UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE BIOCIÊNCIAS
BACHARELADO EM BIOTECNOLOGIA

Matheus de Bastos Balbé e Gutierres

Avaliação da Importância da Informação Estrutural no Processamento de Substratos pelo Proteossomo

TRABALHO DE CONCLUSÃO DE CURSO

Porto Alegre
2016

Matheus de Bastos Balbé e Gutierres

AVALIAÇÃO DA IMPORTÂNCIA DA INFORMAÇÃO ESTRUTURAL NO PROCESSAMENTO DE SUBSTRATOS PELO PROTEOSSOMO

Trabalho de conclusão de curso de graduação apresentado ao Instituto de Biociências da Universidade Federal do Rio Grande do Sul como requisito parcial para a obtenção do título de Bacharel em Biotecnologia.

Área de habilitação: Bioinformática

Orientador: Dr. Gustavo Fioravanti Vieira

Porto Alegre

2016

Este trabalho foi realizado no Núcleo de Bioinformática do Laboratório de Imunogenética no Departamento de Genética da Universidade Federal do Rio Grande do Sul, com apoio financeiro do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

AGRADECIMENTOS

Gostaria de agradecer primeiramente à minha família que permitiu que eu tivesse a oportunidade de realizar este curso: meus avós, Antonio Xavier Balbé e Inah Bastos Balbé e minha mãe Maria Beatriz Bastos Balbé. Também agradeço aos meus colegas de graduação que conviveram comigo ao longo deste período, bem como meu orientador Gustavo Fioravanti Vieira e meus colegas de laboratório Marcus, Martiela, Marcelo e Maurício.

RESUMO

O proteossomo é um sofisticado complexo proteico importante para a geração de peptídeos no citosol, constituindo-se como primeiro passo na via de processamento de antígenos do MHC-I. Sua estrutura evolutivamente conservada consiste em uma partícula central composta por quatro anéis sobrepostos, com ambas extremidades associadas a dois complexos regulatórios que reconhecem proteínas poliubiquitinadas, desenovelam sua estrutura e encaminham-nas para o interior da câmara proteolítica propriamente dita. O interior da partícula central é constituído por três compartimentos internos que isolam o substrato do ambiente citoplasmático. Duas antecâmaras localizam-se entre a junção dos anéis alfa e beta, e uma câmara catalítica contendo treoninas que atuam como nucleófilos localiza-se entre os anéis beta. Ao longo da última década diversos inibidores de proteossomo foram desenvolvidos apresentando uma estruturação semelhante a fita-beta na região de contato próximo à treonina catalítica. Além disso evidências na literatura apontam que proteases em geral reconhecem universalmente substratos em conformação estendida. Contudo é sabido que o interior das paredes da antecâmara interage com o substrato e altera suas propriedades físico-químicas, favorecendo a sua desestruturação. Dadas estas questões, o presente trabalho visa avaliar a influência da estrutura secundária do substrato no processamento proteossômico. Obtivemos 1659 cristais provenientes do PDB e assinalamos a estrutura secundária para cada resíduo, tal qual a probabilidade de ocorrer clivagens por meio de dois preditores independentes. Avaliamos, dentre os aminoácidos com alta probabilidade de corte, quais resíduos estavam presentes em hélices, fitas e voltas. Em comparação ao esperado, calculado em função da composição total do conjunto de dados, verificamos que há uma sobressalência de resíduos presentes em fitas beta em detrimento de voltas. Dentre os aminoácidos com baixa probabilidade de corte, percebemos o comportamento inverso, onde resíduos em estrutura de voltas estão mais presentes do que o esperado ao acaso neste conjunto. Além disso, obtivemos da literatura informações de proteínas digeridas *in vitro,* e utilizamos exclusivamente a probabilidade de formar cada estrutura secundária por resíduo para treinar redes neurais com o algoritmo *backpropagation*. Avaliamos o desempenho das redes por meio de curvas ROC, e para nossas melhores obtivemos um valor de área sob a curva de 0.75 e 0.70 (valores equivalentes aos obtidos pelos outros preditores comparados no estudo), no contexto de imunoproteossomo e proteossomo constitutivo, respectivamente. Estes resultados comparados a preditores utilizando codificações mais complexas demonstram que a estrutura secundária é informativa o suficiente para atingir um

desempenho equivalente. Este trabalho representa uma evidência indireta de que a propensão intrínseca de estruturação dos aminoácidos influencia a clivagem de substratos no interior da câmara catalítica proteossômica.

LISTA DE ABREVIATURAS

MHC Complexo Principal de Histocompatibilidade

MHC-I MHC de classe I

MHC-II MHC de classe II

TCR receptor da célula T

p-MHC complexo peptídeo-MHC

APC células apresentadoras de antígenos

ER retículo endoplasmático

TAP transportador associado à apresentação de antígenos

PLC do inglês *peptide loading complex*

UPS sistema ubiquitina-proteossomo

RP partícula regulatória 19S

CP partícula central 20S

FDA *Food and Drug Administration* (órgão regulador vinculado ao governo americano)

SS estrutura secundária

PSSM matriz de escore com posições específicas

AAA ATPases associadas a diversas atividades celulares

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1. Introdução

1.1 Sistema Imunológico

O termo imunidade é proveniente do latim *immunitas* e refere-se a proteção de senadores romanos contra demandas judiciais sofridas no período de seu mandato, sendo posteriormente adotado com o sentido de proteção a enfermidades. A complexa rede de moléculas, tipos celulares e tecidos que atuam em conjunto para proteger o organismo contra substâncias estranhas e patógenos compõe o chamado sistema imunológico. Este sistema é didaticamente subdividido em duas grandes partes: o sistema imune inato e o sistema imune adaptativo (Figura 1). O sistema imune inato é composto por barreiras químicas e físicas, diversos tipos celulares, como células fagocitárias, dendríticas e assassinas naturais, além de proteínas do sangue e citocinas. Sua resposta inicial representa a primeira linha de defesa do organismo e atua de forma pouco específica, sendo capaz de reconhecer padrões gerais conservados em diversos micro-organismos como por exemplo lipopolissacarídeos nas paredes celulares de bactérias.

Caso algum patógeno consiga evadir ou superar a imunidade inata, o sistema imune adaptativo representa uma maneira mais poderosa para garantir a eliminação do invasor por meio de uma resposta mais específica capaz, inclusive, de gerar memória imunológica. Existem dois tipo de imunidade adaptativa: a resposta humoral e a resposta celular. A imunidade humoral é particularmente útil na eliminação de patógenos extracelulares e é mediada por linfócitos B, que por sua vez sintetizam e secretam anticorpos. Anticorpos são glicoproteínas circulantes capazes de se ligar com alta afinidade a antígenos (moléculas reconhecidas por linfócitos e anticorpos) permitindo desse modo sua neutralização e eliminação por diversos mecanismos efetores. Micro-organismos patogênicos intracelulares, contudo, tornam-se inacessíveis aos anticorpos e para sua eliminação é necessária a resposta celular mediada por linfócitos T. Neste contexto, estas células adaptaram-se a reconhecer subprodutos de degradação de proteínas na superfície de células nucleadas. Este reconhecimento é determinado pela ação de proteínas especializadas denominadas Complexo Principal de Histocompatibilidade (MHC), cuja apresentação abrangem-se duas grandes vias: a do MHC de classe I (MHC-I) e MHC de classe II (MHC-II).

Das diversas populações funcionalmente distintas dos linfócitos T, destacam-se linfócitos T auxiliares CD4+ e linfócitos T citotóxicos CD8+. Linfócitos T auxiliares atuam no reconhecimento de antígenos exógenos apresentados pelo MHC-II e são capazes de secretar citocinas que estimulam diversos mecanismos da imunidade inata e específica. Linfócitos T citotóxicos por sua vez interagem com peptídeos apresentados no contexto do MHC-I e devem

induzir a morte da célula apresentadora quando conteúdos proteicos estranhos forem detectados, discriminando corretamente o próprio do não próprio.

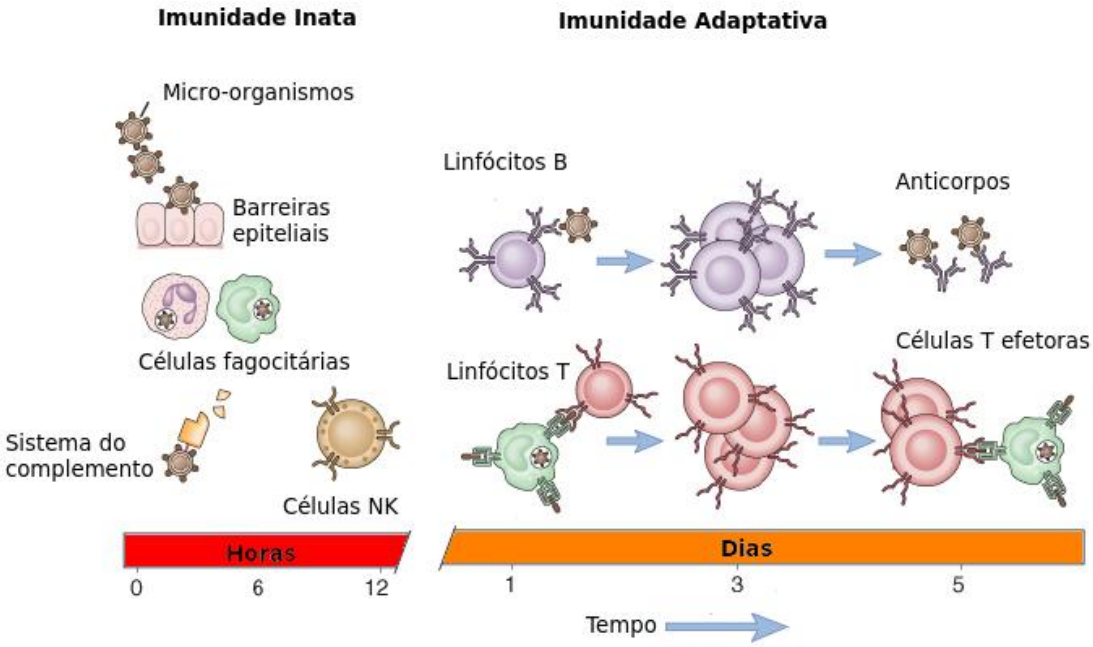

Figura 1. Componentes da Imunidade Inata e Adaptativa. A defesa inicial se dá por meio de mecanismos da imunidade inata. Posteriormente a imunidade adquirida age mediada por linfócitos B e T. Fonte: Modificado de *Cellular and Molecular Immunology, 7th Edition , 2012*

1.2 Vias de Apresentação de Antígenos

O desencadeamento da resposta celular mediada por células T é um evento dependente do resultado da interação entre o receptor da célula T (TCR) e o complexo peptídeo-MHC (p-MHC) na membrana de células apresentadoras de antígenos (APC). O lócus do MHC em humanos localiza-se em uma longa extensão do braço curto do cromossomo 6 (cerca de 3.5 kb) e tanto os genes do MHC-I quanto do MHC-II representam a região mais polimórfica do genoma. Além de diferirem nos tipos celulares em que são expressos, pois genes do MHC-II são principalmente expressos em APC's profissionais, como células dendríticas e macrófagos, enquanto o MHC-I é expresso por todas as células nucleadas, contrastam-se também quanto a procedência dos antígenos obtidos. Na via de apresentação pelo MHC-II antígenos são provenientes do ambiente extracelular, enquanto na via de apresentação pelo MHC-I são predominantemente derivados do ambiente intracelular.

A via do MHC-I (Figura 2) é composta por uma série de etapas bem definidas que

incluem desde a geração dos produtos peptídicos ao processo de maturação do MHC-I no retículo endoplasmático (ER) com subsequente translocação à membrana celular. A primeira etapa consiste na degradação em pequenos fragmentos peptídicos de proteínas intracelulares pelo proteossomo, um grande complexo com capacidade proteolítica. Estes fragmentos são adicionalmente clivados por proteases do citosol, onde a maior parte é destruída. Alguns peptídeos, entretanto, escapam à degradação ao serem transportados para o interior do ER por meio de um transportador embebido em sua membrana denominado transportador associado à apresentação de antígenos (TAP). No interior do ER ocorre a maturação do MHC-I, este que é um heterodímero formado por uma cadeia leve denominada β_2 microglobulina, uma cadeia pesada altamente polimórfica que possui uma fenda capaz de acomodar peptídeos com em torno de 8-10 aminoácidos, fator que é crucial para a estabilidade do complexo. Associado à TAP encontram-se chaperonas, proteínas especializadas e moléculas de MHC-I "vazias" (sem peptídeos em sua fenda), formando o chamado PLC (do inglês *peptide loading complex*). A chaperona tapasina interage com TAP, acoplando o transporte ao interior do ER com o carregamento na fenda. Quando peptídeos estáveis posicionam-se, as chaperonas desligam-se e o então estável complexo p-MHC-I é direcionado à superfície pelo complexo de Golgi.

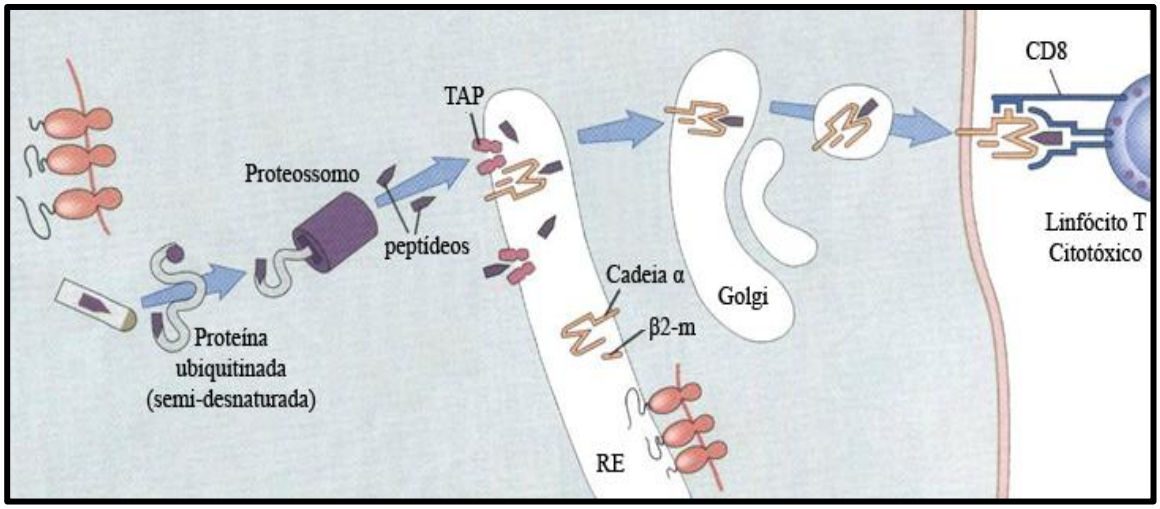

Figura 2. Via de apresentação de antígenos do MHC-1. Proteínas intracelulares são degradadas pelo proteossomo e posteriormente direcionadas a superfície da célula onde ocorre a interação entre o complexo pMHC e o TCR. Fonte: Modificado de *Cellular and Molecular Immunology, 4th Edition , 2000*

1.3 Proteossomo

Conforme já mencionado, como primeiro passo na apresentação de antígenos da via do MHC-I encontra-se o sofisticado complexo do proteossomo, componente central da principal via de degradação de proteínas em eucariotos, o sistema ubiquitina-proteossomo (UPS). Crucial para a homeostase celular, abrange funções desde a remoção de proteínas mal dobradas à regulação de diversos processos como controle da expressão gênica e ciclo celular [1, 2]. Além do proteossomo, o UPS é composto por uma cadeia de ubiquitina ligases que permitem uma regulação efetiva dos substratos proteicos a serem degradados por meio da ligação da pequena proteina ubiquitina de cerca de 8.5 kD a resíduos de lísina como modificação pós-traducional. Normalmente resíduos de ubiquitina ligam-se a outros resíduos de ubiquitina formando cadeias das mais variadas configurações, permitindo dessa forma o reconhecimento por parte do proteossomo como um sinal para destruição (Figura 3) [3].

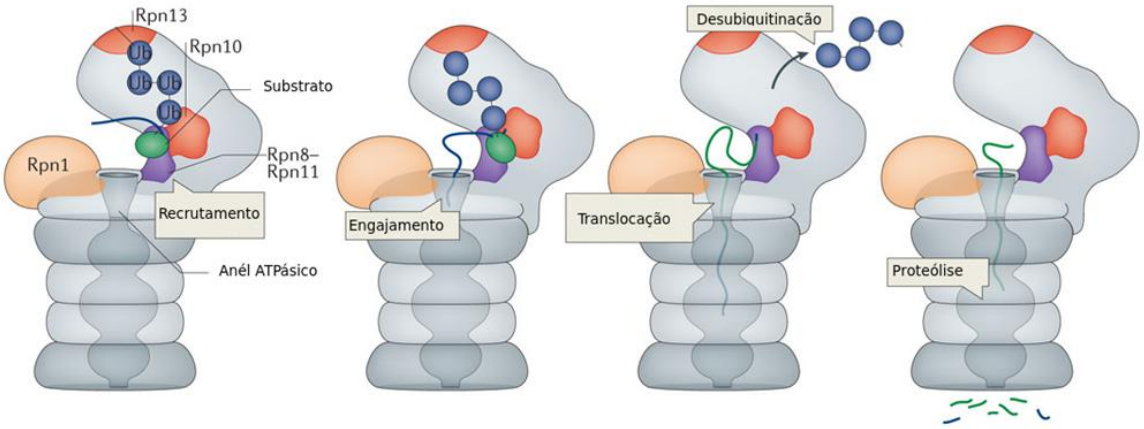

Figura 3. Etapas na via de degradação do proteossomo. A partícula regulatória reconhece proteínas marcadas para degradação ligadas a cadeias de ubiquitinas. Uma pequena região desestruturada é posicionada na entrada da partícula central e com gasto de ATP, a proteína é desestruturada e direcionada para o interior onde ocorre a proteólise propriamente dita. Fonte: Modificado de Sucharita *et al*, 2014

A estrutura evolutivamente conservada do proteossomo 26S de cerca de 2.5-MDa consiste em dois subcomplexos: uma partícula regulatória 19S (RP) associada a uma ou ambas as extremidades de uma partícula central 20S (CP) (Figura 3) [3]. A RP, por sua vez também subdivide-se em dois subcomplexos chamados de base e tampa (*lid).* A base é formada por um anel hetero-hexamérico de APTases da família das AAA ATPases e outras proteínas regulatórias não-ATPásicas cuja principal função é o reconhecimento do substrato, sua desestruturação com gasto de

ATP e translocação para o interior da CP. A tampa é formada pelo restante das proteínas presentes no complexo e à sua função atribui-se o papel de remover e reciclar as cadeias de ubiquitina.

A CP por sua vez é codificada por 14 genes diferentes. Possui estrutura de quatro anéis empilhados, sendo composta por dois anéis externos formados por sete subunidades alfa (α1–α7), e dois anéis internos constituídos por sete subunidades beta (β1–β7), totalizando vinte e oito subunidades com formato semelhante a um barril (α7β7β7α7). O interior deste barril é constituído por três compartimentos internos que isolam o substrato do ambiente citoplasmático, dos quais duas antecâmaras posicionam-se entre a junção dos anéis alfa e beta (α7β7 e β7α7), e uma câmara catalítica localiza-se na intersecção entre os anéis beta (β7β7) (Figura 4) [4].

A capacidade catalítica é proveniente das subunidades β1, β2, e β5, onde resíduos de treoninas na porção N terminal atuam como nucleófilos reativos que catalisam o rompimento da ligação peptídica do substrato. A estas proteínas foram associadas atividades semelhante à caspase (β1), tripsina (β2) e quimiotripsina (β5), por conta da especificidade de clivagem ocorrer após aminoácidos ácidos, básicos e hidrofóbicos, respectivamente.

Na maior parte das células, estresse oxidativo e citocinas pró inflamatórias induzem a substituição das subunidades catalíticas constitutivamente expressas β1, β2, e β5 pelas proteínas análogas β1i, β2i, e β5i, formando o chamado proteossomo imunológico ou imunoproteossomo. Este subtipo de protease apresenta a capacidade de clivar de uma maneira mais efetiva após resíduos hidrofóbicos e básicos. É sabido que imunoproteossomos são capazes de aumentar a produção de peptídeos capazes de serem apresentados pelo MHC-I, apesar de não haver diferença na taxa de processamento entre os proteossomos imunológico e o constitutivo [5]. Inclusive, é discutido que o proteossomo imunológico não seria essencial para uma resposta eficiente, visto que existem epitopos gerados preferencialmente pelo proteossomo constitutivo. Entretanto, não podemos negligenciar a importância do imunoproteossomo por esta evidência, visto que ele exerce diversos papéis na resposta imune, muitos deles ainda desconhecidos.

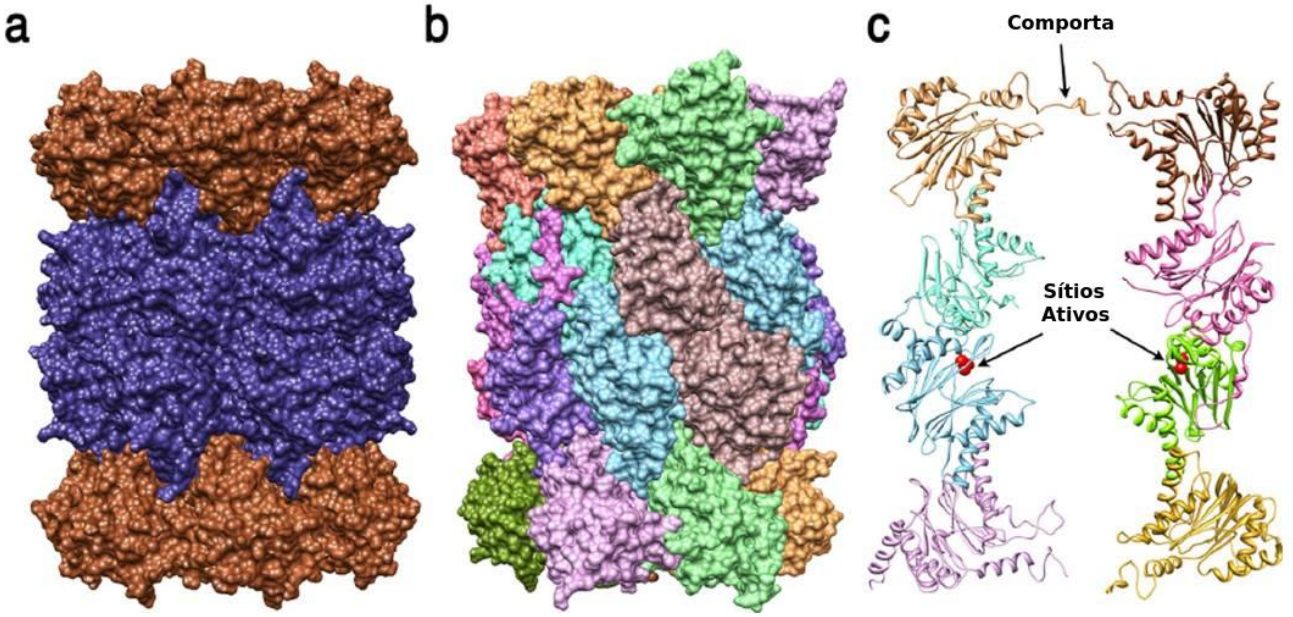

Figura 4. Estrutura da partícula central. a. 20S CP proveniente do archea *T. acidophilum*, evidencia-se os anéis externos formado por subunidades alfa e a os aneis internos formados por subunidades beta. **b.** 20S CP do fungo eucariótico *S. cerevisiae*, nota-se que cada uma das sete subunidades alfa e beta ocupam uma posição específica no complexo. **c.** Visão interna da 20S CP *S. cerevisiae* onde é possível perceber ambas antecâmaras e a câmara catalítica. Fonte: Modificado de Kish-Trier *et al*, 2013

1.4 Inibidores e Estruturação Secundária

Devido ao papel central na homeostase da célula mediando a degradação de proteínas como p53, Bcl-2 e NF-κB, envolvidas na progressão do ciclo celular e apoptose, diversos esforços foram tomados na direção do protossomo como um alvo para terapias de câncer. Em consequência ao rápido acúmulo no interior da célula de proteínas regulatórias incompatíveis, a inibição do proteossomo induz a uma cascata que leva à apoptose celular, e inclusive células tumorais normalmente possuem maior nível de atividade do proteossomo do que células normais, tornando-as mais sensíveis à terapia.

Em 2008 foi aprovado pela *Food and Drug Administration* (FDA, dos Estados Unidos) para o tratamento de mieloma múltiplo o inibidor do proteossomo bortezomib, cujo nome comercial é Velcade. Da classe dos ácidos borônicos, atua reversivelmente nas atividades semelhante a quimiotripsina e caspase, com pouco ou nenhum efeito na atividade semelhante a tripsina. Este foi o primeiro inibidor utilizado clinicamente, contudo uma grande quantidade de outros inibidores de segunda geração vem sendo desenvolvidos. Da classe das epoxi cetonas,

aprovado em 2012 pela FDA para tratamento de mieloma múltiplo em pacientes que receberam ao menos duas terapias prévias incluindo tratamento com Velcade e terapias imunomodulatórias e demonstraram progressão da doença dentro de 60 dias após a última terapia, encontra-se o inibidor carfilzomib, de nome comercial Kyprolis. Em comparação com bortezomib, carfilzomib atua como um inibidor irreversível com maior seletividade da atividade semelhante a quimiotripsina em relação às atividades semelhante a tripsina e caspase [6].

Seguindo a nomenclatura de Schechter e Berger, os aminoácidos do substrato onde ocorre a clivagem são denominados à porção N-terminal de P1 e à porção C-terminal de P1', com a denominação crescente à medida que se afastam da região da clivagem (... P3, P2, P1 | P1', P2', P3'...). Substratos interagem próximos à treonina catalítica preenchendo bolsos nas paredes da câmara que vão de P4 até P1 (região chamada de *non-primed*), e de P1' até P4' (região chamada de *primed*), determinando o sítio de clivagem e conferindo orientação e estabilidade para o ataque nucleofílico. Tanto carfilzomib quanto bortezomib possuem uma porção peptídica que interage com estes bolsos, além de uma porção reativa capaz de formar uma ligação covalente com a treonina catalítica, abolindo desta forma o sítio proteolítico em questão. O paradigma estrutural importante para a extrema maioria de inibidores do proteossomo, incluindo carfilzomib e bortezomib, é o mimetismo estrutural de uma estrutura secundária semelhante a fita-beta na região de contato próximo à treonina catalítica [7].

A estrutura secundária (SS) de uma proteína consiste em padrões estruturais no esqueleto de carbonos resultantes da formação de pontes de hidrogênio ente os aminoácidos que a compõe. A cadeia polipeptídica pode adotar uma conformação de hélice onde o esqueleto de carbonos localiza-se para o interior da hélice enquanto as cadeias laterais posicionam-se para o exterior. Em uma alfa hélice, a estrutura estabiliza-se por meio de pontes de hidrogênio entre o grupamento amina de um aminoácido com o grupamento carboxílico do quarto aminoácido seguinte. Contrastando o formato de hélice, na estrutura denominada folha beta a cadeia polipeptídica estende-se como uma "folha" composta por diversas fitas, e a SS estabiliza-se por meio de pontes de hidrogênio entre o grupamento N-H de uma fita e o grupamento C=O da fita vizinha. Alças ou voltas ocorrem intercalando hélices e fitas e caracterizam-se pela flexibilidade apesar de que quando pequenas podem apresentar elevada rigidez.

Em 2005, por meio de comparações entre mais de 1500 cristais de diversos tipos de proteases complexadas a seus ligantes, desde proteases de serina, cisteína, ácido aspártico, metaloproteases e treonina (proteossomo), foi proposto que compartilhariam um padrão estrutural

de ligação do substrato ou inibidor em conformação de fita beta estendida, seja o ligante em questão um peptídio ou não, de uma maneira universal. Diversas também são as razões para este fenômeno, visto que proteínas com dobramento correto não expõem fitas beta, pois são basicamente compostas por hélices, voltas e folhas, e não seria coerente a célula gastar energia para sintetizar proteínas que seriam rapidamente sujeitas à degradação. Além disso, outros elementos de SS possuem cadeias laterais "protegendo" o esqueleto de carbonos, onde a extensão linear de uma fita beta maximiza a exposição ao solvente e a resíduos das proteases. Inclusive peptídeos são processados principalmente entre elementos de SS e as taxas de clivagem aumentam quando pontes de hidrogênio são rompidas, promovendo a desnaturação do substrato. Por fim, a conformação ligada do substrato no sítio ativo é pequena para acomodar hélices, voltas ou folhas. Também foi predito que clivagens ocorreriam exclusivamente em regiões estendidas, e uma forma de permitir acesso aos sítios então escondidos seria alterar o equilíbro entre substrato estruturado/desestruturado em direção à forma desestruturada [8, 9].

O acesso à CP é controlado por um canal de cerca de 13 angstrons de diâmetro formado pelas subunididades alfa, cuja porção N-terminal pode bloquear ou permitir a entrada do substrato, atuando como uma comporta. Devido ao fato de ser muito estreito, a única forma de o peptídeo entrar na CP é estando desestruturado e linearizado. Apesar disso, tanto a antecâmara (volume de cerca de 59 nm^3) quanto a câmara catalítica (volume de cerca de 84 nm^3) possuem muito espaço em seu interior para formas mais complexas de estruturação. Em 2006, foi observado que as antecâmaras poderiam atuar armazenando substratos enquanto a degradação ocorre na câmara catalítica em situações onde a translocação e desenovelamento são mais lentos do que a proteólise, inclusive sendo demonstrado que a população de substratos ocupando a antecâmara estaria em um estado parcialmente enovelado, resultando em um modelo proposto em que a medida que a proteólise ocorreria na câmara catalítica permitiria o deslocamento dos substratos quando as restrições espaciais então permitissem [10].

Um argumento termodinâmico, entretanto, permite hipotetizar que substratos ocupando a antecâmara tenderiam a uma estabilização de seu estado nativo, o que limitaria tanto os sítios de clivagem acessíveis pelas subunidades catalíticas quanto estancaria a degradação. Em 2010, foi demonstrado que o interior das paredes da antecâmara interage ativamente com o substrato, alterando suas propriedades físico-químicas favorecendo a sua desestruturação, permitindo assim expor os sítios para que a hidrólise ocorra. Então o substrato teria suas propriedades de equilíbrio perturbadas, tentendo à desestruturação, conforme havia sido predito

alguns anos antes pela análise dos cristais. Entretanto, pouco é discutido neste trabalho em questão quanto a influência da SS no momento da clivagem, e motivados pela estrutura dos inibidores aqui apresentados hipotetizamos que a SS talvez seja um fator importante no momento do corte [11].

1.5 Imunoinformática

Imunoinformática compreende o estudo e desenho de algoritmos para mapear potenciais epitopos da imunidade adaptativa, tanto de linfócitos B quanto de linfócitos T, permitindo direcionar a pesquisa em bancada e reduzir o custo e o tempo no desenvolvimento de vacinas. No contexto de epitopos de linfócitos T citotóxicos CD8+, na modelagem *in silico* da via de MHC-I cada etapa apresenta importantes implicações para o sucesso da vacina a ser desenvolvida, desde a predição da estabilidade e afinidade do epitopo ao se ligar ao MHC-I, a translocação de peptídeos ao retículo por TAP, e à sua geração pelo proteossomo. Para todos estes passos existem atualmente uma série de programas disponíveis. Cada uma destas etapas seguem padrões ou regras específicas e com cada vez mais dados de bancada disponíveis alimentando grandes repositórios de dados imunológicos, diversos algoritmos de aprendizado de máquina vem sendo empregados para reconhecer tais padrões e gerar estas ferramentas.

Em relação a clivagens pelo proteossomo, o programa mais utilizado pela comunidade científica é o Netchop 3.1 [12, 13]. Este programa consiste em um conjunto de redes neurais treinadas sobre dados *in vitro* de digestão proteossômica (proteossomo constitutivo) e eluição de ligantes de MHC-I (imunoproteossomo), visto que este último conjunto de peptídeos representam na porção C terminal a informação sobre o local de clivagem. A principal característica de algoritmos de aprendizado de máquina como redes neurais é a capacidade do programa aprender as regras envolvidas no reconhecimento de padrões sem haver sido explicitamente programado para isso. Redes neurais são modelos computacionais inspirados nas redes de neurônios do sistema nervoso, onde neurônios artificiais simulam o comportamento de receber sinal, processar este sinal e disparar um impulso para outros neurônios. Normalmente são constituídas por uma camada de entrada, onde entra o sinal inicial (o dado informativo bruto), e a partir dela neurônios processam o valor de acordo com uma função matemática e o transmitem para neurônios de camadas escondidas, que prosseguem sucessivamente até atingir a camada de saída, geralmente composta por um único neurônio. No exemplo acima citado, as redes foram treinadas com dados *in vitro* (Netchop 20S, chamado constitutivo) e ligantes de MHC-I (Netchop Cterm, chamado imunológico) utilizando três diferentes esquemas para codificar a informação: codificação *sparse*, onde aminoácidos são

representados por 0 ou 1 (Valina: 01000000000000000000, por exemplo), codificação *blosum*, onde o aminoácido é representado pela linha da matriz *blosum* 50, e uma codificação usando modelos ocultos de Markov, onde é construída uma matriz de pesos descrevendo os motivos de clivagem positivos do conjunto de treinamento. Um outro exemplo de preditor de clivagens proteossômicas é o ProteaSMM, o qual é uma matriz de posições específicas construída a partir dos dados de quantificação da digestão proteossômica de três proteínas [14-17]. A matriz é construída a partir de motivos de clivagem de 10 aminoácidos (P6 - P4') minimizando uma função objetiva, que quanto maior a quantidade de fragmentos gerados de um determinado sítio, maior o escore daqueles aminoácidos naquela posição. Dessa forma, para uma sequência de 10 aminoácidos, somando o valor de escore de cada uma das posições é obtido o valor que representa o logaritmo em picomol da quantidade de fragmentos geradas entre as posições 6 e 7. É importante considerar que nenhum destes programas utiliza qualquer tipo de informação estrutural na codificação dos dados para seus respectivos treinamentos.

Não é possível avaliarmos diretamente a estrutura do substrato no interior do proteossomo, entretanto, comparando os dados estruturais originais nas proteínas digeridas e cruzando-os com a informação dos locais de clivagem predita por estes diferentes métodos, podemos avaliar nossa hipótese do impacto da estruturação intrínseca destas sequências de resíduos por meio de análises estatísticas para verificar se há alguma preferência de proteólise em alguma estrutura secundária específica.

2. Objetivos

2.1 Objetivo Geral

O objetivo deste estudo consiste em avaliar se a estruturação secundária do substrato é capaz de influenciar de alguma forma o processamento proteossômico no interior da câmara catalítica.

2.2 Objetivos Específicos

A. Gerar um banco de dados contendo estruturas cristalográficas de proteínas humanas e virais do *Protein Data Bank*.

B. Verificar, através de teste de Qui-quadrado, se os eventos proteolíticos ocorrem ao acaso ou se ocorrem preferencialmente em alguma das estruturas secundárias das proteínas.

C. Buscar na literatura informação de proteínas digeridas *in vitr*o pelo proteossomo.

D. Predizer a estrutura secundária das proteínas deste conjunto de dados, caso não haja cristal disponível.

E. Treinar redes neurais utilizando diferentes tamanhos de janelas de entrada, codificação de dados e parâmetros de treinamento.

F. Avaliar a capacidade de aprendizado das nossas redes geradas, comparando-as com o Netchop e o ProteaSMM.

3. Materiais e Métodos

3.1 Conjunto de Dados

3.1.1 Estruturas Cristalográficas do *Protein Data Bank*

Com o intuito de obter cristais representativos com proteínas que são de fato processadas pelo proteossomo, primeiramente filtramos por proteínas humanas ou virais. Além disso, empregamos uma série de outras restrições: cristais contendo cadeias de DNA e RNA ou híbridos foram excluídos, bem como cristais com ligantes ou resíduos modificados; a resolução deveria apresentar valores númericos inferiores a 2.5 angstroms e, por fim, sequências compartilhando 50% de identidade eram representadas por uma única estrutura. Recuperamos 1419 cristais de proteínas humanas e 573 de proteínas virais. Em ambos os conjuntos, entretanto, o buscador do PDB incluiu diversas estruturas do MHC-I complexado a epitopos. Estas estruturas foram removidas por meio de um programa que reconhecia palavras chaves como "MHC","HLA","COMPLEX" nas descrições de cabeçalho e outras informações dos arquivos pdb. Isto resultou na remoção de 126 cristais provenientes de vírus e 162 provenientes de humanos. Apesar de todo cuidado, 9 cristais eram complexos de proteínas humanas e virais e estavam presentes em ambos os conjuntos e também foram excluidos. Este processo resultou em 1659 cristais compondo o conjunto de dados do PDB para as análises posteriores.

3.1.2 Busca na literatura por ensaios de degradação *in vitro* pelo proteossomo

Até o presente momento existem dados de degradação *in vitro* para quatro proteínas: **beta-caseina** digerida pelo proteossomo 20S constitutivo [15]; **enolase 1 fúngica** digerida por ambas isoformas do proteossomo [14]; **proteína Nef de HIV-1** digerida em um ensaio com ambos os proteossomos presentes e **proteína Príon** recombinante digerida por ambas as isoformas do proteossomo (Tabelas 1 e 2) [18],[16]. Estas proteínas compõe o conjunto de treinamento para construção das redes neurais. O conjunto de validação é composto por 16 peptídeos de 17-30 aminoácidos de extensão provenientes das proteínas GAG e TAT de HIV-1, clivadas tanto por proteossomo constitutivo, quanto imunoproteossomo [19]. Devido a restrição do ProteaSMM de utilizar decâmeros, como tamanho mínimo de sequência para a predição de clivagem (entre os aminoácidos das posições 6 e 7). Neste exemplo os cinco primeiros e os três últimos sítios não são analisáveis. Por esta razão a validação se dá de duas formas: para comparação de nossas redes com Netchop, todos os sítios positivos (139) e sítios negativos (229) no contexto de proteossomo constitutivo são informativos, da mesma forma que no contexto do

imunoproteossomo todos os sítios sítios positivos (140) e sítios negativos (228) são informativos. Para comparações com ProteaSMM, tanto em imunoproteossomo quanto no proteossomo constitutivo os sitivos positivos analisáveis reduzem-se a 99 enquanto os sítios negativos reduzem-se a 140 (Tabelas 3 e 4).

Tabela 1. Mapas de clivagens com proteossomo constitutivo.

Proteína	**Sequência e sítios de clivagem**	**Sítios +**	**Sítios -**
Enolase	AVS|K|VY|AR|SV|YDSRGNPTV|E|VEL|T|TEKG|VFR|SIVPSGA|STG|VHE|A|LE|M|RD|GDKSKWMGKGVLHA|V|KNV|NDVIAP|A|F|V|K|A|N|IDV|KDQKAV|D|D|FLIS|LDGTANKS|KLG|ANAILGV|SLAAS|R|A|AAAEKNV|P|LYK|H|L|AD|LSKSKTS|PYVL|PVPFLN|VL|NGGS|HA|GGA|L|A|LQ|E|F|MIAPTG|A|KTFAE|AL|RI|GS|EV|Y|H|N|L|KSLTKKR|YGAS|AGNVG|D|EGGVAPNI|Q|T|A|E|E|ALDLIVD|AIKA|AGHDGKVKIGLDCA|SSEFF|KDGKYD|L|DFKNPNSDKSKWL|TG|PQ|L|ADLYH|SL|M|KRY|PIV|S|I|EDPFAED|D|WEAW|SH|FFKT|A|GIQI|V|AD|D|L|TVTNPKRI|AT|AIEK|K|A|AD|ALLLKV|NQIGTL|SESIKAA|QDSFAAGW|GVMV|SHRSGETEDTF|I|A|DLV|V|G|L|R|TGQIKTG|APARSE|RLA|KLNQL|LRIE|EE|LGD|N|A|VF|A|GE|N|F|HH|GD|KL	141	294
Caseína	REL|EE|LNVP|G|EIVESLSSSEESITRINKK|IEKF|QSEEQQQTEDELQD|KI|HPFAQ|T|Q|S|L|V|Y|P|FPGPIP|N|SL|PQ|NI|P|PLT|QT|PVVV|PPF|L|Q|PEVM|GVSKVKEA|MAPKHK|EM|PFPKY|PVE|PFT|E|SQSLTL|TDVENL|HL|PLPLL|QS|W|MHQPHQPLPPTVMF|PPQSVL|SL|SQ|S|KVL|PVP|QKA|V|PYP|Q|RD|MPIQ|A|F|LL|Y|Q|E|PVL|G|PV|R|G|PFPIIV	66	142
Prion	SKKRP|KP|GG|G|W|NT|G|GSRY|PGQ|G|S|P|G|GNRY|PPQ|G|GG|G|W|GQPHGGGW|GQPHGGG|W|GQ|PHGGG|W|GQPHGGGGW|GQG|GSH|S|Q|W|NKPSK|PK|T|N|MK|H|V|A|G|A|AA|AG|AVVGGL|GGYM|L|GSVM|S|R|P|LIHF|G|N|D|YEDR|Y|Y|R|E|N|M|Y|RY|PNQV|Y|Y|RPVD|QY|SNQNNFV|HDCVNITVKQH|TVTT|TTKGENF|T|ETDIKI|MERVVEQMCITQ|Y|QRE|S|QA|YY|QR|GA|S	82	128
Nef*	GGKW|SK|SSVVG|W|PTVRE|R|M|R|RAE|PAA|DGV|GA|A|SRDLEKHG|AITS|SNTAAT|N|AACAW|LE|AQEE|E|E|VGF|PV|T|P|Q|V|P|L|RP|M|TY|KA|A|V|DL|SH|F|LKE|K|G|GL|E|GL|IH|SQ|R|RQ|D|I|L|DLW|I|Y|H|T|QGY|F|PD|W|Q|N|Y|TPGPG|VR|YPLTF|GW|CYK|L|V|PVEPDKV|EEANKGENTSLLHPV|SL|HGMD|DPER|E|VL|E|W|RF|DS|RL|A|FH|H|V|A|R|EL|HPEY|F|KN|C	93	111

*Nef foi digerida com ambos proteossomos, não exclusivamente constitutivo.

Tabela 2. Mapas de clivagens com proteossomo imunológico.

Proteína	**Sequência e sítios de clivagem**	**Sítios +**	**Sítios -**
Enolase	AVSKV|YA|R|S|V|YDSRGNPTV|E|VEL|TTEKGVFR|SIVPS G|A| STGVHE|AL|EM|RDGDKSKW|MGKGVL|H|A|VK|N|V|NDV I|APA|F|V|K|A|NIDV|KDQKAVD|DFLIS|L|DGTANKSKL|G| AN|AILGVSL|AAS|RAA|AAEKNVPL|YKHL|ADLSKSK|TS| PYVL|PVPFL|NV|L|NGGSHAGGAL|AL|QEF|MI|APTGA|KT F|AEAL|RI|GSEVY|HNL|K|SLTK KR|Y|GA|SA|GNVG|D|EGGVA|PNI|QT|A|E|E|AL|DLIVD|AI| KA|AGHDGKVK|IGLDCASSEFF|KDGKYDL|DFK|NPNSD KSKW|LTGPQL|ADLY|HSL|MKR|Y|PIVSI|EDPFA|EDDWE AWSH|FF|KTA|GIQ|IV|A|D|DL|TVT|NPKRI|A|TAIEK|KA|A DALL|L|KV|N|QI|GTL|SESI|KA|A|QDSF|AAGWGVMV|SH| RSGETEDTF|I|ADLVVGL|RTGQIKT|G|APA RSERL|AKL|N|QLL|RIEE|EL|GD|NAVF|A|G|ENFHHGDKL	124	311
Prion	SKK|RPK|P|GG|G|W|NT|G|GS|R|Y|P|GQ|GSPGG|N|RY|PPQ| G|GGG|W|GQPHGGG|W|G|Q|PHGGGW|GQPHGGGW|GQP| HGGGG|W|GQ|G|GSH|SQ|W|NKPS|K|PK|T|N|M|K|HV|A|G|A |A|A|AG|A|VV|GGLG|GY|M|L|G|SV|M|S|R|P|LIH|F|G|ND|Y|E DR|Y|Y|RE|N|M|Y|R|Y|P|NQV|Y|Y|R|PVDQ|Y|SN|Q|NN|FV| H|D|CVNIT|VKQH|TVTT|TT|KGE|N|F|TE|TD|IKI|ME|RV|VE QM|CIT|QY|Q|R|E|S|QAYYQR|GAS	107	103

3.2 Anotação e Predição da Estrutura Secundária

3.2.1 Stride

Stride é um programa que assinala a estrutura secundária de uma proteína a partir das coordenadas atômicas da molécula [20]. Para cada estrutura cristalográfica obtidas do PDB, rodamos este programa para obter a estrutura secundária de cada um dos resíduos. A saída do Stride utiliza 7 letras para representar cada tipo de estrutura secundária: H para alfa hélices, G para hélices 3-10, I para hélices Pi, E para conformações estendidas, B para pontes beta isoladas (fitas beta com apenas uma ponte de H), T para *turn* e C para voltas ou nenhuma das anteriores. Para nossas análises agrupamos H + G + I como hélices, E + B como conformações estendidas ou fitas beta e T + C como voltas.

Tabela 3. Mapas de clivagens de peptídeos de validação.

Proteossomo	Peptídeo	Mapa de clivagens	Sítios +	Sítios -	A.P * +	A.P -	SS
I	nat10	PEVIPMF|S|AL|SE|GATPQ|D|L|NTML|NTVGGH	8	20	8	12	Spider
C	nat10	PEVIPMFS|ALSE|GATPQ|D|LNTML|NTVGGH	5	23	5	15	Spider
I	nat11	NNPP|IPVG|E|I|Y|K|R|W|II|L|G|L|N|KI|V	14	8	11	3	Spider
C	nat11	NNPPIPVG|EIY|K|R|W|IIL|G|L|NKI|V	9	13	8	6	Spider
I	nat12	KRWIILGL|NKIVRMYSPVSIL|D	2	19	1	12	Spider
C	nat12	K|R|WIIL|G|L|NK|I|V|R|M|Y|SPV|S|I|L|D	15	6	10	3	Spider
I	nat13	KALGPA|A|TL|EEMM|T|A|CQGVGGPGH	6	17	6	9	Spider
C	nat13	KALGPAATL|EEMM|T|A|CQGVGGPGH	4	19	4	11	Spider
I	p17	YKLK|HI|VW|A|S|RELER|F|AVNPGL|L|E|V|TS|E|GC	13	16	10	11	Spider
C	p17	YKLKHIVWASREL|ERFAVNPGLLEVTSEGC	1	28	1	20	Spider
I	p24	ALSEGATPQD|LNTML|NTV|GGHQA|AMQML	4	23	4	15	Spider
C	p24	ALSE|GATP|Q|DLNTM|L|NTVGGHQA|AMQML	6	21	5	14	Spider
I	p_seq3	RLIY|A|T|R|QL|Q|R|F|A|V|N|PGL|LI|T	13	7	9	3	Spider
C	p_seq3	RLIYA|T|R|QL|Q|R|F|A|V|N|PGL|LI|T	12	8	9	3	Spider
I	p_seq4	YAIP|Q|A|L|N|T|L|L|N|TV|GGHQAA	10	9	8	3	Spider
C	p_seq4	YAIP|Q|A|L|N|T|L|L|N|TV|GGHQAA	10	9	8	3	Spider
I	tat1	MEPVD|PRLEPWKHPG|SQPKTA|CTNC|Y|C|K	6	21	2	17	Spider
C	tat1	MEPVD|PRLEPWKH|PG|SQPKTA|C|TN|C|Y|C|K	9	18	5	14	Spider
I	tat2	CFHCQVC|FI|TK|GLGISY|G|RKK|RR	6	16	5	9	Spider
C	tat2	CFHCQVC|FITK|GLGISY|GRKK|RR	4	18	3	11	Spider
I	nat10mod	R|FII|PXF|T|A|L|SGGRR|A|L|L|Y|GATPY|AI|G	13	13	9	8	Psipred
C	nat10mod	R|FII|PXFT|A|L|SGGRR|A|L|L|Y|GA|TPYA|I|G	13	13	9	8	Psipred
I	nat11mod	RAIPIPA|GTL|L|SGGGR|AIYK|R|W|AI|L|G	9	16	7	10	Psipred
C	nat11mod	R|A|I|PIP|A|GT|LL|SGGGR|A|IY|K|R|W|A|I|L|G	16	9	10	3	Psipred
I	nat12mod	RWL|L|L|GL|NPLV|G|GGR|L|Y|SPTSI|L|G	11	12	6	9	Psipred
C	nat12mod	RWLLLGLNPLV|GGGRLYSPTSILG	1	22	1	14	Psipred
I	nat13mod	R|AL|GPA|A|TL|QTPWTA|SL|GV|G	8	11	4	7	Psipred
C	nat13mod	R|AL|G|PA|A|T|L|Q|T|P|WTA|SLG|VG	12	7	8	3	Psipred
I	p_seq1	FVIH|R|L|EPWL|HPG|SQHI|TA|S|TN	8	13	4	9	Psipred
C	p_seq1	FVIH|R|L|E|PWL|H|PGSQ|H|I|T|A|STN	11	10	8	5	Psipred
I	p_seq2	YVL|F|L|T|K|GL|SI|SY|L|GKK	9	7	5	3	Psipred
C	p_seq2	Y|V|L|F|L|T|K|GL|S|I|SYL|GKK	11	5	5	3	Psipred

* A.P: Analisáveis pelo ProteaSMM. Estão destacados em amarelo os sítios não analisáveis pelo ProteaSMM, o qual necessita minimamente de decâmeros. Dessa maneira os 5 primeiros e os 3 últimos aminoácidos são excluídos.

Tabela 4. Resumo dos dados de validação analisáveis por Netchop

Predição SS	Número de Peptídeos	Proteossomo	Sítios +	Sítios -
Spider	10	Imunológico	82	156
Spider	10	Constitutivo	75	163
Spider + Psipred	16	Imunológico	140	228
Spider + Psipred	16	Constitutivo	139	229

Tabela 5. Resumo dos dados de validação analisáveis por ProteaSMM

Predição SS	Número de Peptídeos	Proteossomo	Sítios +	Sítios -
Spider	10	Imunológico	64	94
Spider	10	Constitutivo	58	100
Spider + Psipred	16	Imunológico	99	140
Spider + Psipred	16	Constitutivo	99	140

3.2.2 Spider e Psipred

Em relação a nosso conjunto de Dados obtidos do PDB, podíamos assinalar a estrutura secundária exata utilizando Stride, entretanto para isso é necessário a estrutura cristalográfica em questão. Apesar de haver cristais para algumas das proteínas do conjunto de treinamento, nosso método de codificação do *input* requere a probabilidade de formação das três principais estruturas secundárias para cada aminoácido da sequência. Dessa maneira optamos por predizer a estrutura secundária por meio do programa Spider para as quatro proteínas do conjunto de treinamento e para os 16 peptídeos do conjunto de validação [21]. Para predizer a estrutura secundária com Spider é necessário construir uma matriz de escore com posições específicas (PSSM) a partir do alinhamento múltiplo de sequências, permitindo reconhecer relações distantes entre proteínas. Para isso, utilizamos PSI-BLAST para construir as PSSMs dos conjuntos de treinamento e validação para alinhar estas sequências com a base de dados proteícos não-redundantes (nr) que contém sequências traduzidas obtidas das bases de dados GenBank, EMBL, DDBJ e sequências proteicas de repositórios como SwissProt e PDB com os seguintes parâmetros: evalue = 10, num_iterations = 2, matrix = blosum45. Para seis peptídeos do conjunto de validação, por terem resíduos modificados, não foi possível gerar as PSSMs, de modo que utilizamos o programa Psipred 3.5, visto que este programa permite predizer a estrutura secundária com apenas uma única sequência, sem a necessidade de construir PSSM [22]. Contudo, esta maneira de

predição não é tão confiável, o que nos motivou a separar as análises de validação utilizando as 10 proteínas preditas pelo Spider e todas as 16 proteínas preditas com Psipred + Spider (Tabelas 4 e 5).

3.3 Teste Qui-quadrado

Para avaliarmos se ocorre a clivagem preferencialmente em alguma estrutura, para cada aminoácido do conjunto de dados do PDB anotamos a estrutura secundária e a probabilidade de ocorrência de corte. Assim, podemos cruzar as informações e formular um teste qui-quadrado que tem como racional medir possíveis discrepâncias entre um conjunto observado e um esperado teórico. Nosso esperado teórico é calculado em função da composição de estrutura secundária de todo o conjunto de dados. Por exemplo, se 50% dos resíduos do conjunto de dados do PDB estiverem em estrutura de alfas hélices, e a clivagem é um fenômeno biológico que não sofre influência da estrutura secundária, esperaríamos encontrar cerca de 50% das clivagens nesta estrutura. Para isso, ordenamos e obtivemos 10% dos resíduos com maior probabilidade de corte, que correspondem ao nosso *observado*. Assim, se estes resíduos estiverem enriquecidos com alguma estrutura secundária em comparação ao esperado teórico, representa uma evidência indireta de que este fenômeno biológico pode influenciar clivagens. De forma análoga, ordenamos e obtivemos 10% dos resíduos com menor probabilidade de corte, assim se este conjunto de aminoácidos que assumimos que não ocorrem clivagens estiver apresentando sobressaliência de alguma estrutura secundária também representará uma forma de avaliar como se dá a composição de estrutura secundária em aminoácidos onde a clivagem é desfavorecida. Para este cálculo usamos a função *chisquare* da biblioteca scipy.

3.4 Construção de Redes Neurais

Com o auxílio da biblioteca Pybrain, podemos construir e treinar redes neurais utilizando Python, de uma forma bastante simples. Desse modo, construímos redes neurais do tipo *feedforward*, onde a informação é unidirecional da camada de entrada à camada de saída, sem conexões recorrentes entre os neurônios formando ciclos. A estrutura das nossas redes é composta por uma camada de entrada variando de 5 a 9 aminoácidos, e cada aminoácido é representado pela probabilidade de formar hélices, fitas ou voltas. Por exemplo, para a sequência de aminoácidos VIPMF, a camada de entrada é composta por 15 neurônios (pois cada aminoácido é representado por 3 neurônios) e a informação que alimentará esta camada é: 0.116, 0.053, 0.259, 0.322, 0.448,

0.672, 0.775, 0.522, 0.453, 0.354, 0.229, 0.173, 0.231, 0.18, 0.187, onde os 5 primeiros valores representam a probabilidade de formar hélices, os 5 conseguintes representam a probabilidade de formar voltas e os cinco últimos representam a probabilidade de formar fitas. Então as camadas de entrada variavam de 15 a 27 neurônios, com 20 neurônios escondidos e um neurônio na camada externa, onde todos os neurônios eram conectados com todos. As redes foram treinadas com algoritmo *backpropagation*, onde quando apresentadas por uma sequência positiva (ocorre clivagem) o neurônio de saíde deveria informar-nos 1, e ao contrário, quando alimentada com uma sequência negativa (não ocorre clivagem) o neurônio deveria informar-nos 0. Alguns dos parâmetros do algoritmo *backpropagation* de treinamento utilizados são momento, taxa de aprendizado e proporção de validação, as redes aqui apresentadas utilizamos momento=0.016, taxa de aprendizado=0.034 e proporção para validação=0.20. Motivados pela recente evidência de que estas isoformas de proteossomo não diferem do ponto de vista qualitativo, isto é, não há sítios exclusivos de imunoproteossomos que proteossomos constitutivos não são capaz de clivar, e vice-versa, consideramos como sítios de clivagem se ao menos uma isoforma clivasse, e como sítios de não-clivagem apenas se não houve corte em ambas (no caso de príon e enolase).

3.5 Validação por Curvas ROC

Na teoria de detecção de sinais, Característica de Operação do Receptor, do inglês *Receiver Operating Characteristic* (ROC) representa uma forma gráfica de avaliar o desempenho de um classificador binário. Neste gráfico, a taxa de verdadeiros positivos (ou sensitividade) é traçada em função da taxa de falsos positivos (ou 1 - especificidade), para cada valor de *threshold* possível do classificador (entre 0 e 1). Em um preditor aleatório, ambas estas taxas crescem de maneira constante, gerando um gráfico próximo a uma reta, caso a sensitividade crescer mais rápido do que a taxa de falsos positivos, o gráfico passa a ter um comportamento de curva, e a área sobre a curva representa uma maneira objetiva de avaliar a qualidade do classificador.

4. Resultados

4.1 Aminoácidos com alta probabilidade de corte apresentam sobressalência de resíduos presentes em fitas beta.

Para cada um dos cristais provenientes do PDB, rodamos os programas ProteaSMM constitutivo, ProteaSMM imunológico, Netchop 20S (constitutivo) e Netchop Cterm (imunoproteossomo), e obtivemos o valor associado a probabilidade de ocorrer clivagem em todas estas situações. Os aminoácidos com maior probabilidade de corte (10% do conjunto total) representam nosso grupo observado de clivagens. Calculamos o esperado teórico em função da proporção de estruturas secundárias de todos os cristais e construímos gráficos com a razão entre o número de eventos observados e esperados para tornar a interpretação intuitiva, visto que quando ambos forem iguais, a razão é igual a 1, quando o observado for maior que o esperado a razão cresce, enquanto diminui quando o observado for menor que o esperado. Além de realizar estas etapas para o ponto onde ocorre a clivagem, recuperamos os aminoácidos adjacentes em uma janela de 10 aminoácidos, quando possível, e realizamos as mesmas análises a fim de verificar o comportamento desta razão nas regiões adajacentes ao corte. Tanto Netchop quanto ProteaSMM, em ambos métodos de modelagem (imunoproteossomo ou constitutivo) apresentam a mesma tendência no comportamento de associar alta probabilidade de corte em aminoácidos presentes em fitas beta em detrimento de aminoácidos em voltas (Figutas 5 e 6).

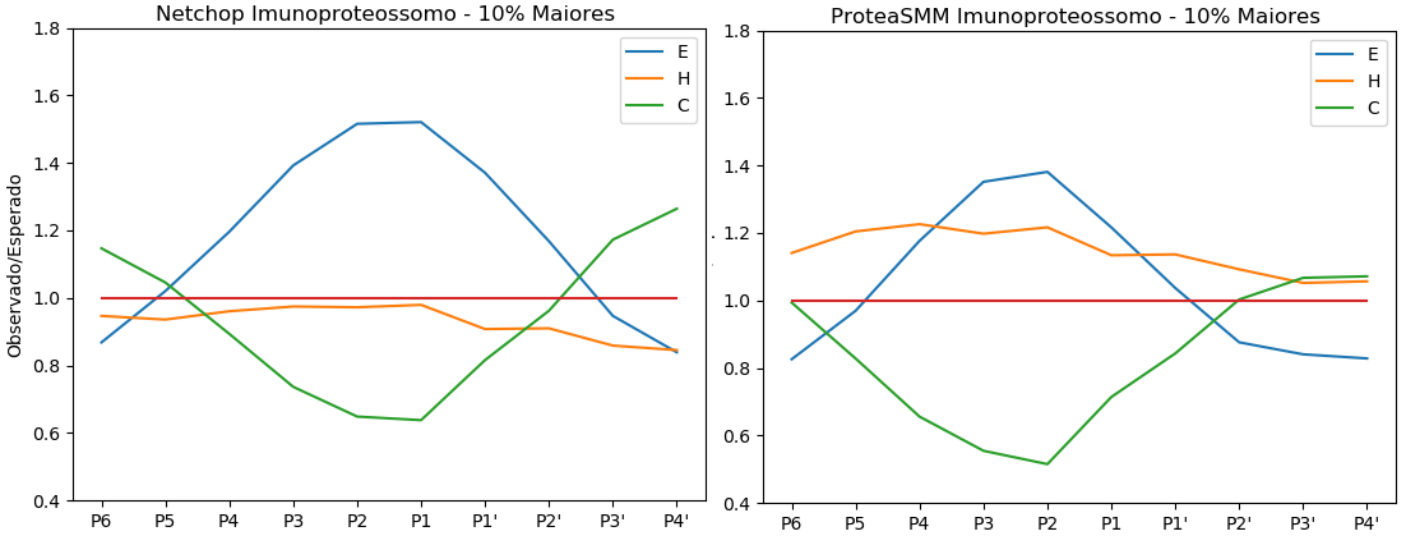

Figura 5. Razão entre observado e esperado em posições de corte (P1|P1') e aminoácidos adjacentes considerando algoritmos que modelam o proteossomo imunológico. E: fitas beta, H: hélices, C: voltas.

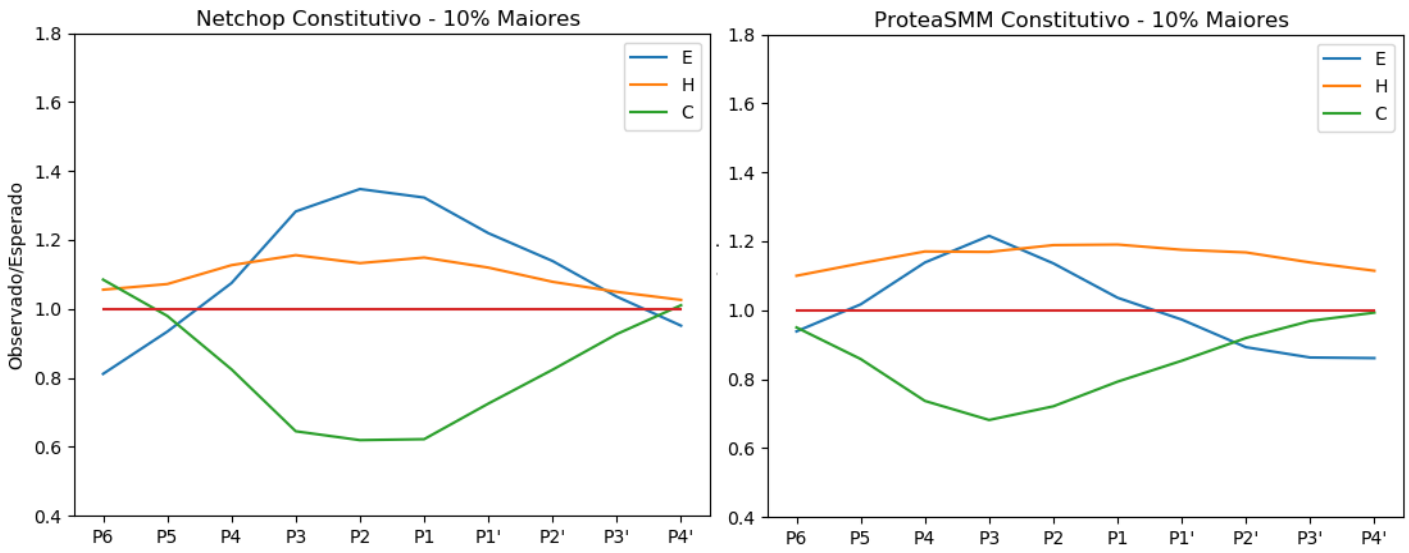

Figura 6. Razão entre observado e esperado em posições de corte e aminoácidos adjacentes considerando algoritmos que modelam proteossomo constitutivo. E: fitas beta, H: hélices, C: voltas.

4.2 Pontos de não clivagem apresentam sobressalência de voltas

Seguindo a mesma lógica empregada, nota-se o comportamento inverso, onde os aminoácidos com baixa probabilidade de corte inferida pelos diferentes preditores ocorrem em voltas mais do que o esperado ao acaso, em detrimento de hélices e fitas.

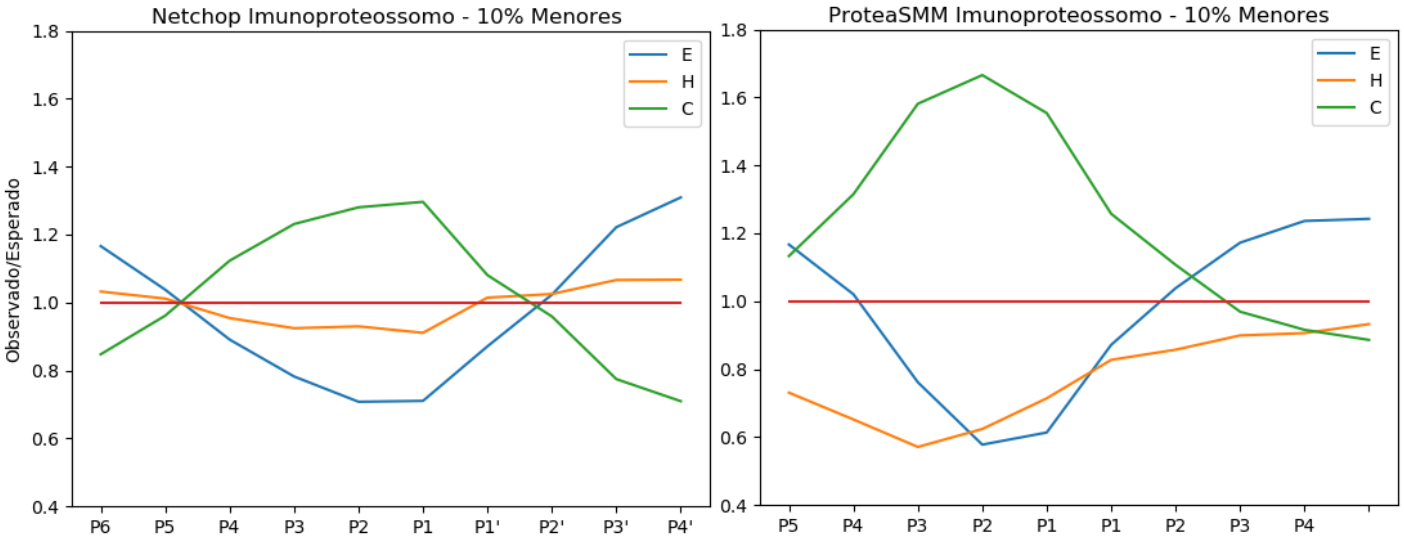

Figura 7. Razão entre observado e esperado em posições de corte e aminoácidos adjacentes considerando algoritmos que modelam o imunoproteossomo. E: fitas beta, H: hélices, C: voltas.

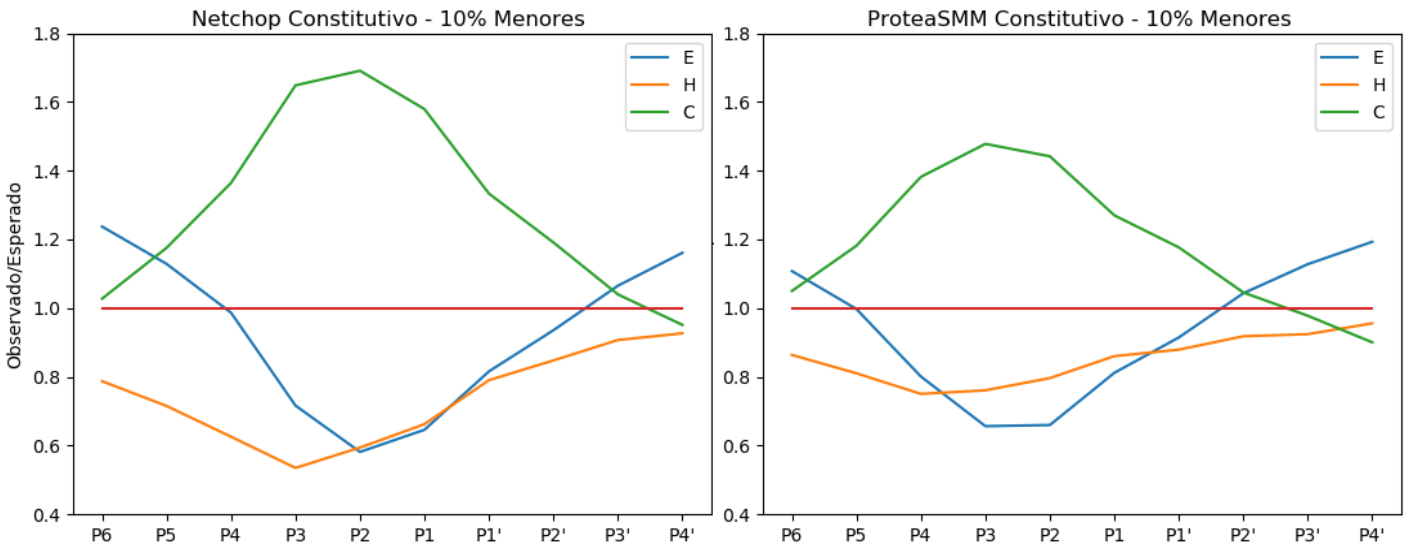

Figura 8. Razão entre observado e esperado em posições de corte e aminoácidos adjacentes considerando algoritmos que modelam o proteossomo constitutivo. E: fitas beta, H: hélices, C: voltas.

4.3 Redes Neurais

Construímos e treinamos uma grande quantidade de redes neurais devido a estocasticidade intrínsica deste método (Tabela 6). Diferente parametrização do apresentado na metodologia (momento=0.016, taxa de aprendizado=0.034, proporção para validação=0.20) não levou a melhores desempenhos, por isso outras redes não foram aqui apresentadas. Devido a restrições na predição de estrutura secundária, possuímos quatro formas de validação: 1, utilizando 10 peptídeos preditos por Spider com todos os sítios (Tabela 6) ou 2, somente os sítios analisáveis pelo ProteSMM (Tabela 8); 3, utilizando 16 peptídeos preditos por Spider ou Psipred com todos os sítios (Tabela 7) ou 4, com apenas os sítios analisáveis pelo ProteaSMM (Tabela 9). As comparações com todos os sítios são possíveis somente entre nossas redes e Netchop e as curvas ROC para as melhores redes destacadas na Tabela 6 estão representadas nas figuras 9 e 10.

Tabela 6. Comparação de Netchop e algumas de nossas redes com seu desempenho calculado sobre dados de validação clivados por imunoproteossomos (I) e proteossomo constitutivo (C) com SS predita por Spider para todos os sítios (10 peptídeos). As duas melhores redes estão ressaltadas em amarelo, e em negrito as que foram superiores ao Netchop.

Algoritmo	Tamanho da Janela	Dados de Validação	Área sob a curva
Netchop 20S	19	I	0.68000
Netchop 20S	19	C	0.68896
Netchop Cterm	19	I	0.64931
Netchop Cterm	19	C	0.64368
Rede-1	**5**	**I**	**0.70716**
Rede-1	5	C	0.66110
Rede-2	6	I	0.62840
Rede-2	6	C	0.56405
Rede-3	7	I	0.66803
Rede-3	7	C	0.61980
Rede-4	**8**	**I**	**0.69305**
Rede-4	8	C	0.62070
Rede-5	**9**	**I**	**0.68527**
Rede-5	9	C	0.65701
Rede-6	5	I	0.56606
Rede-6	5	C	0.60372
Rede-7	6	I	0.67828
Rede-7	6	C	0.60487
Rede-8	**7**	**I**	**0.70349**
Rede-8	**7**	**C**	**0.70082**
Rede-9	**8**	**I**	**0.75203**
Rede-9	**8**	**C**	**0.64368**
Rede-10	**9**	**I**	**0.70294**
Rede-10	9	C	0.58781
Rede-11	5	I	0.62144
Rede-11	5	C	0.57930
Rede-12	6	I	0.65310
Rede-12	6	C	0.66569
Rede-13	7	I	0.63317
Rede-13	7	C	0.61010
Rede-14	8	I	0.69583
Rede-14	8	C	0.61489
Rede-15	9	I	0.65744

Rede-15	9	C	0.63117
Rede-16	**5**	**I**	**0.69071**
Rede-16	5	C	0.64593
Rede-17	6	I	0.67894
Rede-17	6	C	0.59971
Rede-18	7	I	0.64755
Rede-18	7	C	0.56965
Rede-19	8	I	0.67753
Rede-19	8	C	0.66503
Rede-20	**9**	**I**	**0.71029**
Rede-20	9	C	0.65988
Rede-21	5	I	0.67003
Rede-21	5	C	0.63881
Rede-22	**6**	**I**	**0.72162**
Rede-22	6	C	0.63497
Rede-23	7	I	0.67335
Rede-23	7	C	0.62589
Rede-24	**8**	**I**	**0.70689**
Rede-24	8	C	0.59673
Rede-25	**9**	**I**	**0.68965**
Rede-25	9	C	0.60421
Rede-26	**5**	**I**	**0.72471**
Rede-26	5	C	0.59767
Rede-27	6	I	0.62637
Rede-27	6	C	0.52589
Rede-28	**7**	**I**	**0.68824**
Rede-28	7	C	0.64830
Rede-29	**8**	**I**	**0.74175**
Rede-29	**8**	**C**	**0.68245**
Rede-30	**9**	**I**	**0.68007**
Rede-30	9	C	0.60769

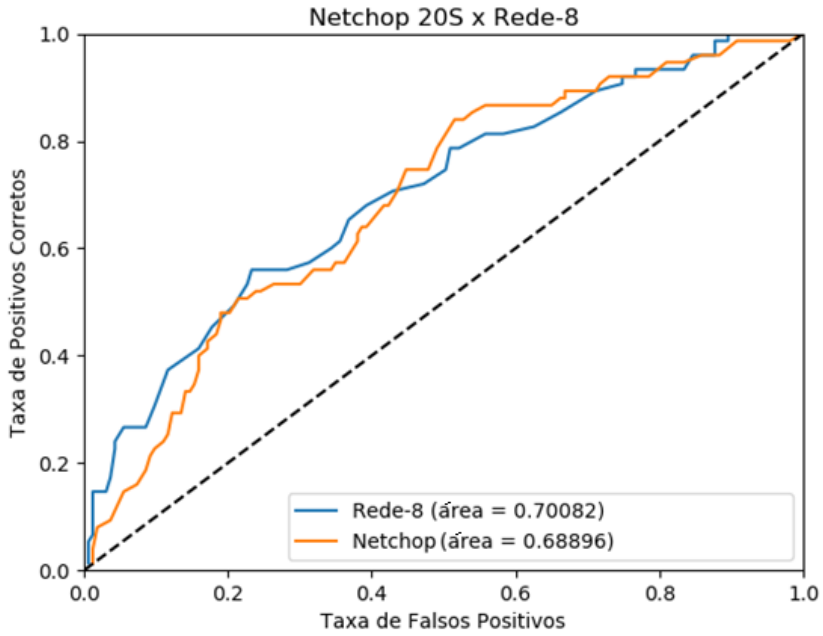

Figura 9. Curva ROC comparando a nossa Rede-8 com Netchop 20S no contexto de validação sobre peptídeos clivados com proteossomo constitutivo.

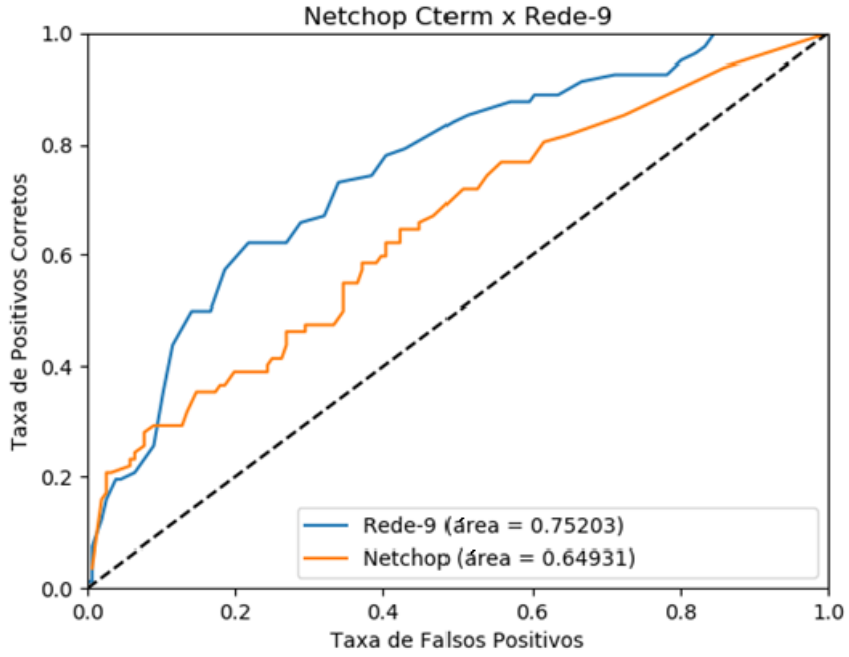

Figura 10. Curva ROC comparando a nossa Rede-9 com Netchop Cterm no contexto de validação sobre peptídeos clivados com imunoproteossomo.

Tabela 7. Área sob a curva de nossas melhores redes em comparação com Netchop utilizando conjunto de validação com estrutura secundária predita por Spider e Psipred para todos os sítios (16 peptídeos).

Algoritmo	**Tamanho da Janela**	**Dados de Validação**	**Área sob a curva**
Netchop 20S	19	I	0.72549
Netchop 20S	19	C	0.68171
Netchop Cterm	19	I	0.71850
Netchop Cterm	19	C	0.63301
Rede-8	7	I	0.68927
Rede-8	7	C	0.66248
Rede-9	8	I	0.71021
Rede-9	8	C	0.62439

Tabela 8. Área sob a curva de nossas melhores redes em comparação com Netchop e ProteaSMM utilizando conjunto de validação com estrutura secundária predita por Spider somente para os sítios analisáveis (10 peptídeos).

Algoritmo	**Tamanho da Janela**	**Dados de Validação**	**Área sob a curva**
Netchop 20S	19	I	0.72781
Netchop 20S	19	C	0.68931
Netchop Cterm	19	I	0.66232
Netchop Cterm	19	C	0.79121
ProteaSMM-c	10	I	0.95869
ProteaSMM-c	10	C	0.69560
ProteaSMM-i	10	I	0.97773
ProteaSMM-i	10	C	0.69069
Rede-8	7	I	0.61245
Rede-8	7	C	0.58879
Rede-9	8	I	0.66132
Rede-9	8	C	0.58095

Tabela 9. Área sob a curva de nossas melhores redes em comparação com Netchop e ProteaSMM utilizando conjunto de validação com estrutura secundária predita por Spider e Psipred somente para os sítios analisáveis (16 peptídeos).

Algoritmo	**Tamanho da Janela**	**Dados de Validação**	**Área sob a curva**
Netchop 20S	19	I	0.75371
Netchop 20S	19	C	0.69794
Netchop Cterm	19	I	0.72487
Netchop Cterm	19	C	0.64646
ProteaSMM-c	10	I	0.96838
ProteaSMM-c	10	C	0.69309
ProteaSMM-i	10	I	0.98156
ProteaSMM-i	10	C	0.68127
Rede-8	7	I	0.61745
Rede-8	7	C	0.59542
Rede-9	8	I	0.62354
Rede-9	8	C	0.55040

5. Discussão

A literatura é pouquíssimo clara quanto à estrutura secundária do substrato influenciar o processamento pelo proteossomo e o presente trabalho representa a primeira investigação que visa responder especificamente esta questão. Devido a escassez de ensaios de digestão proteossômica, visto que além de pequenos peptídeos apenas quatro proteínas foram completamente digeridas, algoritmos de aprendizado de máquina robustos devem ser utilizados para modelar o complexo comportamento do proteossomo partindo destas restrições. Existem diversos preditores de clivagens proteossômicas além de Netchop e ProteaSMM, contudo nenhum foi capaz de superá-los em desempenho. Apenas um preditor de clivagens utiliza informação de estrutura secundária que temos conhecimento [23]. Entretanto não utiliza informação de proteínas clivadas pelo proteossomo, mas sim proteínas clivadas por proteases de um modo geral. Apesar de ProteaSMM e Netchop não haverem sido treinados com esta informação, nossos resultados mostram que estes programas aprenderam a discriminar aminoácidos com estrutura secundária em conformação estendida preferencialmente favorável à proteólise em detrimento de outras estruturas secundárias. Apesar de aminoácidos em estrutura de fitas beta serem extremamente favorecidos em todas as situações (Figuras 6 e 7), hélices parecem também estar levemente sobressalentes quando a clivagem é predita por ProteSMM imunológico e constitutivo, e pelo Netchop 20S. Estes três programas tem em comum o fato de terem sido treinados com ensaios de degradação *in vitro* de proteínas, enquanto Netchop Cterm é treinado com ligantes de MHC-I. Estas situações são diferentes, visto que em ensaios de digestão proteossômica, a única causa de clivagens é o proteossomo, e não qualquer outro fator de dentro da célula. Ligantes de MHC-I são grande parte, mas não exclusivamente provenientes de proteólise proteossômica, o que indica outras etapas da via (ligação à TAP, ligação ao MHC ou outros fatores) estão favorecendo ainda mais que aminoácidos em conformação estendida estejam assinalados como sujeitos a clivagens pelo Netchop Cterm. Além de preferir fitas beta, todos os programas também convergem para o fato de aminoácidos situados em voltas ou *loops* estarem pouco representados em regiões de alta predição de clivagem. Quando avaliamos os aminoácidos que os programas apontam baixa probabilidade de corte, percebemos o comportamento de espelho (o que reforça nossa observação), visto que voltas estão sobressalentes em todas as situações em detrimento das outras estruturas secundárias. Novamente o Netchop Cterm destoa levemente dos outros programas, pois a magnitude da razão entre o observado e esperado é sutilmente menor. É importante ressaltar que esta é uma evidência indireta,

pois a estrutura secundária anotada para os aminoácidos é a presente no cristal, e é sabido que a partícula regulatória gasta energia para desenovelar o substrato, antes dele entrar na partícula central. Contudo, se a preferência intrínseca daquele aminoácido e dos aminoácidos adjacentes formar aquela estrutura (ou outro nível de estruturação permitido por aquela combinação de aminoácidos) for a mesma no contexto de clivagem, esperamos que a informação se mantenha, permitindo a validade de nossas análises. Corroborando para este ponto temos os inibidores de proteossomos, que se ligam com alta afinidade em conformações estendidas, mostrando ao menos que isto é possível. Outro evidência interessante é o comportamento de bolha do gráfico, onde em resíduos próximos a clivagem (P3,P2,P1, região *primed*) a magnitude da razão tende a ser maior, a medida que ao se afastar do ponto de corte tende a ser menor.

Partindo dessas informações, a etapa seguinte foi a construção de redes neurais utilizando única e exclusivamente a probabilidade de se formar estrutura secundária predita com o programa Spider (nem mesmo a informação do aminoácido em cada posição foi informada). Utilizando esta codificação extremamente simples, fomos capazes de desempenho semelhante ou melhor do que o programa mais utilizado na área (NetChop, figuras 9 e 10), nas condições de validação apresentadas. O conjunto de validação formado por 16 peptídeos que já era pequeno, tornou-se ainda menor devido a remoção de 6 peptídeos para os quais não foram possíveis a construção das matrizes para a predição de estrutura secundária pelo Spider. Para contornar este problema e aumentar o grupo de validação, realizamos a predição destes 6 peptídeos com o Psipred. Entretanto, isto foi realizado mais em função da comparação com os outros programas, visto que a utilização de um preditor de estrutura secundária de baixa confiabilidade provavelmente causaria uma queda no desempenho de nossas redes, as quais são diretamente dependentes de uma predição confiável. Isso foi verificado ao olhar para a área sob a curva calculada para Netchop a qual aumenta a medida que os peptídeos foram re-adicionados. Em contrapartida, a nossa melhor rede no contexto de imunoproteossomo, a Rede-9 (Tabela 6 e figura 10) tem sua área sob a curva reduzida de 0.75203 (maior área encontrada) para 0.62354. Como não havíamos treinado redes específicas para imunoproteossomo ou proteossomo constitutivo, pois misturamos a informação considerando sítios negativos somente se não clivados por ambas isoformas, a mesma rede deveria ser validada sobre os conjuntos de validação imunológico ou constitutivo. Percebemos que a maioria de nossas redes eram melhores em predizer clivagens dos peptídeos de validação clivados por imunoproteossomo. A rede-9 é um exemplo disso, visto que a mesma rede que atinge mais de 0.75 na área sob a curva, atinge apenas 0.64368 no contexto de proteossomo constitutivo. O mesmo

comportamento é verificado para o ProteaSMM, que possui uma área quase total quando validado no contexto de imunoproteossomo, tanto ProteaSMM-i, quanto ProteaSMM-c, mas tem uma queda brusca de desempenho quando testado contra os peptídeos clivados pelo imunoproteossomo. É proposto que esta situação advém do fato do imunoproteossomo ser mais fácil de ser predito, com uma tendência maior a clivar aminoácidos hidrofóbicos. Por fim, é importante ressaltar a importância da clivagem proteossômica como primeiro passo para gerar precursores dos epitopos que serão apresentados aos linfócitos T CD8+, visto que houve a adaptação de substituir as subunidades proteolíticas por outras com comportamento muito semelhante, diferindo apenas nas quantidades de fragmentos geradas. Logicamente pensando, o sistema imunológico adaptativo deveria estar direcionado contra regiões conservadas e a geração de epitopos provenientes de regiões de alças não faria muito sentido, visto que estas regiões são mais propensas a mutações, possibilitando mais um mecanismo de escape para os patógenos. Este trabalho representa uma evidência indireta de que a propensão intrínseca de estruturação dos aminoácidos influenciaria a clivagem de substratos no interior da câmara catalítica proteossômica.

6. Referências

1. Bassermann, F., R. Eichner, and M. Pagano, *The ubiquitin proteasome system - implications for cell cycle control and the targeted treatment of cancer.* Biochim Biophys Acta, 2014. **1843**(1): p. 150-62.
2. Yao, T. and A. Ndoja, *Regulation of gene expression by the ubiquitin-proteasome system.* Semin Cell Dev Biol, 2012. **23**(5): p. 523-9.
3. Bhattacharyya, S., et al., *Regulated protein turnover: snapshots of the proteasome in action.* Nat Rev Mol Cell Biol, 2014. **15**(2): p. 122-33.
4. Kish-Trier, E. and C.P. Hill, *Structural biology of the proteasome.* Annu Rev Biophys, 2013. **42**: p. 29-49.
5. Nathan, J.A., et al., *Immuno- and constitutive proteasomes do not differ in their abilities to degrade ubiquitinated proteins.* Cell, 2013. **152**(5): p. 1184-94.
6. Moreau, P., et al., *Proteasome inhibitors in multiple myeloma: 10 years later.* Blood, 2012. **120**(5): p. 947-59.
7. Sledz, P. and W. Baumeister, *Structure-Driven Developments of 26S Proteasome Inhibitors.* Annu Rev Pharmacol Toxicol, 2016. **56**: p. 191-209.
8. Tyndall, J.D., T. Nall, and D.P. Fairlie, *Proteases universally recognize beta strands in their active sites.* Chem Rev, 2005. **105**(3): p. 973-99.
9. Madala, P.K., et al., *Update 1 of: Proteases universally recognize beta strands in their active sites.* Chem Rev, 2010. **110**(6): p. PR1-31.
10. Sharon, M., et al., *20S proteasomes have the potential to keep substrates in store for continual degradation.* J Biol Chem, 2006. **281**(14): p. 9569-75.
11. Ruschak, A.M., et al., *The proteasome antechamber maintains substrates in an unfolded state.* Nature, 2010. **467**(7317): p. 868-71.
12. Nielsen, M., et al., *The role of the proteasome in generating cytotoxic T-cell epitopes: insights obtained from improved predictions of proteasomal cleavage.* Immunogenetics, 2005. **57**(1-2): p. 33-41.
13. Kesmir, C., et al., *Prediction of proteasome cleavage motifs by neural networks.* Protein Eng, 2002. **15**(4): p. 287-96.
14. Emmerich, N.P., et al., *The human 26 S and 20 S proteasomes generate overlapping but different sets of peptide fragments from a model protein substrate.* J Biol Chem, 2000. **275**(28): p. 21140-8.
15. Toes, R.E.M., et al., *Discrete Cleavage Motifs of Constitutive and Immunoproteasomes Revealed by Quantitative Analysis of Cleavage Products.* The Journal of Experimental Medicine, 2001. **194**(1): p. 1-12.
16. Tenzer, S., et al., *Quantitative analysis of prion-protein degradation by constitutive and immuno-20S proteasomes indicates differences correlated with disease susceptibility.* J Immunol, 2004. **172**(2): p. 1083-91.
17. Tenzer, S., et al., *Modeling the MHC class I pathway by combining predictions of proteasomal cleavage,TAP transport and MHC class I binding.* Cellular and Molecular Life Sciences CMLS, 2005. **62**(9): p. 1025-1037.
18. Lucchiari-Hartz, M., et al., *Differential proteasomal processing of hydrophobic and hydrophilic protein regions: contribution to cytotoxic T lymphocyte epitope clustering in HIV-1-Nef.* Proc Natl Acad Sci U S A, 2003. **100**(13): p. 7755-60.
19. Calis, J.J., et al., *Role of peptide processing predictions in T cell epitope identification: contribution of different prediction programs.* Immunogenetics, 2015. **67**(2): p. 85-93.

20. Frishman, D. and P. Argos, *Knowledge-based protein secondary structure assignment.* Proteins, 1995. **23**(4): p. 566-79.
21. Heffernan, R., et al., *Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning.* Sci Rep, 2015. **5**: p. 11476.
22. McGuffin, L.J., K. Bryson, and D.T. Jones, *The PSIPRED protein structure prediction server.* Bioinformatics, 2000. **16**(4): p. 404-405.
23. Li, B.Q., et al., *Prediction of protein cleavage site with feature selection by random forest.* PLoS One, 2012. **7**(9): p. e45854.