bioRxiv preprint doi: https://doi.org/10.1101/116921; this version posted March 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Cysteine proteases of hookworm *Necator Americanus* as virulence factors and implications for future drug design: 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

Conceived and designed the experiments: AB, Performed the experiments: AB, Analyzed the data: AB, Contributed tools: RA, Wrote the paper: AB, RA

bioRxiv preprint doi: https://doi.org/10.1101/116921; this version posted March 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 were also observed to have N terminal signal peptide for extracellular localization, and 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. Active site correlation of the secretory CPs with human pKal that cleaves high molecular weight kininogen (HMWK) to prevent platelet activation is suggestive of the involvement of hookworm CPs in preventing the formation of blood clots via this pathway. *NA* CP1, CP2, CP3, CP5 and CP6 were predicted to bind heparin, which is the glycosaminoglycan molecule that has been demonstrated to aid the 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 that cleaves HMWK. These observations therefore lead to the hypothesis that the functions of the hookworm CPs, which would probably include blood clot prevention, could be assisted by heparin. This study underscores the potential of synthetic heparin analogs as molecular treatment for hookworm infection, which could have implications for future drug design.

bioRxiv preprint doi: https://doi.org/10.1101/116921; this version posted March 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

bioRxiv preprint doi: https://doi.org/10.1101/116921; this version posted March 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

molecules, where in their folded proenzyme form, a self-inhibitory peptide blocks their mature catalytic domain. The removal of the inhibitory peptide upon proteolytic action of other peptidases releases the functional active cysteine proteases. The active sites 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 an amino-terminal domain that is mostly α-helical, and the carboxy-terminal is β sheet dominated. Another clan CA gut-localized and hemoglobin-degrading cathepsin B C1 peptidase - SmCB1 enzyme from *S.mansoni* (Caffrey *et al*, 2004) – is a close structural homologue of the *NA* CPs. The antiparallel β strand rich C-terminal domain shared by these proteases has been implicated to act as hemolysin in lysine-dependent gingipain (Li *et al*, 2010) which is a clan CD cysteine protease having similar mechanisms to clan CA proteases. 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), very 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 cell invasion, 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

bioRxiv preprint doi: https://doi.org/10.1101/116921; this version posted March 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

therapeutics. The approaches to some of the methods adopted here and some discussions pertaining to the physiology of the hookworm are in the context of other relevant cysteine proteases. Implications for the usage of heparin-analogs as inhibitors of the worm 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 and/or deciphered sites of functional importance were mapped onto the aligned CP sequences to denote their positions in the protein sequence.

1.1 N-terminal:

The N-terminal pre-sequences of *NA*-CP1 - CP6 were used to derive PRATT (Jonassen *et al*, 1995) patterns within them. The lack of these patterns was looked for in CP7 by ScanProsite (Sigrist *et al*, 2002) to determine CP1-6 specific N-terminal signature, as such signals often holds clue to protein sorting (Blobel, 1980).

1.2 Signal peptide cleavage:

SignalP (Petersen *et al*, 2011) was used to predict cleavage sites in the proteases, where the signal peptide would be cleaved off to generate the proenzymes.

1.3. Subcellular localization:

The *NA*-CP sequences were submitted for subcellular location 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.

bioRxiv preprint doi: https://doi.org/10.1101/116921; this version posted March 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1.4 Hemoglobinase motif:

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. Canine* CP1 (Q11006), *A. Canine* 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 organisms' proteins in the non-hemoglobin degrading set were: C. elegans_CPR3 (AAA98789.1), C.elegans_CPR4 (AAA98785.1) and L. major cathepsin B (AAB48119.1). Such patterns specific to the hemoglobin degrading enzymes (when found) were scanned in the rest of the *NA* CP sequences.

1.5 Emulation of human plasma kallekrein activity:

Human plasma kallekrein (pKal), a component of the anticoagulation pathway cleaves high molecular weight kininogen (HMWK) to eventually block platelet activation and degranulation (Da'dara *et al*, 2011; Maurer *et al,* 2011). Emulation of human plasma kallekrein activity has been reported for cruzain - cysteine protease of *T.Cruzi* (Del Nery *et al*, 1997; Lima *et al, 2002*). Such function has been suggested also for SmSP1 - serine protease of *S.mansoni* (Mebius *et al*, 2013), as kallikrein-like activity had been reported for a serine protease from *S.mansoni* (Carvalho *et al*, 1998). The pathogens secreting these proteases traverse host bloodstream and therefore have a survival need to prevent clot formation around them, which explains their HMWK-cleaving activity. As *NA* larvae too traces migratory route through blood capillaries of human hosts, the *NA*-CPs were scrutinized for molecular features, which could carry out such functions as cleaving HMWK. Homology of the *NA*-CPs with human plasma kallekrein's (pKal's) active site was studied with the help of PDB structure 4OGX at 2.4Å resolution, which is a co-crystal of pKal and IgG1 monoclonal antibody Fab DX-2930, where the antibody blocks the active site of pKal (Kenniston *et al*, 2014). The antibody-interacting active site residues from kallekrein were determined in Chimera (Pettersen *et al*, 2004). Fab Chain H residues from Ser25 to His31 close to pKal were selected, and then the residues within 4Å of the selection were listed to figure the antibody-interacting pKal residues. The process

bioRxiv preprint doi: https://doi.org/10.1101/116921; this version posted March 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

was repeated with Fab Chain H for the residue stretch Gly102-Glu108, and for Fab Chain L residues Ser52-Val58. The pKal active site residues thus listed from Chimera were mapped onto the pKal – *NA*- CP sequence alignment, done using clustalW (Thompson *et al*, 1994) algorithm. The alignment was generated in PBIL server (Perriere *et al*, 2003), where the scoring matrix was set to BLOSUM, with gap opening penalty of 2.40 and gap extension penalty of 0.15.

2. Heparin binding:

2.1 Heparin binding domain prediction:

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 domains from other organisms, with known heparin binding functionality.

2.2. Heparin binding docking simulation:

Heparin was included in the study to explore the feasibility of its interaction with the *NA*-CPs, as heparin-like glycosaminoglycan (GAGs) displayed on host proteoglycans (Bartlett *et al,* 2010*)*, are most probably encountered by the worm ES products at the host-pathogen interface.

Lack of experimental three-dimensional structure of the cysteine proteases prompted BLAST search for homology model templates, 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% residues in the allowed regions.

bioRxiv preprint doi: https://doi.org/10.1101/116921; this version posted March 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 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:

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 (**Figure 2**). CP7's N-terminal however had a lysosomal targeting pattern [DE]x{3}L[LI], according to LocSigDB. The specific patterns in the pre-sequences are summarized in **Table1**, along with the lysosome targeting peptide in CP7.

1.2 Signal peptide cleavage:

SignalP derived cleavage sites predicted the length and the peptide sequence for the signals contained in the pre-sequences of CP1-6. **Figure 2** shows the positioning of the cleavage sites where the signal peptide would be cleaved off to generate the proenzymes. The lengths of these signal peptides across the *NA* CPs were approximately the same.

1.3 Subcellular localization:

The consensus from the subcellular 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 a lysosome directed protease (**Table 1**).

1.4 Hemoglobinase motif:

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

bioRxiv preprint doi: https://doi.org/10.1101/116921; this version posted March 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

degrading cysteine proteases of the following blood feeders viz: *NA* (CP3), *A. Canine*, *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 3**. 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 Emulation of human plasma kallekrein activity:

Scanning of the *NA*-CPs for their possible emulation of human pKal's function of cleaving HMWK to prevent platelet activation and degranulation was motivated by such cleaving functions reported for cruzain from *T.Cruzi* which survive in host bloodstream as trypomastigotes (Del Nery *et al*, 1997; Lima *et al, 2002*). The scan or sequence homology search was additionally prompted by mouse pKal-homologous SmSP1 from *S.mansoni* which is an obligatory blood fluke and whose serine protease had been reported to exhibit kallikrein-like activity (Carvalho et al, 1998). SmSP1 therefore has been proposed to perform HMWK cleaving functions to prevent blood clots for the survival of *S.mansoni* in blood (Mebius et al, 2013). In case of bloodstream navigating *NA* (larvae stage), though there was minimal identity/similarity between its secretory CPs and human pKal, the structures shared similar domains with antiparallel β strands (**Figure 4**). The overall sequence identity between pKal and CP1-6 were within 15.12% to 19.37%, and the sequence similarity ranged from 20.70% to 24.56%. The antibody-blocked pKal active site corresponded with the Cysteine and Asparagine residues of the *NA* CP catalytic triad in the sequence alignment (**Figure 5A**, **5C**). Conservation with the CPs was noted along the LCG (**Figure 5A**) and WGEG (**Figure 5C**) sequence stretches of the pKal active site.

2. Heparin binding:

2.1 Heparin binding domain prediction:

CP2 and CP5 sequences were predicted by ProDom to have type III fibronectin domain, which is capable of binding glycosaminoglycan such as heparin or heparan sulfate of extracellular proteoglycans (Pankov *et al*, 2002). With CP2 as reference, the fibronectin domain region in the rest of the secretory CPs showed sequence identity within 52.72% to 60.37% and the sequence similarity ranged from 55.21% to 60.26%. The high sequence identity and similarity implied that all the CPs would be predisposed to heparin binding, and so the entire stretch – dominated by antiparallel β strands and loops at the C-terminal (**Figure 4**) –

bioRxiv preprint doi: https://doi.org/10.1101/116921; this version posted March 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

encompassing the predicted fibronectin domain (**Figure 6**) was considered for docking heparin.

2.2. Heparin binding docking simulation:

Heparin on being docked at the putative *heparin-binding* fibronectin domain of the cathepsin B-like CPs showed the best scored conformation to bind surface loops away from the enzymatic cleft (**Figure 7**), 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 basic residues K154 and R235 for binding the negatively charged heparin, mapped close to this study's heparin-interacting K and R residues in most of the *NA*-CPs (derived from cathepsin B - *NA* CP: alignment not shown). This is the closest comparison, which could be drawn, with no structures of cathepsin B – heparin complex available in PDB (**Figure 6**). The binding site residues within 4Å of the ligand are underscored in the sequence alignment (**Figure 6**). Barring CP6, heparin occupied similar region in all the CP homology models. The positioning of heparin was slightly shifted in CP6, albeit away from the active site (**Figure 7**). The overall electrostatics 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. The parasite's physiology involves cysteine proteases for digestive purposes and evading potentially damaging host hemostatic events, only some of which have been characterized – that this study attempts to decode.

1. Sequence alignments and pattern detection:

1.1 N-terminal:

CP1-6 proteins' hydrophobic N-terminal pre-sequence are presumably signal peptides for extracellular localization. N-terminal signals are extremely

bioRxiv preprint doi: https://doi.org/10.1101/116921; this version posted March 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 Signal peptide cleavage:

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 Subcellular localization:

CP1-6 are secreted out as per the subcellular localization predictions from this study, suggesting their presence in the ES products for host-pathogen interactions. CP3, amongst these, has been implicated to be present in the gut of an adult worm (Ranjit *et al*, 2008) and has been shown to be involved in the hemoglobin degradation pathway (Ranjit *et al* 2009). Therefore, CP3 being predicted to be secretory; CP1-6 - as per their predicted localization - could be expected to be in the ES products of *NA* for similar or other supportive functions 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:

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 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 Emulation of human plasma kallekrein activity:

Human plasma Kallekrein (pKal) is a serine protease that cleaves high molecular weight kininogen (HMWK) to produce bradykinin (Da'dara *et al*, 2011; Maurer *et*

bioRxiv preprint doi: https://doi.org/10.1101/116921; this version posted March 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

al, 2011). This molecule stimulates prostacyclin (PGI_2) release from endothelial cells, and PGI_2 in turn is an inhibitor of platelet activation and degranulation pathway (Maurer *et al*, 2011). Cruzain from *T.Cruzi* has been shown, and SmSP1 from *S. mansoni* has been proposed, to play roles in the production of bradykinin by acting on HMWK, the way plasma Kallekrein does (Del Nery *et al*, 1997; Lima *et al,* 2002, Mebius *et al*, 2013), and thereby evading host's primary hemostasis. The cysteine proteases of *NA,* could possibly mimic the active site of pKal, for the purpose of blood thinning to survive in the bloodstream like *T.Cruzi* and *S.mansoni*. The HMWK degradation function could be especially possible with the active sites of pKal and the *NA* CPs corresponding in alignment, encompassing the Asparagine and Cysteine residues of the catalytic triad of the CPs (**Figure 5A**, **5C**). Blood coagulation prevention by thwarting primary hemostasis via this pathway could be employed by cysteine proteases of *NA* larvae, which migrate through blood circulation.

2. Heparin binding:

2.1 Heparin binding domain prediction:

The β sheet and loop dominated C-terminal domains of the *NA*-CPs encompassing the predicted segments of fibronectin domain; structurally resembled the C-terminal domains of cruzain and human cathepsin B (**Figure 4**), which are known to bind heparin (Lima *et al, 2002,* Almeida *et al* 2001, Costa *et al*, 2010). The negatively charged heparin occupied neutral-positive electrostatics patches in most of the different *NA* CPs (**Figure 7**) 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 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 K154 and R235 (**Figure 6**).

2. 2 Heparin binding docking simulation:

Heparin-like glycosaminoglycan present in animal plasma membrane and ECM, form components of the host-pathogen interface. 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 attached to syndecans and glypicans proteoglycans of host cells, harboring important binding sites for parasitic cysteine proteases (Bartlett *et al,* 2010*).* The simultaneous binding of these proteases with GAG and their protein substrate results in the formation of ternary complexes (Lima *et al,* 2002*;* Judice *et al,* 2013)*,* which facilitates the enzymatic action of the proteases. Soluble proteoglycans or GAG chains released upon proteolytic cleavage of ECM or cell surface components can

bioRxiv preprint doi: https://doi.org/10.1101/116921; this version posted March 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

function in similar ways as their immobilized counterpart (Bartlett *et al,* 2010). Furthermore, this type of heparin binding modulates the activity of the cysteine proteases, specifically cathepsin B - which otherwise tends towards alkaline pH induced inactivation. (Almeida *et al* 2001*,* Costa *et al,* 2010). The implications of heparin binding to *NA* CP1, CP2, CP3, CP5 and CP6, as indicated by the docking scores, (**Table 2**) are discussed in the light of heparin binding to other cysteine proteases as follows.

2.2.1 Modulation of catalytic activity:

Human Cathepsin B bound by heparin/heparin-sulfate does not undergo alkaline pH induced loss of catalytic activity. 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 the same 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 6**). 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 of have varying degrees of overall surface electrostatics (**Figure 7**).

2.2.2 RBC lysis and hemoglobin degradation:

Heparan sulfate proteoglycans have been identified on the cell surfaces of Red Blood Cells (Vogt *et al*, 2004). These possibly serve as receptors for parasitic cysteine proteases like Lysine-dependent gingipain with antiparallel β strand rich domain that has been implicated to be a hemolysin (Li *et al,* 2010). *NA* hookworm CPs which share the antiparallel β sheet fold and many of which are believed to be present in the gut of the adult worm to assist blood feeding (Ranjit *et al,* 2008), could be utilizing such proteoglycans as co-receptors (by forming ternary complexes with heparin) to gain access to the red blood cells for hemoglobin degrading purposes.

2.2.3 Activation of anticoagulation pathway:

2.2.3.1 Kininogenonase activity:

Parasitic cysteine protease cruzain from *T.Cruzi* (which traverses capillary vessels and bloodstream in human hosts as trypomastigotes) is known to activate the kinin pathway to produce Lys-bradykinin (Del Nery *et al*, 1997; Lima

bioRxiv preprint doi: https://doi.org/10.1101/116921; this version posted March 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

et al, 2002), which is a potent vasodilator (Carvalho *et al,* 1998) and is also capable of stimulating prostacyclin (PGI_2) production from endothelial cells (Maurer *et al,* 2011). PGI_2, which is another vasodilator is also an inhibitor of platelet activation and degranulation pathway and in effect could trigger the anticoagulation pathway (Mebius *et al,* 2013). Cruzain with its cathepsin B-like substrate specificity has been implicated in the cleavage of high molecular weight kininogen (HMWK) and low molecular weight kininogen (LMWK), via its carboxyl-terminal domain, to produce Lys-bradykinin (Del Nery *et al,* 1997). Heparin induces the kininogenasic activity in cruzain by forming a complex with kininogen and cruzipain, for assisting the proteolytic cleavage of kininogen to release Lys-bradykinin (Lima *et al,* 2002). The docking results from this study suggest high likelihood of heparin binding to most of the cathepsin B-like *NA* CPs, which in turn can bind kininogen as its substrate to form ternary complex. The implications of such interactions underscore hookworm CPs role in possible kininogen cleaving activity for generating vasodilators to aid hookworm larvae's migration through small blood vessels.

2.2.3.2 Antithrombin activity:

The disruption in flow (stasis) of the ingested blood at the adult hookworm's gut would tend to trigger coagulation (Lowe 2003; Bagot *et al,* 2008). Inhibiting such coagulation would be a prerequisite for blood degradation. Heparin is known to serve as co-factors to antithrombin - an inhibitor of thrombin and other coagulation factors in the blood plasma. The molecule by increasing the thrombin binding efficiency of antithrombin by 2,000 to 10,000 fold (Beck *et al,* 1985) becomes extremely effective in preventing blood clots. Soluble free heparin-like GAG chains could be possibly released by *NA* CPs from the proteolytic cleavage of host proteoglycans. The free heparin chains thus released could bind plasma antithrombin to enhance anticoagulation activity of hookworm.

2.2.4 Cell invasion and migration:

Fibronectin is a ubiquitous protein of ECM that forms the scaffolding material for maintaining tissue organization and composition (To *et al,* 2011). Parasitic cysteine protease B secreted from *Leishmania sp* has been shown to degrade ECM fibronectin (Kulkarni *et al,* 2008), aided by heparin (Judice *et al,* 2013) Similarly, heparin-bound cysteine proteases of *NA* can potentially participate in fibronectin degradation, where the ECM proteoglycans could possibly facilitate the proteases' interaction with fibronectin. Though aspartic proteases have been identified as the key players in fibronectin degradation for skin penetration (Brown *et al,* 1999), the implication of parasitic cysteine protease in local migration of *Leishmania* (Kulkarni *et al,* 2008), leaves the scope for hypothesizing that heparin-bound *NA*-CPs could degrade fibronection possibly for tissue invasion (along with their reported fibrinogenolytic activity, possibly for blood clot degradation), during the migratory course of the hookworm larvae from heart-lungs-gastrointestinal tract (Pearson *et al,* 2012).

bioRxiv preprint doi: https://doi.org/10.1101/116921; this version posted March 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2.2.5 Therapy:

NA-CPs' suggestive role in RBC lysis, hemoglobin degradation, anticoagulation and cell invasion/migration - all 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, localized strategically for important 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). The *NA* CPs' active site similarity with pKal, presence of fibronectin domain signature and capability to bind heparin hint towards the CPs' 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 points towards *NA* CPs' heparin-assisted engagement in HMWK-cleaving activity for blood thinning. Such function, verified for the proteases of 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, calls for 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 *NA* cysteine proteases, which could be subjected to further experimental verification.

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

bioRxiv preprint doi: https://doi.org/10.1101/116921; this version posted March 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

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

bioRxiv preprint doi: https://doi.org/10.1101/116921; this version posted March 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

Caffrey, C.R., McKerrow, J.H., Salter, J.P., Sajid, M., 2004. Blood 'n' guts: an update on schistosome digestive peptidases. Trends Parasitol; 20:241-8

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

bioRxiv preprint doi: https://doi.org/10.1101/116921; this version posted March 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

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

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

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.

bioRxiv preprint doi: https://doi.org/10.1101/116921; this version posted March 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

Kenniston, J.A., Faucette R. R., Martik, D., Comeau, S.R., Lindberg, A. P., Kopacz K.J., Conley G. P., Chen, J., Viswanathan, M., Kastrapeli, N., Cosic, J.,

Mason, S., DiLeo, M., Abendroth, J., Kuzmic, P., 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

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

bioRxiv preprint doi: https://doi.org/10.1101/116921; this version posted March 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Kulkarni, M.M., Jones, E.A., McMaster, W.R., McGwire, B.S., 2008. Fibronectin Binding and Proteolytic Degradation by Leishmania and Effects on Macrophage Activation. Infection and Immunity, 76(4): 1738–1747.

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

Li, N., Yun, P., Nadkarni, M.A., Ghadikolaee N.B., Nguyen K.A., Lee M., Hunter N., Collyer C.A., 2010. Structure determination and analysis of a hemolytic adhesin domain from Porphyromonas gingivalis. Mol Microbiol; 76:861-873.80.

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.

bioRxiv preprint doi: https://doi.org/10.1101/116921; this version posted March 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

Perriere, G., Combet, C., Penel, S., Blanchet, C., Thioulouse, J., Geourjon, C., Grassot, J., Charavay, C., Gouy, M., Duret, L., and Dele´age, G., 2003. Integrated databanks access and sequence/structure analysis services at the PBIL. Nucleic Acids Res 31:3393

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

bioRxiv preprint doi: https://doi.org/10.1101/116921; this version posted March 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

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

Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res Nov 11; 22(22): 4673-4680

To, W.S., Midwood, K.S., 2011 Plasma and cellular fibronectin: distinct and independent functions during tissue repair. Fibrogenesis & Tissue Repair, 4:21

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

bioRxiv preprint doi: https://doi.org/10.1101/116921; this version posted March 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Ancylostoma braziliense in the metacarpal footpads of dogs. Z Parasitenkd; 53:263-6.

Vogt, A.M., Winter G., Wahlgren M., Spillmann D., 2004. Heparan sulphate identified on human erythrocytes: a Plasmodium falciparum receptor. Biochem. J. 381: 593–597

bioRxiv preprint doi: https://doi.org/10.1101/116921; this version posted March 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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: The subcellular localization of the *NA* cysteine proteases, and the motifs pertaining to the signals for the localization.

bioRxiv preprint doi: https://doi.org/10.1101/116921; this version posted March 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

NA-CP	Score	Contact Residues	Motif pattern	H-bonding residues
CP1	-21.75	**K270, R271, G272**, I273, Y274, K275, K277, D295, N296, T298, Y300, E313, R318	BBX	R271, K275, N296, R318
CP2	-24.22	Y276, Y277, K278, N279, G280, I281, Y282, M283, E303, N304, V306, R326.		Y276, K278, N279, M283, E303, N304, R326
CP3	-20.33	D266, F269, Y270, E271, K272, G273, V274, Y275, K297, V298, N299, G300, T301, L303		E271, K272, V274, N299, G300
CP4	-3.809	**H268, Y269, K270**, E271, G272, I273, Y274, K275, T277, Y278.	BXB	K270, I273, K275, T277
CP4b	-8.224	E90, R91, R95, K270, E271, G272, G293, T294, E295, N296, G297, Y300, L302, Y309, G312, E313, N314, G315, T316, R318		R91, E271, T294, E295, N296, G297, Y300, E313, N314
CP5	-15.4	**Y272, K273, K274, G275,** V276, Y277, V278, Q298, D299, L301, Y303, L305, G315, D316, E317, R321	XBBX	K273, K274, Q298
CP6	-27.69	F256, Y258, V270, Q271, **K272, A273, G274, K275, R276**, T318, N319, N320, C321, S322, E325	BXXBB	Q271, G274, R276, N319, N320, S322

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

```
33% [8,395]        ....#.L##...#........#L..........L.G.A##-##..#Q.##.#...P.....#K#.#M...##..#.   ........#.....#P-.FDAR
NA_CP1         1   MLLFLTLFVAILAAD----EKILQDAVKKESKALTGHALAEFLRTLQSLFEVKKSEEVPVRMKY-LLPKHFMVKPK---EEDRTKIQLD--KEPPEKFDAR
NA_CP6         1   MLITLALFAFTVA--------LANEGENVDPATLTGHALADYLRKHQTFFKVKESPEADLRMKF-VMDSRFLAIPS---DKDRKEVELD--EEPPERFDAR
NA_CP3         1   -LILIALVVTALAQQPLSLKEYLEQPIPEEAENLSGEAFAEFLNKRQSFFTAKYTPNALNILKMRVMESRFLDNEE---GEMLKEEDMDFSEEIPVSFDAR
NA_CP5         1   MITIITLLLIASTVKSLTVEEYLARPVPEYATKLTGQAYVDYVNQHQSFYKAEYSPLVEQYAKA-VMRSEFMTKPN---QNYVVK-DVDLNINLPETFDAR
NA_CP4         1   MKANFALVVVLLAINQLYADELLHKQESEHG--LSGQALVDYVNSHQSLFKTEYSPTNEQFVKARIMDIKYMTEA----SHKYPRKGINLNVELPERFDAR
NA_CP4b        1   MKANIALVVVLLAINQLYADELLHKQESEHG--LSGQALVDYVNSHQSLFKAEYSPTNEQFVKARIMDIKYMTEA----SHKYPRKGINLNVELPERFDAR
NA_CP7         1   ------------MRTRDDDKDLLDEQIPDFARRLTGQALVDYVNEHQTFFKAEYTPNSGRILKYRLMDLKYVAKP----KKEEILKIEDFDEELPDSFDAR
NA_CP2         1   MLTLAALLISVSLVEPTGIGEFLAQPAPAYARRLTGQALVDYVNSHHSLYKAKYSPDAQERMKSRIMDLSFMVDAEVMMEEMDQQEDIDLAVSLPESFDAR
```

Figure 1: N-terminal alignment of the *NA* CP sequences showing the lack the signal sequence in CP7 (denoted by arrow). The predicted lysosome targeting sequence for the protease is underscored.

```
37% [7,372]        M...#.L##...#........#L..........L.G.A##-##..#QS##.#..SP.....#K..#M...##....   ........#.....#PE.FDAR
NA_CP4         1   MKANFALVVVLLAINQLYADELLHKQESEHG--LSGQALVDYVNSHQSLFKTEYSPTNEQFVKARIMDIKYMTEA----SHKYPRKGINLNVELPERFDAR
NA_CP4b        1   MKANIALVVVLLAINQLYADELLHKQESEHG--LSGQALVDYVNSHQSLFKAEYSPTNEQFVKARIMDIKYMTEA----SHKYPRKGINLNVELPERFDAR
NA_CP2         1   MLTLAALLISVSLVEPTGIGEFLAQPAPAYARRLTGQALVDYVNSHHSLYKAKYSPDAQERMKSRIMDLSFMVDAEVMMEEMDQQEDIDLAVSLPESFDAR
NA_CP5         1   MITIITLLLIASTVKSLTVEEYLARPVPEYATKLTGQAYVDYVNQHQSFYKAEYSPLVEQYAKA-VMRSEFMTKP----NQNYVVKDVDLNINLPETFDAR
NA_CP1         1   MLLFLTLFVAILAAD----EKILQDAVKKESKALTGHALAEFLRTLQSLFEVKKSEEVPVRMKY-LLPKHFMVKPK---EEDRTKIQLD--KEPPEKFDAR
NA_CP6         1   MLITLALFAFTVA--------LANEGENVDPATLTGHALADYLRKHQTFFKVKESPEADLRMKF-VMDSRFLAIPS---DKDRKEVELD--EEPPERFDAR
NA_CP3         1   -LILIALVVTALAQQPLSLKEYLEQPIPEEAENLSGEAFAEFLNKRQSFFTAKYTPNALNILKMRVMESRFLDNEE---GEMLKEEDMDFSEEIPVSFDAR
```

Figure 2: N-terminal alignment of the *NA* CP1-CP6 sequences. The unique pattern M-x(4,5)-L derived from these pre-sequences is boxed. The residues encompassing the cleavage sites for each of the proteases are underscored.

Hemoglobinase motif (Baig *et al*, 2002)

```
                 ..Y..#Y..DK##....#.L....E..I+.EI#.NGPV.A.#.##.DF.#Y..G#Y....G...G.HA#K#IGWG.-  .G..  YWL#ANS#..DWGE .G.FR##RG.N.C.I
NA_CP4      219  FRYPVPYEEDKVFGKNSHILLQDNEARIRQEIFINGPVGANFYVFEDFIHYKEGIYKQTYGKWIGVHAIKLIGWGTE--NGTD--YWLVANSYNYDWGE-NGTFRILRGTNHCLI
NA_CP4b     219  FRYPVPYEEDKVFGKNSHILLQDNETRIRQEIFINGPVGANFYVFEDFIHYKEGIYKQTYGKWIGVHAIKLIGWGTE--NGTD--YWLVANSYNYDWGE-NGTFRFLRGSNHCGI
NA_CP2      229  SDYTVPYNDDRIFGSKTIVL--TGEEKIKREIFNNGPLVATYTVYEDFAYYKNGIYMTGLGRATGAHAVKIIGWGEE--NGVK--YWLIANSWNTDWGE-NGFFRMLRGTNLCDI
NA_CP5      223  LKYNVAYKDDKIYGNDAYSL-PNNETRIMQEIFTNGPVVGSFSVFADFAIYKKGVYVSNGIQQNGAHAVKIIGWGVQ--DGLK--YWLIANSWNNDWGD-EGYVRFLRGDNHCGI
NA_CP1      220  RGYIKPYKKDKFYAKKSYWL-PNDEKEIRLDIMKNGPVQAAFDVYEDFKLYKRGIYKHKEGIQTGGHAVKIIGWGKD--NGTD--YWLIANSWSKDWGE-SGFFRMVRGENDCEI
NA_CP6      215  RGYVKPYEKDKYYAISAYVL-PNDEKAIRREIMKNGPVQAAYFTYEDFKLYDGGIYVQKAGKRTGGHAVKIIGWGEE--KGVK--YWLIANSWNVLWGE-EGYFRMIRGTNNCSL
NA_CP3      220  YKYSKKYADDKYYANSAYRI-PQNETWIKLEIMRNGPVTASFRIYPDFGFYEKGVYVTSGGRELGGHAIKIIGWGTEKVNGTDLPYWLIANSWGTDWGENNGYFRILRGQNHCQI

                 E..#....#.V.
NA_CP4      329  ESQVIATEMIV----------------
NA_CP4b     329  ESAVIATEMIV----------------
NA_CP2      337  ELSATGGTFKV----------------
NA_CP5      332  ESRVVTGTMKV----------------
NA_CP1      329  EDMITAGIMMV----------------
NA_CP6      324  EEMIYAGMMKVD---------------
NA_CP3      334  EQKVIAGMIKVPQPKSAGPPLQPNPSS
```

Derived Hemoglobinase motif

Figure 3: Partial sequence alignment of the *NA*-CPs mapping the region of the derived longer 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], as compared to previously determined hemoglobinase motif by Baig *et al*. *NA* CP2 and CP3 which show the presence of the motif are denoted by arrows.

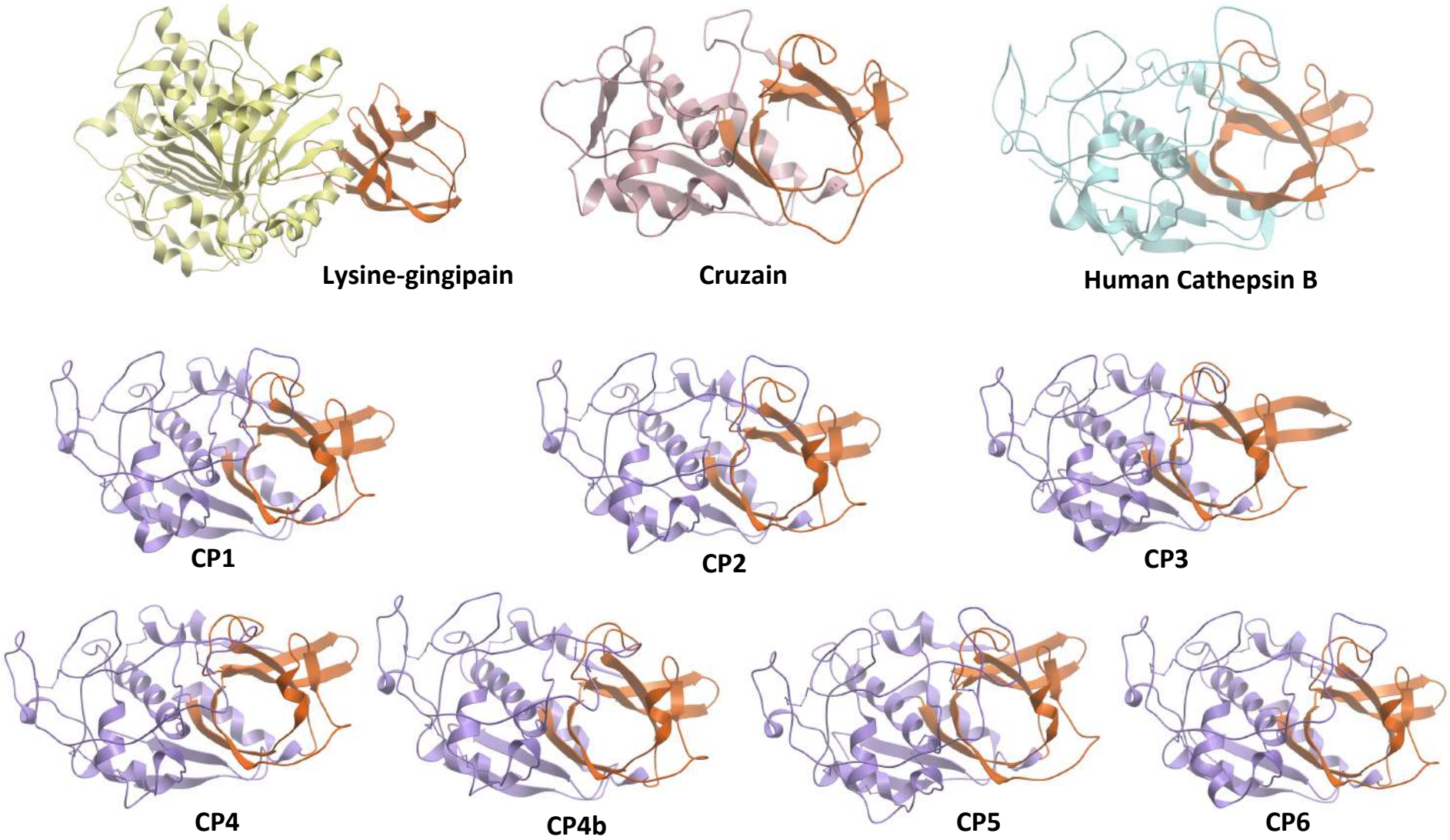

Figure 4: The C-terminal domains (orange) of hemolysis-causing Lysine-gingipain (PDB ID:4RBM), HMWK-cleaving Cruzain (PDB ID: 2OZ2), and human Cathepsin B (PDB ID: 1HUC) compared to: structurally similar C-terminal domains (orange) of the *NA*-CPs dominated by antiparallel beta strands and loops, and harboring the enzymatic residues **H** (Histidine) and **N** (Asparagine). The outer fringes of the *NA* CPs' C-terminal domains (loop regions) make contacts with the docked heparin molecule.

```
NA_CP1        MLLFLTLFVA ILAAD....E KILQDAVKKE SKALTGHALA EFLRTLQSLF
NA_CP6        MLITLALFAF TVA....... .LANEGENVD PATLTGHALA DYLRKHQTFF
NA_CP3        .LILIALVVT ALAQQPLSLK EYLEQPIPEE AENLSGEAFA EFLNKRQSFF
NA_CP4        MKANFALVVV LLAINQLYAD ELLHKQESEH G..LSGQALV DYVNSHQSLF
NA_CP4b       MKANIALVVV LLAINQLYAD ELLHKQESEH G..LSGQALV DYVNSHQSLF
NA_CP5        MITIITLLLI ASTVKSLTVE EYLARPVPEY ATKLTGQAYV DYVNQHQSFY
NA_CP2        MLTLAALLIS VSLVEPTGIG EFLAQPAPAY ARRLTGQALV DYVNSHHSLY
Human_pKal    ...IVGGTES SWGEWP..WQ VSLQVKLTAQ R.HLCGGSLI G....HQWVL

NA_CP1        EVKKSEEVP. .VRMKYLLPK HFMVKPKEED ...RTKIQLD ..KEPPEKFD
NA_CP6        KVKESPEAD. .LRMKFVMDS RFLAIPSDKD ...RKEVELD ..EEPPERFD
NA_CP3        TAKYTPNALN ILKMR.VMES RFLDNEEGEM ...LKEEDMD FSEEIPVSFD
NA_CP4        KTEYSPTNEQ FVKAR.IMDI KYMTEASHKY P..RKGINLN ..VELPERFD
NA_CP4b       KAEYSPTNEQ FVKAR.IMDI KYMTEASHKY P..RKGINLN ..VELPERFD
NA_CP5        KAEYSPLVEQ YAKA..VMRS EFMTKPNQNY V..VKDVDLN ..INLPETFD
NA_CP2        KAKYSPDAQE RMKSR.IMDL SFMVDAEVMM EEMDQQEDID LAVSLPESFD
Human_pKal    TAAHCFDGLP ......LQDV WRIYSGILEL ....SDITKD ....TP..FS

NA_CP1        ARDAWPYCRE IIGHVRDQSR CGSCWAVSAA SVMSDRLCVQ SNGKIKLHVS
NA_CP6        ARDKWPDC.V SIGTIRDQSF CGSCWAVSAA EVMSDRLCIQ SGGRIKLELS
NA_CP3        ARDKWPKC.T SIGFIRDQSH CGSCWAVSSA ETMSDRLCVQ SNGTIKVLLS
NA_CP4        AREKWPHC.A SIGLIRDHSA CGSCWAVSAA SVMSDRLCIQ TNGTNQKILS
NA_CP4b       AREKWPHC.A SIGLIRDQSA CGSCWAVSAA SVMSDRLCIQ TNGTNQKILS
NA_CP5        AREKWPNC.T SIRTIRDQSN CGSCWAVSAA SVMSDRLCIQ SNGTIQSWAS
NA_CP2        AREKWPEC.P SIGLIRDQSA GGGCWAVSSA EVMTDRICIQ SNGTKQVYVS
Human_pKal    ...QIKEI.. IIHQNYKVSE GNHDIALIKL QAPLEYTEFQ KPISLPSKGD
```

Figure 5A: Sequence alignment of human pKal with the *NA* cysteine proteases, where catalytic triad **C** (Cysteine) of the hookworm protease is shown by arrow. The active site residues of the pKal and the corresponding residues of *NA*-CPs in the alignment are shaded in grey.

```
NA_CP1        DTDILACCGE FCGDGCSGGW PFQAWEWVRK YGVCTGGDYR AKGVCKPYAF
NA_CP6        DTDILACCGF QCGSGCEGGY PLQAWRYVME KGVCTGGRYR QKGVCKPYSF
NA_CP3        DTDILACC.P NCGAGCGGGH TIRAWEYFKN TGVCTGGLYG TKDSCKPYAF
NA_CP4        SADILACCGE DCGSGCEGGY PIQAYFYLEN TGVCSGGEYR EKNVCKPYPF
NA_CP4b       SADILACCGE DCGSGCEGGY PIQAYFYLEN TGVCSGGEYR EKNVCKPYPF
NA_CP5        DTDILSCC.W NCGMGCDGGR PFAAFFFAID NGVCTGGPFR EPNVCKPYAF
NA_CP2        ETDILSCCGQ RCGSGCTSGV PRQAFNYAIR KGVCSGGPYG TKGVCKPYPF
Human_pKal    TSTIYTNC.. ....WVTG.. ....WGFSKE K.....GEIQ ..NILQKV.N

NA_CP1        HPCGNHENQV YYGVCPKG.S WPTPRCEKFC QRGYIKPYKK DKFYA.KKSY
NA_CP6        HPCGFKPGQT YYGDCPRK.T WETPKCDKFC RRGYVKPYEK DKYYA.ISAY
NA_CP3        YPCKDES... .YGKCPKD.S FPTPKCRKIC QYKYSKKYAD DKYYA.NSAY
NA_CP4        YPCDGN.... .YGPCPKEGA FDTPKCRKIC QFRYPVPYEE DKVFGKNSHI
NA_CP4b       YPCDGN.... .YGPCPKEGA FDTPKCRKIC QFRYPVPYEE DKVFGKNSHI
NA_CP5        YPCGRHQNQK YFGPCPKE.L WPTPKCRKMC QLKYNVAYKD DKIYG.NDAY
NA_CP2        YPCGYHAHLP YYGPCPDG.M WPTPTCEKAC QSDYTVPYND DRIFG.SKTI
Human_pKal    IPLVTN.... .......... .......EEC QKRY....Q. D..YK.....

NA_CP1        WLPNDEKEIR LDIMKNGPVQ AAFDVYEDFK LYKRGIYKHK EGIQTGGHAV
NA_CP6        VLPNDEKAIR REIMKNGPVQ AAYFTYEDFK LYDGGIYVQK AGKRTGGHAV
NA_CP3        RIPQNETWIK LEIMRNGPVT ASFRIYPDFG FYEKGVYVTS GGRELGGHAI
NA_CP4        LLQDNEARIR QEIFINGPVG ANFYVFEDFI HYKEGIYKQT YGKWIGVHAI
NA_CP4b       LLQDNETRIR QEIFINGPVG ANFYVFEDFI HYKEGIYKQT YGKWIGVHAI
NA_CP5        SLPNNETRIM QEIFTNGPVV GSFSVFADFA IYKKGVYVSN GIQQNGAHAV
NA_CP2        VLTGEEK.IK REIFNNGPLV ATYTVYEDFA YYKNGIYMTG LGRATGAHAV
Human_pKal    ........IT QRMVCAG... .......... .YKEGGKDAC KGDSGGPLVC
```

Figure 5B: Sequence alignment of human pKal with the *NA* cysteine proteases, where the catalytic triad **H** (Histidine) of the hookworm protease is shown by arrow. The active site residues of the pKal and the corresponding residues of *NA*-CPs in the alignment are shaded in grey.

```
NA_CP1          KIIG.WGKD. .NGTD..YWL IANSWSKDWG ES.GFFRMVR GENDCEIEDM
NA_CP6          KIIG.WGEE. .KGVK..YWL IANSWNVLWG EE.GYFRMIR GTNNCSLEEM
NA_CP3          KIIG.WGTEK VNGTDLPYWL IANSWGTDWG ENNGYFRILR GQNHCQIEQK
NA_CP4          KLIG.WGTE. .NGTD..YWL VANSYNYDWG EN.GTFRILR GTNHCLIESQ
NA_CP4b         KLIG.WGTE. .NGTD..YWL VANSYNYDWG EN.GTFRFLR GSNHCGIESA
NA_CP5          KIIG.WGVQ. .DGLK..YWL IANSWNNDWG DE.GYVRFLR GDNHCGIESR
NA_CP2          KIIG.WGEE. .NGVK..YWL IANSWNTDWG EN.GFFRMLR GTNLCDIELS
Human_pKal      KHNGMWRLV. ..GIT..SWG EGCAR...RE QP.GVYTKVA EYMDWILEKT

NA_CP1          ITAGIMMV.. .......... ....
NA_CP6          IYAGMMKVD. .......... ....
NA_CP3          VIAGMIKVPQ PKSAGPPLQP NPSS
NA_CP4          VIATEMIV.. .......... ....
NA_CP4b         VIATEMIV.. .......... ....
NA_CP5          VVTGTMKV.. .......... ....
NA_CP2          ATGGTFKV.. .......... ....
Human_pKal      QSSDGK.... .......... ....
```

Figure 5C: Sequence alignment of human pKal with the *NA* cysteine proteases, where the catalytic triad **N** (Asparagine) of the hookworm protease is shown by arrow. The active site residues of the pKal and the corresponding residues of *NA*-CPs in the alignment are shaded in grey.

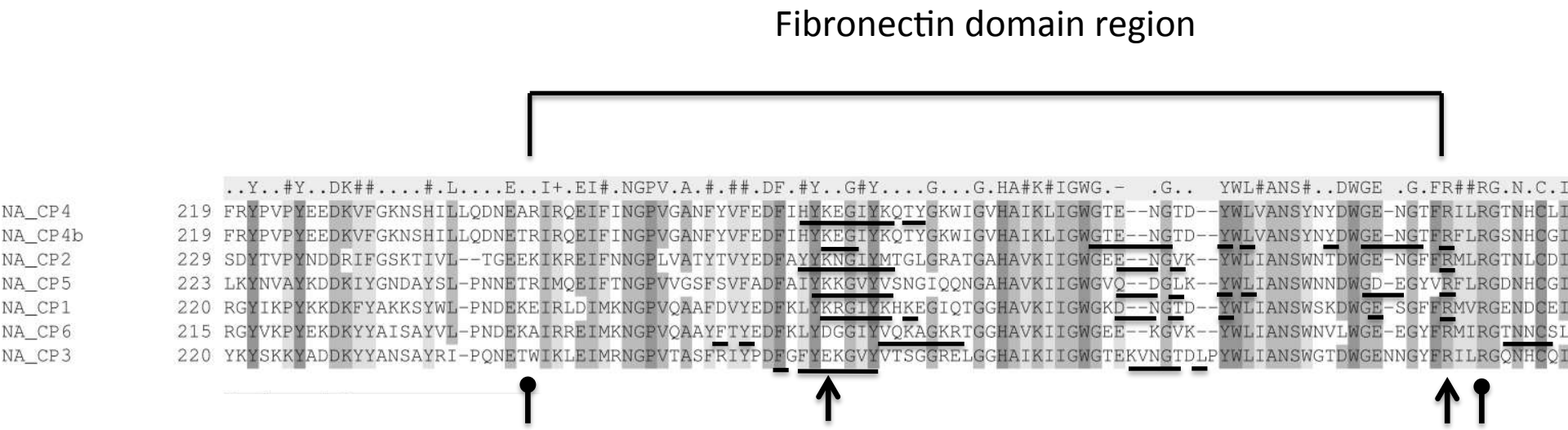

Figure 6: Partial sequence alignment of *NA* CP1-CP6 : region encompassing all the patches of the predicted fibronectin domain (for CP2 and CP5) has been denoted. The residue contacts made by heparin molecule in the proteases are underscored. Some of the N-terminal contacts (few) are not shown for brevity. The K and R residues in contact with heparin from most of the *NA* CP-heparin dockings are shown by arrow. The residues corresponding to human cathepsin B's K154 and R235 (derived from cathepsin B - *NA* CP alignment), which bind heparin according to human Cathepsin B-heparin docking/MD (Costa *et al*, 2010), are shown by spherical arrowheads.

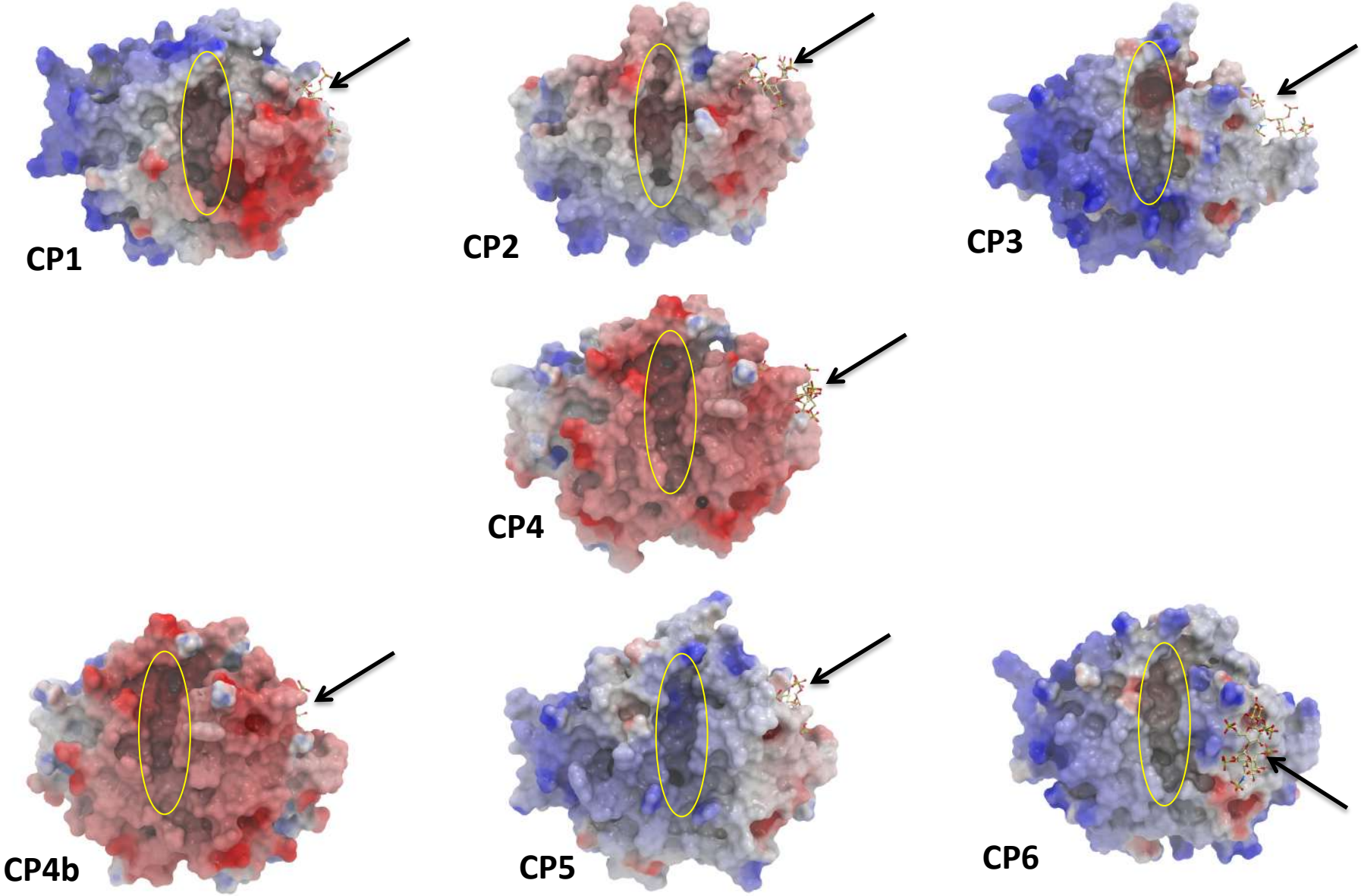

Figure 7: Heparin-bound three dimensional structures of the *NA*-CPs in electrostatic surface represention . The active site of the cysteine proteases which appears as clefts are facing the front (ovaled in yellow) and the location of the heparin (represented in stick form) which bind away from the clefts are indicated by arrows.