PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS
Bacharelado em Ciência da Computação

Thiago Assis de Oliveira Rodrigues

PREDIÇÃO DE FUNÇÃO DE PROTEÍNAS ATRAVÉS DA EXTRAÇÃO DE CARACTERÍSTICAS DA ESTRUTURA PRIMÁRIA E SECUNDÁRIA

Belo Horizonte
2013

Thiago Assis de Oliveira Rodrigues

PREDIÇÃO DE FUNÇÃO DE PROTEÍNAS ATRAVÉS DA EXTRAÇÃO DE CARACTERÍSTICAS DA ESTRUTURA PRIMÁRIA E SECUNDÁRIA

Monografia apresentada ao programa de Bacharelado em Ciência da Computação da Pontifícia Universidade Católica de Minas Gerais, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

Orientador: Cristiane Neri Nobre - PUC Minas

Belo Horizonte
2013

Thiago Assis de Oliveira Rodrigues

PREDIÇÃO DE FUNÇÃO DE PROTEÍNAS ATRAVÉS DA EXTRAÇÃO DE CARACTERÍSTICAS DA ESTRUTURA PRIMÁRIA E SECUNDÁRIA

Monografia apresentada ao programa de Bacharelado em Ciência da Computação da Pontifícia Universidade Católica de Minas Gerais, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

Cristiane Neri Nobre - PUC Minas

João Paulo Domingos Silva - PUC Minas

Wladmir Cardoso Brandão - PUC Minas

Belo Horizonte, 25 de Novembro de 2013

AGRADECIMENTOS

Quero agradecer em primeiro lugar, a Deus, pela força, coragem e principalmente pela proteção que me deu ao longo de toda esta longa caminhada.

Agradeço à Cristiane por sua dedicação em me orientar. Por ter acreditado em mim desde o início. Gostei muito de trabalhar contigo e tenho certeza que ainda faremos muita pesquisa juntos.

Agradeço aos meus pais por todo o apoio, mesmo estando longe. Por todas as horas ao telefone me ouvindo, e sempre dando a palavra certa no momento certo.

Também quero agradecer à tia Giovanina e ao Tio Coelho que me deram todo apoio e condições para que a minha graduação fosse possível.

Agradeço ao Pedro Henrique e ao Caio Eduardo Ribeiro pela ajuda no início do trabalho. Vocês colaboraram muito para que meus objetivos fossem alcançados.

Agradeço enormemente a todos os meus amigos, que ouviram minhas reclamações, que tiveram paciência, que me deram apoio e permitiram que eu terminasse este trabalho. Que acreditaram em mim e que comemoraram cada pequena vitória comigo. Um agradecimento especial ao Róbson Silva por todas as suas orações e ao Kelvin Vianini por ter aguentado as minhas reclamações ao longo deste ano e não ter me deixado desistir em momento algum.

Creio que aprendi muito com todos e que o maior resultado não são apenas páginas, mas sim o fato de hoje eu ser uma pessoa melhor por ter compartilhado um tempo de vida com todos vocês.

RESUMO

As proteínas desempenham uma grande variedade de funções biológicas. O conhecimento de suas características a partir de suas estruturas pode ajudar no entendimento da função desempenhada, e podem ser utilizadas para a descoberta de novas drogas, que podem ser aplicadas em diversos setores, como por exemplo a saúde e a indústria de bioquímicos. Com a finalização do projeto Genoma, o número de novas proteínas descobertas tem crescido muito, mas devido ao alto custo e da demora dos processos de descoberta de função de proteínas, apenas uma pequena parcela das mesmas tem sua função conhecida. Este trabalho apresenta uma metodologia para predição de função de proteínas, através da extração de características de suas estruturas, presentes no banco de dados Sting_DB, da utilização da Transformada Discreta do Cosseno, da codificação da estrutura primária e da utilização de Máquinas de Vetores de Suporte, que são um tipo de aprendizado de máquina. Os resultados foram comparados com outros trabalhos da literatura, e mostraram um aumento de 5% da taxa de precisão, em relação ao trabalho mais recente, demonstrando que esta metodologia foi satisfatória principalmente devido ao fato do grande desbalancamento entre o número de instâncias de cada classe de proteína utilizada. Os valores médios obtidos para as métricas de precisão, sensibilidade, acurácia e especificidade foram respectivamente de 75%, 67%, 71% e 76%.

Palavras-chave: Predição de Função de Proteína. SVM. Transformada Discreta do Cosseno. Codificação de Proteínas. Características de Proteínas. Validação Cruzada.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

AM – Aprendizado de Máquina

CA – Carbono Alfa

CB – Carbono Beta

DNA – Ácido Desoxirribonucleico

EC – *Enzyme Commission*

FN – Falso Negativo

FP – Falso Positivo

GPL – *General Public License*

IUBMB – *International Union of Biochemistry and Molecular Biology*

JPD – *Java Protein Dossier*

LHA – *Last Heavy Atom*

PDB – *Protein Data Bank*

PCA – *Principal Component Analysis*

RCSB – *Research Collaboratory for Structural Bioinformatics*

RNA – Ácido Ribonucleico

SVM – *Support Vector Machine*

TDC – Transformada Discreta do Cosseno

VN – Verdadeiro Negativo

VP – Verdadeiro Positivo

WEKA – *Waikato Environment for Knowledge Analysis*

SUMÁRIO

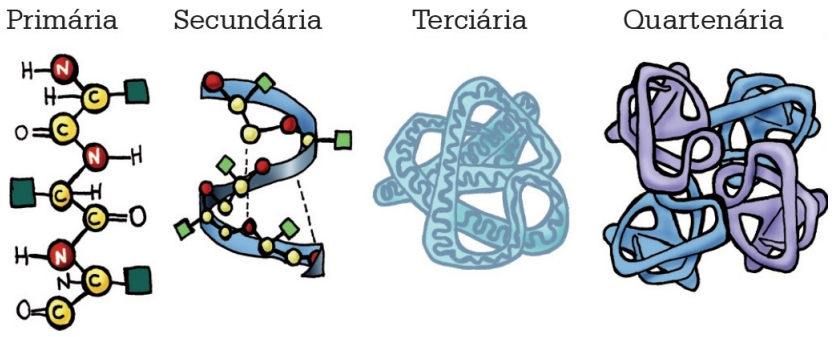
1 INTRODUÇÃO

A bioinformática tem por objetivo o estudo e aplicação de técnicas computacionais a diversas áreas da biologia (PROSDOCIMI et al., 2002). Nesse contexto, a computação pode ser aplicada na resolução de uma série de problemas, tais como: comparação de sequências (DNA, RNA e proteínas), montagem de fragmentos, reconhecimento de genes, identificação e análise da expressão de genes, reconstrução de árvores filogenéticas e determinação da estrutura e função de proteínas (BITTENCOURT, 2005).

Dentre esses problemas, a predição teórica da função de proteínas é um grande desafio da bioinformática. O estudo relativo a esse assunto vem caracterizando uma nova fase para as pesquisas genéticas, denominada proteômica (BITTENCOURT, 2005), que envolve a identificação de todas as proteínas expressas pelo genoma, bem como a determinação de suas funções fisiológicas e patológicas.

As proteínas são as macromoléculas utilizadas como matéria-prima e são componentes funcionais das células, sendo a segunda maior fração em peso, perdendo apenas para a água. Elas são muito diversificadas e, por isso, apresentam várias formas de classificação. Em geral, podemos classificá-las de acordo com suas quatro etapas estruturais: primária, secundária, terciária e quaternária, como pode ser visto na Figura 1.

Figura 1 – Formas Estruturais das Proteínas



Fonte: (ENSINO, 2006).

As funções das proteínas estão diretamente relacionadas à conformação estrutural dada pela composição química de cada um dos vinte aminoácidos e suas respectivas conformações no espaço, a ponto de termos dificuldade de generalizar esse conceito devido à alta sensibilidade ao contexto. Em (PANDEY; KUMAR; STEINBACH, 2006) temos o conceito de que a função da pro-

teína é todo o tipo de atividade que a proteína se envolve, seja celular, molecular ou fisiológica. Alguns exemplos onde as proteínas apresentam importantes funções no organismo são a constituição de órgãos (proteínas estruturais), a catalisação de reações bioquímicas necessárias para o metabolismo (enzimas) e a manutenção das atividades celulares (proteínas da transmembrana). Assim, elas são as mais essenciais e versáteis macromoléculas da vida e o conhecimento de suas funções é crucial para o desenvolvimento de novas drogas, melhores colheitas e até mesmo no desenvolvimento de produtos bioquímicos sintéticos, como os biocombustíveis.

Hoje, existem milhares de proteínas que ainda não têm sua função conhecida (ALVAREZ; YAN, 2010). Com isso, várias metodologias diferentes para a predição de função de proteínas têm sido propostas. Esses métodos utilizam diversas informações sobre as proteínas tais como parâmetros físico-químicos, parâmetros geométricos, superfície de contato (DOBSON; DOIG, 2005), hidrofobicidade, estrutura secundária e solubilidade (PIAO et al., 2008) e informações das sequências de aminoácidos juntamente com as estruturas primárias e secundárias (WATSON; LASKOWSKI; THORNTON, 2005).

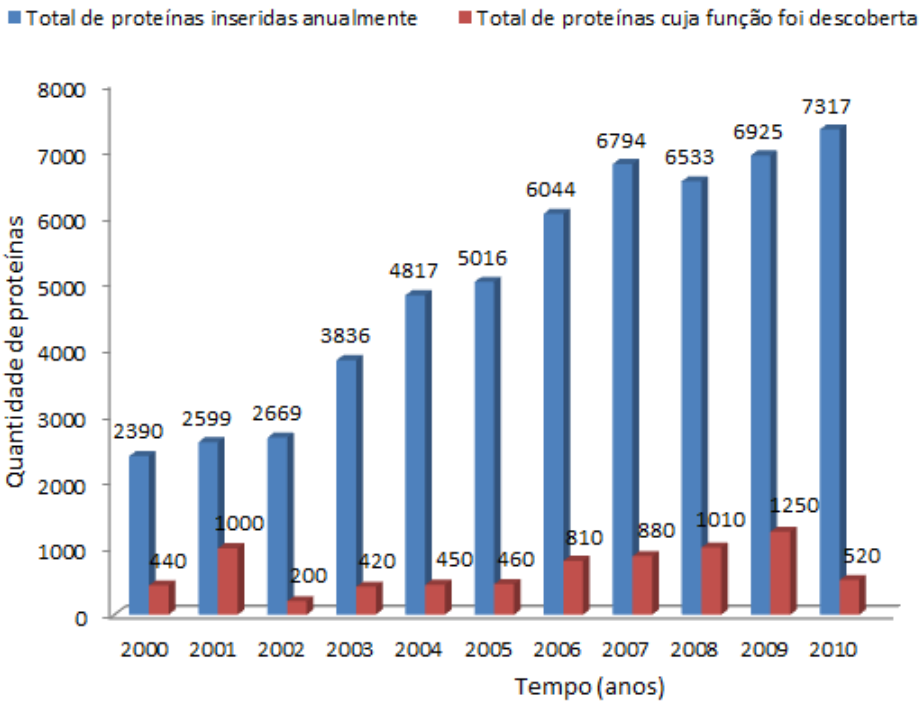
Além disso, são utilizadas técnicas de aprendizado de máquina (AM), por apresentarem capacidade de aprendizado automático a partir de grandes volumes de dados e produção de hipóteses úteis. Essa técnica tem sido consideravelmente explorada na predição automática de funções e estruturas de proteínas como pode ser visto em (BITTENCOURT, 2005; BORRO et al., 2006; RESENDE et al., 2012; OLIVEIRA, 2011; RODRIGUES et al., 2012).

1.1 Justificativa

Com a finalização do Projeto Genoma, o número de proteínas conhecidas tem crescido muito. A Figura 2 mostra que a quantidade de proteínas adicionadas ao *Protein Data Bank* (PDB) anualmente, gira em torno de 7 a 8 mil. Atualmente já passam de 80.000 proteínas inseridas neste banco. Porém, apenas uma pequena parcela destas proteínas possuem sua função conhecida. No trabalho de Nadzirin e Firdaus-Raih (2012), foi feito um levantamento da quantidade de proteínas cuja função foi descoberta ao longo dos últimos anos, e este levantamento também é apresentado na Figura 2.

Esse grande crescimento do número de proteínas gera um grande desafio do ponto de vista laboratorial e computacional. No que diz respeito aos trabalhos laboratoriais, há méto-

Figura 2 – Comparação entre crescimento anual das proteínas no PDB e quantidade de proteínas com função descoberta ao longo dos últimos anos



Fonte: Adaptado de (FILETO et al., 2006) e (NADZIRIN; FIRDAUS-RAIH, 2012)

dos tradicionais de descoberta de função, como a Cristalografia por Difração de Raios-X e a Ressonância Nuclear Magnética (ALBERTS et al., 2002). Porém, essas técnicas exigem grande esforço humano e experimental, gerando altos custos e resultados a longo prazo. Do ponto de vista computacional, um dos maiores desafios é selecionar as características realmente relevantes para que ocorra uma previsão de função de proteínas com exatidão, diminuindo assim a necessidade de testes laboratoriais.

1.2 Problema

Determinar a sequência de uma proteína é relativamente mais fácil do que determinar a sua função, o que leva a uma grande diferença entre o número de sequências e o número de proteínas com suas funções conhecidas. Logo deseja-se saber, qual a forma mais viável de se extrair características de proteínas conhecidas, utilizando métodos computacionais, para classificarmos proteínas com função desconhecida.

1.3 Objetivos

1.3.1 Objetivos Gerais

Neste trabalho, será apresentada uma metodologia de extração de características de proteínas para predição de sua função, que é dada de acordo com a classe a que pertencem, a partir do banco de dados público PDB (BERMAN et al., 2000). A função se restringirá aos aspectos estruturais da proteína, no que diz respeito às suas estruturas primárias e secundárias.

É importante ressaltar que nos limitaremos a utilizar um subgrupo das proteínas, citado por (DOBSON; DOIG, 2005), que utilizou apenas as seguintes superfamílias: Oxidoredutases, Transferases, Hidrolases, Liases, Isomerases e Ligases.

1.3.2 Objetivos Específicos

- Selecionar um conjunto de características que possam ser extraídas das proteínas já catalogadas.
- Selecionar uma codificação para a estrutura primária das proteínas.
- Encontrar o melhor valor para a Transformada Discreta do Cosseno, para que se obtenha o maior número possível de características relevantes para cada proteína.
- Classificar proteínas conhecidas nas seis superfamílias de proteínas utilizadas na literatura.

1.4 Organização do Trabalho

Este trabalho foi organizado da seguinte maneira. O Capítulo 2 apresenta o referencial teórico, mostrando o conceito de proteínas, suas estruturas e a revisão bibliográfica. O Capítulo 3 apresenta a metodologia utilizada. O Capítulo 4 apresenta os resultados dos testes e finalmente o Capítulo 5 apresenta as conclusões e trabalhos futuros.

2 REVISÃO DA LITERATURA

Este capítulo descreve alguns aspectos importantes para o entendimento completo do trabalho. A Seção 2.1 trata da importância das proteínas. A Seção 2.2 explica detalhadamente as quatro formas estruturais presentes nas proteínas. A Seção 2.3 apresenta as principais características das seis classes utilizadas neste trabalho. A Seção 2.4 descreve brevemente os bancos de dados utilizados. A Seção 2.5 introduz o conceito de aprendizado de máquina mostrando as principais técnicas de classificação e descrevendo a técnica escolhida. A Seção 2.6 trata do método de amostragem utilizado. E finalmente a Seção 2.7 apresenta os principais trabalhos relacionados.

2.1 Proteínas

As proteínas são os compostos orgânicos mais comuns em um organismo, os mais abundantes depois da água e também os de maior variedade molecular (ENSINO, 2006). Estão presentes em todas as estruturas celulares, desde a membrana até o núcleo, compondo as substâncias intercelulares, hormônios, anticorpos, dentre outros.

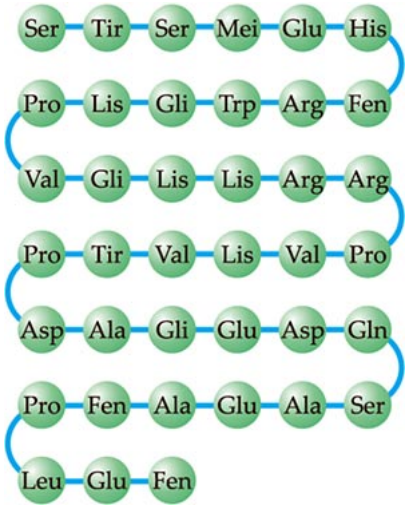
Para obter o completo entendimento das proteínas, devemos entender o conceito dos peptídeos, que estão relacionados à base estrutural das proteínas. Os peptídeos são resultantes do processamento de proteínas, e podem possuir dois ou mais aminoácidos na sua constituição. Chamamos de ligação peptídica a união química que ocorre entre duas moléculas de aminoácidos, a carboxila e a amina, liberando uma molécula de água. O que resta de cada aminoácido designa-se por resíduo de aminoácido, sendo que a cadeia polipeptídica contém de algumas dezenas a várias centenas desses resíduos.

2.2 Estruturas Protéicas

2.2.1 Estrutura Primária

A estrutura primária de uma proteína é simplesmente a sequência linear dos aminoácidos que a compõem, como pode ser visto na Figura 3. Cada proteína apresenta a sua sequência específica de aminoácidos, constituindo assim a sua própria identidade. As propriedades químicas de cada aminoácido determinam as propriedades das proteínas (SILVA, 1999). Durante a formação da proteína, os aminoácidos unem-se entre si por meio de ligações peptídicas que ocorrem entre o grupamento carboxila de um aminoácido com a amina do outro.

Figura 3 – Estrutura Primária de uma Proteína

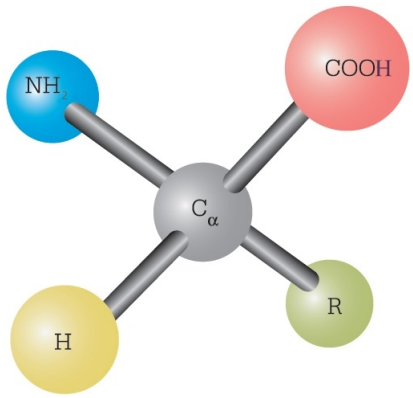


Fonte: (WEB, 2013)

Os aminoácidos são formados por um carbono central ao qual se ligam um hidrogênio (H), um grupo carboxílico ($COOH$) e um grupo amínico (NH_2) comuns a todos os aminoácidos, além de um radical R que os distingue entre si (BITTENCOURT, 2005), como pode ser visto na Figura 4. Esses radicais R podem diferir bastante em relação ao tamanho, estrutura, forma e propriedades químicas, exercendo influência em muitas características dos aminoácidos, como por exemplo, na solubilidade do aminoácido em água.

Na natureza, existem vários tipos de aminoácidos, porém apenas 20 deles estão presentes na estrutura das proteínas. A Tabela 1 apresenta uma lista desses aminoácidos e os classifica de acordo com sua origem, que pode ser do próprio organismo ou obtidos a partir de alimentos vegetais ou animais. Chamamos de aminoácidos essenciais aqueles que não podem ser pro-

Figura 4 – Aminoácido com seus agrupamentos Carboxila, Amina e radical R



Fonte: (ENSINO, 2006)

duzidos pelo corpo humano mas são essenciais para determinadas situações fisiológicas, e de naturais ou não-essenciais, os que podem ser sintetizados pelo corpo humano (ENSINO, 2006).

Tabela 1 – Nomenclatura e simbologia dos 20 aminoácidos encontrados nas proteínas

Classificação Nutricional	Nome	Símbolos	
Naturais	Arginina	Arg	R
	Cisteína	Cys	C
	Glicina	Gly	G
	Glutamina	Gln	Q
	Prolina	Pro	P
	Tirosina	Tyr	Y
	Alanina	Ala	A
	Serina	Ser	S
	Asparagina	Asn	N
	Histidina	His	H
	Ácido Aspártico	Asp	D
Essenciais	Fenilalanina	Phe	F
	Isoleucina	Ile	I
	Leucina	Leu	L
	Lisina	Lys	K
	Metionina	Met	M
	Treonina	Thr	T
	Triptofano	Trp	W
	Valina	Val	V
	Ácido Glutâmico	Glu	E

Fonte: Adaptado de (SILVA, 1999)

2.2.2 Estrutura Secundária

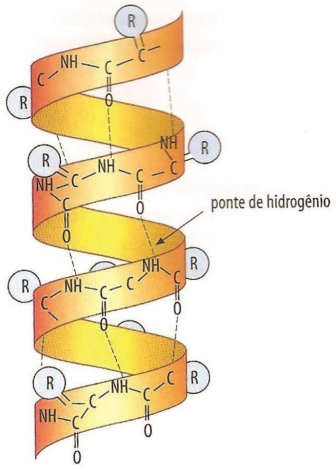
À medida que o comprimento das cadeias polipeptídicas vai aumentando e em função das condições físico-químicas do meio, cria-se a estrutura secundária, que diz respeito à disposição espacial e repetitiva entre os aminoácidos mais próximos entre si na estrutura primária. Os dois padrões de estrutura secundária mais comuns são as alfa-hélices e as folhas beta (SILVA, 1999).

2.2.2.1 Alfa-Hélices

Uma alfa-hélice é uma estrutura semelhante a um bastão, onde a cadeia peptídica principal firmemente helicoizada, forma a parte interna do bastão, e as cadeias laterais se projetam para fora em uma disposição helicoidal. Além disso, o sentido de giro de uma alfa-hélice pode ser para a direita (hélice dextrosa) ou para a esquerda (hélice sinistrosa), sendo que as alfa-hélices encontradas em proteínas são dextrosas (OLIVEIRA, 2011).

A estrutura é estabilizada por pontes de hidrogênio (atrações elétricas entre átomos com carga positiva e um átomo com carga negativa), entre os grupamentos amínicos e carboxílicos da cadeia principal, como pode ser visto na Figura 5. Como exemplo de proteína que está em alfa-hélice temos a ferritina, que auxilia no armazenamento do ferro e possui 75% dos aminoácidos em alfa-hélices.

Figura 5 – Exemplo de uma estrutura do tipo Alfa-Hélice



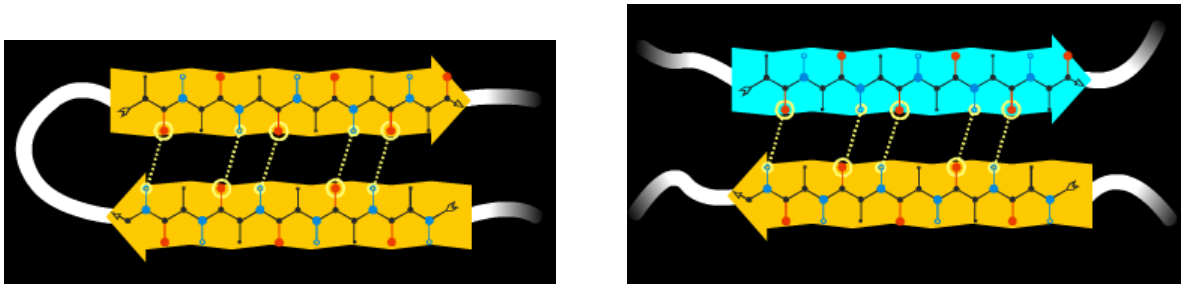
Fonte: (LINHARES; GEWANDSZNAJDER, 2006)

2.2.2.2 Folhas Beta

A estrutura de uma folha beta é formada por uma cadeia polipeptídica e é quase totalmente distendida. A folha beta é estabilizada por pontes de hidrogênio entre grupamentos amínicos (NH) e carboxílicos (CO) em fitas peptídicas diferentes, ao contrário da alfa-hélice cujas pontes de hidrogênio estão entre grupamentos do mesmo filamento (OLIVEIRA, 2011).

De acordo com a orientação relativa dos segmentos das folhas beta, podemos classificá-las em folha beta paralela, quando os segmentos estão todos orientados na mesma direção, e folha beta antiparalela, onde os seguimentos adjacentes são orientados em direções opostas, como pode ser visto na Figura 6.

Figura 6 – Representações de Folhas Beta



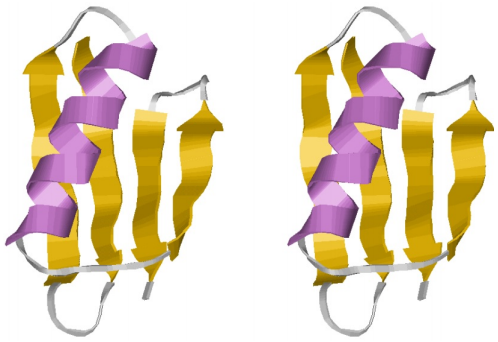
(a) Folha Beta Paralela (b) Folha Beta Antiparalela

Fonte: (QUíMICA, 2013)

2.2.3 Estrutura Terciária

A estrutura terciária de uma proteína descreve todos os aspectos do enovelamento tridimensional de um polipeptídio, que é algo complexo e sem simetria, dizendo como os aminoácidos se agrupam em uma proteína completa, de forma a gerar uma estrutura compacta onde os átomos ocupam posições específicas (BITTENCOURT, 2005). Sua estabilidade é originada por interações químicas entre as cadeias laterais dos aminoácidos como ligações covalentes, interações hidrofóbicas e eletrostáticas, dentre outras. A Figura 7 mostra a estrutura tridimensional da proteína $2GB1$, em estereoscopia, numa visualização conjunta de alfa-hélices e folhas beta (SILVA, 1999).

Figura 7 – Estrutura terciária da proteína 2GB1

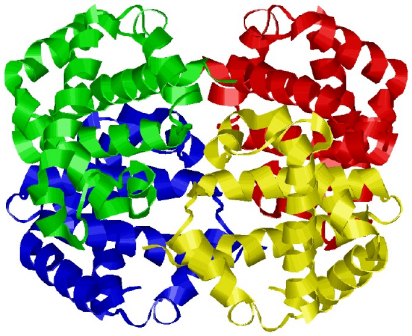


Fonte: (SILVA, 1999)

2.2.4 Estrutura Quaternária

A estrutura quaternária existe apenas quando a proteína é oligomérica, isto é, composta por duas ou mais cadeias polipeptídicas, e consiste nas suas relações e disposições relativas. Dependendo da estrutura quaternária (SILVA, 1999), uma proteína pode ser classificada como fibrosa (cadeias polipeptídicas dispostas ao longo de um eixo, formando uma estrutura alongada) ou globular (cadeias polipeptídicas muito compactas, formando uma estrutura esférica). A Figura 8 mostra a conformação da hemoglobina humana, proteína globular constituída de quatro cadeias polipeptídicas (OLIVEIRA, 2011).

Figura 8 – Estrutura quaternária da hemoglobina humana

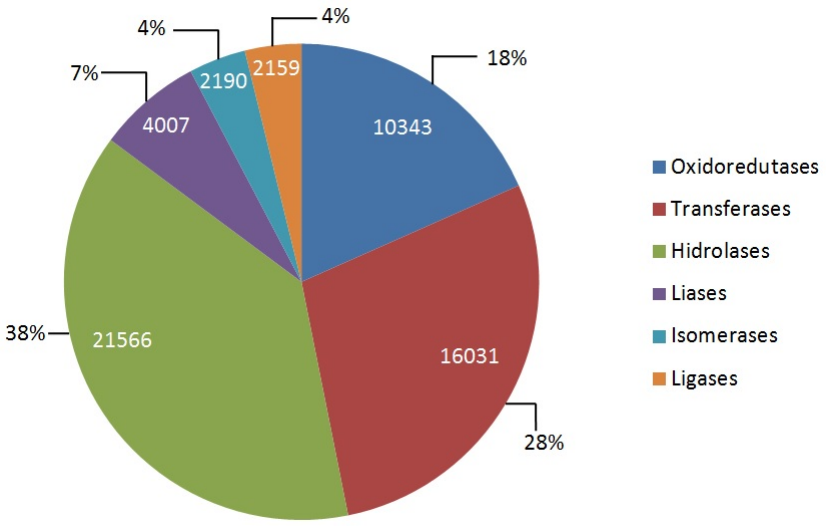


Fonte: (SILVA, 1999)

2.3 Tipos de Enzima

O *International Union of Biochemistry and Molecular Biology* (IUBMB) classificou as enzimas em seis grandes grupos ou classes de proteínas, de acordo com o tipo de reação que catalisam. Cada enzima descrita recebe um número de classificação, conhecido por *Enzyme Commision* (EC). A Figura 9 mostra a quantidade de todas as enzimas conhecidas e que pertecem às classes utilizadas neste trabalho, junto com a porcentagem que cada classe representa, seguida das principais características de cada uma das classes, sendo que os números de 1 a 6, são os valores de EC de cada enzima na seguinte ordem: Oxidoredutases, Transferases, Hidrolases, Liases, Isomerases e Ligases (MOSS, 2013). Essa quantidade foi retirada do banco de dados Sting_DB que será descrito posteriormente.

Figura 9 – Classes de enzimas e sua distribuição



Fonte: (BERMAN et al., 2000)

1. **Oxidoredutases:** também conhecidas como desidrogenases, são todas as enzimas que catalisam reações de oxidação ou de redução, em que os elétrons são transferidos de uma molécula (o redutor) a outra (o oxidante). Elas são enzimas muito importantes, sendo vitais para muitos processos metabólicos, como na respiração aeróbia e anaeróbia, e são divididas em 22 subgrupos (MOSS, 2013).

2. **Transferases:** são enzimas que catalisam a transferência de grupos funcionais entre duas moléculas. A reação básica pode ser representada por $AX + B \rightarrow A + BX$, onde A é o doador, B é o receptor e X é o grupo funcional.

3. **Hidrolases:** catalisam a hidrólise, a quebra de ligações simples através da adição de água. Há uma enorme variedade de enzimas hidrolases, como por exemplo, as enzimas digestivas.

4. **Liases:** são enzimas que adicionam ou removem elementos de água, amônia ou dióxido de carbono.

5. **Isomerases:** catalisam reações de isomerização, ou seja, a modificação de uma única molécula, sem participação de outra.

6. **Ligases:** catalisam reações nas quais há formação de uma molécula a partir da ligação de moléculas já existentes.

2.4 Base de Dados

O banco de dados público de proteínas PDB é um repositório internacional de informações sobre as estruturas 3D de moléculas biológicas, incluindo proteínas e ácidos nucléicos. Ele possui informações relativas a proteínas contendo dados como suas funções, estruturas primária, secundária, terciária e quaternária.

O PDB (BERMAN et al., 2000) foi criado em 1971 pelo *Brookhaven National Laboratory*, e originalmente continha sete estruturas. Em 1998, o *Research Collaboratory for Structural Bioinformatics* (RCSB) tornou-se responsável pela gestão do PDB. Em 2003, foi padronizado um arquivo para manter os dados estruturais macromoleculares e disponibilizá-lo publicamente. Desde então, o PDB é atualizado semanalmente com as novas moléculas encontradas através de testes laboratoriais.

2.4.1 Sting_DB

No Laboratório de Biologia Computacional da Embrapa Informática Agropecuária foi desenvolvido o Sting_DB, o maior banco de dados de características físico-químicas, estruturais e biológicas sobre estruturas protéicas, e que se encontra em sua versão BlueStar Sting.

O Sting é uma suíte de programas com ferramentas para a visualização e análise estru-

tural de proteínas. Estes programas ou módulos estão concentrados em um único pacote que visa oferecer um instrumento completo para estudos das macromoléculas, suas estruturas e as relações entre estrutura e função. Informações como posição dos resíduos de aminoácido na sequência e na estrutura, busca de padrões, identificação de vizinhança, ligações de hidrogênio, ângulos e distâncias entre átomos, são facilmente obtidas, além de dados sobre natureza e volume dos contatos atômicos inter e intracadeias, a conservação e relação entre os contatos e parâmetros funcionais.

Entre os módulos do Sting, há o *Java Protein Dossier* (JPD) (NESHICH et al., 2004), uma ferramenta de visualização que comunica muita informação através de um único gráfico, exibido em formas de cores diferentes de acordo com o valor adotado para cada parâmetro. O JPD fornece aos usuários uma vasta coleção de parâmetros físico-químicos descrevendo a estrutura da proteína, estabilidade e interações com outras macromoléculas. Ao mesmo tempo, o JPD é um passo na direção de compilar uma base de dados diversificada de descritores de estrutura e função, que podem ser usados como uma plataforma para a aquisição de novos conhecimentos. JPD pode mostrar e analisar, simultaneamente, todos os parâmetros físico-químicos de duas estruturas que tenham sido previamente superpostas, permitindo uma comparação direta de parâmetros entre estruturas similares. Além disso, ele permite que sejam salvos os dados de todos os seus parâmetros para cada aminoácido presente em uma determinada cadeia de proteína.

2.5 Aprendizado de Máquina

Segundo (MITCHELL, 1997), aprendizado de máquina (AM) é um termo que engloba um conjunto de metodologias e abordagens, com o objetivo de criar sistemas capazes de reconhecer padrões e comportamentos em dados, que representam exemplos de acontecimentos do mundo real ou experiências passadas.

A maioria dos métodos de AM adquirem experiência estritamente a partir dos dados conhecidos do problema. Assim, melhor será o desempenho dos métodos de AM quanto melhor for a qualidade dos dados (BITTENCOURT, 2005). Diversos aspectos podem influenciar no desempenho de um sistema de AM. Em bases de dados reais, esses aspectos estão relacionados com a presença de valores desconhecidos, pois distorções podem ser introduzidas no conhecimento induzido, ou com a diferença entre o número de instâncias que pertencem a diferentes

classes, pois essa diferença pode ser grande e os sistemas de AM podem ter dificuldade em aprender o conceito relacionado à classe minoritária.

Esses problemas chamaram a atenção dos pesquisadores, e diversas técnicas de pré-processamento de dados têm sido criadas para contorná-los, como pode ser visto em (PRATI; BATISTA; MONARD, 2003; BATISTA, 2003; SCHIAVONI, 2010).

2.5.1 Técnicas de Classificação

O problema de classificação consiste em, dado um conjunto de elementos divididos em classes e uma instância desse conjunto, atribuir uma classe a essa instância de acordo com as suas características que se assemelham com os demais membros da mesma classe. O AM pode ser dividido em dois tipos principais, supervisionada e não supervisionada, de acordo com os dados disponíveis para a realização do processo de indução.

- **Supervisionada:** Nesse tipo de classificação há um conjunto de instâncias, onde cada instância é formada por um conjunto de atributos de entrada e um conjunto de atributos de saída. O objetivo é construir um classificador que possa determinar corretamente a classe de novas instâncias ainda não classificadas. Como principais exemplos desse tipo de classificação podem ser citados: as Árvores de Decisão, $k$vizinhos Mais Próximos, *Naïve Bayes*, Máquinas de Vetores Suporte (LORENA; ANDRé; CARVALHO, 2007) e Redes Neurais Artificiais (DIAS, 2007).

- **Não Supervisionada:** É realizada quando, para cada instância, apenas os atributos de entrada estão disponíveis. Esse tipo de aprendizado é utilizado quando o objetivo é encontrar, em um conjunto de dados, padrões ou tendências que auxiliem o entendimento dos dados. Como exemplo de técnicas desse tipo de aprendizado podem ser citados, dentre outros: k-médias e agrupamento hierárquico.

O tipo de aprendizado abordado neste trabalho é o supervisionado, pois suas técnicas tem sido utilizadas na literatura para problemas similares ao deste trabalho, e tem obtido os melhores resultados. Para a predição de função de proteínas foi utilizada a técnica de Máquinas de Vetores de Suporte, que será descrita a seguir.

2.5.2 Máquinas de Vetores de Suporte

As Máquinas de Vetores de Suporte (SVMs, do inglês *Support Vector Machines*), foram desenvolvidas por (VAPNIK, 1999) e sua função original era para classificação binária, porém hoje já existem SVMs que trabalham com mais de duas classes simultaneamente. Elas constituem uma técnica de aprendizado baseada no fato de que, em altas dimensões do espaço de características, todos os problemas se tornam linearmente separáveis. Algumas das principais características que tornam seu uso atrativo são (LORENA; ANDRé; CARVALHO, 2007):

- **Boa capacidade de generalização:** os classificadores gerados por uma SVM em geral alcançam bons resultados de generalização. A capacidade de generalização de um classificador é medida por sua eficiência na classificação de dados que não pertençam ao conjunto utilizado em seu treinamento. Na geração de preditores por SVMs, portanto, é evitado o *overfitting*, situação na qual o preditor se torna muito especializado no conjunto de treinamento, obtendo baixo desempenho quando confrontado com novos padrões.

- **Robustez em grandes dimensões:** as SVMs são robustas diante de objetos de grandes dimensões. Comumente há a ocorrência de *overfitting* nos classificadores gerados por outros métodos inteligentes sobre esses tipos de dados.

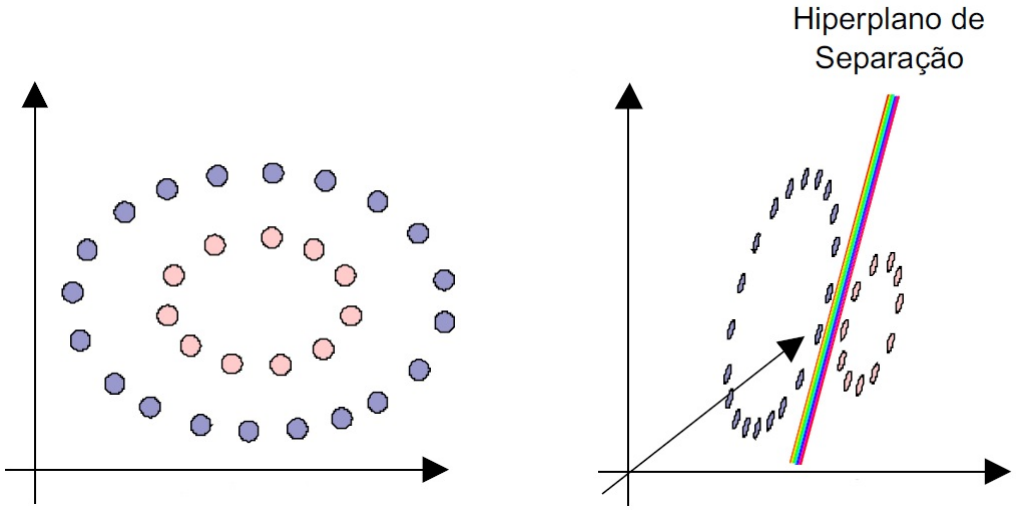
- **Teoria bem definida:** as SVMs possuem uma base teórica bem estabelecida dentro da matemática e estatística.

Entre as características citadas, o destaque das SVMs está em sua capacidade de generalização. Estes resultados foram apresentados por Vapnik e Chernovenkis através da Teoria de Aprendizado Estatístico, proposta por estes autores nas décadas de 60 e 70 (CORTES; VAPNIK, 1995). Apesar de sua teoria ser relativamente antiga, as primeiras aplicações práticas das SVMs são mais recentes e datam da década de 90.

Quando as SVMs binárias são utilizadas, elas dividem o espaço de decisão em duas classes apenas, criando-se um hiperplano que ajuda a colocar os dados que não sejam linearmente separáveis, em maiores dimensões para que possam ser separados, assim como pode ser visto na Figura 10, que fez a projeção de um espaço bidimensional em um espaço tridimensional.

Dado um conjunto de treinamento S que contém pontos de duas ou mais classes, uma SVM separa as classes através de um hiperplano determinado por certos pontos de S chamados

Figura 10 – Projeção em um espaço de maior dimensão

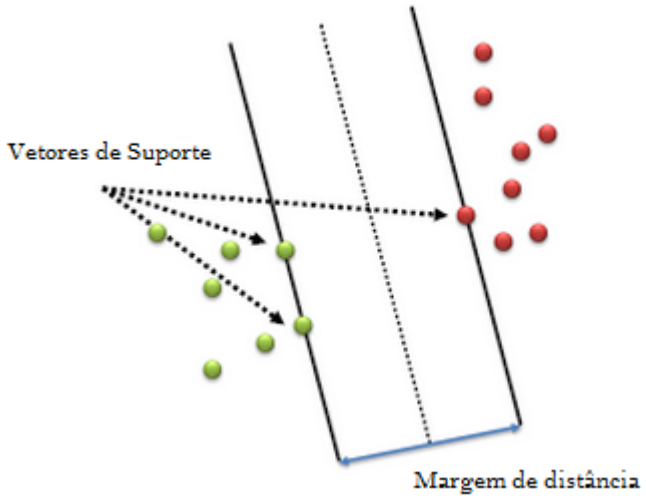


(a) Bidimensional (b) Tridimensional

Fonte: (DIAS, 2007)

de vetores de suporte. Caso as classes sejam separáveis, este hiperplano maximiza a margem de distância entre as classes, e todos os vetores que caem na mesma distância mínima a partir do hiperplano serão os vetores de suporte utilizados pelo classificador. Porém, caso as classes sejam não separáveis, os vetores de suporte serão obtidos da solução de um problema de otimização com restrições, cuja solução é um compromisso controlado por um parâmetro de regularização entre a maior margem e o menor número de erros. Estes vetores de suporte e a margem de distância estão apresentados na Figura 11.

Figura 11 – Exemplo de localização dos vetores de suporte a partir da margem de distância



Fonte: Adaptado de (GOMES, 2002)

As SVMs utilizam funções denominadas de *kernels* que são capazes de mapear um conjunto de dados em diferentes espaços, possibilitando a utilização dos hiperplanos. Existem três tipos de *kernel* para SVMs: Linear, Polinomial e RBF (*Radial-Basis Function*). Cada *kernel* (LORENA; ANDRé; CARVALHO, 2007), possui os seus respectivos parâmetros, apresentados na Tabela 2, e são altamente sensíveis, variando os seus valores de acordo com o problema a ser tratado e influenciando diretamente os resultados. Estes parâmetros tem seus melhores valores encontrados através da utilização de *scripts* disponibilizados pelos autores das SVMs, ou de outras técnicas como por exemplo algoritmos genéticos utilizados por Resende et al (2012).

Tabela 2 – Funções *Kernel* utilizadas pelas SVMs

Tipo de Kernel	Parâmetros
Linear	-
Polinomial	d
RBF	γ

Fonte: Adaptado de (LORENA; ANDRé; CARVALHO, 2007)

O *kernel* linear é o mais simples, que utiliza uma constante c opcional em seus cálculos e basicamente não usa nenhum parâmetro. O *kernel* polinomial é muito útil quando os dados estão normalizados e utiliza o parâmetro d como grau do polinômio a ser utilizado. Por fim o *kernel* RBF é um dos mais utilizados, tendo sua base na teoria Gaussiana e utiliza o parâmetro γ para determinar melhor os hiperplanos de separação das SVMs.

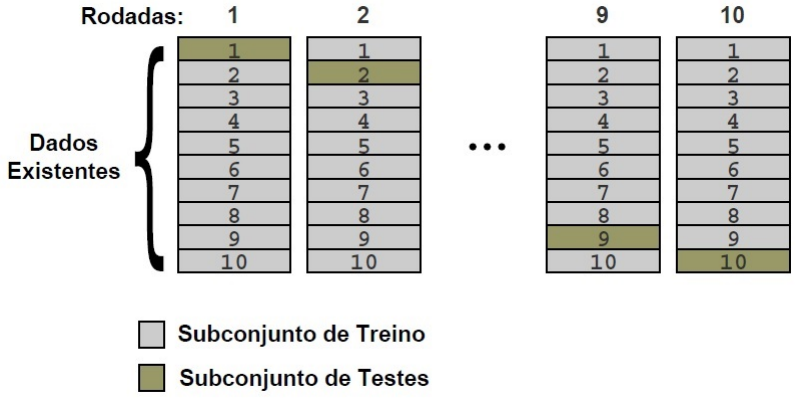
2.6 Método de Amostragem

Um método de amostragem deve ser capaz de determinar se uma hipótese vai ser suficiente para prever os dados futuros de forma eficaz. A Figura 12, ilustra o funcionamento da técnica de amostragem Validação Cruzada (*Cross-Validation*).

No *k-fold Cross-Validation* inicialmente os dados disponíveis são divididos em k subconjuntos disjuntos e de tamanhos aproximadamente iguais. Essa divisão separa a mesma quantidade de dados, de cada uma das classes utilizadas, para cada subconjunto. O valor de k pode variar livremente, mas $k = 5$ ou $k = 10$ são geralmente usados. No exemplo da Figura 12, vemos o funcionamento desta técnica para $k = 10$.

Após a divisão dos conjuntos de treino e teste, inicia-se um processo iterativo, onde em cada rodada um subconjunto é utilizado para teste, enquanto os outros são utilizados para

Figura 12 – Método de Validação Cruzada



Fonte: (DIAS, 2007)

treinamento. E isso trás uma grande vantagem, pois garante que todos os dados foram utilizados tanto para treino, quanto para testes. No final, é feita a média dos resultados encontrados.

2.7 Trabalhos Relacionados

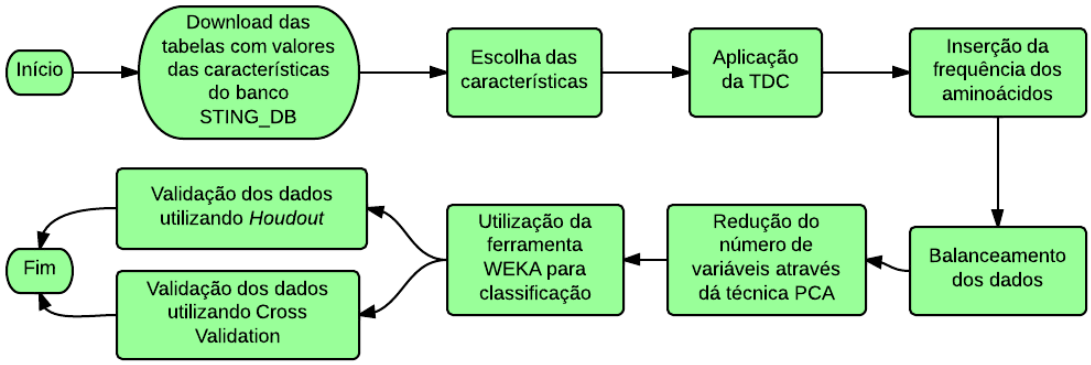
Dobson e Doig (2005) apresentaram um método para predição de classes de proteínas a partir de seus dados estruturais, utilizando atributos simples, tais como o conteúdo da estrutura secundária, propensões de aminoácidos e propriedades de superfície, facilitando o entendimento sobre a relação entre estrutura e função. A ideia do método proposto é, a partir de um grupo de atributos estruturais, classificar cada proteína dentro de uma das seis classes protéicas, conhecidas como superfamílias. Eles utilizaram o banco de dados ASTRAL SCOP (MURZIN et al., 1995) e obtiveram uma precisão de 35% usando SVMs (DOBSON; DOIG, 2005).

No trabalho de Oliveira et al (2006), o objetivo foi melhorar o processo de seleção de parâmetros para aumentar a precisão do modelo de classificação de proteínas. Por meio de uma abordagem híbrida que utilizou recursos da matemática e da estatística, os autores utilizaram um conjunto de estruturas de proteínas e abordaram três desafios presentes na classificação de proteínas: o ruído presente nos parâmetros, o grande número de variáveis e o número não balanceado de membros por classe.

A Figura 13 apresenta todas as etapas da metodologia proposta neste artigo (OLIVEIRA et al., 2006). Os autores utilizaram o banco de dados Sting_DB, que opera como uma coleção de dados retirados do banco de dados públicos PDB e contêm diversos valores de características

de cada uma de suas proteínas. Em seguida, utilizaram a transformada discreta do cosseno, para eliminar o problema de se trabalhar com tamanhos diferentes de cadeias de aminoácidos. Depois, efetuaram um balanceamento entre as classes de proteínas utilizadas, pois o fato de uma classe possuir muito mais elementos do que a outra pode afetar a precisão do classificador. Para diminuir o número de variáveis utilizadas, os autores aplicaram o algoritmo de análise de componentes principais (PCA, do inglês *Principal Component Analysis*). E por último, utilizaram quatro métodos de aprendizado de máquina da ferramenta WEKA (HALL et al., 2009): Árvore de Decisão, Modelo Bayesiano, Redes Neurais e SVM.

Figura 13 – Fluxograma da metodologia de classificação de proteínas segundo Oliveira et al (2006)



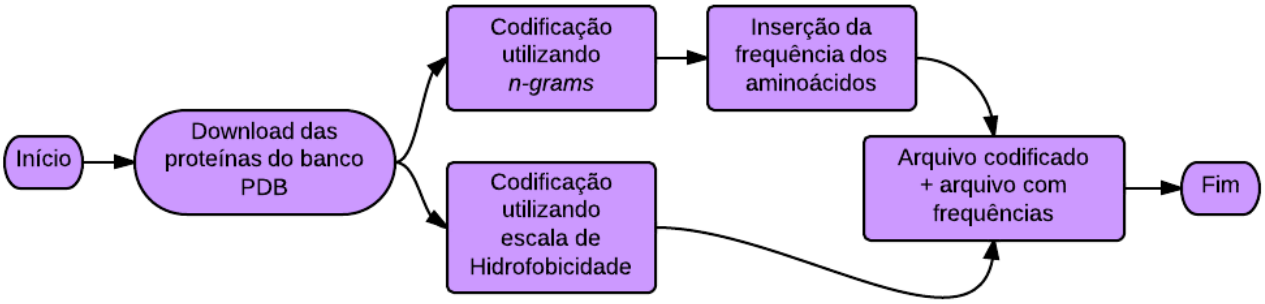
Fonte: Adaptado de (OLIVEIRA et al., 2006)

Os primeiros testes foram realizados, utilizando ao todo 241 variáveis, sem a aplicação da técnica de PCA, e obteve-se uma taxa de 68,41% de precisão para o método de Árvore de Decisão, seguido do método de Redes Neurais com 64,49%, SVM com 42,60% e método Bayesiano com 41,42%. Posteriormente, foram realizados testes utilizando a redução de variáveis, que reduziu o número de parâmetros para 80. Nesse caso, o método baseado em Árvore de Decisão obteve novamente o melhor resultado com precisão igual a 69,01%, seguido do método de Redes Neurais com 60,35%, SVM com 59,17% e o método Bayesiano com 49,11% de precisão.

Em (ROSSI; BRUNETTO, 2006), foi discutida a questão do pré-processamento dos dados antes da etapa de classificação. Foram analisadas duas metodologias de codificação do alfabeto que representa os aminoácidos, como pode ser visto no diagrama da Figura 14, com o intuito de verificar qual delas apresentaria o menor tempo de processamento.

A primeira metodologia é chamada de $n - gram$, e é utilizada em indexação de textos, para efetuar o casamento de padrões. Ela verifica quantas vezes uma determinada sequência

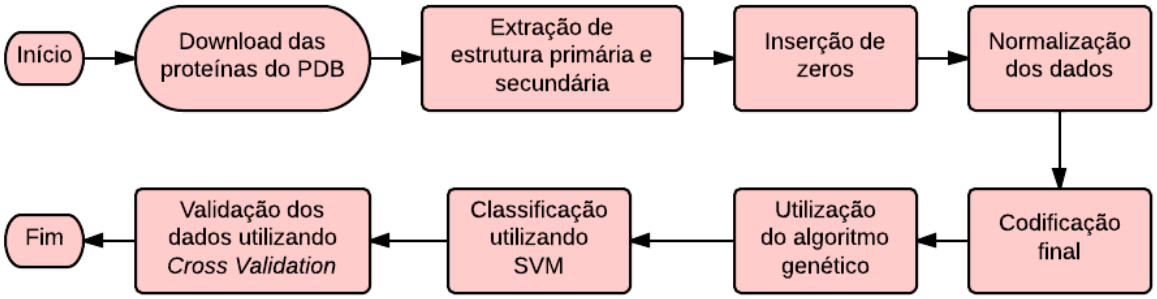
Figura 14 – Fluxograma de metodologias de codificação proposto por Rossi e Brunetto (2006)



Fonte: Adaptado de (ROSSI; BRUNETTO, 2006)

de caracteres $(1..n)$ ocorre em todo o texto, além de verificar a frequência dos aminoácidos em cada proteína. A segunda metodologia utilizou a escala de hidrofobicidade para criar uma codificação para os aminoácidos dividindo-os em três categorias: hidrofóbicos, neutros e hidrofílicos. Com base no tempo de execução de cada metodologia foi observado que a que utiliza escala de hidrofobicidade foi a mais eficiente.

Figura 15 – Fluxograma da metodologia utilizada por Resende et al (2012)



Fonte: Adaptado de (RESENDE et al., 2012)

O trabalho de Resende et al (2012) foi uma modificação do trabalho de Nascimento, Yoshioka e Calanzans (2011), e utilizou a metodologia da Figura 15. Primeiramente a base de dados foi criada com as mesmas proteínas do PDB utilizadas por (DOBSON; DOIG, 2005). Feito isso, foi feito um pré-processamento dos dados com o objetivo de extrair as características primárias e secundárias das proteínas. Porém, o problema de termos diferentes tamanhos para as cadeias de aminoácidos deveria ser resolvido, e para isso foram inseridos zeros até que as cadeias atingissem o tamanho da maior cadeia presente no banco de dados. Depois, foi feita uma normalização dos dados, com base no menor e maior valor utilizados, juntamente com uma codificação que utilizou tabelas *hash*. Por fim, foi utilizado um algoritmo genético para detectar

os melhores valores para os parâmetros do classificador SVM. Os resultados dessa metodologia foram os seguintes: 79,74% de acurácia, 70,31% de sensibilidade, 70,06% de precisão e 96,72% de especificidade.

Em (DIAS, 2007) é proposto um novo modelo de predição, que tem como objetivo sugerir um conjunto de prováveis funções de uma proteína, dada sua estrutura, utilizando o conceito de suas funções moleculares como parâmetros estruturais, calculados a partir da conformação espacial da própria proteína, retirada do Sting_DB, através da técnica de aprendizado de máquina SVM. O modelo utilizou as mesmas características propostas por Oliveira et al (2006) , porém difere do paradigma comum de predição, por não ser necessário calcular similaridades através de alinhamentos entre a proteína que se deseja prever a função e as proteínas de função conhecida. Foi criado um classificador para cada função escolhida, e caso a proteína execute a função desejada, ele devolve uma resposta do tipo sim ou não. Em seguida, foi criado um classificador global, que reúne todas as funções trabalhadas. O autor utilizou as métricas de precisão e sensibilidade que obtiveram, respectivamente, 98% e 93%.

3 METODOLOGIA

Este capítulo descreve a metodologia adotada para a realização deste trabalho. A Tabela 3 apresenta as classes utilizadas, seguida de uma breve descrição de suas funções e da quantidade de proteínas de cada classe. Foram utilizadas as mesmas proteínas coletadas pelos autores dos seguintes trabalhos: (DOBSON; DOIG, 2005), (OLIVEIRA et al., 2006) e (RESENDE et al., 2012).

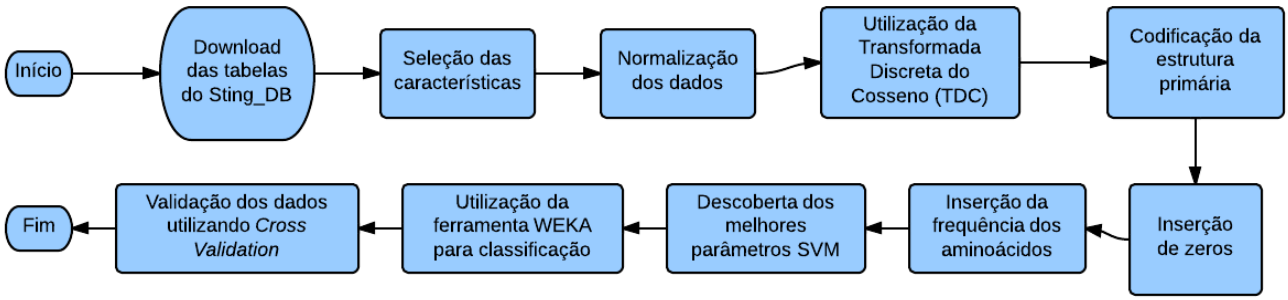
Tabela 3 – Tabela de classes de proteínas e suas quantidades

Classes	Função	Quantidade
Oxidorredutases	Cataliza reações de redução e oxidação.	76
Transferases	Cataliza reações de transferências de grupos funcionais.	120
Hidrolases	Cataliza reações de hidrólise (transferência de grupos funcionais para água).	161
Liases	Cataliza a quebra de ligações $C-C$, $C-O$ e $C-N$.	60
Isomerases	Cataliza a transferência de grupos dentro da mesma molécula para formar isômeros.	57
Ligases	Cataliza a ligação do tipo $C-C$, $C-S$, $C-O$ e $C-N$.	18

Fonte: Adaptado de (RESENDE et al., 2012)

O diagrama da Figura 16, mostra todas as etapas presentes nesta metodologia e serão descritas a seguir.

Figura 16 – Fluxograma da Metodologia Proposta



Fonte: Elaborado pelo autor

3.1 *Download* das tabelas

Apesar de este trabalho utilizar as mesmas proteínas do PDB que vários autores já utilizaram, não trabalhamos diretamente com os arquivos PDB de suas respectivas proteínas. Neste trabalho, foi criado um *script* que utilizou o mesmo nome das proteínas anteriormente utilizadas, mas realizou o *download* das tabelas que possuem as diversas características do banco de dados Sting_DB. Essas tabelas também mantinham a estrutura primária das proteínas, porém forneciam diversas outras informações sobre as características das proteínas, que não estão presentes nos arquivos do PDB.

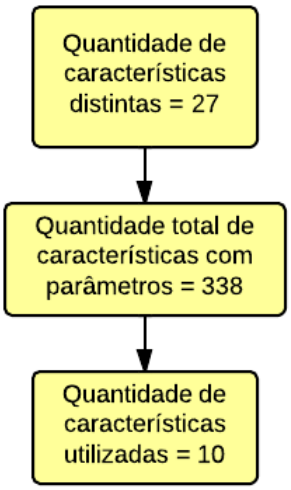
Além disso, estre trabalho utilizou cada cadeia de proteína como se fosse uma proteína em si, pois uma proteína pode ser formada por diversas cadeias, que apresentam suas características independentemente uma das outras. Porém, verificamos se existiam algumas cadeias dentro da mesma classe e dentro de classes diferentes, que apresentavam a mesma sequência de aminoácidos, para evitar redundâncias na metodologia. Feito isso, foram extraídas as características que serão descritas na próxima seção.

3.2 Seleção das principais características

O Sting_DB disponibiliza 27 características distintas para cada aminoácido de cada proteína, classificadas nos seguintes grupos: Densidade, Contatos, Estrutural, Físico-Químicas e Geométricas. Porém, a maioria dessas características pode ter o seu valor alterado de acordo com os parâmetros que são utilizados. Por exemplo, a característica Densidade de cada aminoácido, pode variar de acordo com o número de vizinhos considerados ($3, 4, 5, 6$ ou 7), com o local de seu centro atômico que pode ser no Carbono Alfa (CA) ou no Último Átomo Pesado (LHA do inglês *last heavy atom*), ou com outros parâmetros. Sendo assim, ao invés de obtermos apenas um valor para a densidade, obtemos 39 diferentes valores para apenas uma característica. Com isso, o Sting_DB disponibiliza ao todo 338 valores para todas as possíveis permutações dos parâmetros das 27 características. A Figura 17 mostra um diagrama que representa melhor a quantidade de características distintas, seguido da quantidade total de características disponibilizas e da quantidade final de características utilizadas neste trabalho.

Para selecionar quais seriam as melhores características a serem utilizadas, utilizamos

Figura 17 – Diagrama de representação da quantidade de características disponibilizadas pelo Sting_DB e quantidade utilizada neste trabalho



Fonte: Dados da Pesquisa

como base os trabalhos de Oliveira et al (2006) e Dias (2007). No primeiro trabalho, foram realizados testes que tinham como objetivo reduzir a redundância dos dados das características do banco Sting_DB. Na época existiam apenas 18 características distintas, e os autores concluíram que dessas 18, apenas 11 não eram redundantes entre si, ou seja, apresentavam dados diferentes. Já o trabalho de Dias (2007) utilizou 6 dessas 11 características selecionadas por Oliveira et al (2006), junto com 4 novas características que foram inseridas no banco de dados nesse intervalo de 1 ano entre os dois trabalhos.

De 2007 até o presente momento, foram inseridas mais 5 novas características, totalizando assim 27 características distintas disponíveis atualmente. Este trabalho, utilizou algumas características que estes dois trabalhos já haviam utilizado, além de algumas destas novas características acrescentadas nos últimos anos. Dentre todas as 27 caraterísticas, as 10 características escolhidas para serem utilizadas neste trabalho serão descritas nas seções seguintes.

Também devemos ressaltar que nenhum trabalho da literatura mencionou quais os parâmetros das características que foram utilizados. A única informação passada foi o nome das características. Ou seja, nós sabemos quais das 27 características distintas que foram utilizadas, mas não sabemos quais das 338 que foram utilizadas. Com isso, nós decidimos utilizar apenas um valor de parâmetro para cada característica, e ao invés de pegarmos todos os possíveis valores para uma característica, como foi descrito no início desta seção no exemplo da característica de Densidade, pegamos apenas 1.

3.2.1 Potencial Eletrostático

Os átomos que compõem os resíduos de aminoácido das proteínas podem, em determinadas condições, apresentar carga elétrica, que interagem com outras regiões carregadas da própria proteína ou ainda com outras moléculas e/ou íons de seu ambiente. Portanto, em um determinado ponto do espaço é possível calcular o potencial eletrostático devido a cargas presentes nas macromoléculas ao redor do ponto.

O potencial eletrostático é uma pressão elétrica que quando varia, produz um campo capaz de atrair ou repelir partículas eletricamente carregadas. Conhecer essa característica é importante, pois ela interfere diretamente na estabilidade de uma ligação entre a proteína e o seu ligante, determinando muitos processos biológicos, a estabilidade das proteínas e o mapeamento de canais, podendo ser usadas para análise das estruturas e a função das proteínas. O seu cálculo é feito com base nos valores da Tabela 4 que apresenta os tipos de contato e a respectiva energia de cada um.

Tabela 4 – Valores para energias de contato

Tipos de Contato	**Energia em Kcal/mol**
Van der Waals	0,08
Interações Hidrofóbicas	0,6
Contatos dos anéis aromáticos	1,5
Pontes de Hidrogênio	2,6
Pontes Salinas	10,0
Pontes Dissulfídicas	85,0

Fonte: (DIAS, 2007)

3.2.2 Hidrofobicidade

A maior parte das proteínas cuja estrutura foi resolvida e depositada no PDB são hidrofílicas. Porém, nem todos os tipos de aminoácidos que constituem as proteínas possuem em sua cadeia lateral átomos de nitrogênio e oxigênio capazes de estabelecerem ligações de hidrogênio com as moléculas de água. Para esses aminoácidos é associado o termo hidrofílico enquanto que para o restante dos aminoácidos é associado o termo hidrofóbico.

A característica hidrofóbica é associada ao favorecimento energético de átomos apolares (em especial o átomo de carbono) a estarem juntos espacialmente, reduzindo assim a área de contato com solvente polar. Dessa forma, os átomos de nitrogênio e oxigênio ficam mais expostos ao solvente.

3.2.3 Hot−Spots

Esta característica indica a existência de manchas hidrofóbicas nas superfícies das proteínas. Elas são potencialmente importantes para a identificação de porções superficiais que podem se envolver nas interações das proteínas.

3.2.4 Curvatura

Uma superfície de uma proteína não é uma superfície plana. Nela estão presentes áreas côncavas e convexas. As porções côncavas das superfícies podem ser indicadores de prováveis interações com outras proteínas. Durante o processo de ancoragem, duas moléculas devem encaixar geometricamente porções opostas das superfícies. A curvatura média dessa porção de superfície deve ser igual a duas superfícies, de sinais opostos.

3.2.5 Ordem de Cross Link

Devido ao dobramento da sequência de resíduos de aminoácido na estrutura tridimensional, resíduos são colocados próximos no espaço, e podem, portanto interagir entre si. O parâmetro *Cross Link* presente no JPD leva essa característica em conta, e é definida em relação ao número dos contatos estabelecidos entre seguimentos de resíduos de aminoácido de no mínimo 15 resíduos na estrutura primária da proteína.

3.2.6 Ordem de Cross Presence

Seguindo a mesma definição de Cross Link, esse valor é calculado através da contagem de todos os resíduos de aminoácidos que estão dentro da sonda esférica centrada no CA, Carbono Beta(CB) ou no LHA, mesmo que os resíduos de aminoácido não estejam estabelecendo nenhum contato entre si.

3.2.7 Densidade

A densidade local de cada resíduo de aminoácido é calculada utilizando uma abordagem de sonda esférica. Para cada aminoácido, uma sonda esférica de raio variável (entre 3 e 7) é centrada no CA e no LHA. As massas dos átomos internos à sonda esférica são somadas e divididas pelo volume da sonda esférica.

3.2.8 Distância do Centro de Gravidade

Representa a distância entre o CA de cada resíduo e o centro de massa da cadeia (baricentro).

3.2.9 Esponjicidade

A Esponjicidade é uma medida do espaço vazio do nano-ambiente de cada aminoácido e segue a mesma abordagem descrita para a densidade.

3.2.10 Ocupação Múltipla

A presença de dois ou mais conjuntos de coordenadas para o mesmo átomoresíduo no arquivo PDB é devido à interpretação do mapa de densidade de elétrons onde o experimento

registra uma difração a partir dos cristais de congelamento da mesma molécula, mas com as diferentes posições de espaço para um determinado aminoácido.

3.3 Normalização

O propósito da normalização é minimizar os problemas oriundos do uso de unidades e dispersões distintas entre as variáveis. Isso evita que uma dimensão se sobreponha em relação às outras, prevenindo assim que o aprendizado fique estagnado.

A técnica de normalização utilizada neste trabalho foi a $Max - MinEqualizada$, que utiliza os valores máximo e mínimo para normalizar linearmente os dados entre $[0, 1]$, através da Equação 3.1, onde $novo\_x$ é o novo valor de x para um determinado número x que será normalizado, $min(x)$ é o menor valor de x presente nos dados e $max(x)$ é o maior valor de x.

$$novo\_x = \frac{x - min(x)}{max(x) - min(x)} \tag{3.1}$$

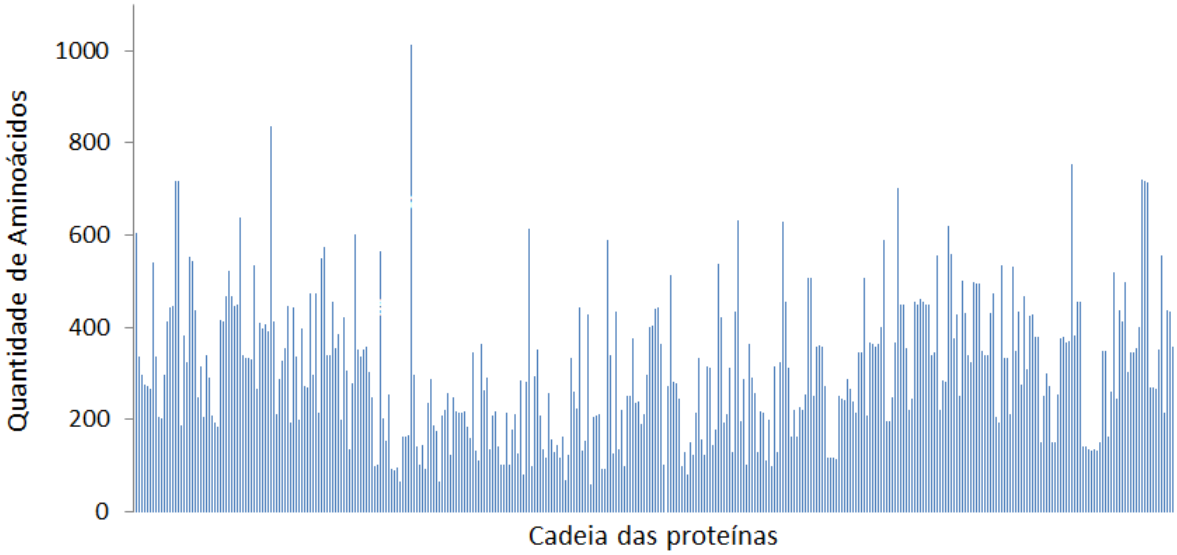
Para que a aplicação da normalização fosse possível, foi feito um pré-processamento dos dados, de forma a obter os maiores e menores valores para cada uma das características extraídas das proteínas. Em seguida, foi feita a normalização dos dados por característica, ou seja, para cada característica foram obtidos os maiores e menores valores para normalização daqueles dados.

3.4 Transformada Discreta do Cosseno

Uma vez que todas as características já haviam sido selecionadas e normalizadas, ainda existia um problema a ser resolvido: o problema da diferença da quantidade de aminoácidos de cada proteína, que pode ser visto na Figura 18. Logo, para que a utilização dos classificadores fosse possível, o tamanho de todos os vetores de entrada deveria ser o mesmo.

Para solucionar este problema foi utilizada a técnica da Transformada Discreta do Cosseno (TDC) (AHMED; NATARAJAN; RAO, 1974), também utilizada nos trabalhos de Oliveira et al (2006) e Ulisses (2007). A TDC foi escolhida pois é uma transformação que preserva as normas e os ângulos dos vetores, e é uma transformada para números reais, ao contrário da

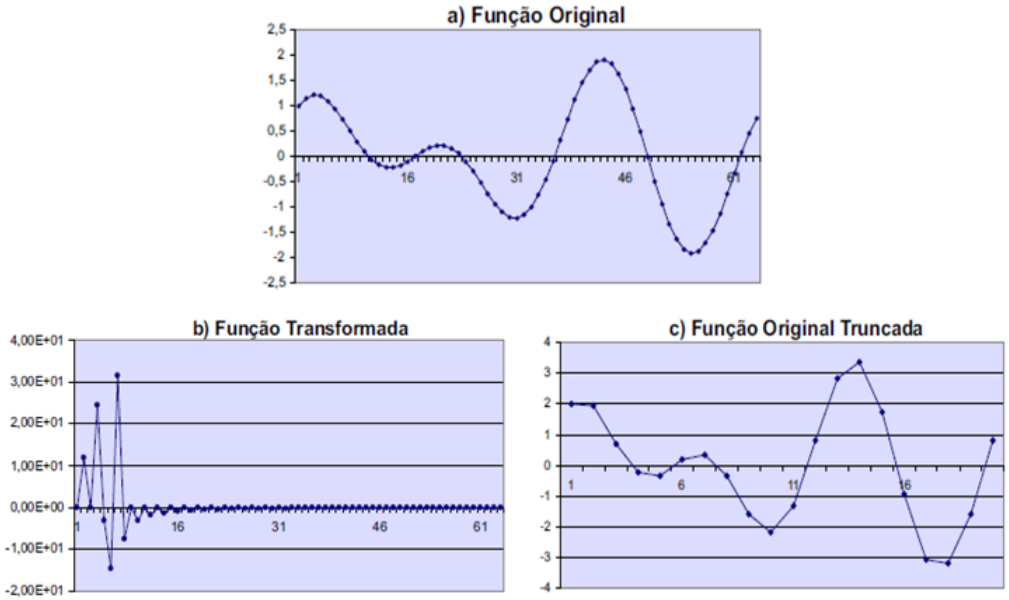
Figura 18 – Diferença entre a quantidade de aminoácidos das cadeias protéicas



Fonte: Dados da Pesquisa

Transformada Discreta de Fourier, que é uma transformada definida sobre o corpo dos números complexos. Esta técnica representa um vetor com os dados originais, como apresentado na Figura 19 (a) em um sinal no domínio da frequência, Figura 19 (b). Uma vez que os dados estão transformados, grande parte da informação está localizada nos primeiros coeficientes do vetor transformado. Ao passo que nos últimos coeficientes, estão as pequenas variações que ocorrem na sequência original, que possivelmente serão os ruídos.

Figura 19 – Exemplo de utilização da Transformada Discreta do Cosseno



Fonte: (DIAS, 2007)

Ao selecionarmos apenas os primeiros n coeficientes, estamos dizendo que a parte mais relevante da informação foi obtida. Para verificar quais são os valores originais correspondentes deve-se usar a Transformada Inversa do Cosseno. O resultado serão valores próximos aos valores originais, visto que ao transformá-los há uma pequena perda de informação. Mesmo com essa perda de informação, o comportamento da curva é preservado, como pode ser visto na Figura 19 (c).

3.5 Codificação

As estruturas das proteínas, descritas no Capítulo 2, são fonte de muita informação que pode ser utilizada durante o processo de classificação. Para que elas possam ser utilizadas, deve ser utilizada alguma técnica de codificação para que seja possível mapear suas informações de forma numérica. Uma vez que o mapeamento foi feito, essa informação foi acrescentada às informações de características que já haviam sido extraídas nos passos anteriores.

No trabalho de Wagner e Heitor (2003) foi proposta uma codificação da estrutura primária das proteínas, também utilizada neste trabalho, coluna Valor Real da Tabela 5, que utiliza os valores da escala de Hidrofobicidade proposta por Kyte e Doolittle (1982), também denominada escala KD, presentes na coluna de Valor K.D..

3.6 Inserção de Zeros

Após a etapa de codificação da estrutura primária das proteínas, o problema de vetores de diferentes tamanhos voltou a ocorrer. Mas, devido à grande quantidade de informações contida na estrutura primária, a utilização da TDC se tornou inviável, pois apesar de serem selecionados os valores mais significativos, a ordem dos aminoácidos na cadeia seria perdida.

Para contornar este problema foi feita a inserção de zeros, semelhante à feita por (RESENDE et al., 2012) em todos os vetores com a estrutura primária, de forma a preencher os campos que faltavam até chegar ao tamanho do maior vetor da base de dados, que no caso foi de 1014 aminoácidos.

Tabela 5 – Valores de Hidrofobicidade e codificação adotada em valores reais

Aminoácido (símbolo)	Valor K.D	Valor Real	Categoria
I	+4,5	0,05	Hidrofóbico
V	+4,2	0,10	Hidrofóbico
L	+3,8	0,15	Hidrofóbico
F	+2,8	0,20	Hidrofóbico
C	+2,5	0,25	Hidrofóbico
M	+1,9	0,30	Hidrofóbico
A	+1,8	0,35	Hidrofóbico
G	-0,4	0,40	Neutro
T	-0,7	0,45	Neutro
S	-0,8	0,50	Neutro
W	-0,9	0,55	Neutro
Y	-1,3	0,60	Neutro
P	-1,6	0,65	Neutro
H	-3,2	0,70	Hidrofílico
Q	-3,5	0,75	Hidrofílico
N	-3,5	0,80	Hidrofílico
E	-3,5	0,85	Hidrofílico
D	-3,5	0,90	Hidrofílico
K	-3,9	0,95	Hidrofílico
R	-4,0	1,00	Hidrofílico

Fonte: (WEINERT; LOPES, 2003)

3.7 Frequência dos Aminoácidos

Após termos selecionado as principais características e realizado a codificação da estrutura primária, foi decidido realizar a inserção da frequência de cada aminoácido, assim como proposto em alguns trabalhos da literatura, como Rossi e Brunetto (2006) e Oliveira et al (2006).

3.8 Seleção dos Melhores Parâmetros da SVM

Ao utilizar um algoritmo de aprendizado de máquina, é necessário decidir quais os melhores parâmetros a serem utilizados. Cada classificador possui os seus respectivos parâmetros, e se tratando das SVMs, existem dois parâmetros principais que são c e γ. No trabalho de Resende et al (2012) foi utilizado um algoritmo genético que dizia quais os melhores valores para esses parâmetros. Neste trabalho foi utilizado um *script* disponibilizado pelos autores do LibSVM (CHANG; LIN, 2011).

3.9 Ferramenta de Classificação

Para classificar os dados, foi utilizada a ferramenta WEKA, que é um pacote desenvolvido pela Universidade de Waikato em 1993, com o intuito de agregar algoritmos de aprendizado de máquina para mineração de dados na área de Inteligência Artificial, como por exemplo: Modelo Bayesiano, Redes Neurais, Regressão Linear, Árvores de Decisão, IB1, *Bagging*, *LogistBoot*, etc (HALL et al., 2009). Dentre todos os algoritmos de classificação disponibilizados, foi escolhido a SVM, visto que têm obtido os melhores resultados da literatura. Além disso, essa ferramenta disponibiliza as métricas de precisão e sensibilidade, que serão apresentadas na próxima seção, juntamente com outras métricas utilizadas neste trabalho.

3.10 Métricas de Avaliação

Para avaliar o desempenho do método proposto, foram utilizadas 4 métricas: Precisão, Sensibilidade, Acurácia e Especificidade (RESENDE et al., 2012). Porém, para o completo entendimento dessas métricas, é necessário o conhecimento dos seguintes conceitos:

- **Verdadeiro Positivo (VP):** quantidade de proteínas corretamente classificadas na classe em questão.
- **Falso Negativo (FN):** quantidade de proteínas da classe analisada, erroneamente classificadas.
- **Falso Positivo (FP)**: proteínas que não são da classe considerada, mas que foi classificada nesta classe.
- **Verdadeiro Negativo (VN)**: quantidade de proteínas pertencentes a outras classes, classificada como a classe em questão.

A métrica Precisão tem como objetivo medir o quanto das predições positivas estão corretas. É a taxa de instâncias corretamente classificadas como pertencentes a classe em questão dentre todos os que foram classificados na classe em questão, conforme a Equação 3.2.

$$P = \frac{VP}{VP + FP} \tag{3.2}$$

A Sensibilidade é a taxa de instâncias corretamente classificadas como pertencentes a classe em questão, dentre todos os que realmente são da classe em questão, como apresentado na Equação 3.3.

$$S = \frac{VP}{VP + FN} \tag{3.3}$$

A Especificidade é a taxa de proteínas de outras classes classificadas corretamente, sobre o número proteínas classificadas em outras classes, conforme Equação 3.4.

$$E = \frac{VN}{VN + FP} \tag{3.4}$$

A Acurácia é a taxa total de instâncias corretamente classificadas e é definida pela Equação 3.5.

$$A = \frac{VP + VN}{VP + VN + FP + FN} \tag{3.5}$$

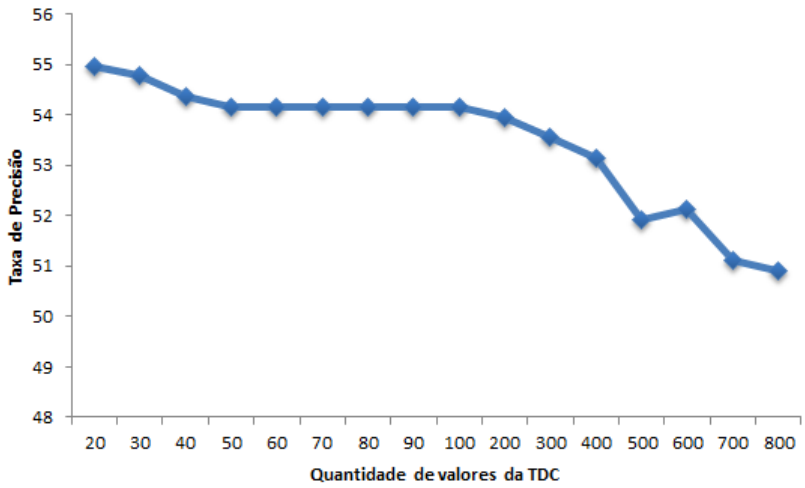
Foram implementados alguns métodos na linguagem de programação Java, que foram utilizados em algumas das etapas da metodologia proposta, principalmente na parte de pré-processamento dos dados das tabelas, na geração dos arquivos de dados utilizados como entrada para a ferramenta Weka e no cálculo dos valores das métricas de avaliação.

4 RESULTADOS E DISCUSSÕES

Este capítulo tem como objetivo apresentar os resultados dos testes realizados para a metodologia proposta. Todos os testes foram realizados utilizando o classificador LibSVM (CHANG; LIN, 2011), capaz de classificar os dados em diferentes classes. Este classificador está integrado com a ferramenta WEKA, que também disponibiliza muitos outros algoritmos de classificação, que não foram tratados neste trabalho. Além disso, foi utilizada a técnica de amostragem *k-fold Cross Validation* descrita na Seção 2.6, com valor de $k = 10$.

Os primeiros testes foram realizados com o objetivo de determinar qual o melhor valor da TDC a ser utilizado. Para isso, foram realizados testes variando-se o número de coeficientes entre 20 e 800, com intervalo de 10, e medindo-se a precisão encontrada, como pode ser visto na Figura 20. O motivo da escolha do valor máximo para 800 foi porque o maior valor possível para utilização da TDC neste trabalho seria igual ao número de características utilizadas, no caso 10, multiplicado pelo tamanho da menor cadeia de aminoácidos da base utilizada, no caso 80.

Figura 20 – Exemplo de utilização da Transformada Discreta do Cosseno



Fonte: Dados da Pesquisa

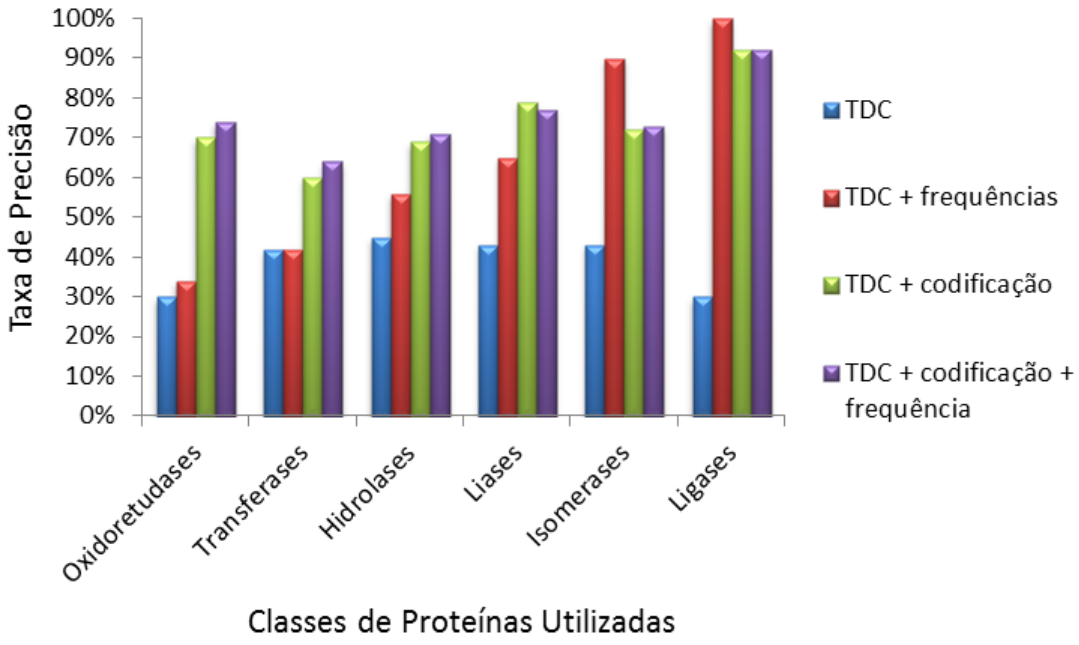
A figura mostra que, assim como no trabalho de (OLIVEIRA et al., 2006) o melhor valor de coeficientes a ser utilizado foi 20, que obteve precisão de 55%, e à medida que aumentamos a quantidade de coeficientes utilizados, verificou-se que a taxa de precisão cai. Isso aconteceu pois, como explicado na seção anterior, a TDC seleciona os coeficientes mais significativos e

os coloca nas primeiras posições, deixando para as outras apenas as informações de ruídos. Quando selecionamos mais valores, acabamos pegando valores de ruídos, e isso atrapalha o classificador.

Uma vez que o valor da TDC foi escolhido, o nosso objetivo foi verificar qual a melhor combinação de metodologias para a predição de função de proteínas. Para isso, realizamos testes para cada uma das etapas descritas no Capítulo 3. Para cada teste, foi utilizado o *script* de seleção dos melhores parâmetros da SVM, buscando assim os melhores resultados possíveis para cada teste. Observou-se que a utilização dos valores de c e γ retornados pelo *script*, aumentou em média 10% em todas as métricas.

Nós analisamos os dados com base na precisão e na sensibilidade, porém iremos apresentar nesta parte apenas as informações da precisão, que nos informa o quanto das predições positivas estão corretas, visto que o valor de sensibilidade cresceu nas mesmas proporções que o da precisão. Primeiramente avaliou-se a precisão do classificador considerando-se apenas a TDC sobre as 10 características extraídas do banco Sting_DB. Em seguida, testamos os resultados para a TDC + frequência dos aminoácidos. Depois com a TDC + codificação da estrutura primária das proteínas. E por fim, quais seriam os resultados utilizando-se toda a metodologia. A Figura 22 apresenta os resultados de todos estes testes.

Figura 21 – Resultados Parciais da Metodologia Utilizada



Fonte: Elaborado pelo autor

Observa-se pela figura que o classificador obteve uma precisão de 39% quando utilizou-

se apenas a TDC sobre as 10 características utilizadas. Este resultado indica que o classificador encontrou muitos falsos positivos, reduzindo-se assim a precisão para cada classe. Já ao acrescentarmos a informação da frequência dos aminoácidos, observou-se um aumento de 25% da média de precisão do classificador, que no total obteve 64%, aumentando assim a taxa de verdadeiros positivos. Além disso, percebemos que para a classe das Ligases, o classificador obteve 100% de precisão, que significa que nenhuma outra classe chegou a classificar alguma instância erroneamente nesta classe.

Já para os testes que utilizaram a codificação da estrutura primária, os resultados foram melhores. Para testes utilizando a TDC + codificação da estrutura primária, foi obtida uma média de 73% de precisão, que foi 9% mais alta do que a metodologia que utilizou apenas a TDC + frequência dos aminoácidos. E para os testes finais, que utilizaram toda a metodologia, ou seja, TDC + codificação da estrutura primária + frequência dos aminoácidos, obtivemos uma precisão média de 75%, utilizando os valores de $c = 512$ e $\gamma = 0.00048828125$ para a função RBF da SVM.

Para avaliarmos a metodologia proposta através das métricas descritas no Capítulo 3, é necessária a visualização da Matriz de Confusão, exibida na Tabela 6 que mostra o número de classificações corretas em contraste com o número de classificações preditas para cada classe, sendo assim uma medida efetiva do classificador.

Tabela 6 – Matriz de Confusão.

	O	**T**	**H**	**Lia**	**I**	**Lig**	**Total**
Oxidoredutases **(O)**	50	9	14	0	2	1	76
Transferases **(T)**	5	88	19	4	4	0	120
Hidrolase **(H)**	6	21	127	3	4	0	161
Liase **(Lia)**	3	13	8	34	2	0	60
Isomerase **(I)**	3	6	9	2	37	0	57
Ligase **(Lig)**	1	1	2	1	2	11	18

Fonte: Dados da Pesquisa

Ao analisar a matriz de confusão, podemos observar que o classificador obteve um melhor desempenho com a classe de menor número de instâncias, as Ligases. Também podemos ver que, como existe um desbalanceamento entre as classes, muitas instâncias foram classificadas em classes diferentes, sendo que a maior parte dos falsos positivos ocorreu para as duas classes com maior quantidade de proteínas, que são as Hidrolases e as Transferases, respectivamente. A Tabela 7, nos mostra os resultados de todas as 4 métricas avaliadas neste trabalho.

Analisando a sensibilidade, que nos informa exatamente a porcentagem de verdadeiros

Tabela 7 – Tabela de resultados das métricas utilizadas.

Classes	Precisão	Sensibilidade	Acurácia	Especificidade
Oxidoredutases	0,74	0,66	0,71	0,76
Transferases	0,64	0,73	0,66	0,58
Hidrolase	0,71	0,79	0,73	0,68
Liase	0,77	0,57	0,70	0,83
Isomerase	0,73	0,65	0,70	0,75
Ligase	0,92	0,61	0,78	0,94
Média	0,75	0,67	0,71	0,76

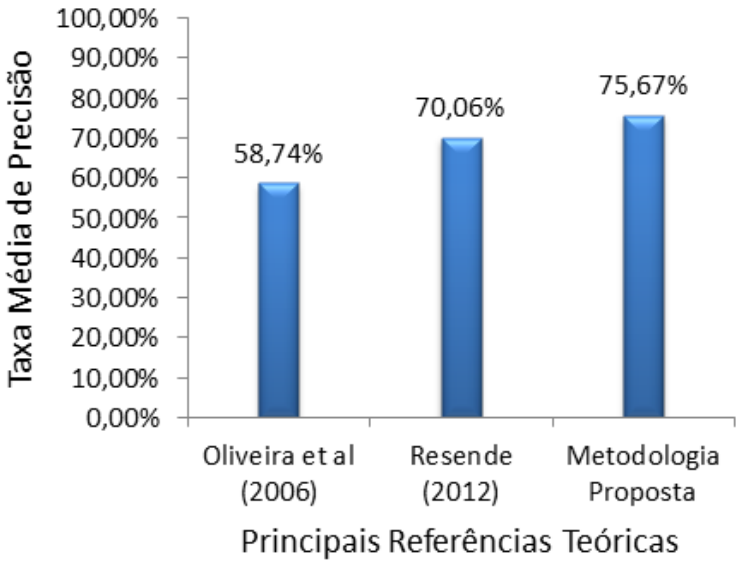
Fonte: Dados da Pesquisa

positivos dentre todas as instâncias da classe, podemos observar que foi a métrica que obteve o menor valor geral, com média de apenas 67%, devido principalmente ao grande desbalanceamento entre o número de instâncias das classes. Assim como em outros trabalhos da literatura, as classes com maior número de instâncias obtiveram um maior número de verdadeiros positivos, alcançando sensibilidade de 79% para as Hidrolases que de 161 instâncias classificou corretamente 127, e de apenas 57% para as Liases que de 60 instâncias classificou apenas 34 corretamente.

Já para a métrica de precisão, percebemos que dentre todas as predições realizadas nas classes, 75% estão corretas. Essa métrica nos mostra a porcentagem de acertos, dentre a quantidade de instâncias que foram classificadas para a classe em questão. O maior valor encontrado foi para a classe das Ligases, que das 12 predições acertou 11, e o menor valor encontrado foi para classe das Transferases que de 138 valores acertou apenas 88, obtendo precisão igual a 64%. A explicação para essa enorme quantidade de falsos positivos é novamente o fato de termos grande diferença entre o número de instâncias entre as classes. Para a acurácia, que mede a taxa total de instâncias corretamente classificadas, obtivemos uma média de 71% e para a especificidade uma média de 76%.

Além de analisarmos cada uma das métricas, vale a pena comparar nossos resultados com outros trabalhos similares, que utilizaram a mesma base de dados e o mesmo classificador. A Figura 22 apresenta a média da precisão para os trabalhos de Oliveira et al (2006) e Resende et al (2012). Nela podemos ver que, a média da metodologia proposta neste trabalho, supera a média de precisão dos trabalhos anteriores.

Figura 22 – Comparação entre resultados de Precisão de diversos autores



Fonte: Dados da Pesquisa

5 CONCLUSÃO

Neste trabalho foi apresentada uma metodologia para predição de função de proteínas, que combina recursos da matemática (Transformada Discreta do Cosseno), com aprendizado de máquina (Máquinas de Vetor de Suporte), dados da estrutura primária e secundária e diversas características disponibilizadas pelo banco de dados Sting_DB.

Os resultados de cada etapa da metodologia foram comparados entre si, de forma a escolher a melhor combinação de métodos. Os resultados médios obtidos foram de 75% de precisão, 67% de sensibilidade, 71% de acurácia e 76% de especificidade. Além disso, comparamos os resultados com alguns trabalhos relevantes da literatura e observou-se que, diferentemente de todos os trabalhos anteriores, que obtiveram dados ruins para as classes com poucas instâncias, o presente trabalho conseguiu aumentar a sensibilidade e a precisão da classe de Ligases, que possui 17 instâncias, para 61% e 92% ao contrário do trabalho de Resende et al (2012) que obteve 20% e 13% respectivamente.

Porém, o problema do desbalanceamento das classes, que é causado pela diferente de quantidade de proteínas em cada uma das classes utilizadas, ainda é crítico, e observou-se que houve uma alta taxa de falsos positivos para as classes de Hidrolases e Transferases, que são as que apresentam maior número de instâncias. O motivo de termos escolhido utilizar a mesma base de dados de outros autores foi unicamente para que a comparação de resultados pudesse ser feita, porque atualmente o banco de dados PDB apresenta um número muito maior de instâncias para cada uma das classes, e acredita-se que ao trabalhar com as classes balanceadas os resultados possam melhorar significativamente.

Como trabalhos futuros, propomos que sejam utilizadas em conjunto com a presente metodologia, as características das estruturas secundárias, terciárias e quaternárias das proteínas, visto que, elas disponibilizam diversas informações que podem aumentar a eficiência do classificador. Elas não foram utilizadas neste trabalho pois o foco maior foi na seleção das características disponibilizadas pelo Sting_DB, que são calculadas utilizando-se apenas a estrutura primária e secundária das proteínas. Porém, através do banco de dados PDB é possível acessar essas informações de cada proteína.

Também acreditamos que possa ser realizado um trabalho apenas para selecionar quais as melhores características, dentre todas as 338 disponibilizadas pelo banco de dados Sting_DB,

que devem ser utilizadas para predição de função de proteínas. Esse é um problema que por si só é desafiador, visto que a análise a ser feita deve ser fatorial, com o objetivo de verificar qual a combinação de características ideal.

REFERÊNCIAS

AHMED, N.; NATARAJAN, T.; RAO, K. Discrete cosine transfom. *IEEE Transactions on Computers*, IEEE Computer Society, Los Alamitos, CA, USA, v. 23, n. 1, p. 90–93, 1974.

ALBERTS, B. et al. *Fundamentos da biologia celular: uma introdução à biologia molecular da célula*. Artmed, 2002.

ALVAREZ, M.; YAN, C. Exploring structural modeling of proteins for kernel-based enzyme discrimination. In: *Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2010 IEEE Symposium*. 2010. p. 1–5.

BATISTA, G. E. de A. P. A. *Pré-processamento de Dados em Aprendizado de Máquina Supervisionado*. Dissertação (Mestrado) — USP - São Carlos, 2003.

BERMAN, H. M. et al. The protein data bank. *Nucleic acids research*, Oxford University Press, Research Collaboratory for Structural Bioinformatics (RCSB), USA, v. 28, n. 1, p. 235–242, jan. 2000.

BITTENCOURT, V. G. *Aplicações de Técnicas de Aprendizado de Máquina no Reconhecimento de Classes Estruturais de Proteínas*. Dissertação (Mestrado) - Universi - dade Federal do Rio Grande do Norte, 2005.

BORRO, L. C. et al. Predicting enzyme class from protein structure using bayesian classification. *Genet Mol Res*, v. 5, n. 1, p. 193–202, 2006.

CHANG, C.-C.; LIN, C.-J. LIBSVM: A library for support vector machines. *ACM Transa - ctions on Intelligent Systems and Technology*, v. 2, p. 27:1–27:27, 2011.

CORTES, C.; VAPNIK, V. Support-vector networks. *Machine Learning*, Kluwer Academic Publishers, Hingham, MA, USA, v. 20, n. 3, p. 273–297, set. 1995.

DIAS, U. M. *Predição da Função das Proteínas Sem Alinhamentos Usando Máquinas de Vetor de Suporte*. Dissertação (Mestrado) - Universidade Federal de Alagoas, 2007.

DOBSON, P. D.; DOIG, A. J. Predicting enzyme class from protein structure without alignments. *Journal of Molecular Biology*, v. 345, n. 1, p. 187 – 199, 2005.

ENSINO, S. de Apoio ao. *Bioquímica: Proteínas*. Mar. 2006. Acesso em: 18 Abr. 2013. Disponível em: <http://www.iesde.com.br/pai/arquivos/EM_1S_BIO_003.pdf>.

FILETO, R. et al. Pdb-metrics: a web tool for exploring the pdb contents. *Genet. Mol. Res*,v. 5, n. 2, p. 333–341, 2006.

GOMES, F. de C. *Máquinas de Vetores Suporte na Classificação de Impressões Digitais*. Dissertação (Mestrado) — Universidade Federal do Ceará, 2002.

HALL, M. et al. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, ACM, New York, NY, USA, v. 11, n. 1, p. 10–18, nov. 2009. Disponível em: <http://doi.acm.org/10.1145/1656274.1656278>.

LINHARES, S.; GEWANDSZNAJDER, F. *Biologia*. 1ª. ed. [S.l.]: Ática, 2006.

LORENA, A. C.; ANDRé; CARVALHO, C. P. L. F. de. Uma introdução às support vector machines. *RITA*, v. 14, n. 2, p. 43–67, 2007.

MITCHELL, T. *Machine Learning* McGraw-Hill Education (ISE Editions), 1997. Paperback.

MOSS, G. P. *Classification and Nomenclature of Enzymes by the Reacti-ons they Catalyse*. Outubro 2013. Acesso em: 03 Nov. 2013. Disponível em: <http://www.chem.qmul.ac.uk/iubmb/enzyme/rules.html>.

MURZIN, A. G. et al. Scop: a structural classification of proteins database for the investigation of sequences and structures. *Journal of molecular biology*, MRC Laboratory of Molecular Biology and Centre for Protein Engineering, Cambridge, England., v. 247, n. 4, p. 536–540, abr. 1995.

NADZIRIN, N.; FIRDAUS-RAIH, M. Proteins of unknown function in the protein data bank (pdb): An inventory of true uncharacterized proteins and computational tools for their analysis. *International Journal of Molecular Sciences*, v. 13, n. 10, p. 12761–12772, 2012. Disponível em: <http://www.mdpi.com/1422-0067/13/10/12761>.

NESHICH, G. et al. Java protein dossier: a novel web-based data visualization tool for comprehensive analysis of protein structure. *Nucleic Acids Research*, v. 32, p. 595–601, 2004.

OLIVEIRA, L. L. de. *Uso de Estratégias Baseadas em Conhecimento para Algoritmos Genéticos Aplicados à Predição de Estruturas Tridimensionais de Proteínasa*. Dissertação (Mestrado) — Universidade de São Paulo, 2011.

OLIVEIRA, S. d. M. et al. Uma metodologia para seleção de parâmetros em modelos de classificação de proteínas. *Embrapa Informática Agropecuária. Boletim de pesquisa e desenvolvimento*, Campinas: Embrapa Informática Agropecuária., v. 14, 2006.

PANDEY, G.; KUMAR, V.; STEINBACH, M. *Computational Approaches for Protein Function Prediction: A Survey*, 2006.

PIAO, M. et al. Hydrophobic amino acid composition patterns over secondary structure elements of proteins. In: *Bioinformatics and Biomedical Engineering, 2008. ICBBE 2008. The 2nd International Conference on*. [S.l.: s.n.], 2008. p. 148–151.

PRATI, R. C.; BATISTA, G. E. A. P. A.; MONARD, M. C. Uma experiência no balanceamento artificial de conjuntos de dados para aprendizado com classes desbalanceadas utilizando análise roc. In: *IV Workshop de Inteligência Artificial (ATAI'2003)*. [S.l.: s.n.], 2003.

PROSDOCIMI, F. et al. Bioinformática: manual do usuário. *Biotecnologia Ciência & Desenvolvimento*, v. 29, p. 12–25, 2002.

QUíMICA, I. de. *Estrutura Secundária: Folha Beta*. Abr. 2013. Acesso em: 18 Abr. 2013. Disponível em: <http://www.iq.usp.br/bayardo/softwares/proteina/basic/cap4-2/main4-2.html>.

RESENDE, W. et al. The use of support vector machine and genetic algorithms to predict protein function. In: *Systems, Man, and Cybernetics (SMC), 2012 IEEE International Conference on*. [S.l.: s.n.], 2012. p. 1773 –1778.

RODRIGUES, L. et al. Parallel and distributed kmeans to identify the translation initiation site of proteins. In: *Systems, Man, and Cybernetics (SMC), 2012 IEEE International Conference on*. [S.l.: s.n.], 2012. p. 1639–1645.

ROSSI, A. L. D.; BRUNETTO, M. A. de O. C. Métodos de codificação de proteínas para uso com redes neurais artificiais. *X Congresso Brasileiro de Informática em Saúde*, Florianópolis: Anais do X Congresso Brasileiro de Informática em Saúde., 2006.

SCHIAVONI, A. S. *Um Estudo Comparativo de Métodos para Balanceamento do Conjunto de Treinamento em Aprendizado de Redes Neurais Artificiais*. Dissertação (Mestrado) —Universidade Federal de Lavras, 2010.

SILVA, S. G. O. da. *Estrutura Secundária de Proteínas Utilizando Redes Neuronais*. Dissertação (Mestrado) — Universidade de Lisboa, 1999.

VAPNIK, V. *The Nature of Statistical Learning Theory (Information Science and Statistics)*. 2nd. ed. Springer, 1999. Hardcover. Disponí-vel em: <http://www.amazon.com/exec/obidos/redirect?ag=citeulike07-20&path=ASIN/0387987800>.

WATSON, J. D.; LASKOWSKI, R. A.; THORNTON, J. M. Predic-ting protein function from sequence and structural data. *Current Opinion in Structural Biology*, v. 15, n. 3, p. 275 – 284, 2005. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0959440X05000825>.

WEB, V. *Composição Química da Célula - Proteínas*. Abr. 2013. Acesso em: 19 Abr. 2013. Disponível em: <http://www.vestibulandoweb.com.br/biologia/teoria/estrutura-das-proteinas.asp>.

WEINERT, W. R.; LOPES, H. S. Aplicação de um sistema neural ao problema de classificação de proteínas. *VI Congresso Brasileiro de Redes Neurais*, São Paulo: Centro Universitário da FEI., p. 85–90, 2003.