Cysteine proteases of hookworm *Necator Americanus* as virulence factors and implications for future drug design with heparin analogs: A bioinformatics-based study.

Arpita Banerjee❖, Ruben Abagyan❖

❖ Skaggs School of Pharmacy & Pharmaceutical Sciences,
9500 Gilman Drive.
University of California, San Diego
San Diego, CA – 92093-0719
U.S.A
Email: a4banerjee@ucsd.edu, arpita.005@gmail.com, ruben@ucsd.edu

Keywords: Cysteine proteases, Necator Americanus, Hookworm, Bioinformatics, Hemoglobinase motif, Kallikrein-like activity, Heparin binding.

Author Contributions:
Design of analysis: AB
Formation of analysis: AB
Software development: RA
Original writing and draft preparation: AB
Writing review and editing: RA

Abbreviations:

NA:	Necator Americanus
CPs:	Cysteine proteases.
SmSP1:	Schistosoma mansoni Serine Protease1
HMWK	High Molecular Weight Kininogen
T.Cruzi	Trypanosoma Cruzi
S.mansoni:	Schistosoma mansoni
P.falciparum	Plasmodium falciparum
P.Westermani	Paragonimus westermani
C. elegans	Caenorhabditis elegans
L. major	Leishmania major
A. Caninum	Ancylostoma caninum
S.japonicum	Schistosoma japonicum
H. Contortus	Haemonchus contortus
O. ostertagi	Ostertagia ostertagi
A. Suum	Ascaris suum

Abstract:

Human hookworm *Necator Americanus* causes iron deficiency anemia, as the parasite ingests blood from the gastrointestinal tract of its human host. The virulence factors of this blood feeding nematode have not been researched extensively. This bioinformatics based study focuses on eight of the cathepsin B like cysteine proteases (CPs) of the worm, which could have immense pathogenic potential. The role of the individual CPs remain vaguely determined except for CP3 which has been shown to act as globinase in the hemoglobin degradation pathway. In this study, the cysteine proteases were subjected to predictive molecular characterizations viz: probability of extracellular secretion to the interface between pathogen and host, ability for hemoglobin degradation, and/or interaction with host plasma proteins. CP1 - CP6, which harbored the active site cysteine and were observed to have N terminal signal peptide for extracellular localization, were relevantly predicted to be secretory. Amongst these, CP2 and CP3 showed the presence of hemoglobinase motif derived in this study that could be a prerequisite for globin or hemoglobin degradation. CP1 and CP6 that shared similar S2 subsite features with kininogen-cleaving cathepsin B and cruzain, suggests the possible involvement of these two hookworm CPs in binding and cleaving High Molecular weight kininogen (HMWK). CP1, CP2, CP3, CP5 and CP6 were predicted to bind heparin, which is the glycosaminoglycan molecule that has been demonstrated to aid the substrate-cleaving functionality of other cysteine proteases like human cathepsin B and cruzain. Heparin docked onto the *NA* CPs at the C terminal domain, away from the active site, similar to what has been shown for heparin binding to cathepsin B and cruzain. These observations therefore lead us to the hypothesize that the functions of the hookworm CPs, which could include preventing blood clot formation, might be assisted by heparin. Also, this study underscores the potential of synthetic heparin analogs as molecular treatment for hookworm infection, which could have implications for future drug design.

Introduction:

Hookworm infection in humans is a neglected tropical disease that affects over 700 million people worldwide, mostly in the developing countries of the tropical and subtropical regions. (Hotez *et al*, 2004; Hotez *et al*, 2010) *Necator Americanus (NA)* species of hookworm constitutes the majority of these infections (~85%) (Loukas *et al*, 2011).

The clinical manifestation of the disease includes anemia, malnutrition in pregnant women, and cognitive and/or physical development impairment in children (Diemert *et al*, 2008). These helminth blood feeders on reaching maturity can feed up to 9ml of blood per day in an infected individual by attaching themselves to the intestinal mucosa of the host, through cutting plates as in *NA* (Pearson *et al* 2012). Iron deficiency anemia is the direct effect of the hookworm's blood feeding (Kassebaum *et al*, 2014), resulting in other subsidiary consequences of the hookworm disease (Sakti *et al* 1999; Hotez *et al*, 2008).

The infective larval stage (L3) worm penetrates into the host skin from soil (Vetter *et al*, 1977; Vetter *et al*, 1977) and then invades the circulatory system to reach heart and lungs, wherefrom it migrates to alveoli and then to trachea. The parasite eventually reaches gastrointestinal tract as fourth stage larvae (L4) to develop into blood feeding adult stage hookworms (Hotez *et al*, 2004).

An array of diverse enzymes and molecules in *NA*'s biomolecule repertoire facilitate the pathogen's survival in the host for up to seven years or longer during the different stages of its lifecycle (Pearson *et al* 2012). The most important therapeutic targets are the enzymes involved in interaction with host and in nutrient acquisition. These are often found in the excretory-secretory (ES) products of the worm. The ES proteins have been shown to engage in crucial functions like tissue degradation for host invasion (Pearson *et al* 2012, Brooker *et al*, 2004), fibrinogen degradation (Brown *et al*, 1995) for preventing blood clots, hemoglobin degradation (Brown *et al*, 1995) for nutrient acquisition, and evasion of the host immune system (Bungiro *et al*, 2011). The enzymes in the ES potpourri of *NA* are not yet completely characterized. However, Brown *et al*, had reported cysteine and serine proteases from ES products to degrade hemoglobin and fibrinogen, and detected the presence of at least two cysteine proteases operating at different pH optima (Brown *et al*, 1995).

The cysteine proteases (CPs) in *NA* form the ninth most gut-expressed abundant gene family (Ranjit *et al* 2006) These are most similar to *H.Contortus* (a blood feeding ruminant) CPs that dominate (~16%) intestinal transcriptome of the barber pole worm (Jasmer *et al*, 2001), highlighting the importance of pathogenic CPs in host blood degradation. The CPs are synthesized as precursor molecules, where in their folded proenzyme form, a self-inhibitory peptide blocks their catalytic domain. The removal of the inhibitory peptide upon proteolytic

action of other peptidases releases the mature cysteine proteases, which are active. The enzymatic clefts of the NA-CPs contain the catalytic triad residues: Cysteine, Histidine and Asparagine. Classification of cysteine proteases relies on the sequence homology spanning the catalytic residues (Sajid *et al*, 2002). CPs of parasitic organisms are divided into clans CA and CD. The clans are further classified into families in which Cathepsin B-like proteases belong to C1. The *NA*-CPs which are cathepsin B-like, therefore belong to the clan CA C1 family of cysteine proteases, which have a mostly α-helical amino-terminal known as the L domain, and the antiparallel β strands dominated carboxy-terminal is known as the R domain. The active site for proteolytic degradation is at the interface of these L and R domains. The entire cascade of hemoglobin degradation in hookworm has not been elucidated. However, certain key enzymes have been shown to participate in the degradation pathway. Aspartic protease *NA*-APR1 acts on hemoglobin, whereas metalloprotease MEP1 and cysteine protease CP3 degrade globin fragments (Ranjit *et al* 2009). While ingestion and digestion of blood, anticoagulant proteins are secreted to prevent clot formation (Stanssens *et al*, 1996; Harrison *et al*, 2002; Furmidge *et al*, 1995). Cysteine proteases, from *NA* (Brown *et al*, 1995) and from phylogenetically close *H. contortus* (Cox *et al*, 1990) have been implicated to have anticoagulation properties (Brown *et al*, 1995). *NA* adopts a number of complementary strategies to evade host procoagulation system (Furmidge *et al*, 1995), few of which have been elucidated so far.

Taken together, cysteine proteases from different pathogenic organisms perform diverse functions pertaining to blood feeding. This study focuses on eight cysteine proteases viz: CP1, CP2, CP3, CP4, CP4b, CP5, CP6, CP7 encoded by *NA* genome; of which CP2, CP3, CP4 and CP5 genes are reportedly expressed in abundance in the gut tissue of the adult worm (Ranjit *et al* 2008). Only CP3 amongst these has been characterized as a globinase (Ranjit *et al* 2009). The expression of the other CP genes in the *NA* gut is suggestive of their involvement in digestive or other assistive functions. Despite *NA*-CPs' importance in the parasite's physiology, they are under-researched and not much is known about the individual proteases and which of these constitute the ES products of *NA*. This bioinformatics based study on the molecular characterization of the CPs probes into those aspects of the proteases, which could have pathogenic potential such as hemoglobin degradation and blood clot prevention. Several bioinformatics methodologies have been applied ranging from sequence-based predictive methods, homology modeling, docking, motif derivation from sequence patterns, and mapping of molecular interactions to elucidate the role of the CPs as possible virulence factors and hence target for therapeutics. The approaches to some of the methods adopted here are in the context of other relevant cysteine proteases. Implications for the usage of heparin-analogs as worm-inhibitors have been outlined based on docking of heparin in the *NA*-CPs.

Materials and Methods:

1. Sequence alignments and pattern detection:

The following *NA* cysteine protease sequences were retrieved for analyses from Uniprot protein sequence database (The UniProt Consortium, 2015)(Uniprot ID in parentheses): Necpain or CP1 (Q9U938), CP2 (A1YUM4), CP3 (A1YUM5), CP4 (A1YUM6), CP4b (W2TRZ7), CP5 (A1YUM7), CP6 (W2T0C4) and CP7 (W2SQD9) (the organism code part of the ID is omitted for brevity). The alignments were done in ICM (Abagyan *et al*, 1994) by BLOSUM62 scoring matrix, with gap opening penalty of 2.40 and gap extension penalty of 0.15. Patterns of relevance and predicted or deciphered sites of functional importance were mapped onto the aligned CP sequences to denote their positions in the protein sequences.

1.1 N-terminal signal for protein localization:

The N-terminals of protein sequences often hold clue to protein sorting (Blobel, 1980) which determine the localization of the proteins. CP1-6 harbored pre-sequences at their N-terminals unlike CP7. These pre-sequences were used to derive PRATT (Jonassen *et al*, 1995) pattern unique to CP1-6, which was then searched in CP7 by ScanProsite (Sigrist *et al*, 2002) to see and confirm if the protease lacks the pattern. This was done to delineate the N-terminal signal of CP1-6 from that of CP7, which could lead to the proteases' differential localization.

1.2 Cleavage of the N-terminal signal peptide:

The N-terminal signal peptides of cysteine proteases are cleaved off to generate the zymogens (Turk *et al*, 2012). SignalP (Petersen *et al*, 2011) was used to predict cleavage sites in the hookworm *NA* CPs to see and confirm if the N-terminal cleavage products span the signal peptide pre-sequences of the proteases.

1.3. Localization guided by signal peptide:

The *NA*-CP sequences were submitted for localization prediction to TargetP (Emanuelsson *et al*, 2007), iPSORT (Bannai *et al*, 2002), TMHMM (Krogh *et al*, 2001), LocSigDB (King *et al*, 2007), Bacello (Pierleoni *et al*, 2006), Protein Prowler (Bodén *et al*, 2005), Cello (Yu *et al*, 2006) and PrediSi (Hiller *et al*, 2004) webservers to determine which of the proteases would be prone to secretion. While the algorithms for most of these programs take N-terminal signals into account, Bacello predicts localization on the basis of the information contained in the folded protein and LocSigDB is a signature pattern database derived from proteins whose localization has been confirmed by experiments on bench.

1.4 Hemoglobinase motif for hemoglobin-degradation functionality:

The incompletely elucidated hemoglobin degradation pathway in *NA* describes the role of only CP3 as a globinase, amongst other CPs in the cysteine protease repertoire. In an effort to investigate the involvement of the rest of the *NA* CPs in hemoglobin degradation, cysteine protease sequences from other organisms - known to degrade hemoglobin - were taken along with *NA*-CP3 to derive conserved patterns unique to these proteins. Those organisms, the proteins, and their genbank (Benson *et al*, 2013) accession numbers (in parenthesis) are: *Necator Americanus* CP3 (ABL85237.1), *A. Caninum* CP1 (Q11006), *A. Caninum* CP2 (AIG62903.1), *S.mansoni* CB1 (3QSD_A), *P.falciparum* falcipain2 (AAK06665.1), *P.falciparum* falcipain 3(KOB61544.1), *S.japonicum* Cathepsin B (P43157.1), *H. Contortus* AC3 (Q25032), *H.Contortus* AC4 (Q25031), *P.Westermani* CP1 (AAF21461.1) and *O. ostertagi* CP1 (P25802.3) and *A. Suum* CP (AAB40605.1) Conserved patterns from the aforementioned proteins were derived in PRATT (Jonassen *et al*, 1995) and the motifs were scanned against some other non-hemoglobin degrading proteins in ScanProsite (Sigrist *et al*, 2002) to pinpoint patterns specific to the hemoglobin degrading enzymes. The set of non-hemoglobin degrading organisms and their relevant proteins were: *C. elegans*_CPR3 (AAA98789.1), *C.elegans*_CPR4 (AAA98785.1) and *L. major* cathepsin B (AAB48119.1). Such derived patterns (when found) specific to the hemoglobin degrading enzymes were scanned in the rest of the *NA* CP sequences.

1.5 Kallikrein-like activity of cleaving kininogen for implicated vasodilation and anticoagulation:

Kallikrein-like activity of cleaving high molecular weight kininogen (HMWK) has been reported for cruzain, which is a cysteine protease of *T.Cruzi* – a parasite that traverses blood capillary vessels as trypomastigotes (Del Nery *et al*, 1997; Lima *et al,* 2002). Such kininogen-cleaving activity of cruzain generates Lys-bradykinin, a potent vasodilator that releases PGI_2 - which inhibits platelet activation and degranulation (Maurer *et al*, 2011). As *NA* larvae too traces migratory route through blood capillaries of human hosts, the cysteine proteases of the human hookworm were scrutinized for molecular features which could be responsible for kallikrein-like activity. This was done by aligning the cathepsin B-like *NA* CPs with cathepsin B (PDB: 1HUC) and cruzain, a cathepsin L (PDB: 2OZ2) – both of which share similar substrate specificity that has been attributed to a critical Glu placed at their S2 subsites (Gillmor *et al*, 1997), and both cleave HMWK (Barros *et al*, 2004, Del Nery *et al*, 1997). The sequences were aligned in ICM with the same parameters as mentioned before. The distances between the C-alpha atoms of relevant S2 subsite residues within the enzymatic cleft of the protease structures were measured in ICM for estimating the possible kininogen-cleaving role of the crucial residues (if any) in the *NA* CPs.

2. Heparin binding for transportation of indigenous worm glycosaminoglycans, and molecular recognition of host proteins:

Glycosylation with glycosaminoglycans (GAGs) like heparin is an important post-translational modification in cysteine proteases of many pathogenic organisms (Sajid *et al*, 2002). The CPs of *NA* hookworm were therefore scanned for sequence motifs that could undergo such modification. Also, glycosaminoglycan molecules remain displayed on host proteoglycans (Bartlett *et al,* 2010*)* and are probably encountered for recognition by the worm ES products at the host-pathogen interface. The feasibility of such recognition of host proteins through the latter's GAG interaction with the hookworm's CPs was explored by *NA* CP - heparin dockings.

2.1 Heparin-binding motif for post-translational glycosylation and transportation of worm GAGs to extracellular matrix:

The sequences were scanned in Scanprosite (Sigrist *et al*, 2002) for searching putative glycosylation motifs, which could be the co-valent attachment sites for pathogenic glycosaminoglycan (indigenous to the worm) like heparin. Such glycosylation sites were searched to explore the possibility of (worm) heparin, with anticoagulation properties, being transported out to extracellular matrix while attached to secretory *NA* CPs (if any), for impairing coagulation.

2.2 Heparin binding domain for molecular recognition of host proteins:

The sequences were subjected to query by ProDom (Bru *et al*, 2005), a protein domain family database derived from Uniprot knowledgebase. The search was carried out for the purpose of finding any heparin-binding protein domains on the *NA* CPs, as such sites would be predisposed to non-covalently bind heparin attached to host proteins

2.3. Heparin binding docking simulation for molecular recognition of host proteins:

Homology models of the *NA*_cysteine proteases_were built to carry out the CP – heparin dockings, due to lack of_experimental three-dimensional structure of the CPs. BLAST search for finding homology model templates was performed against Protein Data Bank (PDB) (Berman *et al,* 2000). 3QSD - mature CathepsinB1 enzyme from *Schistosoma Mansoni* - was chosen from the search results for building models as it aligned well at the active site and had a resolution of 1.3 Å. Also, this structure had co-ordinates for the two occluding loop residues near the active site, which have been designated crucial for the exopeptidase activity for this class of cathepsin B like proteases (Illy *et al,* 1997). Homology models were built within the internal co-ordinates mechanics protocol of ICM software (Abagyan *et al,* 1994). The sequence alignment between the template and the model sequence was generated by using BLOSUM62 matrix, with gap opening penalty of 2.40 and gap extension penalty of 0.15. Further, for generating reliable models, the alignment around the active site was edited wherever necessary, according to conservation propensity of residues, and for modeling the occluding loop residues. Loops were sampled for the alignment

gaps where the template did not have co-ordinates for the model sequence. The loop refinement parameters were used according to the default settings of the procedure. Acceptance ratio during the simulation was 1.25. The *NA*-CP model structures were then built within the full refinement module of the software. The quality of the homology models were checked using PROCHECK (Laskowski *et al*, 1993) which showed 100% of the residues from most of the CPs to lie within the allowed regions of the Ramachandran plot. CP2 and CP5 were the exceptions, which had 99.5% of residues in the allowed regions.

The co-ordinates for heparin were taken from its complex deposited in PDB (ID: 5D65) and saved as SDF formatted ligand. The CP homology models were converted to ICM formatted receptors for docking the heparin molecule. The sequence stretch of the CPs encompassing the predicted fibronectin domain, which can putatively bind heparin like molecules (Pankov *et al,* 2002) was selected for docking the heparin tetrasachharide which had alternating units of N, O6-disulfo-glucosamine (PDB ID: SGN) and 2-O-sulfo-alpha-L-idopyranuronic acid (PDB ID: IDS). The receptor maps were generated with grid step of 0.50. The dockings for non-covalent binding were performed with a thoroughness level of 3, with the generation of three initial ligand conformations for each simulation.

The heparin-bound *NA* CP models were rendered in electrostatic surface representation by ICM (Abagyan *et al,* 1994), where the potential scale was set to 5.0 along with the assignment of simple charges, for the purpose of viewing the electrostatics of the protein sites occupied by the negatively charged heparin.

Results:

1. Sequence alignments and pattern detection:

1.1 N-terminal signal for protein localization:

The *NA*-CP alignment showed that the CP7 lacked the N-terminal signal pre-sequence present in the other proteases (**Figure1**). The PRATT derived motif unique to CP1-6's pre-sequence was M-x(4,5)-L. CP7's N-terminal however had a lysosomal targeting pattern [DE]x{3}L[LI], according to LocSigDB. The specific residues in the pre-sequences are summarized in **Table1**, along with the lysosome targeting peptide in CP7.

1.2 Cleavage of the N-terminal signal peptide:

SignalP derived cleavage sites predicted the length and the peptide sequence for the signals contained in the pre-sequences of CP1-6. **Figure 1** shows the positioning of the cleavage sites where the signal peptide would be cleaved off to release the proenzymes or zymogens. The lengths of these signal peptides across the *NA* CP1-6 were approximately the same and spanned the entire length of the signal pre-sequences. CP7 did not show any such cleavage site at equivalent position to the other proteases, confirming its lack of such signal.

1.3 Localization guided by the signal peptide:

The consensus from the localization prediction methods deemed CP1-6 to be secretory proteins, with the aforementioned presence of pre-sequences, which signal for the proteases' extracellular localization. CP7 was predicted to be non-secretory with its lysosome targeting signal peptide (**Table 1**).

1.4 Hemoglobinase motif for hemoglobin-degradation functionality:

The hemoglobinase motif Y-[WY]-[IL]-[IV]-x-N-S-W-x-[DEGNQST]-[DGQ]-W-G-E-x(1,2)-G-x-[FI]-[NR]-[FILM]-x(2)-[DG]-x-[DGNS] was derived from the hemoglobin degrading cysteine proteases of the following blood feeders viz: *NA* (CP3), *A. Caninum*, *S.mansoni*, *S.japonicum*, *H. Contortus*, *O. ostertagi* and *A. Suum*. The pattern detected here is longer than the previously reported Y-W-[IL]-[IV]–A-N-SW-X–X–D-W-G-E motif by Baig *et al*, (Baig *et al*, 2002). See comparison in **Figure 1**. The training data set for deriving this new motif additionally included falcipain2 and falcipain3 of *P.falciparum* and CP1 of *P.westermani* and the derived motif was absent in the cysteine proteases of the non-blood feeders viz: *C. elegans* and *L. major*. The observations from this study are similar to the previous study in the context of the presence of the motif in blood feeding proteases and its absence in the non-blood feeding proteases. The derived motif when searched in the *NA*-CPs (excluding CP3 of the training dataset) was detected only in CP2.

1.5 Kallikrein-like activity of cleaving kininogen for implicated vasodilation and anticoagulation :

NA CP1 and CP6 were observed to have an aspartic acid and a glutamic acid respectively at the S2 subsite of their enzymatic pocket similar to human cathepsin B's and cruzain's (both of which cleave HMWK) critical Glu, **(Figure 2),** which forms an important specificity determinant for substrates having Arg or Phe (Gillmor *et al*, 1997). *NA* CP1 and CP6, due to their similarity in enzymatic-pocket molecular feature with cruzain and human cathepsin B, are likely to bind human HMWK as their substrate in the region Leu^{373}-Gly-Met-Ile-Ser-Leu-Met-Lys-Arg-Pro-Pro-Gly-Phe-Ser-Pro-Phe-Arg-Ser-Ser-Arg-Ile^{393} of the kininogen; the same region that cruzain cleaves between Met^{379}-Lys^{380} and Arg^{389}-Ser^{390} to generate the vasodilator Lys - bradykinin (Del Nery *et al*, 1997). With the catalytic cysteine of the cysteine protease enzymes located between S1' and S1 subsites, where the peptide bond hydrolysis occurs, the cleavage sites of HMWK indicate that Arg^{381} and Phe^{388} of the kininogen are likely to get accommodated at the S2 subsite of cruzain. The side chain of the Glu207 at its S2 site could form salt-bridge interaction with the charged Arg^{381} of the substrate. Next, when S2 subsite is occupied by the hydrophobic Phe^{388} of the kininogen, the aforementioned critical glutamate of the enzyme could turn away to face the solvent. Therefore, the proteolytic action on HMWK by *NA*-CP1 and CP6 - which possess acidic

Asp246 and Glu245 near their S2 pockets - could possibly release vasoactive peptides.

The enzymatic pocket of cathepsin B, cruzain, *NA* CP1 and CP6 shared an alanine at their S2 subsite (**Figure 2**). Cruzain's Ala137 and human cathepsin B's Ala172 at their S2 subsites, located near their critical Glu residue, corresponded with Ala175 of CP1 and Ala174 of CP6. The distances between the C-alpha atoms of this conserved Ala and the critical acidic Asp/Glu residues of the relevant enzymes are listed in **Supplementary Table 1.** Such measurements were made in order to get an estimate of the distance of the critical Asp/Glu from the active site cleft, given the lack of substrate/ligand in the structures (except cruzain). The Ala and Asp/Glu around the S2 site of *NA* CP1 and CP6 were spaced further away compared to the distance between the relevant Ala and Glu in cruzain and cathepsin B. However, in the dynamic *NA* CP proteins, the side chains of the charged acidic residues could probably move closer (more so Glu245 of CP6, because of its longer side chain) to make the salt-bridge interaction with charged Arg^{381} of HMWK substrate.

2. Heparin binding for transportation of indigenous worm glycosaminoglycans, and molecular recognition of host proteins:

2.1 Heparin-binding motif for post-translational glycosylation and transportation of worm GAGs to extracellular matrix:

Asn glycosylation 'NGTD' motifs were detected in *NA* CP1, CP3, CP4, CP4b and CP7, as per prosite entry PS00001. The glycosylation site region in the alignment (**Figure1**) showed CP2 to retain at least the Asn residue for N-glycosylation in its 'NGVK' stretch.

2.2 Heparin binding domain for molecular recognition of host proteins:

CP2 and CP5 sequences were predicted by ProDom to have fibronectin domain type III (entry: PDC9H7K4), which is known to harbor N-linked and O-linked glycosylation sites (Pankov *et al*, 2002) and has been shown to bind heparin (Takahashi *et al*, 2007). The N-glycosylation sites mentioned earlier were within the region spanning this predicted fibronectin III domain in the *NA* CPs. The region did not have any cysteine residues in any of the proteases for disulfide bond formation, which is again a feature of type III repeat (Pankov *et al*, 2002). The domain region was 74 amino acids in length. CP3 having the longest sequence had 79 residues. CP5 showed the RGD motif bordering the predicted fibronectin region (See alignment in **Figure 1**), which is another signature for type III_{10} repeat (Pankov *et al*, 2002). Structurally, however, the domain did not exactly resemble the beta sandwich fold typical of fibronectin III. With CP5 as reference, the fibronectin domain-like region in the rest of the secretory CPs showed sequence identity within 62.5% to 68.75% and the sequence similarity ranged from 56.45% to 71.43%. (**Supplementary Table 2**). The high sequence

identity and similarity implied that all the CPs would be predisposed to heparin binding, and so the entire stretch of the predicted fibronectin domain was considered for docking heparin.

2.3. Heparin binding docking simulation for molecular recognition of host proteins:

Heparin on being docked at the putative *heparin-binding* fibronectin III – like domain of the cathepsin B-like CPs showed the best scored conformation to bind surface loops (**Figure 3**), away from the enzymatic cleft (**Figure 4**), at a site similar to what has been reported in an earlier docking/MD simulation study on human cathepsin B - heparin interaction (*Costa et al, 2010*). The crucial interactions made there by the human cathepsin B's mentioned basic residues Lys157 and Arg234 for binding the negatively charged heparin, mapped close to this study's heparin-interacting Lys and Arg residues in most of the *NA*-CPs (derived from the alignment of human cathepsin B with *NA* CPs: **Figure 2**). This is the closest comparison (**Figure 1**), which could be drawn, with no structures of cathepsin B – heparin complex available in PDB. The binding site residues within 4Å of the ligand are underscored in the sequence alignment. Barring CP6, which showed the least sequence similarity with the reference sequence CP5 in the fibronectin-like domain region (**Supplementary Table 2**), heparin occupied similar sites in the CP homology models **(Figure 3)**. The positioning of heparin was slightly shifted in CP6, albeit away from the active site (**Figure 4**). The overall electrostatics (**Figure 4**) at the heparin-bound sites of the CPs varied from predominantly neutral to positive (except CP2, CP4 and CP4b), depending on the site-residues and long-range electrostatic effects from residues beyond. The negative sulfate groups in the glycosaminoglycan molecule mostly interacted with basic/neutral residues at the binding site. Some of the heparin-contacts of the CPs showed the patterns: BBX, XBBX, BXB, and BXXBB, which were entirely or partially in conformity with previously reported heparin-binding motifs (Forster *et al,* 2006; Proudfoot *et al,* 2001*;* Mann *et al,* 1994*;* Fromm *et al,* 1997), where B is a basic residue and X is any amino acid. The highly negative binding-site electrostatics of CP4 and CP4b contributed to unfavorable docking scores for these proteases. The rest of the CPs showed scores for good binding of heparin. **Table 2** summarizes the scores, contact residues, sequence patterns, and H-bonding interactions of the highest scored conformations of heparin.

Discussion:

NA infection and survival in human hosts requires a repertoire of proteolytic enzymes in the ES products of the worm. The parasite's physiology involves cysteine proteases for digestive purposes (only CP3 has been characterized as such), and for evading potentially damaging host hemostatic events. This study is an attempt to decode the individual cysteine proteases of *NA* in terms of their functionality.

1. Sequence alignments and pattern detection:

1.1 N-terminal signal for protein localization:

CP1-6 proteins' hydrophobic N-terminal pre-sequence are presumably signal peptides for extracellular localization. N-terminal signals are extremely degenerate across various proteins. The conserved hydrophobic M-x(4,5)-L sequence pattern in the *NA*-CPs therefore forms part of the unique signal peptide for these proteins.

1.2 Cleavage of the N-terminal signal peptide:

CP7, which lacked the N-terminal pre-sequence, was not predicted to have any signal peptide cleavage site. The observation implies and re-emphasizes that this protease gets synthesized without the signal for extracellular localization, unlike the other CPs.

1.3 Localization guided by signal peptide:

CP1-6 are secretory as per the localization predictions and could be available for important host-pathogen interactions. CP3 amongst these has been implicated to be present in the gut of an adult worm at the host-pathogen interface (Ranjit *et al*, 2008), and has been shown to be involved in the hemoglobin degradation pathway (Ranjit *et al* 2009). Therefore, CP1, CP2, CP4, CP4b, CP5 and CP6 could also be expected to perform similar or other supportive functions as CP3, pertaining to blood feeding. Whereas, CP7 that lacks the active site cysteine residue is predicted to reside in lysosome for unknown purposes.

.

1.4 Hemoglobinase motif for hemoglobin-degradation functionality:

The cysteine proteases from the *NA* ES products of the parasite had been demonstrated to cleave hemoglobin (Brown *et al*, 1995). However, only CP3 in the worm has been characterized to have such role. The molecular features pertaining to such function in *NA* CPs has not been researched. Hemoglobin degrading activity in Cathepsin B like proteases from blood feeding helminths were attributed to Y-W-[IL]-[IV]–A-N-SW-X–X–D-W-G-E sequence motif by Baig *et al*. In this study, a specific sequence pattern generic to hemoglobin degrading enzymes was sought, without emphasis on a particular family of cysteine proteases. This study therefore included hemoglobin-cleaving non-cathepsin B enzymes from *P.falciparum* and *P.Westermani*, which were not taken into account by Baig *et al*. Despite adopting a different methodology (mentioned in materials and methods) from the previous study for deriving the motif, a pattern unique to the blood degrading enzymes emerged. The derived hemoglobinase motif Y-[WY]-[IL]-[IV]-x-N-S-W-x-[DEGNQST]-[DGQ]-W-G-E-x(1,2)-G-x-[FI]-[NR]-[FILM]-x(2)-[DG]-x-[DGNS] which is being reported here is a longer pattern, and working with such motif is advantageous in terms of avoiding false positives. The derived motif could be investigated in different enzymes across blood feeding

pathogens. CP2 was the only protein in the repertoire of the *NA* cysteine proteases to have the motif, other than the already established globinase CP3. This observation is suggestive of CP2's involvement in hemoglobin degradation, along with CP3.

1.5 Kallikrein-like activity of cleaving kininogen for implicated vasodilation and anticoagulation:

Human plasma kallikrein (pKal) is a component of the anticoagulation pathway that cleaves HMWK - a protein involved in blood coagulation and fibrinolysis - to produce a potent vasodilator bradykinin (Maurer *et al*, 2011) that stimulates prostacyclin (PGI_2) release from endothelial cells. PGI_2 in turn is an inhibitor of platelet activation and degranulation pathway (Maurer *et al*, 2011). Emulation of plasma kallikrein activity has been reported for cruzain (Del Nery *et al*, 1997; Lima *et al,* 2002) – a cysteine protease from *T.Cruzi* that traverses blood capillary vessels. Such function has been suggested also for SmSP1 - serine protease of blood fluke *S.mansoni* (Mebius *et al*, 2013), as kallikrein-like activity had been reported for a serine protease from *S.mansoni* (Carvalho *et al*, 1998, Da'dara *et al*, 2011). The blood stream navigating physiology of the mentioned pathogens explains their need for cleaving HMWK to generate bradykinin for vasodilation and blood clot prevention.

The cysteine protease cruzain, although being cathepsin L-like, shares similar substrate specificity with cathepsin B. The S2 pocket of the enzyme is the major specificity determinant, which has a glutamic acid residue at its bottom, same as what is found in cathepsin B's S2 subsite. This particular feature is responsible for similar substrate bindings of cruzain and cathepsin B, (Gillmor *et al*, 1997) which includes HMWK - the cleavage of which by these proteases generates vasoactive bradykinin (Del Nery *et al*, 1997, Barros *et al*, 2004), which dilates blood vessels and eventually thwarts the blood coagulation cascade. Cathepsin B has been implicated to impair coagulation and fibrinolysis under pathological conditions, (Herszenyi *et al*, 1997) Cathepsin B-like extracellular *NA* CPs are therefore likely to exhibit HMWK-cleaving kininogenase activity. Especially CP1 and CP6 of the *NA* hookworm could possibly generate bradykinin by the dint of their critically positioned Asp246 and Glu245. The hookworm might be exploiting this component of the anticoagulation pathway (amongst others) to evade host blood clots for easing migration and/or facilitating feeding.

2. Heparin binding for transportation of indigenous worm glycosaminoglycans, and molecular recognition of host proteins:

2.1 Heparin-binding motif for post-translational glycosylation and transportation of worm GAGs to extracellular matrix:

The prediction of N-glycosylation motifs within the predicted fibronectin type III domain region of the *NA* CPs hint that these proteases can have GAG molecules

covalently attached to these sites as post-translational modifications. Such modifications could include heparin, which are reported to occupy N-glycosylation sites, and even sites without the Asn attachment point (Fenaille *et al*, 2007). N-glycosylation sites harboring *NA,* CP1, CP2, CP3, CP4 and CP4b, could be transporting the hookworm's indigenous heparin (if present) to the host-pathogen interface. CP5 and CP6, which did not show any N-motif in the region, could also be glycosylated with heparin. Heparin when released at the interface, by deglycosylation of the secreted *NA* CPs, could possibly assist in evading blood coagulation (by serving as co-factors to antithrombin - an inhibitor of thrombin and other coagulation factors in the blood plasma), as has been implicated for indigenous worm heparin/heparan sulfate found on the tegumental surface of blood fluke *S.mansoni* (Mebius *et al*, 2013).

2.2 Heparin binding domain prediction for molecular recognition of host proteins:

Fibronectin can non-covalently bind a number of biologically important ligands that includes glycosaminoglycan such as heparin or heparan sulfate, which are present on the surfaces of extracellular proteoglycans (Pankov *et al*, 2002). The molecule's type II domain forms the primary binding site for heparin. However, fibronectin fragments with type III_{7-10} and type III_{10-12} repeats have been shown to bind heparin (Takahashi *et al*, 2007). In the cathepsin B-like *NA*-CPs, the predicted fibronectin type III-like domain overlapped with the β sheet and loop dominated R domains; structurally resembling the R domains of human cathepsin B and cruzain (**Figure 3**), where heparin has been implicated to bind (Costa *et al*, 2010, Lima *et al,* 2002). The negatively charged heparin occupied neutral-positive electrostatic patches in most of the different *NA* CPs (**Figure 4**) at about the same location as the low-energy heparin-docked region of human cathepsin B (Costa *et al*, 2010). Similar results were obtained in this study despite adopting a different methodology of pattern-based prediction method to specify the region for non-covalently docking heparin, as compared to the Costa *et al* study, where they relied on protein patches having positive electrostatic potential. Heparin bound to the *NA*-CPs, away from the active site, making the K and R contacts, as made by human cathepsin B's K157 and R234 (**Figure 1**).

2.3 Heparin binding docking simulation for molecular recognition of host proteins:

Heparin-like glycosaminoglycan present in animal plasma membrane and ECM, form components of the host-pathogen interface (Bartlett *et al,* 2010). The GAGs serve as recognition factors for molecular interactions, controlling activities like cell adhesion and parasitic infection (Bartlett *et al,* 2010*;* Lima *et al,* 2002*;* Judice *et al,* 2013*).* These negatively charged molecules are covalently linked to syndecans and glypicans proteoglycans of host cells, and help to recognize and bind parasitic cysteine proteases (Bartlett *et al,* 2010*)* such as cruzain (Lima *et al,* 2002*;* Judice *et al,* 2013*).* The simultaneous binding of the proteases with their protein substrate and the latter's co-valently linked GAG results in the formation

of ternary complexes (Lima *et al,* 2002*;* Judice *et al,* 2013)*,* which facilitates the enzymatic action of the proteases. Further implication of heparin binding to cathepsin B-like *NA* CP1, CP2, CP3, CP5 and CP6 (with good docking scores - **Table 2**) in terms of assisted functionality is discussed next in the light of heparin-binding to human cathepsin B.

2.3.1 Prevention of alkaline pH-induced inactivation:

Human Cathepsin B bound by heparin/heparin-sulfate does not undergo alkaline pH induced loss of catalytic activity (Almeida *et al* 2001*,* Costa *et al,* 2010). Inactivation of cathepsin B under alkaline conditions occurs due to disruption of electrostatic interactions like the thiolate-imidazolium ion pair of the active site, and breakage of crucial salt bridge interactions. Heparin binding prevents the loss of such interactions and helps stabilize the protease's structure at alkaline pH by maintaining the helical content of the active enzyme (Almeida *et al* 2001). This allosteric mechanism mediated by heparin's binding away from the active site has been corroborated by computational studies (Costa *et al,* 2010*). NA*-CPs being cathepsin B-like are likely to be affected by heparin binding in a similar way. The molecule docked on to the *NA*-CPs, at a region similar to what has been reported by Costa *et al* in their cathepsin B–heparin dockings (**Figure 1**). This is suggestive of possible allosteric control that could be exercised by heparin over the enzymatic action of the *NA*-CPs to make them functional even at alkaline pH. Such controls could be crucial for the survival of hookworm, which encounters different pH conditions within the host, and the cysteine proteases of which have varying degrees of overall surface electrostatics (**Figure 4**).

2.4 Implications for future drug design with heparin analogs:

NA-CPs' suggestive role in hemoglobin degradation and anticoagulation, aided by heparin-like molecules as mechanistically described before, probably requires abolishment of GAG binding to the *NA*-CPs for preventing the proteases' pathogenic activities. Blocking the putative GAG binding site on the parasite's CPs, (see residues in **Table 2**) could help thwart hookworm infection. Soluble heparin/heparin sulfate usage as competitive inhibitors have the chances of causing adverse physiological effects due to their anticoagulant/immunogenic properties (Bartlett *et al,* 2010). However, synthetic GAG mimetics with limited biological activity (minimum active structure) or polysaccharides (Vann *et al*, 1981*;* Copeland *et al,* 2008) targeted at GAG binding site, could be used for inhibiting hookworm infection, in conjunction with active-site inhibitors.

This study was initiated to explore the role of *NA* CPs in human hookworm pathogenesis*,* as the proteases remain uncharacterized despite their demonstrated pathogenic potential (Brown *et al*, 1995). Bioinformatics based analyses of the *NA*-CPs is indicative of the presence of CP1, CP2, CP3, CP4, CP4b, CP5 and CP6 in the ES content of the hookworm for host-pathogen

interactions. The hemoglobinase motif derived here is harbored by CP2 and CP3, which suggests hemoglobinase activity for these two proteins, of which only CP3 has been confirmed to degrade globin (Ranjit *et al*, 2009). Possible HMWK – cleaving activity by *NA* CP1 and CP6 hint towards the proteases' role in evading the host hemostatic system for preventing blood clots, in order to facilitate feeding and survival. The heparin docking results from this study is speculative of *NA* CPs' heparin-assisted HMWK-cleaving activity. Such function, already verified for the proteases from other parasites like *T.Cruzi* (Del Nery *et al*, 1997; Lima *et al,* 2002) and *S.mansoni* (Carvalho *et al*, 1998) that navigate blood capillaries, leads to the hypothesis of similar survival strategy adoption by larvae-stage *NA* for migrating through blood capillaries. The predicted extracellular localizations of CP1-CP6 and their predicted pathogenic roles render these CPs to be possibly multi-targeted for heparin-analog binding, which could inhibit hookworm infection. This study attempts to provide molecular level information based on computational predictions, decoding previously unreported possible functions of the individual *NA* cysteine proteases, which could be subjected to further experimental validation.

Acknowledgement:
The authors acknowledge useful discussions with Prof. Conor Caffrey, UCSD. Previous work on hookworm by UCSD visiting student Kevin Weidmer is also acknowledged.

References:

Abagyan, R.A., Totrov, M.M., Kuznetsov, D.A., 1994. ICM: A New Method For Protein Modeling and Design: Applications To Docking and Structure Prediction From The Distorted Native Conformation J. Comp. Chem. 15, 488-506.

Almeida, P.C., Nantes, I.L., Chagas, J.R., Rizzi C.C.A., Faljoni-Alario A, Carmona E., Julianoi, L, Nader, H.B., Tersariol, I. L. S., 2001. Cathepsin B activity regulation heparin-like glycosaminoglycans protect human cathepsin B from alkaline pH-induced inactivation. Vol. 276, No. 2, Issue of January 12: 944–951.

Baig, S., Damian R.T., Peterson, D. S., 2002. A novel cathepsin B active site motif is shared by helminth bloodfeeders. Experimental Parasitology 101: 83–89.

Bagot, C.N., Arya, R., 2008. Virchow and his triad: A question of attribution. Br J Haematol 143: 180–190.

Bannai, H., Tamada, Y., Maruyama, O., Nakai, K., and Miyano, S., 2002. Extensive feature detection of N-terminal protein sorting signals, Bioinformatics, 18(2) 298-305.

Barros, N. M.T., Tersariol, I. L., Oliva, M. L., Araujo, M. S., Sampaio, C. A., Juliano, L. & Motta, G. 2004. High molecular weight kininogen as substrate for cathepsin B. Biol. Chem. 385, 551–555

Bartlett, A. H., Park, P. W., 2010. Proteoglycans in host–pathogen interactions: molecular mechanisms and therapeutic implications. Expert reviews in molecular medicine: 1-25

Beck, W.S., 1985. Hematology. Cambridge: The MIT Press. 496

Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Sayers, E.W., 2013. GenBank. Nucleic Acids Res. Jan; 41 (Database issue): D36-42.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., 2000. The Protein Data Bank Nucleic Acids Research, 28: 235-242.

Blobel, G., 1980. Intracellular protein topogenesis. PNAS. Mar; 77(3): 1496–1500

Bodén, M., Hawkins, J., 2005 Prediction of subcellular localization using sequence-biased recurrent networks. Bioinformatics. 21(10): 2279-2286.

Brooker, S., Bethony, J., and Hotez P.J., 2004. Human Hookworm Infection in the 21st Century. Adv Parasitol; 58: 197–288.

Brooker S.J., Murray C.J. L., 2014. A systematic analysis of global anemia burden from 1990 to 2010. Blood 123:615-624

Brown, A., Burleigh J.M., Billett, E.E., and Pritchard D.I., 1995. An initial characterization of the proteolytic enzymes secreted by the adult stage of the human hookworm *Necator americanus* Parasitology. 110: 555-563

Brown A, Girod, N., Billett, E.E., Pritchard, D.I., 1999. Necator Americanus (human hookworm) aspartyl proteinases and digestion of skin macromolecules during skin penetration. Am. J. Trop. Med. Hyg 60(5): 840–847

Bru, C., Courcelle, E., Carrère, S., Beausse, Y., Dalmar S., Kahn, D., 2005. The ProDom database of protein domain families: more emphasis on 3D. Nucleic Acids Res. 33: D212-D215

Bungiro, R., Cappello, M., 2011. Twenty-first century progress toward the global control of human hookworm infection. Curr Infect Dis Rep; 13:210-7.

Carvalho, W.S, Lopes, C.T, Juliano, L., Coelho, P.M., Cunha-Melo, J.R, Beraldo W.T., Pesquero, J.L., 1998. Purification and partial characterization of kininogenase activity from Schistosoma mansoni adult worms. Parasitology 117: 311–319

Copeland, R., Balasubramaniam, A., Tiwari,V., Zhang, F., Bridges, A., Linhardt R.J., Shukla, D., Liu, J., 2008. Using a 3-O-sulfated heparin octasaccharide to inhibit the entry of herpes simplex virus type 1. Biochemistry 47:5774-5783.

Costa, M. G.S., Batista, P.R., Shida, C.S., Robert, C.H., Bisch, P.M., Pascutti, P.G., 2010. How does heparin prevent the pH inactivation of cathepsin B? Allosteric mechanism elucidated by docking and molecular dynamics. BMC Genomics 11(Suppl 5): S5

Cox, G. N., Pratt, D., Hageman, R., Goisvenue, R. J., 1990. Molecular cloning and primary sequence of a cysteine protease expressed by Haemonchus contortus adult worms. Molecular and Biochemical Parasitology 41: 25-34

Da'dara, A., Skelly, P.J., 2011. Manipulation of vascular function by blood flukes? Blood Rev 25: 175–179.

Del Nery, E., Juliano, M.A., Lima, A.P., Scharfstein, J., Juliano, L., 1997. Kininogenase Activity by the Major Cysteinyl Proteinase (Cruzipain) from Trypanosoma cruzi. J Biol Chem Vol. 272(41): 25713–25718

Diemert, D.J., Bethony, J.M., Hotez, P.J, 2008. Hookworm vaccines. Clin Infect Dis 46:282-288.

Emanuelsson, O., Brunak, S., Heijne, G.V., Nielsen, H., 2007. Locating proteins in the cell using TargetP, SignalP, and related tools Nature Protocols 2, 953-971.

Fenaille, F., Mignon, M.L., Groseil, C., Ramon, C., Riand´e, S., Siret, L., Bihoreau, N., 2007. Glycobiology vol. 17 (9): 932–944

Forster, M., Mulloy B. 2006. Computational approaches to the identification of heparin-binding sites on the surface of proteins. Biochem Soc Trans; 34:431–434

Fromm, J.R., Hileman, R.E., Caldwell, E.E.O., Weiler, J.M., Linhardt, R.J., 1997. Pattern and spacing of basic amino acids in heparin binding sites. Arch Biochem Biophys; 343:92–100.

Furmidge, B. A., Horn, L. A., Pritchard, D. I., 1995. The anti-haemostatic strategies of the human hookworm *Necator americanus.* Parasitology, 112: 81-87

Gillmor, S., Craik, C.S., Fletterick, R.J. 1997. Structural determinants of specificity in the cysteine protease cruzain. Protein Science 6:1603-1611

Harrison L. M., Nerlinger, A., Bungiro, R. D., Cordova J.L., Kuzmic, P., Cappello, M., 2002. Molecular Characterization of Ancylostoma Inhibitors of Coagulation Factor Xa; Journal of biological chemistry. Vol. 277(8), Issue of February 22: 6223–6229.

Heijne, G.V., 1985. Signal sequences. The limits of variation. Journal of molecular biology 184: 99-105

Herszenyi, L., Plebani., M., Carraro P., De Paoli, M., Cardin, R., Di Mario F., Kusstatscher S., Naccarato R., Farinati, F. 1997. Impaired fibrinolysis and increased protease levels in gastric and duodenal mucosa of pateints with active duodenal ulcer. Am j. gastroenterology. 92: 843-847

Hiller, K., Grote A., Scheer M., Mu¨nch, R., Jahn D., 2004. Prediction of signal peptides and their cleavage positions. Nucleic Acids Research. 32: W375–W379.

Hotez, P.J., Brooker, S., Bethony, J.M., Bottazzi, M.E., Loukas A., Xiao, S.H, 2004. Hookworm infection. N Engl J Med; 351:799–807.

Hotez P., 2008. Hookworm and poverty Ann NY Acad Sci;1136:38–44

Hotez, P.J., Bethony, J.M., Diemert, D.J., Pearson, M. & Loukas, A, 2010 Developing vaccines to combat hookworm infection and intestinal schistosomiasis. Nat. Rev. Microbiol. 8, 814–826

Illy, C., Quraishi, O., Wang, J., Purisima, E., Vernet, T., & Mort, J.S., 1997. Role of the Occluding Loop in Cathepsin B Activity J. Biol. Chem. 272, 1197-1202

Jasmer, D.P., Roth, J., Myler, P.J., 2001. Cathepsin B-like cysteine proteases and Caenorhabditis elegans homologues dominate gene products expressed in adult Haemonchus contortus intestine. Mol. Biochem. Parasitol. 116:159–169

Jonassen I., Collins, J.F., Higgins, D., 1995. Finding flexible patterns in unaligned protein sequences Protein Science,4(8):1587-1595

Judice, W. A. S., Manfredi, M.A., Souza, G. P., Sansevero, T.M., Almeida, P. C., Shida, C.S., Gesteira, T.F., Juliano, L., Westrop, G.D., Sanderson, S.J., Coombs, G.H., Tersariol, I. L. S., 2013. Heparin Modulates the Endopeptidase Activity of Leishmania mexicana Cysteine Protease Cathepsin L-Like rCPB2.8 PLoS ONE 8(11): e80153. doi: 10.1371/journal.pone.0080153.

Kassebaum N.J., Jasrasaria R., Naghavi M., Wulf S. K., Johns N., Lozano R, Regan M., Weatherall D, Chou D.P., Eisele T.P., Flaxman S.R., Pullan R.L., Brooker, S.J., Murray CJ., 2014. A systematic analysis of global anemia burden from 1990 to 2010. Blood. 123:615-624

King, B.R., Guda, C., 2007. ngLOC: an n-gram-based Bayesian method for estimating the subcellular proteomes of eukaryotes. Genome biology.;8(5):R68.

Krogh, A., Larsson, B., Heijne, G.V., Sonnhammer, E.L., 2001 Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol. Jan 19; 305(3): 567-80

Laskowski, R.A., MacArthur, M.W., Moss, D.S., 1993. PROCHECK: a program to check stereochemical quality of protein structures. J. Appl. Cryst, 26: 283-291

Lima, A.P. C. A., Almeida, P.C., Tersariol, I.L. S., Schmitz, V., Schmaier, A.H., Juliano L., Hirata, I.Y., Muller-Esterl, W., Chagas, J.R., Scharfstein, J., 2002. Heparan sulfate modulates kinin release by T.Cruzi through the activity of cruzipain. JBC Vol. 277, No. 8, Issue of February 22: 5875–5881.

Loukas, A. Gaze S., Mulvenna J. P., Gasser R.B., Brindley P.J., Doolan D.L., Bethony J.M., Jones M.K., Gobert G.N., Driguez P, McManus D.P., and Hotez P.J, 2011. Vaccinomics for the major blood feeding helminthes of humans. OMICS 15(9): 567-577

Lowe, G.D., 2003. Virchow's triad revisited: Abnormal flow. Pathophysiol

Haemost Thromb 33: 455–457.

Mann, D.M., Romm, E., Migliorini, M., 1994. Delineation of the glycosaminoglycan binding site in human inflammatory response protein Lactoferrin. J Biol Chem; 269:23661–23667.

Mason S., DiLeo M., Abendroth J., Kuzmic, Petr., Ladner, R. C., Edwards, T.E., TenHoor, C., Adelman, B.A., Nixon, A.E., Sexton, D.J., 2014. Inhibition of Plasma Kallikrein by a Highly Specific Active Site Blocking Antibody. Journal of Biological Chemistry Vol. 289, No. 34: 23596–23608.

Maurer, M., Bader, M., Bas, M., Bossi, F., Cicardi, M., Cugno, M., Howarth, P., Kaplan, A., Kojda, G., Leeb-Lundberg, F., Lötvall, J., Magerl, M., 2011. New topics in bradykinin research. Allergy 66: 1397–1406.

Mebius, M.M., van Genderen, P.J.J., Urbanus, R.T., Tielens, A.G.M., de Groot, P.G., 2013. Interference with the Host Haemostatic System by Schistosomes. PLoS Pathog 9(12): e1003781.

Pankov, R., Yamada K. M., 2002. Fibronectin at a glance. Journal of Cell Science 115: 3861-3863

Pearson M. S., Tribolet L., Cantacessi C., Periago M.V., Valerio M. A., Jariwala A.R., Hotez P.J, Diemert D, Loukas A, Bethony J., 2012. Molecular mechanisms of hookworm disease: Stealth, virulence, and vaccines. J allergy clin immunol volume 130, Number 1

Petersen, T. N., Brunak, S., Heijne, G.V., Nielsen, H., 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions Nature Methods, 8:785-786.

Pettersen, E.F, Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., Ferrin, T.E., 2004 UCSF Chimera - A visualization system for exploratory research and analysis. J Comput Chem (13): 1605-12

Pierleoni, A., Martelli, P.L., Fariselli, P., Casadio, R., 2006. BaCelLo: a balanced subcellular localization predictor. Bioinformatics 22 (14): e408-e416.

Proudfoot, A.E.I., Fritchley, S., Borlat, F., Shaw, J.P., Vilbois, F., Zwahlen, C., Trkola, A., Marchant, D., Clapham, P.R., Wells, T.N.C., 2001. The BBXB motif of RANTES is the principal site for heparin binding and controls receptor selectivity. J Biol Chem;276: 10620–10626.

Ranjit, N., Jones, M.K., Stenzel, D.J., Gasser, R.B., Loukas, A., 2006. A survey of the intestinal transcriptomes of the hookworms, Necator americanus and Ancylostoma caninum, using tissues isolated by laser microdissection microscopy. International Journal for Parasitology 36: 701–710.

Ranjit, N., Zhan, B., Stenzel, D.J., Mulvenna, J., Fujiwara, R., Hotez, P.J., Loukas, A., 2008. A family of cathepsin B cysteine proteases expressed in the gut of the human hookworm, Necator Americanus. Molecular & Biochemical Parasitology 160:90–99

Ranjit, N., Zhan, B., Hamilton B., Stenzel D., Lowther, J., Pearson, M., Gorman, J., Hotez, P., Loukas, A., 2009. Proteolytic Degradation of Hemoglobin in the intestine of the Human Hookworm Necator americanus The Journal of Infectious Diseases; 199:904 –12

Sajid, M., McKerrow, J.H., 2002. Cysteine proteases of parasitic organisms. Molecular & Biochemical Parasitology 120: 1–21

Sakti H., Nokes C., Hertanto W.S., Hendratno S., Hall A., Bundy D.A., Satoto, 1999. Evidence for an association between hookworm infection and cognitive function in Indonesian school children. Trop Med Int Health; 4(5): 322–34.

Sigrist, C.J.A., Cerutti, L., Hulo, N., Gattiker, A., Falquet, L., Pagni, M., Bairoch, A., Bucher, P., 2002. PROSITE: a documented database using patterns and profiles as motif descriptors. Brief Bioinform. 3:265-274.

Stanssens, P., Bergumt, P.W., Gansemanst, Y., Jesperst L., Larochet, Y., Huangt, S., Makit S., Messenst, J., Lauwereyst, M., Cappello, M., Hotez, P.J., Lasterst, I., Vlasuk, G.P., 1996. Anticoagulant repertoire of the hookworm Ancylostoma caninum. Proc. Natl. Acad. Sci. USA Vol. 93, pp. 2149-2154.

Takahashi, S., Leiss, M., Moser, M., Ohashi, M., Kitao, T., Heckmann, D., Pfeifer, A., Kessler, H., Takagi, J., Erickson, H.P., Fässler, R. 2007. The RGD motif in fibronectin is essential for development but dispensable for fibril assembly. The Journal of Cell Biology, Vol 178(1): 167–178

The UniProt Consortium. 2015. UniProt: a hub for protein information. Nucleic Acids Research, Vol. 43, Database issue.

Turk, V., Stoka, V., Vasiljeva, O., Renko, M., Sun, T., Turk, B., Turk, D. 2012. Cysteine cathepsins: From structure, function and regulation to new frontiers. Biochimica et Biophysica Acta: 68-88

Yu, C.S., Chen, Y.C., Lu, C.H., Hwang, J.K., 2006: Prediction of protein subcellular localization. Proteins: Structure, Function and Bioinformatics, 64:643-651.

Vann, W.F., Schmidt, M.A., Jann, B., Jann, K., 1981. The structure of the capsular polysaccharide (K5 antigen) of urinary-tractinfective Escherichia coli 010:K5:H4. A polymer similar to desulfo-heparin. European Journal of Biochemistry 116: 359-364

Vetter, J.C., van der Linden ME. 1977.Skin penetration of infective hookworm larvae. I. The path of migration of infective larvae of Ancylostoma braziliense in canine skin. Z Parasitenkd; 53:255-62.

Vetter, J.C, van der Linden ME. 1977.Skin penetration of infective hookworm larvae. II. The path of migration of infective larvae of Ancylostoma braziliense in the metacarpal footpads of dogs. Z Parasitenkd; 53:263-6.

NA Cysteine Protease	Localization	Signal motif	Motif specific to protea
CP1	Secretory	M-x(4,5)-L	MLLFLTL
CP2	Secretory	M-x(4,5)-L	MLTLAAL
CP3	Secretory	M-x(4,5)-L	LILIAL
CP4	Secretory	M-x(4,5)-L	MKANFAL
CP4b	Secretory	M-x(4,5)-L	MKANIAL
CP5	Secretory	M-x(4,5)-L	MITIITL
CP6	Secretory	M-x(4,5)-L	MLITLAL
CP7	Non-secretory (Lysosomal)	[DE]x{3}L[LI]	DDKDLL

Table1: Localization of the *NA* cysteine proteases, and the corresponding motifs pertaining to the localization signals.

NA-CP	Score	Contact Residues	Motif pattern	H-bonding residues
CP1	-21.75	**K186, R187, G188**, I189, Y190, K191, K193, D211, N212, T214, Y216, E229, R234	BBX	R187, K191, N212, R234
CP2	**-**24.22	Y182, Y183, K184, N185, G186, I187, Y188, M189, E209, N210, V212, R232.		Y182, K184, N185, M189, E209, N210, R232
CP3	-20.33	D176, F179, Y180, E181, K182, G183, V184, Y185, K207, V208, N209, G210, T211, L213		E181, K182, V184, N209, G210
CP4	-3.809	**H180, Y181, K182**, E183, G184, I185, Y186, K187, T189, Y190.	BXB	K182, I185, K187, T189
CP4b	-8.224	E2, R3, R7, K182, E183, G184, G205, T206, E207, N208, G209, Y212, L214, Y221, G224, E225, N226, G227, T228, R230		R3, E183, T206, E207, N208, G209, Y212, E225, N226
CP5	-15.4	**Y183, K184, K185, G186,** V187, Y188, V189, Q209, D210, L212, Y214, L216, G226, D227, E228, R232	XBBX	K184, K185, Q209
CP6	-27.69	F175, Y177, V189, Q190, **K191, A192, G193, K194, R195**, T237, N238, N239, C240, S241, E244	BXXBB	Q190, G193, R195, N238, N239, S241

Table 2: Docking scores, contact residues, sequence motif patterns, and H-bonding interactions for the highest scored conformations of heparin molecule, docked onto the *NA* mature cysteine proteases.

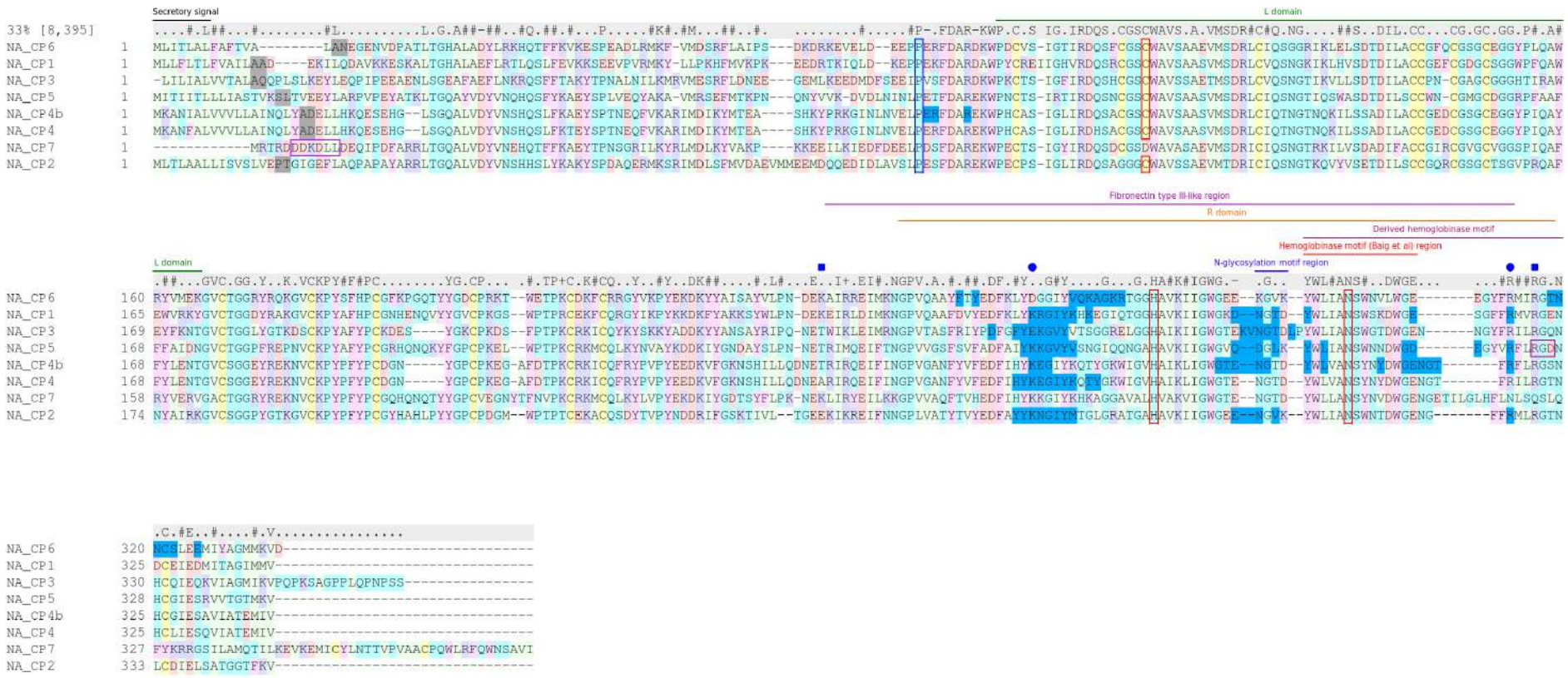

Figure 1 : Sequence alignment of the *NA* CPs with annotated functional regions. The begin position of the mature proteases is boxed in blue and the active-site catalytic triad residues are within red box . The residues spanning the cleavage sites between the extracellular signal peptide and the zymogen are shaded in grey. The lysosome-targeting peptide of *NA* CP7 is boxed. The derived hemoglobinase motif (region indicated in dark red) appears in *NA* CP2 and CP3 . The RGD motif signature in CP5 for fibronectin type III-like domain is boxed. The residues in contact with the docked heparin molecule are shaded in blue. The K and R residues contacting heparin in most of the docked complexes are indicated by blue spheres to denote their positions with respect to the residues corresponding to human cathepsin B's K157 and R234, (shown by blue squares) that bind heparin according to human Cathepsin B-heparin docking/MD (Costa *et al*, 2010).

Figure 2 : Sequence alignment of the NA CPs with HMWK-cleaving and heparin-binding cruzain and human cathepsin B. The S2 subsite residues are indicated by pink spheres. The conserved Ala and the crucial acidic residues of the S2, shared between the cathepsins and two of the hookworm CPs, are boxed. The K157 and R234 residues hypothesized to be crucial for heparin-binding in human cathepsin B (Costa et al, 2010) are boxed in blue.

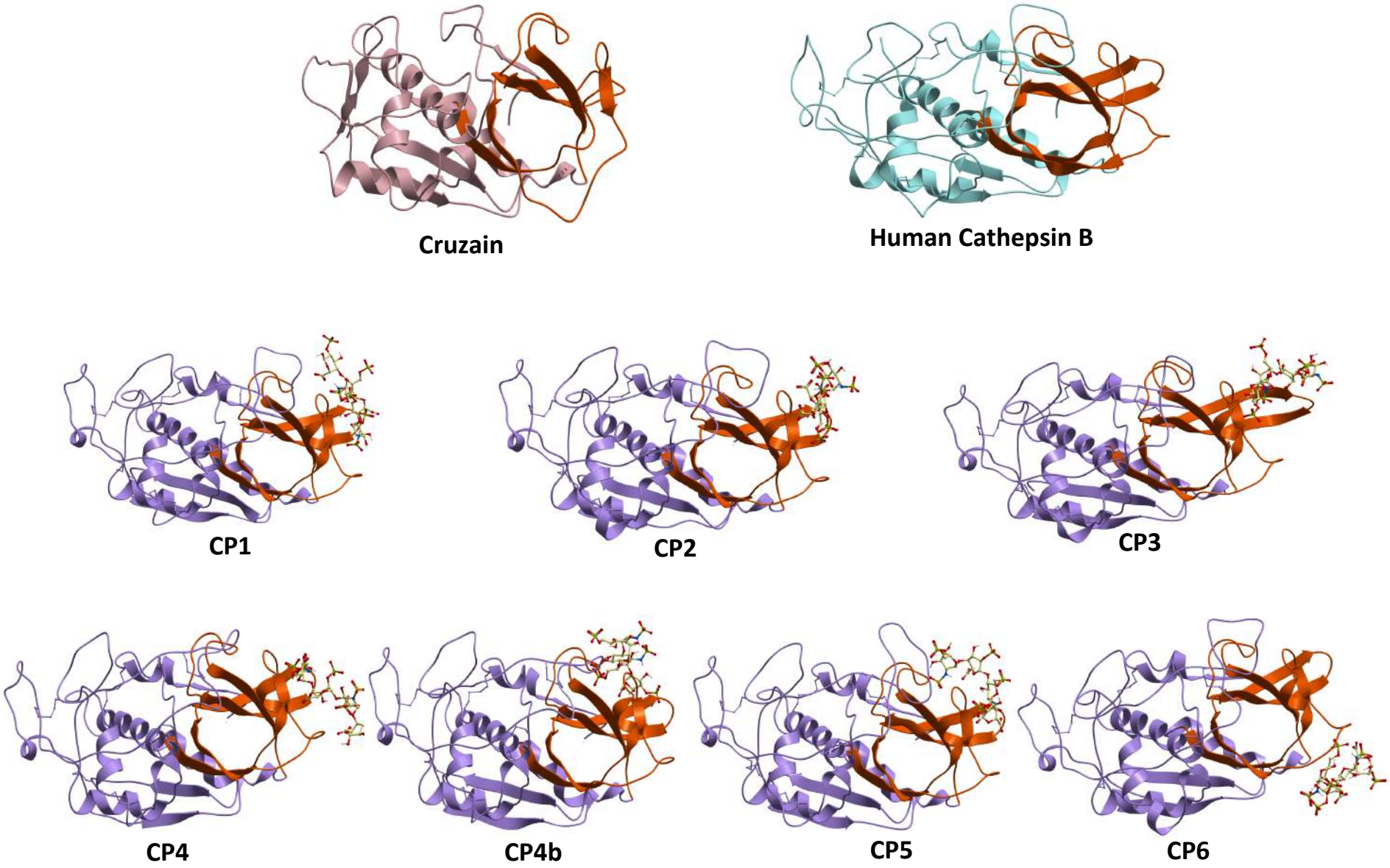

Figure 3: The R domains (orange) of Cruzain (PDB ID: 2OZ2) and human Cathepsin B (PDB ID: 1HUC) implicated to bind heparin, compared to structurally similar R domains (orange) of the *NA*-CP homology models. The antiparallel beta strand and loop dominated domain overlaps with the fibronectin type III-like region at the C-terminus of the *NA* CPs. The docked heparin molecule (represented in stick) mostly makes contact with the loops of the highlighted domain.

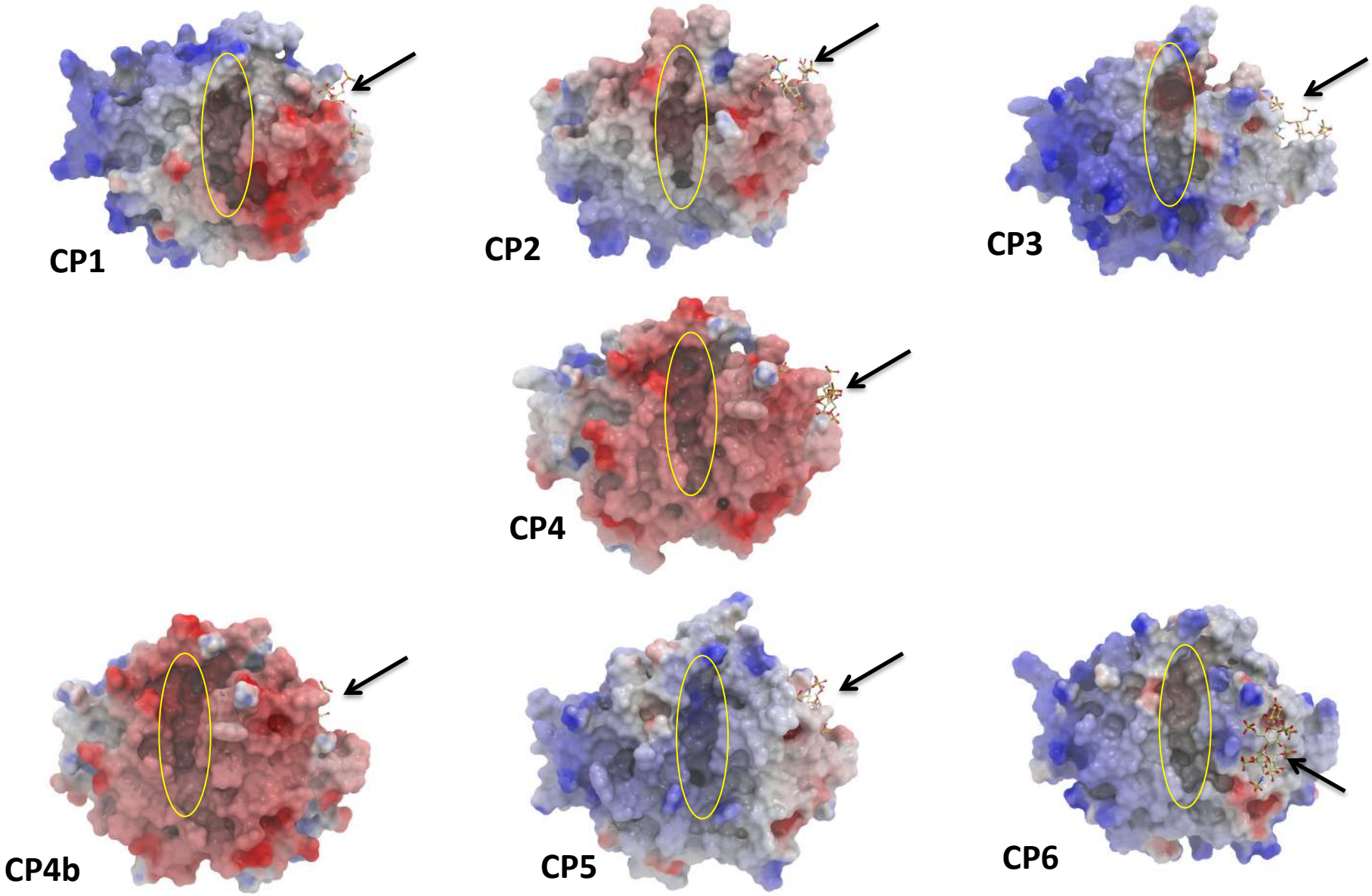

Figure 4: Heparin-docked homology models of the *NA*-CPs in electrostatic surface represention. The active sites of the cysteine proteases, appearing as clefts, are shown within yellow ovals. The location of the docked heparin (represented in stick) which bind away from the active site are indicated by arrows.