UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE CIÊNCIAS EXATAS E NATURAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

PAULO VITOR RODRIGUES CARDOSO

Identificação de Regiões de Ubiquitinação Utilizando Nondominated Sorting Genetic Algorithm II

Dissertação

Belém

2016

UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE CIÊNCIAS EXATAS E NATURAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

PAULO VITOR RODRIGUES CARDOSO

Identificação de Regiões de Ubiquitinação Utilizando Nondominated Sorting Genetic Algorithm II

Dissertação apresentada para obtenção de grau de Mestre em Ciência da Computação, elaborada sob orientação do Prof. Dr. Claudomiro de Souza de Sales Junior e co-orientado pela Profa.Dra. Regiane Silva Kawasaki Francês.

Belém

2016

PAULO VITOR RODRIGUES CARDOSO

Identificação de Regiões de Ubiquitinação Utilizando Nondominated Sorting Genetic Algorithm II

Dissertação apresentada para obtenção de grau de Mestre em Ciência da Computação, elaborada sob orientação do Prof. Dr. Claudomiro de Souza de Sales Junior e co-orientado pela Profa.Dra. Regiane Silva Kawasaki Francês.

Trabalho aprovado. Belém, 26 de novembro de 2016:

Prof. Dr. Claudomiro de Souza de Sales Júnior
Orientador - PPGCC/UFPA

Prof. Dr. Ronnie Cley de Oliveira Alves (UFPA)
Membro Interno - PPCGG/UFPA

Profa. Dra. Regiane Kawasaki Silva Francês (UFPA)
Membro Externo - FACOMP/UFPA

Profa. Dra. Danielle Costa Carrara Couto (UNIFESSPA)
Membro Externo - UNIFESSPA

Belém
2016

Este trabalho é dedicado aos meus pais que tanto batalharam para que pudesse ter as melhores oportunidades possíveis.

Agradecimentos

A Deus, por todas as oportunidades desta vida, seja pelas coisas agradáveis, que são um grande alento, mas em especial pelas coisas que a nossos olhos não são tão agradáveis, mas que na realidade servem para nos fazer crescer e por isso nos tornam felizes. Agradeço por tudo que tenho em minha vida, meus projetos, meus sonhos, minhas realizações. Tudo que aprendi te ofereço de livre vontade, para que aquilo que busco não seja vazio, mas que sirva para ti e consequentemente para o bem de todos.

Agradeço a Minha família, estes que em tudo me apoiam. Me ensinaram o verdadeiro valor das virtudes obtidas através do trabalho fiel, sem estes valores jamais teria conseguido perseverar até o fim, não teria o equilíbrio necessário nem a sabedoria e serenidade suficiente nos momentos mais difíceis.

Agradeço a minha futura esposa Mara Chagas por mais uma vez estar ao meu lado, muito obrigado pela participação ativa em mais esta conquista, sem a sua dedicação amorosa não teria sido possível alcançar muitas coisas. Obrigado por todo o amor e carinho dedicado em todos estes anos, sem você as coisas seriam mais difíceis, obrigado e por tudo.

Agradeço ao professor Claudomiro por todos esses anos de parceria e trabalho conjunto, neles aprendi grandes valores de simplicidade, companheirismo e humildade. Para mim sua postura como professor junto a seus alunos serve de grande exemplo a ser alcançado, obrigado por sua confiança neste trabalho e sua dedicação a me ajudar da melhor forma possível. Da mesma forma agradeço a professora Regiane Kawasaki, que também é para mim uma referência de profissional da educação e relacionamento com os alunos. Obrigado por todo o estímulo e pela grade ajuda em meu trabalho, também agradeço pelo apoio em todas as dificuldades, nossos encontros foram sempre muito estimulantes para a caminhada, e por isso sou muito grato.

Obrigado aos colegas e amigos que ganhei neste caminho, em especial ao Reginaldo e Afonso por toda a ajuda, correções e discussões sobre este e diversos outros projetos e mais diversos temas, vocês foram de grande importância para mim.

Também agradeço a todos os amigos que me acompanharam nesse tempo de luta e muita dedicação, foram muitos afazeres, preocupações e tarefas a cumprir, vocês tornaram tudo isso muito mais leve para mim, em especial agradeço os amigos sempre presentes no meu cotidiano Átila e Taline Bandeira, Igor Vianna e Ana Luiza Espada, Lucas e André Fonseca e todos os demais que por ventura não tenha citado aqui, vocês são presentes de Deus na minha vida. E por todos aqueles que por algum motivo não se encontram aqui, mas que participaram em tudo através do seu apoio em diversos momentos.

”A pesquisa científica leva ao conhecimento de verdades sempre novas sobre o homem e o universo. O verdadeiro bem da humanidade, acessível na fé, abre o horizonte no qual se deve mover o seu caminho de descoberta. Assim a fé, vivida realmente, não entra em conflito com a ciência, aliás, coopera com ela, oferecendo critérios basilares a fim de que promova o bem de todos, pedindo-lhe que renuncie apenas àquelas tentativas que — opondo-se ao desígnio originário de Deus — podem produzir efeitos que se voltam contra o próprio homem.”

Joseph Ratzinger, Papa emérito Bento XVI

Resumo

A ubiquitinação se constitui como um importante mecanismo biológico, que tem entre suas finalidades a regulação de proteínas através da marcação daquelas que são indesejadas e que posteriormente são degradadas por organelas conhecidas como proteassoma. A ubiquitina também atua no mecanismo de processos inflamatórios, alem de ter relação importante com doenças neuro-degenerativas e câncer. Haja vista sua evidente importância dentro dos organismos eucariotos, seu entendimento colabora com investigações a respeito da mesma. Entretanto, para melhor entender o funcionamento da ubiquitina, é importante compreender como o processo de ubiquitinação ocorre. A identificação de regiões de ubiquitinação em proteínas é fundamental para ter um conhecimento completo a respeito dos mecanismos moleculares deste sistema. Métodos tradicionais possuem um custo elevado de experimentação, alem de também possuir um alto custo de tempo e esforço. Nesses quesitos, métodos de aprendizado de máquina se mostram como uma alternativa promissora, tanto por possuir um baixo custo comparado com a experimentação convencional, quanto por conseguir bons resultados na identificação de regiões de ubiquitinação agregado a um baixo custo de tempo e esforço. Estes métodos conseguem identificar potenciais regiões de ubiquitinação através da análise das propriedades das sequencias genéticas de fragmentos de proteínas, ou do proteoma inteiro. Este trabalho utiliza metodologias de aprendizado de máquina como Non-dominated Sorting Genetic Algorithm II (NSGA-II) para identificar e predizer regiões de ubiquitinação baseadas em propriedades físico-químicas e na composição de aminoácidos. O NSGA-II constrói regras de classificação baseadas nas propriedades das sequencias de fragmentos, a partir destas regras é possível levantar hipóteses a respeito da importância de cada um dos parâmetros utilizados. Além de buscar resultados na identificação de regiões de ubiquitinação, este trabalho também reuniu e integrou informações relevantes para pré-processamento e geração de base de dados de fragmentos ubiquitináveis através da integração de várias bases de dados de informações em uma base integrada chamada UPIS. Como resultado da pesquisa, obteve-se uma base de dados com cerca de 60.016 regiões ubiquitinadas de 4 organismos distintos: *Arabidopsis thaliana*, *Saccharomyces cerevisiae*, *Mus musculus* e *Homo sapiens*. Os classificadores utilizados obtiveram um valor de acurácia igual a 76.79%, 84.17%, 71.69% e 69.71% respectivamente. Um teste com todas as regiões também foi realizado, obtendo um valor de acurácia igual a 68.76%.

Palavras chave: Ubiquitinação, Biologia molecular, Aprendizado de máquina, Algoritmo Genético, NSGA-II, Algoritmo Multi-objetivo.

Abstract

The ubiquitylation is constituted as an important biological mechanism, which has among its objectives the proteins regulation by marking those that are unwanted and are subsequently degraded by organelles known as proteasome. Ubiquitin is also active in inflammation mechanism, besides having important relationship with neuro-degenerative diseases and cancer. Given its obvious importance within the eukaryotic organisms, their understanding collaborates with research about the same. However, to better understand the operation of ubiquitin, it is important to understand how the ubiquitination process occurs. The identification of ubiquitylation regions in proteins is critical to have a complete knowledge about the molecular mechanisms of the ubiquitylation system. Traditional methods have a high cost of experimentation, time and effort. In this area, machine learning methods prove to be a promising alternative, it have a low cost compared with conventional experimentation methods, and by achieving good results in identifying regions of ubiquitylation at a low cost of time and effort. These methods can identify potential areas of ubiquitylation by analyzing the properties of the genetic sequences of protein fragments, or the whole proteome. This paper uses machine learning methods as Non-dominated Sorting Genetic Algorithm II (NSGA-II) to identify and predict ubiquitylation sites based on the physicochemical properties and amino acid composition. The NSGA-II builds classification rules based on the properties of fragments sequences, from these rules, is possible construct hypotheses about the importance of each used. Besides seeking results in the identification of ubiquitylation sites, this work also gathered and incorporated relevant information for preprocessing and generation of ubiquitylation sites database, integrating various information databases in an integrated database called UPIS. As research result, we obtained a database of approximately 60.016 ubiquitinated regions from 4 different organisms: Arabidopsis thaliana, Saccharomyces cerevisiae, Mus musculus and Homo sapiens. The classifiers obtained an accuracy value of 76.79%, 84.17%, 71.69% and 69.71% respectively. A test with all regions was also carried out, obtaining a value of accuracy of 68.76%.

Keywords: Ubiquitylation, Molecular Biology, Machine Learning, Genetic Algorithm, NSGA-II, MultiObjective Algorithm.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

AG	Algoritmo Genético
AAC	Amino Acid Composition
Ac	Acurácia (Ac)
Sb	Sensibilidade (Sb)
Es	Especificidade (Es)
Ub	Ubiquitina
ROC	Receiver operating characteristic
MCC	Matthew's correlation coefficient
EM	Espectrometria de massa
KNN	k-nearest neighbor
SVM	Support Vector Machine
NSGA-II	Nondominated Sorting Genetic Algorithm - II

Sumário

1 Introdução

O avanço das técnicas utilizadas na biologia molecular, permite a ciência obter cada vez mais conhecimento a respeito do funcionamento dos processos biológicos, químicos e físicos que ocorrem dentro dos seres viventes. Este avanço tem se intensificado ano após ano, chegando a um ponto onde se faz necessário a utilização de outra ferramenta para nos auxiliar na análise dos dados gerados: a Bioinformática.

Antes, amplamente utilizada apenas em análises quantitativas, agora vem ganhando cada vez mais importância, e um papel de destaque na análise qualitativa da grande quantidade de dados geradas por ferramentas da biologia molecular, isto graças a capacidade das chamadas técnicas de aprendizagem de máquina, que conseguem aprender padrões, agrupar valores e características de dados, além de fornecer outras diversas funcionalidades, agindo de forma similar a capacidade de aprendizado da mente humana.

Com a grande quantidade de dados obtidos a partir de diversos tipos de análises e a utilização da bioinformática na análise desses dados, a ciência é capaz de acelerar a velocidade do aprendizado e compreensão a respeito de muitas funções orgânicas, o que consequentemente permite com que nos especializemos e concentremos cada vez mais esforço no estudo de engrenagens menores dentro do organismo. Um grande exemplo dessa especialização, são os estudos direcionados ao entendimento do sistema de regulação de proteínas dentro do organismo e a relação deste sistema com outros processos do organismo.

Um dos sistemas que está relacionado com esta renovação de proteínas é o sistema proteassoma-ubiquitina, onde o processo de ubiquitinação tem um papel fundamental. O sistema proteassoma-ubiquitina desempenha um papel crucial na regulação de uma grande variedade de processos biológicos, como o ciclo de divisão celular, a resposta imunológica e o processo de inflamação (CHEN et al., 2013). A ubiquitina (Ub) por sua vez, também possui um importante papel em inúmeros processos que envolvem cascatas de sinalização (UDESHI et al., 2013a). A ubiquitinação ocorre quando a ubiquitina (uma pequena proteína composta por aproximadamente 76 aminoácidos) se liga a uma proteína alvo (CAI et al., 2012). A partir dai, se inicia o processo de identificação dessas proteínas marcadas com a ubiquitina, e a renovação ou destruição dessas proteínas danificadas.

A identificação de regiões de ubiquitinação em proteínas ajuda a obter um maior entendimento a respeito dos papeis regulatórios da ubiquitinação e dos mecanismos moleculares do sistema ubiquitinável (CAI et al., 2012; NALEPA; ROLFE; HARPER, 2006). A utilização de métodos experimentais convencionais para identificar essas potenciais regiões de ubiquitinação, possui um elevado custo de tempo e esforço, mesmo existindo

métodos de purificação mais eficientes que corroboram na identificação dessas regiões (CHEN et al., 2014).

Contudo, os métodos computacionais podem detectar possíveis novas regiões de ubiquitinação de forma mais barata e com menos esforço através da análise das características moleculares e do contexto das sequências genéticas de regiões já identificadas experimentalmente. A entrada utilizadas por esses preditores computacionais, normalmente são propriedades retiradas de fragmentos de sequências com uma lisina (K) central de interesse na investigação.

Grande parcela dos estudos mais conhecidos na área utilizam algoritmos como *Support Vector Machine* (SVM) (CHEN et al., 2013; WALSH; Di Domenico; TOSATTO, 2014; TUNG; HO, 2008; CHEN et al., 2013), *Radial Basis Function* (LEE et al., 2011) e *Random Forest* (RADIVOJAC; VACIC, 2010).

Entre as possibilidades de algoritmos de aprendizado de máquina, encontra-se o algoritmo genético multi-objetivo. Este é aplicado em diferentes áreas, onde decisões precisam ser tomadas considerando-se conflitos de escolhas entre dois ou mais objetivos conflitantes. A habilidade de manusear problemas complexos, envolvendo características como descontinuidades, multimodalidades, avaliações de funções com ruídos, entre outras, reforça o potencial efetivo dos algoritmos genéticos (AG) em problemas de otimização, oferecendo assim uma alternativa na busca pelas regiões de ubiquitinação. O uso deste tipo de algoritmo também abre novas perspectivas e possibilidades para o entendimento e identificação das regiões de ubiquitinação. Este trabalho utiliza o algoritmo *Non-Dominated Genetic Algorithm* (NSGA-II) para identificar regiões de ubiquitinação. No NSGA-II, a população é modelada como regras de classificação, o algoritmo busca otimizar os valores de precisão e o *recall* de cada uma delas.

1.1 Motivação

Este trabalho busca colaborar nas pesquisas científicas do processo de ubiquitinação, devido a sua importância dentro de processos que originam doenças neurodegenerativas e auto-imunes como Parkinson ou mesmo o câncer.

Este trabalho também busca testar e demonstrar o potencial que o algoritmo genético possui ao atuar como um algoritmo de classificação. E também demonstrar que a sua capacidade e independência da utilização de outros algoritmos vai além do que normalmente é retratado em pesquisas científicas, onde o mesmo é utilizado apenas para otimizar parâmetros de outros algoritmos classificatórios, sendo assim sub-utilizado e atuando como um algoritmo coadjuvante.

1.2 Trabalhos Relacionados

Diversos estudos foram realizados utilizando técnicas *in silico* na tentativa de identificar possíveis regiões de ubiquitinação e classificadores que consigam bons resultados na identificação dessas regiões em novas bases de dados. Entre os principais e também considerado um dos percursores entre os trabalhos publicados na área encontra-se o UbiPred (TUNG; HO, 2008). A partir deste trabalho, muitos outros foram desenvolvidos.

No trabalho de Tung (2008), foi estabelecido uma base de dados de sequencias genéticas retiradas de 105 proteínas. A partir destas sequencias foram identificadas todas as lisinas (K), e estabelecido a partir daí um tamanho de sequencia a ser considerado a partir da lisina central. Este conjunto de sequencias com uma lisina no meio foram armazenados, e identificadas como ubiquitinadas e não-ubiquitinadas, considerando diversos outros trabalhos que comprovaram esta classificação experimentalmente.

Dessas sequencias foram extraídas as informações físico-químicas das moléculas e montada uma base de dados final, que serviria de entrada para os classificadores utilizados no trabalho. Três classificadores foram utilizados na análise da base de dados: *k-nearest neighbor* (KNN), Naive Bayes e *Support Vector Machine* SMV, sendo este ultimo o que conseguiu melhores resultados de classificação. A partir dos primeiros resultados alcançados, uma segunda carga de testes foi realizada, entretanto, antes do inicio do processo, utilizando um algoritmo genético, os parâmetros de entrada do SVM foram otimizados, o que permitiu que os resultados fossem melhorados consideravelmente.

Os demais trabalhos relacionados também contribuíram estabelecendo outros tipos de informações que poderiam ser analisadas utilizando o mesmo tipo de produto base, as sequencias genéticas centralizadas em uma lisina, estes trabalhos começaram a sobrepor os resultados alcançados por Tung(2008) com a utilização das propriedades físico-químicas das moléculas, utilizando no lugar dessas propriedades, informações baseadas nas sequencias ou codificação das cadeias genéticas. Esta propriedade é comumente chamada de Composição de amino ácidos (*Amino Acid Composition* - AAC).

Em (CHEN et al., 2011), foi construído um preditor baseado em SVM chamado CKSAAP_UbSite. Este preditor detecta regiões de ubiquitinação baseado nas cadeias das proteínas. A propriedade de entrada do preditor é chamada de codificação *k-spaced amino acid pairs* (CKAAP), esta codificação representa a sequencia ao redor das regiões de ubiquitinação, esta propriedade reflete as interações de curto alcance entre os resíduos. O SVM foi treinado utilizando esta codificação, as métricas utilizadas para avaliar a qualidade das predições: Acurácia (Ac), Sensibilidade (Sb), Especificidade (Es) e o coeficiente de correlação de Matthew (MCC). A acurácia média alcançada pelo algoritmo foi de 73,40% de acurácia.

O estudo descrito por (CAI et al., 2012), utiliza o algoritmo KNN combinado

com um algoritmo de seleção de propriedades (PENG; SCHWARTZ; ELIAS, 2003) para construir o preditor. Vinte e seis (26) parâmetros foram utilizados para descrever cada um dos aminoácidos que circulam a região da lisina. O estudo utilizou o MCC como métrica na identificação das lisinas ubiquitináveis, e o resultado obtido foi de 0,142.

No trabalho de (CHEN et al., 2013) é especificado um preditor estabelecido para classificar especificamente ubiquitinação presente no organismo de humanos, utilizando para isso uma integração de diversos classificadores complementares. Foi utilizado um SVM baseado na codificação CKAAP. Além deste SVM, outros três classificadores baseados em SVM foram construídos. O preditor resultante foi chamado de hCKSAAP_UbSite e alcançou um resultado de 0,770 de área sob a curva (AUC). Quando este foi testado em uma base de dados independente e balanceada com 3419 regiões de ubiquitinação, hCKSAAP_UbSite alcançou uma área sob a curva de 0,757

UbiProber é o nome do classificador originado do trabalho de (CHEN et al., 2013) e o mesmo foi construído para classificar tanto regiões de ubiquitinação especificas (de apenas uma espécie) quanto para base de dados com diversos organismos. Três tipos de propriedades foram extraídas dos dados de treinamento e combinado utilizando SVM para realizar predições, estas propriedades foram retiradas de um algoritmo KNN, propriedades fisico químicas (PCP), e da composição de aminoácidos (AAC). As propriedades do KNN capturam a similaridade das sequencias locais, PCP e AAC refletem o ambiente bioquímico das regiões vizinhas à região de ubiquitinação (lisina). Dados de ubiquitinação de três espécies foram coletadas: *Homo sapiens*, *Mus musculus* e *Saccharomyces cerevisiae*. UbiProber alcançou áreas sob a curva de 80.94% (H. sapiens), 82.11% (M. musculus) e 83.40% (S. cerevisiae).

Rapid UBIquitylation (RUBI) é apresentado por (WALSH; Di Domenico; TOSATTO, 2014). O método é um preditor baseado na sequencia de aminoácidos, desenvolvido para ser uma aplicação rápida. Esse preditor é baseado em dois métodos diferentes de aprendizado de máquina, um SVM e uma rede neural bidirecional. A rede neural aprende sobre o contexto em que se encontra a lisina e os aminoácidos que a circula. A base de dados foi composta a partir de um conjunto de dados de 1.054 motifs positivos de *H. sapiens* retirados de 4.273 proteínas. RUBI foi treinado com 3.705 sequencias testadas em um conjunto independente, obtendo uma área sob a curva igual a 0,745 em dados independentes.

1.3 Justificativa

Os trabalhos relacionados não apresentam de forma satisfatória os modelos de classificação encontrados para classificar as bases de dados utilizadas, possibilitando até mesmo críticas a respeito dos resultados alcançados que foram alegados em seus trabalhos.

Além do mais, é necessário trazer a discussão, a utilização de algoritmos que possuam potencial dentro da área de classificação desses dados, e que saiam do pequeno círculo de algoritmos já utilizados, já que a maioria dos trabalhos se concentra na utilização uma quantidade ínfima de algoritmos diante de uma grande quantidade de algoritmos existentes. A utilização de algoritmos ainda não utilizados na classificação destes dados,trará novas perspectivas a respeito do problema.

Estes algoritmos possibilitariam observar como os parâmetros de avaliação da classificação se comportam diante da variedade de possibilidades de valores e utilização de parâmetros informativos das sequencias genéticas. Analisar passo-a-passo a evolução dos indivíduos da população dentro do algoritmo também permite observar novas perspectivas do problema e determinar em quais pontos e parâmetros são cruciais na identificação do objeto de pesquisa.

1.4 Objetivos

O objetivo geral deste trabalho é identificar novas regiões de ubiquitinação na base de dados de proteínas a partir das regiões já identificadas como ubiuqitinadas. Como objetivos específicos temos:

- Analisar o desempenho do NSGA-II comparado a outros algoritmos de classificação menos utilizados na área.
- Estabelecer uma base de dados mais robusta que as encontradas e difundidas em pesquisas semelhantes do referencial bibliográfico.
- Apresentar e comparar os diferentes resultados a partir de propriedades utilizadas na classificação.

1.5 Estrutura da Dissertação

Este trabalho é composto por 6 capítulos incluindo este capítulo introdutório. Os mesmos estão dispostos a seguir:

Capítulo 2: Ubiquitina Este capítulo tratará exclusivamente de explicar a importância biológica do tema principal do estudo, alem de apresentar informações importantes a respeito das características analisadas no processo de classificação. Neste capítulo também serão apresentados alguns das principais pesquisas relacionadas ao trabalho aqui realizado.

Capítulo 3: Algoritmos de classificação Este capítulo apresenta os algoritmos de aprendizado de máquina utilizados neste trabalho para classificar os dados de regiões de ubiquitinação e suas propriedades.

Capítulo 4: Metodologia Este capítulo apresentará toda a estrutura construída para realizar as investigações relacionadas a identificação das regiões ubiquitinadas. Aqui também será explicado a construção da base de dados, seus processos agregados, alem de apresentar mais detalhadamente a metodologia de construção dos algoritmos utilizados.

Capítulo 5: Resultados Aqui será apresentado os principais resultados obtidos da pesquisa, bem como a discussão a respeito dos mesmos, dificuldades encontradas e melhorias empregadas.

Capítulo 6: Considerações finais Capítulo de apresentação final dos resultados e considerações a respeito do trabalho e dos principais resultados alcançados e perspectivas de trabalhos futuros.

2 Ubiquitina

A ubiquitina é um polipeptídeo composto por 76 aminoácidos, que age em modificações pós-translacionais de substratos de proteínas através de uma ligação covalente a um resíduo de lisina localizada na proteína alvo. O sistema celular da ubiquitina afeta praticamente todos os eventos celulares mais complexos (CLAGUE; HERIDE; URBÉ, 2015).

A ubiquitina se faz presente de forma altamente conservada em organismos eucariotos, e está presente em praticamente todos os tipos de células desses organismos (LIU et al., 2015). Nos mamíferos, a maioria das ubiquitinas são simples (monoubiquitinadas) (CLAGUE; HERIDE; URBÉ, 2015). Porém, existe mais de um formato de ubiquitina. As proteínas podem se ligar a ubiquitinas de cadeia simples (monoubiquitina) ou a moléculas de ubiquitina com múltiplas cadeias (poli-ubiquitina), resultando em uma ampla variedade de formatos e consequentemente processos celulares diferenciados (WALSH; Di Domenico; TOSATTO, 2014).

A diferença principal está na forma como as ubiquitinas se ligam a uma proteína. Na monoubiquitinação, apenas uma ubiquitina se liga a uma lisina do substrato alvo, regulando assim a atividade de muitas proteínas celulares. Já na poliubiquitinação, uma molécula de ubiquitina se liga a um substrato de lisina na proteína alvo, e posteriormente, outras moléculas de ubiquitina vão se ligando em cadeia, a partir da primeira ubiquitina (HERRMANN; LERMAN; LERMAN, 2007).

Na figura 1 é possível identificar os formatos mais comuns da ubiquitina e suas respectivas funções celulares. A monoubiquitina (a) exibida na imagem 1 está relacionada a função de endocitose, reparo no DNA, brotamento de vírus. A monoubiquitina (b) está relacionada a endocitose. Já a poliubiquitina (c) está relacionada ao reparo do DNA, endocitose e ativação da quinase da proteína. Por fim, o ultimo dos tipos de cadeias de poliubiquitina está diretamente relacionado a degradação proteossomal (HAGLUND; DIKIC, 2005).

2.1 Importância e relação com doenças

Quanto as funções do sistema ubiquitina no organismo, historicamente, o mesmo foi primeiramente associado com a degradação de proteínas através do proteasoma, que ganhou bastante notoriedade. A degradação mediada pelo proteassoma é um mecanismo comum por onde as células renovam suas proteínas intracelulares. Neste processo, a ubiquitina se liga a uma proteína para marca-la para degradação. O sistema ubiquitina é

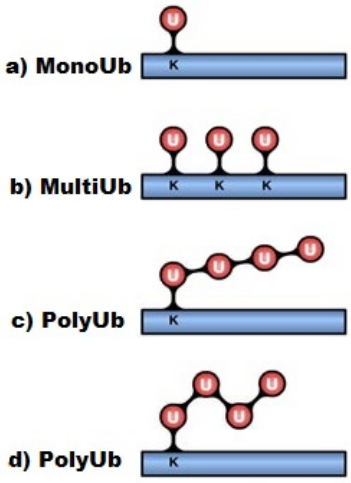

Figura 1 – Formatos da ubiquitina.

o principal coordenador da fisiologia celular através da regulação tanto da degradação de proteínas quanto através das vias de sinalizações (CLAGUE; HERIDE; URBÉ, 2015).

O sistema da ubiquitina também desempenha um papel crucial em diversos outros processos biológicos, entre eles o ciclo de divisão celular, resposta imunológica, inflamação e tradução de sinais ou sinalização alem do sistema de degradação de proteínas (CHEN et al., 2013).

Devido a sua importância, a ubiquitina também está relacionada a diversos tipos de doenças, normalmente causadas por mudanças no sistema de ubiquitinação (OKATSU et al., 2015; LIU et al., 2015). Em geral, essas são doenças neurodegenerativas como parkinson (OKATSU et al., 2015) ou doenças como o câncer (WALSH; Di Domenico; TOSATTO, 2014; LIU et al., 2015). Em células cancerígenas por exemplo, a estabilidade e o equilíbrio entre as oncoproteinas e proteínas supressoras do tumor podem ser perturbado em parte por conta de inconsistências no processo de degradação. Isto pode levar tanto a uma estabilização das oncoproteínas quanto ao aumento da degradação do supressor do tumor, contribuindo assim à progressão do câncer.

Em contrapartida, o entendimento e a busca por conhecer melhor o processo de ubiquitinação e sua atuação multi-funcional pode ser utilizado em benefício da formulações de terapias e drogas para combater doenças (WALSH; Di Domenico; TOSATTO, 2014). Em estudos recentes, tem sido desenvolvidas drogas de combate ao câncer (LIU et al., 2015) e outras doenças com resultados promissores no combate a doença. Nestes trabalhos, são investigados formas de atacar um dos componentes presentes no sistema de regulação, e assim, entender melhor o sistema de ubiquitinação e desenvolver drogas que possam regular tal sistema, assim será possível utilizar este sistema na geração de tratamentos e drogas.

2.2 Processo de Ubiquitinação

O processo de ubiquitinação é o processo em que uma ubiquitina é ligada a um resíduo de lisina de uma proteína alvo, existindo duas formas de estrutura na ubiquitinação, a monoubiquitinação e a poliubiquitinação. O processo de ubiquitinação acontece através de uma série de passos que requerem a ação de 3 enzimas principais (LIU et al., 2015). A primeira é a enzima de ativação E1, a proteína de conjugação E2 e por fim a enzima ligase E3, estas 3 enzimas tem a responsabilidade de controlar a conjugação entre as proteínas e as ubiquitinas eucariotas.

Inicialmente, as moléculas de ubiquitina são ativadas pela enzima E1 via associação a um resíduo de cisteína da enzima E1. Então, a molécula de ubiquitina ativada é transferida ao resíduo de cisteína da enzima E2 que é subsequentemente recrutada pelas ligases da enzima E3. Por fim, as ligações complexas da enzima E3 entre o substrato e a proteína ativam as formas da E2 e a proteína alvo é marcada para ser degradada, alem de facilitar a ligação entre as ubiquitinas e as proteínas marcadas (LIU et al., 2015).

A figura 2, exibe o processo de ubiquitinação e a ação das enzimas E1, E2 e E3. Nesta figura fica visível o processo e a ordem em que as enzimas atuam até que a ubiquitina esteja ligada a seu alvo. A figura demonstra a cadeia de eventos que ocorrem a partir da ação das enzimas E1, E2 e E3. Primeiro a molécula de ubiquitina é ativada pela enzima E1, a enzima E2 entra no processo ligando-se à molécula ativa, a enzima E2 é posteriormente chamada pela enzima E3 que está ligada a proteína alvo. Após o evento de ligação da enzima E3 a molécula de ubiquitinação, a molécula da ubiquitina é ligada diretamente ao substrato da proteína alvo, finalizando o processo.

2.3 Identificação de regiões de ubiquitinação

Atualmente, entre os métodos experimentais utilizados na identificação de regiões de ubiquitinação está o *site directed mutagenesis* que é um método que faz mudanças específicas e intencionais a sequencia do DNA, e a espectrometria de massa (TRUNTZER et al., 2014; UDESHI et al., 2013a; UDESHI et al., 2013b) utilizando métodos diferenciados e mais eficientes de purificação, como a utilização de anticorpos ou a utilização de proteínas que se liguem a ubiquitina além de outros trabalhos que tentam aperfeiçoar técnicas existentes ou gerar novas técnicas (GROU et al., 2015, 2015; TUCKOW et al., 2013). A técnica de espectrometria de massa é um método que consiste em identificar moléculas de interesse por meio da medição de sua massa e da caracterização de sua estrutura química através de análise física da molécula.

A espectrometria é particularmente adequada para a identificação de regiões de ubiquitinação em larga escala (CHEN et al., 2014). Entretanto, a ubiquitinação é um

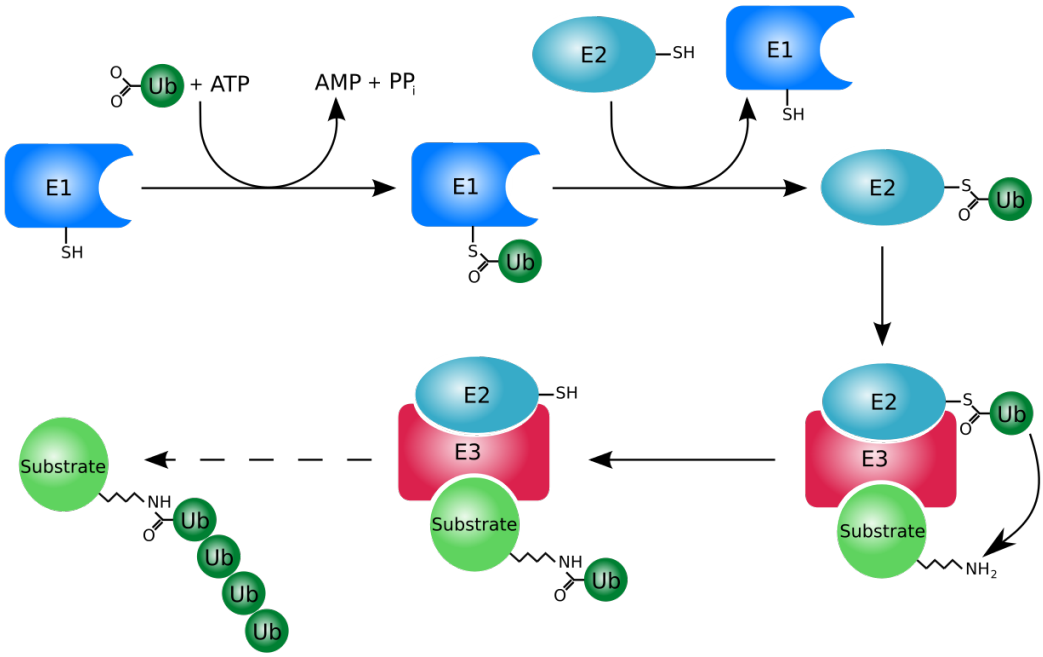

Figura 2 – Processo de ubiquitinação.

processo rápido, que além de ser uma modificação pós-translacional reversível.

As proteínas ubiquitináveis também possuem um baixo valor estequiométrico além de existir uma grande diversidade de cadeias de ubiquitina resultantes do processo de ubiquitinação, o que torna o processo de identificação bastante desafiador (UDESHI et al., 2013a). Por isso, a identificação em larga escala de proteínas ubiquitináveis e suas regiões de ubiquitinação se torna cara, custosa e demorada.

Recentemente, técnicas mais robustas e assertivas como a espectrometria de massa global, são capazes de processar milhares de regiões de ubiquitinação. Entretanto, essas técnicas experimentais frequentemente variam, capturando diferentes propriedades das proteínas (WALSH; Di Domenico; TOSATTO, 2014). Novas estratégias e métodos vem sendo desenvolvidos (TRUNTZER et al., 2014; UDESHI et al., 2013a; UDESHI et al., 2013b) ao longo do tempo como alternativas para aumentar a capacidade de captura de peptídeos modificados e conseguir identificar a existência de ubiquitinação em larga escala.

O trabalho que vem sendo desenvolvido pela comunidade mundial em completar o sequenciamento do genoma humano, seguido dos avanços na bioquímica e na bioinformática, vem gerando um inventário e novas informações a respeito dos diversos componentes do sistema relacionado a ubiquitina (CLAGUE; HERIDE; URBÉ, 2015).

Em paralelo a identificação experimental das regiões ubiquitináveis, a predição computacional das regiões de ubiquitinação vem se mostrando uma estratégia útil no trabalho de completar a anotação proteômica, na priorização de candidatos a substratos ubiquitináveis e *design* experimental de hipóteses dirigidas (CHEN et al., 2014).

Os métodos *in silico* podem predizer novas regiões de ubiquitinação através do aprendizado das propriedades extraídas do contexto das sequencias que envolvem as lisinas (pontos de ubiquitinação na cadeia da proteína) através de algoritmos de classificação e aprendizado de máquina (CHEN et al., 2014).

Essas ferramentas de aprendizado de máquina em geral são utilizadas para guiar os experimentos *in vivo*, essas ferramentas devem ser rápidas, com alta especificidade (uma taxa mínima de falso-positivos) e uma sensibilidade significante (alto índice de verdadeiro-positivo). Preditores eficientes e altamente específicos também são necessários para executar testes de hipóteses (WALSH; Di Domenico; TOSATTO, 2014).

Um número crescente de métodos computacionais vem sendo desenvolvidos recentemente. Tung e Ho(2008) são considerados os percussores, quando desenvolveram a primeira ferramenta de predição de regiões de ubiquitinação com o nome UbiPred, onde foi utilizado um SVM para identificar as possíveis regiões e também para gerar hipóteses (CHEN et al., 2014). Desde então, outras técnicas vem sendo implementadas e disponibilizadas para a comunidade que oferecem a funcionalidade de identificar regiões ubiquitinadas.

3 Algoritmos de classificação

Abaixo serão descritos os algortimos de classificação utilizados neste trabalho para identificar regiões de ubiquitinação.

3.1 Nondominated Sorting Genetic Algorithm - II

Non-dominated genetic algorithm II (DEB, 2011) é um algoritmo evolucionário multi-objetivo (MOEA, do inglês *Multi-Objective Evolutionary Algorithm*). Enquanto problemas de otimização simples possuem um único objetivo como meta, para os algoritmos multi-objetivos não existe uma única solução ótima, constituindo assim um espaço multi-dimensional. O NSGA-II é baseado em pelo menos 3 características importantes: princípio elitista; no mecanismo de preservação de diversidade (distância de multidão) e; nas soluções não dominadas (frente de Pareto). Todas as solução são comparadas baseadas nos valores obtidos de suas funções objetivo.

Quanto ao conceito de dominância, é dito que uma solução S_1 domina outra solução S_2, se ambas as condições são verdadeiras: **1.** A solução S_1 não é pior que a solução S_2 em nenhum objetivo. **2.** A solução S_1 é melhor que a solução S_2 em pelo menos um objetivo. Após a avaliação de cada solução dentro do NSGA-II, as solução são agrupadas de acordo com a quantidade de soluções que a dominam, a esses grupos chamamos de frentes. Ao final, esses grupos são ordenados, e o grupo com as soluções que não são dominadas por nenhuma outra solução é chamada de *frente de pareto*, esse grupo possui as soluções com os melhores conjuntos de resultados de todo o universo de soluções do NSGA-II.

No NSGA-II, duas ou mais funções objetivos devem ser utilizadas, estas funções objetivo avaliam as soluções e fornecem um valor de avaliação para cada objetivo determinado. Na geração de regras de classificação, vários objetivos podem ser considerados no processo de extração de informações, estes objetivos são utilizados para obter um conjunto de regras com maior valor agregado em acurácia e informação a respeito das informações das propriedades da base.

3.2 K-Nearest Neighbor

Uma das formas mais simples do aprendizado de máquina é a chamada memorização plana, ou aprendizado de rota. Uma vez que um conjunto de instâncias de treino foram memorizadas, ao chegar uma nova instância, o conjunto de treino é lido e a instância de treino mais próxima da instância nova é encontrada, e a nova instância começa a

compartilhar a mesma classe que a instância mais semelhante (WITTEN; FRANK, 2011).

O *K-Nearest Neighbor* (KNN) é um modelo baseado em instância. Neste modelo, todo o processamento é realizado no processo de classificação em si, não havendo esforço em pré-processamentos como o processamento de treino do classificador. Neste classificador, cada nova instância é comparada com todas as demais instâncias presentes na base de dados, isto é realizado para que se possa calcular uma distância entre a instância nova e as instâncias mais próximas, que serão utilizadas para classificar a nova instância.

Este classificador usa o conceito de k vizinhos próximos para efetuar a classificação, ou seja, se o parâmetro k for igual a 5, então o classificador irá procurar as 5 instância mais aproximadas dos valores da nova instância, para que a classe da maioria das instâncias retornadas seja utilizada para classificar a nova instância.

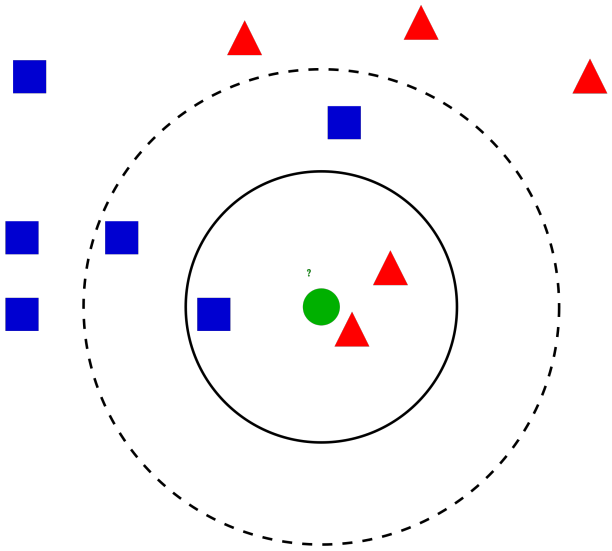

Figura 3 – Representação do algoritmo KNN, onde o círculo central representa um novo objeto que será classificado, já os triângulos vermelhos e quadrados azuis sao objetos já conhecidos, dependendo da quantidade de objetos próximos escolhidos, a classe que será atribuída ao círculo verde será a classe predominante entre os objetos mais próximos.

3.3 Random Forest

A técnica *Random Forest* é um modelo composto por um conjunto de árvores de classificação e implementa uma metodologia chamada *Ensemble learning*, que é uma técnica voltada para realização do aprendizado através da combinação de vários classificadores. Os classificadores individuais votam, e o objeto que está sendo classificado recebe sua classificação baseado na quantidade de votos para cada possível classe (HAN; KAMBER; PEI, 2012).

Os modelos de classificação em conjuntos como o *Random Forest* tendem a ser mais assertivos que seus componentes individuais (CHEN et al., 2014). Consideremos que cada um dos classificadores em um *ensemble* é uma árvore de decisão, e que as árvores são geradas utilizando uma seleção aleatória de atributos em cada nó para determinar a divisão, este conjunto de árvores formam o *Random Forest*.

A técnica *Random Forest* é bastante utilizada na área de aprendizado de máquinas, esta técnica obtém bons resultados onde é utilizada, incluindo na área de identificação de regiões ubiquitinadas já tendo sido utilizada em diversos trabalhos como em (ZANGOOEI; HABIBI; ALIZADEHSANI, 2014), e seguindo a este, os trabalhos (RADIVOJAC; VACIC, 2010; LEE et al., 2011; CHEN et al., 2014).

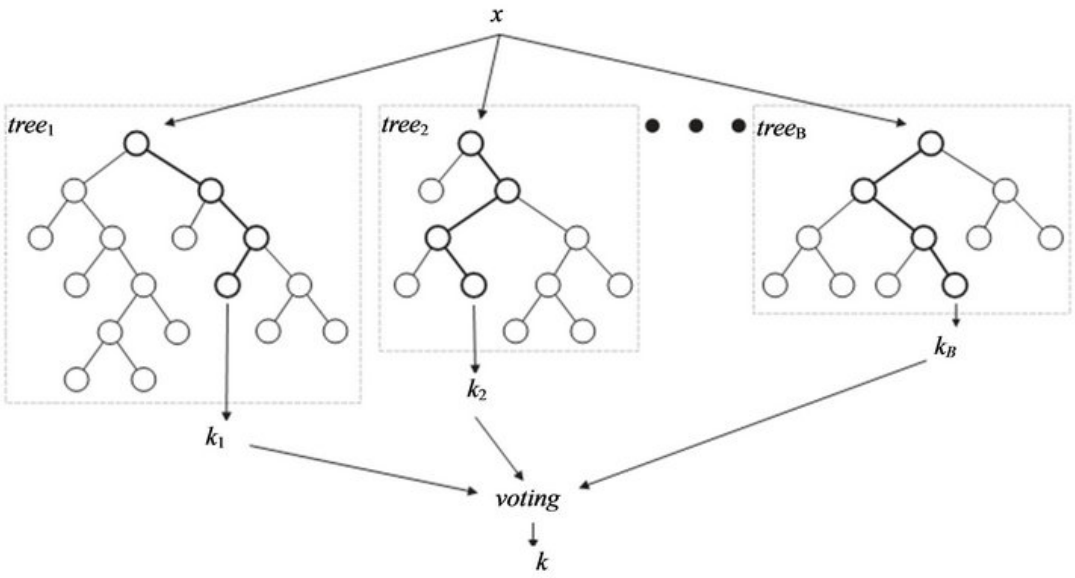

Figura 4 – Representação do algoritmos *Random Forest*, na imagem podes ser vistas as árvores que formam a floresta, a partir destas o algoritmo gera seu resultado final para cada entrada.

4 Metodologia

4.1 Base de dados

Neste trabalho, utilizamos sequencias de peptídeos retirados das proteínas de 4 organismos, a partir destas sequencias foram extraídas as propriedades utilizadas pelos classificadores para identificar as regiões ubiquitinadas. Cada uma das sequencias, possuem uma lisina (**k**) centralizada dentro de um fragmento **w** de aminoácidos, onde metade dos aminoácidos está a esquerda e outra metade a direita. Cada sequencia possui um tamanho igual a **w** caracteres. As sequencias que possuem uma quantidade de aminoácidos inferior a **w** foram complementadas com '-' (gap's) no lado com menor quantidade de caracteres, até alcançar uma quantidade de aminoácidos igual a **w**.

Todas as sequencias proteicas foram obtidas originalmente da base de dados *Uniprot Knowledgebase* (CONSORTIUM, 2014). Esta base de dados possui diversas informações e a sequencia genética de diversas proteínas (disponível em http://www.uniprot.org).

Em paralelo a obtenção das informações de proteínas, foram reunidas informações sobre regiões de ubiquitinação de 4 organismos diferentes: *S. cerevisiae*, *H. sapiens*, *M. musculus* e *A. thaliana*. Estas regiões originam de 5 trabalhos diferentes (UDESHI et al., 2013b; KIM et al., 2013; MERTINS et al., 2013; STARITA et al., 2012; WAGNER et al., 2012). Nestes trabalhos, as regiões de ubiquitinação foram amplamente estudadas e experimentadas. As informações originadas nestes trabalhos possuem diversas informações sobre os experimentos, informações sobre anti-corpos utilizados e também o percentual de chance das regiões serem ubiquitinadas.

Todas as lisinas presentes nas sequencias de proteínas dos 4 organismos foram mapeadas e suas posições anotadas. A principios, todas as lisinas mapeadas foram marcadas como não ubiquitinadas, a partir das informações obtidas das bases de dados de regiões ubiquitinadas, as lisinas foram marcadas com sua classificação correta (Ubiquitinável ou não ubiquitinável). Ao final deste processo, foi possível montar um conjunto de posições das regiões de ubiquitinação, as regiões positivas (ubiquitinadas) são as que foram comprovadas através de experimentos de outros trabalhos, já as demais regiões são consideradas não ubiquitináveis.

É importante notar uma dificuldade de identificar negativos absolutos, por mais que existam regiões não ubiquitináveis comprovadas, a grande maioria das regiões ainda não foi experimentada para que se comprove que são de fato não-ubiquitináveis, entretanto todas foram consideradas neste trabalho.

O mapeamento destas regiões ubiquitinadas forneceu o suporte necessário para

construir 5 conjuntos básicos de dados: 4 destes conjuntos possuíam informações específicas de um dos organismos, e 1 conjunto era a união dos outros conjuntos. Assim, 4 bases de dados são organismo-específicas e 1 base geral. As bases de dados foram nomeadas de acordo com a informação do organismo que continha, ficando assim: S.dataset (*S. cerevisiae*), H.dataset (*H. sapiens*), M.dataset (*M. musculus*) and A.dataset (*A. thaliana*) e G.dataset (base geral). Todas as bases de dados são balanceadas, possuindo uma quantidade igual de regiões ubiquitinadas e não ubiquitinadas. Na tabela 1 são mostradas as quantidades de amostras de cada conjunto de dados.

Tabela 1 – Número de proteínas e regiões ubiquitinadas.

Dataset	Número de proteínas ubiquitinadas	Regiões ubiquitinadas
A. Thaliana	167	668
S. Cerevisiae	418	749
M. Musculus	4040	17.945
H. Sapiens	3657	40.654

Após obter todas as informações sobre as posições das regiões de ubiquitinação, o próximo passo foi indicar qual o tamanho do fragmento **w** que deve ser utilizado para complementar a região a esquerda e a direita da lisina **k** centralizada. Após indicar o valor deste fragmento, a sequencia é extraída da proteína e adicionada em um arquivo externo.

4.2 Propriedades de Classificação

Para cada sequência na base de dados, foram obtidos dois tipos de informação baseado na estrutura da sequencia dos aminoácidos: Composição de aminoácidos (AAC - *Amino Acid Composition*) (RADIVOJAC; VACIC, 2010; LEE et al., 2011; CHEN et al., 2013) e propriedades físico químicas (PCP - physiochemical properties) baseadas na base de dados AAIndex (KAWASHIMA; POKAROWSKI, 2008). Estas duas propriedades foram amplamente utilizadas em trabalhos que tratavam da identificação *in silico* de regiões de ubiquitinação.

4.2.1 Propriedades Físico Químicas

AAIndex (KAWASHIMA; POKAROWSKI, 2008) é uma base de dados de índices de aminoácidos e matrizes de mutação. Atualmente existem 544 PCPs diferentes cadastradas na base de dados. Cada propriedade possui um índice de aminoácidos que contém valores numéricos, e cada um desses valores representa o valor da propriedade para cada aminoácido. A imagem 5 representa uma dessas propriedades, todos os índices da propriedade ARGP820101 para cada aminoácido estão representados.

A estrutura e função de cada proteína é amplamente dependente de várias propriedades da estrutura dos aminoácidos que a compõem. As PCPs foram aplicadas com

```
H ARGP820101
D Hydrophobicity index (Argos et al., 1982)
R LIT:0901079b PMID:7151796
A Argos, P., Rao, J.K.M. and Hargrave, P.A.
T Structural prediction of membrane-bound proteins
J Eur. J. Biochem. 128, 565-575 (1982)
C JOND750101    1.000  SIMZ760101    0.967  GOLD730101    0.936
  TAKK010101    0.906  MEEJ810101    0.891  ROSM880104    0.872
  CIDH920105    0.867  LEVM760106    0.865  CIDH920102    0.862
  MEEJ800102    0.855  MEEJ810102    0.853  ZHOH040101    0.841
  CIDH920103    0.827  PLIV810101    0.820  CIDH920104    0.819
  LEVM760107    0.806  NOZY710101    0.800  GUYH850103   -0.808
  PARJ860101   -0.835  WOLS870101   -0.838  BULH740101   -0.854
I    A/L     R/K     N/M     D/F     C/P     Q/S     E/T     G/W     H/Y     I/V
    0.61    0.60    0.06    0.46    1.07     0.    0.47    0.07    0.61    2.22
    1.53    1.15    1.18    2.02    1.95    0.05    0.05    2.65    1.88    1.32
```

Figura 5 – Representação da tabela de índices da propriedade ARGP820101 extraída da base de dados AAIndex.

sucesso na predição de tais modificações executadas pela ubiquitina. Diversos trabalhos (TUNG; HO, 2007; RADIVOJAC; VACIC, 2010) demonstram que este tipo de propriedade é um tipo valioso de informação, que colabora na predição de regiões de ubiquitinação.

Das 544 PCPs existentes cadastradas no AAIndex, 31 foram escolhidas para ser utilizadas neste trabalho, esta escolha foi feita a partir do trabalho de Tung (2008) que utilizou um algoritmo genético para selecionar as propriedades de maior contribuição para serem utilizadas em um SVM, estas mesmas propriedades estão presentes na tabela 2. Neste trabalho, após experimentação, foi constatado que estas 31 propriedades possuem maior potencial quando utilizadas na identificação das regiões de ubiquitinação. Por isso, apenas estas propriedades foram utilizadas nos testes, o que contribuiu na velocidade de processamento e utilização de recursos já que não foram utilizadas todas as 544 propriedades. Estas 31 propriedades também foram utilizadas por outros trabalhos que também utilizaram Tung(2008) como referência.

Na tabela 2, a coluna "Identificador AAindex"é o código de identificação dentro da base de dados AAIndex, em seguida temos a descrição da propriedade e por último um código de identificação que foi adotado neste trabalho para facilitar o entendimento, ele também doi utilizado dentro do NSGA-II como referência.

Como explicado, cada aminoácido em torno da lisina possui um valor específico, para calcular as propriedades físico químicas dos fragmentos das sequencias utilizadas na base de dados é preciso tirar a média dos valores de dos aminoácidos presentes no fragmento. Assim, na base de dados final utilizada pelos classificadores, teremos para cada fragmento 31 propriedades adjacentes que representam cada uma das propriedades físico químicas.

Tabela 2 – Propriedades fisico químicas e bioquímicas presentes na base de dados AAIndex.

Identificador AAindex	Description	ID
GUYH850105	Apparent partition energies calculated from Chothia index	C0
GUYH850104	Apparent partition energies calculated from Janin index	C1
GRAR740102	Polarity	C2
CORJ870101	NNEIG index	C3
HUTJ700103	Entropy of formation	C4
GEOR030108	Linker propensity from helical (annotated by DSSP) dataset	C5
CEDJ970102	Composition of amino acids in anchored proteins (percent)	C6
NAKH920101	AA composition of CYT of single-spanning proteins	C7
RACS770102	Average reduced distance for side chain	C8
FAUJ880101	Graph shape index	C9
NAKH900101	AA composition of total proteins	C10
HARY940101	Mean volumes of residues buried in protein interiors	C11
NADH010102	Hydropathy scale based on self-information values in the two-state model	C12
CEDJ970103	Composition of amino acids in membrane proteins (percent)	C13
BULH740102	Apparent partial specific volume	C14
CEDJ970105	Composition of amino acids in nuclear proteins (percent)	C15
FAUJ830101	Hydrophobic parameter pi	C16
LEVM780101	Normalized frequency of alpha-helix, with weights	C17
LEVM780106	Normalized frequency of reverse turn, unweighted	C18
QIAN880129	Weights for coil at the window position of -4	C19
BROC820102	Retention coefficient in HFBA	C20
BROC820101	Retention coefficient in TFA	C21
MEIH800102	Average reduced distance for side chain	C22
MEIH800103	Average side chain orientation angle	C23
ZIMJ680102	Bulkiness	C24
ROSM880102	Side chain hydropathy, corrected for solvation	C25
CHOC760103	Proportion of residues 95% buried	C26
KRIW790102	Fraction of site occupied by water	C27
EISD840101	Consensus normalized hydrophobicity scale	C28
ZHOH040102	The relative stability scale extracted from mutation experiments	C29
PUNT030102	Knowledge-based membrane-propensity scale from 3D_Helix in MPtopo databases	C30

4.2.2 Composição de Aminoácidos

A composição de aminoácidos (AAC) representa o número de ocorrências dos aminoácidos em torno da lisina centralizada. Cada fragmento da sequencia possui propriedades adjacentes representando cada um dos aminoácidos, o valor dessas propriedades é a quantidade de vezes que o aminoácido em questão aparece dentro da sequencia.

Esta propriedade se mostrou muito importante, uma vez que os resultados dos

Tabela 3 – Propriedades fisico químicas e bioquímicas presentes na base de dados AAIndex.

Propriedades
C2 - Polarity
C3 - NNEIG index
C0 - Apparent partition energies calculated from Chothia index
C4 - Entropy of formation
C5 - Linker propensity from helical (annotated by DSSP) dataset
C1 - Apparent partition energies calculated from Janin index
C7 - AA composition of CYT of single-spanning proteins
C6 - Composition of amino acids in anchored proteins (percent)

classificadores melhoraram consideravelmente quando utilizaram esta propriedade conforme demonstrado em trabalhos diversos (CHEN et al., 2011; LEE et al., 2011; CHEN et al., 2013; CHEN et al., 2014; TUNG; HO, 2008; CHEN et al., 2013). Outras propostas de metodologias semelhantes também são implementadas nestes trabalhos como em (CHEN et al., 2011) onde não é apenas utilizado o número de ocorrências simples dos aminoácidos nos fragmentos, mas também a ocorrência dos pares de aminoácidos dentro dos fragmentos. Entretanto esta metodologia não foi implementada neste trabalho, onde utilizamos apenas a frequência simples dos aminoácidos pois preferimos iniciar utilizando parâmetros mais simples de serem trabalhados e entendidos.

4.3 Base de dados UPIS

Devido a grande quantidade de dados utilizados no trabalho, e devido o fato dos mesmos possuírem diversas origens diferentes, foi necessário modelar e construir um meio de organizar e padronizar tudo em um único local. A partir desta necessidade foi construída uma base de dados que unifica diversas fontes de dados, a esta base de dados chamou-se UPIS (*Ubiquitin-Protein Integration Service*), esta base de dados também está integrada com uma série de ferramentas de auxílio a essas informações. A figura 6 representa todo o *workflow* do trabalho com a base de dados, desde a obtenção dos dados, pré-processamento, armazenamento na base de dados UPIS, extração dos fragmentos de ubiquitinação, calculo das propriedades e classificação dos dados em ubiquitinável e não-ubiquitinável. Já a figura 7 representa o primeiro passo de construção da base que é a coleta e integração de dados.

Inicialmente, os dados são importados para a base de dados UPIS, nessa importação de dados estão incluídos dados da base Uniprot, AAIndex e dados de regiões de ubiquitinação. A partir desta importação para a base UPIS, são gerados as bases de dados com as regiões de ubiquitinação, com fragmentos centralizados na lisina principal. Os arquivos contendo as propriedades PCP e AAC são gerados a partir dos fragmentos de proteínas com uma lisina centralizada. Por fim, os arquivos com as propriedades PCP e AAC são processados pelos algoritmos de classificação e os resultados das classificações e

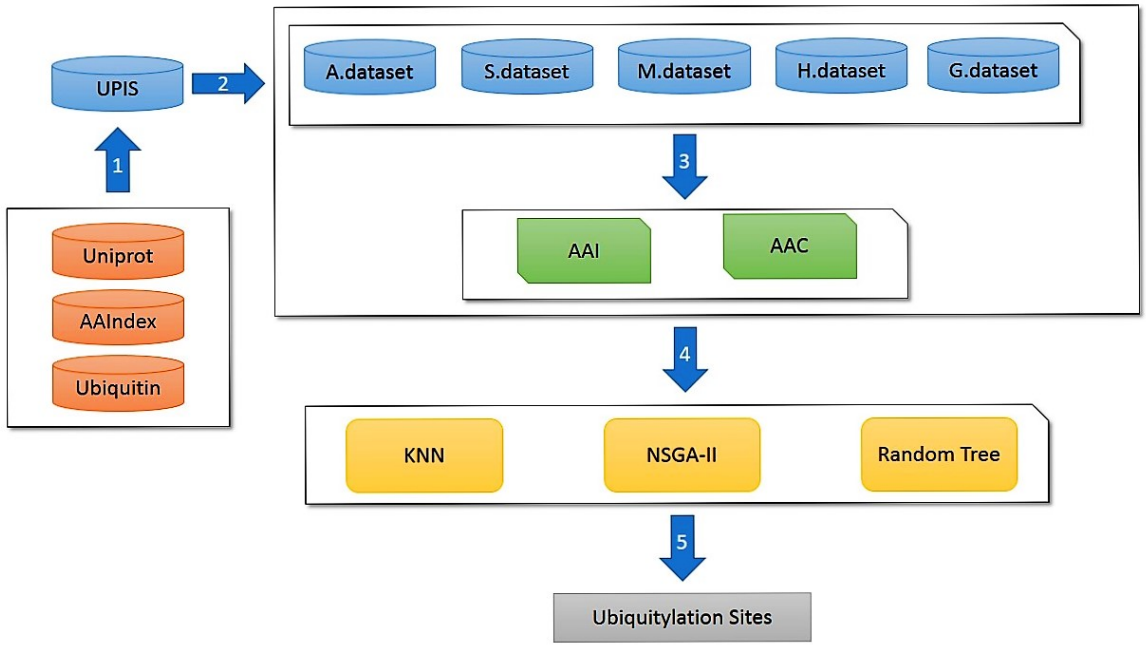

Figura 6 – Workflow do trabalho

das novas possíveis regiões de classificação são geradas.

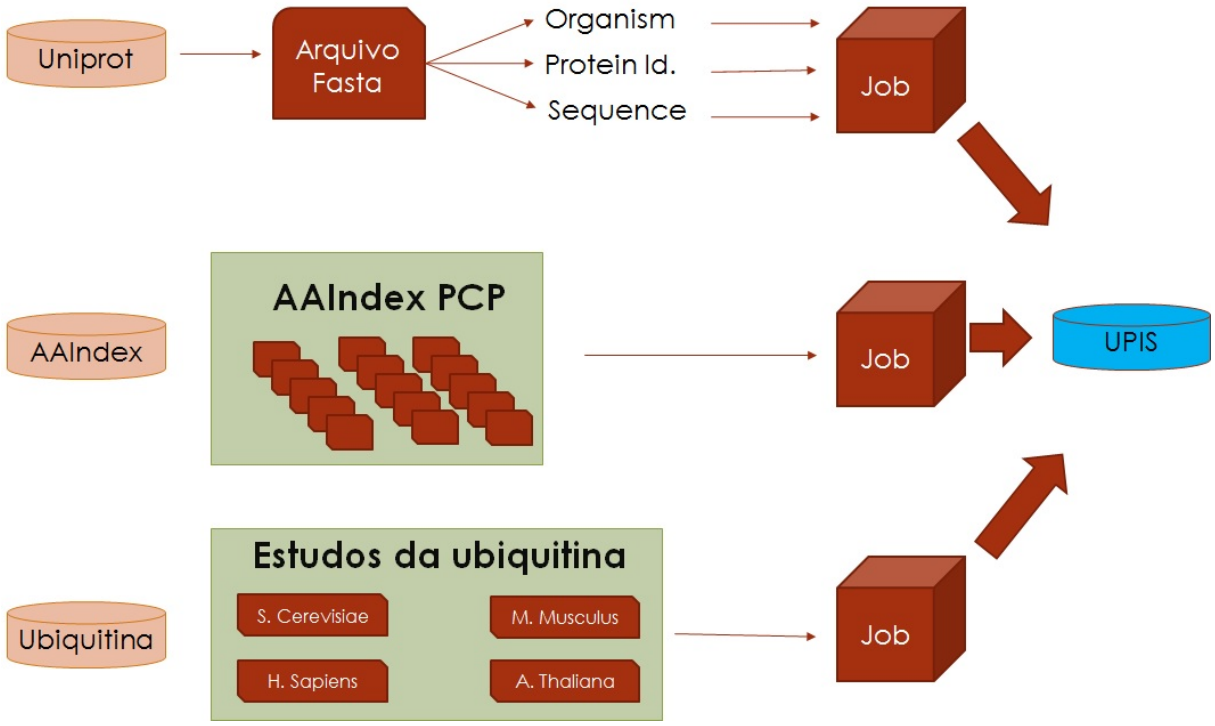

Figura 7 – Coleta e integração de dados ao UPIS.

4.4 Configuração do NSGA-II

4.4.1 Representação do cromossomo

Neste trabalho, a versão original de (DEB, 2011) foi adaptada para o problema de classificação. Entre as adaptações realizadas, estão a configuração e composição dos indivíduos do AG e os operadores genéticos. Estes foram configurados de acordo com o tipo de informação dos indivíduos que foram configuradas para representarem regras de classificação. Também fazem parte da adaptação, as funções objetivas utilizadas na otimização.

Cada indivíduo (também chamado cromossomo) no conjunto da população representa uma regra de classificação, a estrutura de cada cromossomo é representada na figura 8. Estes cromossomos possuem uma estrutura interna chamada *gene*, cada *gene* representa uma informação na base de dados. O número de *genes* dentro de cada indivíduo é igual ao número de propriedades consideradas na classificação. Na configuração utilizada, cada *gene* possui três propriedades diferentes chamadas *alelo*. O primeiro *alelo* da configuração representa um ponto de corte, durante o processamento e teste das regras de classificação geradas por cada indivíduo, caso o valor deste alelo esteja abaixo do ponto de corte, o gene inteiro é desconsiderado na geração da regra de classificação final. O segundo *alelo* representa o operador de comparação que será utilizado para o valor do gene. E por fim, o terceiro e ultimo *alelo* é o valor da regra.

Gene 1	Gene 2	[...]	Gene **n**
8 | $\geq$ | 0.23	5 | $<$ | 0.54	... | ... | ...	10 | $<$ | 0.75

Figura 8 – Representação do indivíduo: cada cromossomo possui esta configuração no NSGA-II

A partir de cada indivíduo, será gerada uma regra de classificação, as regras serão formadas pela união dos *alelos* no menor nível, e depois pela união dos *genes*. Para exemplificar a formação das regras, a seguir é descrita a regra gerada a partir do indivíduo da figura 8.

IF *Propriedade. 1* $\geq$ *0.23* **And** *Propriedade n* $<$ *0.75*
THEN Ubiquitinável.

Considerando que o valor do ponto de corte seja igual a 8, a regra extraída do indivíduo irá conter apenas os genes em que o *alelo* que representa o valor de corte, seja maior ou igual a 8, portanto, apenas os *genes* **1** e **n** serão considerados na regra. A seguir, os genes que não foram retirados formarão a regra do tipo **IF-THEN**. Para entender

melhor o uso da regra, conforme explicado anteriormente, cada gene representa uma propriedade dentro da base de dados, portanto, a regra sugere que são ubiquitináveis, todos os objetos na base de dados onde a propriedade 1 seja maior ou igual a 0.23 e a propriedade **n** seja menor que 0.75.

4.4.2 Operadores Genéticos

Abaixo são descritos todos os operadores genéticos utilizados no AG:

- **Seleção dos pais:** O método utilizado foi o método de torneio, onde 3 indivíduos são aleatoriamente selecionados da população descendente, e o melhor entre os 3 é selecionado para participar do processo de *cruzamento* que dará origem a próxima geração.
- **Cruzamento:** Os indivíduos na população são selecionados 2 a 2. Para cada *alelo* de cada gene, seus valores são utilizados para determinar os valores dos novos indivíduos utilizando a operação $N_n = a_1 + (a_2 - a_1) * r$, onde N_n é o valor do *alelo* do novo indivíduo, a_1 é o valor do *alelo* do indivíduo 1, a_2 é o valor do *alelo* do indivíduo 2, e r é o valor aleatório de distribuição gaussiana.
- **Mutação:** Para cada gene do indivíduo, existe uma probabilidade de o *alelo* sofrer mutação. O novo valor do *alelo* será determinado pela distribuição gaussiana.
- **Seleção de sobrevivência:** Será utilizada a seleção elitista. Neste método, todos os indivíduos, tanto dos indivíduos pais quanto os indivíduos descendentes, são realocados e ordenados em *fronts* e ordenados pela valor de dominância e sua distância de multidão, e apenas os $\boldsymbol{p}$ melhores serão selecionados para a próxima geração, onde $\boldsymbol{p}$ é o tamanho da população.

5 Resultados

Três algoritmos de classificação foram utilizados para analisar os dois tipos de informação da base de dados. Os resultados alcançados pelos algoritmos, assim como as demais informações obtidas a partir dos testes serão descritas a seguir.

5.1 Testes Realizados

5.1.1 Ambiente

O ambiente de testes utilizado foi um PC utilizando windows 10 como sistema operacional, possuindo 8 Gb de memória RAM, possuindo um processador Intel CORE i7. O algoritmo genético foi implementado em Java, já os demais algoritmos foram retirados do pacote WEKA (HALL et al., 2009).

A base de dados foi construída utilizando o sistema de banco de dados SQL SERVER 2008. Para executar a tarefa de integrar os dados retirados da base AAIndex, UNIPROT e dos demais trabalhos de onde foram obtidas as localizações das regiões ubiquitinadas, foram construídos scripts em Python para automatizar a integração. A análise dos dados, resultados e gráficos criados foram realizadas através de uma biblioteca construída na linguagem R.

5.1.2 Utilização da base de dados

Para avaliar a performance de predição dos algoritmos, foram executados testes utilizando o modelo de validação cruzada ou *k-folds*. Neste modelo, os dados iniciais são particionados em K subconjuntos ou *folds*, cada *fold* possui aproximadamente o mesmo tamanho que as demais. Após a divisão em *folds*, são executados tantas iterações de treino quanto a quantidade de *folds*. A cada iteração, um dos *fold* é escolhido para ser o conjunto de testes, e as demais $(k-1)$ *folds* o conjunto de treino.

Durante a montagem da base de dados, também foi avaliado o impacto do tamanho das janelas de seleção nas predições. Esta janela determina quantos aminoácidos a direita e a esquerda da lisina central são consideradas para montar a sequencia a ser avaliada. Cinco tamanhos de janelas diferentes foram consideradas nos testes (20, 21, 25, 27 e 41), para determinar os tamanhos que seriam testados, levou-se em consideração trabalhos similares.

Este procedimento de divisão e utilização da base de dados foi realizado para todas as 5 bases, 4 bases específicas de cada organismo, e uma geral contendo todas as outras.

5.1.3 Execução dos testes

Todos os testes foram realizados com os dois tipos de informações das bases de dados (informações físico-químicas e de composição do aminoácido). Os testes foram feitos em todas as bases de dados especificas e a base de dados completa com todas as sequencias. No total, o número de sequencias distintas analisadas somaram 120.032 sequencias. Dessas, 60.016 sequencias eram ubiquitinadas, e 60.016 não-ubiquitinadas.

A partir dos dois tipos de base de dados (uma contendo informações físico-químicas e outra da composição de aminoácidos), da construção de bases com tamanho de janelas distintas para cada tipo e da divisão em *k-folds* das mesmas, os algoritmos foram treinados.

Para obter os resultados do NSGA-II, tirou-se uma média de resultados de 10 execuções para cada base de dados gerada para os testes. Este processo foi necessário pois o NSGA-II possui uma natureza estocástica e não determinística.

O NSGA-II foi testado em muitos conjuntos de parâmetros, entre os parâmetros que variaram está a quantidade de gerações, tamanho da população, probabilidade de mutação e probabilidade de cruzamento, conforme demonstrado na tabela 4, onde alguns entre os principais conjuntos de parâmetros são exibidos. Cada conjunto de parâmetros foi testado pelo menos 10 vezes, dessas execuções foram retirados a maioria dos resultados e análises. Quanto ao Random Forest e o KNN, os parâmetros do Random Forest foram mantidos com seus valores padrões no WEKA, e no KNN, o valor K foi testado com os valores: 1,3,5,7,9,11,15,21 em cada execução.

Tabela 4 – Parâmetros de testes utilizados nos experimentos do NSGA-II.

Gerações	População	Probabilidade de Mutação	Probabilidade de Cruzamento
100	100	10%	70%
200	200	20%	60%
300	350	25%	80%
500	500	10%	75%

Realizados os testes, os resultados foram armazenados para posterior comparação entre os resultados dos algoritmos.

5.1.4 Métricas de Avaliação

Diversas métricas de avaliação de classificadores podem ser utilizadas. Entretanto, neste trabalho, apenas algumas delas foram empregadas por se tratar de indicadores bastante conhecidos, e já terem sido utilizadas anteriormente em diversos trabalhos, o que facilita a comparação de resultados.

Problemas de otimização multi-objetivo envolve mais de uma função objetivo que deve ser minimizada ou maximizada. Neste estudo, as funções objetivas utilizadas no

NSGA-II foram o Recall (Rec., descrita na equação 5.1) e Precisão (Pr., equação 5.2). Estas duas funções objetivo foram maximizadas no NSGA-II.

Também utilizamos a função de acurácia (Acc., equação 5.3) como um método de validação, avaliação e de comparação entre os resultados. Embora este atributo não tenha sido utilizado como uma função objetiva no NSGA-II, a mesma também foi calculada para cada individuo. Esta função é amplamente utilizada em trabalhos como este, o que facilita a comparação de resultados com outros algoritmos.

Estas medidas serão definidas a seguir, as notações TP, TN, FP, FN utilizadas nas equações são respectivamente: Verdadeiro Positivo (*True Positive*), Verdadeiro Negativo (*True Negative*), Falso Positivo (*False Positive*), Falso Negativo (*False Negative*) respectivamente:

$$\mathbf{Rec} = \frac{TP}{TP + FN} \tag{5.1}$$

$$\mathbf{Pr} = \frac{TP}{TP + FP} \tag{5.2}$$

$$\mathbf{Ac} = \frac{TP + TN}{TP + FP + TN + FN} \tag{5.3}$$

Recall (Eq. 5.1) é o percentual do total de amostras positivas que foram corretamente classificadas como positivas pelos classificadores. Em um contexto médico, o *Recall* é considerado um objetivo primário pois seu objetivo é identificar todos os casos positivos (ZANGOOEI; HABIBI; ALIZADEHSANI, 2014). Precisão (Eq. 5.2) indica o percentual de casos classificados como positivos que realmente são positivos. Já a Acurácia (Eq. 5.3) é a proporção dos casos corretamente classificados pelo total de casos analisados, é uma medida estatística de quão bem o classificador identifica corretamente ou exclui corretamente cada caso.

5.2 Análise dos resultados

5.2.1 Resultados obtidos por tipo de propriedade

Para avaliar os classificadores foi utilizado o valor de acurácia fornecido pelos algoritmos, o aplicativo Weka já fornece esse e outros valores ao final de sua execução, já para o NSGA-II, durante a execução dos testes, esta métrica era calculada junto com as funções objetivo.

Os resultados apresentados abaixo foram retirados dos testes que retornaram os melhores resultados. A tabela 5 mostra os resultados obtidos por essas execuções para

as bases de dados com informações físico-químicas apenas, e a tabela 6 da base de dados com informações de composição de aminoácido.

Tabela 5 – Acurácia obtida pelos algoritmos utilizando base de dados com informações físico-químicas.

Dataset	Random Forest	KNN	NSGA-II
G.Dataset	68.76%	64.46%	63.00%
S.Dataset	74.36%	71.36%	78.40%
H.Dataset	63.88%	67.63%	63.25%
M.Dataset	65.66%	69.71%	64.14%
A.Dataset	70.95%	70.95%	73.87%

Tabela 6 – Acurácia obtida pelos algoritmos utilizando base de dados com informações de composição de aminoácidos.

Dataset	Random Forest	KNN	NSGA-II
G.Dataset	67.21%	63.00%	62.50%
S.Dataset	84.17%	78.83%	81.60%
H.Dataset	69.71%	65.34%	62.60%
M.Dataset	71.69%	66.41%	66.09%
A.Dataset	76.79%	70.73%	73.00%

Todos os algoritmos utilizados alcançaram bons resultados. Entretanto, para cada uma das bases de dados e tipo de informação, um dos algoritmos alcançou melhores resultados que outros. Como mencionado por (CHEN et al., 2014), não existe um classificador universal para as regiões de ubiquitinação para todas as espécies. Para cada espécie é recomendado um tipo de classificador. Por exemplo, para determinar novas possíveis regiões de ubiquitinação no organismo *S. Cerevisiae*, a melhor escolha é usar *Random Forest* ou NSGA-II. Para o organismo *A. Thaliana*, o NSGA-II é recomendado, mas para *M. Musculus* o usuário deve usar o Random Forest.

Os preditores também obtiveram diferentes resultados para cada um dos tipos de informações utilizadas como entradas. Como já indicado anteriormente por (LEE et al., 2011; CAI et al., 2012; CHEN et al., 2013; CHEN et al., 2014), a utilização da composição de aminoácidos provê melhores resultados que as propriedades físico-químicas extraídas do AAIndex. Apesar dos classificadores também obterem bons resultados com as propriedades fisico-químicas, estes resultados ainda são inferiores aos obtidos pela composição de aminoácidos.

Inicialmente, apenas as informações físico-químicas foram extraídas e utilizadas a partir das bases de dados. Entretanto, posteriormente as informações de composição de amino-ácido foram introduzidas no experimento. A utilização dessas informações se mostrou eficiente no que se diz respeito a performance das classificações, com isso melhores resultados foram obtidos na classificação. Estes fatos são consistentes com o que é afirmado

por outros trabalhos como em (LEE et al., 2011; CAI et al., 2012; CHEN et al., 2013; CHEN et al., 2014).

Em algumas bases de dados o NSGA-II obteve resultados pouco inferiores aos resultados obtidos pelo melhor algoritmo. Parte dos possíveis motivos para que isto tenha ocorrido está relacionado a codificação utilizada nos operadores genéticos do algoritmo. Vários outros operadores podem ser utilizados, estes podem ser vistos em problemas similares, onde se usa algoritmo genético. Entre essas possibilidades de operadores, pode-se existir algum que se encaixe melhor no problema de classificação.

Outra possível solução que pode ser empregada diz respeito aos valores dentro da base de dados. Por conter valores contínuos como parâmetro de entrada para os classificadores, o NSGA-II pode ter tido dificuldades para sintonizar os valores contidos nas regras de classificação e por isso não ter alcançado melhores resultados. Uma possível solução que pode ser implementada em trabalhos futuros é a discretização dos valores de entrada, substituindo por exemplo intervalos existentes entre os dados de uma propriedade por valores discretos, ou ao menos reduzir o número de casas decimais consideradas no treino. Esta solução diminuiria o impacto das grandes distâncias entre os valores de cada objeto na base de dados.

5.2.2 NSGA-II

Neste tópico sera tratado especificamente dos resultados obtidos pelo NSGA-II e das observações possíveis de serem feitas após uma análise mais aprofundada dos resultados obtidos.

5.2.2.1 Regras de classificação

Ao final de seu processamento, o NSGA-II gera um conjunto de modelos de classificação (regras IF-THEN). A figura 9 mostra os resultados alcançados pelo NSGA-II em uma de suas execuções utilizando a base de dados *S.dataset*. Na figura, são exibidas todas as regras presentes na frente de pareto da ultima geração da execução em questão.

Na tabela estão sendo exibidas as regras de acordo com seus valores de *Recall* e Precisão, é importante notar que algumas das regras possuem o valor máximo de *Recall*, em outros casos algumas das regras possuem o valor máximo de precisão, mas também existe um grupo que está em uma área intermediária entre os valores máximos de *Recall* e precisão.

Esta variedade de soluções permite um maior potencial de escolha para tomada de decisões, também permite a possibilidade de analisar a formação das regras, quais são as diferenças entre as mesmas e de também criar hipóteses a respeito da base de dados e das propriedades de classificação.

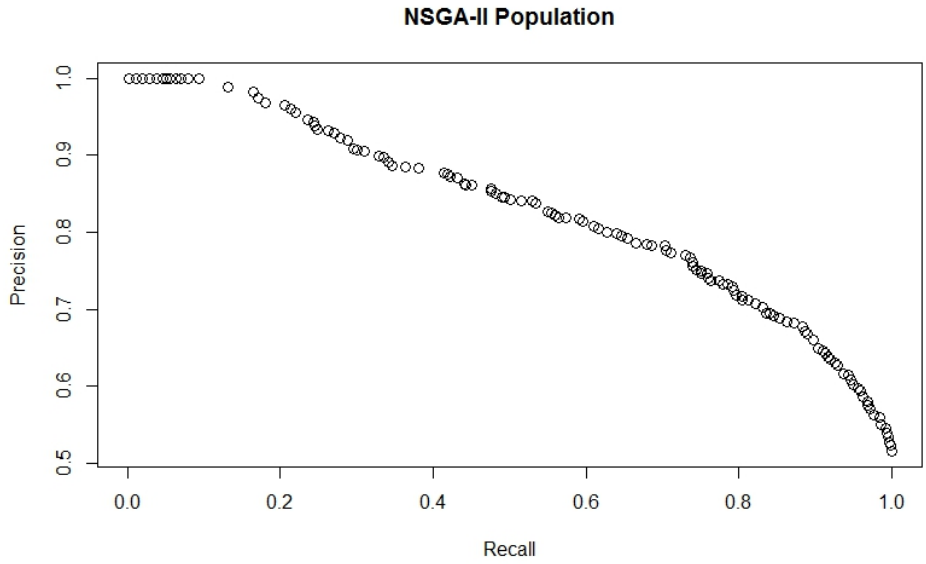

Figura 9 – População do NSGA-II com 150 indivíduos. Cada indivíduo possui uma combinação única de *recall* e precisão entre sua população. A população exibida também é a frente de pareto.

Um exemplo do potencial de tomada de decisão é a escolha de qual regra aplicar, caso o interesse na pesquisa seja de ter mais precisão e certeza quanto as regiões ubiquitinadas que serão investigadas, pode-se escolher regra que possuam um valor de precisão maior que as demais, caso o interesse seja simplesmente abranger a busca e encontrar a maior quantidade possível, considerando uma possibilidade de acerto maior que uma escolha aleatória, podem-se selecionar uma das regras que possuem um valor de *recall* máximo.

Para melhor exemplificar as regra de classificação geradas pelo NSGA-II, a partir da mesma população exibida na figura 9, foram selecionadas 3 regras de classificação exibidas na tabela 7. Essas regras são mostradas em sua forma completa com seus respectivos resultados de acurácia, *recall* e precisão. A partir da tabela é possível verificar por exemplo quais são os parâmetros que diferenciam uma regra com maior *recall* e uma que possui maior precisão, e uma terceira com um valor intermediário entre os dois resultados. A primeira regra possui uma acurácia de 81,60%, entretanto seu valor de recall e precisão é de 79,37% e 83,33% respectivamente, já as demais regras, embora possuam uma acurácia semelhante a primeira, os valores de *recall* e precisão são diferenciados. Novamente o poder de decisão sobre qual regra melhor irá resolver o problema em questão fica a cargo do problema biológico a ser analisado.

Entre as maiores contribuições do NSGA-II está a variedade de soluções providas ao final e durante a execução do algoritmo, alem do mais, o modelo da solução possui uma baixa complexidade tanto para a análise da solução proposta quanto para a aplicação da mesma. Esta baixa complexidade é permitida graças a capacidade de adaptação do

Tabela 7 – Regras extraídas da frente de pareto de uma das execuções do NSGA-II. Essas regras foram usadas para identificar regiões de ubiquitinação na base de dados S.database.

Regra	Acurácia	Precisão	Recall
IF *C2 < 0.12* **AND** *C7 < 0.34* **AND** *C8 < 0.62* **AND** *C12 < 0.42* **AND** *C13 < 0.58* **AND** *C15 < 0.32* **AND** *C18 < 0.22* **AND** *C21 < 0.75* **AND** *C22 < 0.34* **THEN Ubiquitylated**	81.6%	83.33%	79.37%
IF *C2 < 0.13* **AND** *C7 < 0.34* **AND** *C8 < 0.44* **AND** *C9 < 0.51* **AND** *C12 < 0.36* **AND** *C13 < 0.40* **AND** *C15 < 0.32* **AND** *C18 < 0.23* **AND** *C21 < 0.70* **AND** *C22 <0.37* **THEN Ubiquitylated**	79.2%	89.36%	66.67%
IF *C2 < 0.13* **AND** *C7 < 0.33* **AND** *C8 < 0.62* **AND** *C12 < 0.41* **AND** *C13 < 0.57* **AND** *C15 < 0.39* **AND** *C18 < 0.24* **AND** *C21 < 0.85* **AND** *C22 < 0.39* **THEN Ubiquitylated**	80%	79.69%	80.95%

algoritmo genético que é modelado de acordo com as necessidades de resolução do problema. Neste caso, a saída do NSGA-II é a própria regra de classificação.

Outra contribuição do algoritmo é o filtro natural de parâmetros aplicado em suas soluções. Durante a evolução dos indivíduos, os parâmetros que não colaboram ou possuem um baixo nível de colaboração tornando-os dispensáveis a regra, são automaticamente removidos da solução final. Isto ocorre pois o AG durante sua evolução consegue perceber que com uma regra sem determinado parâmetro consegue obter o mesmo resultado ou um resultado superior a uma regra com o parâmetro, ou então quando o parâmetro prejudica a classificação e prejudica o resultado final das regras objetivo. Com isso, o AG pode fornecer regras mais simple e com menos níveis de validação necessário para executar uma classificação.

Quanto ao consumo de recursos do algoritmo. O NSGA-II necessita que seus parâmetros sejam bem configurados, e dependendo desta configuração, dos operadores genéticos, do tamanho da população do algoritmo e da base de dados utilizada, o algoritmo pode executar em minutos ou levar grandes quantidades de horas para finalizar. Os parâmetros escolhidos influenciam diretamente no processamento, alem de também influenciar nos resultados. Neste trabalho o algoritmo levou cerca de 7 minutos para ser executado em alguns testes, já em outros levou cerca de 6 horas.

5.2.2.2 Importância dos parâmetros

A partir das das regras de classificação geradas pelo do NSGA-II, os parâmetros das regras foram analisados. Um filtro foi aplicado as regras de classificação dos testes realizados utilizando as bases de dados com informações físico-químicas, e apenas as que possuíam um índice acima de 65% de acurácia foram consideradas. A partir desta seleção, foram analisadas as frequências em que cada um dos parâmetros da base de dados apareceram nas regras. Um *ranking* desta frequência foi construído como mostrado na figura 10.

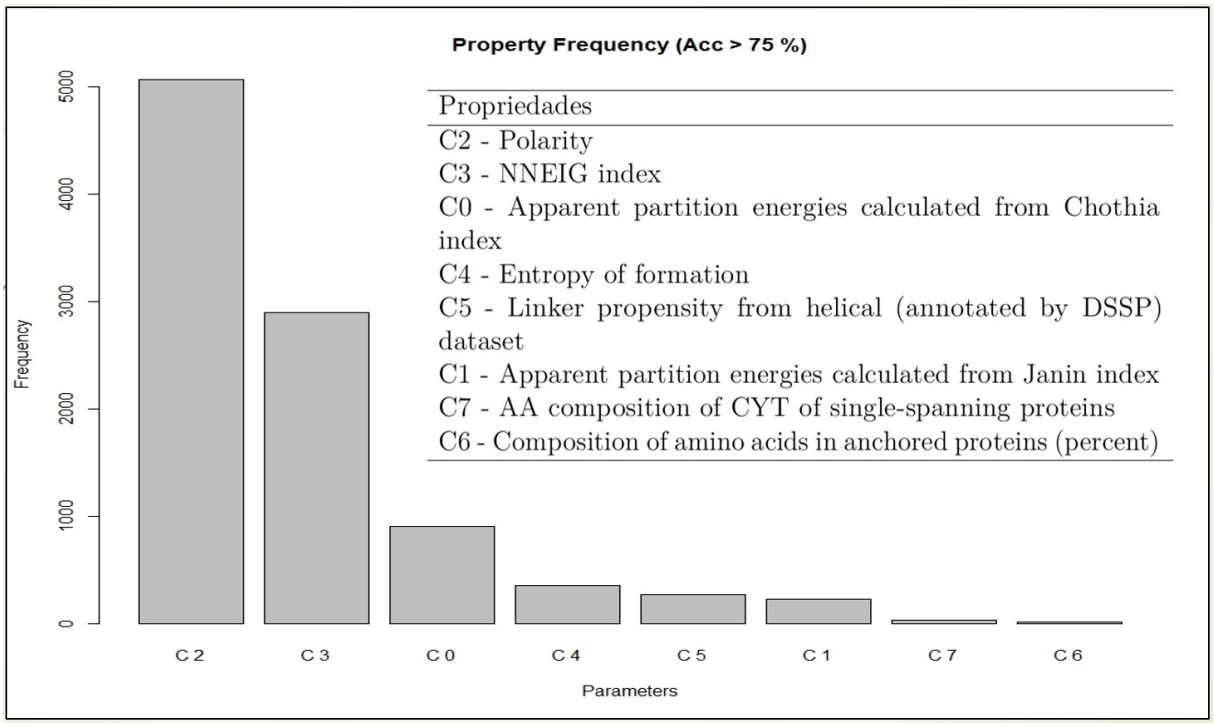

Figura 10 – Parâmetros físico-químicos: frequencia das propriedades físico-químicas extraídas das regras com valor maior que 65% de acurácia.

Como mostrado, as propriedades mais frequentes nas regras de classificação são as propriedades C2, C3, C0, C4, C5, C1, C7 e C6. Esta frequência nos leva a crer que estas propriedades são as que possuem um maior valor informativo dentro do processo de classificação e portanto mais relevantes para a investigação das regiões ubiquitinadas.

5.2.3 Base de dados Gerada

Entre os resultados obtidos deste trabalho, está o pacote denominado *Ubiquitin-Protein Integration Service* ou **UPIS**. Este pacote inclui uma série de funcionalidades de integração de dados e a base de dados contendo todas as informações necessárias para a realização deste trabalho.

A construção das funcionalidades e a constituição e modelagem da base de dados presentes no UPIS, se originou a partir da dificuldade de encontrar bases de dados atualizadas que contenham as informações necessárias para desenvolver este trabalho, além da falta de um padrão estabelecido para acessar estas informações. A base de dados comumente utilizada pelos demais trabalhos é a base Ubiprot, entretanto, haja vista que existiu uma grande dificuldade de exportar as informações la contidas, e também da falta de atualização das informações da base em questão, tornou-se inviável a utilização da mesma.

O pacote UPIS foi desenvolvido para que, além de contribuir na execução deste trabalho, pudesse propor uma base de dados com uma ferramenta que facilite o acesso a estas informações, alem de construir um padrão. Entre os principais objetivos desta base está a proposta de facilitar primeiramente o acesso as informações nela contidas. O usuário interessado pode criar padrões de exportação para o formato e *layout* de seu interesse, ou seja, o mesmo pode escolher quais informações quer obter, aplicando um filtro para autor que publicou a base de dados, organismo, que tipo de arquivo de saída o mesmo deseja, entre outros.

Outro ponto importante, está na atualização das informações contidas na base. Esta ferramenta foi construída visando a possibilidade de escolher quais dados utilizar, incluindo inclusive a possibilidade de se escolher o trabalho que originou os dados, consequentemente favorecendo os próprios autores, dando-lhes possibilidade de ter seu trabalho citado a partir de sua base de dados.

Por isso, o próprio autor pode submeter seus dados para a base a partir de alguns requisitos como a publicação. Esta funcionalidade foi construída visando resolver duas problemáticas: a primeira é a de atualização da base de dados, oferecendo aos autores algo de interesse, o segundo problema é a de falta de padronização, pois a ferramenta também solicitará ao autor que os dados sejam submetidos seguindo um padrão pré-estabelecido, solicitando obrigatoriamente por exemplo, a posição da lisina ubiquitinada e o percentual de certeza da ubiquitinação.

Entre as funcionalidades presentes, está a integração com a base de dados de propriedades físico-químicas AAIndex. Todas as informações e valores necessários para o cálculo dessas propriedades a partir da sequencia dos aminoácidos estão na base de dados, alem da ferramenta que calcula automaticamente as propriedades selecionadas a partir da sequencia.

Este pacote também possui a capacidade de ler alguns dos padrões mais utilizados na área de biologia molecular como o formato FASTA. Neste trabalho, esta ferramenta foi utilizada principalmente para integrar as informações de proteínas obtidas a partir da base de dados Uniprot á base de dados UPIS.

Foram escritos *scripts* em python para integrar as bases de dados que continham as informações sobre as regiões de ubiquitinação de outros trabalhos. Para cada uma dessas bases de dados, foi construído um script para integrar as informações oriundas de formatos diversos, entre eles: CSV, TXT e FASTA. Estes *scripts* também normalizaram as informações no padrão estabelecido na base de dados do UPIS.

O trabalho que foi efetuado após a integração de todas estas informações foi de extrair as sequencias contidas na base de dados. Para isto, foram estabelecidas mais funcionalidades, a primeira de selecionar a posições das lisinas ubiquitinadas e não ubiquitinadas da base, depois determinar o tamanho w da janela de aminoácidos próximos a ser considerado. Após exportar estas sequencias, as outras duas funcionalidades já geravam os valores das informações desejadas, sejam estas as informações físico-químicas, sejam as de composição do aminoácido.

6 Conclusão

Este trabalho foi iniciado com a proposta de realizar um estudo para avaliar o desempenho do NSGA-II como um método de classificação e posteriormente gerar hipóteses a respeito da relação entre as diversas informações disponíveis sobre as propriedades físico químicas e como as mesmas se comportavam de acordo com a classe que lhe era atribuída. Inicialmente, trabalhamos apenas com a base de dados gerada pelo trabalho de Tung e Ho(2008), que era composta por 302 amostras de um único organismo (*Saccharomyces cerevisiae*), o que posteriormente se mostrou pequena comparado com as diversas possibilidades disponíveis na área. Após uma extensa investigação, diversas bases de dados foram encontradas e sugeridas, entretanto as mesmas se assemelharam muito, sendo muitas vezes apenas uma extensão da base de Tung e Ho (2008). Quando não eram semelhantes, as informações eram de difícil acesso ou mesmo estavam desatualizadas.

A partir daí, deu-se início a um processo de estudo e obtenção de informações para que se pudesse montar uma base de dados melhor estruturada, que oferecesse mais de um organismo e uma quantidade de dados maior do que a já obtida. Alguns testes e estudos foram realizados e por fim uma metodologia nova de trabalho e uma estrutura de base de dados surgiu.A esta base de dados deu-se o nome de UPIS.

O principal objetivo a ser alcançado com a construção desta base, foi obter uma ferramenta que oferecesse uma padronização de dados e possuísse uma robusta estrutura a fim de facilitar o acesso as informações. Assim seria possível resolver um dos maiores problemas no desenvolvimento do trabalho, a obtenção de dados e o esforço necessário para obter as informações, o que inviabilizava e dificultava o trabalho de obtenção das informações. Buscou-se na construção da base de dados, resolver esses problemas, oferecendo ferramentas para atualização de dados e uma maneira de carregar os dados de forma a se evitar a necessidade de trabalho manual na preparação dos dados antes de utiliza-los. A ideia é que ao se carregar as informações para o banco de dados, os mesmos já estivessem prontos para ser utilizados.

A base de dados UPIS e as ferramentas de apoio permitiram um aumento substancial da produtividade deste trabalho, pois entre as integrações realizadas, já estavam incluídas as bases de dados de proteínas *UniprotKB*, informações de propriedades físico químicas *AAIndex* e principalmente as informações de ubiquitinação. O desafio maior porém foi a carga das informações de ubiquitinação, pois cada autor disponibiliza as informações de seus trabalhos em um formato, não havendo uma padronização. Após integrar as informações porém, o UPIS permitiu organizar todas as informações e assim facilitar o uso das mesmas.

A ideia é que futuramente, após a estrutura da base alcançar mais maturidade, a mesma seja disponibilizada junto com suas ferramentas, para que assim outros interessados tenham um acesso facilitado as informações de trabalhos sobre a ubiquitina.

Durante a criação do UPIS, foram obtidas informações de mais 3 outros organismos(*Arabidopsis thaliana*, *Mus musculus* e *Homo sapiens*), alem de obter novos dados do organismo já existente. A quantidade de regiões ubiquitinadas aumentou consideravelmente, tendo ao final do processo, um número muito superior de regiões ubiquitinadas e possíveis regiões de ubiquitinação originalmente obtidas.

Na metodologia inicial, também era previsto que fosse trabalhado com apenas um tipo de propriedade: as propriedades físico químicas. Mas com a criação da estrutura de dados, foi possível aumentar a quantidade de tipos de informação utilizadas no trabalho devido a facilidade de se fazer isso com uma estrutura bem organizada. Assim, também geramos informações sobre a composição de aminoácidos.

Outros dois algoritmos também foram inseridos no trabalho, estes foram utilizados para se ter uma base de comparação do desempenho do NSGA-II com a nova base. Estes dois algoritmos são o KNN e o *Random Forest*.

O motivo de escolher estes algoritmos, se deve ao fato de que não existe uma grande variedade de algoritmos sendo utilizados nesta área de investigação, e a ideia por traz da escolha, era expandir as possibilidades de algoritmos sendo utilizados e assim possibilitar a geração de novas hipóteses a respeito da ubiquitinação. A maioria dos trabalho nesta área se limita a utilizar em geral o SVM, e em apenas alguns trabalhos houve variedade na utilização de algoritmos como *Random Forest* ou KNN.

Se tratando de possibilitar novas hipóteses, com as regras geradas pelas execuções do NSGA-II, pode ser possível por exemplo, gerar matrizes de correlação entre as regras, categorizar os tipos de ubiquitina a partir da relação entre regiões classificadas com as regras geradas. Pode-se avaliar por exemplo, a relação de cada uma das regras com o tipo de ubiquitina. Determinada regra gerada pelo NSGA-II pode estar apenas conseguindo classificar monoubiquitinas ou poliubiquitinas, levantar hipóteses sobre a relação de cada regra, juntar regras em uma única árvore de decisão, entre outras alternativas que podem fornecer uma melhor ideia do funcionamento do mecanismo da ubiquitinação.

Neste trabalho, o *ranking* gerado a partir da frequência das propriedades nas regras de classificação, é uma hipótese a respeito da influência de cada propriedade dentro do contexto dos fragmentos das proteínas. Além do mais, os resultados alcançados pelas classificações realizados pelos algoritmos alcançaram bons resultados, mostrando assim que os objetivos deste trabalho foram alcançados, pois os resultados são comparáveis com outros trabalhos na área, mostrando que estes algoritmos podem ser utilizados como alternativas, alem de gerar novas hipóteses a respeito do funcionamento da ubiquitinação

e suas características.

A metodologia de construção da base de dados também se mostrou eficiente, pois no final foi gerado um resultado que serviu a seu propósito e permitiu agregar mais informações como as informações de composição de aminoácidos, devido a facilidade de se trabalhar com a base.

Em trabalhos futuros, pode-se explorar muito mais as propriedades informativas utilizadas neste trabalho. Como exemplo, pode-se juntar os dois tipos de informação (Físico química e de composição), e utilizar todas estas propriedades ao mesmo tempo na classificação. Pode-se incluir também, métodos diferenciados que já vem sendo utilizados em outros trabalhos como a composição de aminoácidos em pares, onde a frequência dos pares de aminoácidos também são consideradas.

O NSGA-II também pode ser modificado para tentar obter outros tipos de informação, existem operadores genéticos que podem funcionar para analisar as posições em que os aminoácidos estão ocupando na janela de proximidade da lisina central. Este método fornecerá hipóteses que não são comuns aos demais algoritmos de classificação, pois geralmente não analisam a posição dos aminoácidos, apenas analisam dados contínuos.

Neste trabalho pôde-se concluir que os algoritmos de aprendizado de máquina podem e devem ser utilizados como uma alternativa para identificar novas regiões de ubiquitinação, além de direcionar o esforço gasto em experimentação *in vivo* das regiões de ubiquitinação. Esse tipo de ferramenta se apresenta como uma alternativa facilitadora da experimentação *in vivo* que pode auxiliar no aumento da produtividade e direcionar esforços da experimentação, alcançando assim mais eficiência.

Também foi possível concluir que as técnicas aqui apresentadas, em especial o NSGA-II e a *Random Forest* são métodos competitivos comparados aos algoritmos normalmente utilizados em ferramentas de classificação de regiões de ubiquitinação.

Aqui também notou-se o mesmo fato percebido em outros trabalhos na área, a composição do aminoácido fornece informações muito mais valiosas a respeito do padrão encontrado em regiões de ubiquitinação, e este tipo de informação ainda possui um grande potencial a ser explorado, como utilizar, por exemplo, a composição de aminoácidos em pares, o que pode gerar hipóteses muito mais assertivas a respeito da composição das regiões ubiquitináveis.

Um *ranking* de propriedades físico químicas foi gerado, o que pode ajudar no direcionamento das investigações a respeito da relação dessas propriedades com a ocorrência da ubiquitinação, e também da relação com os tipos de cadeias de ubiquitinação.

Referências

CAI, Y. et al. Prediction of lysine ubiquitination with mRMR feature selection and analysis. *Amino acids*, v. 42, n. 4, p. 1387–95, apr 2012. ISSN 1438-2199.

CHEN, X. et al. Incorporating key position and amino acid residue features to identify general and species-specific Ubiquitin conjugation sites. *Bioinformatics (Oxford, England)*, v. 29, n. 13, p. 1614–22, jul 2013. ISSN 1367-4811.

CHEN, Z. et al. Prediction of ubiquitination sites by using the composition of k-spaced amino acid pairs. *PloS one*, v. 6, n. 7, p. e22930, jan 2011. ISSN 1932-6203.

CHEN, Z. et al. hCKSAAP_UbSite: Improved prediction of human ubiquitination sites by exploiting amino acid pattern and properties. *... et Biophysica Acta (BBA)-Proteins and ...*, Elsevier B.V., v. 1834, n. 8, p. 1461–7, aug 2013. ISSN 0006-3002.

CHEN, Z. et al. Towards more accurate prediction of ubiquitination sites: a comprehensive review of current methods, tools and features. *Briefings in bioinformatics*, sep 2014. ISSN 1477-4054.

CLAGUE, M. J.; HERIDE, C.; URBÉ, S. *The demographics of the ubiquitin system*. 2015.

CONSORTIUM, T. G. O. UniProt: a hub for protein information. *Nucleic Acids Research*, v. 43, n. Database issue, p. D204–12, 2014. ISSN 0305-1048.

DEB, K. Multi-Objective Optimization Using Evolutionary Algorithms : An Introduction. p. 1–24, 2011.

GROU, C. P. et al. The de novo synthesis of ubiquitin: identification of deubiquitinases acting on ubiquitin precursors. *Scientific reports*, v. 5, p. 12836, 2015. ISSN 2045-2322. Disponível em: <http://www.nature.com/srep/2015/150803/srep12836/full/srep12836.html>.

HAGLUND, K.; DIKIC, I. Ubiquitylation and cell signaling. *The EMBO journal*, v. 24, n. 19, p. 3353–9, 2005. ISSN 0261-4189.

HALL, M. et al. The WEKA Data Mining Software : An Update. *SIGKDD Explorations*, v. 11, n. 1, p. 10–18, 2009. ISSN 19310145.

HAN, J.; KAMBER, M.; PEI, J. *Data mining: concepts and techniques*. [S.l.: s.n.], 2012. ISBN 9780123814791.

HERRMANN, J.; LERMAN, L. O.; LERMAN, A. *Ubiquitin and ubiquitin-like proteins in protein regulation*. 2007. 1276–1291 p.

KAWASHIMA, S.; POKAROWSKI, P. AAindex: amino acid index database, progress report 2008. *Nucleic acids ...*, 2008.

KIM, D.-Y. et al. Advanced proteomic analyses yield a deep catalog of ubiquitylation targets in Arabidopsis. *The Plant cell*, v. 25, n. May, p. 1523–40, 2013. ISSN 1532-298X.

LEE, T.-Y. et al. Incorporating distant sequence features and radial basis function networks to identify ubiquitin conjugation sites. *PloS one*, v. 6, n. 3, p. e17331, jan 2011. ISSN 1932-6203.

LIU, J. et al. Targeting the ubiquitin pathway for cancer treatment. *Biochimica et Biophysica Acta - Reviews on Cancer*, Elsevier B.V., v. 1855, n. 1, p. 50–60, 2015. ISSN 18792561. Disponível em: <http://dx.doi.org/10.1016/j.bbcan.2014.11.005>.

MERTINS, P. et al. Integrated proteomic analysis of post-translational modifications by serial enrichment. *Nature methods*, v. 10, n. 7, p. 634–637, 2013. ISSN 1548-7105.

NALEPA, G.; ROLFE, M.; HARPER, J. W. Drug discovery in the ubiquitin-proteasome system. *Nature reviews. Drug discovery*, v. 5, n. 7, p. 596–613, 2006. ISSN 1474-1776.

OKATSU, K. et al. Phosphorylated ubiquitin chain is the genuine Parkin receptor. *Journal of Cell Biology*, v. 209, n. 1, p. 111–128, 2015. ISSN 15408140.

PENG, J.; SCHWARTZ, D.; ELIAS, J. A proteomics approach to understanding protein ubiquitination. *Nature ...*, v. 21, n. 8, p. 921–6, aug 2003. ISSN 1087-0156.

RADIVOJAC, P.; VACIC, V. Identification, analysis, and prediction of protein ubiquitination sites. *Proteins: Structure, ...*, v. 78, n. 2, p. 365–80, feb 2010. ISSN 1097-0134.

STARITA, L. M. et al. Sites of ubiquitin attachment in Saccharomyces cerevisiae. *Proteomics*, v. 12, n. 2, p. 236–40, jan 2012. ISSN 1615-9861.

TRUNTZER, C. et al. Comparison of classification methods that combine clinical data and high-dimensional mass spectrometry data. *BMC bioinformatics*, v. 15, n. 1, p. 385, nov 2014. ISSN 1471-2105.

TUCKOW, A. P. et al. Identification of ubiquitin-modified lysine residues and novel phosphorylation sites on eukaryotic initiation factor 2B epsilon. *Biochemical and biophysical research communications*, Elsevier Inc., v. 436, n. 1, p. 41–6, jun 2013. ISSN 1090-2104.

TUNG, C.-W.; HO, S.-Y. POPI: predicting immunogenicity of MHC class I binding peptides by mining informative physicochemical properties. *Bioinformatics (Oxford, England)*, v. 23, n. 8, p. 942–9, apr 2007. ISSN 1367-4811.

TUNG, C.-W.; HO, S.-Y. Computational identification of ubiquitylation sites from protein sequences. *BMC bioinformatics*, v. 9, p. 310, jan 2008. ISSN 1471-2105.

UDESHI, N. D. et al. Large-scale identification of ubiquitination sites by mass spectrometry. *Nature protocols*, Nature Publishing Group, v. 8, n. 10, p. 1950–60, oct 2013. ISSN 1750-2799.

UDESHI, N. D. et al. Refined preparation and use of anti-diglycine remnant (K-ϵ-GG) antibody enables routine quantification of 10,000s of ubiquitination sites in single proteomics experiments. *Molecular & cellular proteomics : MCP*, v. 12, p. 825–31, 2013. ISSN 1535-9484.

WAGNER, S. a. et al. Proteomic analyses reveal divergent ubiquitylation site patterns in murinetissues. *Molecular & Cellular Proteomics*, p. 1578–1585, 2012. ISSN 1535-9476.

WALSH, I.; Di Domenico, T.; TOSATTO, S. C. E. RUBI: rapid proteomic-scale prediction of lysine ubiquitination and factors influencing predictor performance. *Amino acids*, v. 46, n. 4, p. 853–62, apr 2014. ISSN 1438-2199.

WITTEN, I.; FRANK, E. *Data Mining: Practical machine learning tools and techniques*. [S.l.: s.n.], 2011. ISBN 9780123748560.

ZANGOOEI, M. H.; HABIBI, J.; ALIZADEHSANI, R. Disease Diagnosis with a hybrid method SVR using NSGA-II. *Neurocomputing*, Elsevier, v. 136, p. 14–29, 2014. ISSN 0925-2312.