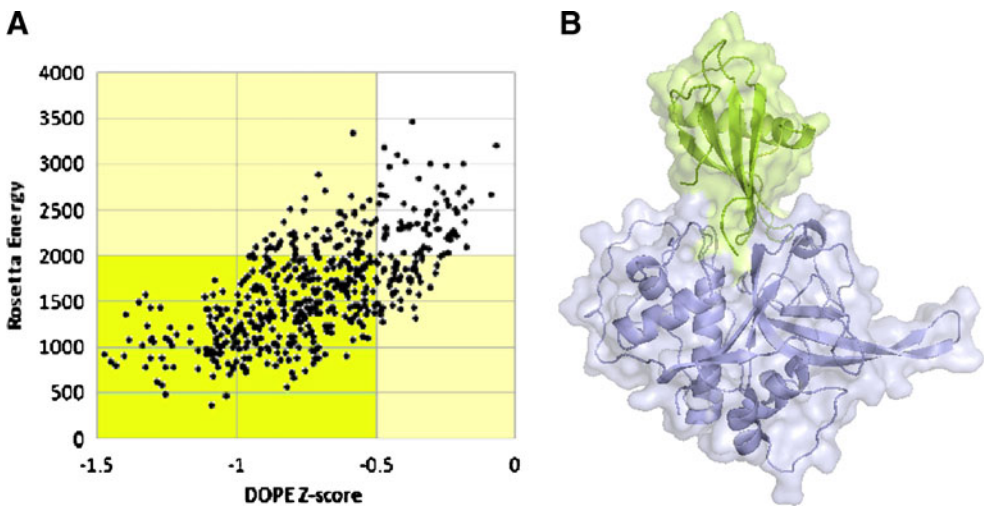

Fig. 5 Study Set energy distributions. (**A**) The relation between Rosetta energy and DOPE Z score for the models in the study set. Models falling into the yellow area have Rosetta energy below 2000 and DOPE Z score below -0.5. Models falling in the pink area satisfy one of the two criteria. (**B**) Cartoon representation of falcipain-2 (blue, below) and cystatin-A (green, above) model as an example

Table 4 Average DOPE Z scores and Rosetta energies for study set models. The proteases of the complexes are given on the left and the inhibitors are given on the top. The DOPE Z scores are on the left and Rosetta energies on the right in ***bold italic***

	Cystatin-A		Cystatin-B		Cystatin-C		Cystatin		Tarocystatin		Kininogen-1		Kininogen-2		Kininogen-3		Chagasin		Falstatin	
Cathepsin-B	-1.25	***1652***	-1.09	***1498***	-0.79	***1923***	-0.90	***1983***	-0.95	***1218***	-0.71	***1932***	-0.59	***2101***	-0.59	***2117***	-1.09	***1687***	-0.73	***2712***
Cathepsin-H	-1.09	***1379***	-1.03	***1077***	-0.58	***1835***	-0.83	***1272***	-0.78	***1162***	-0.56	***1752***	-0.38	***2021***	-0.48	***1369***	-1.04	***981***	-0.65	***2049***
Cathepsin-L	-1.30	***1053***	-1.26	***651***	-0.85	***1227***	-1.05	***971***	-1.04	***619***	-0.76	***1282***	-0.61	***1518***	-0.67	***1351***	-1.24	***699***	-0.87	***1664***
Cathepsin-S	-1.33	***1459***	-1.25	***1053***	-0.89	***1727***	-1.05	***1391***	-1.04	***1283***	-0.78	***1746***	-0.62	***2014***	-0.64	***1786***	-1.26	***1070***	-0.86	***2241***
Falcipain-2	-1.05	***1039***	-0.96	***1138***	-0.64	***1480***	-0.76	***1414***	-0.76	***787***	-0.55	***1369***	-0.38	***1697***	-0.45	***1547***	-1.03	***651***	-0.77	***1420***
Falcipain-2p	-1.00	***1410***	-0.93	***1285***	-0.61	***1467***	-0.75	***1352***	-0.76	***847***	-0.56	***1367***	-0.37	***1679***	-0.43	***1683***	-1.00	***847***	-0.71	***1720***
Falcipain-3	-0.94	***1495***	-0.76	***1762***	-0.44	***1977***	-0.53	***1942***	-0.62	***1225***	-0.36	***1973***	-0.24	***2133***	-0.32	***1857***	-0.80	***1563***	-0.46	***2328***
Vivapain-2a	-0.91	***1604***	-0.85	***1518***	-0.41	***2249***	-0.66	***1825***	-0.65	***1450***	-0.42	***1858***	-0.33	***2039***	-0.33	***2226***	-0.91	***1209***	-0.64	***1851***
Vivapain-2b	-0.85	***2012***	-0.68	***2196***	-0.42	***2330***	-0.55	***2207***	-0.55	***1557***	-0.30	***2417***	-0.22	***2344***	-0.21	***2540***	-0.80	***1630***	-0.50	***2616***
Vivapain-3	-0.82	***2091***	-0.81	***1540***	-0.52	***2075***	-0.63	***2047***	-0.66	***1396***	-0.36	***2332***	-0.26	***2261***	-0.36	***2048***	-0.85	***1533***	-0.46	***2905***
Papain	-1.43	***909***	-1.36	***795***	-0.96	***1168***	-1.10	***991***	-1.16	***744***	-0.91	***1019***	-0.73	***1325***	-0.84	***1015***	-1.36	***673***	-0.94	***1877***

10. Using the DOPE Z score the best model is of papain with cystatin-A with a score of -1.475 which again is better than some of the self-models. The worst score is 0.069 for a model of kininogen-2 with vivapain-3. Figure 5A shows that both metrics should be calculated as they measure different aspects of model quality.

Conclusions

Reliable prediction of protein complexes is not easy and current high accuracy protein-protein docking methods are computationally expensive. Homology modeling is a promising new approach for predicting protein complex structures. In this study, homology based docking approach was applied and detailed accuracy analysis was done for cysteine proteases and their inhibitor structures as a case study. The conserved interaction between cysteine proteases and inhibitors, and the availability of crystal structures within the family, were important factors making the method feasible. Thus, this approach would be applicable to other families of protein complexes satisfying these conditions.

The study was performed in two parts. In the evaluation set, we constructed models of complexes with known crystal structures, and identified factors indicating model accuracy, namely sequence coverage, sequence identity and energy scores. It was also found that hybrid models are as good as those using a simple template. This is important since it gives a much wider choice of template, making it easier to meet the criteria for a high accuracy model.

In the study set, we constructed novel models of human and malaria parasite cysteine protease-inhibitor complexes. Cysteine proteases are crucial for the survival of the malaria parasite as well as of other parasites. Since they function differently from homologous enzymes in human, the comparative analysis of cysteine protease-inhibitor structures between human and malarial parasites would help to understand the function of parasite cysteine proteases and how they differ from human ones. Falcipain-2 of *P. falciparum* is a validated drug target. Available crystal structures of falcipain-2 with cystatin and chagasin, made it possible to calculate vivapain-2, homologous enzyme in *P. vivax*, complexes with high accuracy. Some other examples of novel high accuracy complexes are falcipain-2 with human cystatin-A (Fig. 5B) and human cathepsin-L with human cystatin-A. The detailed analysis of all high accuracy complexes of the study set will be reported elsewhere.

Acknowledgments MK thanks Rhodes University and National Research Foundation (NRF) for financial support. We thank Prof Anna Tramontano for comments.

References

1. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H (2000) The protein data bank. Nucleic Acids Res 28:235–242
2. Kundrotas PJ, Alexov E (2006) Predicting 3D structures of transient protein-protein complexes by homology. Biochim Biophys Acta 1764:1498–1511
3. Kundrotas PJ, Vakser IA (2010) Accuracy of protein-protein binding sites in high-throughput template-based modeling. PLoS Comput Biol 6:e1000727
4. Movshovitz-Attias D, London N, Schueler-Furman O (2010) On the use of structural templates for high-resolution docking. Proteins 78:1939–1949
5. Atkinson HJ, Babbitt PC, Sajid M (2009) The global cysteine peptidase landscape in parasites. Trends Parasitol 25:573–581
6. Rosenthal PJ (2004) Cysteine proteases of malaria parasite. Int J Parasitol 34:1489–1499
7. Lecaille F, Kaleta J, Brömme D (2002) Human and parasitic papain-like cysteine proteases: their role in physiology and pathology and recent developments in inhibitor design. Chem Rev 102:4459–4488
8. Vasiljeva O, Reinheckel T, Peters C, Turk D, Turk V, Turk B (2007) Emerging roles of cysteine cathepsins in disease and their potential as drug targets. Curr Pharm Des 13:387–403
9. Rosenthal PJ, Mckerrow JH, Aikawa M, Nagasawa H, Leech JH (1988) A malarial cysteine proteinase is necessary for hemoglobin degradation by Plasmodium falciparum. J Clin Invest 82:1560–1566
10. Rosenthal PJ, Sijwali PS, Singh A, Shenai BR (2002) Cysteine proteases of malaria parasites: targets for chemotherapy. Curr Pharm Des 8:1659–1672
11. Cazzulo JJ, Stoka V, Turk V (1997) Cruzipain, the major cysteine proteinase from the protozoan parasite Trypanosoma cruzi. Biol Chem 378:1–10
12. Mottram JC, Coombs GH, Alexander J (2004) Cysteine peptidases as virulence factors of Leishmania. Curr Opin Microbiol 7:375–381
13. Mahmoudzadeh-Niknam H, McKerrow JH (2004) Leishmania tropica: cysteine proteases are essential for growth and pathogenicity. Exp Parasitol 106:158–163
14. Rosenthal PJ (2002) Hydrolysis of erythrocyte proteins by proteases of malaria parasites. Curr Opin Hematol 9:140–145
15. Goulet M, Dallaire C, Vaillancourt L, Khalf M, Badri AM, Preradov A, Duceppe M, Goulet C, Cloutier C, Michaud D (2008) Tailoring the specificity of a plant cystatin toward herbivorous insect digestive cysteine proteases by single mutations at positively selected amino acid sites. Plant Physiol 146:1010–1019
16. Monteiro AC, Abrahamson M, Lima AP, Vannier-Santos MA, Scharfstein J (2001) Identification, characterization and localization of chagasin, a tight-binding cysteine protease inhibitor in Trypanosoma cruzi. J Cell Sci 114:3933–3942
17. Pandey KC, Singh N, Arastu-Kapur S, Bogyo M, Rosenthal PJ (2006) Falstatin, a cysteine protease inhibitor of Plasmodium falciparum, facilitates erythrocyte invasion. PLoS Pathog 2:e117
18. Wang SX, Pandey KC, Scharfstein J, Whisstock J, Huang RK, Jacobelli J, Fletterick RJ, Rosenthal PJ, Abrahamson M, Brinen LS, Rossi A, Sali A, McKerrow JH (2007) The structure of chagasin in complex with a cysteine protease clarifies the binding mode and evolution of an inhibitor family. Structure 15:535–543
19. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410
20. The UniProt Consortium (2010) The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Res 38:D142–148
21. Pei J, Kim BH, Grishin NV (2008) PROMALS3D: a tool for multiple protein sequence and structure alignments. Nucleic Acids Res 36:2295–2300
22. Shatsky M, Nussinov R, Wolfson HJ (2004) A method for simultaneous alignment of multiple protein structures. Proteins 56:143–156
23. Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234:779–815
24. Zemla A, Venclovas C, Moult J, Fidelis K (1999) Processing and analysis of CASP3 protein structure predictions. Proteins Suppl 3:22–29
25. Zhang Y, Skolnick J (2004) Scoring function for automated assessment of protein structure template quality. Proteins 57:702–710
26. Shen M, Sali A (2006) Statistical potential for assessment and prediction of protein structures. Protein Sci 15:2507–2524