BÁRBARA MONTEIRO DE CASTRO ATAÍDE

ANÁLISE DE RESÍDUOS CONSERVADOS E CORRELACIONADOS DAS CISTEÍNO PROTEASES

Belo Horizonte

Minas Gerais – Brasil

Maio – 2019

BÁRBARA MONTEIRO DE CASTRO ATAÍDE

ANÁLISE DE RESÍDUOS CONSERVADOS E CORRELACIONADOS DAS CISTEÍNO PROTEASES

Orientadora: Dra Rafaela Salgado Ferreira

Co-orientador: Dr Lucas Bleicher

Dissertação submetida ao Departamento de Bioquímica e Imunologia do Instituto de Ciências Biológicas da Universidade Federal de Minas Gerais, como requisito parcial para a obtenção do grau de Mestre em Bioquímica e Imunologia

Universidade Federal de Minas Gerais

Belo Horizonte

Maio – 2019

Universidade Federal de Minas Gerais
Curso de Pós-Graduação em Bioquímica e Imunologia ICB/UFMG
Av. Antônio Carlos, 6627 – Pampulha
31270-901 – Belo Horizonte – MG
e-mail: pg-biq@icb.ufmg.br (31)3409-2615

ATA DA DEFESA DA DISSERTAÇÃO DE MESTRADO DE BÁRBARA MONTEIRO DE CASTRO ATAÍDE. Aos vinte e seis dias do mês de fevereiro de 2019 às 09:00 horas, reuniu-se no Instituto de Ciências Biológicas da Universidade Federal de Minas Gerais, a Comissão Examinadora da dissertação de Mestrado, indicada *ad referendum* do Colegiado do Curso, para julgar, em exame final, o trabalho intitulado ""Análise de resíduos conservados e correlacionados de cisteíno proteases"", requisito final para a obtenção do grau de Mestre em Bioquímica e Imunologia, área de concentração: Bioquímica. Abrindo a sessão, o Presidente da Comissão, Prof. Rafaela Salgado Ferreira, da Universidade Federal de Minas Gerais, após dar a conhecer aos presentes o teor das Normas Regulamentares do Trabalho Final, passou a palavra à candidata para apresentação de seu trabalho. Seguiu-se a arguição pelos examinadores, com a respectiva defesa da candidata. Logo após a Comissão se reuniu, sem a presença da candidata e do público, para julgamento e expedição do resultado final. Foram atribuídas as seguintes indicações: Dr. Erich Birelli Tahara (Universidade Federal de Minas Gerais), aprovada; Dra. Mariana Torquato Quezado de Magalhães (Universidade Federal de Minas Gerais), aprovada; Dr. Lucas Bleicher - Coorientador (Universidade Federal de Minas Gerais), aprovada; Dra. Rafaela Salgado Ferreira - Orientador (Universidade Federal de Minas Gerais), aprovada. Pelas indicações a candidata foi considerada:

(X) APROVADA
( ) REPROVADA

O resultado final foi comunicado publicamente à candidata pelo Presidente da Comissão. Nada mais havendo a tratar, o Presidente da Comissão encerrou a reunião e lavrou a presente Ata que será assinada por todos os membros participantes da Comissão Examinadora. Belo Horizonte, 26 de fevereiro de 2019.

Dr. Erich Birelli Tahara (UFMG)

Dra. Mariana Torquato Quezado de Magalhães (UFMG)

Dr. Lucas Bleicher - Coorientador (UFMG)

Dra. Rafaela Salgado Ferreira - Orientador (UFMG)

Profª Lêda Quércia Vieira
Coordenadora do Curso de Pós Graduação
em Bioquímica e Imunologia
ICB - UFMG

Agradecimentos

Agradeço primeiramente a Deus, por ter me sustentado e me guiado nessa longa trajetória, por ter dado força nos momentos de maior dificuldade, por ter colocado anjos em meu caminho que me ajudaram de inúmeras formas e por ter permitido que esse sonho se tornasse realidade! Sei que Ele é o braço invisível que me conforta, que me sustenta e me guia e sem O qual eu não chegaria até aqui... que eu sempre seja digna de Sua proteção!

Agradeço à minha família, aos meus irmãos Filipe, Carolina e Thiago, por acreditarem em mim e me darem apoio, e especialmente à minha mãe, Rute, por ter me encorajado nos momentos difíceis, por ter estado sempre presente (mesmo estando fisicamente distante). Mãe, te agradeço pelas palavras de incentivo, pelas vezes em que secou minhas lágrimas me fazendo rir... Que Deus continue guardando a senhora por todos os dias de sua vida!

Ao meu grande amor, Célio, por me apoiar, por compreender os momentos nos quais eu não podia estar ao seu lado por ter que me dedicar ao estudo, por aceitar pacientemente todas as privações de lazer as quais me impus. Obrigada meu amor por secar minhas lágrimas, por estar ao meu lado nos momentos de dor (inclusive física), por não me deixar fraquejar nem desanimar... Obrigada pelo leite quente, pelo abraço reconfortante, pelo colo e pelas brincadeiras que me faziam sorrir quando o que eu mais queria era chorar... Obrigada por trazer alegria aos meus dias e por existir em minha vida e ser quem você é! Você é o meu maior presente!!!

Aos meus amigos Mariana e Suelen, que mesmo distantes me incentivaram e me deram força... Mariana, seus conselhos me ajudaram demais! Aos amigos que conquistei no Hospital: Gislene, Paulo e Ronaldo... Gislene, obrigada pelas risadas (até as fora de hora – rsrsrs), pela amizade e carinho, pelos momentos em que discordamos, por tudo! Você me faz crescer muito! Paulo, obrigada pelos conselhos, pelo apoio, pelo carinho, pelo seu jeito sempre alegre e pela amizade! Você é um grande amigo! Ronaldo, obrigada pelos conselhos, pelo apoio e auxílio e por sempre me encorajar! Muito obrigada pela amizade!

À minha amiga e segunda mãe, Ilca, agradeço pelos ensinamentos de vida que sempre me deu! Obrigada pelos conselhos, pelo amor e por sempre acreditar em mim! Obrigada por me ensinar a acreditar em mim! Você é muito especial!

Aos colegas do Pronto Socorro do Hospital Odilon Behrens que trocaram plantões para que eu pudesse ir às aulas, fazer os trabalhos bem como cumprir com as demais atividades na UFMG.

À minha ex-coordenadora Rachel e à coordenadora Gabriela por me auxiliarem com os plantões, permitindo que eu conseguisse atender aos meus compromissos do Mestrado.

Ao professor Flávio Amaral por ter me aceitado inicialmente em seu laboratório... Obrigada pela confiança e receptividade e pela compreensão quando optei por outra linha de pesquisa!

À professora Ângela Ribeiro por ter me acolhido no LANEC, por ter acreditado no meu projeto inicial e ter me dado a liberdade de escolha... Obrigada pelo apoio quando adoeci e pela compreensão quando tive que deixar o LANEC!

Ao professor Fabrício de Araújo, por ter sido meu co-orientador no projeto de efeito placebo. Foi uma experiência maravilhosa e enriquecedora! Obrigada pelo apoio quando adoeci e por se disponibilizar a me auxiliar de alguma forma nesse momento tão difícil!

Ao professor Jader Cruz, pela carta de indicação, por sempre acreditar em mim e por se prontificar a me orientar quando eu possuía limitações de trabalho na pesquisa! Muito obrigada por tudo!

Aos meus professores e orientadores Rafaela Salgado e Lucas Bleicher por me aceitarem como aluna. Agradeço imensamente por me acolherem em um momento extremamente delicado e difícil para mim! Perdoem-me pelas dúvidas bobas e por algo que eu tenha feito que tenha desagradado de alguma forma... Se não fosse vocês certamente eu não teria chegado até aqui! Vocês possuem a minha eterna e profunda gratidão! Que Deus os recompense grandemente!

Aos colegas dos laboratórios por onde passei, muito obrigada pelo acolhimento, pelos bons momentos vividos e pelos ensinamentos!

Aos colegas do laboratório de Biologia Computacional de Proteínas, Viviane e Marcelo, que sem ao menos me conhecer estenderam-me as mãos, acolheram-me e sempre se mostraram disponíveis, auxiliando-me em minhas dúvidas e dificuldades! Que Deus os abençoe sempre!

Ao Departamento de Bioquímica e Imunologia por me propiciar um conhecimento tão rico e singular! Muito obrigada pela oportunidade de poder aprender um pouco mais e também poder contribuir para o crescimento da ciência!

Enfim, não foi fácil! Surgiram imprevistos e acontecimentos desagradáveis que quase me fizeram desistir. Mas o objetivo que me animava era mais forte e eu só sabia que não poderia regressar... Valeu à pena e agradeço a todos que de alguma forma me auxiliaram, me incentivaram e participaram dessa trajetória!

"Fica sempre um pouco de perfume nas mãos que oferecem rosas, nas mãos que sabem ser generosas!"

Alberto Costa

A Força Que Nunca Seca

Já se pode ver ao longe

A senhora com a lata na cabeça

Equilibrando a lata vesga

Mais do que o corpo dita

O que faz é equilíbrio cego

A lata não mostra

O corpo que entorta

Pra lata ficar reta

Pra cada braço uma força

De força não geme uma nota

A lata só cerca, não leva

A água na estrada morta

E a força nunca seca

Pra água que é tão pouca

Composição: Chico César/Vanessa da Mata

Resumo

As proteínas evoluem ao longo do tempo, devido ao acúmulo de mutações, inserções e deleções nos genes que as codificam. Posições fortemente conservadas geralmente referem-se a resíduos que são estritamente necessários para a estrutura da proteína ou para sua função. A metodologia DRCN (*Decomposition of Residue Coevolution Networks*) permite estudar coevolução de resíduos em famílias de proteínas. Para tanto, consiste em cinco etapas básicas: filtragem do alinhamento, cálculo de correlações, decomposição da rede, geração de arquivos de visualização auxiliares e anotação de posições por busca automática no UniProt. As cisteíno proteases são um grande e diverso grupo de enzimas que pertencem ao clã CA e à família C1 (família da papaína) e podem ser encontradas em todos os animais e reinos de plantas bem como em muitos vírus e organismos procariotas. O presente trabalho tem como objetivo analisar e discutir a importância dos resíduos correlacionados e conservados da família das cisteíno proteases a partir do alinhamento múltiplo de sequências. Para tanto utilizamos, o *software* PFSTATS, que emprega a metodologia DRCN. Foram encontrados 34 resíduos correlacionados dispostos em 5 comunidades de resíduos que coevoluem. A comunidade 1 é a maior comunidade, sendo formada por 22 resíduos que coevoluem tanto por questões estruturais, como por exemplo Cys22 e Cys63 e Cys56 e Cys101, que realizam pontes dissulfeto; quanto por questões funcionais, como Pro2, Phe28, Gly35 e Tyr89, que se localizam em sítios alostéricos preditos da cruzaína. A comunidade 2 é composta por 6 resíduos, dentre eles resíduos da díade catalítica, Cys25 e His162, e resíduos Gly23 e Gly65, que coevoluem por questões estruturais. As comunidades 3, 4 e 5 possuem apenas 2 resíduos cada. O resíduo Cys155 da comunidade 3 possui importância estrutural, bem como os resíduos Val13 e Lys181 (comunidade 4) e Gln51 (comunidade 5). Os resíduos da comunidade 5, Leu48 e Gln51 possuem importância funcional e o resíduo Ser55 (comunidade 3) ainda não apresenta sua função descrita na literatura. Também foram encontrados 10 resíduos que são altamente conservados evolutivamente nessa família. Enquanto alguns resíduos, incluindo os que compõem a díade catalítica (Cys25 e His162), são conservados por questões funcionais, outros o são por questões estruturais (Ser49, Gly168, Gly189 e Gly192).

Palavras chave: cisteíno proteases, cruzaína, alinhamento múltiplo de sequências, correlação, coevolução.

Abstract

Proteins evolve over time, due to the accumulation of mutations, insertions and deletions in the genes that encode them. Highly conserved positions generally refer to residues that are strictly necessary for the structure of the protein or for its function. The DRCN (Decomposition of Residue Coevolution Networks) methodology allows the study of coevolution of residues in protein families. To do this, it consists of five basic steps: filtering the alignment, calculating correlations, decomposing the network, generating auxiliary visualization files and annotating positions by automatic search in UniProt. Cysteine proteases are a large and diverse group of enzymes belonging to the CA clan and the C1 family (papain family) and can be found in all animals and plant kingdoms as well as in many viruses and prokaryotes. The present work aims to analyze and discuss the importance of correlated and conserved residues of the cysteine protease family from the multiple sequence alignment. For this, we use the PFSTATS software, which employs the DRCN methodology. There were 34 correlated residues found in 5 coevoluted communities. Community 1 is the largest community, consisting of 22 residues coevolved by structural issues, such as Cys22 and Cys63 and Cys56 and Cys101, which carry disulfide bridges; as well as for functional issues, such as Pro2, Phe28, Gly35 and Tyr89, which are located in predicted allosteric sites of cruzain. Community 2 is composed of 6 residues, among them residues of the catalytic dyad, Cys25 and His162, and residues Gly23 and Gly65, which coevolve for structural reasons. Communities 3, 4 and 5 have only 2 residues each. The Cys155 residue of community 3 has structural importance, as well as residues Val13 and Lys181 (community 4) and Gln51 (community 5). The residues of community 5, Leu48 and Gln51 have functional importance and the residue Ser55 (community 3) still does not present its function described in the literature. Also 10 residues were found that are highly conserved evolutionarily in this family. While some residues, including those that make up the catalytic dyad (Cys25 and His162), are conserved for functional reasons, others are for structural reasons (Ser49, Gly168, Gly189 and Gly192).

Key words: cysteine proteases, cruzain, multiple sequence alignment, correlation, coevolution.

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

BLOSUM: Blocks of Amino Acid Substitution Matrix

DNA: Ácido Desoxirribonucleico

DRCN: Decomposition of Residue Coevolution Networks

IFN-γ: Interferon gama

IL-1β: Interleucina 1 beta

IL-6: Interleucina 6

LDL: Lipoproteína de Baixa Densidade

mRNA: Ácido Ribonucleico mensageiro

MSA: Multiple Sub Alignment

NO: Óxido Nítrico

PDB: Protein Data Bank

PFSTATS: Protein Families Statistics

Pfam: Protein families

SNC: Sistema Nervoso Central

TNF-α: Fator de Necrose Tumoral alfa

UniProt: Universal Protein

Índice

1. Introdução

1.1. Evolução das Proteínas

A evolução é um processo no qual a seleção para manutenção de determinados traços em detrimento de outros ocorre a partir de uma pressão seletiva do ambiente (EIJSINK, V. G. H. et al, 2005). As proteínas evoluem ao longo do tempo, devido ao acúmulo de mutações, inserções e deleções nos genes que as codificam. Também pode ocorrer duplicação gênica, que por sua vez pode ser seguida de mutação e seleção resultando em divergência. Assim, esses mecanismos evolutivos podem combinar-se culminando em uma proteína com diferenças em relação à proteína inicial (KINCH, L. N., GRISHIN, N. V., 2002).Para que não haja perda da função proteica, essas mutações devem ocorrer em resíduos que não são críticos, direta ou indiretamente, para a sua atividade (GÖBEL, U. et al, 1994). No caso de enzimas, mutações nos resíduos que constituem o sítio catalítico ocasionam inatividade da proteína. Mutações em resíduos que não fazem parte do sítio catalítico, mas que ocasionem alterações estruturais, resultantes de interações entre resíduos que outrora não existiam ou que passarão a existir também podem ocasionar inatividade da proteína uma vez que podem alterar a sua estrutura.

Portanto, dentro de rol de mutações que podem ocorrer durante o processo evolutivo de uma proteína, uma pequena quantidade é permitida e não ocasiona perda da função proteica. Outras mutações podem gerar perda total ou parcial da sua atividade. Posições fortemente conservadas geralmente referem-se a resíduos que são estritamente necessários para a estrutura da proteína ou para sua função. Sabe-se que a estrutura primária de uma proteína determina a forma como ela se enovelará e, consequentemente, a sua função. Contudo, há proteínas que possuem sequências muito semelhantes e, no entanto, apresentam conformações distintas.Tais casos demonstram que a estrutura terciária também pode evoluir e sofrer alterações (KINCH, L. N., GRISHIN, N. V., 2002).

Embora tenha sido sugerido que os resíduos em uma proteína interagem entre si e influenciam como cada resíduo dessa proteína vai evoluir, por insuficiência de dados que pudessem comprovar tal teoria, tradicionalmente pensava-se que os resíduos evoluíam independentemente uns dos outros. No entanto, com o avanço da

bioinformática bem como com a crescente disponibilidade de dados genômicos, tornou-se possível analisar a fundo o processo de coevolução das proteínas (LITTLE, D. Y., CHEN, L., 2009).

A hipótese de covariação levantada por Fitch e Markowitz postula que durante o processo evolutivo de uma proteína, a qualquer momento, apenas uma pequena fração de resíduos apresentam um grau de liberdade para variar. Haveria dentro da proteína dois grupos de resíduos: um grupo no qual os resíduos poderiam sofrer mutações sem causar danos à proteína tanto em questões estruturais quanto em questões funcionais, e um outro grupo no qual os resíduos não poderiam sofrer mutação sob pena de causarem danos à proteína. Quando um resíduo X que integra o grupo de resíduos que podem sofrer mutações sofre uma mutação, automaticamente ele pode influenciar outro resíduo, Y, deste mesmo grupo. O resíduo Y pode deixar de integrar o grupo de resíduos que podem sofrer mutações, migrando para o grupo de resíduos que não podem mais ser alterados. Ou seja, a mutação de um único resíduo possui influência na possibilidade, ou não, de mutação de outros resíduos da proteína.

Tal conceito de coevolução é mais amplo e assim, a hipótese de covariação pode ser reafirmada da seguinte maneira: a qualquer momento durante o processo evolutivo de uma proteína, apenas uma pequena fração de mutações é admissível e quando há mutação em uma determinada posição as forças seletivas associadas a outros pontos passíveis de mutação podem ser alteradas, culminando numa alteração do conjunto de mutações admissíveis para essa proteína. Portanto, considerando-se um par de resíduos, a variabilidade de um deles é dependente do estado no qual se encontra o outro resíduo (LITTLE, D. Y., CHEN, L., 2009).

As sequências das proteínas que se relacionam evolutivamente são selecionadas para preservar uma função e, portanto, devem ser parcialmente conservadas. Nesse contexto, há resíduos que se correlacionam e se acoplam pela evolução (coevoluem). E os pares de resíduos em suas respectivas posições ao longo de uma sequência proteica podem apresentar fortes correlações decorrentes tanto de questões funcionais quanto de questões estruturais (SUTTO, L., et al, 2015).Portanto, resíduos conservados em uma proteína geralmente possuem

importância funcional e/ou estrutural, sendo críticos, direta ou indiretamente, para o funcionamento adequado da proteína.

1.2. Alinhamentos Múltiplos de Sequências de Proteínas

Para estudar as sequências proteicas e analisar a relação evolutiva entre elas faz-se necessário o uso de ferramentas. Essas ferramentas possibilitam explorar as sequências proteicas de forma detalhada, analisar os resíduos que são conservados, os que coevoluem bem como os resíduos que sofreram mutações em diferentes proteínas. A partir dessas informações pode-se avaliar o impacto dessas mutações bem como a importância da conservação de determinados resíduos para a função biológica da proteína.

Além disso, o sequenciamento do DNA e o consequente avanço e aperfeiçoamento das técnicas utilizadas para tal ocasionou o aumento substancial do número de genomas disponíveis nos bancos de dados. Houve também a necessidade de aumento da capacidade computacional para processar e armazenar adequadamente esses dados. Assim, a criação e aprimoramento de algoritmos se fizeram extremamente necessários e as técnicas de alinhamento de sequências despontaram como ferramentas essenciais para a análise das moléculas.

O alinhamento de sequências de proteínas é uma técnica de comparação entre duas ou mais sequências, realizada por um algoritmo, e tem como objetivo buscar caracteres individuais que se encontram em posições equivalentes nas sequências analisadas. Cada resíduo é considerado um caracter e as sequências são organizadas em linhas e os caracteres em colunas e a partir daí os algoritmos buscarão a melhor correspondência para as sequências sob análise. Para comparar sequências de interesse frequentemente determina-se a identidade entre elas, que quantifica os caracteres idênticos entre as sequências em valores de porcentagem. Importante destacar que a identidade diz respeito a caracteres que se repetem (ou não) nas mesmas posições em diferentes sequências enquanto a homologia refere-se à ancestralidade comum para diferentes sequências (BENTON, D., 1996; JUNQUEIRA, M. D., BRAUN, R. L., VERLI, H., 2014).

Em um alinhamento de sequências algumas vezes torna-se necessária a adição de lacunas ou *gaps*, que são representadas por “-“ e se caracterizam como

um ou mais eventos de inserção ou deleção. Tais eventos, denominados “indels” (*in* significa inserção e *del* significa deleção) decorrem de processos mutagênicos e dificultam o alinhamento, consequentemente tornando as interpretações mais complexas. Em casos de análises evolutivas e filogenéticas, por exemplo, há eliminação das sequências que possuem um número elevado de *gaps* e *indels*, pois há um nível maior de incerteza quanto ao alinhamento das sequências (DO, C. B., KATOH, K., 2008; JUNQUEIRA, M. D., BRAUN, R. L., VERLI, H., 2014).

Para avaliar a significância de um alinhamento múltiplo de sequências utilizam-se matrizes como a BLOSUM (*Blocks of Amino Acid Substitution Matrix*), que é uma matriz de substituição que pontua alinhamentos entre sequências de proteínas divergentes. A matriz BLOSUM baseia-se em análises envolvendo regiões conservadas de famílias de proteínas, que não possuem lacunas no alinhamento, e no cálculo da frequência relativa de aminoácidos bem como as probabilidades de substituição. Assim, são calculadas pontuações para as substituições que possam ocorrer a partir dos 20 aminoácidos. Como em um alinhamento são comparadas sequências, procura-se verificar se em uma determinada posição os resíduos são os mesmos em diferentes sequências. E a comparação dos diferentes possíveis alinhamentos para um conjunto de sequências é realizada a partir de uma pontuação dada a cada substituição possível tendo como base o alinhamento das proteínas relacionadas. Assim, os *matches* são as correspondências de caracteres observadas em uma mesma coluna de diferentes sequências e são pontuados positivamente enquanto os *mismatches* são as desigualdades de caracteres em uma mesma coluna e são pontuados negativamente (JUNQUEIRA, D. M., BRAUN, R. L., VERLI, H., 2014). Pontuações positivas também são atribuídas a substituições mais prováveis de ocorrer enquanto pontuações negativas são dadas às substituições menos prováveis.

Os alinhamentos de sequências de proteínas geralmente são gerados a partir de um número muito grande de sequências e pode ser denominado como MSA (*Multiple Sequence Alignment*). Ao se fazer o alinhamento múltiplo de sequências de proteínas homólogas, por exemplo, tem-se que resíduos em posições equivalentes podem apresentar a mesma função, e na medida em que se conhece a função de alguns resíduos de uma determinada proteína de uma família pode-se inferir a função dos resíduos equivalentes de outras proteínas do alinhamento.

Existem diversos bancos de dados que utilizam algoritmos para a classificação das sequências de proteínas, tanto por similaridade estrutural, quanto por similaridade de sequência. Um exemplo de banco de dados relevante e muito utilizado é o Pfam (*Protein families*), que agrupa conjuntos de sequências protéicas em famílias a partir do alto grau de identidade entre elas. Cada família de proteínas presente no Pfam possui um alinhamento gerado previamente e a organização dessas proteínas em famílias gera uma maior organização para o banco de dados. Nos últimos dois anos houve uma reorganização substancial no Pfam e uma mudança importante é que ele passou a ser baseado principalmente na referência de proteomas UniProt (*Universal Protein*) (FINN, R. D. et al, 2016).

Portanto, o avanço cada vez mais significativo das tecnologias referentes ao estudo das proteínas, desde técnicas cristalográficas até ferramentas de bioinformática, tem permitido uma maior exploração e entendimento dessas macromoléculas tão cruciais para a vida. Há diversos métodos para o estudo das famílias de sequências proteicas. O fato de que a proteína deve ter sua estrutura mantida de forma a preservar sua função restringe, de certa forma, a evolução da sequência proteica. Isso pode ser utilizado para a interpretação de mutações correlacionadas que são observadas em uma família de proteínas. Um método que foi desenvolvido para tal consiste na análise de correlações entre os diferentes resíduos que sofrem mutação em um alinhamento múltiplo de sequência. Tais correlações podem ser utilizadas para previsão de mapas de contato entre as proteínas e esses mapas podem auxiliar no cálculo da estrutura terciária dessas proteínas. O grau de correlação entre dois resíduos relaciona-se à força de contato entre esses resíduos e é mensurado a partir de um coeficiente de correlação mutacional. Os resíduos são alocados em *clusters* onde um resíduo deve estar correlacionado com pelo menos algum outro resíduo desse *cluster*. Os resíduos podem estar alocados em mais de um *cluster* e esses *clusters* podem se sobrepor (GÖBEL, U. et al, 1994). No entanto, tal método de estudo das proteínas apresenta limitação de utilização devido ao fato de não apresentar elevada fidedignidade, ocasionando falsos positivos.

Outra técnica que foi desenvolvida para o estudo das sequências proteicas consiste no mapeamento das interações energéticas que ocorre entre os resíduos de uma proteína a partir dos dados evolutivos de uma família. Muitas proteínas

podem propagar de forma eficiente a energia através de sua estrutura terciária, o que também pode ser a base para propriedades biológicas como alosterismo e transmissão de sinais. Estudos mutacionais podem confirmar essas previsões, corroborando que a análise estatística da energia entre os resíduos é um bom indicador do acoplamento termodinâmico entre os resíduos de uma proteína (LOCKLESS, S. W., RANGANATHAN, R., 1999). Contudo observou-se que essa análise energética entre os resíduos não apresenta resultados tão satisfatórios no que se refere ao estudo das proteínas.

Uma metodologia muito interessante e com excelentes resultados para o estudo de famílias de proteínas foi proposta por Bleicher, Lemke & Garrat (2011) e é denominada DRCN (*Decomposition of Residue Coevolution Networks*). Ela consiste em cinco etapas básicas que são: 1. filtragem de um alinhamento múltiplo de sequências que represente a família a ser estudada, 2. cálculo de correlações para os pares de resíduos e construção de rede de correlações, 3. decomposição da rede de correlações gerada em grupos de resíduos que coevoluem, 4. geração de arquivos de visualização auxiliares que facilitam o entendimento dos resultados e 5. anotação de posições de resíduos por busca automática no UniProt com a relação de todas as referências existentes para os resíduos das proteínas da família. Tal metodologia também possibilita o estudo das famílias proteicas com profundidade já que são gerados dados como: comunidades de resíduos que coevoluem, resíduos que são conservados e resíduos que se anticorrelacionam.

1.3. Cisteíno Proteases

A partir das ferramentas citadas anteriormente, alinhamento múltiplo de sequências,análise de conservação/coevolução e análise estrutural, pode-se realizar um estudo mais completo das sequências protéicas que constituem uma família. Torna-se possível a caracterização de famílias proteicas, buscando averiguar quais resíduos são conservados e quais se correlacionam pela evolução, e a partir daí buscar o entendimento da importância desses resíduos para a função biológica dessas proteínas.

As proteases ou peptidases são enzimas proteolíticas que catalisam a hidrólise de ligações peptídicas de outras proteínas (STOKA, TURK & TURK, 2005). Sabe-se que muitas proteínas passam por modificações pós-traducionais reversíveis, como por exemplo, fosforilação, no entanto a proteólise é um processo irreversível. Portanto, quando uma proteína é hidrolisada em alguma de suas ligações peptídicas, é necessária a tradução de uma nova proteína a partir do respectivo mRNA (CHAPMAN, RIESE & SHI, 1997). A proteólise é um mecanismo empregado por praticamente todas as células e é importante na regulação da função e do destino metabólico de suas proteínas. Nesse sentido as enzimas proteolíticas estão envolvidas na regulação de processos cruciais e irreversíveis como coagulação, digestão, maturação de citocinas e hormônios, apoptose e clivagem de proteínas intracelulares (CHAPMAN, RIESE & SHI, 1997). Além disso, a proteólise possui papel importante na renovação das proteínas. Assim, proteínas que já desempenharam suas funções devem ser degradadas para que os resíduos que as constituem sejam reaproveitados em proteínas que serão formadas (BERG, J. M; TYMOCZKO, J. L.; STRYER, L, 2010).

Cisteíno proteases são um grande e diverso grupo de enzimas que foram classificadas por Rawlings e Barrett em clãs e famílias (CYGLER, MORT, 1997). Segundo a base de dados de peptidases MEROPS (https://www.ebi.ac.uk/merops/), o clã contém todas as peptidases que surgiram evolutivamente a partir de uma única origem, o que engloba uma ou mais famílias que apresentam relação evolutiva através da similaridade de suas estruturas terciárias. Quando a estrutura terciária dessas peptidases não está disponível pode-se realizar a análise dos resíduos que compõem o seu sítio ativo ou dos motivos em torno dos resíduos do sítio ativo para alocá-las no clã adequado.

Cada clã é identificado por duas letras, sendo que a primeira representa o tipo de resíduo catalítico das famílias que estão contidas neste clã. A letra “P” é utilizada quando um clã contém famílias distintas que, por sua vez, apresentam tipos de resíduos catalíticos também distintos, como por exemplo, no caso da Serina, Treonina e Cisteína. Muitos clãs são subdivididos em subclãs porque há evidências de que houve uma divergência evolutiva antiga dentro do clã. Portanto, há peptidases dos seguintes clãs: Aspartato, Cisteína, Glutamina, Metalo, Asparagina,

Serina, Treonina, peptidases pertencentes a um clã não determinado e peptidases pertencentes a um clã com famílias variáveis.

Ainda segundo MEROPS, família é um conjunto de enzimas proteolíticas homólogas. A homologia é evidenciada por uma significativa similaridade na sequência de aminoácidos de uma enzima em relação a uma outra enzima da família. Cada família é identificada por uma letra que representa o tipo de resíduo no sítio ativo (catalítico) da enzima junto com um único número. Algumas famílias são divididas em subfamílias porque há evidências de que houve uma divergência evolutiva antiga dentro da família. Assim sendo, há peptidases das seguintes famílias com as suas respectivas representações: Aspartato (A), Cisteína (C), Glutamina (G), Metalo (M), Asparagina (N), Serina (S), Treonina (T), peptidases que apresentam resíduo catalítico desconhecido (U) e peptidases que apresentam resíduo catalítico variável (P).

Enquanto a família é considerada um grupo de enzimas onde qualquer das enzimas que o compõem deve apresentar uma relação de evolução com pelo menos alguma outra enzima integrante dessa família,o clã compreende um grupo de famílias para as quais há uma relação evolutiva apesar da falta de similaridade de sequências estatisticamente significativa (CYGLER, MORT, 1997).

As cisteíno proteases pertencem ao clã CA e à família C1 (família da papaína) e podem ser encontradas em todos os animais e reinos de plantas bem como em muitos vírus e organismos procariotas (WIEDERANDERS, KAULMANN & SCHILLING, 2003). Conforme dados da MEROPS há na família C1 15702 sequências catalogadas. Exemplos muito conhecidos e bem estudados de cisteíno proteases incluem a cruzaína, enzima principal do protozoário *Trypanosoma cruzi*, e a papaína, enzima presente no mamão, *Carica papaya*. As cisteíno proteases podem ser classificadas como exopeptidases e endopeptidases (RZYCHON, M., CHMIEL, D., STEC-NIEMCZYK, J., 2004).

Nos últimos anos as peptidases semelhantes à papaína e presentes nos animais passaram a ser denominadas como catepsinas para distingui-las das peptidases encontradas em outras classes de organismos (NOVINEC, M., LENARCIC, B., 2013). Peptidases semelhantes à papaína são proteínas monoméricas globulares com peso molecular médio de 25 a 35kDa (NOVINEC, M.,

LENARCIC, B., 2013), sendo que a catepsina humana C é uma exceção por ser tetramérica (TURK, V., TURK, B., TURK, D., 2001). Estas proteínas possuem uma estrutura composta por dois domínios, tendo o sítio ativo localizado em uma fenda entre eles. O domínio L é composto por α-hélices e o domínio R é composto por folhas β antiparalelas (TURK, V., TURK, B., TURK, D., 2001). A figura 1, representada abaixo, ilustra a estrutura terciária da cruzaína, uma importante cisteíno protease. Pode-se verificar a extremidade N-terminal (domínio L) constituído por α-hélices,bem como a extremidade C-terminal, constituída por folhas β antiparalelas.

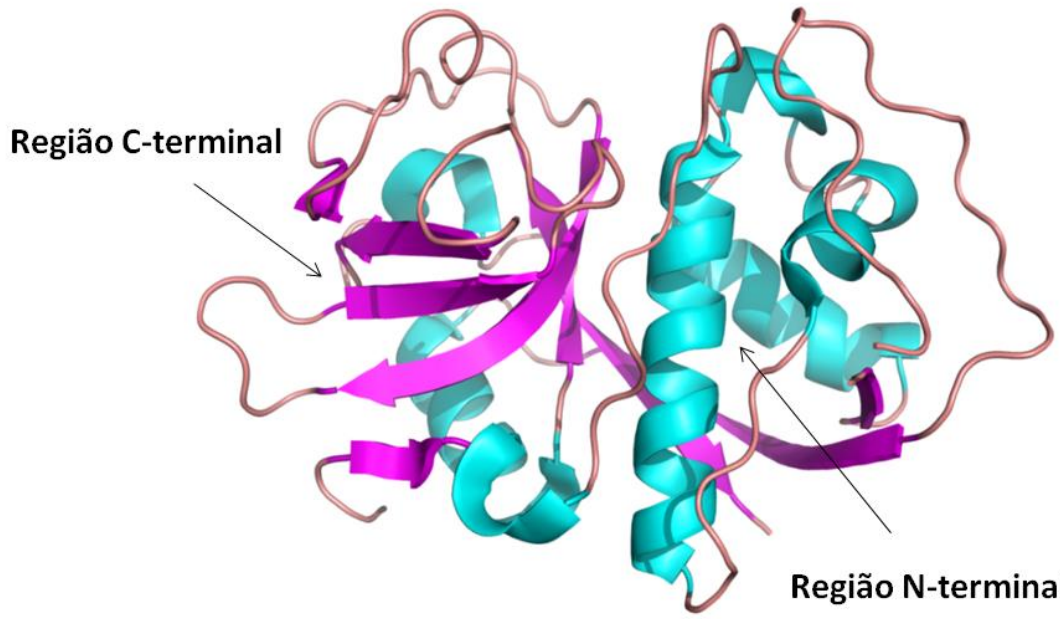

Figura 1. Estrutura da cruzaína com destaque das estruturas secundárias. Destacada a estrutura da cruzaína(PDB: 3KKU) com a região N-terminal constituída majoritariamente por α-hélices, região C-terminal constituída por folhasβ antiparalelas. α-hélices em ciano, folhas β em magenta, alças em rosa claro. Figura preparada com o programa Pymol (Schrödinger, LCC, 2015).

A atividade das cisteíno proteases pode ser regulada por diversas formas e tal regulação constitui-se em um balanço entre a quantidade de enzimas ativas presentes e a quantidade de inibidores ativos presente. Uma desregulação, seja de enzimas ativas, seja de inibidores ativos, pode ocasionar consequências deletérias para o organismo. Além da regulação que advém da expressão gênica, há outros fatores que atuam no controle da atividade das cisteíno proteases. São eles:

1. pH. Muitas cisteíno proteases apresentam instabilidade e/ou pouca atividade em um ambiente neutro ao passo que suas funções são otimizadas em ambiente ácido, encontrado no interior de vesículas intracelulares.
2. Potencial redox. O sítio ativo das cisteíno proteases é um ambiente oxidativo, portanto tais enzimas são mais ativas em um ambiente com

maior potencial redox. Os endossomos que acumulam em seu interior as cisteíno proteases conseguem manter tal ambiente.

3. Síntese das enzimas em precursores inativos (zimógenos). Todas as cisteíno proteases requerem ativação proteolítica e tal ativação requere um ambiente com pH ácido, de forma a prevenir uma ativação indiscriminada dessas enzimas.
4. Direcionamento das enzimas para endossomos e lisossomos. As cisteíno proteases possuem sítios de N-glicosilação e quando esses sítios estão glicosilados eles se ligam a receptores de manose-6-fosfato, que são os principais receptores dos lisossomos que auxiliam no direcionamento das enzimas.
5. Presença de inibidores das cisteíno proteases. Todos os fatores citados anteriormente contribuem para a compartimentalização da atividade das cisteíno proteases. Além disso, os inibidores atuam para controlar a atividade enzimática (CHAPMAN, H. A., RIESE, R. J., SHI, G. P., 1997).

Portanto, as formas mais importantes de regulação da atividade das cisteíno proteases são a ativação dos zimógenos e a inibição por inibidores proteicos endógenos. Similarmente a outras proteases, as catepsinas são sintetizadas como precursores inativos e, portanto, devem ser ativadas por remoção proteolítica do pró-peptídeo N-terminal (TURK, V., TURK, B., TURK, D., 2001).

O sítio ativo das cisteíno proteases é localizado na interface entre ambos os domínios, no topo da molécula, em uma fenda em forma de “V”. O mecanismo catalítico das cisteíno proteases é muito similar ao mecanismo catalítico das serino proteases, uma classe muito conhecida de peptidases. Enquanto as serino proteases possuem uma tríade catalítica formada pelos resíduos de Ser, His e Asp, as cisteíno proteases possuem uma díade catalítica, formada pelos resíduos de Cys25 e His162. No entanto alguns outros resíduos como Asn175 e Gln19 (numeração da papaína) parecem ser necessários para estabilização do adequado posicionamento dos resíduos da díade catalítica (NOVINEC, M., LENARCIC, B., 2012). A figura 2, mostrada abaixo, representa a estrutura da cruzaína destacando os aminoácidos que compõe a díade catalítica, bem como os que possuem provável envolvimento no processo de catálise.

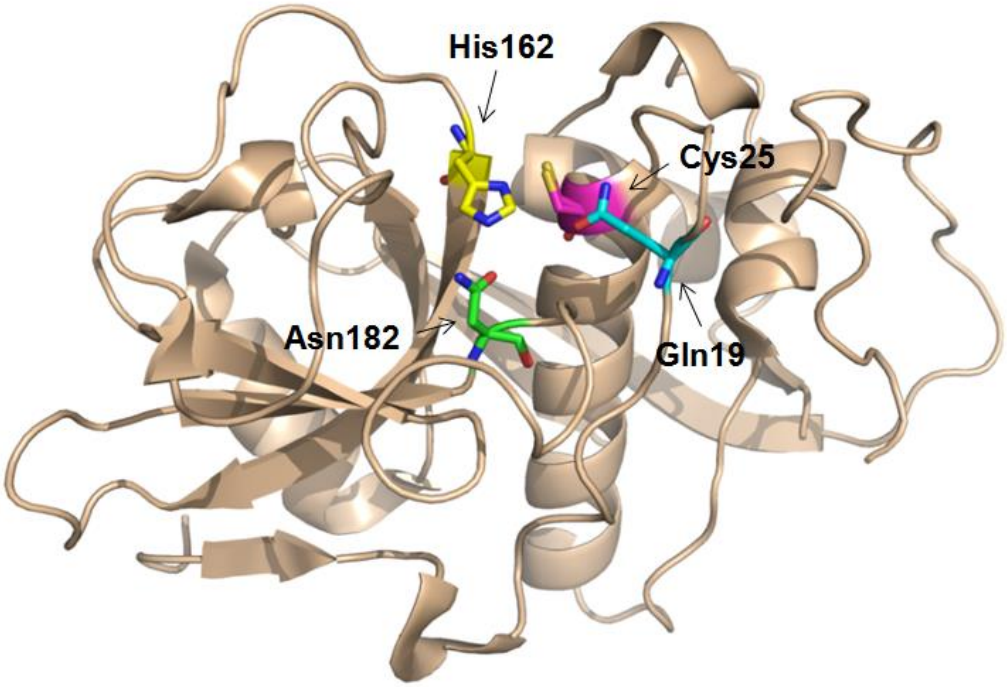

Figura 2. Modelo estrutural da cruzaína com destaque dos aminoácidos que compõe a díade catalítica e possuem envolvimento no processo de catálise. Destacada a estrutura da cruzaína (PDB: 3KKU) com α-hélices, folhas β e alças em rose claro. Nitrogênio em azul escuro, oxigênio em vermelho e enxofre em amarelo. Figura preparada com o programa Pymol (Schrödinger, LCC, 2015).

Assim como ocorre no processo catalítico das serino proteases, a catálise das cisteíno proteases envolve a acilação e a desacilação da enzima, a entrada de uma molécula de água e a formação de dois intermediários tetraédricos. Uma etapa crucial do processo de catálise das cisteíno proteases é a formação de um par iônico reativo composto pelos resíduos Cys25 e His162, o que envolve a transferência de prótons entre eles.A figura 3, ilustrada abaixo, mostra o mecanismo catalítico da cruzaína, tomada como exemplo para as demais proteínas da família.

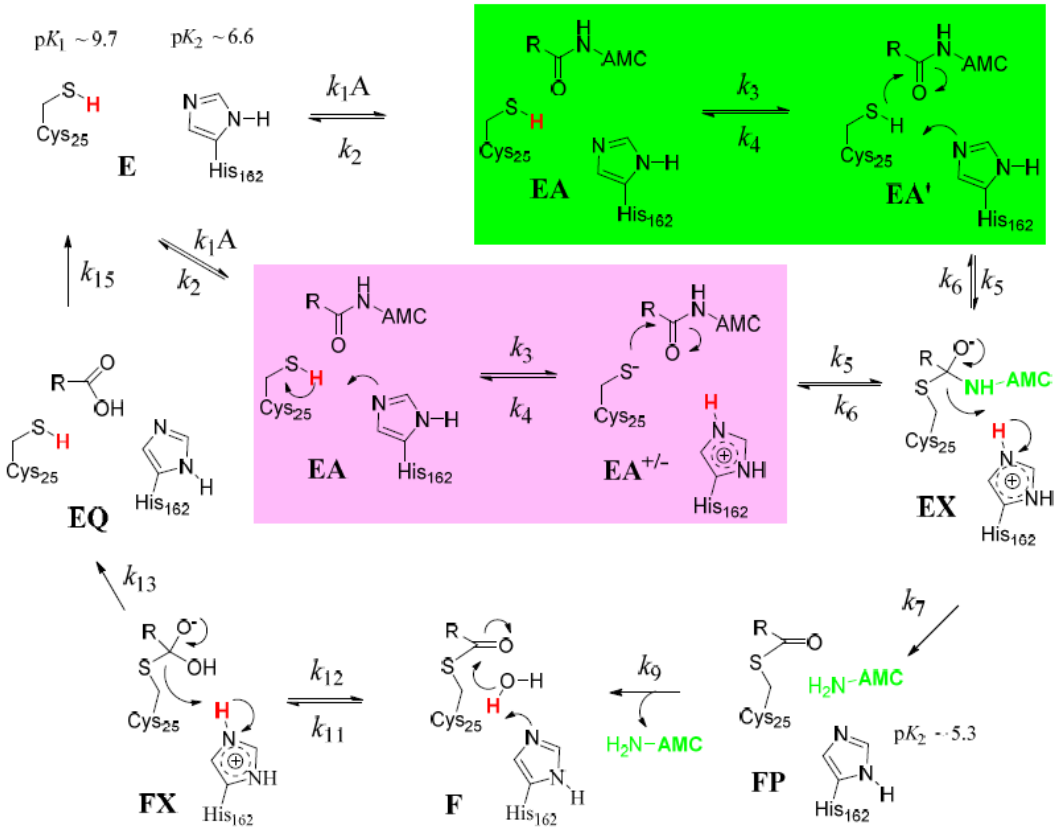

Figura 3. Mecanismo catalítico da cruzaína. Pode-se observar a representação das oito etapas que compõem o mecanismo catalítico da cruzaína, como exemplo do que ocorre para as demais cisteíno proteases. Há dois possíveis mecanismos para a formação do primeiro intermediário tetraédrico. O primeiro mecanismo está representado no quadro verde enquanto o segundo está representado no quadro rosa. A catálise das cisteíno proteases ocorre mediante a acilação e a desacilação da enzima, além de necessitar da entrada de uma molécula de água. Também são formados dois intermediários tetraédricos. Uma etapa crucial do processo de catálise é a formação de um par iônico tiolato-imidazólico reativo composto pelos resíduos Cys25 e His162, o que envolve a transferência de prótons entre eles. E, EA, EA', $EA^{+/-}$, EX, FP, F, FX, EQ representam as etapas do processo de catálise. Figura extraída de: ZHAI, X., MEEK, T. D., (2018).

Na primeira etapa do mecanismo catalítico das cisteíno proteases (E) verificam-se os resíduos de Cys25 e His162 em sua forma neutra. A partir daí dois mecanismos distintos podem ser observados. No primeiro mecanismo (quadro verde), ocorre a ligação da enzima ao seu substrato, caracterizando a segunda etapa (EA). Em seguida ocorre a transferência de prótons de Cys25 para His162 de forma combinada com o ataque do tiolato recém-formado à carbonila da amida do substrato (EA'), culminando na formação do primeiro intermediário tetraédrico

(EX).No segundo mecanismo (quadro rosa) o ataque da carbonila pelo tiolato acontece em duas etapas. Também ocorre a ligação da enzima ao seu substrato (EA) e em seguida ($EA^{+/-}$) ocorre a transferência de prótons de Cys25 para His162, com consequente formação do par imidazólico-tiolato. Subsequentemente há o ataque do tiolato à carbonila da amida, levando à formação do primeiro intermediário tetraédrico (EX). A reação de acilação é completada após o colapso do intermediário tetraédrico com transferência de prótons do anel imidazólico para o nitrogênio da ligação peptídica que será hidrolisada. Ocorre a clivagem da ligação peptídica e observa-se a ligação da porção carbonílica do substrato à Cys25, formando a acil-enzima (FP). A reação de desacilação é a mesma para os dois mecanismos. Há a desprotonação de uma molécula de água (F) pela His162 neutra e a formação do segundo intermediário tetraédrico (FX). A seguir o intermediário tetraédrico sofre colapso e consequentemente há formação do produto carboxilato e restauração da Cys25 e His162 em suas formas neutras (EQ) (ZHAI, X., MEEK, T. D., 2018).

Com relação às interações entre enzima e substrato, Schechtner e Berger propuseram que as cisteíno proteases podem interagir com sete resíduos do substrato e os locais de interação tanto da enzima com o substrato quanto do substrato com a enzima são designados pelas letras “S” e “P”, respectivamente. Nas cisteíno proteases quatro desses sítios de interação estão na porção N-terminal da ligação peptídica a ser clivada (S4 a S1) enquanto os outros três sítios de interação estão na porção C-terminal da ligação peptídica a ser clivada (S1’ a S3’). Os sítios de interação correspondentes no substrato também são em número de sete e são nomeados de P4 a P3’ (NOVINEC, M., LENARCIC, B., 2013).

A figura 4, representada abaixo, ilustra os sítios de ligação da enzima e do substrato a partir da nomenclatura proposta por Schechtner e Berger.

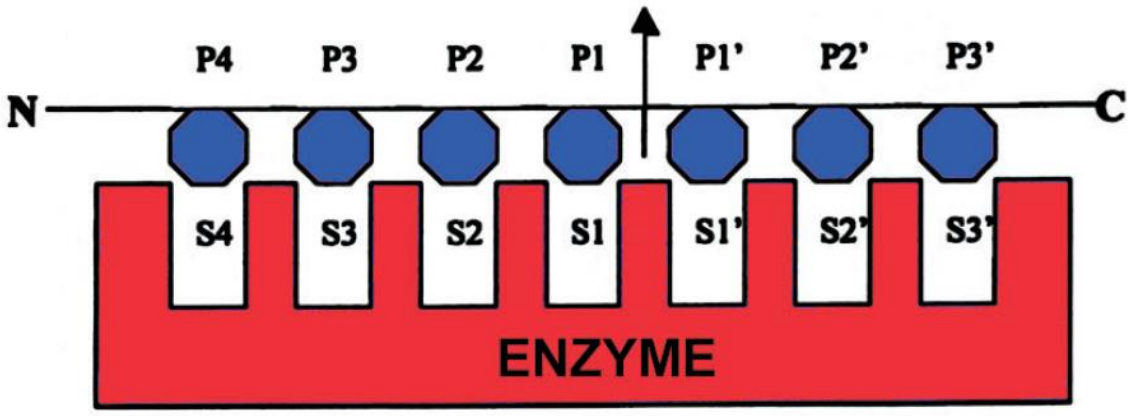

Figura 4. Representação dos sítios de interações das cisteíno proteases e seus substratos. A enzima, representada em vermelho possui sete sítios de interação com o substrato. A seta vertical representa a ligação peptídica a ser clivada e na porção N-terminal enzimática encontram-se quatro sítios de interação com o substrato (S4 a S1) enquanto na porção C-terminal encontram-se os três sítios de interação com o substrato (S1' a S3'). Em azul estão representados os sete sítios de interação do substrato com a enzima, P4 a P1 na porção N-terminal e P1' a P3' na porção C-terminal. Os sete sítios do substrato apresentam correspondência aos sítios enzimáticos S4 a S3'. Figura extraída de: RZYCHON, M., CHMIEL, D., STEC-NIEMCZYK, J., (2004).

O alosterismo pode ser definido como uma modulação da função proteica que ocorre secundariamente a uma interação entre uma molécula, denominada modulador alostérico, e um outro sítio da proteína, denominado sítio alostérico. O sítio alostérico é distinto do sítio catalítico e o alosterismo não é uma propriedade exclusiva das proteínas, podendo ser descrito como uma característica inerente de muitas macromoléculas. Nesse sentido, o alosterismo configura-se como outra forma de controlar a população de proteínas ativas e inativas, uma vez que a ligação de um modulador alostérico no respectivo sítio alostérico pode levar a mudanças conformacionais na estrutura proteica, e estas podem causar sua ativação ou inativação (ALVAREZ, L. H. et al, 2019). Portanto, moduladores alostéricos podem causar o aumento ou a diminuição da atividade de uma determinada macromolécula.Com relação aos sítios alostéricos das enzimas da família das cisteíno proteases ainda não existem estudos experimentais que tenham demonstrado a existência desses sítios para a cruzaína. Estudos computacionais realizados tomando como base a catepsina humana K foram capazes de predizer sítios alostéricos potenciais, que servem para regulação da atividade enzimática.

Assim, foram preditos 7 sítios alostéricos distintos para a catepsina K (NOVINEC, M. et al, 2014). Estudos mais recentes evidenciaram que a atividade da catepsina K é modulada por glicosaminoglicanos que se ligam à enzima em diversos locais. Tais glicosaminoglicanos podem produzir efeitos ativadores e inibidores da catepsina K e inclusive podem conduzir a enzima para a formação de um dímero que apresenta atividade de hidrólise do colágeno aumentada. A condroitina-4-sulfato, um glicosaminoglicano, tem sido implicada no papel de acelerar a atividade biológica da catepsina K. Um trabalho realizado com a catepsina K mostrou que efetores se ligam a um sítio alostérico na superfície da proteína e que os seguintes resíduos são importantes para a ligação entre o efetor e a proteína: Ala1, Asp3, Lys119, Lys122, Arg123, Tyr169, Lys176, Arg198 e Asn199. Tais resíduos são importantes ao realizarem ligações de hidrogênio que auxiliam na ligação do efetor com a enzima (NOVINEC, M., REBERNIK, M., LENARCIC, B., 2016).

Também foi demonstrado efeito da condroitina-4-sulfato na modulação alostérica da catepsina S. Foram realizados experimentos com vários glicosaminoglicanos, no entanto, apenas a condroitina-4-sulfato possui um padrão molecular capaz de reduzir significativamente o nível de hidrólise de colágeno tipo IV pela catepsina S. A condroitina-4-sulfato também retardou a maturação da pró-catepsina S em pH 4 de maneira dose-dependente por interferir na via de processamento molecular. Ela também induziu mudanças conformacionais sutis na catepsina S madura. Foram identificados na catepsina S 3 prováveis sítios específicos para ligação da condroitina-4-sulfato (SAGE, J. et al, 2013).

Com relação à catepsina B, foi verificado um efeito protetor da heparina de forma a prevenir a inativação da catepsina B decorrente de alterações de pH. Em um ambiente ácido a estrutura da catepsina B apresenta-se estável, no entanto em ambiente alcalino pode-se observar pequenas flutuações na estrutura da enzima. A heparina parece estabilizar essas flutuações de forma global, o que inclui a estabilização de regiões como o loop de oclusão e o sítio ativo. Além disso, em ambiente alcalino pode-se observar uma maior separação entre os domínios R e L da enzima, o que pode influenciar sobremaneira na sua capacidade catalítica, uma vez que o sítio ativo localiza-se em uma fenda entre os dois domínios. Na presença da heparina essa separação aumentada não ocorre, o que mostra um efeito protetor

da heparina sobre a estabilidade da catepsina B em ambientes de pH alcalino (COSTA, M. GS. et al, 2010).

Com relação à cruzaína, um estudo computacional realizou a predição de 7 cavidades na superfície desta proteica, consideradas como potenciais sítios alostéricos, predizendo também os resíduos que integram cada sítio. O sítio 1 é composto por 18 resíduos, os sítios 2 e 4 são compostos por 10 resíduos cada, os sítios 3 e 7 são compostos por 16 resíduos cada, o sítio 5 é composto por 17 resíduos e o 6 por 19 resíduos (ALVAREZ, L. H., 2017). Contudo, dados recentes advindos de outro estudo computacional destacaram que, dentre esses 7possíveis sítios alostéricos preditos, os sítios 1 e 3 parecem ser os mais prováveis para a cruzaína (ALVAREZ, L. H. et al, 2019). A tabela 1 e as figuras5a 11 representados abaixo mostram os resíduos que compõe os 7 prováveis sítios alostéricos da cruzaína.

Sítios alostéricos preditos para a cruzaína	
Nome do sítio	**Resíduos que o compõem**
Primeiro	Met52, Ser55, Cys56, Asp57, Lys58, Asp60, Gly62, Trp74, Glu78, Asn79, Asn80, Gly81, Ala82, Tyr84, Tyr89, Cys101, Thr102, Thr103
Segundo	Thr153, Ser154, Cys155, Ser157, Glu158, Gly199, Ser200, Asn201, Gln202, Leu204
Terceiro	Thr14, Ala15, Val16, Lys17, Asp18, Phe28, Glu35, Leu45, Thr46, Asn47, Leu48, Glu50, Thr85, Glu86, Asp87, Tyr91
Quarto	Gln51, Tyr89, Pro90, Ala92, Ser93, Glu95, Ile97, Pro99, Pro100, Thr102
Quinto	Ser143, Thr146, Tyr147, Thr148, Gly149, Gly150, Val151, Met152, Thr153, Ser154, Cys155, Val156, Ala197, Lys198, Gly199, Ser200, Asn201
Sexto	Ala1, Pro2, Ala3, Ala4, Asp121, Glu122, Ala123, Gln124, Ala126, Ala127, Tyr169, Asp171, Ser172, Pro176, Lys198, Gly199, Ser200, Asn201, Lys206
Sétimo	Trp7, Gln37, Leu40, Ala41, Gly42, Val116, Glu117, Leu118, Pro119, Trp128, Val131, Asn132, Ala212, Val213, Val214, Gly215

Tabela 1. Sítios alostéricos preditos computacionalmente para a cruzaína. Listados os 7 potenciais sítios alostéricos para a cruzaína com os respectivos resíduos que os compõe (ALVAREZ, L. H., 2017).

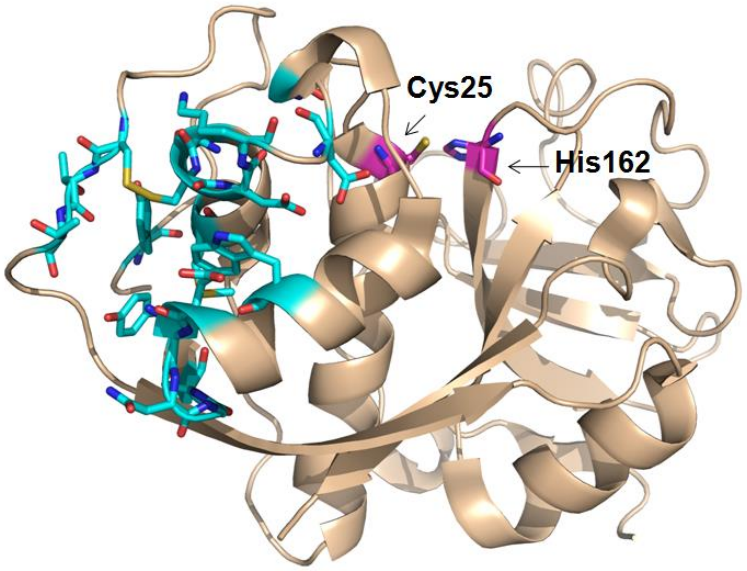

Figura 5. Estrutura da cruzaína com a representação do primeiro sítio alostérico predito computacionalmente. Estrutura da cruzaína (PDB: 3KKU) com os resíduos que compõem a díade catalítica, Cys25 e His162, destacados com carbonos em magenta. Resíduos do primeiro sítio alostérico em ciano.Nitrogênio em azul escuro, oxigênio em vermelho e enxofre em amarelo. Figura preparada com o programa Pymol (Schrödinger, LCC, 2015).

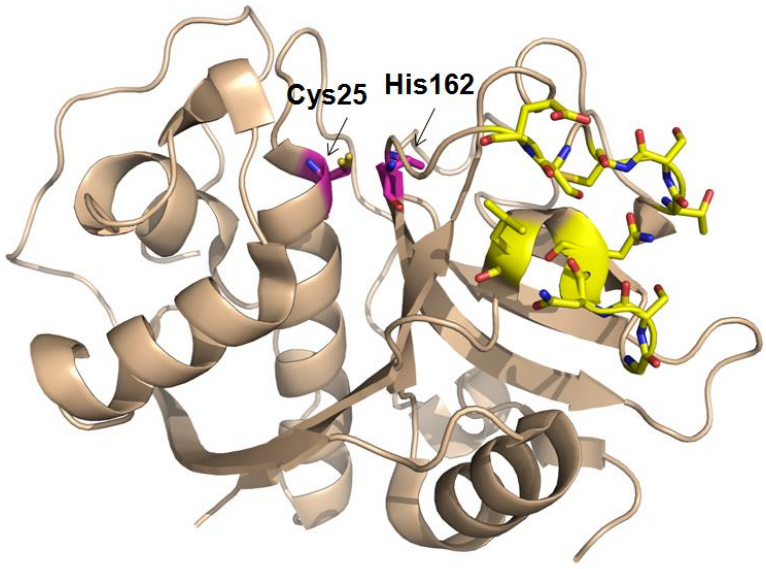

Figura 6. Estrutura da cruzaína com a representação do segundo sítio alostérico predito computacionalmente. Estrutura da cruzaína (PDB: 3KKU) com os resíduos que compõem a díade catalítica, Cys25 e His162, destacados com carbonos em magenta. Resíduos do segundo sítio alostérico em amarelo.Nitrogênio em azul escuro, oxigênio em vermelho e enxofre em amarelo. Figura preparada com o programa Pymol (Schrödinger, LCC, 2015).

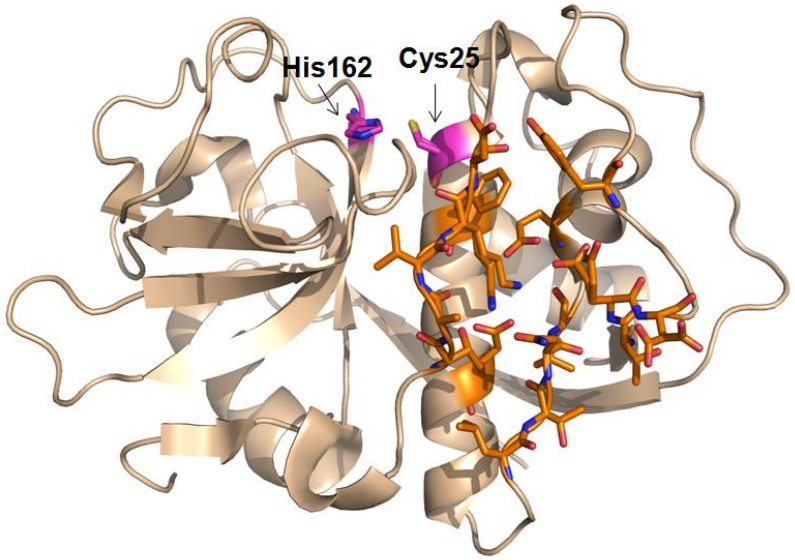

Figura 7. Estrutura da cruzaína com a representação do terceiro sítio alostérico predito computacionalmente. Estrutura da cruzaína (PDB: 3KKU) com os resíduos que compõem a díade catalítica, Cys25 e His162, destacados com carbonos em magenta. Resíduos do terceiro sítio alostérico em laranja.Nitrogênio em azul escuro, oxigênio em vermelho e enxofre em amarelo. Figura preparada com o programa Pymol (Schrödinger, LCC, 2015).

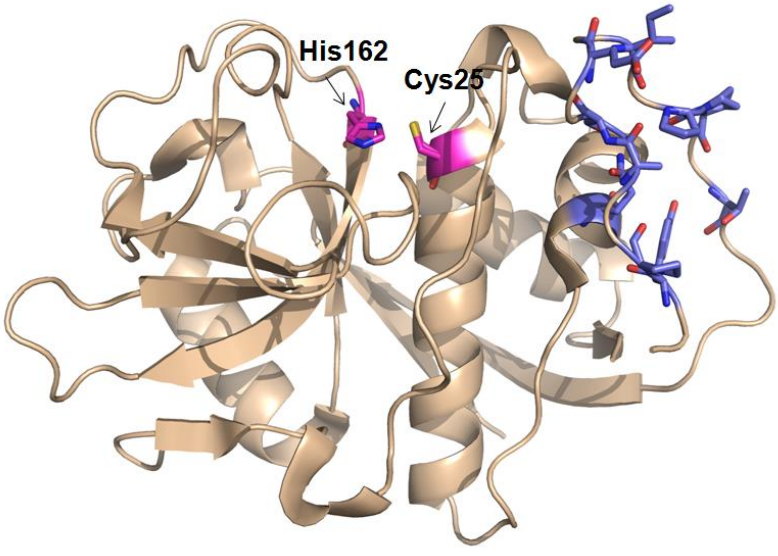

Figura 8. Estrutura da cruzaína com a representação do quarto sítio alostérico predito computacionalmente. Estrutura da cruzaína (PDB: 3KKU) com os resíduos que compõem a díade catalítica, Cys25 e His162, destacados com carbonos em magenta. Resíduos do quarto sítio alostérico em roxo.Nitrogênio em azul escuro, oxigênio em vermelho e enxofre em amarelo. Figura preparada com o programa Pymol (Schrödinger, LCC, 2015).

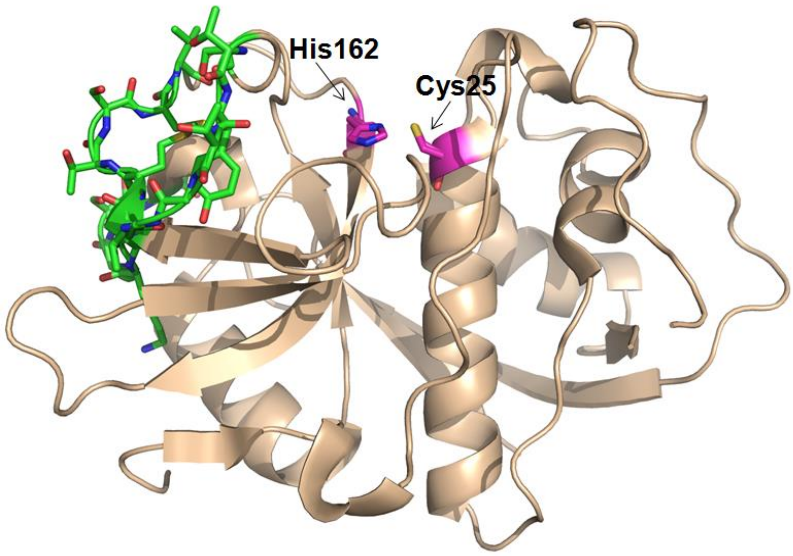

Figura 9. Estrutura da cruzaína com a representação do quinto sítio alostérico predito computacionalmente. Estrutura da cruzaína (PDB: 3KKU) com os resíduos que compõem a díade catalítica, Cys25 e His162, destacados com carbonos em magenta. Resíduos do quinto sítio alostérico em verde. Nitrogênio em azul escuro, oxigênio em vermelho e enxofre em amarelo. Figura preparada com o programa Pymol (Schrödinger, LCC, 2015).

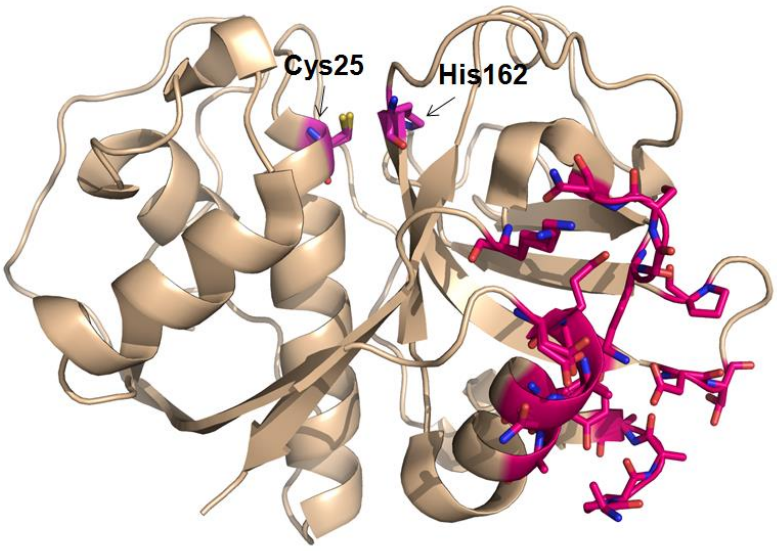

Figura 10. Estrutura da cruzaína com a representação do sexto sítio alostérico predito computacionalmente. Estrutura da cruzaína (PDB: 3KKU) com os resíduos que compõem a díade catalítica, Cys25 e His162, destacados com carbonos em magenta.Resíduos do sexto sítio alostérico em rosa shock. Nitrogênio em azul escuro, oxigênio em vermelho e enxofre em amarelo. Figura preparada com o programa Pymol (Schrödinger, LCC, 2015).

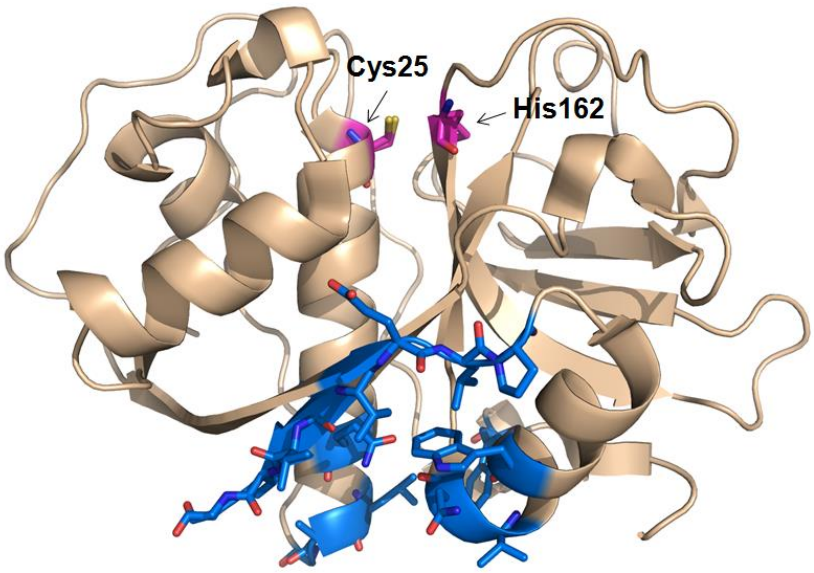

Figura 11. Estrutura da cruzaína com a representação do sétimo sítio alostérico predito computacionalmente. Estrutura da cruzaína (PDB: 3KKU) com os resíduos que compõem a díade catalítica, Cys25 e His162, destacados com carbonos em magenta. Resíduos do sétimo sítio alostérico em azul escuro. Nitrogênio em azul escuro, oxigênio em vermelho e enxofre em amarelo. Figura preparada com o programa Pymol (Schrödinger, LCC, 2015).

Portanto, apesar de não haver evidências experimentais que possam confirmar a existência desses 7 sítios alostéricos bem como dos resíduos que os compõem, estes estudos computacionais representam um ponto de partida para pesquisas futuras que possam validar a existência desses sítios alostéricos na enzima cruzaína.

Dentre as inúmeras enzimas pertencentes à família das cisteíno proteases algumas se destacam por serem alvo de muitas pesquisas, bem como por possuírem envolvimento em processos biológicos relevantes. Assim, enzimas como a cruzaína, papaína e as catepsinas humanas B, C, F, H, K, Z, O, S, W, L e V são extremamente relevantes. Abaixo segue uma descrição do envolvimento e importância dessas enzimas em diversos processos biológicos. Cumpre ressaltar que no presente trabalho a enzima cruzaína será utilizada como um exemplo representativo da família das cisteíno proteases.

1.3.1. Cruzaína

A cruzaína é uma importante enzima do protozoário *Trypanosoma cruzi*, agente etiológico da doença de Chagas, e é expressa em todos os estágios de vida do parasita: amastigota, tripomastigoa e epimastigota, sendo que nos epimastigotas a sua expressão é maior (SAJID, M. et al, 2011). Ela localiza-se em vesículas intracelulares durante o estágio de epimastigota do *Trypanosoma cruzi*, em um ambiente levemente ácido. Durante o estágio de amastigota do ciclo de vida do parasita a cruzaína passa a se localizar em sua superfície, onde o pH do hospedeiro está na faixa da neutralidade (pH: 7,4) (ARAFET, K., et al, 2017).

A cruzaína originalmente foi denominada cruzipaína, pois é a principal enzima da família da papaína do *Trypanosoma cruzi*. Assim, o termo cruzipaína refere-se à enzima nativa derivada do parasita, enquanto cruzaína refere-se à forma recombinante da proteína, que possui a região C-terminal truncada. Na literatura ambos os termos, cruzaína e cruzipaína, são utilizados. No entanto, neste trabalho será utilizado apenas o termo cruzaína para descrever ambas as formas. A cruzaína é uma enzima constituída por 467 resíduos e possui a arquitetura típica de domínios comum às demais enzimas da família da papaína. Ela possui um peptídeo sinal, formado pelos resíduos nas posições 1 a 18, responsável pelo direcionamento da proteína para o retículo endoplasmático. A pró-região, composta pelos resíduos 19 a 122, possui 3 funções: atua como uma chaperona intramolecular auxiliando no dobramento correto da enzima nascente, possui atividade inibitória enzimática potente e é um elemento essencial para o tráfico intracelular. O domínio catalítico possui 215 resíduos (123 a 337) e a porção C-terminal é formada pelos resíduos 338 a 467 e contém um número considerável de modificações pós-traducionais (SAJID, M. et al, 2011). Na figura 12, representada abaixo, estão indicadas as regiões da cruzipaína e da cruzaína.

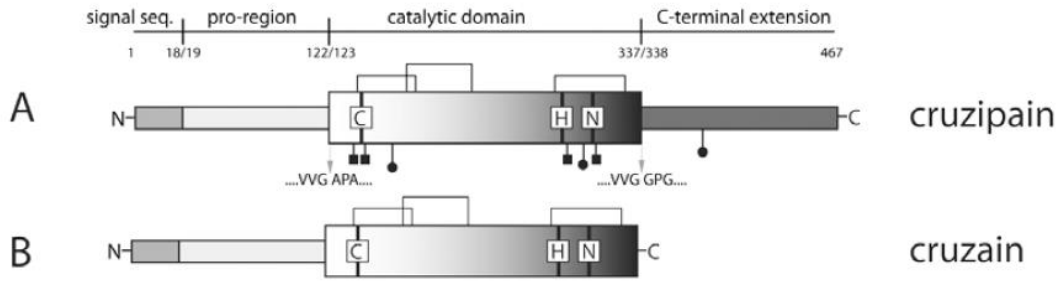

Figura 12. Representação esquemática da cruzipaína e da cruzaína com sinalização das regiões que as compõem. Em A, estrutura esquemática da cruzipaína composta pelo peptídeo sinal (resíduos 1 a 18), pró-região (resíduos 19 a 122), domínio catalítico (resíduos 123 a 337) e extensão C-terminal (resíduos 338 a 467). Em B, estrutura esquemática da cruzaína recombinante utilizada em estudos bioquímicos e estruturais. Nessa estrutura não é observada a extensão C-terminal, no entanto, encontram-se as demais regiões descritas para a cruzipaína. N: domínio N-terminal, C: domínio C-terminal. C, H, N dentro do domínio catalítico representam Cys25, His162 e Asn182, respectivamente (numeração da cruzaína). Os quadrados sinalizados acima do domínio catalítico representam pontes dissulfeto internas. Os sítios de N-glicosilação estão representados por quadrados preenchidos e os sítios de O-glicosilação estão representados pelos círculos preenchidos. Os locais de clivagem entre a pró-região e o domínio catalítico e entre o domínio catalítico e a extensão C-terminal estão destacados com uma seta cinza e os resíduos que cercam a ligação de cisão estão representados. Figura extraída de: SAJID, M. et al (2011).

O domínio catalítico da cruzaína possui duas regiões, uma formada por α-hélices (L-domínio) e outra formada por folhas β antiparalelas (R-domínio). A díade catalítica da cruzaína é composta pelos resíduos Cys25 e His162. Estudos mostram que a cruzaína é uma enzima crítica para a sobrevivência do parasita (LIMA, A. P, C. A. et al, 2001) e possui importância crucial para invasão celular e evasão do sistema imunológico do hospedeiro, além de atuar nas funções de metabolismo e reprodução do parasita (SOUZA, A. S., OLIVEIRA, M. T., ANDRICOPULO, A. D., 2017).

Devido ao fato da cruzaína ser uma enzima extremamente relevante para a sobrevida do *Trypanossoma cruzi*, ela configura-se como um importante alvo terapêutico para controle da doença de Chagas. A doença de Chagas, conhecida também como tripanossomíase americana, é uma doença potencialmente fatal e encontra-se principalmente em 21 países da América Latina, onde é transmitida

majoritariamente por vetores, sendo que o principal vetor é o triatomíneo conhecido como “barbeiro”. Segundo a Organização Mundial de Saúde estima-se que cerca de 8 milhões de pessoas estejam infectadas em todo o mundo, e a maior concentração de pessoas infectadas é na América Latina. Também são estimados mais de 10 mil óbitos decorrentes das manifestações clínicas da doença, sendo que mais de 25 milhões de pessoas apresentam risco iminente de contrair a doença. A doença muitas vezes não é curável, a não ser que seja realizado o diagnóstico precoce.

1.3.2. Papaína

A papaína é uma cisteíno protease isolada do látex do mamão, *Carica papaya L*, a partir do corte da sua pele verde, e quanto mais verde a fruta mais ativa é a papaína. Esta enzima é importante em muitos processos biológicos vitais e apresenta extensa atividade proteolítica, sendo aplicada extensivamente na medicina e na indústria alimentícia. Ela cliva ligações peptídicas envolvendo aminoácidos básicos, como arginina e lisina e resíduos localizados após a fenilalanina (AMRI, E., MAMBOYA, F., 2012).

A papaína pode ser usada na medicina como agente debridante, sem apresentar efeitos nocivos sobre os tecidos sadios (FLINDT, M. L., 1979), como auxiliar na remoção da cárie dentária (BEELEY, J. A., YIP, H. K., STEVENSON, A. G., 2000), como tratamento adjuvante de lesões esportivas e outras lesões traumáticas (DIETRICH, R. E., 1965). A papaína também tem sido utilizada com bons resultados no tratamento de alergias relacionadas a problemas intestinais, além de possuir atividade analgésica e anti-inflamatória significativa no tratamento da sinusite alérgica aguda (MANSFIELD, L E. et al, 1985).

Na indústria alimentícia é utilizada como amaciante de carnes e em cervejaria, podendo atuar também como agente clarificante. Também pode ser utilizada na fabricação de alguns doces e queijo Nabulsi (ABU-ALRUZ, K. et al., 2009).

1.3.3. Catepsinas Humanas

As catepsinas humanas totalizam 11 e, com exceção da catepsina S, são encontradas em uma ampla variedade de células. Acreditava-se que sua função primária era a degradação de proteínas de forma não seletiva no interior dos lisossomos. No entanto, elas podem ser encontradas externamente aos lisossomos, em situações especiais que muitas vezes caracterizam condições patológicas (TURK, B., TURK, D., TURK, V., 2000). Essas 11 catepsinas podem ser divididas da seguinte forma (tabela 2 mostrada abaixo): grupo das catepsinas semelhantes à catepsina B, grupo das catepsinas semelhantes à catepsina L, catepsina X, catepsina C e catepsina W (sendo que essas 3 últimas catepsinas não fazem parte de nenhum grupo ou subgrupo específico).

Catepsinas Humanas		
Grupos	**Subgrupo**	**Componentes**
Catepsinas semelhantes à B	_____	Catepsina B
Catepsinas semelhantes à L	Catepsinas semelhantes à L	Catepsinas L, V, S, K
	Catepsinas semelhantes à F	Catepsina F
	Catepsinas semelhantes à H	Catepsina H
	_____	Catepsina O
Outros	_____	Catepsinas X, C, W

Tabela 2. Divisão das catepsinas humanas em grupos e subgrupos. Listados os 2 grupos de catepsinas humanas com os respectivos subgrupos (quando houverem) e as enzimas que os compõem. As catepsinas humanas X, C e W não fazem parte de nenhum grupo ou subgrupo específico.

A divisão das catepsinas nos grupos de catepsinas semelhantes à B e à L foi realizada inicialmente por Karrer et al (1993) considerando-se a presença ou não do motivo conservado ERFNIN. O motivo ERFNIN (Glu-Arg-Phe-Asn-Ile-Asn) é uma sequência de consenso de resíduos altamente conservados encontrados na região pró-peptídeo da enzima. Assim, as catepsinas semelhantes à L possuem tal motivo em suas regiões pró-peptídeo enquanto as catepsinas semelhantes à B não possuem tal motivo. Ademais, a distribuição filogenética sugere que esses dois grupos foram estabelecidos no início do processo evolutivo da família das cisteíno proteases. As proteases semelhantes à L são encontradas em organismos que variam desde protozoários a mamíferos enquanto provavelmente as proteases semelhantes à B evoluíram antes da divergência dos platelmintos (KARRER, K. M., PEIFFER, S. L., DITOMAS, M. E., 1993).

As catepsinas semelhantes à B compõem um importante grupo e apresentam uma região pró-peptídeo um pouco mais longa (62 resíduos na catepsina humana B) que contém duas α-hélices curtas. Essas enzimas possuem um padrão conservado de seis pontes dissulfeto e um loop oclusivo característico que determina sua atividade exoproteolítica (NOVINEC, M., LENARCIC, B., 2013). As enzimas que compõem este grupo foram isoladas e caracterizadas a partir de outros organismos além dos mamíferos, como em plantas, onde estão envolvidas em uma forma de apoptose secundária à invasão de patógenos, em mecanismos de defesa basais e em processos de senescência (GILROY, E. M. et al, 2007; MCLELLAN, H. et al, 2009). Foram encontradas catepsinas semelhantes à B em espécies de *Leishmania*, onde elas estão envolvidas em apoptose (EL-FADILI, A. K. et al, 2010) e também no parasita *Toxoplasma gondii*, onde apresentam função crítica para a invasão do hospedeiro pelo parasita (QUE, X. et al, 2002).

Dentro do grupo de catepsinas semelhantes à B, há apenas uma catepsina humana, a catepsina B.A atividade da catepsina B é mais amplamente investigada em algumas patologias, dentre elas doenças reumáticas (LENARCIC, B. et al, 1988; BAICI, A. et al, 1995; BAICI, A. et al, 1995) e diversos tipos de câncer, como carcinomas de mama e cólon, onde ela encontra-se geralmente associada aos estágios de progressão tumoral bem como com malignidade. Assim, a expressão da catepsina B está regulada positivamente em numerosos tumores (MOHAMED M. M., SLOANE, B. F., 2006). Foram demonstrados efeitos benéficos da catepsina B sobre

cognição, neurogênese hipocampal e memória espacial em ratos. Em humanos observou-se uma correlação positiva entre níveis elevados de catepsina B, atividade física e memória com envolvimento do hipocampo (MOON, H. Y. et al, 2016). Também foi demonstrado, em experimentos realizados com ratos, que a catepsina B possui importante papel na clivagem da proteína β amilóide, que é encontrada de forma acumulada no meio extracelular de pacientes com doença de Alzheimer (WANG, C. et al, 2012).

As catepsinas do tipo L constituem o maior grupo e as enzimas que o compõem são caracterizadas por um padrão conservado de 3 pontes dissulfeto, embora algumas enzimas possam apresentar pontes dissulfeto adicionais. Elas também possuem o maior pró-peptídeo, com mais de 100 resíduos que se dobra predominantemente em um domínio de α-hélice que contém os motivos que as distingue do subgrupo de catepsinas semelhantes à B. O grupo de catepsinas semelhantes à L é dividido em outros subgrupos: catepsinas semelhantes à L (contém as catepsinas humanas L, V, S e K), catepsinas semelhantes à F (contém apenas a catepsina humana F), catepsinas semelhantes à H (contém apenas a catepsina humana H) e a catepsina O (não está alocada em nenhum subgrupo) (tabela 2) (NOVINEC, M., LENARCIC, B., 2013).

As proteínas do subgrupo das catepsinas semelhantes à L estão envolvidas em numerosas condições fisiológicas e patológicas e dependendo da situação podem agir sozinhas ou conjuntamente com outras peptidases (NOVINEC, M., LENARCIC, B., 2013). A secreção das catepsinas L, V, S e K a partir de macrófagos ativados em processos inflamatórios foi demonstrada, indicando que essas enzimas sejam um fator causal principal para os danos teciduais ocorridos decorrentes da inflamação crônica (PUNTURIERI, A. et al, 2000; REDDY V. Y., ZHANG, Q. Y., WEISS, S. J., 1995).As catepsinas semelhantes à L apresentam atividade de degradação da elastina e tal atividade está relacionada ao desenvolvimento de condições patológicas decorrentes da excessiva degradação de componentes da matriz extracelular, como doenças cardiovasculares e pulmonares (CHENG, X. W. et al, 2011; VEILLARD, F., LECAILLE, F., LALMANANCH, G., 2008). Essas catepsinas também estão implicadas em outras condições inflamatórias, tais como pancreatite (LYO, V. et al, 2012), bem como em diversos tipos de câncer (MOHAMED M. M., SLOANE, B. F., 2006), além de estarem provavelmente relacionadas na regulação

do processo de geração de células adiposas, podendo ocasionar obesidade e diabetes (HAN, J. et al, 2009; YANG, M. et al, 2008; YANG M. et al, 2007; TALEB, S. et al, 2006).

A catepsina S é a peptidase mais importante no processo de apresentação de antígenos (NAKAGAWA, T. Y. et al, 1999; SHI G. P. et al, 1999; DRIESSEN, C. et al, 1999). Ela é seletivamente expressa em células apresentadoras de antígenos, como os linfócitos B, células dendríticas, macrófagos e microglia. Portanto a catepsina S possui um envolvimento significativo na resposta imune adaptativa participando do processamento de peptídeos de proteínas estranhas. Ela também parece possuir papel importante na artrite reumatóide, atuando no processo de degradação dos joelhos, bem como na esclerose múltipla (CLARK, A. K., MALCANGIO, M., 2012).

Por outro lado, a catepsina K é a principal peptidase com atividade colagenolítica envolvida no processo de remodelamento ósseo (NOVINEC, M., REBERNIK, M., LENARCIC, B., 2016). Mutações envolvendo o gene que codifica a catepsina K dão origem a uma doença autossômica recessiva rara denominada picnodisostose, que se caracteriza por severas anormalidades ósseas (GELB, B. D. et al, 1996). A atividade excessiva da catepsina K tem sido associada com a ocorrência de osteoporose (NOVINEC, M., REBERNIK, M., LENARCIC, B., 2016, BOONEN, S. et al, 2012) e ela também possui envolvimento na ocorrência de osteosarcomas, inclusive em situações de metástases (HUSMANN, K. et al, 2008). A catepsina K também se relaciona à esquizofrenia e à sinalização envolvendo receptores semelhantes ao Toll, bem como à homeostase cardiovascular (LENDECKEL, U. et al, 2009; ASAGIRI, M. et al, 2008; LECAILLE, F. et al, 2007).

A catepsina L é extremamente importante para a homeostase da pele (ROTH, W. et al, 2000) e pode ser substituída em alguns casos pela catepsina V. Ela pode ser encontrada no núcleo das células, onde participa da regulação da progressão do ciclo celular mediante processamento de fatores de transcrição (GOULET, B. et al, 2004; GOULET, B., TRUSCOTT, M., NEPVEU, A., 2006).

A catepsina V, por sua vez, é encontrada em vesículas secretórias no córtex cerebral e hipocampo e é responsável pela produção dos peptídeos neurotransmissores encefalina e neuropeptídeo Y (FUNKELSTEIN, L. et al, 2012). Ela foi primariamente identificada no timo e testículo e está associada à miastenia

grave, diabetes tipo 1 e doenças neurológicas. É expressa em altos níveis nos tumores colorretais e mamários e também se relaciona à progressão do câncer. A sua atividade está relacionada a N-glicosilação de seus sítios (NIWA, Y. et al, 2012).

A catepsina F distingue-se das demais peptidases semelhantes à papaína por possuir um pró-domínio anormalmente longo, que contempla o pró-peptídeo e também uma sequência de aproximadamente 100 resíduos, similarmente às cistatinas (NAGLER, D. K., SULEA, T., MENARD, R., 1999). Ela possui o motivo altamente conservado ERFNAQ (Glu-Arg-Phe-Asn-Ala-Gln) em sua região pró-peptídeo, assim como as demais catepsinas semelhantes à L, no entanto, ao invés de possuir o motivo ERFNIN, possui o motivo ERFNAQ (NOVINEC, M., LENARCIC, B., 2013). A especificidade e atividade da catepsina F é similar às das catepsinas semelhantes à L com as quais também se assemelha no que tange ao envolvimento nos mesmos processos fisiopatológicos. A catepsina F está envolvida na apresentação de antígenos (SHI, G. P. et al, 2000) e também tem sido encontrada em placas ateroscleróticas onde está implicada na degradação da lipoproteína de baixa densidade (LDL) (OORNI, K. et al, 2004).

A catepsina H foi uma das primeiras catepsinas a serem identificadas. Ela está envolvida no processamento de vários neuropeptídeos (BRGULJAN, P. M. et al, 2003; LU, W. D. et al, 2012), bem como no processamento da granzima B, além de ser importante, porém não indispensável, para a maturação da proteína B surfactante nos pulmões (BUHLING et al, 2011). Estudos recentes mostram que a catepsina H apresenta contribuição importante em patologias inflamatórias e metástase. Os níveis de catepsina H encontram-se elevados nos processos inflamatórios decorrentes de lesões pulmonares agudas, pancreatite, miopatia inflamatória e aterogênese e esses níveis elevados correlacionam-se diretamente com a gravidade do processo inflamatório nessas doenças. Os níveis de catepsina H encontram-se aumentados em áreas do cérebro na doença de Huntington,portanto, tal catepsina provavelmente correlaciona-se positivamente com processos neuroinflamatórios na patogênese de doenças neurológicas. Estudos *in vitro* demonstraram que as citocinas pró-inflamatórias (TNF-α, IL-1β, IL-6, IFN-γ) induzem a liberação de catepsina H pela microglia e que a catepsina H induz ativação microglial caracterizada pela liberação de óxido nítrico (NO) e várias outras citocinas inflamatórias. Tais efeitos levam a um círculo vicioso que pode levar a uma

inflamação crônica. Também foi demonstrado um efeito tóxico da catepsina H para os neurônios, com consequente morte neuronal (FAN, K. et al, 2015).

Com relação à catepsina O cumpre destacar que há poucos estudos avaliando a sua função. Apesar dela ser amplamente expressa, ainda não foi profundamente investigada. Ela integra o grupo das catepsinas semelhantes à L, no entanto, é a que menos se assemelha às demais enzimas do grupo (NOVINEC, M., LENARCIC, B., 2013).

A catepsina X pode também ser denominada como catepsina Z, catepsina B2, carboxipeptidase ou catepsina IV. Ela é altamente expressa no sistema imunológico, desempenhando importante função na regulação do comportamento celular bem como na sua diferenciação (OBERMAJER, N. et al, 2008; OBERMAJER, N. et al, 2008). A catepsina X também parece possuir papel na regulação da adesão celular, juntamente com integrinas de superfície celular, além de realizar ligação aos proteoglicanos da superfície celular (LECHNER, A. M. et al, 2006). Além disso,está associada a doenças neurodegenerativas, como a doença de Alzheimer, e possíveis alvos dessa enzima no Sistema Nervoso Central (SNC) seriam as α e γ-enolases (WENDT, W. et al, 2007; OBERMAJER, N. et al, 2009). Outros possíveis substratos que foram identificados são alguns hormônios peptídicos como a bradicinina (NAGLER, D. K. et al, 2010). A catepsina X é regulada positivamente na mucosa gástrica inflamada em infecções ocasionadas por *Helicobacter pylori*, bem como em câncer gástrico, sendo assim um promissor marcador biológico para tais patologias (KRUEGER, S. et al, 2005).

A catepsina C, também denominada dipeptidil peptidase I, é a única dentre as peptidases similares à papaína que é encontrada sob a forma de tetrâmero; as demais são monômeros. As mutações que ocasionam perda da função do gene responsável pela codificação da catepsina C apresentam como consequência o desenvolvimento de um distúrbio genético de caráter recessivo, a síndrome de Papillon-Lefèvre. Tal síndrome caracteriza-se por hiperceratose das palmas das mãos e da sola dos pés, periodontite grave com perda prematura dos dentes por volta dos 20 anos de idade, além do aumento da suscetibilidade às infecções, que pode ocorrer em alguns casos (HART, T. C. et al, 1999). A catepsina C é expressa de forma onipresente e possui também papel importante no sistema imunológico,

atuando como ativadora de serino peptidases efetoras, tais como granzimas, elastases e mieloblastinas (MCGUIRE, M. J., LIPSKY, P. E., THIELE, D. L., 1993; PHAM, C. T., LEY, T. J., 1999; ADKISON, A. M. et al, 2002). Devido a essas funções, tem sido considerada um importante alvo terapêutico para patologias nas quais há uma atividade excessiva dessa peptidase, tais como em doenças inflamatórias e auto-imunes (GUAY, D., BEAULIEU, C., PERCIVAL, M. D., 2010).

A catepsina W é encontrada no retículo endoplasmático e parece estar expressa apenas em células T citotóxicas e células *natural killer*, sem ser essencial para a citotoxicidade dessas células (STOECKLE, C. et al, 2009). Um estudo recente mostrou que a catepsina W é um fator intracelular necessário para a entrada do vírus da gripe influenza A nas células alvo, bem como para o ciclo de vida viral. Ela também é importante para a fusão entre as membranas virais e endossômicas. Também foi demonstrado que a atividade proteolítica da catepsina W é necessária para o desempenho de sua função “pró-viral” tornando-a assim um alvo potencial para o desenvolvimento de medicamentos antivirais (EDINGER, T. O., 2015).

A partir da breve descrição de algumas cisteíno proteases pode-se concluir que todas apresentam funções relevantes e que, no caso das catepsinas humanas, por exemplo, qualquer descontrole na atividade e expressão dessas enzimas pode ocasionar sérias consequências para o equilíbrio e a saúde humana.

2. Justificativa

As cisteíno proteases são um grande grupo de proteínas que fazem parte do clã CA e da família C1 (família da papaína). Elas possuem uma ampla distribuição em vários seres vivos, sendo encontradas em todos os animais e reinos de plantas bem como em muitos vírus e organismos procariotas (WIEDERANDERS, KAULMANN & SCHILLING, 2003). Segundo a base de dados da MEROPS, há na família C1 15702 sequências catalogadas.

Algumas proteínas dessa família possuem destaque para a pesquisa como possíveis alvos para desenvolvimento de fármacos, como por exemplo, a cruzaína em relação à doença de Chagas, a catepsina B em relação ao câncer, a catepsina S em relação ao controle da apresentação de antígenos e possivelmente atenuação da resposta imune, a catepsina K em relação ao controle da osteoporose e a catepsina L em relação ao processamento de antígeno e metabolismo de células tumorais (MEROPS).

Assim, devido ao fato dessas proteínas possuírem envolvimento em patologias importantes e/ou em mecanismos fisiológicos significativos, faz-se necessário o conhecimento mais aprofundado das mesmas. No presente trabalho estudamos a família de cisteíno proteases a partir da realização de um alinhamento múltiplo de sequências e análise dos resíduos correlacionados e resíduos conservados dessas proteínas. A partir desses dados será discutida a importância e função de cada comunidade de resíduos, assim como de resíduos específicos, mediante análises de bioinformática e de referências encontradas na literatura. Cumpre destacar que as proteínas que serão tomadas como referência da família de cisteíno proteases nesse trabalho são: a cruzaína, a papaína e as catepsinas humanas.

Diante do exposto, a importância do presente trabalho justifica-se pela necessidade de obter um melhor entendimento da estrutura das cisteíno proteases, dos seus resíduos correlacionados e conservados, bem como da importância destes para a estrutura e função dessas cisteíno proteases. Portanto, tal trabalho contribui para o corpo de conhecimentos da ciência básica na medida em que busca entender de uma forma mais detalhada e aprofundada a família das cisteíno proteases.

3. Objetivos

3.1. Objetivo Geral

Analisar e discutir a importância dos resíduos correlacionados e conservados da família das cisteíno proteases a partir do alinhamento múltiplo de sequências tendo como referência as enzimas cruzaína, papaína e catepsinas humanas.

3.2. Objetivos Específicos

- Determinar a conservação e a correlação de resíduos que coevoluem nas cisteíno proteases tendo como referência as enzimas cruzaína, papaína e catepsinas humanas mediante aplicação da metodologia DRCN proposta por Bleicher, Lemke & Garratt (2011).
- Discutir a importância e a função de resíduos conservados e que coevoluem nas cisteíno proteases e em cada comunidade.

4. Materiais e Métodos

4.1. Decomposição de Redes de Correlação de Resíduos

A análise de sequências das cisteíno proteases foi realizada empregando-se a metodologia DRCN (*Decomposition of Residue Coevolution Networks*), que foi proposta por Bleicher, Lemke & Garratt (2011) e consiste nos seguintes passos:

1. Filtragem do alinhamento: considerando-se que a única entrada para tal método seja um alinhamento múltiplo de sequências, é necessário que ele constitua-se uma amostra representativa da família. Portanto, o *software* elimina tanto sequências de tamanho menor que o esperado para a proteína em questão quanto sequências muito parecidas (redundantes).

2. Cálculo de correlações: é realizado um cálculo de correlação, denominado escore de correlação, para cada par de resíduos (com as suas respectivas posições). Escores positivos indicam correlação, ou seja, a presença de um resíduo em uma posição aumenta a ocorrência de outro resíduo em outra posição. Escores negativos indicam anti-correlação, ou seja, a presença de um resíduo em uma posição diminui a ocorrência de outro resíduo em outra posição. A partir do cálculo de todos os escores possíveis para um alinhamento é construída uma rede de correlações (a partir dos escores mais significativos, acima de um determinado valor de corte).

3. Decomposição da Rede: a rede de correlações é decomposta em comunidades e cada comunidade é composta por resíduos altamente conectados entre si, mas não ao resto da rede. A espessura das linhas que liga dois resíduos é diretamente proporcional ao grau de correlação entre eles. As comunidades representam grupos de resíduos que tendem a aparecer simultaneamente em um subconjunto de uma família de proteínas (resíduos que coevoluem).

4. Geração de arquivos de visualização auxiliares: são gerados arquivos que facilitam o entendimento dos resultados (tabelas de resíduos conservados, matrizes de auto-correlação, figuras representando as redes de correlação, entre outros).

5. Anotação de posições por busca automática no UniProt: o *software* realiza uma busca na base de dados UniProt de todas as referências existentes às posições encontradas nas proteínas de uma família e as lista.

Portanto, tal ferramenta possibilita uma análise mais aprofundada de uma família de proteínas e por isso foi escolhida para atender aos objetivos propostos para este trabalho.

4.2. Obtenção e Filtragem do Alinhamento

A obtenção do alinhamento de todas as proteínas que integram a família das cisteíno proteases foi realizada a partir do Pfam, que promove a anotação das famílias de proteínas a partir dos seus motivos estruturais. Para realização da análise do alinhamento da família das cisteíno proteases foi utilizado o programa PFSTATS (*Protein Families Statistics*) (FONSECA-JÚNIOR, N. J. et al, 2018) que analisa as famílias de proteínas através de métodos estatísticos e calcula a conservação e a correlação de resíduos específicos bem como as suas posições na proteína.

Foi utilizado o código de acesso Pfam das cisteíno proteases, PF00112, o que gerou inicialmente 20698 sequências. A partir daí foi realizada filtragem do alinhamento pelos filtros de mínima cobertura e máxima identidade. Um filtro de mínima cobertura com valor 0.8 foi aplicado ao alinhamento, o que significa que foram removidas do alinhamento inicial sequências que apresentavam menos de 80% das posições esperadas para domínios cisteíno-protease. O filtro de máxima identidade é aplicado para evitar o viés de redundância, que levaria à detecção de posições equivocadamente detectadas como conservadas ou coevoluídas devido à sua presença em grande número de sequências parecidas. O valor utilizado foi de

0,8, portanto a cada vez que um par de sequências com uma identidade maior que 80% era detectado, uma dessas sequências era eliminada.

Tais filtros resultam em um alinhamento de maior qualidade, que possibilita análises de correlação mais consistentes. A filtragem do alinhamento foi realizada a partir da interface do PFSTATS.

4.3. Cálculo do Sub-Alinhamento Mínimo

Para determinar o número mínimo de sequências que caracterizam o menor sub-alinhamento com representatividade estatística, define-se o sub-alinhamento mínimo, que é gerado após a aplicação dos filtros, utilizando uma versão adaptada do procedimento descrito por Dima & Thirumalai (2006). Para tal é realizada uma nova amostragem na qual a média da entropia de Shannon, também denominada medida da variabilidade, para todas as posições do alinhamento é calculada, enquanto há remoção de sequências aleatórias do alinhamento. O cálculo da entropia de Shannon é realizado através da equação 1 descrita abaixo:

$$\mathrm{H} = \sum_{i=1}^{M} \mathrm{P_i} \log_2 \mathrm{P_i}$$

Equação 1. Cálculo da entropia de Shannon.Fórmula utilizada para o cálculo da entropia de Shannon. Em uma determinada posição do alinhamento múltiplo de sequências tem-se que i é um índice que representa um determinado tipo de resíduo; M é o número total de resíduos e Pi é a frequência do resíduo i em sua respectiva posição.

Um gráfico bidimensional é construído com o valor médio de entropia para todas as posições do alinhamento no eixo das ordenadas e o tamanho do alinhamento no eixo das abscissas. O sub-alinhamento mínimo é um número

aproximado de sequências para o qual a retirada de sequências aleatórias ainda não causa alterações significativas da média da entropia de Shannon em relação ao alinhamento original.

Tanto o cálculo do sub-alinhamento mínimo quanto a formação e visualização do gráfico bidimensional da média da entropia de Shannon foram realizados a partir da interface do PFSTATS.

4.4. Construção de Rede de Correlações e Anticorrelações

Após o cálculo do sub-alinhamento mínimo é necessário definir a correlação e anticorrelação de pares de resíduos e suas respectivas posições no alinhamento múltiplo. Considerando-se dois resíduos, tem-se que a presença de um deles em uma determinada posição pode aumentar ou reduzir a freqüência do outro em outra posição em um valor mínimo.

Assim, uma correlação provável entre dois pares de resíduos e suas respectivas posições, como por exemplo, o resíduo X na posição A e o resíduo Y na posição B é definida quando há uma frequência mínima de Y na posição B no sub-alinhamento que contém todas as sequências que possuem X na posição A. A frequência esperada de sequências que contém Y na posição B para um sub-alinhamento que contém todas as sequências com X na posição A deveria ser a mesma observada para todo o alinhamento.

A partir daí pode-se calcular a probabilidade de ocorrer uma variação na frequência para um determinado alinhamento mediante a aplicação da distribuição binomial cumulativa tanto para as correlações quanto para as anticorrelações. O escore de correlação é definido pelo logaritmo na base 10 da probabilidade binomial cumulativa de um resíduo X ser encontrado em uma posição A, dada a sua frequência esperada, com sinal invertido. Uma probabilidade binomial cumulativa de 10^{-10} indica um escore de correlação de 10. Para anticorrelações o cálculo é realizado da mesma forma, no entanto, a probabilidade binomial cumulativa é calculada de forma a encontrar um valor menor ou igual ao número de sequências que contém X na posição A, considerando-se a frequência de X na posição A.

As equações 2 e 3 abaixo, desenvolvidas por Bleicher, Lemke & Garratt (2011), descrevem o cálculo das correlações e anticorrelações entre pares de resíduos em um sub-alinhamento mínimo.

$$P_{corr} = \sum_{n=n_{B/A}}^{n_{A!}} \frac{n_{A!}}{n!(n_a - n)!} \; X \left(\frac{n_B}{N}\right)^n X \left(1 - \frac{n_B}{N}\right)^{n_A - n}$$

$$P_{anticorr} = \sum_{n=0}^{n_{B/A}} \frac{n_{A!}}{n!(n_a - n)!} \; X \left(\frac{n_B}{N}\right)^n X \left(1 - \frac{n_B}{N}\right)^{n_A - n}$$

Equações2 e 3. Cálculo da distribuição binomial cumulativa para a definição de correlações e anticorrelações válidas. Fórmulas utilizadas para o cálculo da distribuição binomial cumulativa para definição de correlações (equação superior) e anticorrelações (equação inferior). n_A representa o número de sequências que contém o resíduo na posição A; n_B representa o número de sequências que contém o resíduo na posição B; N é o número total de sequências no alinhamento. O primeiro termo, somatório, na primeira equação é aquele no qual a probabilidade cumulativa para correlações será calculada e definida a partir da probabilidade de serem observadas o número de sequências que contém A e B ($n_{B/A}$) até o número total de sequências do sub-alinhamento (N). Na segunda equação, anticorrelações, o termo somatório é calculado a partir da probabilidade de ser observada nenhuma sequência que contenha A e B até o número observado de sequências que contenha A e B ($n_{A/B}$). Os demais termos que são comuns a ambas equações se referem a (considerando-se a sequência de aparecimento dos termos da esquerda para a direita): combinações possíveis de sequências que possuem um certo número de sequências que contenham B (n) dentro do sub-alinhamento n_A; a frequência observada de B no alinhamento múltiplo inicial elevada a um determinado número de sequências contendo B (n) dentro do sub-alinhamento n_A (frequência esperada de n_B), que representa a probabilidade de encontrar B e o inverso da frequência observada de B no alinhamento múltiplo inicial elevado ao número de sequências que não contém B (n) dentro do sub-alinhamento n_A, que representa a probabilidade de não encontrar B.

Para indicar que a presença de um resíduo em uma determinada posição diminui a probabilidade de ocorrência de um outro resíduo em uma determinada

posição utilizam-se números negativos. Esses números negativos servem para definir anticorrelação. Portanto, em uma probabilidade binomial cumulativa de 10^{-10} o escore de anticorrelação é -10. Nesse caso é exigido um valor absoluto de escore mínimo de 10 para definir uma correlação e uma anticorrelação válidas. As equações 4 e 5 desenvolvidas por Bleicher, Lemke & Garratt (2011) e mostradas abaixo evidenciam tal conceito.

$$Escore\ corr = -log_{10}p_{corr}$$

$$Escore\ anticorr = log_{10}p_{corr}$$

Equações 4 e 5. Cálculo dos escores de correlação e anticorrelação. Fórmulas utilizadas para o cálculo dos escores de correlação (fórmula superior) e de anticorrelação (fórmula inferior). O cálculo desses escores é realizado a partir do logaritmo da probabilidade binomial cumulativa. Para anticorrelação o escore é indicado por números negativos.

O cálculo das correlações e anticorrelações para a família das cisteíno proteases foi realizado a partir da interface do PFSTATS e o escore mínimo utilizado para o cálculo de correlações e anticorrelações foi de 10. Para que uma correlação entre dois resíduos seja válida é necessário que a variação da frequência de um resíduo em um sub-alinhamento definido por outro resíduo seja maior que o *cutoff* previamente definido. O tamanho mínimo desse sub-alinhamento também deve ser maior que o que foi calculado previamente e a probabilidade binomial cumulativa da correlação ter ocorrido ao acaso deve ser menor que 10^{-10}.

Portanto, os pares de resíduos são selecionados desde que a presença de um resíduo ocasione um aumento na frequência de um outro resíduo para 80% ou mais para correlações, ou uma diminuição na frequência do outro resíduo para 20% ou menos para anticorrelações.

4.5. Geração e Visualização de Dados

Após a construção da rede de correlações e anticorrelações o PFSTATS gera uma gama de resultados que são expressos em tabelas que contemplam os resíduos conservados, matrizes de correlação entre os resíduos, tabelas de resíduos correlacionados separados por comunidade, figuras contendo as redes de correlação e anticorrelação, dentre outros dados. É também gerada uma lista com referências obtidas no UniProt para todas as posições de resíduos encontradas nas proteínas da família.

4.6. Análise dos Dados e Buscas Bibliográficas

Foi realizada extensa busca bibliográfica para subsidiar a análise dos dados referentes aos resíduos correlacionados de cada comunidade bem como acerca dos resíduos conservados, descrevendo sua importância e função dentro da proteína. Para tal foram consideradas as enzimas cruzaína, papaína e as catepsinas humanas. Para os resíduos sobre os quais ainda não existem dados na literatura que sejam capazes de sustentar uma análise concreta foram realizadas inferências a partir das estruturas analisadas no Pymol.

Foram preparadas figuras a partir do programa Pymol (Schrödinger, LCC, 2015) representando as comunidades encontradas e os respectivos resíduos que as constituem, usando como modelo representante da família de cisteíno proteases a cruzaína. Também foi preparada uma figura contendo os resíduos conservados que foram encontrados na família de cisteíno proteases usando como modelo a enzima cruzaína.

4.7. Diagrama de Ramachandran

Para a obtenção dos diagramas de Ramachandran para a cruzaína (PDB: 3KKU) foi utilizado o servidor MolProbity, disponibilizado pela *Duke University School of Medicine*, através do endereço eletrônico http://molprobity.biochem.duke.edu/. Além dos gráficos obtidos para todos os resíduos (*general case*), também foram

gerados gráficos para os resíduos de isoleucina e valina, glicina, pré-prolina e prolina (configurações *cis* e *trans*). Ademais, foi gerada uma tabela contendo os ângulos φ e ψ para todos os resíduos da enzima.

5. Resultados

5.1. Obtenção e Filtragem do Alinhamento

O alinhamento inicial foi realizado a partir de 20698 sequências de cisteíno proteases, extraídas do UniProt, utilizando o programa PFSTATS. Em seguida foi realizada filtragem do alinhamento mediante aplicação dos filtros de mínima cobertura e máxima identidade. O emprego de *cutoffs* de 0,8 para ambos os filtros resultou em remoção de 6377 sequências pelo filtro de mínima cobertura e de 9303 sequências pelo filtro de máxima identidade. Portanto, o número de sequências remanescentes no alinhamento final filtrado foi de 5018, conforme representado na Tabela 3.

	Alinhamento Inicial	Sequências removidas por filtro de mínima cobertura	Sequências removidas por filtro de máxima identidade	Alinhamento final (filtrado)
Total de sequências	20698	6377	9303	5018

Tabela 3. Número de sequências filtradas a partir do alinhamento múltiplo da família das cisteíno proteases contendo todas as sequências do UniProt. O alinhamento múltiplo foi obtido pelo *software* PFSTATS.

5.2. Cálculo do Sub-Alinhamento Mínimo

Após análise do gráfico da média da entropia de Shannon para definição dos sub-alinhamentos gerados a partir da remoção aleatória de sequências foi definido um valor de 33% (1656 sequências) para o sub-alinhamento mínimo. Tal valor de sub-alinhamento mínimo é estatisticamente relevante e mantém a

representatividade. A figura 13, representada abaixo, mostra o gráfico com a média da entropia de Shannon e o tamanho do sub-alinhamento mínimo.

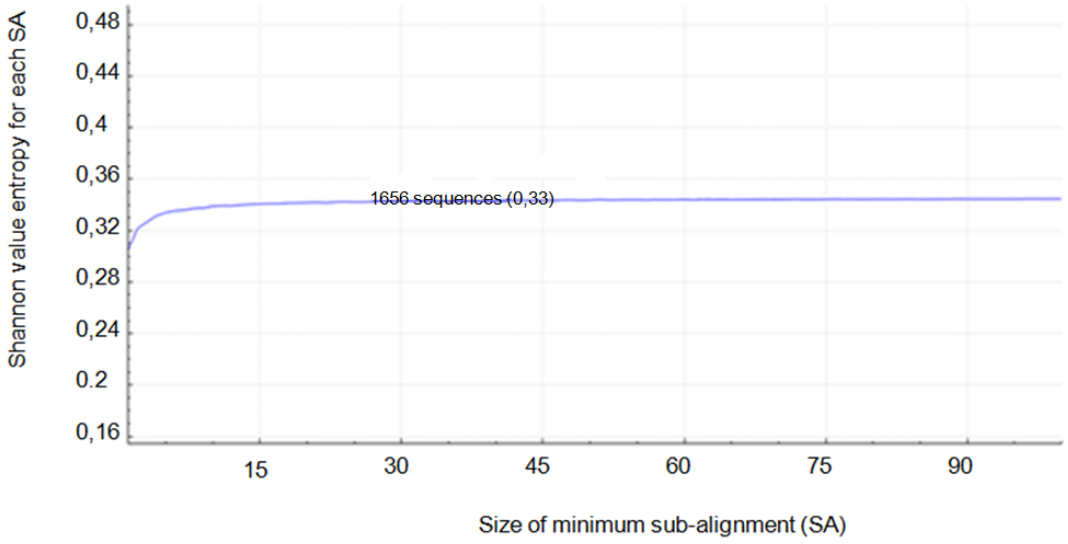

Figura 13. Média da entropia de Shannon para cada sub-alinhamento gerado a partir da remoção de sequências aleatórias. No eixo das abscissas tem-se o tamanho do sub-alinhamento mínimo e no eixo das ordenadas o valor da entropia de Shannon para cada sub-alinhamento mínimo.

5.3. Decomposição dos Resíduos em Comunidades

A partir da análise de correlação da família de cisteíno proteases foram encontrados um total de 34 resíduos correlacionados dispostos em 5 comunidades. Cada comunidade é composta por um grupo de resíduos que coevoluem, e não estão positivamente correlacionados de forma significativa com os demais resíduos das outras comunidades. Os resíduos de cada comunidade foram numerados a partir do seu posicionamento na estrutura da cruzaína, tomada como representante da família de cisteíno proteases. Portanto serão apresentados todos resíduos com a sua numeração correspondente referente à cruzaína.

A comunidade 1 é a maior comunidade, sendo composta por 22 resíduos listados em ordem crescente de posicionamento: Pro2, Asp6, Arg8, Cys22, Trp26, Phe28, Glu35, Cys56, Gly62, Cys63, Gly66, Tyr89, Cys101, Pro134, Tyr147, Gly150, Val167, Asn170, Tyr177, Trp178, Tyr193 e Cys203.

A comunidade 2 possui 6 resíduos listados em ordem crescente de posicionamento: Gln19, Gly23, Cys25, Gly65, Tyr91 e His162.

As comunidades 3, 4 e 5 são compostas por apenas 2 resíduos cada, listados em ordem crescente de posicionamento. Comunidade 3: Ser55 e Cys155; comunidade 4: Val13 e Lys181; comunidade 5: Leu48 e Gln51.

5.4. Construção da Rede de Correlações

A partir da obtenção de uma rede de correlações entre os resíduos torna-se possível a segregação dos mesmos em comunidades nas quais eles apresentam correlações entre si. Assim, os resíduos que compõem uma mesma comunidade são correlacionados, mas não se correlacionam aos resíduos que fazem parte das outras comunidades da rede. Na figura 14, representada abaixo, pode-se observar que as linhas que unem os resíduos possuem espessura diretamente proporcional ao grau de correlação entre esses resíduos. Portanto, quanto mais espessa for uma linha unindo dois resíduos, maior o grau de correlação entre esses dois resíduos.

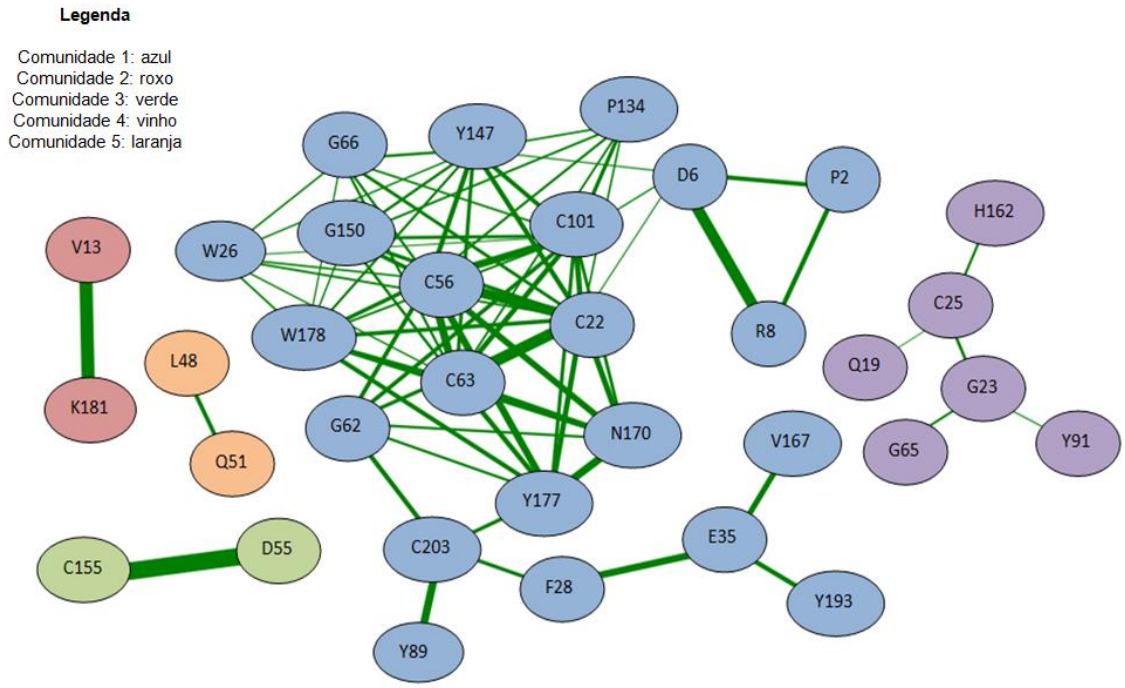

Figura 14. Rede de correlação da família de cisteíno proteases obtida a partir do alinhamento múltiplo realizado pelo PFSTATS. As linhas verdes representam a correlação entre os resíduos e a espessura das mesmas está diretamente relacionada ao grau de correlação entre esses resíduos. Após realização do alinhamento múltiplo foram apontadas cinco comunidades de resíduos que coevoluem. Resíduos da comunidade 1estão indicados em azul, resíduos da comunidade 2 em roxo, resíduos da comunidade 3 em verde, resíduos da comunidade 4 em vinho e resíduos da comunidade 5 em laranja.

5.5. Resíduos Correlacionados por Comunidade

Os resíduos correlacionados de cada comunidade foram representados na estrutura da cruzaína para uma visão geral do posicionamento dos mesmos na estrutura da proteína e dentro da sua respectiva comunidade. As figuras15a19representadas abaixo correspondem às comunidades 1 a 5, respectivamente, e representam a estrutura da cruzaína com os resíduos correlacionados de cada comunidade.

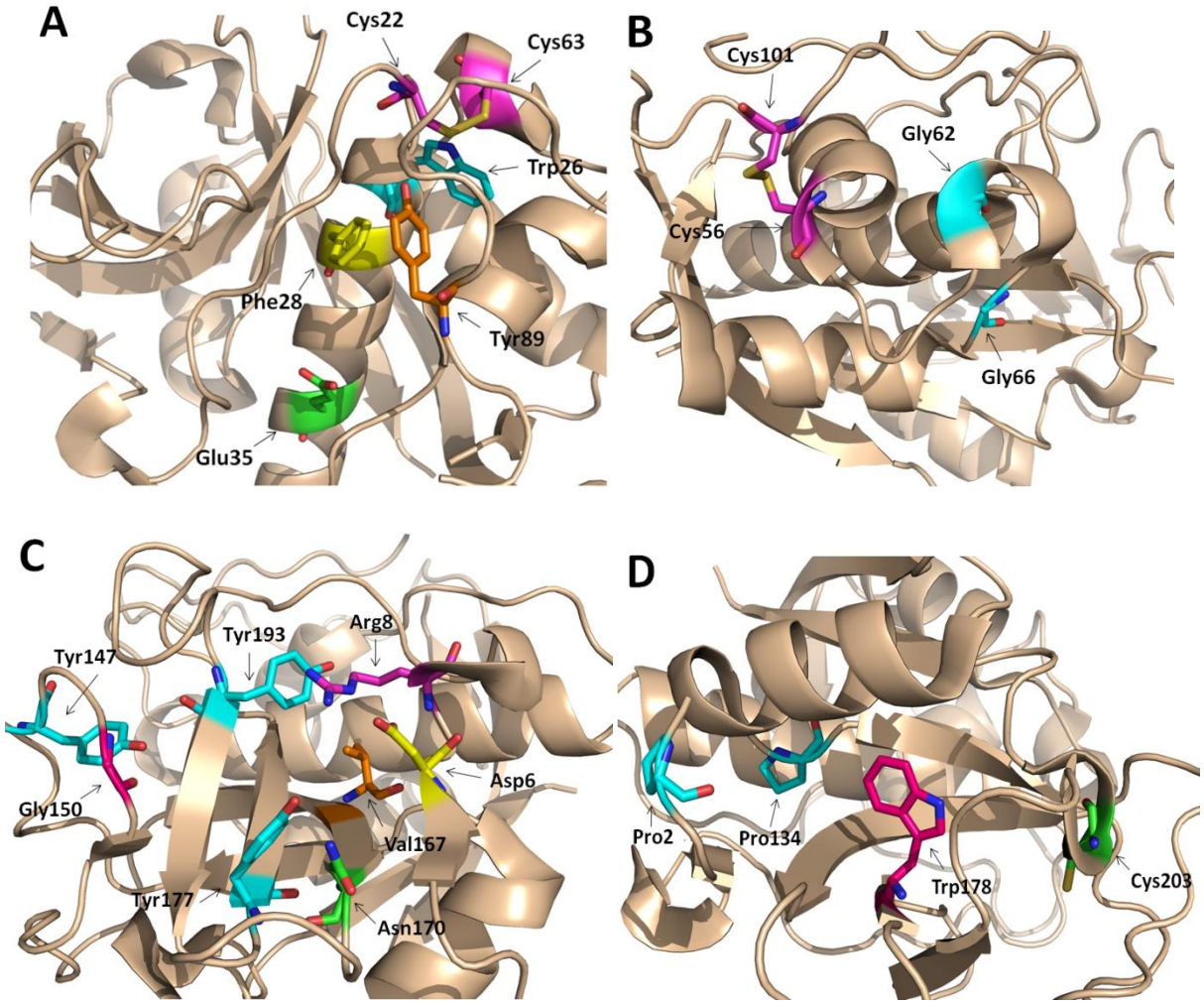

Figura 15. Resíduos correlacionados dacomunidade1 destacados na estrutura da cruzaína. Destacada a estrutura da cruzaína (PDB: 3KKU) em *cartoon* rosa claro com os resíduos representados na forma de bastões. Nitrogênio em azul escuro, oxigênio em vermelho e enxofre em amarelo. Os painéis de A a D representam todos os resíduos que integram a comunidade 1 e esses resíduos estão alocados em uma mesma figura por proximidade na estrutura. Figura preparada com o programa Pymol (Schrödinger, LCC, 2015).

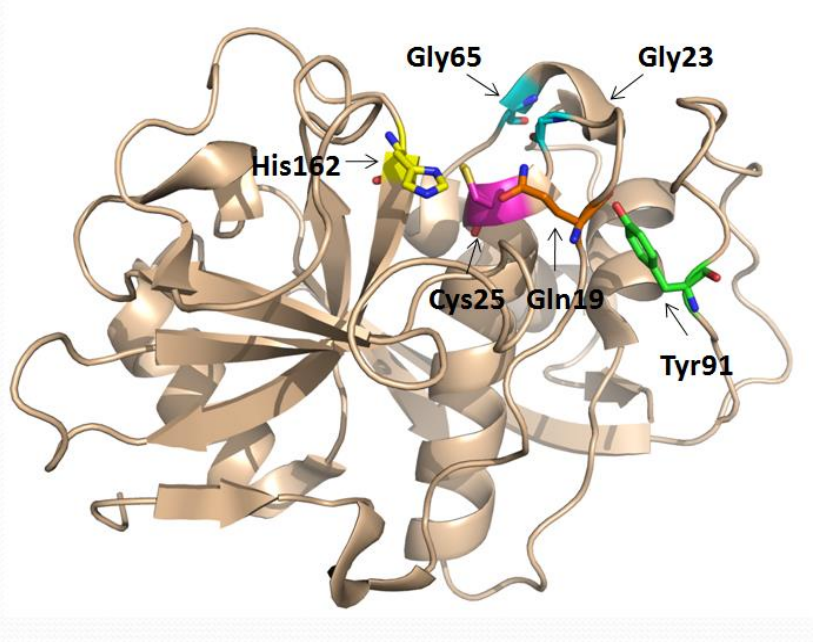

Figura 16. Resíduos correlacionados da comunidade 2 destacados na estrutura da cruzaína. Destacada a estrutura da cruzaína (PDB: 3KKU) em *cartoon* rosa claro com os resíduos representados na forma de bastões. Nitrogênio em azul escuro, oxigênio em vermelho, enxofre em amarelo. Figura preparada com o programa Pymol (Schrödinger, LCC, 2015).

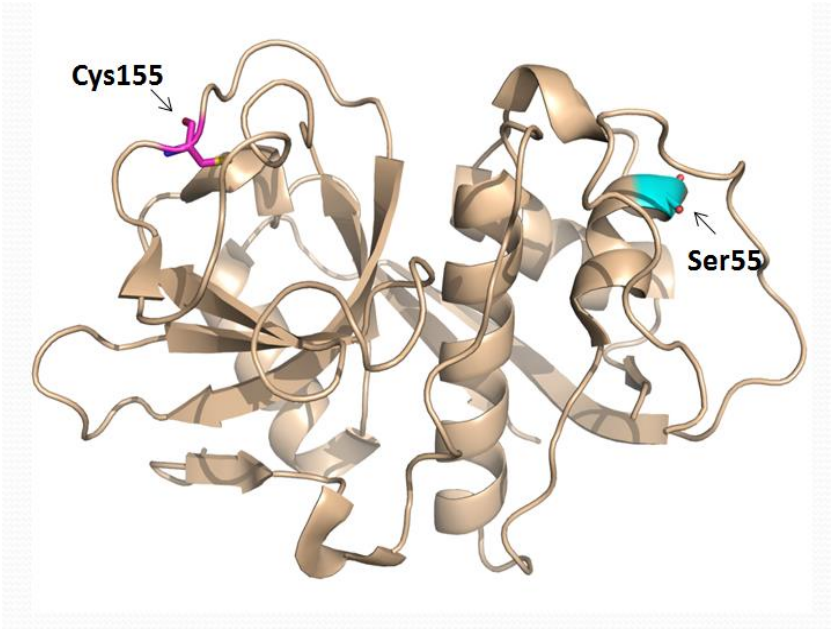

Figura 17. Resíduos correlacionados dacomunidade3 destacados na estrutura da cruzaína. Destacada a estrutura da cruzaína (PDB: 3KKU) em *cartoon* rosa claro com os resíduos representados na forma de bastões. Nitrogênio em azul escuro, oxigênio em vermelho, enxofre em amarelo. Figura preparada com o programa Pymol (Schrödinger, LCC, 2015).

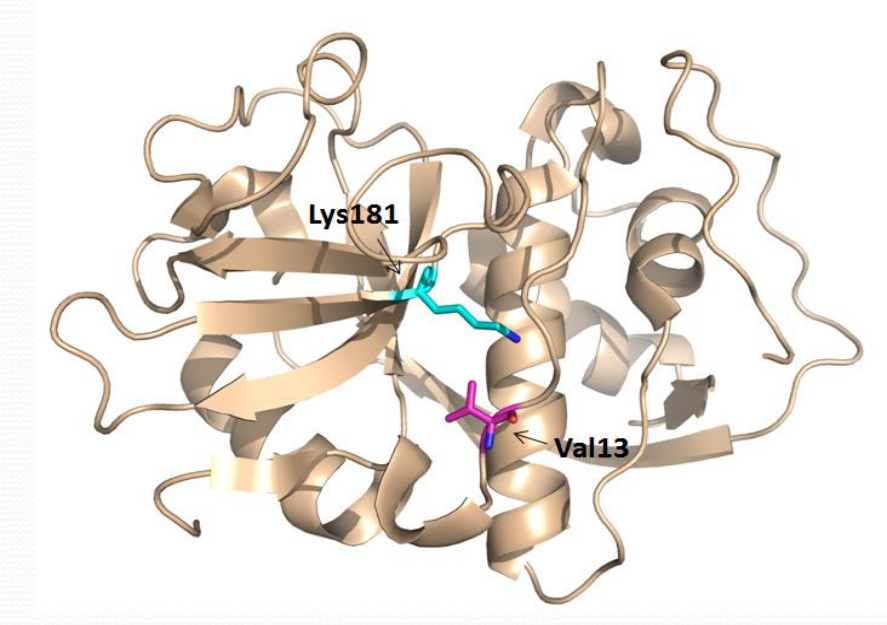

Figura 18. Resíduos correlacionados da comunidade 4 destacados na estrutura da cruzaína. Destacada a estrutura da cruzaína (PDB: 3KKU) em *cartoon* rosa claro com os resíduos representados na forma de bastões. Nitrogênio em azul escuro, oxigênio em vermelho, enxofre em amarelo. Figura preparada com o programa Pymol (Schrödinger, LCC, 2015).

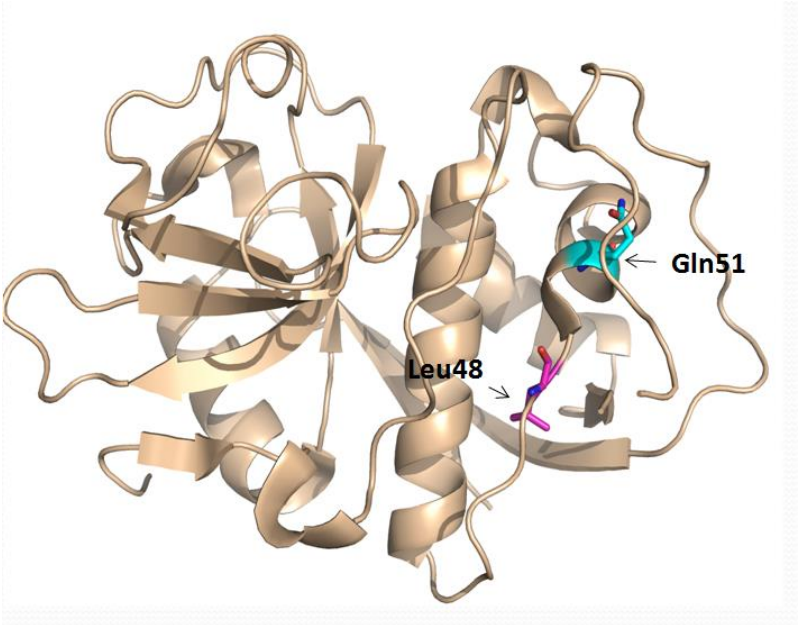

Figura 19. Resíduos correlacionados da comunidade 5 destacados na estrutura da cruzaína. Destacada a estrutura da cruzaína (PDB: 3KKU) em *cartoon* rosa claro com os resíduos representados na forma de bastões. Nitrogênio em azul escuro, oxigênio em vermelho, enxofre em amarelo. Figura preparada com o programa Pymol (Schrödinger, LCC, 2015).

5.6. Matrizes de Correlação

Outro dado importante obtido a partir da análise de correlação entre os resíduos das comunidades da família das cisteíno proteases diz respeito à matriz de correlação entre eles. A matriz de correlação evidencia a probabilidade de ocorrência entre dois resíduos em determinada posição e pode ser expressa em porcentagem. A primeira coluna (coluna "*all*") indica a porcentagem de ocorrência de determinado resíduo no alinhamento global. As colunas seguintes da matriz relacionam-se com as linhas e considerando-se um determinado resíduo X (sinalizado em uma coluna) em uma dada posição a probabilidade de ocorrência de um resíduo Y (sinalizado em uma linha) em determinada posição, é expressa em porcentagem. Os dados da matriz de correlação permitem que sejam realizadas análises de correlação entre os resíduos das comunidades e tais análises são expressas numericamente. Portanto, foram obtidas matrizes de correlação para as 5 comunidades de resíduos conservados que estão representadas nas tabelas 4 a 8.

Na matriz de correlação da comunidade 1 (Tabela 4) pode-se observar a porcentagem de ocorrência dos resíduos em todas as proteínas do alinhamento bem como a porcentagem de correlação entre todos os resíduos da comunidade. Para elucidar o significado de uma matriz de correlação será analisado o resíduo de Cys56. A Cys56 está presente em 84,77% de todas as proteínas da família e quando há um resíduo de cisteína na posição 101 a ocorrência do resíduo de cisteína na posição 56 aumenta de 84,77% para 95,36%. Ou seja, a ocorrência de um determinado resíduo em uma posição específica aumenta a ocorrência de um outro resíduo em outra posição também específica. O mesmo tipo de análise pode ser realizado para quaisquer outros resíduos que fazem parte da matriz de correlação. A partir da análise dos dados dessa matriz pode-se dizer que os índices de correlação entre os resíduos da comunidade 1 são altos, acima de 65% e majoritariamente em torno de 70 a 80%, sendo que entre alguns resíduos são valores ainda mais altos, em torno de 90%.

	All	C203	F28	E35	V167	Y193	G62	C56	C63	C101	C22	D6	P2	R8	Y147	G150	P134	W178	N170	Y177	G66	W26	Y89
C203	69,96	0	80,11	78,56	73,15	78,43	81,92	79,74	78,56	78,96	79,33	74,99	72,87	75,52	78,3	76,36	75,13	76,47	79,19	80,36	75,89	76,72	87,47
F28	73,28	83,92	0	88,61	79,87	80,09	78,37	78,05	76,72	78,32	78,18	76,49	74,56	77,53	77,11	75,24	72,55	76,57	75,05	76,63	77,01	78,53	82,65
E35	67,87	76,21	82,07	0	81,91	80,58	70,56	68,69	67,95	68,4	68,81	71,51	67,69	71,98	69,6	68,15	66,79	68,05	64,41	67	68,91	70,2	77,77
V167	69,44	72,61	75,69	83,8	0	80,64	71,08	69,14	68,86	69,92	69,07	70,42	68,89	68,68	70,07	69,34	69,84	67,78	67,01	67,57	69,69	68,92	72,91
Y193	68,39	76,67	74,74	81,2	79,42	0	73,86	69	68,31	70,28	68,29	71,37	70,58	72,11	70,24	68,98	70,25	67,97	67,45	69,08	69,81	69,75	76,46
G62	72,25	84,6	77,27	75,1	73,95	78,02	0	80,84	79,78	80,86	80,08	77,07	76,46	76,24	79,57	77,62	78,48	76,81	80,53	80,71	77,35	75,27	86,87
C56	84,77	96,62	90,29	85,79	84,4	85,52	94,85	0	95,15	95,36	96,66	88,15	87,96	88,58	93,43	91,83	90,68	93,14	96,6	97,28	90,58	90,42	96,51
C63	86,52	97,16	90,59	86,61	85,8	86,42	95,54	97,12	0	95,36	98,63	88,44	88,56	88,76	93,87	92,3	91,8	93,54	96,85	96,68	92,27	91,45	96,06
C101	83,62	94,37	87,09	84,27	84,2	85,93	93,59	94,07	92,16	0	93,01	88,39	87,71	88,6	91,79	89,83	91,29	89,67	92,65	93,09	88,66	87,76	95,18
C22	84,45	95,77	90,1	85,62	84	84,33	93,62	96,29	96,28	93,94	0	88,59	88,53	89,22	93,03	91,13	89,57	92,87	96	95,78	91,09	90,54	94,97
D6	82,21	88,12	85,81	86,61	83,37	85,78	87,7	85,48	84,03	86,9	86,23	0	92,48	98,07	86,37	84,92	85,38	83,96	84,86	84,99	84,85	85,58	89,02
P2	72,98	76,02	74,25	72,79	72,4	75,32	77,24	75,73	74,7	76,56	76,51	82,1	0	84,63	76,6	74,86	76,48	77,45	77,76	77,17	76,18	75,7	76,77
R8	74,14	80,02	78,43	78,62	73,32	78,17	78,23	77,46	76,06	78,55	78,32	88,44	85,97	0	78,05	76,7	76,95	78,43	78,28	79,4	77,12	78,23	81,1
Y147	85,03	95,17	89,47	87,2	85,8	87,33	93,64	93,71	92,26	93,34	93,67	89,33	89,24	89,51	0	90,75	90,12	91,43	93,64	93,03	90,86	90,29	95,21
G150	88,57	96,68	90,94	88,93	88,43	89,33	95,16	95,94	94,49	95,15	95,57	91,49	90,85	91,63	94,53	0	93,75	93,32	96,27	96,21	91,04	90,74	95,51
P134	85,55	91,87	84,7	84,18	86,03	87,88	92,93	91,51	90,76	93,39	90,72	88,85	89,65	88,79	90,67	90,55	0	88,15	91,28	91,6	87,75	86,15	92,84
W178	81,19	88,75	84,83	81,4	79,24	80,7	86,32	89,21	87,78	87,07	89,29	82,93	86,16	85,89	87,3	85,54	83,66	0	92,27	91,91	85,62	87,38	87,44
N170	72,5	82,07	74,25	68,8	69,97	71,51	80,82	82,62	81,16	80,34	82,42	74,85	77,25	76,56	79,85	78,81	77,36	82,39	0	87,6	77,21	78,71	81,68
Y177	72,47	83,23	75,77	71,53	70,51	73,2	80,96	83,16	80,97	80,67	82,18	74,92	76,63	77,61	79,28	78,72	77,6	82,03	87,55	0	76,8	78,3	82,98
G66	86,98	94,35	91,4	88,31	87,29	88,78	93,12	92,94	92,76	92,23	93,81	89,77	90,79	90,48	92,94	89,41	89,22	91,72	92,62	92,18	0	92,2	93,51
W26	79,26	86,93	84,94	81,99	78,67	80,84	82,58	84,54	83,78	83,19	84,98	82,52	82,21	83,64	84,17	81,21	79,83	85,31	86,04	85,65	84,02	0	86,14
Y89	65,55	81,96	73,93	75,1	68,82	73,28	78,81	74,62	72,77	74,61	73,71	70,98	68,95	71,71	73,39	70,68	71,14	70,59	73,84	75,06	70,47	71,23	0

Tabela 4. Matriz de correlação dos resíduos da comunidade 1 gerada a partir do alinhamento múltiplo das cisteíno proteases. Os dados são apresentados em porcentagem e na tabela constam todos os resíduos que coevoluem na comunidade 1. Tabela gerada pelo PFSTATS como parte dos resultados obtidos a partir do alinhamento da família de cisteíno proteases.

Na matriz de correlação da comunidade 2(tabela 5) pode-se observar que a Cys25, por exemplo, está presente em 89,11% de todas as sequências da família, enquanto na presença de Gly23 a porcentagem de ocorrência da Cys25 aumenta de 89,11% para 95,56%. A partir da análise dos dados dessa matriz pode-se dizer que os índices de correlação entre os resíduos da comunidade são bem altos, com valores acima de 84% e majoritariamente acima de 90%.

	All	**C25**	**G23**	**G65**	**Y91**	**H162**	**Q19**
C25	89,11	0	95,56	91,93	90,94	92,58	91,29
G23	82,84	88,84	0	88	86,38	85,5	84,82
G65	88,49	91,3	94	0	90,67	90,02	89,25
Y91	87,75	89,56	91,5	89,91	0	88,36	88,33
H162	95,45	99,17	98,51	97,1	96,1	0	96,99
Q19	95,21	97,55	97,48	96,02	95,83	96,75	0

Tabela 5. Matriz de correlação dos resíduos da comunidade 2 gerada a partir do alinhamento múltiplo das cisteíno proteases. Os dados são apresentados em porcentagem e na tabela constam todos os resíduos que coevoluem e se correlacionam na comunidade 2. Tabela gerada pelo PFSTATS como parte dos resultados obtidos a partir do alinhamento da família de cisteíno proteases.

No caso da comunidade 3 (Tabela 6) tem-se que a porcentagem de ocorrência tanto da Cys155 quanto da Ser55 em todas as proteínas do alinhamento é em torno de 50%, um valor mais baixo quando se compara com a porcentagem de ocorrência dos resíduos em todas as proteínas do alinhamento das comunidades 1 e 2 (que fica em torno de 70 a 80%, podendo chegar em alguns casos a valores em torno de 90%). Uma outra observação importante a se fazer é que dentre as proteínas referência neste estudo apenas a cruzaína e a catepsina C possuem na posição 55 ou na posição correspondente um resíduo de serina. A catepsina B possui uma treonina e as demais proteínas que são referência neste estudo possuem um aspartato em tal posição. A incidência de cada resíduo na presença do

outro entre os dois resíduos da comunidade 3 apresenta percentuais altos, acima de 80%, ainda que individualmente eles estejam presentes pouco acima de 50%.

	All	C155	S55
C155	55,03	0	86,1
S55	52,35	81,9	0

Tabela 6. Matriz de correlação dos resíduos da comunidade 3 gerada a partir do alinhamento múltiplo das cisteíno proteases. Os dados são apresentados em porcentagem e na tabela constam todos os resíduos que coevoluem na comunidade 3. Tabela gerada pelo PFSTATS como parte dos resultados obtidos a partir do alinhamento da família de cisteíno proteases.

Na comunidade 4 a porcentagem de ocorrência tanto da Lys181 quanto da Val13 em todas as proteínas do alinhamento também é em torno de 50%, como observado para a comunidade 3. Os índices de correlação entre os dois resíduos da comunidade 4 também são valores altos, em torno de 80%(Tabela 7).

	All	K181	V13
K181	52,03	0	80,03
V13	52,07	80,09	0

Tabela 7. Matriz de correlação dos resíduos da comunidade 4 gerada a partir do alinhamento múltiplo das cisteíno proteases. Os dados são apresentados em porcentagem e na tabela constam todos os resíduos que coevoluem na comunidade 4. Tabela gerada pelo PFSTATS como parte dos resultados obtidos a partir do alinhamento da família de cisteíno proteases.

Finalmente, na comunidade 5 a porcentagem de ocorrência tanto da Leu48 quanto da Gln51 em todas as proteínas do alinhamento é em torno de 68%, um valor mais alto que o observado nas comunidades 3 e 4 mas ainda menor quando se

compara com a porcentagem de ocorrência dos resíduos em todas as proteínas do alinhamento das comunidades 1 e 2. Assim como nas demais comunidades descritas anteriormente, os índices de correlação entre os dois resíduos da comunidade 5 são valores altos, em torno de 80% (Tabela 8).

	All	L48	Q51
L48	68,19	0	80,25
Q51	68,75	80,9	0

Tabela 8. Matriz de correlação dos resíduos da comunidade 5 gerada a partir do alinhamento múltiplo das cisteíno proteases. Os dados são apresentados em porcentagem e na tabela constam todos os resíduos que coevoluem na comunidade 5. Tabela gerada pelo PFSTATS como parte dos resultados obtidos a partir do alinhamento da família de cisteíno proteases.

5.7 Resíduos Conservados

Os resíduos conservados da família das cisteíno proteases foram obtidos a partir da determinação de um *cutoff* mínimo de conservação de 80% e estão descritos na tabela 9, representada abaixo com a numeração dos resíduos da cruzaína e com os respectivos índices de ocorrência em toda a família de cisteíno proteases. Foram encontrados 10 resíduos que são conservados evolutivamente nesta família. Assim, o resíduo Gln19, por exemplo, é encontrado em 87% de todas as proteínas do alinhamento.

Resíduos conservados										
	Q19	**C25**	**S49**	**H162**	**G168**	**N182**	**S183**	**W184**	**G189**	**G192**
Índice de conservação	87%	84%	84%	89%	94%	92%	86%	83%	85%	91%

Tabela 9. Tabela de resíduos conservados gerada a partir do alinhamento múltiplo de cisteíno proteases. Os índices de conservação de cada resíduo são apresentados em porcentagem e na tabela constam todos os resíduos conservados na família de cisteíno proteases. A numeração dos resíduos apresentada é a da cruzaína, que nesse estudo é a representante da família de cisteíno proteases. A numeração dos resíduos das demais proteínas dessa família são correspondentes às da cruzaína. Tabela gerada pelo PFSTATS como parte dos resultados obtidos a partir do alinhamento da família de cisteíno proteases.

A figura 20 representa todos os resíduos conservados da família de cisteíno proteases na estrutura da cruzaína, para uma visão geral do posicionamento dos mesmos.

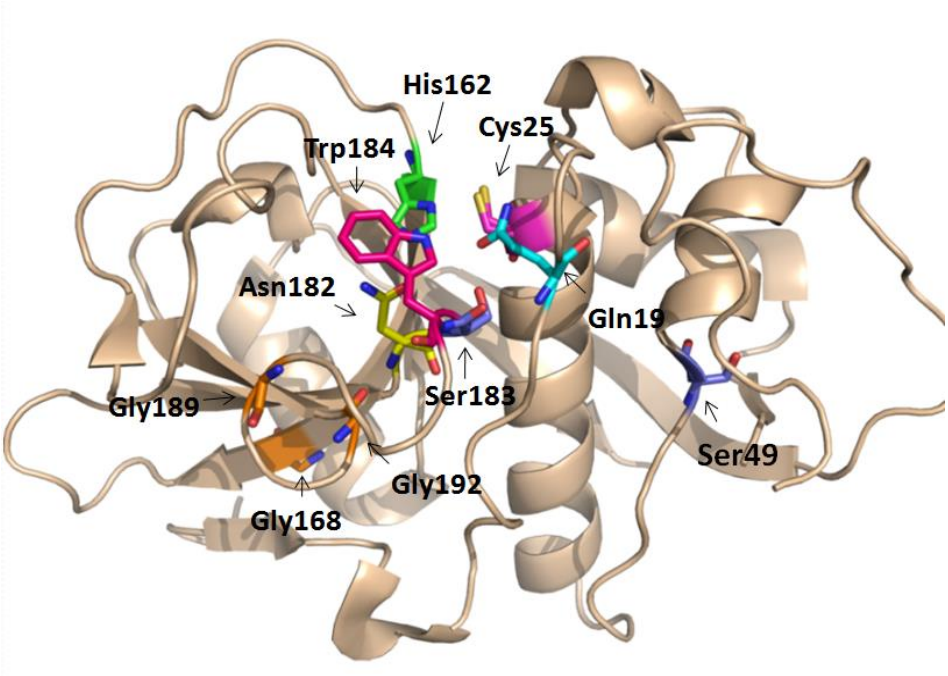

Figura 20. Resíduos conservados da família de cisteíno proteases destacados na estrutura da cruzaína. Destacada a estrutura da cruzaína (PDB: 3KKU) em *cartoon* rosa claro. Resíduos conservados representados na forma de bastões. Nitrogênio em azul escuro, oxigênio em vermelho, enxofre em amarelo. Figura preparada com o programa Pymol (Schrödinger, LCC, 2015).

5.8 Diagrama de Ramachandran

Para definir os resíduos que se encontram em posições energeticamente favoráveis e analisar os ângulos φ (phi) e ψ (psi) de alguns resíduos dentro da sequência tais como prolina e glicina, foi gerado o Diagrama de Ramachandran para a cruzaína através do servidor MolProbity. Na figura 21 observam-se todos os resíduos da cruzaína, todas as glicinas e todas as prolinas nas conformações *cis* e *trans*. Pode-se observar que 97,8% (222 dos 227) dos resíduos estão localizados em regiões favoráveis do diagrama.

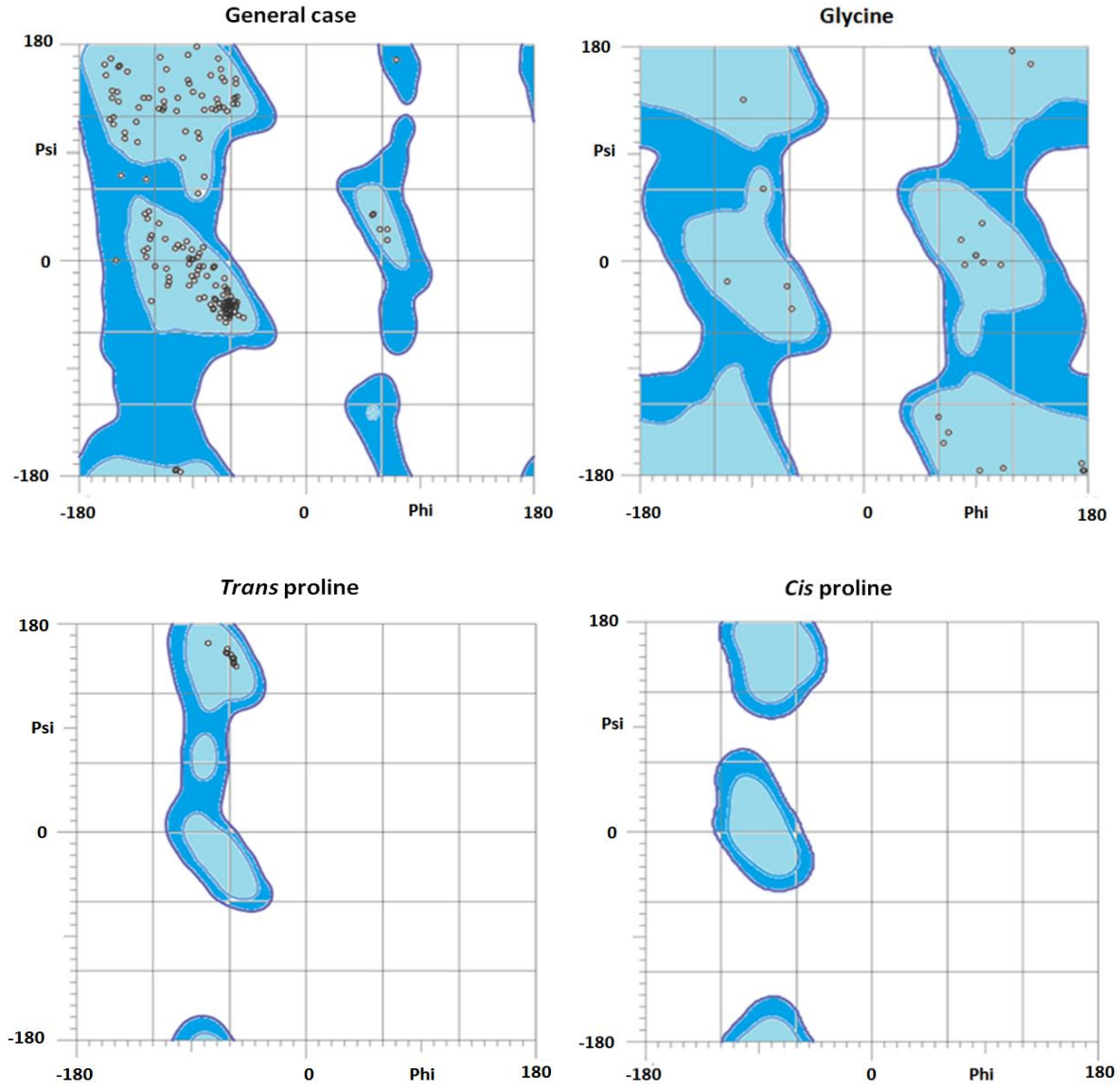

Figura 21. Diagrama de Ramachandran para a estrutura da cruzaína. Representados os ângulos φ (phi) e ψ (psi) de todos os resíduos da cruzaína no canto superior esquerdo, para todos os resíduos de glicina no canto superior direito, para todos os resíduos de prolina na conformação *trans* no canto inferior esquerdo e para todos os resíduos de prolina na conformação *cis* no canto inferior direito. As regiões favorecidas estão representadas em azul claro, as regiões permitidas, porém com algum impedimento energético estão representadas em azul escuro e as regiões não permitidas estão representadas em branco. Diagrama gerado pelo servidor MolProbity.

6. Discussão

6.1 Comunidade 1

A comunidade 1 é a maior comunidade de resíduos correlacionados identificada neste estudo, sendo composta por 22 resíduos. Dentre eles encontram-se cinco resíduos de cisteína: Cys22, Cys56, Cys63, Cys101 e Cys203, destacados na figura 22 abaixo. Todos eles apresentam ocorrência mediana ou elevada na família, sendo que Cys203 é o resíduo encontrado com menor frequência. Assim, Cys22 é encontrada em 84,45% das sequências, Cys56 em 84,77%, Cys63 em 86,52%, Cys101 em 83,62% e Cys203 em 69,96%. Cys22 forma uma ponte dissulfeto com Cys63 e está a uma distância de 8Å de Cys25, e ambos os resíduos, Cys22 e Cys63, são classificados como resíduos proximais ao sítio ativo. Cys56 forma uma ponte dissulfeto com Cys101 e encontra-se a uma distância de mais de 15Å de Cys25. Cys56 e Cys101 são classificados como resíduos remotos, devido à sua maior distância em relação a Cys25 (LEE, G. M., et al, 2012). Cys22 e Cys101 se encontram em alças separadas e Cys56 e Cys63 em pequenas α-hélices distintas. O resíduo Cys203 realiza ponte dissulfeto com o resíduo Cys155, sendo que este último integra a comunidade 3. Cys 203 é encontrado em uma pequena α-hélice e é classificado como resíduo remoto por estar a uma distância de mais de 15Å de Cys25 (LEE, G. M., et al, 2012).

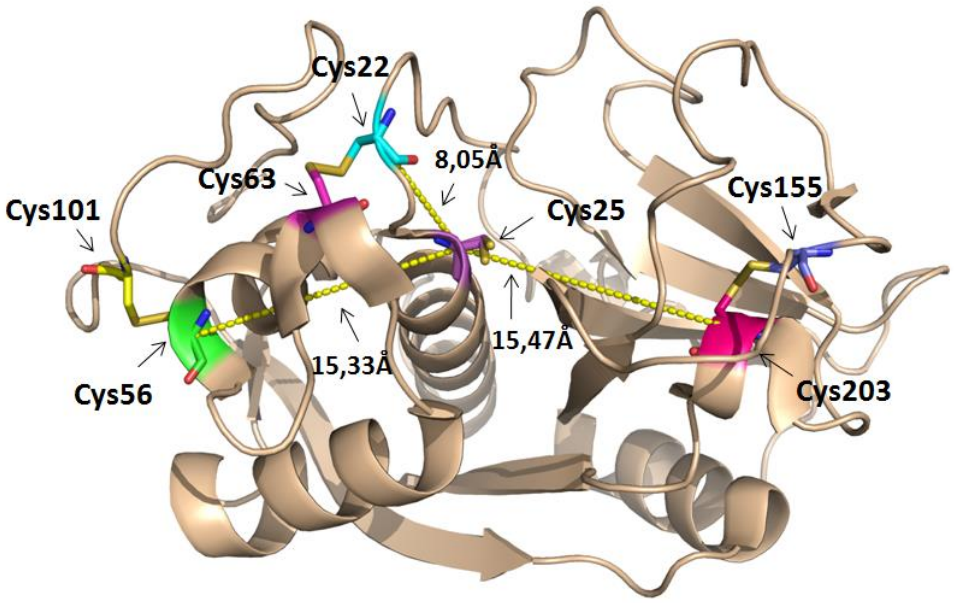

Figura 22. Representação da estrutura da cruzaína destacando os resíduos de cisteína que fazem parte da comunidade 1 e suas distâncias em relação à Cys25. Destacada a estrutura da cruzaína (PDB: 3KKU) em *cartoon* rosa claro. Os resíduos estão representados na forma de bastões coloridos,nitrogênio em azul escuro e oxigênio em vermelho. O resíduo Cys22 realiza uma ponte dissulfeto com o resíduo Cys63, o resíduo Cys56 realiza uma ponte dissulfeto com Cys101 e o resíduo Cys203 realiza uma ponte dissulfeto com Cys155 (sendo que esse último faz parte da comunidade 3). Representado o resíduo de Cys25, que faz parte da díade catalítica, e destacadas em linhas tracejadas amarelas as distâncias entre as cisteínas da comunidade 1 e a Cys25. Figura preparada com o programa Pymol (Schrödinger, LCC, 2015).

Coerente com as ligações dissulfeto descritas acima, a análise de correlação entre os resíduos de cisteína dessa comunidade apontam correlações altas entre os resíduos Cys22, Cys56, Cys63 e Cys101. Considerando-se, por exemplo, o resíduo Cys na posição 63 a sua ocorrência aumenta de 86,52% para 98,63% quando há uma cisteína na posição 22. Da mesma forma, observa-se um aumento na ocorrência da Cys56, de 84,77% para 95,36%, quando há um resíduo de cisteína na posição 101. Assim, se verifica correlações muito fortes entre os resíduos que formam as pontes dissulfeto. Com relação ao resíduo Cys203 a correlação deste com as demais cisteínas também é alta, em torno de 78%. Portanto, de forma geral observa-se que a ocorrência de cada cisteína da comunidade 1 nas proteínas da família das cisteíno proteases sofre aumento na presença dos demais resíduos de cisteína dessa comunidade.

Na estrutura da cruzaína são encontrados dois conjuntos de resíduos com ocorrência elevada, situados no lado da proteína em frente ao seu sítio ativo. O primeiro conjunto é composto pelos resíduos Tyr91, Pro90, Glu86, Tyr89, Gln51 e Ser49 e o segundo conjunto é composto pelos resíduos Tyr193, Arg8, Val16, Gly11, Asp6 e Val13. Ambos os grupos de resíduos estão localizados em um longo sulco raso, formado principalmente por várias alças desordenadas que atravessam a superfície da proteína. Apesar dessas alças serem desordenadas, elas são mantidas em uma posição rígida pela estabilização dos resíduos que compõe os dois conjuntos citados (DURRANT, J. D., et al, 2010). A figura 23, representada abaixo,ilustra esses dois grupos de resíduos destacados na estrutura da cruzaína.

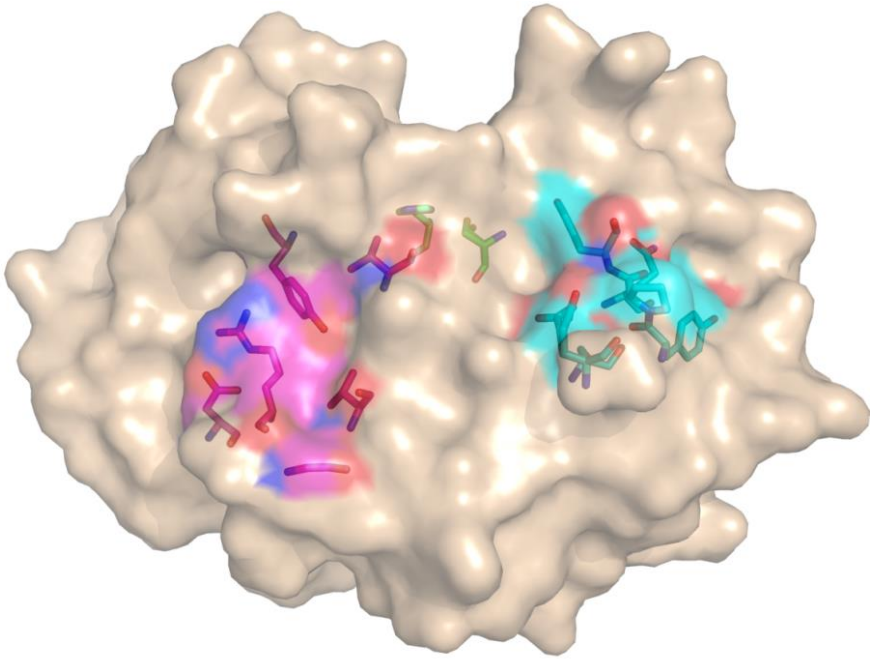

Figura 23. Representação de dois conjuntos de resíduos importantes para estabilização de alças na estrutura da cruzaína. Destacada a estrutura da cruzaína (PDB: 3KKU) em superfície transparente rosa claro como uma enzima representante da família de cisteíno proteases. Os resíduos estão representados na forma de bastões coloridos de acordo com o tipo atômico (Carbono colorido de acordo com a legenda determinada para cada grupo, hidrogênio em branco, nitrogênio em azul escuro, oxigênio em vermelho, enxofre em amarelo). Primeiro grupo destacado em ciano e composto pelos resíduos: Tyr91, Pro90, Glu86, Tyr89, Gln51, Ser49. Segundo grupo destacado em magenta e composto pelos resíduos Tyr193, Arg8, Val16, Gly11, Asp6, Val13. Ambos os grupos possuem papel importante para estabilização de várias alças que atravessam a estrutura da cruzaína. Resíduos que compõem a díade catalítica, Cys25 e His162, foram destacados em verde. Figura preparada com o programa Pymol (Schrödinger, LCC, 2015).

Os resíduos Asp6 e Arg8 integrantes do segundo conjunto e da comunidade 1 são importantes para manter a rigidez e organização das alças desordenadas que atravessam a estrutura da proteína. O resíduo Arg8 realiza ligações de hidrogênio com o resíduo Gly11 e ambos ancoram parte da alça formada pelos resíduos de 11 a 23. Uma pequena α-hélice formada pelos resíduos 7 a 10, incluindo Arg8, é mantida em sua posição adequada mediante múltiplas ligações de hidrogênio entre os resíduos Asp6 e Arg8 (DURRANT, J. D., et al, 2010). A figura 24, representada abaixo, ilustra a longa alça que é formada pelos resíduos 11 a 23 e também as interações que ocorrem entre os resíduos Asp6, Arg8 e Gly11. Provavelmente tais interações entre esses resíduos são críticas para manutenção da estrutura adequada da proteína e para sua função. Ademais, Asp6 é análogo ao resíduo Asp236 da catepsina C e pacientes com mutações Asp236Tyr desenvolvem a Síndrome Papillon-Lefèvre, o que sugere disfunção da catepsina C (ALLENDE, L. M., 2001).Os resíduos Asp6 e Arg8 possuem ocorrência elevada e mediana na família, sendo encontrados em 82,21 e 74,14% das cisteíno proteases. As interações observadas entre estes resíduos sugerem correlação por questões estruturais. Além das interações descritas anteriormente entre ambos os resíduos, pode-se observar a partir da figura 24 a ocorrência de uma ponte salina entre eles.

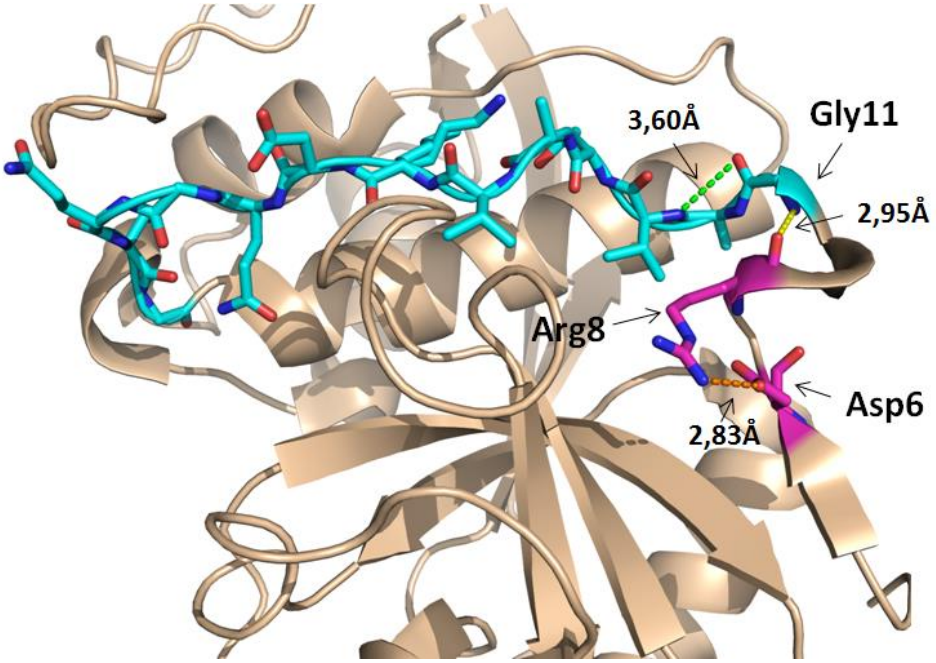

Figura 24. Representação de parte da estrutura da cruzaína destacando interações polares entre resíduos que auxiliam no ancoramento da alça contendo os resíduos 11-23. Destacada a estrutura da cruzaína (PDB: 3KKU) em *cartoon* rosa claro. Os resíduos estão representados na forma de bastões coloridos de acordo com o tipo atômico (Carbono colorido de acordo com a legenda determinada para cada grupo, nitrogênio em azul escuro, oxigênio em vermelho). O resíduo Arg8 realiza ligação de hidrogênio com o resíduo Gly11 e uma ponte salina com o resíduo Asp6. Esses resíduos auxiliam no ancoramento de uma longa alça formada pelos resíduos das posições 11 a 23 (destacada com carbonos em ciano). A linha tracejada em amarelo representa uma ligação de hidrogênio, a linha tracejada em laranja representa uma ponte salina e a linha tracejada em verde representa uma interação entre resíduos que auxilia no ancoramento da longa alça. Figura preparada com o programa Pymol (Schrödinger, LCC, 2015).

Os resíduos Phe28 e Glu35 fazem parte de uma mesma α-hélice longa e apresentam ocorrência mediana na família, sendo encontrados em 73,28 e 67,87% de todas as cisteíno proteases, respectivamente. Ambos são descritos como integrantes do terceiro sítio alostérico predito computacionalmente para a cruzaína. O resíduo Glu35 forma uma ponte salina com o resíduo Lys17, conforme demonstrado na figura 25 abaixo, e ambos atuam como um “portão funcional” da cavidade interna do terceiro sítio alostérico predito para a cruzaína. Cumpre destacar que a forma mais fechada do terceiro sítio é a mais prevalente nas estruturas cristalinas descritas para a cruzaína, já que geralmente Lys17 e Glu35 encontram-se a uma distância ideal para formação da ponte salina (cerca de 3,5Å

em estruturas cristalinas) (ALVAREZ, L. H., 2017). A correlação entre Phe28 e Glu35 é alta e devido ao fato de ambos fazerem parte do terceiro sítio predito presume-se que estejam envolvidos na regulação alostérica da enzima (DURRANT, J. D., et al, 2010). Tal envolvimento poderia explicar os altos valores de correlação encontrados entre eles.

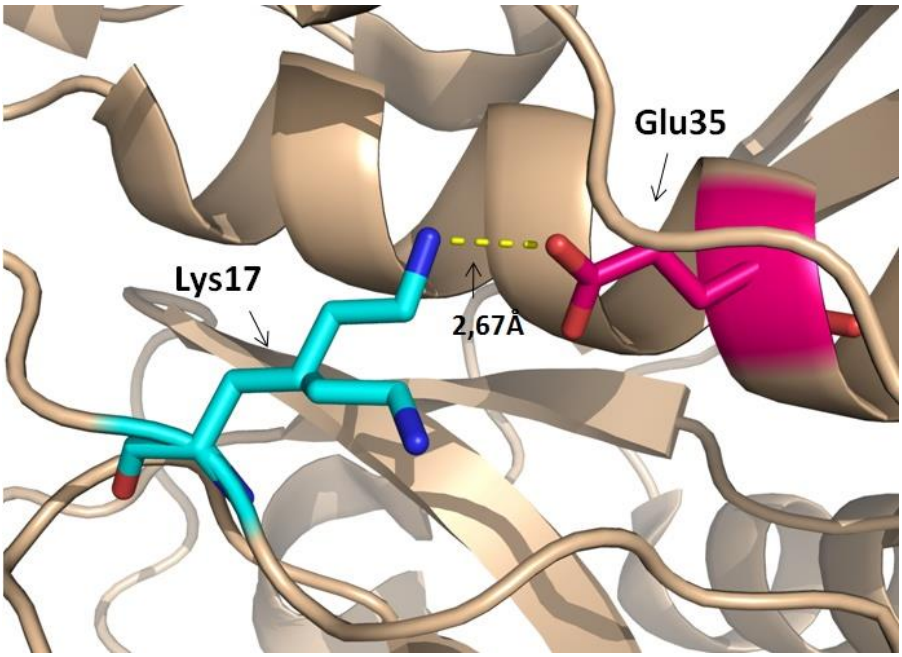

Figura 25. Representação de parte da estrutura da cruzaína destacando uma interação entre os resíduos Glu35 e Lys17. Destacada a estrutura da cruzaína (PDB: 3KKU) em *cartoon* rosa claro. Os resíduos estão representados na forma de bastões, nitrogênio em azul escuro e oxigênio em vermelho. O resíduo Glu35 realiza uma ponte salina com o resíduo Lys17 e ambos atuam como um "portão funcional" da cavidade interna do terceiro sítio alostérico predito para a cruzaína.A linha tracejada em amarelo representa a ponte salina entre os átomos dos resíduos envolvidos nessa interação. Figura preparada com o programa Pymol (Schrödinger, LCC, 2015).

Os resíduos Tyr147 e Gly150 ocupam uma extensa alça que faz parte do emaranhado de alças que atravessam a superfície da proteína. Ambos possuem ocorrência elevada na família, sendo encontrados em 85,03 e 88,57% de todas as cisteíno proteases. Não foram encontrados dados na literatura acerca do resíduo Tyr147 da cruzaína, no entanto o resíduo correspondente da catepsina K (Tyr145)

parece ser importante para interação da enzima com a condroitina-4-sulfato (LI, Z. et al, 2008).

Também não foram encontrados na literatura dados acerca do resíduo Gly150 da cruzaína, porém o resíduo correspondente da pró-catepsina K (Gly148), forma uma alça composta pelos resíduos Ser138 a Asn156. Ademais Gly148 forma a quinta margem da maior fita β do domínio C-terminal, juntamente com Tyr150 (SIVARAMAN, J., et al, 1999). A análise de correlação entre Tyr147 e Gly150 é alta e o fato de tais resíduos estarem conjuntamente envolvidos em questões estruturais corrobora os altos valores de correlação encontrados.

Os resíduos Gly62 e Gly66 apresentam ocorrência mediana e elevada na família, respectivamente. Gly62 é encontrada em 72,25% de todas as sequências enquanto Gly66 é encontrada em 86,98%. Gly62 encontra-se em uma pequena α-hélice e Gly66 em uma pequena alça que liga duas α-hélices. O resíduo Gly65 da catepsina S, correspondente a Gly62 da cruzaína, forma juntamente com átomos da cadeia principal de Asn64, um sulco na parede do domínio esquerdo do sítio de ligação S1 desta proteína (MCGRATH, M. E., et al, 1998). Por outro lado, Gly66 possui participação na realização de ligações de hidrogênio com o arcabouço de alguns inibidores análogos de hidroximetil cetona da cruzaína, contribuindo assim para sua estabilização (HUANG, L., BRINEN, L. S., ELLMAN, J. A., 2003). A análise de correlação dos resíduos Gly62 e Gly66 evidencia valores medianos e elevados de correlação entre esses resíduos e os demais resíduos da comunidade 1.

A análise do Diagrama de Ramachandran para os resíduos de glicina da comunidade 1 da cruzaína mostra que todos eles, Gly62, Gly66 e Gly150 se encontram em regiões permitidas, conforme demonstrado na figura 26 abaixo. Os resíduos Gly62 e Gly66 se encontram em regiões que são também permitidas para os outros tipos de resíduos. Por outro lado, o resíduo Gly150, na estrutura secundária localiza-se em uma alça, e encontra-se em uma região não permitida para os outros tipos de resíduos (φ = +93,7°, ψ = - 176°).Desta forma,um outro resíduo na posição 150 não poderia assumir os mesmos ângulos torcionais, causando alteração conformacional na proteína.

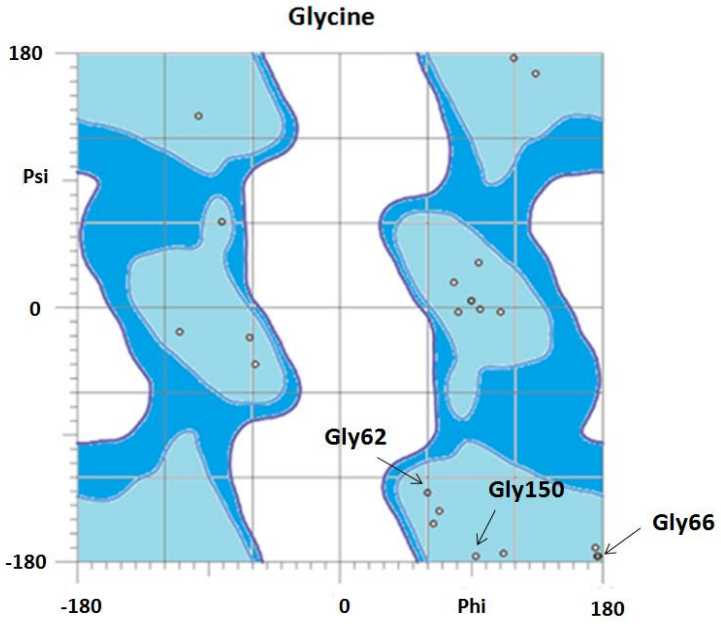

Figura 26. Diagrama de Ramachandran para a estrutura da cruzaína destacando as glicinas que fazem parte da comunidade 1. Representados os ângulos φ (phi) e ψ (psi) de todos os resíduos de glicina. As regiões favorecidas estão representadas em azul claro, as regiões permitidas, porém com algum impedimento energético estão representadas em azul escuro e as regiões não permitidas estão representadas em branco. Foram indicadas as glicinas da comunidade 1com suas respectivas posições: Gly62(φ = +60,7°, ψ = -131,8°), Gly66 (φ = +176,4°, ψ = -177°) e Gly150 (φ = +93,7°, ψ = -176°). Diagrama gerado pelo servidor MolProbity.

O resíduo Tyr89 faz parte das alças desordenadas que atravessam a superfície da proteína e que são mantidas em suas posições rígidas por dois conjuntos de resíduos. Conforme mencionado anteriormente, o primeiro conjunto é formado pelos resíduosTyr91, Pro90, Glu86, Tyr89, Gln51 e Ser49 e o segundo conjunto é formado pelos resíduos Tyr193, Arg8, Val16, Gly11, Asp6 e Val13. Esses dois conjuntos de resíduos possuem importante função de estabilização na estrutura da proteína (DURRANT, J. D., et al, 2010). Tyr89 apresenta ocorrência mediana, sendo encontrada em 65,55% das cisteíno proteases. Tyr89 integra uma alça formada pelos resíduos 88 a 109 e também integra o primeiro e o quarto sítios alostéricos preditos computacionalmente para a cruzaína (ALVAREZ, L. H., 2017).

Os resíduos Tyr177 e Trp178 apresentam ocorrência mediana e elevada na família das cisteíno proteases, respectivamente, sendo encontrados em 72,47% e 81,19% das sequências. A análise da localização dos resíduos Tyr177 e Trp178 na

estrutura da cruzaína através do Pymol mostra que ambos fazem parte de uma fita β. Não foram encontradas informações adicionais na literatura acerca dos resíduos Tyr177 e Trp178 da cruzaína, tampouco acerca dos resíduos correspondentes nas proteínas que foram referência neste estudo.

O resíduo Trp26 apresenta ocorrência mediana na família das cisteíno proteases, sendo encontrado em 79,26% de todas as sequências. Trp26 localiza-se em uma longa α-hélice e apresenta correlação mediana e elevada com os demais resíduos da comunidade 1.

Com relação ao resíduo Val167, não foram encontrados dados na literatura acerca de tal resíduo da cruzaína, tampouco na papaína ou em catepsinas humanas. Val167 é encontrada em 69,44% de todas as proteínas da família e conforme análise da estrutura da cruzaína verificou-se que se localiza em uma extensa fita β. O resíduo de Val167 possui valores de correlação medianos e elevados em relação aos demais resíduos da comunidade 1.

O resíduo Tyr193 faz parte do segundo conjunto de resíduos, mencionado anteriormente, que auxiliam na estabilização da estrutura da proteína (figura 23) (DURRANT, J. D., et al, 2010). A análise da localização de Try193 na estrutura da cruzaína no Pymol evidencia que tal resíduo localiza-sena transição entre uma alça e uma fita β. Tyr193 é encontrada em 68,39% das sequências e possui correlação mediana e elevada em relação aos demais resíduos da comunidade 1. Portanto, Tyr193 possui importância apenas estrutural na família.

Os resíduos de prolina que fazem parte da comunidade 1, Pro2 e Pro134, estão localizados em pequenas voltas. A partir da análise da estrutura da cruzaína no Pymol verificou-se que Pro2 se localiza no início de uma volta que antecede uma fita β enquanto Pro134 se localiza em outra volta situada entre uma α-hélice e uma fita β. Enquanto Pro2 apresenta ocorrência mediana na família sendo encontrada em 72,98% das sequências, Pro134 possui ocorrência elevada, sendo encontrada em 85,55% das sequências. Segundo ALVAREZ, L. H. (2017), Pro2 faz parte do sexto sítio computacionalmente predito como alostérico da cruzaína, apresentando, portanto, importância funcional na proteína. Com relação à Pro134 não foram encontrados dados na literatura para nenhuma das proteases referência. A análise do Diagrama de Ramachandran para os resíduos de prolina mostra que na estrutura

da cruzaína são encontradas prolinas apenas na configuração *trans* e que todas as prolinas, incluindo Pro2 e Pro134, estão localizadas em regiões permitidas (figura 27 abaixo).

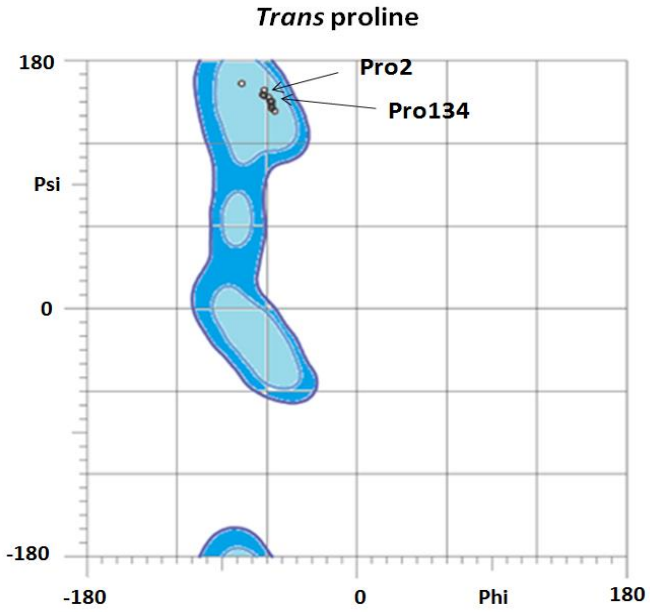

Figura 27. Diagrama de Ramachandran para a estrutura da cruzaína destacando as prolinas que fazem parte da comunidade 1. Representados os ângulos φ (phi) e ψ (psi) de todos os resíduos de prolina. Todas as prolinas da cruzaína encontram-se na configuração *trans*. As regiões favorecidas estão representadas em azul claro, as regiões permitidas, porém com algum impedimento energético estão representadas em azul escuro e as regiões não permitidas estão representadas em branco. Foram indicadas as prolinas da comunidade 1 com suas respectivas posições: Pro2 (φ = -62,6°, ψ = +159,3°) ePro134 (φ = -57,4°, ψ = +149°). Diagrama gerado pelo servidor MolProbity.

Por fim, Asn170, localiza-se na transição entre a ponta de uma fita β e uma longa alça e apresenta ocorrência mediana na família, sendo encontrada em 72,5% das sequências. Asn170 é um sítio de N-glicosilação e as duas isoformas da cruzaína, cruzipaína 1 e 2, possuem três prováveis sítios de glicolisação, sendo que um deles é o resíduo Asn170, encontrado em ambas as isoformas. Os demais sítios são Asn47, encontrado apenas na cruzipaína1, e Asn255, encontrado em ambas isoformas (SCHARFSTEIN, J. in KELLY, J. M., 2003). As demais proteínas que são referência neste estudo possuem na posição 170 ou correspondente resíduos de glicina enquanto apenas a catepsina O possui um resíduo de aspartato.

Portanto, a partir da análise de todos os resíduos da comunidade 1 pode-se concluir que nela encontram-se resíduos que se correlacionam evolutivamente por questões distintas, tanto funcionais quanto estruturais. Assim, os resíduos Pro2, Phe28, Glu35 e Tyr89 correlacionam-se por questões funcionais, provavelmente alosterismo. Por outro lado os resíduos Cys22, Cys56, Cys63, Cys101 e Cys203 estão implicados em questões estruturais já que realizam ligações dissulfeto (Cys22 e Cys63, Cys56 e Cys101, Cys203 realiza ligação dissulfeto com Cys155, que por sua vez faz parte da comunidade 3), assim como os demais resíduos: Asp6, Arg8, Trp26, Gly62, Gly66, Tyr89, Tyr147, Gly150 e Tyr193, que parecem coevoluir também por questões estruturais. Com relação aos resíduos Pro134, Val167, Tyr177 e Trp178 ainda não há dados na literatura que descrevam sua função e/ou importância para a proteína.

6.2 Comunidade 2

Dentre os resíduos correlacionados que compõem a comunidade 2, Cys25 e His162 formam a díade catalítica e apresentam ocorrência elevada na família, sendo encontrados em 89,11 e 95,45% das sequências, respectivamente.A atividade proteolítica de todas as cisteíno proteases ocorre apenas devido à presença dos resíduos de Cys e His no sítio ativo da enzima e nesse sentido a etapa crucial do processo de catálise é a formação de par iônico reativo tiolato/imidazólico, que resulta da transferência de prótons entre Cys25 e His162 (RZYCHON, M.; CHMIEL, D.; STEC-NIEMCZYK, J., 2004).

O resíduo Gln19, que também faz parte da comunidade 2 apresenta conservação alta na família sendo encontrado em 95,21% das sequências. Ele parece possuir importante papel na catálise, auxiliando na manutenção do caráter nucleofílico dos resíduos da díade catalítica (NOVINEC, M.; LENARCIC, B., 2013). A análise de correlação entre os resíduos Gln19, Cys25 e His162 aponta uma correlação alta entre eles. Quando a glutamina está na posição 19 e a histidina está na posição 162, a ocorrência de Cys25 aumenta de 89,11% para 91,29% e 92,58%, respectivamente. Da mesma forma, são observados aumentos na ocorrência de Gln19 e His162 quando se tem os pares de resíduos: Cys25 e His162 e Gln19 e Cys25, respectivamente. Os aumentos de tais valores de ocorrência desses

resíduos na presença dos demais confirmam a relevância e a interdependência dos mesmos no processo de catálise. Portanto, pode-se afirmar que eles correlacionam-se evolutivamente por questões funcionais. A figura 28, mostrada abaixo, destaca os resíduos de Gln19, Cys25 e His162.

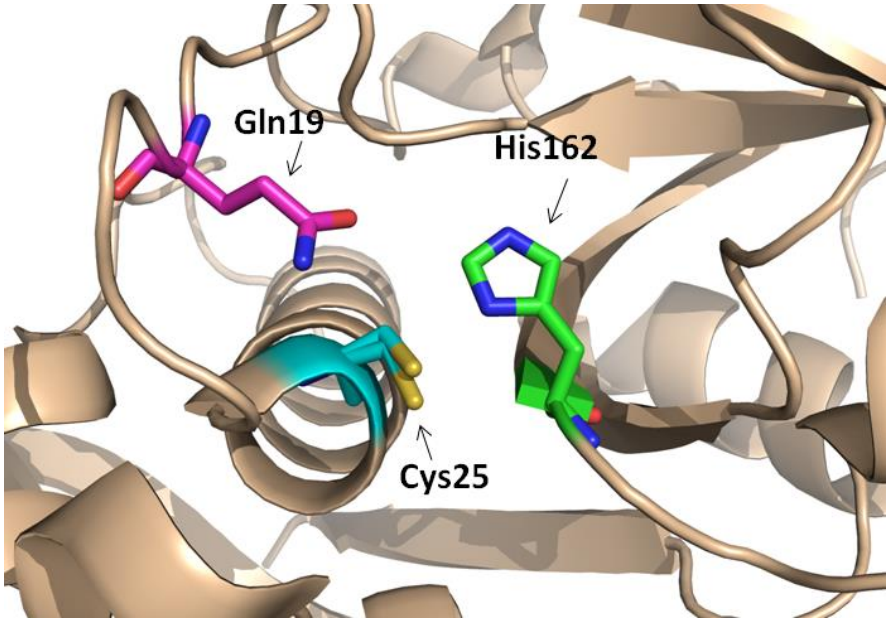

Figura 28. Representação de parte da estrutura da cruzaína destacando resíduos importantes, direta e indiretamente, para o processo de catálise e que fazem parte da comunidade 2. Destacada a estrutura da cruzaína (PDB: 3KKU) em *cartoon* rosa claro, nitrogênio em azul escuro e oxigênio em vermelho. Os resíduos Cys25 e His162 fazem parte da díade catalítica e o resíduo Gln19 auxilia na estabilização do oxiânion de Cys25 além de auxiliar na manutenção do caráter nucleofílico da díade. Figura preparada com o programa Pymol (Schrödinger, LCC, 2015).

Os resíduos Gly23 e Gly65 também apresentam ocorrência alta, sendo encontrados em 82,84 e 88,49% das sequências da família, respectivamente. Gly65 integra a formação de uma ampla volta juntamente com os resíduos Cys63, Asn64 e Gly66 que por sua vez é contínua a uma cadeia lateral do sítio de ligação S1. Gly23 faz parte de uma alça, juntamente com os resíduos Ser21 e Cys22, que se configura como o local de ligação da cadeia lateral de P1. Tanto a volta quanto a alça são reticuladas no topo da proteína mediante uma ponte dissulfeto entre Cys22 e Cys63 (TURK, D., et al, 1998).

A análise do Diagrama de Ramachandran para os resíduos de glicina da comunidade 2 da cruzaína mostra que os resíduos Gly23 e Gly65 se encontram em regiões permitidas, conforme demonstrado na figura 29 abaixo. O resíduo Gly23 se encontra em uma região também permitida para os outros tipos de resíduos. Já o resíduo Gly65 encontra-se em uma região não permitida para os demais resíduos (φ = +119,3°, ψ = +177,1°), de forma que a presença de glicina nessa posição é importante para assegurar a conformação da proteína.

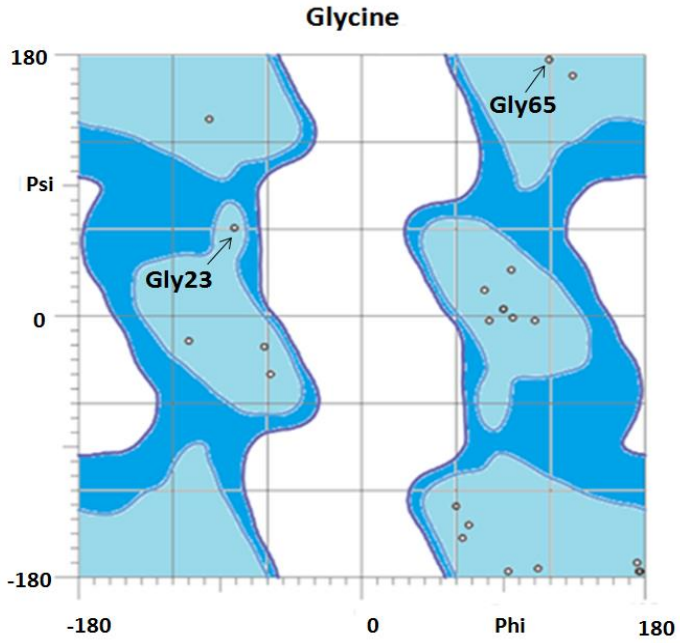

Figura 29. Diagrama de Ramachandran para a estrutura da cruzaína destacando as glicinas que fazem parte da comunidade 2. Representados os ângulos φ (phi) e ψ (psi) de todos os resíduos de glicina. As regiões favorecidas estão representadas em azul claro, as regiões permitidas, porém com algum impedimento energético estão representadas em azul escuro e as regiões não permitidas estão representadas em branco. Foram indicadas as glicinas da comunidade 2 com suas respectivas posições: Gly23 (φ = -81,8°, ψ = +61,6°) e Gly65 (φ = +119,3°, ψ = +177,1°). Diagrama gerado pelo servidor MolProbity.

Outro resíduo correlacionado que compõem a comunidade 2 é o resíduo Tyr91, encontrado em 87,75% das sequências da família. O resíduo Tyr91 faz parte de uma longa alça que atravessa a superfície da proteína, que por sua vez é formada pelos resíduos Tyr86 a Thr101. Tyr91 também faz parte do grupo de

resíduos do terceiro sítio alostérico predito da cruzaína, juntamente com os resíduos Phe28 e Glu35 da comunidade 1 e Leu48 da comunidade 5 (ALVAREZ, L. H., 2017). Além disso, Tyr91 integra o primeiro conjunto de resíduos, juntamente com os resíduos Pro90, Glu86, Tyr89, Gln51 e Ser49, que possuem importância na estabilização da estrutura da proteína (figura 12) (DURRANT, J. D., et al, 2010).

Portanto, a partir da análise dos resíduos da comunidade 2 pode-se concluir que nela encontram-se resíduos que se correlacionam evolutivamente por questões funcionais como catálise, resíduos Gln19, Cys25 e His162, e possivelmente alosterismo, resíduo Tyr91, enquanto há resíduos que se correlacionam por questões estruturais, como é o caso dos resíduos Gly23, Gly65 e Tyr91.

6.3 Comunidade 3

Os resíduos Ser55 e Cys155 são correlacionados e fazem parte da comunidade 3. O resíduo Ser55 encontra-se localizado em uma pequena α-hélice e possui conservação mediana na família, sendo encontrado em 52,35% das sequências. Dentre as proteínas elencadas como referência neste estudo apenas a cruzaína e a catepsina C possuem um resíduo de serina na posição 55 ou correspondente. A catepsina B possui uma treonina enquanto as demais proteínas referência possuem resíduos de aspartato nesta posição. Os resíduos de serina e treonina são polares neutros e possuem em sua cadeia lateral grupos que tendem a formar ligações de hidrogênio enquanto o resíduo de aspartato é ácido e sua cadeia lateral é formada pelo grupamento carboxilato.

Dados do alinhamento mostram que diversas proteínas de diversos tipos de organismos, tais como animais, plantas, vírus e organismos procariotas possuem resíduo de serina ou resíduo de aspartato na posição 55 ou correspondente. Portanto, a ocorrência de resíduos de serina ou de aspartato nesta posição parece não se relacionar a funções específicas de um determinado grupo de proteínas. Por outro lado, a substituição de resíduos de serina por resíduos de aspartato ou vice-versa parece não afetar substancialmente a estrutura da proteína. Não foram encontrados na literatura informações referentes à importância e/ou função dos resíduos presentes nesta posição.

O resíduo Cys155 é encontrado em 55,03% de todas as sequências, localiza-se em uma longa alça e realiza ponte dissulfeto com o resíduo Cys203 (LEE, G. M., et al, 2012), que integra a comunidade 1, conforme demonstrado na figura 30 abaixo. O resíduo Cys155 é classificado como resíduo remoto por localizar-se a uma distância de mais de 15Å de Cys25 (LEE, G. M., et al, 2012). Portanto, o resíduo Cys155 possui importante papel estrutural na proteína. Com relação ao resíduo Ser55 ainda não existem dados conclusivos na literatura sobre a sua importância e/ou função. Assim, são necessários estudos que possam verificar o papel que Ser55 desempenha na estrutura da proteína. A análise de correlação entre os resíduos desta comunidade aponta uma correlação alta entre eles.

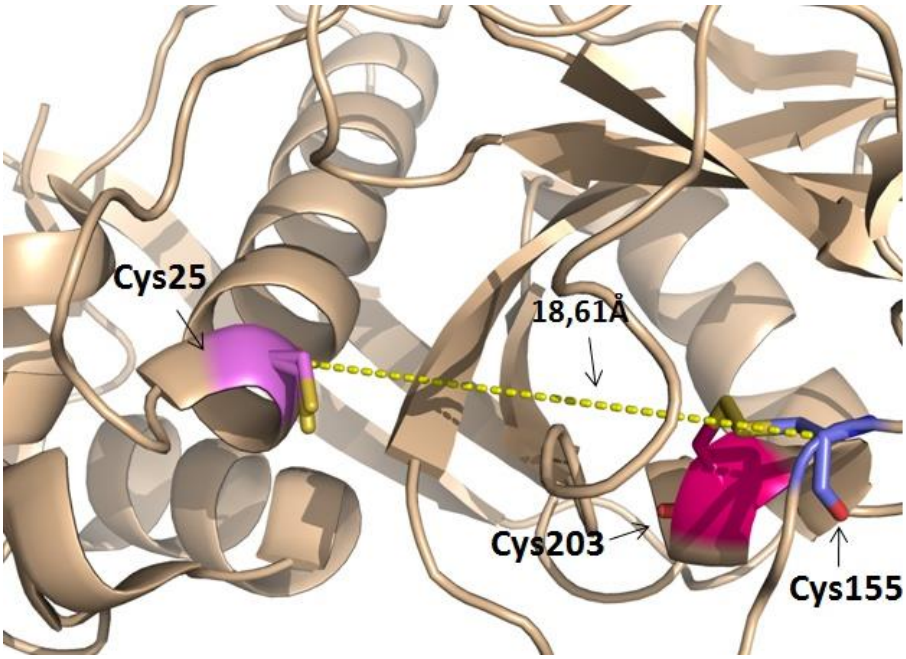

Figura 30. Representação de parte da estrutura da cruzaína destacando o resíduo de Cys155 e sua distância em relação à Cys25. Destacada a estrutura da cruzaína (PDB: 3KKU) em *cartoon* rosa claro, nitrogênio em azul escuro e oxigênio em vermelho. O resíduo de Cys155 realiza ponte dissulfeto com o resíduo Cys203 (sendo que esse último faz parte da comunidade 1). Representado o resíduo de Cys25, que faz parte da díade catalítica, e destacada em linha tracejada amarela a distância entre a cisteína da comunidade 3 e a Cys25. Figura preparada com o programa Pymol (Schrödinger, LCC, 2015).

6.4 Comunidade 4

Os resíduos Val13 e K181 correlacionam-se e fazem parte da comunidade 4. O resíduo Val13 integra um grupo de resíduos descritos como importantes para estabilização da estrutura da cruzaína, juntamente com os resíduos Tyr186, Arg8, Val16, Gly11 e Asp6 (figura 12). O resíduo Val13 possui ocorrência mediana na família, sendo encontrado em 52,07% das sequências (DURRANT, J. D., et al, 2010).

O resíduo Lys181 também possui conservação mediana na família, 52,03% de todas as sequências e localiza-se na porção C-terminal da proteína, no início de uma alça que une duas fitas β. Não foram encontrados na literatura informações referentes à importância e/ou função do resíduo Lys181 nas proteínas referência analisadas. Como o resíduo Val13 está implicado em questões estruturais pode-se cogitar a hipótese de que também o resíduo Lys181 poderia estar envolvido em questões estruturais, uma vez que ambos correlacionam-se pela evolução. A análise dos resíduos de Val13 e Lys181 na estrutura da cruzaína no Pymol permite identificar uma provável interação hidrofóbica entre a cadeia lateral da valina e a porção hidrofóbica da lisina, conforme destacado na figura 31. Assim, a distância observada entre essas porções da valina e da lisina é de 3,73Å, o que pode indicar uma interação entre esses resíduos corroborando a hipótese do envolvimento da Lys181 em questões estruturais. No entanto, são necessários estudos para analisar a função e a importância desse resíduo para estrutura e/ou função da proteína.A análise de correlação entre os resíduos desta comunidade aponta uma correlação alta entre eles.

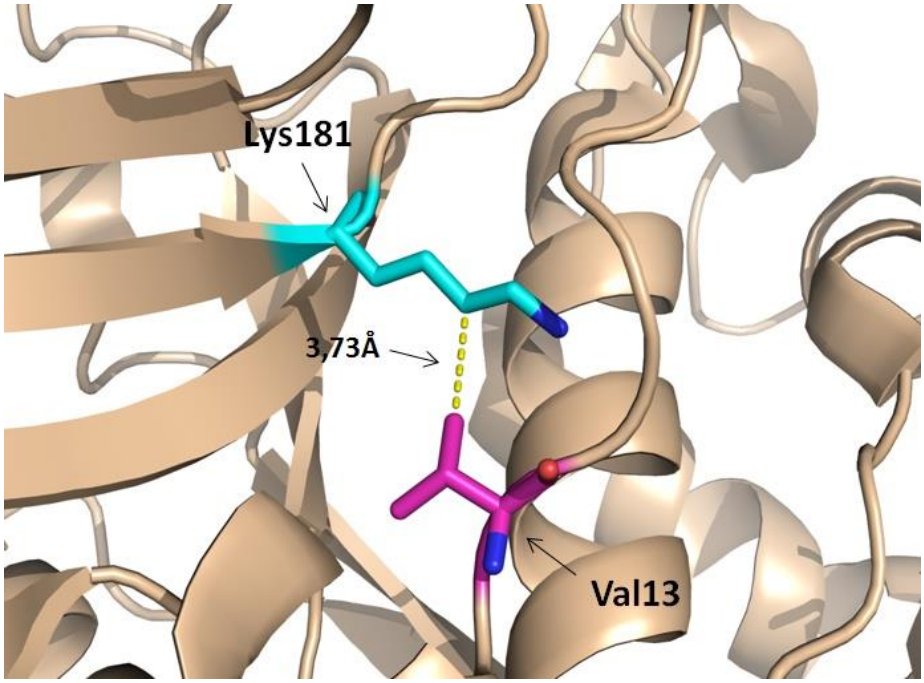

Figura 31. Representação de parte da estrutura da cruzaína destacando uma provável interação hidrofóbica entre resíduos da comunidade 4. Destacada a estrutura da cruzaína (PDB: 3KKU) em *cartoon* rosa claro, nitrogênio em azul escuro, oxigênio em vermelho. Pode-se verificar uma provável interação hidrofóbica entre a cadeia lateral da Val13 e a porção hidrofóbica da Lys181. A linha tracejada em amarelo representa a interação hidrofóbica entre os átomos dos resíduos envolvidos nessa interação. Figura preparada com o programa Pymol (Schrödinger, LCC, 2015).

6.5 Comunidade 5

Os resíduos Leu48 e Gln51 compõem a comunidade 5. O resíduo Leu48 possui ocorrência mediana na família, sendo encontrado em 68,19% das sequências. Ele situa-se em uma longa alça entre uma α-hélice longa e uma α-hélice curta e realiza ligações de hidrogênio com os resíduos Lys17 e Glu35, conforme demonstrado na figura 32 abaixo. Leu48 faz parte do grupo de resíduos do terceiro sítio alostérico predito computacionalmente para a cruzaína. As posições dos resíduos dentro deste terceiro sítio alostérico predito são importantes para a interação adequada com os ligantes e nesse sentido o resíduo Leu48 parece não se configurar como extremamente relevante uma vez que não é conservado, o que já ocorre com os resíduos Lys17 e Phe28 (ALVAREZ, L., H., 2017). Assim sendo, o resíduo Leu48 parece ser importante no que diz respeito a questões funcionais (alosterismo).

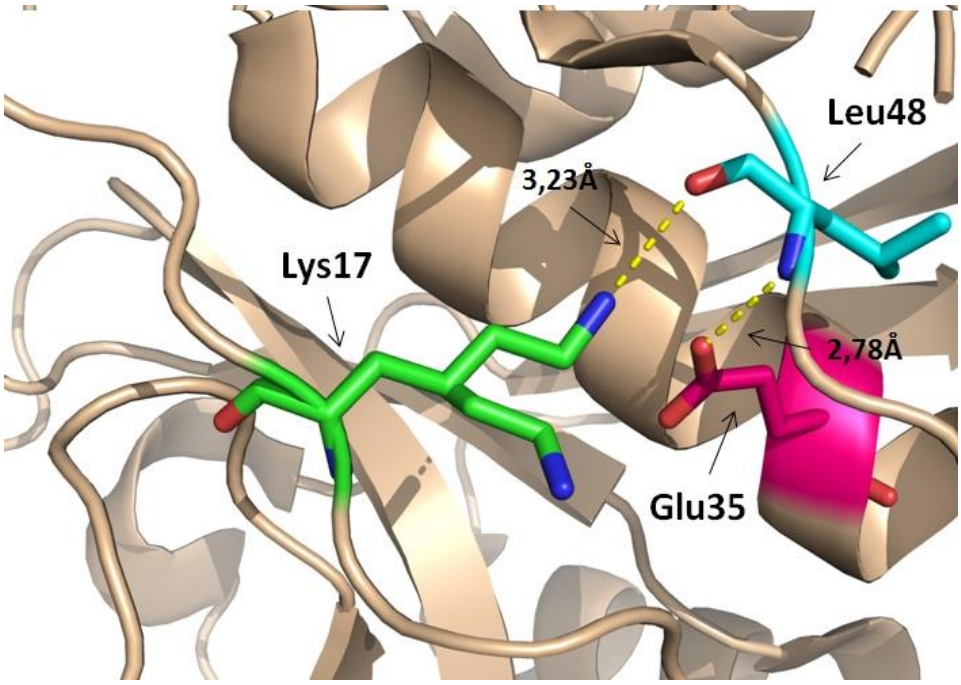

Figura 32. Representação de parte da estrutura da cruzaína destacando interações polares entre o resíduo de Leu48 e os resíduos Lys17 e Glu35. Destacada a estrutura da cruzaína (PDB: 3KKU) em *cartoon* rosa claro, nitrogênio em azul escuro e oxigênio em vermelho. O resíduo de Leu48 realiza ligação de hidrogênio com os resíduos de Lys17 e Glu35.As linhas tracejadas em amarelo representam as ligações de hidrogênio entre os átomos dos resíduos envolvidos nessas interações. Figura preparada com o programa Pymol (Schrödinger, LCC, 2015).

O resíduo Gln51 integra um conjunto de resíduos da cruzaína, juntamente com Tyr88, Pro87, Glu83, Tyr86 e Ser49, que possui papel importante na estabilização de alças da proteína (figura 23). Gln51 possui ocorrência mediana, sendo encontrado em 68,75% das sequências e pertence a uma α-hélice estável que é formada pelos resíduos de Ser49 a Leu56. Além disso, o átomo de oxigênio do grupamento carbonílico da cadeia lateral de Gln51 forma duas ligações de hidrogênio, uma com átomos da cadeia principal do resíduo Ala92 e outra com o grupo hidroxila da cadeia lateral do resíduo Ser93, conforme demonstrado na figura 33 abaixo. O grupamento amino da cadeia lateral de Gln51 também realiza ligações de hidrogênio com o grupamento hidroxila da cadeia lateral do resíduo Ser93 (DURRANT, J. D., et al, 2010). Gln51 também integra o quarto sítio alostérico da cruzaína que foi predito computacionalmente (ALVAREZ, L., H., 2017).

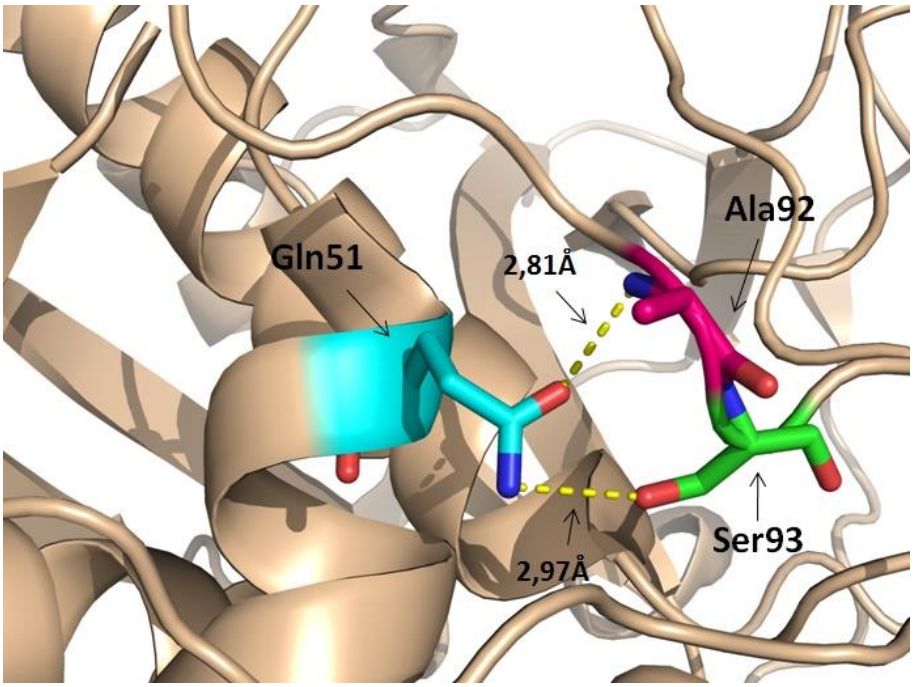

Figura 33. Representação de parte da estrutura da cruzaína destacando interações polares entre o resíduo de Gln51 e os resíduos Ala92 e Ser93. Destacada a estrutura da cruzaína (PDB: 3KKU) em *cartoon* rosa claro, nitrogênio em azul escuro, oxigênio em vermelho. O resíduo Gln51 realiza ligações de hidrogênio com Ala92 e Ser93.As linhas tracejadas em amarelo representam as ligações de hidrogênio entre os átomos dos resíduos envolvidos nessas interações. Figura preparada com o programa Pymol (Schrödinger, LCC, 2015).

Dados de estudos de mutagênese da catepsina C apontam papel importante do resíduo Gln51. A mutação Gln286Arg da catepsina C, correspondente a Gln51Arg da cruzaína, ocasiona o desenvolvimento da síndrome de Haim-Munk (SULÁK, A. et al, 2015). A síndrome de Haim-Munk é uma doença autossômica recessiva, extremamente rara e de queratinização. Caracteriza-se clinicamente por hiperqueratose palmoplantar, periodontite grave de início precoce, onicogrifose (hipertrofia que pode produzir unhas que se assemelham a garras ou chifre de carneiro), pé plano, aracnodactilia e acrosteólise (reabsorção das falanges ósseas distais que pode estar associada a alterações mínimas ou isquêmicas da pele que podem resultar em necrose digital). Tal síndrome foi descrita inicialmente na Índia e embora alguns achados tenham sido sugestivos da síndrome Papillon-Lefèvre, a presença de outras características clínicas como aracnodactilia, onicogrifose,

acrosteólise e pé plano confirmou a hipótese de que a síndrome de Haim-Munk era uma entidade clínica distinta (HART, T. C. et al, 2000; JANJUA, S. A. et al, 2008; PAHWA, P. et al, 2010).

Portanto pode-se concluir que o resíduo Gln51 possui uma importante função na estrutura da proteína por fazer parte de um grupo de resíduos que podem desempenhar papéis na regulação alostérica bem como na estabilidade estrutural (DURRANT, J. D., et al, 2010).A análise de correlação entre os resíduos desta comunidade aponta uma correlação alta entre eles.

6.6 Resíduos Conservados

A partir do alinhamento realizado foram encontrados dez resíduos que são altamente conservados evolutivamente na família de cisteíno proteases. Desses, três resíduos coevoluem e integram a comunidade 2, discutida anteriormente, e estão envolvidos no processo de catálise sendo críticos para o funcionamento da proteína: Gln19, Cys25 e His162. Além desses três resíduos, Asn182 também é um resíduo altamente conservado que possui importância no processo catalítico. Asn182 localiza-se em uma extensa alça que conecta duas fitas β.

Cys25 e His162 formam a díade catalítica e possuem conservação elevada, sendo encontrados em 84 e 89% de todas as sequências da família, respectivamente. Asn182 e Gln19 também são altamente conservados e possuem importante papel na manutenção do adequado posicionamento dos resíduos Cys25 e His162, que por sua vez são cruciais para a catálise. (NOVINEC, M., LENARCIC, B., 2013). Enquanto Asn182 é encontrado em 92% de todas as sequências da família, Gln19 é encontrado em 87% das sequências. A figura 34, mostrada abaixo, destaca os resíduos Gln19, Cys25, His162 e Asn182, bem como os resíduos Ser173 e Trp184 que serão discutidos posteriormente e possuem interações com resíduos envolvidos na catálise.

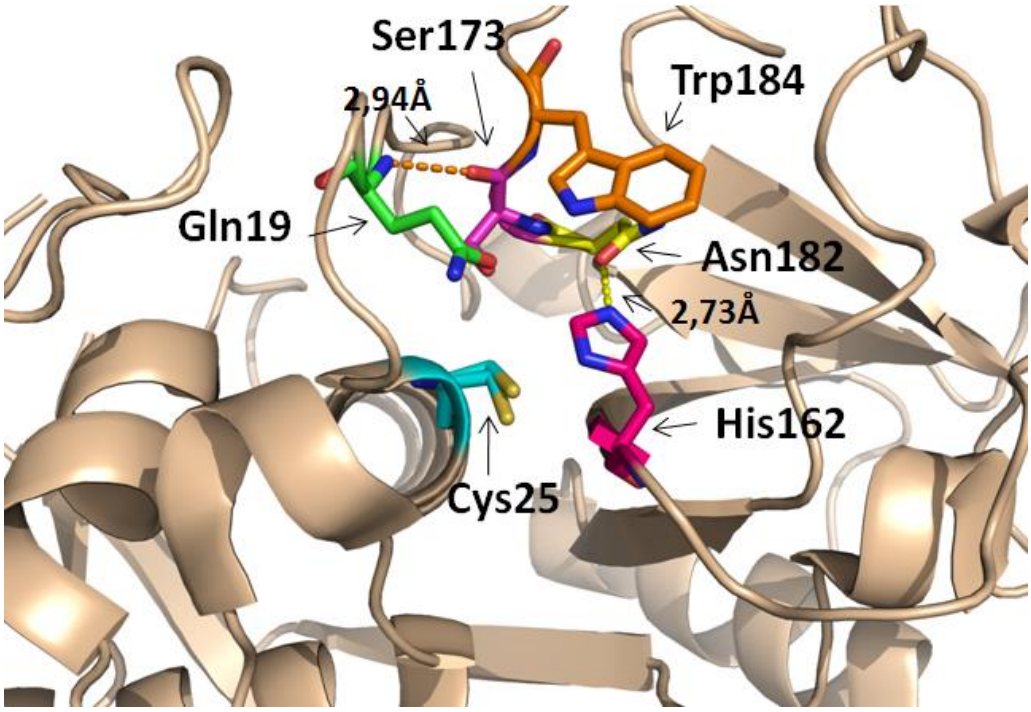

Figura 34. Representação de parte da estrutura da cruzaína destacando resíduos da díade catalítica bem como resíduos que possuem algum tipo de interação com resíduos envolvidos na catálise. Destacada a estrutura da cruzaína (PDB: 3KKU) em *cartoon* rosa claro, nitrogênio em azul escuro, oxigênio em vermelho. Os resíduos Cys25 e His162 fazem parte da díade catalítica e são mantidos em posicionamento adequado devido a interações que ocorrem com outros resíduos. O resíduo de Gln19, além de auxiliar na estabilização do oxiânion de Cys25, realiza ligação de hidrogênio com o resíduo de Ser173.Ser173 parece possuir um papel importante na orientação da interação entre Gln19 e o substrato. Asn182 realiza ponte uma ponte salina com His162 auxiliando na manutenção do correto posicionamento da díade catalítica. Trp184 realiza ligações de hidrogênio e também interações hidrofóbicas com o resíduo de His162. A linha tracejada em amarelo indica uma ligação de hidrogênio entre Asn182 e His162 enquanto a linha laranja indica a ligação de hidrogênio entre Gln19 e Ser173. Figura preparada com o programa Pymol (Schrödinger, LCC, 2015).

Como a Cys25 é um resíduo crítico para a catálise, foi realizada uma análise das proteínas do alinhamento que não possuem resíduos de cisteína na posição 25 e foi identificado que as proteínas que não possuem a Cys25 possuem quaisquer dos 19 aminoácidos, exceto triptofano e metionina. Assim, os resíduos que aparecem na posição 25 em substituição à cisteína em um maior número de proteínas são: serina, glicina, alanina e aspartato. Um exemplo de proteína com Ser25 ocorre no protozoário *Trypanosoma congolense,* agente etiológico da doença nagana que acomete bovinos, ovelhas, porcos, cabras, cavalos, camelos e cães.

Foram identificadas nesse parasita duas subfamílias de genes que codificam cisteíno proteases: seis genes codificam cisteíno proteases com a díade catalítica clássica (Cys e His) enquanto sete genes codificam proteínas nas quais a cisteína é substituída por serina. Esta substituição é perfeitamente conservada nesses sete genes e foi demonstrado que pelo menos uma catepsina B que possui o resíduo de Ser25 é expressa na corrente sanguínea do hospedeiro, o que sugere que tal proteína desempenhe uma função no ciclo de vida extracelular do parasita (MENDOZA-PALOMARES, C. et al, 2008). Outros exemplos de proteínas que possuem Ser25 incluem uma peptidase da subfamília C1A da *Giardia intestinalis*, uma peptidase da família C1 do *Schistosoma mansoni* e uma catepsina B do *Schistosoma japonicum*. Não foram encontradas informações acerca da função dessas proteínas. Também foram encontradas muitas proteínas ainda não caracterizadas de organismos ciliados, insetos e organismos marinhos.

Os resíduos Ser183 e Trp184 possuem conservação elevada na família, sendo encontrados em 86 e 83% de todas as sequências, respectivamente, e fazem parte da mesma longa alça formada pelos resíduos Lys180 a Tyr193 da qual também fazem parte os resíduos Gly189 e Gly192. A cadeia lateral do resíduo Ser176 da papaína (correspondente à Ser183 da cruzaína) realiza ligação de hidrogênio com o oxigênio da carbonila da cadeia lateral de Gln19 e parece possuir um papel importante de orientação da interação entre Gln19 e o substrato (figura 34). Em um estudo de mutagênese sítio dirigida foi realizada uma mutação Ser176Ala com o objetivo de avaliar a cinética da enzima após tal mutação e foi observado um efeito relativamente pequeno no funcionamento catalítico da enzima. Parâmetros cinéticos de hidrólise obtidos para a enzima mutante em pH 6,5 são levemente diferentes dos encontrados para a papaína. Valores de k_{cat} e K_m para a enzima mutante são de 24/s e 0,30mM, respectivamente, enquanto para a papaína os valores de k_{cat} e K_m são de 52/s e 0,42mM. A constante de especificidade k_{cat}/K_m para a enzima mutante é de 8,1 x 10^4/M/s enquanto para a papaína é de 1,2 x 10^5/M/s, o que representa uma redução de duas vezes na atividade da enzima mutante (MÉNARD, R., et al, 1991).

Um estudo realizado para o desenvolvimento de farmacóforos a partir de inibidores não covalentes da cruzaína mapeou algumas interações intermoleculares nos subsítios da cruzaína (S1', S2 e S3). Portanto, dentre os resíduos analisados foi

verificado que Trp184 parece integrar um grupo de resíduos formado por Leu160, Asp161 e Gly66, que realiza interações com os resíduos do sítio ativo da enzima. Assim, Leu160, Asp161 e His162 atuariam como aceptores de prótons na ligação de hidrogênio dessa interação, Gly66 e Trp184 atuariam como doadores de prótons na ligação de hidrogênio e His162 e Trp184 apresentariam interações hidrofóbicas secundárias a seus anéis aromáticos (figura 34). Trp184 localiza-se no sítio S1' juntamente com His162 (SOUZA, A. S., OLIVEIRA, M. T., ANDRICOPULO, A. D., 2017).

O resíduo Ser49 possui alta conservação sendo encontrado em 84% de todas as sequências. A partir da análise da estrutura da cruzaína verificou-se que Ser49 se localiza na interface entre uma α-hélice pequena e uma longa alça que por sua vez se conecta a uma longa α-hélice. Ser49 integra um conjunto de resíduos da cruzaína situado próximo ao seu sítio ativo juntamente com os resíduos Tyr91, Pro90, Glu86, Tyr89 e Gln51 (figura 12). Como esse grupo de resíduos possui um papel importante para a estabilização de várias alças, pode-se concluir que Ser49 também é importante para estabilização e manutenção do posicionamento adequado das alças que atravessam a superfície da proteína (DURRANT, J. D., et al, 2010). A análise do resíduo Ser49 na estrutura do Pymol não evidencia a possibilidade de interações entre tal resíduo e os demais resíduos conservados, uma vez que a menor distância entre eles é superior a 10Å.

Dados de mutagênese acerca da catepsina C humana evidenciam que uma mutação Ser284Asn nesta enzima,em posição equivalente a Ser49 da cruzaína, ocasiona o desenvolvimento da Síndrome Papillon-Lefèvre. Tal síndrome decorre de disfunção da catepsina C, portanto, depreende-se que Ser49 é um resíduo importante para manutenção da estrutura adequada da proteína (DURRANT, J. D., et al, 2010).

Com relação aos resíduos conservados de glicina, Gly168, Gly189 e Gly192 são encontrados em respectivamente 94, 85 e 91% de todas as sequências. A partir da análise estrutural verificou-se que Gly168 localiza-se em uma longa fita β enquanto Gly189 e Gly192 fazem parte de uma longa alça formada pelos resíduos Lys180 a Tyr193. A análise do Diagrama de Ramachandran para esses resíduos mostra que Gly168, Gly189 e Gly192 se encontram em regiões que são permitidas

para os demais tipos de resíduos, conforme mostrado na figura 35 abaixo. Foi realizada uma análise dos resíduos de Gly168, Gly189 e Gly192 na estrutura da cruzaína no Pymol a fim de verificar o tamanho das cadeias laterais dos resíduos que se encontram próximos a essas glicinas. A partir dessa análise estrutural verificou-se que esses resíduos possuem cadeias laterais maiores que a da glicina, sendo que algumas são formadas inclusive por anéis aromáticos. Assim, foi possível perceber que nas posições 168, 189 e 192, há um espaço relativamente reduzido para a acomodação de resíduos que porventura possuam cadeias laterais volumosas. Portanto, a presença dos resíduos de Gly168, Gly189 e Gly192em suas respectivas posições parece ser crítica para o adequado enovelamento da proteína, uma vez que resíduos com cadeias laterais maiores nessas posições poderiam causar problemas para o enovelamento proteico. Não foram encontradas informações na literatura acerca dos resíduos Gly168, Gly189 e Gly192 da cruzaína bem como dos resíduos de glicina nas posições correspondentes das enzimas que foram elencadas como referência nesse estudo.

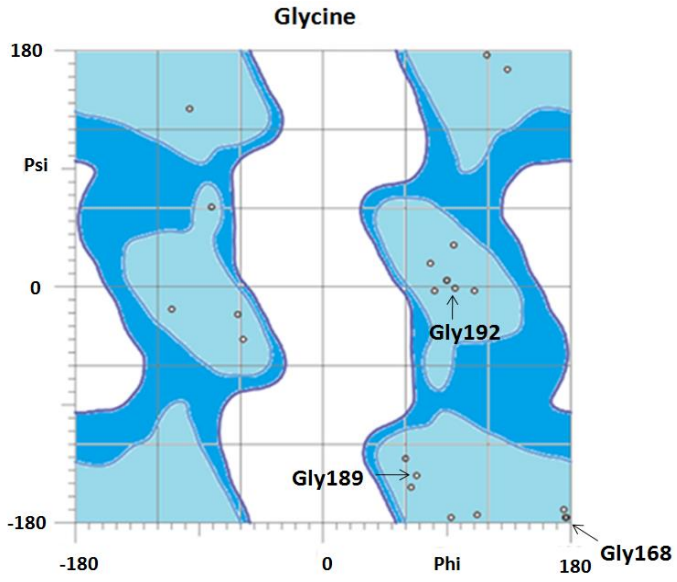

Figura 35. Diagrama de Ramachandran para a estrutura da cruzaína destacando as glicinas conservadas. Representados os ângulos φ (phi) e ψ (psi) de todos os resíduos de glicina. As regiões favorecidas estão representadas em azul claro, as regiões permitidas, porém com algum impedimento energético estão representadas em azul escuro e as regiões não permitidas estão representadas em branco. Foram indicadas as glicinas conservadas com suas respectivas posições: Gly168 (φ = +177,3°, ψ = -176,4°), Gly189 (φ = +68,4°, ψ = -144,2°) e Gly192 (φ = +97°, ψ = -1,1°). Diagrama gerado pelo servidor MolProbity.

Assim, dentre o grupo de resíduos conservados da família das cisteíno proteases pode-se concluir que alguns resíduos são conservados por questões funcionais, com envolvimento direto ou indireto na catálise, a saber: Gln19, Cys25, His162, Asn182, Ser183 e Trp184. Os demais resíduos, Ser49, Gly168, Gly189 e Gly192, parecem ser conservados por questões estruturais.

7. Conclusão

A partir da realização e análise do alinhamento múltiplo de sequências da família das cisteíno proteases e dos dados gerados acerca dos resíduos correlacionados e conservados foi possível obter maiores informações sobre tais proteínas. Mediante a realização da busca bibliográfica acerca de tais resíduos tornou-se possível obter um maior entendimento de sua importância para essas proteínas.

Assim, verificou-se que há muitos resíduos com importância estrutural e funcional claramente definida, ao passo que há resíduos que precisam ser investigados de forma mais profunda para que sua função seja realmente confirmada, como no caso dos resíduos situados nos sítios alostéricos preditos computacionalmente para a cruzaína. No entanto, os dados encontrados no presente trabalho certamente constituem-se um avanço no sentido de um entendimento maior da estrutura das proteínas da família de cisteíno proteases. Diante da importância das proteínas dessa família, espera-se que possam surgir novos estudos que busquem averiguar a importância dos resíduos que foram encontrados nas 5 comunidades bem como dos resíduos conservados.

8. **Trabalhos futuros**

O presente trabalho abre perspectivas para um maior aprofundamento acerca da importância dos resíduos correlacionados das 5 comunidades encontradas, bem como dos resíduos conservados da família das cisteíno proteases. Nesse sentido, fazem-se necessários estudos experimentais que possam confirmar a existência dos sítios alostéricos preditos para a cruzaína, mapeando os resíduos que os integram. Estudos de mutagênese sítio dirigida poderiam ser empregados para verificação da função de resíduos para os quais os dados da literatura ainda são insuficientes, permitindo avaliar o impacto de mutações sobre a estabilidade proteica e sobre sua atividade catalítica, por exemplo.

9. Referências Bibliográficas

ABU-ALRUZ, K., MAZAHREH, A. S., QUASEM, J. M., HEJAZIN, R. K., EL-QUDAH, J. M. Effect of Proteases on Meltability and Stretchability of Nabulsi Cheese. Am. J. Agric. Biol. Sci., 4:173-178, 2009.

ADKISON, A. M., RAPTIS, S. Z., KELLEY, D. G., PHAM, C. T. Dipeptidyl peptidase I activates neutrophil-derived serine proteases and regulates the development of acute experimental arthritis. J Clin Invest, 109:363-371, 2002.

ALLENDE, L. M., GARCIA-PEREZ, M. A., MORENO, A., CORELL, A., CARASOL, M., MARTÍNEZ-CANUT, P., ARNAIZ-VILLENA, A. Cathepsin C gene: First compound heterozygous patient with Papillon-Lefevre syndrome and a novel symptomless mutation. Hum Mutat. 17: 152–153, 2001.

ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W., Lipman, D. J. Basic local alignment search tool. Journal of Molecular Biology, 215(3):403-410, 1990.

ALVAREZ, L. H. Identification and Characterization of Cruzain Allosteric Inhibitors: A Computer-Aided Approach. 2017. 104f. Dissertação de Mestrado, UNESP, São José do Rio Preto, 2017.

ALVAREZ, L. H., GOMES, D. E. B., GONZÁLEZ, J. E. H., PASCUTTI, P. G. Dissecting a novel allosteric mechanism of cruzain: A computer-aided approach. PLOS ONE, 14(1), e0211227, 2019.

AMRI, E., MAMBOYA, F. Papain, a plant enzyme of biological importance: a review. American Journal of Biochemistry and Biotechnology. 8(2): 99-104, 2012.

ARAFET, K., FERRER, S., MOLINER, V. A Computational Study of the Catalytic Mechanism of the Cruzain Cysteine Protease.*ACS Catal.*, 2016.

ARAFET, K., FERRER, S., GONZÁLEZ, F. V., MOLINER, V. Quantum mechanics/molecular mechanics studies of the mechanism of cysteine protease inhibition by peptidyl-2,3-epoxyketones. Phys. Chem. Chem. Phys., 2017.

ASAGIRI, M., HIRAI, T., KUNIGAMI, T., KAMANO, S., GOBER, H. J., OKAMOTO, K., NISHIKAWA, K., LATZ, E., GOLENBOCK, D. T., AOKI, K., OHYA, K., IMAI, Y., MORISHITA, Y., MIYAZONO, K., KATO, S., SAFTIG, P., TAKAYANAGI, H. Cathepsin K-dependent toll-like receptor 9 signaling revealed in experimental arthritis. Science, 319:624-627, 2008.

BAICI, A., LANG, A., HORLER, D., KISSLING, R., MERLIN, C. Cathepsin B in

osteoarthritis: cytochemical and histochemical analysis of human femoral head cartilage. Ann Rheum Dis; 54:289-297, 1995.

BAICI, A., HORLER, D., LANG, A., MERLIN, C., KISSLING, R. Cathepsin B inosteoarthritis: zonal variation of enzyme activity in human femoral head cartilage. Ann Rheum Dis, 54:281-288, 1995.

BEELEY, J. A., YIP, H. K., STEVENSON, A. G. Chemochemical caries removal: A review of the techniques and latest developments. Br. Dent. J. 188:427-430, 2000.

Benton, D. Bioinformatics - principles and potential of a new multidisciplinary tool. Trends in Biotechnology, 14(8), 261–272, 1996.

BERG, J. M; TYMOCZKO, J. L.; STRYER, L. Bioquímica.6.Ed. Rio de Janeiro: Guanabara Koogan, 2010.

BLEICHER, L., LEMKE, N., GARRATT, R. C. Using amino acid correlation and community detection algorithms to identify functional determinants in protein families. Plos one, 6(12):1-11, 2011.

BOONEN S., ROSENBERG, E., CLAESSENS, F., VANDERSCHUEREN, D., PAPAPOULOS, S. Inhibition of cathepsin K for treatment of osteoporosis. CurrOsteoporos Rep, 10:73-79, 2012.

BRGULJAN, P. M., TURK, V., NINA, C., BRZIN, J., KRIZAJ, I., POPOVIC, T. Human brain cathepsin H as a neuropeptide and bradykinin metabolizing enzyme. Peptides, 24:1977-1984, 2003.

BUHLING, F. KOUADIO, M., CHWIERALSKI, C. E., KERN, U., HOHLFELD, J. M., KLEMM, N., FRIEDRICHS, N., ROTH, W., DEUSSING, J. M., PETERS, C.,REINHECKEL, T. Gene targeting of the cysteine peptidase cathepsin H impairs lung surfactant in mice. PLoS One, 6:e26247, 2011.

CHAPMAN, H. A., RIESE, R. J., SHI, G. P. Emerging roles for cysteine proteases in human biology. Annu. Rev. Physiol. 59:63-88, 1997.

CHENG, X. W., HUANG, Z., KUZUYA, M., OKUMURA, K., MUROHARA, T. Cysteine protease cathepsins in atherosclerosis-based vascular disease and its complications. Hypertension, 58:978-986, 2011.

CLARK, A. K., MALCANGIO, M. Microglial signalling mechanisms: Cathepsin S and Fractalkine. Experimental Neurology, 234:283-292, 2012.

COSTA, M. GS., BATISTA, P. R., SHIDA, C. S., ROBERT, C. H., BISCH, P. M., PASCUTTI, P. G. How does heparin prevent the pH inactivation of cathepsin B? Allosteric mechanism elucidated by docking and molecular dynamics. BMC Genomics, 11(Suppl 5), S5, 2010.

CYGLER, M., MORT, J. S. Proregion structure of members of the papain superfamily. Mode of inhibition of enzymatic activity.Biochimic, 79:645-652, 1997.

DIMA, R. I., THIRUMALAI, D. Determination of network of residues that regulate allostery in protein families using sequence analysis. Protein Science, 15(2), 258–268, 2006.

DIETRICH, R. E. Oral proteolytic enzymes in the treatment of athletic injuries: a double-blind study. Pennsyl. Med. J., 68:35-37, 1965.

DO, C. B., KATOH, K. Protein multiple sequence alignment. Methods in molecular biology (Clifton, N. J.), v. 484, p. 379-413, 2008.

DRIESSEN, C., BRYANT, R. A., LENNON-DUMENIL, A. M., VILLADANGOS, J. A., BRYANT, P. W., SHI, G. P., CHAPMAN, H. A., PLOEGH, H. L. Cathepsin S controls the trafficking and maturation of MHC class II molecules in dendritic cells. J Cell Biol, 147:775-790, 1999.

DURRANT, J. D., KERANEN, H., WILSON, B. A., MCCAMMON, J. A. Computational Identification of Uncharacterized Cruzain Binding Sites. Plos Neglected Tropical Diseases. 4(5): 1-11, 2-2010.

EDINGER, T. O., POHL, M. O., YÁNGÜEZ, E., STERTZ, S. Cathepsin W Is Required for Escape of Influenza A Virus from Late Endosomes. mBio, 6(3), 2015.

EIJSINK, V. G. H., GASEIDNES, S., BORCHERT, T. V., VAN DEN BURG, B. Directed evolution of enzyme stability. Biomolecular Engineering, 22(1-3), 2005.

EL-FADILI, A. K., ZANGGER, H., DESPONDS, C., GONZALES, I. J., ZALILA, H., SCHAFF, C., IVES, A., MASINA, S., MOTTRAM, J. C., FASEL, N. Cathepsin B-like

and cell death in the unicellular human pathogen Leishmania.Cell Death Dis, 1:e71, 2010.

FAN, K., LI, D., ZHANG, Y., HAN, C., LIANG, J., HOU, C., XIAO, H., IKENAKA, K., MA, J.The induction of neuronal death by up-regulated microglial cathepsin H in LPS-induced neuroinflammation.Journal of Neuroinflammation, 12(1), 2015.

FINN, R. D., COGGILL, P., EBERHARDT, R. Y., EDDY, S. R., MISTRY, J., MITCHELL, A. L., POTTER, S. C., PUNTA, M., QURESHI, M., SANGRADOR-VEGAS, A., SALAZAR, G. A., TATE, J., BATEMAN, A.. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Research, 44(D1):D279--D285, 2016.

FLINDT, M. L. Allergy to alpha-amylase and papain. Lancet, 1:1407-1408, 1979.

FONSECA-JÚNIOR, N. J., AFONSO, M. Q. L., OLIVEIRA, L. C., BLEICHER, L. PFstats: A Network-Based Open Tool for Protein Family Analysis. J Comput Biol. 25(5):480-486, 2018.

FUNKELSTEIN, L., LU, W. D., KOCH, B., MOSIER, C., TONEFF, T., TAUPENOT, L., O'CONNOR, D. T., REINHECKEL, T., PETERS, C., HOOK, V. Human cathepsin V protease participates in production of enkephalin and NPY neuropeptide neurotransmitters. J BiolChem, 287:15232-15241, 2012.

GELB, B. D., SHI, G. P., CHAPMAN, H. A., DESNICK, R. J. Pycnodysostosis, aLysosomal disease caused by cathepsin K deficiency. Science, 273:1236-1238, 1996.

GILROY, E. M., HEIN, I., VAN DER HOORN, R., BOEVINK, P. C., VENTER, E., MCLELLAN, H., KAFFARNIK, F., HRUBIKOVA, K., SHAW, J., HOLEVA, M.,LOPEZ, E. C., BORRAS-HIDALGO, O., PRITCHARD, L., LOAKE, G. J., LACOMME, C., BIRCH, P. R. Involvement of cathepsin B in the plant diseaseresistance hypersensitive response. Plant J, 52:1-13, 2007.

GÖBEL, U., SANDER, C., SCHNEIDER, R., VALENCIA, A. Correlated mutations and residue contacts in proteins. Proteins: Structure, Function, and Genetics, 18(4), 309–317, 1994.

GOULET, B., BARUCH, A., MOON, N. S., POIRIER, M., SANSREGRET, L. L., ERICKSON, A., BOGYO, M., NEPVEU, A. A cathepsin L isoform that is devoid of a signal peptide localizes to the nucleus in S phase and processes the CDP/Cux transcription factor. Mol Cell, 14:207-219, 2004.

GOULET, B., TRUSCOTT, M., NEPVEU, A. A novel proteolytically processed CDP/Cux isoform of 90 kDa is generated by cathepsin L. BiolChem, 387:1285-1293, 2006.

GUAY, D., BEAULIEU, C., PERCIVAL, M. D. Therapeutic utility and medicinal chemistry of cathepsin C inhibitors. Curr Top Med Chem, 10:708-716, 2010.

HAN, J., LUO, T., GU, Y., LI, G., JIA, W., LUO, M. Cathepsin K regulates adipocyte differentiation: possible involvement of type I collagen degradation. Endocr J, 56:55-63, 2009.

HART, T. C., HART, P. S., BOWDEN, D. W., MICHALEC, M. D., CALLISON, S. A., WALKER, S.J., ZHANG, Y., FIRATLI, E. Mutations of the cathepsin C geneare responsible for Papillon-Lefevre syndrome. J Med Genet, 36:881-887, 1999.

HART, T. C., HART, P. S., MICHALEC, M. D., ZHANG, Y., FIRATLI, E., VAN DYKE, T. E., STABHOLZ, A., ZLOTOGORSKI, A., SHAPIRA, L. SOSKOLNE, W. A. Haim-Munk Syndrome and Papillon-Lefevre syndrome are allelic mutations in cathepsin C. J Med Genet. 37: 88–94, 2000.

HO, B. K; THOMAS, A.; BRAUSSER, R. Revisiting the Ramachandran plot: hard-sphere repulsion, electrostatics, and H-bonding in the α-helix. Protein Science. 12:2508-2522, 2003.

HUANG, L., BRINEN, L. S., ELLMAN, J. A. Crystal Structures of Reversible Ketone-Based Inhibitors of the Cysteine Protease Cruzain. Bioorganic & Medicinal Chemistry 11:21–29, 2003.

HUSMANN, K., MUFF, R., BOLANDER, M. E., SARKAR, G., BORN, W., FUCHS, B. Cathepsins and osteosarcoma: expression analysis identifies cathepsin K as an indicator of metastasis. Mol Carcinog, 47:66-73, 2008.

JANJUA, S. A., IFTIKHAR, N., HUSSAIN, I., KHACHEMOUNE, A. Dermatologic, periodontal, and skeletal manifestations of Haim-Munk syndrome in two siblings. J Am Acad Dermatol. 58(2): 339-344, 2008.

JUNQUEIRA, D. M., BRAUN, R. L., VERLI, H. Alinhamentos. In: VERLI, V. Bioinformática da Biologia à flexibilidade Molecular. 1. Ed. São Paulo: SBBq. 2014. Cap 3, p. 39-61.

KARRER, K. M., PEIFFER, S. L., DITOMAS, M. E. Two distinct gene subfamilies within the family of cysteine protease genes. Proceedings of the National Academy of Sciences, 90(7), 3063–3067, 1993.

KINCH, L. N., GRISHIN, N. V. Evolution of protein structures and functions. Current Opinion in Structural Biology, 12(3), 400–408, 2002.

KRUEGER, S., KALINSKI, T., HUNDERTMARK, T., WEX, T., KUSTER, D., PEITZ, U., EBERT, M., NAGLER, D. K., KELLNER, U., MALFERTHEINER, P., NAUMANN, M., ROCHEN, C., ROESSNER, A. Up-regulation of cathepsin X in Helicobacter pylori gastritis and gastric cancer. J Pathol, 207:32-42, 2005.

LECALILLE, F., VANDIER, C., GODAT, E., HERVE-GREPINET V., BROMME D., LALMANACH, G. Modulation of hypotensive effects of kinins by cathepsin K. Arch. Biochem Biophys, 459:129-136, 2007.

LECHNER, A. M., ASSFALG-MACHLEIDT, I., ZAHLER, S., STOECKELHUBER, M., MACHLEIDT, W., JOCHUM, M., Nagler, D. K. RGD-dependent binding of procathepsin X to integrin α v β 3 mediates cell-adhesive properties. J BiolChem, 281:39588-39597, 2006.

LEE, G. M., BALOUCH, E., GOETZ, D. H., LAZIC, A., MCKERROW, J. H., CRAIK, C. S. Mapping Inhibitor Binding Modes on an Active Cysteine Protease via Nuclear Magnetic Resonance Spectroscopy. Biochemistry. 51:10087−10098, 2012.

LENARCIC IC, B., GABRIJELCIC, D., ROZMAN, B., DROBNIC-KOSOROK, M., TURK, V. Human cathepsin B and cysteine proteinase inhibitors (CPIs) in inflammatory and metabolic joint diseases. BiolChem Hoppe-Seyler, 369 Suppl: 257-261, 1988.

LENDECKEL, U., KAHNE, T., TEN HAVE, S., BUKOWSKA, A., WOLKE, C., BOGERTS, B., KEILHOFF, G., BERNSTEIN, H. G. Cathepsin K generates enkephalin from β -endorphin: a new mechanism with possible relevance for schizophrenia. NeurochemInt, 54:410-417, 2009.

LIMA, A. P, C. A., REIS, F. C. G., SERVEAU, C., LALMANACH, G., JULIANO, L., MÉNARD, R., VERNET, T., THOMAS, D. Y., SOTRER, A. C., SCHARFSTEIN, J. Cysteine protease isoforms from Trypanosoma cruzi, cruzipain 2 and cruzain, present different substrate preference and susceptibility to inhibitors. Molecular & Biochemical Parasitology, 114:41-52, 2001.

LITTLE, D. Y., CHEN, L. Identification of Coevolving Residues and Coevolution Potentials Emphasizing Structure, Bond Formation and Catalytic Coordination n Protein Evolution. Plos one. 4(3): 1-14, 2009.

LOCKLESS, S. W., RANGANATHAN, R. Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families. Science, 286(5438), 295–299, 1999.

LU, W. D., FUNKELSTEIN, L., TONEFF, T., REINHECKEL, T., PETERS, C., HOOK, V. Cathepsin H functions as an aminopeptidase in secretory vesicles for production of enkephalin and galanin peptide neurotransmitters. J Neurochem, 122:512-522, 2012.

LYO, V., CATTARUZZA, F., KIM, T. N., WALKER, A. W., PAULICK, M., COX, D., CLOYD, J., BUXBAUM, J., OSTROFF, J., BOGYO, M., GRADY, E. F., BUNNET, N. W., KIRKWOOD, K. S. Active cathepsins B, L, and S in murine and human pancreatitis. Am J Physiol, 303:G894 – 903, 2012.

MANSFIELD, L. E., TING, S., HAVERLY, R. W., YOO, T. J. The incidence and clinical implications of hypersensitivity to papain in an allergic population, confirmed by blinded oral challenge. Ann. Allergy., 55:541-543, 1985.

MCGHIRE, M.J., LIPSKY, P. E., THIELE, D. L. Generation of active myeloid and lymphoid granule serine proteases requires processing by the granule thiol protease dipeptidyl peptidase I. J BiolChem, 268:2458–2467, 1993.

MCGRATCH, M. E., EAKIN, A. E., ENGEL, J. C., MCKERROW, J. H., CRAIK, C. S., FLETTERICK, R. J. The crystal structure of cruzaín: a therapeutic target for Chagas' disease. J. Mol. Biol, 247:251-259, 1995.

MCGRATH, M. E., PALMER, J. T., BROMME, D., SOMOZA, J. R. Crystal structure of human cathepsinS. Protein Science*.* 7:1294-1302, 1998.

MCLELLAN, H., GILROY, E. M., YUN, B. W., BIRCH, P. R., LOAKE, G. J. Functional redundancy in the Arabidopsis Cathepsin B gene family contributes to basal defence, the hypersensitive response and senescence. New Phytol, 183:408-418, 2009.

MENARD, R., PLOUFFE, C., KHOURI, H. E., DUPRAS, R., TESSIER, D. C., VERNET, T., THOMAS, D. Y., STORER, A. C. Removal of an inter-domain hydrogen bond through site-directed mutagenesis: role of serine 176 in the mechanism of papain. Protein Engineering.4(3):307-311, 1991.

MENDOZA-PALOMARES, C., BITEAU, N., GIROUD, C., COUSTOU, V., COETZER, T., AUTHIE, E., BOULANGE, A., BALTZ, T. Molecular and Biochemical Characterization of a Cathepsin B-Like Protease Family Unique to *Trypanosoma congolense*. Eukaryotic Cell, 7(4), 684–697, 2008.

MOHAMED M. M, SLOANE B. F. Cysteine cathepsins: multifunctional enzymes in cancer. Nat Rev Cancer; 6: 764-775, 2006.

MOON, H. Y., BECKE, A., BERRON, D., BECKER, B., SAH, N., BENONI, G., JANKE, E., LUBEJKO, S. T., GREIG, N. H., MATTISON, J. A., DUZEL, E., VAN PRAAG, H. Running-Induced Systemic Cathepsin B Secretion Is Associated with Memory Function. Cell Metabolism, 24:1-9, 2016.

NAGLER, D.K., KRAUS, S., FEIERLER, J., MENTELE, R., LOTTSPEICH, F., JOCHUM, M., FAUSSNER, A. A cysteine-type carboxypeptidase, cathepsin X,generates peptide receptor agonists. IntImmunopharmacol, 10:134-139, 2010.

NAGLER, D. K., SULEA, T., MENARD, R. Full-length cDNA of human cathepsin F predicts the presence of a cystatin domain at the N-terminus of the cysteine protease zymogen. BiochemBiophys Res Commun, 257:313-318, 1999.

NAKAGAWA, T. Y., BRISSETTE, W. H., LIRA, P. D., GRIFFITHS, R. J., PETRUSHOVA, N., STOCK, J., MCNEISH, J. D., EASTMAN, S. E., HOWARD, E. D., CLARKE, S. R., ROSLONIEC, E. F., ELLIOTT, E. A., RUDENSKY, A. Y. Impaired invariant chain degradation and antigen presentation and diminished collagen-induced arthritis in cathepsin S null mice. Immunity; 10:207-217, 1999.

NIWA, Y., SUZUKI, T., DOHMAE, N., UMEZAWA, K., SIMIZY, S. Determination of cathepsin V activity and intracellular trafficking by N-glycosylation. FEBS Letters, 586:3601-3607, 2012.

NOVINEC, M., REBERNIK, M., LENARCIC, B. An allosteric site enables fine-tuning of cathepsin K by diverse effectors. FEBS Letters, 590(24), 4507–4518, 2016.

NOVINEC, M., KORENC, M., CAFLISCHL, A., RANGANATHAN, R., ENARCIC, B., BAICI, A.A novel allosteric mechanism in the cysteine peptidase cathepsin K discovered by computational methods. Nature Comunications, 5(3287):1-10, 2014.

NOVINEC, M., LENARCIC, B. Papain-like peptidases: structure, function, and evolution. BioMol Concepts, 4(3):287-308, 2013.

OBERMAJER, N., SVAJGER, U., BOGYO, M., JERAS, M., KOS J. Maturation of dendritic cells depends on proteolytic cleavage by cathepsin X. J Leukoc Bio, 84:1306-1315, 2008.

OBERMAJER, N., REPNIK, U., JEVNIKAR, Z., TURK, B., KREFT, M., KOS, J. Cysteine protease cathepsin X modulates immune response via activation of β 2 integrins. Immunology, 124:76-88, 2008.

OBERMAJER, N., DOLJAK, B., JAMNIK, P., FONOVIC, U. P., KOS, J. Cathepsin X cleaves the C-terminal dipeptide of α - and γ -enolase and impairssurvival and neuritogenesis of neuronal cells. Int J Biochem Cell Biol, 41:1685-1696, 2009.

OORNI, K., SNECK, M., BROMME, D., PENTIKAINE, M. O., LINDSTEDT, K. A., MAYRANPAA, M., AITIO, H., KOVANEN, P. T. Cysteine protease cathepsin F is expressed in human atherosclerotic lesions, is secreted by cultured macrophages, and modifies low density lipoprotein particles in vitro. J BiolChem, 279:34776-34784, 2004.

PAHWA, P., LAMBA, A. K., FARAZ, F., TANDON, S. Haim-Munk syndrome.J Indian SocPeriodontol. 14(3): 201-203.

PHAM, C. T., LEY, T.J. Dipeptidyl peptidase I is required for the processing and activation of granzymes A and B in vivo. Proc Natl AcadSci USA, 96:8627-8632, 1999.

PUNTURIERI, A., FILIPPOV, S., ALLEN, E., CARAS, I., MURRAY, R., REDDY, V., WEISS, S. J. Regulation of elastinolytic cysteine proteinase activity in normal and cathepsin K-deficient human macrophages. J Exp Med, 192:789-799, 2000.

QUE, X., NGO, H., LAWTON, J., GRAY, M., LIU, Q., ENGEL, J., BRINEN, L., GHOSH, P., JOINER, K. A., REED, S. L. The cathepsin B of Toxoplasma gondii, toxopain-1, is critical for parasite invasion and rhoptry protein processing. J BiolChem, 277:25791-25797, 2002.

REDDY, V. Y., ZHANG, Q. Y., WEISS, S. J. Pericellular mobilization of the tissue-destructive cysteine proteinases, cathepsins B, L, and S, by human monocyte-derived macrophages. Proc Natl AcadSci USA, 92:3849-3853, 1995.

ROTH, W., DEUSSING, J., BOTCHKAREV, V. A., PAULY-EVERS, M., SAFTIG, P., HAFNER, A., SCHMIDT, P., SCHMAHL, W., SCHERER, J., ANTON-LAMPRECHT, I., VON FIGURA, K., PAUS, R., PETERS, C. Cathepsin L deficiency as molecular defect of furless: hyperproliferation of keratinocytes and pertubation of hair follicle cycling. FASEB J, 14:2075-2086, 2000.

RZYCHON, M., CHMIEL, D., STEC-NIEMCZYK, J. Modes of inhibition of cysteine proteases. Acta Biochimica Polonica. 51(4):861-873, 2004.

SAGE, J., MALLÈVRE, F., BARBARIN-COSTES, F., SAMSONOV, S. A., GEHRCKE, J-P., PISABARRO, PERRIER, E., SCHNEBERT, S., ROGET, A., LIVACHE, T., NIZARD, C., LALMANACH, G., M. T., LECAILLE, F. Binding of Chondroitin 4-Sulfate to Cathepsin S Regulates Its Enzymatic Activity. Biochemistry, 52(37), 6487–6498, 2013.

SAJID, M., ROBERTSON, S. A., BRINEN, L. S., MCKERROW, J. H. Cruzain: the path from target validation to the clinic. In: ROBINSON, M. W., DALTON, J. P. Cysteine proteases of pathogenic organisms. USA: Springer, 2011. Cap 7, p. 100-111.

SCHARFSTEIN, J. Activation of bradykinin-receptors by Trypanosoma cruzi: a role for cruzipain in microvascular pathology. In: KELLY, J. M.Molecular Mechanisms of Pathogenesis in Chagas' Disease. USA: Springer, 2003. Cap. 8.

SHI, G. P., VILLADANGOS, J. A., DRANOFF, G., SMALL, C., GU, L., HALEY, K. J., RIESE, R., PLOEGH, H. L., CHAPMAN, H. A. Cathepsin S required fornormal MHC class II peptide loading and germinal center development. Immunity, 10:197-206, 1999.

SHI, G. P., BRYANT, R. A., RIESE, R., VERHELST, S., DRIESSEN, C., LI, Z., BROMME, D., PLOEGH, H. L., CHAPMAN, H. A. Role for cathepsin F in invariant chain processing and major histocompatibility complex class II peptide loading by macrophages. J Exp Med, 191:1177-1186, 2000.

SIVARAMAN, J., LALUMIÈRE, M., MÉNARD, R., CYGLER, M. Crystal structure of wild-type human procathepsin K. Protein Science.8:283–290, 1999.

SOUZA, A. S., OLIVEIRA, M. T., ANDRICOPULO, A. D. Development of a pharmacophore for cruzain using oxadiazoles as virtual molecular probes: quantitative structure–activity relationship studies. J. Comput. Aided. Mol. Des., 2017.

STOECKLE, C., GOUTTEFANGEAS, C., HAMMER, M., WEBER, E., MELMS, A., TOLOSA, E. Cathepsin W expressed exclusively in CD8 + T cells and NK cells, is secreted during target cell killing but is not essential for cytotoxicity in human CTLs. Exp Hem, 37:266-275, 2009.

STOKA, V., TURK, B., TURK, V. Lysosomal cysteine proteases: structural features and their role in apoptosis. IUBMB Life, 57(4/5):347-353, 2005.

SULÁK, A., TÓTH, L., FARKAS, K., TRIPOLSZKI, K., FÁBOS, B., KEMÉNY, L., VALYI, P., NAGY, K., NAGY, N., SZELL, M. One mutation, two phenotypes: a single nonsense mutation of theCTSCgene causes two clinically distinct phenotypes. Clinical and Experimental Dermatology, 41(2), 190–195, 2015.

SUTTO, L., MARSILI, S., VALENCIA, A., GERVASIO, F. L.From residue coevolution to protein conformational ensembles and functional dynamics.PNAS.1-6, 2015.

TALEB, S., CANCELLO, R., CLEMENT, K., LACASA, D. Cathepsin S promotes human preadipocyte differentiation: possible involvement of fibronectin degradation. Endocrinology, 147:4950-1959, 2006;

TURK, D., GUNCAR, G. PODOBNIK, M., TURK, B. Revised Definition of Substrate Binding Sites of Papain-Like Cysteine Proteases. Biol. Chem. 379:137 -147, 1998.

TURK, B., TURK, D., TURK, V. Lysosomal cysteine proteases: more than scavengers. Biochimica et Biophysica Acta, 1477:98-111, 2000.

TURK, V., TURK, B., TURK, D. Lysosomal cysteine proteases: facts and opportunities. The EMBO Journal, 20(17):4629-4633, 2001.

VEILLARD, F., LECAILLE, F., LALMANACH, G. Lung cysteine cathepsins: intruders or unorthodox contributors to the kallikrein-kinin system? Int J Biochem Cell Biol; 40:1079-1094, 2008.

WANG, C., SUN, B., ZHOU, Y., GRUBB, A., GAN, L. Cathepsin B Degrades Amyloid-βin Mice Expressing Wild-type Human Amyloid Precursor Protein. The Journal of Biological Chemistry, 287(47):39834-39841, 2012.

WENDT, W., ZHU, X. R., LUBBERT, H., STICHEL, C. C. Differential expression of cathepsin X in aging and pathological central nervous system of mice. ExpNeurol, 204:525-540, 2007.

WIEDERANDERS, B., KAULMANN, G., SCHILLING, K. Functions of propeptide parts in cysteine proteases. Current Protein and Peptide Science, 4:309-326, 2003.

YANG, M., SUN, J., ZHANG, T., LIU, J., ZHANG, J., SHI, M. A., DARAKHSHAN, F., GUERRE-MILLO, M. CLEMENT, K., GELB B. D., DOLGNOV, G., SHI, G. P. Deficiency and inhibition of cathepsin K reduce body weight gain and increase glucose metabolism in mice. Arterioscler Thromb Vasc Biol, 28:2202-2208, 2008.

YANG, M., ZHANG, Y., PAN, J., SUN, J., LIU, J., LIBBY, P., SUKHOVA, G. K., DORIA, A., KATUNUMA, N., PERONI, O. D., GUERRE-MILLO, M., KAHN, B. B., CLEMENT, K., SHI, G. P. Cathepsin L activity controls adipogenesis and glucose tolerance. Nat Cell Biol, 9:970-977, 2007.

Zhai, X., Meek, T. D. Catalytic Mechanism of Cruzain from Trypanosoma cruzi As Determined from Solvent Kinetic Isotope Effects of Steady-State and Pre-Steady-State Kinetics. Biochemistry, 57(22), 3176–3190, 2018.

ZHOU, A. Q., O'HERN, C. S., REGAN, L. Revisiting the Ramachandran plot from a new angle. Protein Science. 20:1166-1171, 2011.