Um algoritmo genético para a seleção de características utilizadas na predição de função de proteínas

Larissa Fernandes Leijôto, Cristiane Neri Nobre (Orientadora)

Pontifícia Universidade Católica de Minas Gerais,
Departamento de Ciência da Computação,
Av. Dom José Gaspar. 500, 30535-901 Belo Horizonte, Brasil
larissa.leijoto@sga.pucminas.br, nobre@pucminas.br
http://www.pucminas.br/pos/informatica

Resumo Proteínas são macromoléculas que possuem alto peso molecular, e formam, ao lado da água, a maior fração das células. As funções que elas desempenham são extremamente importantes, tais como a catálise de reações bioquímicas, a formação do citoesqueleto e o transporte e armazenamento de substâncias. Com o término do sequenciamento do genoma, a descoberta de proteínas tem crescido exponencialmente, e os métodos laboratoriais para a resolução de suas funções não têm conseguido acompanhar esse crescimento. Devido a este fato, torna-se necessária a criação de métodos que auxiliem esse processo de descoberta. Dessa forma, este trabalho propõe uma metodologia de seleção de características físico-químicas, calculadas por meio das estruturas que compõem a proteína. Esta etapa tem por finalidade a escolha de um subconjunto de característica, dentre todas as disponíveis. Uma característica é considerada relevante se com ela a máquina consegue criar uma capacidade de discernimento entre as diferentes classes de proteínas. Para a seleção deste conjunto, foi proposto o uso de um algoritmo genético simples. Os resultados obtidos com a metodologia proposta foram satisfatórios e são tão bons ou melhores do que as referências citadas.

Keywords: Algoritmo Genético, Seleção de Características, Classificação, Proteínas, Support Vector Machine

1 Introdução

A união entre a biologia e a ciência da computação criou uma nova área denominada bioinformática. Essa consiste no desenvolvimento de técnicas computacionais que possuem capacidade de inferir e derivar predições importantes e relevantes a partir de dados adquiridos através da biologia molecular [19].

Proteínas são macromoléculas que possuem alto peso molecular, e formam ao lado da água a maior fração das células. Sendo assim, responsáveis por desempenhar os mais importantes papéis biológicos. Devido ao seu grande conjunto de funcionalidades, o conhecimento de sua função é fundamental para a realização

de diversas aplicações na biologia [22]. Essas aplicações vão desde o desenvolvimento de novas drogas para o tratamento de doenças até aplicações na área agropecuária. Após o término do sequenciamento do genoma humano, a taxa do crescimento de proteínas com função desconhecida tem aumentado significativamente. Os métodos laboratoriais para a descoberta da sua função, tais como a ressonância magnética nuclear e a cristalografia por difração de raios x, possuem um alto custo e demandam muito tempo, por esta razão não conseguem acompanhar esse crescimento. A partir disso, torna-se necessário o desenvolvimento de técnicas computacionais, capazes de auxiliar na definição da função de uma proteína, reduzindo assim a necessidade de testes laboratoriais.

Para essa finalidade existem três abordagens que são comumente usadas. A primeira, realiza a predição de acordo com a similaridade da sequência primária. Essa abordagem é amplamente utilizada devido à alta quantidade de sequências descobertas. Entretanto, falha pois a estrutura primária é a menos conservada no aspecto evolutivo das estruturas. Ou seja, proteínas podem possuir alta similaridade entre suas cadeias, mas desempenhar funções totalmente diferentes. A segunda abordagem está relacionada com a estrutura terciária [24], que é muito mais conservada do que a sequência. A função de uma proteína é diretamente relacionada com essa estrutura. Apesar disso, tem sido observado que a similaridade estrutural nem sempre corresponde à semelhança catalítica. A terceira abordagem baseia-se no uso de características físico-químicas para representar os aminoácidos presentes na estrutura primária. Essas são calculadas através das interações entre todas as estruturas da proteína [27] [17] [8].

Oliveira et al. [7], Borro et al. [3], Dias [10] e Rodrigues [9] utilizaram as características presentes no Sting_DB [21], um dos maiores bancos de dados de características físico-químicas, estruturais e biológicas de proteínas. Para a seleção das melhores características foram utilizados recursos da estatística e métodos de mineração de dados. Entretanto, não se pode garantir que as características selecionadas por eles seja o melhor subconjunto presente na base de dados. O objetivo deste trabalho é desenvolver uma metodologia que, baseada em um algoritmo genético (AG) capaz de selecionar características físico-químicas representativas, dado um conjunto de proteínas. O intuito é predizer a suas funções utilizando aprendizado de máquina. A técnica de aprendizado escolhida para a realização deste trabalho foi a Máquina de Vetor Suporte (SVM -*Support Vector Machine*), amplamente utilizada para a resolução de problemas desta natureza [13] [5] [24] [20] [4]. Essa possui uma boa capacidade de generalização, e uma teoria bem fundamentada dentro de conceitos matemáticos e estatísticos.

Na Seção 2 são detalhados os principais conceitos utilizados neste trabalho, e que são primordiais para o entendimento do mesmo. Os trabalhos relacionados estão descritos na Seção 3. A Seção 4 detalha a metodologia proposta. Os resultados são apresentados na Seção 5. Finalmente, a Seção 6 traz as conclusões e considerações finais.

2 Conceitos Básicos

2.1 Proteínas

As reações químicas que ocorrem no interior dos aminoácidos formam uma proteína, e determina como essa proteína esta organizada tridimensionalmente. Para compreender as propriedades de uma proteína, é necessário descrever como os aminoácidos são formados. Sua estrutura é constituída por um carbono central ou carbono alfa, que se liga a quatro grupos: o grupo amina (NH2), o grupo carboxílico (COOH), um hidrogênio e uma cadeia que é denominada cadeia lateral, essa cadeia é onde eles se diferem [23] [18] [19]. Proteínas possuem quatro níveis de organização: primária, secundária, terciária, e quaternária.

- **Estrutura Primária** é a sequência de aminoácidos ao longo de sua cadeia.
- **Estrutura Secundária** consiste na relação espacial entre os aminoácidos que estão próximos na estrutura primária. Nas proteínas, as unidades básicas da estrutura secundária são: as alfas hélices e as folhas betas.
- **Estrutura Terciária** são como os átomos de uma cadeia polipeptídica estão organizados em espaço tridimensional.
- **Estrutura Quaternária** são as interações entres as diversas cadeias de aminoácidos presentes em uma proteína.

Enzimas são proteínas catalisadoras, ou seja, que aceleram a velocidade das reações bioquímicas. A união internacional de bioquímica e biologia molecular (*International Union of Biochemistry and Molecular Biology*, IUBMB), desenvolveu um sistema no qual as enzimas são divididas em seis classes, que podem ser vistas na Tabela 1.

Tabela 1. Classes de enzimas e suas respectivas funções

EC	Classe	Função
1	Oxidoreductase	Reações de transferências de elétrons.
2	Transferases	Transferência de grupos entre duas moléculas.
3	Hidrolases	Reações de hidrólise de várias ligações covalentes.
4	Liases	Quebra de ligações covalentes e remoção de moléculas de água, amônia e gás carbônico.
5	Isomerases	Modificações de uma única molécula, sem partição de outra.
6	Ligases	Reações de formação de uma nova molécula a partir da ligação entre outras duas.

Fonte: Elaborado pelo autor

2.2 Máquinas de Vetores Suporte

Atualmente, existem diversas técnicas de aprendizado de máquina. Dentre estas técnicas, a que mais tem se destacado é a SVM, propostas por Vapnik [6]. Essas procuram minimizar a probabilidade de se classificar equivocadamente padrões ainda não vistos, através de uma distribuição de probabilidade [25]. A SVM mapeia os exemplos de treinamento para um espaço de maior dimensionalidade, partindo do pressuposto que para altas dimensões do espaço todos os problemas se tornam separáveis. A partir disso, encontra os vetores de suporte sobre as bordas de um hiperplano, que distingue as diferentes características maximizando a distância entre as classes, o que torna o problema uma decisão entre essas classes. Na Tabela 2 são apresentados os principais *kernels* utilizados no processo de classificação da SVM [26]. Esses possuem o objetivo de formular o plano de melhor separação dos dados. Neste trabalho foi utilizado a LibSvm [26], que é uma biblioteca que utiliza máquinas de vetores de suporte para a classificação e regressão de padrões. Ela dá suporte a vários sistemas operacionais, e contém implementações em várias linguagens. Sua utilização é facilmente integrada com o ambiente do WEKA [12], utilizado neste trabalho.

Tabela 2. *kernels* utilizados na SVM

Tipo da Função	Equação	Parâmetros
Linear	$X_i^T X_j$	-
Polinomial	$(\gamma X_i^T X_j + r)^d$	r, d
Gaussiano(RBF)	$exp(\gamma \|X_i - X_j\|^2)$	γ
Sigmoidal	$\tanh(\gamma X_i^T X_j + r)$	γ, r

Fonte: Hsu et al. [26]

2.3 Base de dados e Extração de características

O conjunto de proteínas utilizado neste trabalho foi o mesmo utilizado por Dobson [11], Oliveira et al. [8], Borro et al. [3] e Rodrigues [9]. Esse conjunto foi utilizado para posteriormente ser possível a comparação dos resultados obtidos. As proteínas foram extraídas do *Protein Data Bank*(PDB) [2], o maior e mais completo repositório de proteínas existente. A Tabela 3 apresenta a quantidade de enzimas utilizadas neste trabalho. Outro banco de dados utilizado foi o Sting_DB, desenvolvido pelo laboratório de Biologia Computacional da Embrapa Informática. Todas as características utilizadas neste trabalho, na etapa de seleção, foram extraídas deste repositório.

Tabela 3. Quantidade de enzimas extraídas

Classe de Enzimas	Quantidade
Oxidoreductase	76
Transferases	120
Hidrolases	161
Liases	60
Isomerases	57
Ligases	18

Fonte: Dados da pesquisa

As características extraídas do Sting_BD podem ser agrupadas em:

- **Evolutivas** são calculadas por meio da mudança das proteínas, ou seja, o quanto a sequência delas evoluíram ao longo do tempo.
- **Contatos inter atômicos** são calculados por meio do contato entre os átomos presentes em cada resíduo da proteína.
- **Físico-químicas** são obtidas por meio das atrações ocorridas pelos diversos tipos de ligações entre os aminoácidos.
- **Estrutura geométrica** são calculadas por meio da estrutura tridimensional da proteína.
- **Superfície** são calculadas através das cavidades contidas na superfície de uma proteína, de onde os ligantes se acoplam.

2.4 Seleção de Características

A seleção de característica é uma importante etapa de pré-processamento, pois a partir dela são escolhidos os atributos que servirão de entrada para o algoritmo de aprendizado. Esta etapa tem por finalidade a escolha de um subconjunto de atributos, dentre todos os disponíveis. Um atributo é considerado relevante se com ele a máquina consegue criar uma capacidade de discernimento entre as diferentes classes. Dentre os algoritmos utilizados para a seleção de características, podemos citar os algoritmos exponenciais, que fazem uma busca exaustiva no conjunto de soluções para determinar qual a melhor. Esse método é inviável, pois o tempo computacional cresce exponencialmente. Outra técnica é a seleção sequencial, que utiliza heurísticas como a *forward selection* e a *backward elimination* [7]. Sua desvantagem é não levar em consideração a interação das características. Dois algoritmos evolucionários fazem parte dos métodos de busca randômicos: Algoritmo Genético [16] e o *particle swarm optimization*(PSO) [15]. A vantagem dos algoritmos randômicos em relação aos métodos sequenciais é que eles tratam a questão das interações entre as características.

2.5 Algoritmo Genético

O AG, inicialmente proposto por Holland [14], baseia-se na teoria da evolução de Darwin e pertence ao grupo dos algoritmos evolucionários. Esses partem de

uma população inicial, em que indivíduo é associado a uma solução em potencial dentro de todo o seu conjunto de soluções. Cada indivíduo possui um valor de *fitness* que determina o quanto um indivíduo está adaptado ao ambiente e suas chances de sobrevivência. Após um processo de seleção (baseado nesta *fitness*), os indivíduos escolhidos para permanecerem na população são então recombinados através dos operadores genéticos, cruzamentos e mutações. A partir daí o processo se repete, esperando-se obter um melhor valor de *fitness* a cada população gerada. O Algoritmo 1 descreve o funcionamento de um algoritmo genético simples. O uso do AG é altamente justificável para o problema proposto neste trabalho devido a sua capacidade de gerar amostras representativas do conjunto de soluções em poucas execuções.

Entrada: Conjunto de características
Saída: Melhor indivíduo
1 Inicializar população aleatória;
2 **repita**
3 | avaliar os indivíduos da população;
4 | executar seleção;
5 | executar cruzamento;
6 | executar mutação;
7 **até** *(Alcançar critério de convergência)*;

Algoritmo 1: Algoritmo Genético Simples

3 Trabalhos Relacionados

Dobson e Doig [11] propuseram uma metodologia para a predição de enzimas a partir de seu dados estruturais. O método tem o objetivo de classificar as enzimas em uma das seis super famílias, a partir de um grupo de atributos estruturais da proteína. O valor médio da precisão das classes analisadas foi de 35%, usando a SVM.

Oliveira et al. [8] propuseram uma metodologia, que tem o objetivo de melhorar o processo de seleção de parâmetros, para aumentar a precisão do modelo de classificação de proteínas. Para contornar o problema da dimensão do vetor de características passadas para o classificador, foi utilizada a Transformada Discreta do Cosseno (*Discrete Cosine Transform* - DCT). Outro problema encontrado foi o desbalanceamento das classes utilizadas, já que a maior classe possui 160 proteínas e a menor apenas 18. Com o intuito de resolver esse problema foram utilizados recursos da estatística, como a correlação de variáveis e a amostragem com reposição. Com a metodologia proposta foi possível obter uma precisão de aproximadamente 70%.

Dias [10] utilizou a SVM para a predição de função de proteínas. Para representar os aminoácidos das proteínas foram utilizadas características físico-químicas contidas no Sting_DB, as mesmas utilizadas por Oliveira et al. [8] e

Borro [3]. Juntamente com essas características foram utilizados dados da GO *(Gene Ontology)*. Para extrair as características relevantes das proteínas que serviriam de entrada na rede neural, utilizou-se a DCT. Após esta etapa, foram criados vinte e três classificadores binários locais, capazes de afirmar ou negar uma função específica. Posteriormente, foi criado um classificador global que agrupa todas as funções pré determinadas. Com a metodologia utilizada, encontraram 98% de precisão e 93% de sensibilidade.

Huang e Wang [16] propuseram a utilização de um algoritmo genético para a seleção de características utilizadas na classificação de padrões em banco de dados, utilizando uma função de avaliação baseada na acurácia da SVM. Os principais problemas encontrados em se tratando de classificação são: escolha de um conjunto de características para formar os vetores de entrada da SVM e quais os melhores parâmetros do *kernel* escolhido para o aprendizado. Esses problemas estão intimamente relacionados, pois a escolha dos parâmetros está condicionada às características que serão utilizadas no processo de aprendizagem e vice versa. Devido a esse fato, é fundamental que a escolha dos parâmetros e das características sejam feitas simultaneamente. Eles compararam duas formas de seleção dos parâmetros para o *kernel* da SVM. Primeiramente foi usando o próprio algoritmo genético utilizado na seleção de característica e depois foi utilizando uma heurística chamada *grid search*. Foram testados 11 contextos diferentes, que estão disponíveis no repositório da *University of California, Irvine*(UCI). Para avaliar o resultado obtido, foram utilizadas as métricas sensibilidade, especificidade e acurácia. Com esta proposta de otimização os autores chegaram à conclusão que o algoritmo genético melhora os resultados em relação a acurácia, mas o *grid search* tem uma pequena vantagem quanto ao tempo de execução.

Rodrigues [9] propôs uma metodologia de predição de função de proteínas. Para essa tarefa foram utilizadas características extraídas do Sting_DB. O conjunto de características utilizadas foi o mesmo de Borro et al. [3], Oliveira et al. [8] e Dias [10]. Com o objetivo de solucionar o problema das dimensões desiguais das proteínas, foi utilizado a DCT que é uma técnica de extração de características. Após a seleção dos coeficientes, optou-se por adicionar a frequência dos aminoácidos de cada proteína. A última característica adicionada foi uma codificação do nível de hidrofobicidade de cada aminoácido contido na estrutura primária da proteína. Entretanto, com a adição dessa codificação as proteínas voltaram a ter tamanhos desiguais, afim de solucionar esse problema foram adicionados zeros no final do vetor de características até que todas as proteínas ficassem com o mesmo tamanho. Com a metodologia proposta foi obtido uma precisão média de 75% e sensibilidade média de 67%.

4 Metodologia

Uma das dificuldades no processo da classificação de proteínas está na forma como os aminoácidos serão codificados, uma vez que a SVM não aceita variáveis nominais. No banco de dados Sting_DB são oferecidos 338 características físico-

químicas das proteínas contidas no PDB. Esta metodologia tem como objetivo determinar por meio de um algoritmo genético quais as características mais adequadas para classificação de enzimas nas suas respectivas classes. A Figura 1 mostra a metodologia proposta neste trabalho, e as seções seguintes descrevem cada etapa.

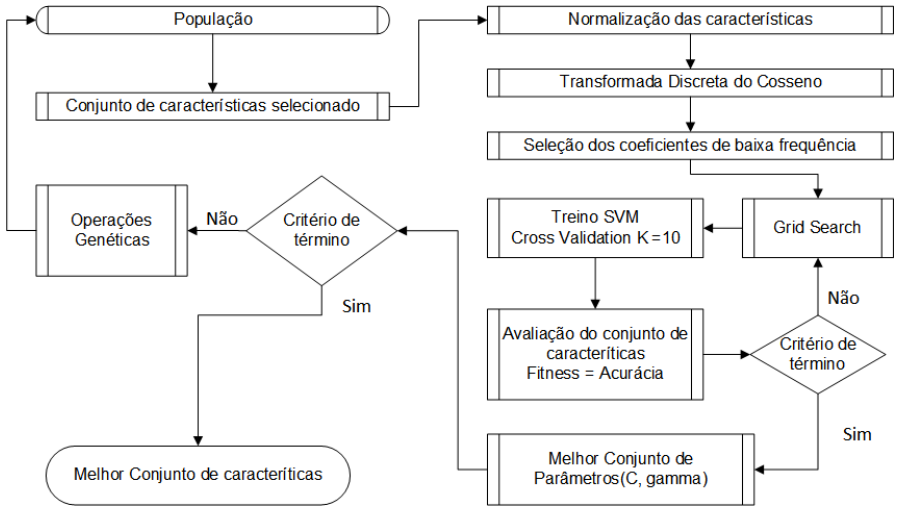


Figura 1. Metodologia Proposta

4.1 Codificação do indivíduo

No algoritmo proposto o indivíduo é representado por um vetor que possui 11 variáveis inteiras, que podem assumir valores de 1 até N, sendo N o número de características presentes no banco de dados. Cada valor que compõe este vetor representa uma característica contida nos arquivos extraídos do Sting_DB.

4.2 Normalização

Antes do processo de treinamento da SVM, são necessárias etapas para o pré processamento das características selecionadas. A normalização dos dados, tem o objetivo de evitar que atributos com valores de intervalo maior dominem aqueles que estão em intervalos menores. Outra finalidade é a de evitar grandes dificuldades durante os cálculos numéricos realizados pelo *kernel* escolhido da SVM. Cada característica é normalizada, e assim representada em um intervalo entre [0,1]. A Equação 1 é usada para o processo de normalização, onde X é o valor original da característica, *max* e *min* é o maior e menor valor dessa característica, respectivamente.

$$X' = \frac{X - max}{max - min} \tag{1}$$

4.3 Transformada Discreta do Cosseno

Proteínas possuem quantidades diferentes de aminoácidos, portanto, caso os aminoácidos fossem codificados sequencialmente, os vetores de entrada para a SVM possuiriam tamanho diferentes. Assim, é necessário fazer o uso de alguma técnica para contornar essa situação, pois a SVM só aceita vetores de características com a mesma dimensão. Para solucionar o problema de dimensionalidade, foi usada a DCT [1], cuja fórmula está descrita na Equação 2. A DCT é uma técnica de extração de características que transforma os dados do domínio do tempo para o domínio frequência. Neste processo, as frequências são colocadas em ordem decrescente, onde as primeiras frequências são as que guardam informações mais relevantes contidas no seu conjunto de dados. As frequências baixas são consideradas ruídos.

$$C_k = \alpha_k \sum_{n=0}^{N-1} X_n \cos\left[\frac{\pi}{N}\left(n + \frac{1}{2}\right)k\right], n > 0 \tag{2}$$

onde $\alpha_k = \frac{1}{\sqrt{N}} para, k = 0$ e $\alpha_k = \sqrt{\frac{2}{N}} para, k = 1...N$

4.4 Validação cruzada e Grid Search

A função de avaliação determina o quanto a solução é apropriada para a resolução de um problema. Neste trabalho o objetivo foi maximizar a taxa média de acerto da classificação das proteínas. Para isso, foi utilizada o processo de validação cruzada e a heurística *grid search*.

O *kernel* mais utilizado, que inclusive é utilizado como padrão na biblioteca LibSvm, é o RBF (*Radial Basic Function*). Com ele é possível resolver problemas multi classe, através do mapeamento das características para um espaço de maior dimensão. No *kernel* RBF, dois parâmetros podem ser variados para se obter um melhor aprendizado do classificador, são eles: γ (gamma) e C (custo). O parâmetro γ tem o objetivo de determinar a largura da curva gaussiana, já o parâmetro C é a penalidade da função, um tipo de tolerância aos erros presentes em um problema de classificação. O *grid search* busca otimizar a classificação através da execução da SVM e permite a análise dos resultados obtidos com os ajustes dos parâmetros, testando sequências exponenciais para eles.

Neste trabalho, o processo de otimização dos parâmetros foi utilizado juntamente com a validação cruzada, que consiste em uma técnica estatística para o particionamento do conjunto de teste e treino. Para o particionamento do conjunto foi usado $k = 10$. Com isso a base de dados é dividida em k subconjuntos, onde k-1 são usados para o processo de construção do modelo, onde é realizado

o treinamento e o restante é utilizado para teste. Esse processo é feito repetidamente k vezes, e a cada uma é usado um conjunto de teste diferente. Dessa forma essa técnica procura otimizar o aprendizado da máquina para que ela possa aprender o máximo possível para generalizar o modelo, e assim prever o comportamento dos dados para entradas futuras.

4.5 Operadores Genéticos

Uma parte muito importante para a eficiência de AGs é a escolha adequada dos operadores de cruzamento e mutação. Com a mudança da codificação ou para determinados tipos de problemas os operadores tradicionais são ineficientes.

- **Seleção:** A seleção escolhida foi a Torneio. Nela K indivíduos disputam a sua permanência na população, sendo que o indivíduo que possui maior valor de *fitness* permanece na população.
- **Cruzamento:** Devido às restrições na codificação utilizada, foi necessário utilizar um tipo de cruzamento especial, denominado cruzamento de mapeamento parcial (*Partially Mapped Crossover*- PMX). Consiste em um cruzamento de dois cortes, onde os valores são mapeados para garantir que não haja repetição entre as variáveis do indivíduo. A repetição de variáveis no processo de seleção de característica é totalmente indesejável, uma vez a sua repetição não agrega no processo de classificação. A taxa de cruzamento escolhida foi de 0.65.
- **Mutação:** Para manter a variabilidade da população foi escolhido uma taxa de mutação de 0.01.

4.6 Solução retornada

Ao final da execução da metodologia proposta, obtêm-se um subconjunto de características. Essas fizeram com que o indivíduo se adaptasse melhor ao ambiente, ou seja, que produziu a melhor acurácia no processo de classificação.

4.7 Métricas de avaliação

As métricas de avaliação abaixo serão utilizadas para avaliar o melhor conjunto de características obtido com a metodologia.

- **Acurácia** é a proporção das instâncias de teste que são classificadas corretamente pelo classificador

$$Acurácia = \frac{VP + VN}{VP + FP + VN + FN} \tag{3}$$

- **Precisão** número de instância classificadas como positivas são realmente positivas

$$Precisão = \frac{VP}{VP + FP} \tag{4}$$

– **Sensibilidade** avalia os acertos em cada classe.

$$Sensibilidade = \frac{VP}{VP + FN} \tag{5}$$

– **Especificidade** avalia se as instância que não são de uma determinada classe foram realmente classificados como não pertencentes àquela classe.

$$Especificidade = \frac{VN}{VN + FP} \tag{6}$$

5 Resultados e Discussões

Para a seleção da melhor quantidade de coeficientes para representar a proteína, foram realizados experimentos com as características utilizadas por Oliveira et al. [8], Dias [10] e Borro et al. [3]. Observa-se na Figura 2 que a taxa de precisão cresce à medida que a quantidade de coeficientes selecionados aumentaaumenta; em compensação, a taxa de sensibilidade decresce. Uma vez que as duas métricas são importantes, foi necessário escolher uma quantidade de coeficientes na qual obtivesse o melhor equilíbrio entra elas. Portanto, a quantidade de coeficientes escolhida foi 75.

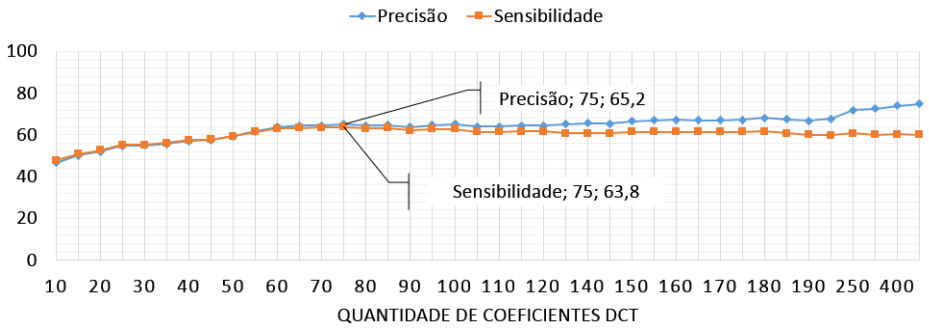


Figura 2. Média da precisão e sensibilidade por número de coeficientes selecionados

Após ser feito a etapa de escolha da quantidade de coeficiente, foram feitas 7 execuções do algoritmo genético com o objetivo de encontrar um melhor conjunto de características. O número de gerações e indivíduos foi limitada em 50 e 10, respectivamente. Essa limitação deve-se ao fato da enorme demanda de tempo do algoritmo a medida que se aumenta esses parâmetros. Após o término das execuções, o conjunto de características que obteve melhor acurácia foi selecionado. Os parâmetros da SVM selecionados pelo *grid search* para esse conjunto de características foram C: 32.0, γ: 4.8828125E-4.

As características selecionadas pelo AG estão descritas a seguir:

- ***3DEntropyINT(8)*** número de resíduos da mesma cadeia que são utilizados para calcular a entropia relativa, em seguida, dividida pelo volume da esfera.
- ***3DEntropyLHAsw(9,7)*** entropia relativa dos aminoácidos encontrados no domínio de um raio igual a 7, e utilizando uma janela deslizante com 9 resíduos. Em seguida, dividida pelo volume da esfera.
- ***ACCC*** acessibilidade do resíduo ao solvente calculada com as proteínas unidas.
- ***ACCI*** acessibilidade do resíduo ao solvente calculada para cada proteína isolada.
- ***ContactsEnergyAllsw(true,3)*** média da energia de todos os contatos do resíduo e 2 vizinhos, considerando também contatos com moléculas de água.
- ***Curvature*** curvatura média do resíduo quando as proteínas estão em complexo.
- ***DensityCAsw(3,6)*** densidade calculada com esfera centrada no Cα , raio 6, utilizando uma janela deslizante com 3 resíduos.
- ***DensityCAsw(5,4)*** densidade calculada com esfera centrada no Cα , raio 4, utilizando uma janela deslizante com 5 resíduos.
- ***DensityCAsw(9,5)*** densidade calculada com esfera centrada no Cα , raio 5, utilizando uma janela deslizante com 9 resíduos.
- ***DensityLHAsw(3,4)*** densidade calculada com esfera centrada no LHA(último átomo da cadeia lateral com exceção do hidrogênio) , raio 4, utilizando uma janela deslizante com 3 resíduos.
- ***DistanceCG*** representa a distância entre o Cα de cada resíduo e o centro de massa da cadeia (baricentro).

A Tabela 4 apresenta a precisão e a sensibilidade obtida em cada classe e a média de todas as classes, utilizando as características selecionadas pelo algoritmo genético. Com o objetivo de melhorar a predição e comparar com outras metodologias proposta na literatura, foi adicionada a frequência dos aminoácidos de cada proteína ao vetor de características. Os resultados desta adição são apresentados também na Tabela 4. Pode-se observar que com a inserção da frequência foi obtido uma melhora na média da precisão e sensibilidade de 3% e 4%, respectivamente.

Tabela 4. Resultados sem e com o acréscimo da frequência dos aminoácidos

	Sem Frequência		Com Frequência	
	Precisão	Sensibilidade	Precisão	Sensibilidade
Oxidoreductase	0,677	0,579	0,735	0,658
Transferases	0,598	0,658	0,621	0,725
Hidrolases	0,654	0,764	0,767	0,758
Liases	0,733	0,55	0,621	0,6
Isomerases	0,745	0,667	0,755	0,702
Ligases	0,818	0,5	0,786	0,611
Média	0,67	0,663	0,708	0,703

Fonte: Dados da pesquisa

Os resultados das métricas de avaliação são obtidos através dos dados contidos na Matriz de Confusão, exibida na Tabela 5. Essa apresenta o número de classificações corretas sobreposto ao número de classificações preditas para cada classe.

Analisando a matriz de confusão, podemos perceber que algumas instâncias foram classificadas em classes diferentes. Uma explicação possível pode ser devido ao desbalanceamento entre as classes. A maior parte dos falsos positivos ocorreu para as duas classes majoritárias, que são as Hidrolases e as Transferases, respectivamente.

Tabela 5. Matriz de Confusão

	Oxi	Tra	Hid	Lia	Iso	Lig
Oxidoreductase **(Oxi)**	50	9	11	1	4	1
Transferases **(Tra)**	6	87	13	10	3	1
Hidrolases **(Hid)**	5	26	122	6	1	1
Liases **(Lia)**	1	14	7	36	2	0
Isomerases **(Iso)**	6	3	5	3	40	0
Ligases **(Lig)**	0	1	1	2	3	11

Fonte: Dados da pesquisa

A Tabela 6 apresenta os resultados obtidos com as métricas apresentadas na Seção 4.7. A sensibilidade obteve uma média de 70%. Novamente a possível explicação para este fato é desbalanceamento entre as classes. As instâncias tendem a serem classificadas nas classes majoritárias, devido ao grande número de exemplos que o classificador possui para aprender sobre elas.

A precisão obteve uma média de 71%, onde a maior taxa de precisão veio da classe ligase. Isso ocorreu, pois dentre as 14 instâncias classificada como sendo dessa classe, 11 estavam corretamente classificadas. As classes que obtiveram a

Tabela 6. Métricas utilizadas para a avaliação do melhor conjunto

	Precisão	Sensibilidade	Acurácia	Especificidade
Oxidoreductase	0,73	0,66	0,71	0,76
Transferases	0,62	0,72	0,64	0,56
Hidrolases	0,77	0,76	0,76	0,77
Liases	0,62	0,60	0,62	0,63
Isomerases	0,75	0,70	0,74	0,77
Ligases	0,79	0,61	0,72	0,83
Média	0,71	0,70	0,70	0,72

Fonte: Dados da pesquisa

menor taxa de precisão foi a transferase e liase, obtendo precisão de 62%. Na transferase, das 140 instâncias classificadas como pertencentes a ela, somente 87 pertenciam de fato. Já para as liases, a taxa de 62% deve-se ao fato do número significativo de instâncias das classes majoritárias classificadas erroneamente como sendo da liase. Isso pode ser atribuído novamente ao desbalanceamento das classes. Como o número de instâncias da classe liase dada ao classificador era menor, ele não conseguiu aprender o suficiente para diferenciá-la das classes majoritárias. Para a acurácia, obtivemos uma média de 70% e a especificidade obteve uma média de 72%.

A Figura 3 apresenta a comparação da precisão entre o conjunto de características encontrado com a metodologia proposta e as características usadas por Rodrigues [9]. Pode-se observar que, a metodologia proposta obteve uma melhora significativa em todas as classes de enzimas.

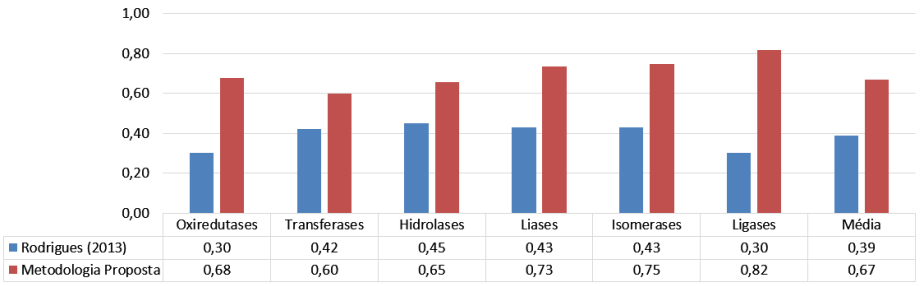


Figura 3. Taxa da precisão somente do conjunto de características

A Figura 4 apresenta o resultado das características com a inserção da frequências dos aminoácidos presente na estrutura primária. Na média a metodologia proposta ainda foi melhor, mas devido alta taxa de precisão obtida por Rodrigues [9] nas classes isomerase e ligase o ganho na média não foi tão alto quanto da utilização das características apenas.

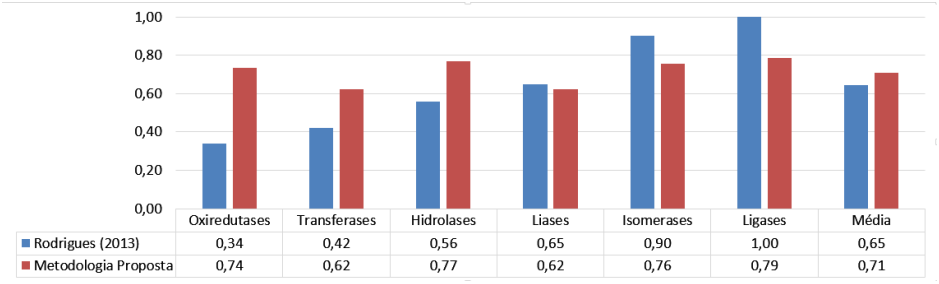


Figura 4. Taxa da precisão com o conjunto de características adicionando a frequência de aminoácidos

Como citado anteriormente, Rodrigues [9] utilizou outra característica além das contidas no Sting_DB. Com a inserção dessa característica o problema da dimensão entre as proteínas voltou a ocorrer. Para a resolução deste problema foram inseridos zeros ao final do vetor de característica até todas as proteínas ficarem com tamanhos iguais. Isso não foi feito no presente trabalho, pois o objetivo era utilizar somente as características presentes no Sting_DB. Além disso a inserção de zeros pode ocasionar *overfitting* no classificador, ou seja, ele pode atribuir a classe devido a quantidade de zeros que as proteínas delas possuem. Portanto, o classificador se ajusta em demasiado ao conjunto de amostra, não criando uma boa representação da realidade.

A Figura 5 apresenta a comparação entre a taxa de precisão obtida pelo vetor de características com a adição da codificação por Rodrigues [9] e a metodologia proposta. Pode-se observar que o resultado obtido na metodologia proposta é bastante similar ao obtido na metodologia de Rodrigues [9], entretanto vale destacar que a metodologia proposta não adicionou nenhuma características além das contidas no Sting_DB.

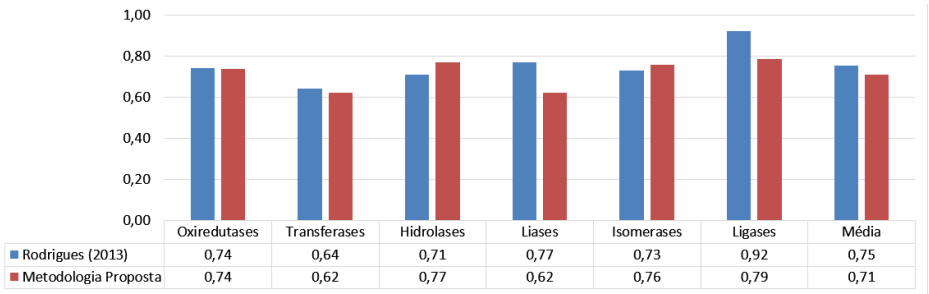
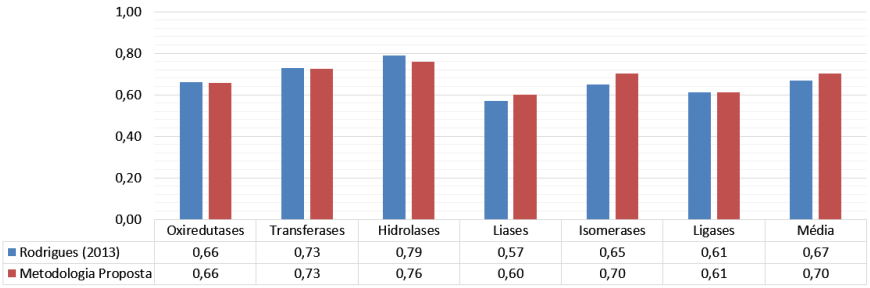


Figura 5. Comparação entre a precisão das metodologias

A Figura 6 apresenta a comparação entre a sensibilidade obtida em cada classe de enzima. Pode-se observar que a metodologia proposta foi em média 3% melhor do que a metodologia de Rodrigues [9].



	Oxiredutases	Transferases	Hidrolases	Liases	Isomerases	Ligases	Média
Rodrigues (2013)	0,66	0,73	0,79	0,57	0,65	0,61	0,67
Metodologia Proposta	0,66	0,73	0,76	0,60	0,70	0,61	0,70

Figura 6. Comparação entre a sensibilidade das metodologias

6 Conclusões

Neste trabalho foi proposta uma metodologia de seleção de características para o problema de predição de função de proteínas. Esta metodologia tem como base uma estratégia evolucionária, que é usada no processo de seleção das características. Juntamente com esse processo, também foram utilizados recursos matemáticos para a resolução do problema de dimensionalidade das proteínas e aprendizado de máquina para classificá-las em suas devidas classes. Os resultados obtidos com a metodologia proposta foram satisfatórios e são tão bons ou melhores do que as referências citadas. Acredita-se que com o aumento do número de gerações e indivíduos no AG, ele possa achar características ainda melhores para a distinção entre as classes.

Como trabalhos futuros, propomos o desenvolvimento de um AG paralelo e distribuído. Com isso poderíamos aumentar o espaço de soluções abordadas pelo AG. Outro ponto importante é a função de *fitness* do AG, neste trabalho procuramos maximizar apenas a acurácia, por isso a implementação de um algoritmo genético simples atendeu ao nosso propósito. É desejável que não só a acurácia seja maximizada, mas também a precisão e a sensibilidade. Para isso pode ser utilizado o AG multiobjetivo, onde é possível efetuar mais de uma otimização ao mesmo tempo.

Referências

1. N. Ahmed, T. Natarajan, and K. R. Rao. Discrete cosine transfonn. *Computers, IEEE Transactions on*, C-23:90–93, (1974).

2. Helen M. Berman, John Westbrook, Zukang Feng, Gary Gililand, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The protein data bank. *Nucleic Acids Research*, 28:235–242, (2000).
3. Luiz César Borro, Stanley Robson de Medeiros Oliveira, Michel Eduardo Beleza Yamagishi, Adaulto L. Mancini, José G. Jardine, Ivan Mazoni, Edgar Henrique do Santos, Roberto H. Higa, Paula Regina Kuser Falcão, and Goran Neshich. Predicting enzyme class from protein structure using bayesian classification. *Genetic and Molecular Research*, 1:193–202, (2006).
4. C.Z. Cai, L.Y. Han, Z.L. Ji, and Y.Z. Chen. Enzyme family classification by support vector machines. *Proteins: Structure, Function, and Bioinformatics*, 55:66–76, (2004).
5. C.Z. Cai, W.L. Wang, L.Z. Sun, and Y.Z. Chen. Protein function classification via support vector machine approach. *Mathematical Biosciences*, 185:111–122, (2003).
6. Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, September (1995).
7. José Geraldo de Carvalho Pereira. Caracterização dos aminoácidos da interface proteína-proteína com maior contribuição na energia de ligação e sua predição a partir dos dados estruturais. Master's thesis, Universidade Estadual de Campinas, (2012).
8. Stanley Robson de Medeiros Oliveira, Michel Eduardo Beleza Yamagishi, Luiz César Borro, Paula Regina Kuser Falcão, Edgar Henrique do Santos, Fábio Danilo Vieira, Ivan Mazoni, José Gilberto Jardine, and Goran Neshich. Uma metodologia para a seleção de parâmetros em modelos de classificação. Technical report, Embrapa Informática Agropecuária, (2006).
9. Thiago Assis de Oliveira Rodrigues. Predição de função de proteínas através da extração de características da estrutura primária e secundária. (2013).
10. Ulisses Martins Dias. Predição da função das proteínas sem alinhamento usando máquinas de vetor de suporte. Master's thesis, Universidade Federal de Alagoas, Março (2007).
11. Paul D. Dobson and Andrew J. Doig. Predicting enzyme class from protein structure without alignments. *Molecular Biology*, 345:187–199, (2004).
12. Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: An update. *ACM SIGKDD Explorations Newsletter*, 11:10–18, (2009).
13. L. Y. Han, C. Z. Cai, Z. L. Ji, Z. W. Cao, J. Cui, and Y. Z. Chen. Predicting functional family of novel enzymes irrespective of sequence similarity: a statistical learning approach. *Nucleic*, 32:6437–6444, (2004).
14. John Holland. Adaptation in natural and artificial systems. *The University of Michigan Press, Ann Arbor.*, (1975).
15. Cheng Lung Huang and Jian-Fan Dun. A distributed pso–svm hybrid system with feature selection and parameter optimization. *Applied Soft Computing*, 8:1381–1391, (2008).
16. Cheng-Lung Huang and Chieh-Jen Wang. A ga-based feature selection and parameters optimization for support vector machines. *Expert System with Application*, 31:231–240, (2006).
17. Bum Ju Lee, Jong Yun Lee, Heon Gu Lee, and Keun Ho Ryu. Classification of enzyme function from protein sequence based on feature representation. In *Bioinformatics and Bioengineering, 2007. BIBE 2007. Proceedings of the 7th IEEE International Conference on*, pages 741–747, Oct (2007).
18. Albert Lehninger, David L. Nelson, and Michael M. Cox. *Lehninger Principles of Biochemistry*. Macmillan, (2008).

19. Arthur M Lesk. *Introduction to Bioinformatics*. Oxford University Press, (2005).
20. Lingyi Lua, Ziliang Qian, Yu-Dong Cai, and Yixue Li. Ecs: An automatic enzyme classifier based on functional domain composition. *Computational Biology and Chemistry*, 31(3):226–232, jun (2007).
21. A. L. Mancini, R. H. Higa, A. Oliveira, F. Dominiquini, P. R. Kuser, M. E. B. Yamagishi, R. C. Togawa, and G. Neshich. Sting contacts: a web-based application for identification and analysis of amino acid contacts within protein structure and across protein interfaces. *Bioinformatics*, 20:2145–2147, (2004).
22. Gaurav Pandey, Vipin Kumar, and Michael Steinbach. Computational approaches for protein function prediction: A survey. Technical report, Department of Computer Science and Engineering University of Minnesota, October (2006).
23. João Setubal and João Meidanis. *Introduction to computational Molecular Biology*. Pws, (1997).
24. Mohammad Tabrez Anwar Shamim, Mohammad Anwaruddin, and H.A. Nagarajaram. Support vector machine-based classification of protein folds using the structural properties of amino acid residues and amino acid residue pairs. *Bioinformatics*, 23:3320–3327, (2007).
25. Vladimir N. Vapnik. An overview of statistical learning theory. *Transactions on Neural Networks*, 10:988–999, (1999).
26. Chih wei Hsu, Chih chung Chang, and Chih jen Lin. A practical guide to support vector classification. Technical report, National Taiwan University, Taiwan, (2010).
27. Ang Yang, Renfa Li, Wen Zhu, and Guangxue Yue. A novel method for protein function prediction based on sequence numerical features. *MATCH Communication in Mathematical and in Computer Chemistry*, 67:833–843, (2012).