The protease inhibitor chagasin of *Trypanosoma cruzi* adopts an immunoglobulin-type fold and may have arisen by horizontal gene transfer

Daniel J. Rigden*, Ana C.S. Monteiro, M. Fatima Grossi de Sá

National Centre of Genetic Resources and Biotechnology, Cenargen/Embrapa, S.A.I.N. Parque Rural, Final W5 Norte, 70770-900 Brasilia, Brazil

Received 14 May 2001; accepted 23 July 2001

First published online 6 August 2001

Edited by Takashi Gojobori

Abstract Chagasin, a protein from *Trypanosoma cruzi*, is the first member of a new family of cysteine protease inhibitors. Despite its lack of significant sequence identity with known proteins, convincing structural models, using variable light chain templates, could be constructed on the basis of threading results. Experimental support for the final structure came from inhibition data for overlapping oligopeptides spanning the chagasin sequence. Chagasin therefore exemplifies a new protease inhibitor structural class and a new natural use for an immunoglobulin-like domain. Limited sequence resemblance suggests that chagasin may represent the result of a rare horizontal gene transfer from host to parasite. © 2001 Federation of European Biochemical Societies. Published by Elsevier Science B.V. All rights reserved.

Key words: Chagasin; Protease inhibitor; Immunoglobulin-like domain; Threading; Horizontal gene transfer; *Trypanosoma cruzi*

1. Introduction

Cysteine proteases are ubiquitous enzymes. The causative agent of Chagas' disease, *Trypanosoma cruzi*, contains a cathepsin L-like protease called cruzipain [1]. The three-dimensional structure of the mature, cleaved form of cruzipain has been determined by X-ray crystallography [2]. In contrast to a simple digestive role, cysteine proteases of parasites seem to perform multiple tasks in different hosts and environmental settings and therefore represent possible targets for drug intervention. Several classes of inhibitors of cysteine proteases are known, the best studied being the cystatin superfamily of sequentially homologous proteins [3].

A search for cystatin-type inhibitors in *T. cruzi* instead revealed a novel inhibitor, named chagasin, expressed in all life stages and exhibiting a stronger affinity for cruzipain than for other cysteine proteases [4]. Chagasin shares no significant sequence similarity with other protease inhibitors, or indeed with any other protein, and so represents the first member of a new class of protease inhibitor. Here we show by threading and model-building studies that chagasin adopts an immunoglobulin-type fold, the first protease inhibitor known to do so. Furthermore, a small number of significant sequence characteristics conserved between chagasin and variable light chain domain sequences suggest that chagasin arose as a consequence of a rare host to parasite horizontal gene transfer.

2. Materials and methods

Attempts to locate structural homologues using PSI-BLAST [5] failed to produce any significant results. Threading experiments were therefore carried out using several WWW servers. The methods of Fischer and Eisenberg [6] and Rice and Eisenberg [7] were applied at the UCLA fold recognition server (http://fold.doe-mbi.ucla.edu). The GenTHREADER program [8] was run at http://globin.bio.warwick.ac.uk/psipred. Fold recognition through the use of hidden Markov models [9] was carried out at http://www.cse.ucsc.edu/research/compbio/HMM-apps. An initial alignment of the chagasin sequence with templates was constructed with CLUSTAL W [10], manually modified to reflect the threading results and displayed with ALSCRIPT [11].

Protein models were constructed using the MODELLER-4 package [12]. Fifteen models were made and evaluated for each alignment tested. Several programs were used for the rigorous evaluation of the protein models. PROCHECK [13] was used to monitor stereochemical quality while PROFILER_3D [14] and PROSA II programs [15] were used to measure overall protein quality in terms of packing and solvent exposure. These programs both produce both overall scores for a given structure which, for a given protein length, should be in known positive and negative ranges for PROFILER_3D and PROSA II respectively, and running profiles to enable the localisation of errors resulting from, for example, incorrect alignments.

The program O [16] was used for inspection and manipulation of models and for secondary structure definition of the models. The secondary structure of chagasin was predicted using the PHD program [17] at http://www.embl-heidelberg.de/predictprotein. Three-dimensional superposition of proteins was carried out using LSQMAN [18].

3. Results

Threading results for the chagasin sequence unanimously suggested that, of the structures contained in the present PDB, an immunoglobulin-like fold, and more specifically the fold of the antibody light chain variable domain, would provide the best template for model construction. By the methods of Fischer and Eisenberg [6], only the variable light chain domains of 1lmk and 1vfa, scoring 6.97 and 5.92 respectively, exceeded the quoted significance threshold of 4.8 ± 1.0. These were used as templates. Using GenTHREADER [8], the 10 quoted results were all variable light chain domains, the top four of which, with probabilities of 0.87–0.79, were interpreted as being of 'high' confidence. The best two scoring hits, 2hrp and 1fvc, were added to the templates. Similarly, the only variable light chain domain apparently present in the HMM database [9], 1nqb, came top of the results with a score

*Corresponding author. Fax: (55)-61-340 3658.
E-mail address: daniel@cenargen.embrapa.br (D.J. Rigden).

PII: S0014-5793(01)02753-3

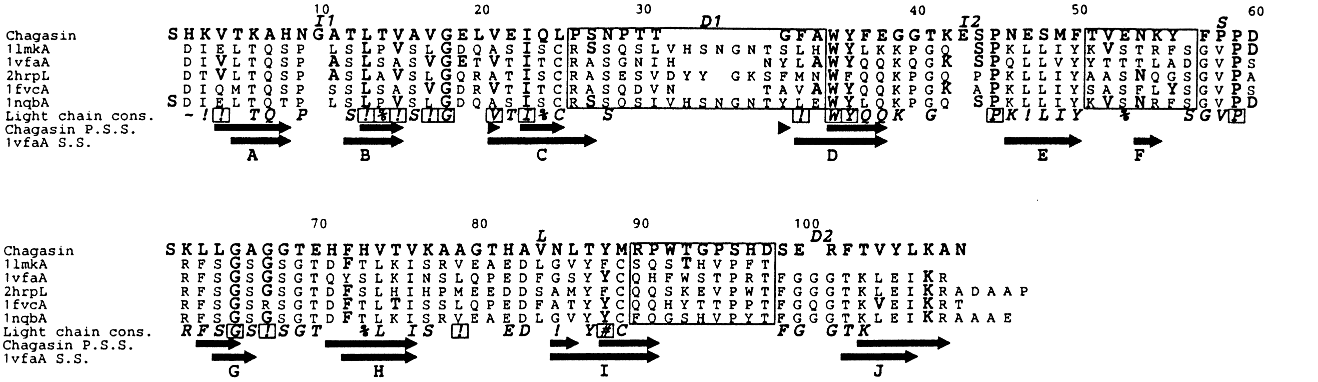
10
20
30
40
50
60
I1
D1
I2
S
Chagasin SHKVTKAHNGATLTVAVGELVEIQLPSNPTT GFAWYFEGGTKESPNESMFTVENKYFPPD
1lmkA DIELTQSP LSLPVSLGDQASISCRSSQSLVHSNGNTSLHWYLKKPGQ SPKLLIYKVSTRFSGVPD
1vfaA DIVLTQSP ASLSASVGETVTITCRASGNIH NYLAWYQQKQGK SPQLLVYYTTTLADGVPS
2hrpL DTVLTQSP ASLAVSLGQRATISCRASESVDYY GKSFMNWFQQKPGQ PPKLLIYAASNQGSGVPA
1fvcA DIQMTQSP SSLSASVGDRVTITCRASQDVN TAVAWYQQKPGK APKLLIYSASFLYSGVPS
1nqbA SDIELTQTP LSLPVSLGDQASISCRSSQSIVHSNGNTYLEWYLQKPGQ SPKLLIYKVSNRFSGVPD
Light chain cons. ~!! TQ P S!%!S!G VTI%C S ! WYQQK G PKILIY % SGVP
Chagasin P.S.S.
1vfaA S.S.
A
B
C
D
E
F
70
80
90
100
L
D2
Chagasin SKLLGAGGTEHFHVTVKAAGTHAVNLTYMRPWTGPSHDSE RFTVYLKAN
1lmkA RFSGSGSGTDFTLKISRVEAEDLGVYFCSQSTHVPFT
1vfaA RFSGSGSGTQYSLKINSLQPEDFGSYYCQHFWSTPRTFGGGTKLEIKR
2hrpL RFSGSGSGTDFSLHIHPMEEDDSAMYFCQQSKEVPWTFGGGTKLEIKRADAAP
1fvcA RFSGSRSGTDFTLTISSLQPEDFATYYCQQHYTTPPTFGQGTKVEIKRT
1nqbA RFSGSGSGTDFTLKISRVEAEDLGVYYCFQGSHVPYTFGGGTKLEIKRAAAE
Light chain cons. RFSGS!SGT %L IS ! ED ! Y#C FG GTK
Chagasin P.S.S.
1vfaA S.S.
G
H
I
J

Fig. 1. Sequence alignment of chagasin with template proteins. The chagasin sequence and positions that agree with it in the templates are emboldened. The variable light chain sequence consensus is shown beneath the alignment (~ signifies a charged residue, ! a hydrophobic residue, % a hydroxyl-containing residue and # an aromatic residue) with positions at which the chagasin sequence agrees boxed. Below are the chagasin secondary structure prediction and the actual secondary structure of the template determined to highest resolution (1vfa). Regions of the alignment corresponding to CDRs in the templates are boxed. Above the alignment are labelled regions in which the backbone of the chagasin model differs significantly from the templates (see text and also Fig. 2).

of −8.79 and was added to the templates. The H3P2 method [7] also favoured immunoglobulin folds with all the quoted 30 folds being immunoglobulin domains and the top 10 all variable light chain domains. The existing templates also ranked highly in H3P2 output so that no new templates were deemed necessary. We also searched for other variable light chain structures that might prove better templates for model construction in the chagasin regions corresponding to complementarity determining regions (CDRs) of the antibody domains, but found none. The five chosen templates were all solved recently (1992–1997) by X-ray crystallography to reasonable resolution (1.8–2.6 Å).

The sequence alignment of chagasin with the five variable light chain domains showed an excellent agreement between the predicted secondary structure for chagasin and the observed secondary structure in the templates, just one insertion and one deletion, and some other intriguing features. First, the tryptophan, invariant among variable light chains, and many other classes of immunoglobulin fold [19], is present in the chagasin sequence (Trp 35). At seven other positions where the consensus sequence specifies a particular amino acid, the chagasin sequence conforms. Two of these residues, Ile 23 and Gly 65, pack against Trp 35, while another, Gly 18, has a positive ϕ angle and occupies a structurally important position in a β-turn. There are a further nine positions where the chagasin sequence conforms to a particular type of amino acid in the variable light chain consensus. The agreement of the chagasin sequence with the variable light chain consensus is better towards the N-terminus with 11 of the 17 agreements in the first third of the sequence. Among the notable differences between the chagasin and consensus sequences was the lack of the consensus invariant disulphide bond in the chagasin sequence.

Based on these results, we embarked on the comparative protein modelling of chagasin. The 12–17% sequence identity observed between chagasin and the templates meant that the alignment needed to be subject to careful scrutiny in order to produce the best possible model [20]. We made three alignment improvements using multiple model construction and evaluation of the models, by PROSA II energy profiles, to validate each one [20,21]. Overall PROSA II scores and rankings also improved throughout this process. The first alignment modification made was a readily accommodated deletion, relative to the templates, at position 101 (*D2* in Figs. 1 and 2). As well as improving the packing characteristics of the C-terminus, this removed from the structural alignment a conserved template glycine occupying a disallowed area of the Ramachandran plot. The second change was an alignment shift of five residues around position 58 (*S* in Figs. 1 and 2). This aligned the PD sequence in chagasin with the PD sequence seen in two templates and introduced a small stretch of α-helix into the model (Fig. 2). We next shifted the alignment near the N-terminus by introducing an insertion, relative to the templates, between strands A and B (*I1* in Figs. 1 and 2). Inspection of the templates showed that this could be accommodated either after position 9 or after position 10. We therefore compared the models deriving from both alignment variants and found the latter better.

Three remaining problems were successfully addressed by database loop searches using O or MODELLER-4. These were the unusual position of Thr 32 in a Ramachandran plot, the *cis* amino acid Ser 97 resulting from alignment opposite a conserved *cis* proline, and the poor packing around position 82. We replaced the region around 82 with a particularly favourable stretch found in another cycle of model construction (*L* in Figs. 1 and 2). The resulting stretch had an almost ideal type I (α_R,α_R) turn [22] at positions $_{80}$GTHA$_{83}$ that enabled the favourable packing of the following Val 84 against valines 21, 76 and 104. Finally we dealt with Thr 70 which occupied a disallowed area of the Ramachandran plot due to its alignment with a conserved glycine in the templates. This glycine was the second residue in a β-turn. The type of β-turn at $_{68}$GTEH$_{71}$ in the chagasin model was therefore changed to type I (α_R,α_R), this being the most common type overall [22] and having Thr and Glu as commoner than average amino acids for the second and third positions respectively [23]. Alternatively, it would have been possible to align the GTE sequence with the GTD region seen in four templates. However, examination of the structure showed that the following Phe 72 appeared to be crucial for hydrophobic packing [24] so that this was seen as an example of a misleading local sequence alignment [25].

The best PROSA II scoring structure from this round of model generation was taken as the final model. Several key

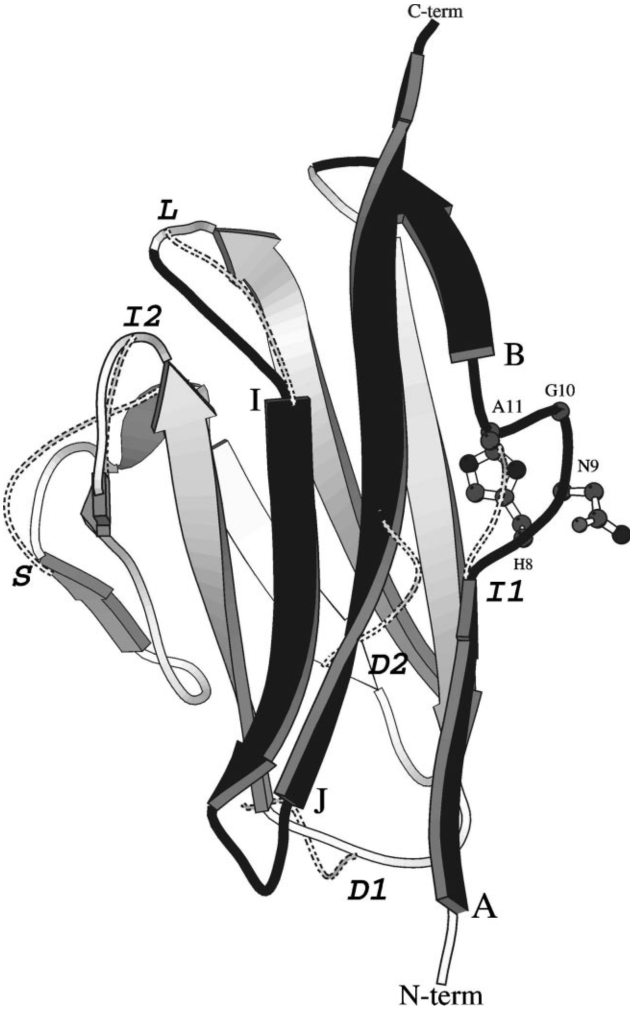

Fig. 2. Molscript [42] diagram of the final chagasin model. Parts of the sequence corresponding to the best three inhibitory peptides are drawn black and the β-strands within them labelled. Where the main chain of the chagasin model differs significantly from the templates the template main chain is shown as thin dotted tubes. These regions are labelled as in Fig. 1. Side chains of residues of the tetrapeptide ${}_8$HNGA$_{11}$, also seen in the propeptide of cruzipain, are shown in ball and stick representation and labelled, as are the N- and C-termini.

protein quality indicators for the final model are compared with the ranges seen for the template structures in Table 1. All of these indicators either fall in the range seen for templates or approach the templates' scores. Val 52 is the chagasin model residue in a disallowed area of the Ramachandran plot. The corresponding residue in the templates is always disallowed and may be characteristic of the fold. It is worth noting that the worst template performers in the different categories are distributed among three proteins so that there is not one overall structure of poor quality. Previous threading-based models have yielded protein quality indicators comparable to those of NMR structures [26,27], falling some way behind structures determined by X-ray crystallography, so these results are encouraging.

Of 13 overlapping 15-residue peptides spanning the chagasin sequence, six inhibited cruzipain with true K_i values in the range 1.7–9.9 μM (A.C.S. Monteiro, unpublished data). Examination of the final model showed that the three peptide sequences with the lowest K_i values are localised together on one face of the molecule (Fig. 2). They comprise the four β-strands A, B, I and J along with some loop regions. The peptide with the fourth highest affinity overlaps partly with the best inhibitory peptide while the weakest binding sequences are localised on other faces of the molecule. The better binding peptides thus define a putative site of interaction of chagasin with proteases.

The tetrapeptide sequence HNGA, present in chagasin from positions 9 to 12, is also present in cruzipain's own inhibitory propeptide and has been shown to be essential for inhibition by propeptide-derived sequences [28]. Given that the HNGA-containing peptide tested here also showed inhibitory activity, it seemed possible that this sequence might bind to the enzyme in the same way, irrespective of chagasin or propeptide context. Also supporting this idea is the highly exposed position of this region in the final model (Fig. 2). This arises largely from the insertion *I1*, but this is fully justified by comparison of models with and without the insertion. The structures of cruzipain available in the PDB [2] are of the mature, cleaved enzyme. We carried out threading experiments with the full cruzipain sequence but different methods produced different alignments for the sequence HNGA. Thus it appears that the propeptide-containing protease crystal structures of the PDB may not provide a good structural model for this part of the cruzipain propeptide. It is therefore difficult to say whether the presence of HNGA in both propeptide and chagasin is anything more than coincidence.

4. Discussion

Both the unanimity of the threading results and the strong performance of the final chagasin model, against a battery of protein verification tools, support the assignment of an immunoglobulin-type fold to chagasin. The model gains additional experimental support from the peptide inhibition data presented which also define a putative site of interaction of chagasin with proteases.

The variety of functions of immunoglobulin-like domains has long been apparent [19], with more recent additions to the list being a DNA-binding protein [29], a G-protein modulator [30], cell adhesion molecules [31], nerve proteins [32] and muscle proteins [33,34]. Chagasin is the first protease inhibitor known to adopt an immunoglobulin-like domain and there-

Table 1
Comparison of key protein quality scores for the final chagasin model with the ranges of scores seen among the templates

	Chagasin structure	Range in templates	
		worst	best
PROSA II scores			
Overall	−6.75	−6.21	−7.56
Pair potential	−3.37	−4.15	−5.13
Surface potential	−6.17	−5.05	−6.24
PROFILER_3D score	37.9	41.3	59.1
PROCHECK			
Ramachandran plot			
% of residues in core areas	87	86	92
Number in disallowed areas	1	1	1
G-factors			
Main chain	−0.80	−0.68	−0.35
Overall	−0.30	−0.40	0.24
Side chains			
Number of χ_1,χ_2 outliers	2	3	0

fore represents a further addition to the list of immunoglobulin-like domain functions.

When compared with variable light chain domains and other immunoglobulin structures, the predicted chagasin structure is unusual in several ways. The length of the chagasin loop corresponding to CDR1 in the templates is nine residues, in contrast to the 10–17 residues seen in variable light chain structures. This reduced length can be smoothly accommodated in the fold (*D1* in Fig. 2). Absence of a disulphide bond is unknown among natural variable light chain domains, although not among immunoglobulin folds as a whole [19], but recent crystallographic results show that the variable light chain fold can survive intact when the disulphide bridge is artificially removed by mutation [35]. The binding site for cruzipain suggested by the peptide inhibition data involves mainly β-strands and only a small part of the third loop aligning with CDR3. Classically, β-sheets have been observed to be used for interactions between immunoglobulin domains and loop regions for other ligands [19], although this binding repertoire continues to expand [36].

Given the vestiges of sequence similarity observed between chagasin and variable light chains, and their concentration at structurally important positions, the most likely origin for the chagasin gene appears to be a host animal. Host to parasite horizontal gene transfer has not often been reported but at least one example, involving a semiparasitic mite, is known [37] and viruses have been highlighted as possible mediators of such transfers [38]. For trypanosomatids, chagasin appears to be the first known example, although horizontal transfer from procaryotes to trypanosomatids has been strongly suggested in the cases of cytosolic glyceraldehyde-3-phosphate dehydrogenase [39] and glycerol-3-phosphate dehydrogenase [40]. Interestingly, eucaryote to bacterium horizontal gene transfer has been suggested for the immunoglobulin module of bacterial sialidase [41].

References

[1] Murta, A.C.M., Persechini, P.M., Padron, T. de S., de Souza, W., Guimarães, J.A. and Scharfstein, J. (1990) Mol. Biochem. Parasitol. 43, 27–38.

[2] McGrath, M.E., Eakin, A.E., Engel, J.C., McKerrow, J.H., Craik, C.S. and Fletterick, R.J. (1995) J. Mol. Biol. 247, 251–259.

[3] Calkins, C.C. and Sloane, B.F. (1995) Biol. Chem. Hoppe Seyler 376, 71–80.

[4] Monteiro, A.C.S., Scharfstein, J., Abrahamson, M. and Grossi de Sá, M.F. (2000) in: Abstr. 6th Int. Congress Plant Mol. Biol., pp. S16–S27.

[5] Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Nucleic Acids Res. 25, 3389–3402.

[6] Fischer, D. and Eisenberg, D. (1996) Protein Sci. 5, 947–955.

[7] Rice, D.W. and Eisenberg, D. (1997) J. Mol. Biol. 267, 1026–1038.

[8] Jones, D.T. (1999) J. Mol. Biol. 287, 797–815.

[9] Karplus, K., Barrett, C. and Hughey, R. (1998) Bioinformatics 14, 846–856.

[10] Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) Nucleic Acids Res. 11, 4673–4680.

[11] Barton, G.J. (1993) Protein Eng. 6, 37–40.

[12] Šali, A. and Blundell, T.L. (1993) J. Mol. Biol. 234, 779–815.

[13] Laskowski, R., MacArthur, M., Moss, D. and Thornton, J. (1993) J. Appl. Crystallogr. 26, 283–290.

[14] Lüthy, R., Bowie, J.U. and Eisenberg, D. (1992) Nature 356, 83–85.

[15] Sippl, M.J. (1993) Proteins 17, 355–362.

[16] Jones, T.A., Zou, J.-Y., Cowan, S.W. and Kjeldgaard, M. (1991) Acta. Crystallogr. A 47, 110–119.

[17] Rost, B. (1996) Methods Enzymol. 266, 525–539.

[18] Kleywegt, G.J. (1996) Acta Crystallogr. Sect. D Biol. Crystallogr. 52, 842–857.

[19] Bork, P., Holm, L. and Sander, C. (1994) J. Mol. Biol. 242, 309–320.

[20] Martin, A.C.R., MacArthur, M.W. and Thornton, J.M. (1997) Proteins 1 (Suppl.), 14–28.

[21] Sánchez, R. and Šali, A. (1997) Proteins 1 (Suppl.), 50–58.

[22] Wilmot, C.M. and Thornton, J.M. (1990) Protein Eng. 3, 479–493.

[23] Wilmot, C.M. and Thornton, J.M. (1988) J. Mol. Biol. 203, 221–232.

[24] Poupon, A. and Mornon, J.P. (1998) Proteins 33, 329–342.

[25] Saqi, A.S.M., Russell, R.B. and Sternberg, M.J.E. (1998) Protein Eng. 11, 627–630.

[26] Rigden, D.J. and Carneiro, M. (1999) Proteins 37, 697–708.

[27] Rigden, D.J., Mello, L.V. and Bertioli, D.J. (2000) Proteins 41, 133–143.

[28] Lalmanach, G., Lecaille, F., Chagas, J.R., Authie, E., Scharfstein, J., Juliano, M.A. and Gauthier, F. (1998) J. Biol. Chem. 273, 25112–25116.

[29] Ghosh, G., van Duyne, G., Ghosh, S. and Sigler, P.B. (1995) Nature 373, 303–310.

[30] Heep, N.H., Barnes, M., Barsukov, I., Badii, R., Lian, L.Y., Segal, A.W., Moody, P.C.E. and Roberts, G.C.K. (1997) Structure 5, 623–633.

[31] Jones, E.Y., Harlos, K., Bottomley, M.J., Robinson, R.C., Driscoll, P.C., Edwards, R.M., Clements, J.M., Dudgeon, T.J. and Stuart, D.I. (1995) Nature 373, 539–544.

[32] Shapiro, L., Doyle, J.P., Hensley, P., Colman, D.R. and Hendrickson, W.A. (1996) Neuron 17, 435–449.

[33] Improta, S., Politou, A.S. and Pastore, A. (1996) Structure 4, 323–337.

[34] Fong, S., Hamill, S.J., Proctor, M., Freund, S.M., Benian, G.M., Chothia, C., Bycroft, M. and Clarke, J. (1996) J. Mol. Biol. 264, 624–639.

[35] Uson, I., Bes, M.T., Sheldrick, G.M., Scheider, T.R., Hartsch, T. and Fritz, H.J. (1997) Fold. Des. 2, 357–361.

[36] Newton, J.P., Buckley, C.D., Jones, E.Y. and Simmons, D.L. (1997) J. Biol. Chem. 272, 20555–20563.

[37] Houck, M.A., Clark, J.B., Peterson, K.R. and Kidwell, M.G. (1991) Science 253, 1125–1128.

[38] Damian, R.T. (1997) Parasitology 115, S169–S175.

[39] Michels, P.A., Marchand, M., Kohl, L., Allert, S., Wierenga, R.K. and Opperdoes, F.R. (1991) Eur. J. Biochem. 198, 421–428.

[40] Kohl, L., Drmota, T., Thi, C.D., Callens, M., Van Beeumen, J., Opperdoes, F.R. and Michels, P.A. (1996) Mol. Biochem. Parasitol. 76, 159–173.

[41] Gaskell, A., Crennell, S. and Taylor, G. (1995) Structure 3, 1197–1205.

[42] Kraulis, J. (1991) J. Appl. Crystallogr. 24, 946–950.