PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS

Codificação para Predição de Função de Proteínas *

Coding to Protein Function Prediction

Guilherme Padilha Maia[1]
Cristiane Neri Nobre (Orientadora)

Resumo

As proteínas são macromoléculas que estão presentes no organismo de todos os seres vivos e são responsáveis por diversas funções. Muitas proteínas ainda não têm suas funções conhecidas e pelas maneiras hoje existentes fica inviável o processo para descobrir suas funções. Com o avanço da bioinformática começaram a serem pesquisadas formas de predizer a função de uma proteína usando algoritmos de classificação. O objetivo deste trabalho foi desenvolver codificações que conseguissem ter um bom resultado na tarefa de predição de função. As codificações foram testadas usando o classificador Support Vector Machine (SVM) e depois comparadas. O melhor resultado obtido foi de 70,94% de acurácia média para todas as classes, utilizando apenas a quantidade de cada tipo de aminoácido presente na proteína.

Palavras-chave: Proteínas, Codificação, Predição de função, Support Vector Machine.

* Artigo apresentado ao Instituto de Ciências Exatas e Informática da Pontifícia Universidade Católica de Minas Gerais como pré-requisito para obtenção do título de Bacharel em Sistemas de Informação.

[1] Instituto de Ciências Exatas e Informática da PUC Minas, Brasil – gpmaia@pucminas.br

Abstract

Proteins are large biological molecules that are present in all living organisms and are responsible for a vast array of functions. Many proteins yet have not known functions and the currently existing forms to discover their functions not is a feasible process. With the advancement of bioinformatics began to be researched ways to predict the function of a protein using classifiers algorithms. The objective of this study was to develop encodings that could have a good result in the function prediction task. The encodings were tested using the classifier Support Vector Machine (SVM) and then compared. The best result was found of correct classification was of 70.94% of the base using only the amount of each type of amino acid present in each protein.

Keywords: Proteins, Encoding, Function prediction, Support Vector Machine.

1 INTRODUÇÃO

Os seres vivos possuem em seus organismos milhares biomoléculas, ou macromoléculas, denominadas proteínas, cuja formação é composta por uma sequência de aminoácidos determinados por sua sequência genética.Esta sequência de aminoácidos é composta por vinte aminoácidos, responsáveis por executar as mais diversificadas funções dentro do organismo, como catalisação de reações metabólicas, resposta a estímulos e o transporte de moléculas de um local para outro, além de participarem de quase todos os processos celulares do organismo (UZUNIAN E BIRNER, 2008).

As proteínas podem ter quatro tipos de estruturas: 1) *Estrutura primária*, é a sequência linear dos aminoácidos; 2) *Estrutura secundária*, nessa estrutura as proteínas assumem duas principais formas, a alfa-hélice e folha-beta. Além destas duas principais formas ainda há outras formas que são conhecidas como laços. 3) *Estrutura terciária*, resultante do enrolamento da hélice ou da folha pregueada, assumindo sequências diferentes umas das outras, refletindo em estruturas e funções diferentes. 4) *Estrutura quaternária*, ocorre quando há duas ou mais proteínas de estrutura terciária ligadas e sua formação ocorre como as de estrutura terciária.

Devido às várias possibilidades de estruturas podemos ter formações de proteínas com as mais diversificadas funções. Apesar da tecnologia e ciência evoluírem cada vez mais, ainda existem milhares de proteínas cujas funções são desconhecidas (SUN, HU, 2007).

Para identificar a função das proteínas, a estrutura e as características são de suma importância, porém, devido à grande quantidade de características contidas em uma proteína, a classificação utilizando algoritmos pode ser comprometida (SAEYS, INZA, LARRAÑAGA, 2007). A fim de melhorar a classificação, é necessário o planejamento de uma codificação da proteína utilizando as características mais relevantes que resultem em um melhor resultado de classificação.

Devido ao grande número de proteínas cujas funções não são conhecidas e ao custo e dificuldade para identificação dessas funções, cria-se a necessidade de descoberta de um método mais eficaz para predição de sua função.

Apesar dos métodos não computacionais existentes serem capazes de fazer a predição de função, eles possuem um alto custo e o tempo para solução elevado. Existem ainda codificações que juntamente com algoritmos computacionais são capazes de realizar a predição de função.

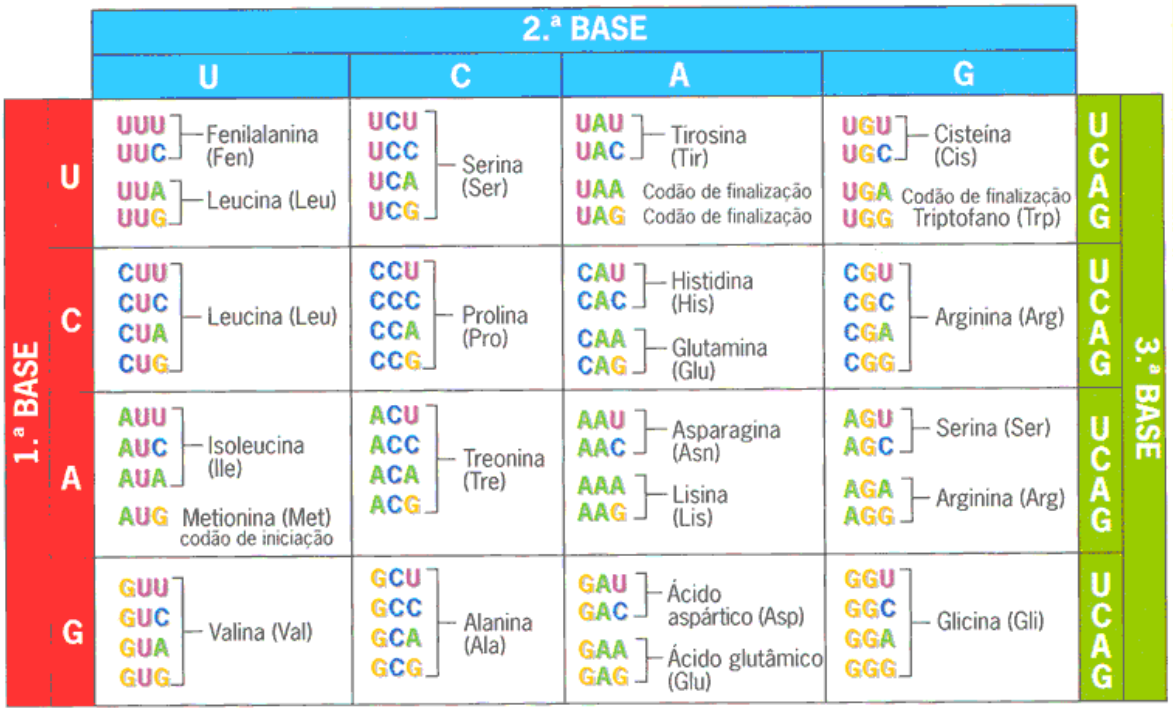
Com objetivo de encontrar uma solução melhor para o problema propõem se o estudo de técnicas utilizadas na área, através da Revisão Sistemática da Literatura e também planejar e implementar uma codificação para as proteínas de maneira que com um baixo custo de processamento consiga realizar a predição de função usando o classificador Support Vector Machine (SVM).

2 REFERENCIAL TEÓRICO

2.1 Proteínas

As proteínas são formadas por ligações de aminoácidos. Existem vinte tipos de aminoácidos, e cada aminoácido é formado por uma trinca de ácidos nucleicos, que podem estar ligados de maneiras diferentes, porém, formando um mesmo aminoácido. O código genético apresentado na Figura 1 ilustra o aminoácido resultante da cada uma das ligações de ácidos nucleicos possíveis.

Figura 1 – Código genético



Fonte: BORGES-OSÓRIO E ROBINSON, 2013.

2.2 Estrutura das proteínas

As proteínas possuem várias formas que são classificadas em quatro tipos de estruturas (Figura 2).

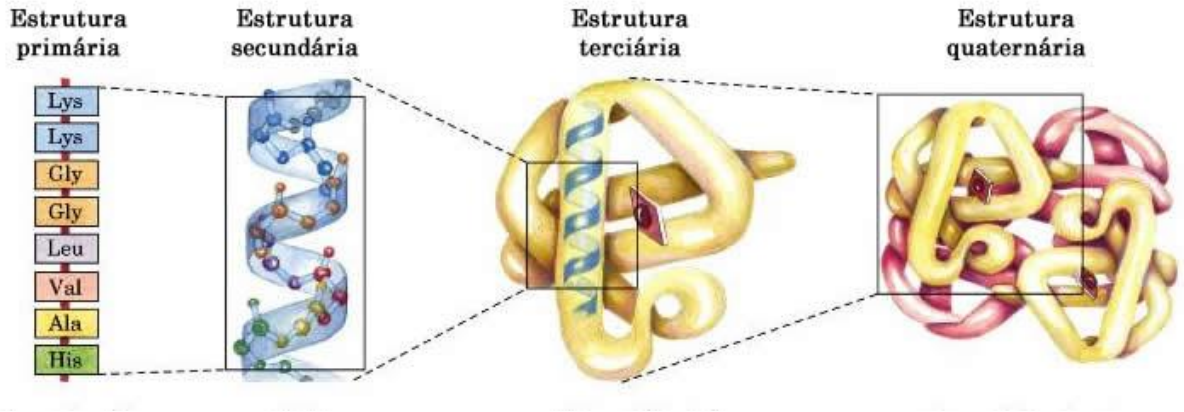
Estrutura primária: É o nível de estrutura mais simples e mais importante. Esta sequência de aminoácidos é determinada geneticamente. Nesta estrutura as moléculas de proteínas são como fios esticados.

Estrutura secundária: Podem ter duas formas básicas: a alfa-hélice e a folha-beta. Esta estrutura é mantida por pontes de hidrogênio entre átomos de aminoácidos que estão próximos ao longo da cadeia.

Estrutura terciária: É resultante da atração entre radicais de aminoácidos localizados em regiões distantes das moléculas, levando dobramento da estrutura secundária sobre si mesma, dando a molécula um aspecto esférico.

Estrutura quaternária: É originada da união de duas ou mais cadeias peptídicas, iguais ou diferentes formando uma única molécula proteica.

Figura 2 – Estruturas das proteínas



Adaptado de: LEHNINGER, 2002.

2.3 Enzimas

As enzimas são substâncias orgânicas que desempenham uma função catalisadora no interior e fora das células vivas. São divididas em seis principais classes:

Oxidoredutases: Possuem a função de óxido-redução nos sistemas biológicos, relacionadas aos processos de respiração e fermentação.

Transferases: São enzimas responsáveis pela transferência de grupos de um composto para outro.

Hidrolases: Realizam a divisão de material orgânico através da utilização da água.

Liases: São as enzimas que atuam removendo elementos de água, amônia ou dióxido de carbono do substrato.

Isomerases: São enzimas que catalisam reações de isomerização.

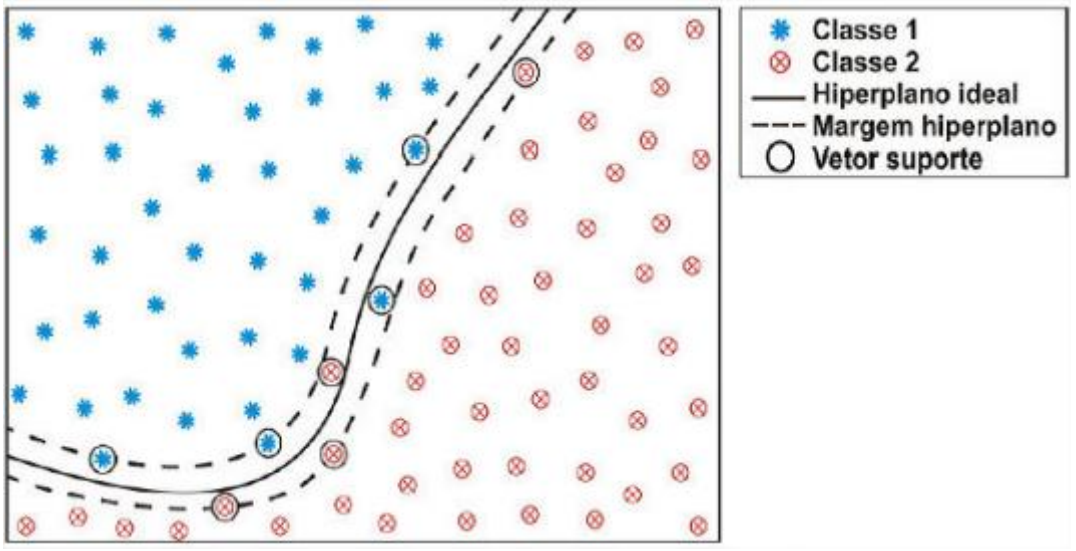
Ligases: São enzimas que causam a degradação da molécula de ATP (Adenosina Trifosfato), usando a energia liberada nesta reação para síntese de novos compostos, unindo duas moléculas.

2.4 Support Vector Machine

A Support Vector Machine – SVM - (ou Máquina de Vetor de Suporte) é um dos mais populares algoritmos de máquinas de aprendizado. Ele foi introduzido por Vapnik em 1992 e a partir daí tornou-se cada vez mais popular, principalmente por prover um bom desempenho de classificação em comparação a outros algoritmos de classificação (MARSLAND, 2009).

Para classificar determinados dados, a SVM necessita de uma base de dados de treinamento e uma base de teste. A base de treinamento deve conter dados já classificados corretamente, para que ele possa aprender a classificar de acordo com a codificação usada e, a partir dela classificar os dados desconhecidos. O SVM seleciona um hiperplano que maximiza a margem de separação entre as amostras das classes dos demais hiperplanos, conforme Figura 3. Se os exemplos não forem perfeitamente separáveis no espaço de características pela função de *kernel*, um parâmetro C é usado para os conflitos de treinamento (o número de exemplos de treinamento erroneamente classificados) contra a margem para os exemplos de treinamento corretamente classificados (MUPPIRALA, HONAVAR e DOBBS, 2011).

Figura 3 – Hiperplano SVM



Fonte: COSTA, ZEILHOFER E RODRIGUES, 2010.

As funções usadas para projetar os dados do espaço de entrada para o espaço de alta dimensão são chamadas de *kernels*. Diferentes *kernels* têm sido propostos na literatura, são eles: lineares, polinomiais, gaussianas (mais comumente chamadas de funções de bases radial) e sigmoides. Diferentes definições da função *Kernel* e seus respectivos parâmetros provocam alterações nos resultados fornecidos por uma SVM (SOUSA, TEXEIRA, SILVA, 2009).

3 REVISÃO SISTEMÁTICA DA LITERATURA

A Revisão Sistemática da Literatura, (SLR) (COOPER, 1998), é o processo adotado por alguns pesquisadores para verificar a qualidade da revisão ou executá-la de modo a obter fontes confiáveis para a pesquisa desenvolvida, além da investigação de trabalhos e pesquisas executados abordando o mesmo tema.

Através desta técnica obtém-se uma fonte confiável de dados dentro de determinado objetivo, para dar base e argumentação para o desenvolvimento de uma abordagem do tema.

Para esta pesquisa, foi realizado um levantamento de todos os artigos científicos publicados nos periódicos e conferências listadas abaixo, no período de 1998 a 2014, em língua portuguesa ou inglesa. Dentre os artigos encontrados, foram excluídos os que tratavam o mesmo tema, sendo mantido o mais recente.

3.1 Questões de pesquisa

Abaixo estão descritas as questões de pesquisas utilizadas na SLR.

- [Q1]: Qual a informação utilizada para predição de função de proteínas? Primária? Secundária? Terciária? Combinação delas? Características físico-químicas?
- [Q2]: Qual a codificação utilizada na informação dos aminoácidos?
- [Q3]: Qual a precisão média na predição de função de proteínas obtida através da codificação?
- [Q4]: Qual a influência da codificação no desempenho do classificador?

3.2 Processo de pesquisa

A pesquisa constitui-se na leitura e avaliação de diversos artigos científicos publicados nos periódicos e conferências listados abaixo que contenham as palavras chave: função da proteína, codificação, predição, no idioma português ou inglês.

- Advances in Bioinformatics (http://www.hindawi.com/journals/abi/)
- BMC Bioinformatics (www.biomedcentral.com/bmcbioinformatics/)
- PubMed (http://www.ncbi.nlm.nih.gov/pubmed)

Dentre os artigos científicos encontrados, foram aplicados métodos de exclusão divididos em fases, a fim de melhorar a precisão. A seguir encontram-se listados as respectivas fases com seus métodos de filtragem e eliminação por julgamento de assuntos não condizentes com o os abordados nesta pesquisa:

- Fase 1: Título
- Fase 2: Resumo
- Fase 3: Leitura diagonal
- Fase 4: Leitura Completa

Após a aplicação dos critérios de pesquisa e análise dos artigos encontrados foram obtidos os resultados em suas respectivas fases conforme apresentado na Tabela 1. Devido a quantidade dos artigos ter ficado dentro do número esperado não foi necessário a aplicação de critérios de qualidade.

Tabela1 – Quantitativo de artigos por fase de revisão

	Pesquisa inicial	1ª Fase	2ª Fase	3ª Fase	4ª Fase
Advances in Bioinformatics	6	5	4	3	2
BMC Bioinformatics	557	33	20	8	7
PubMed	792	32	17	10	6
Total	**1355**	**70**	**41**	**21**	**15**

Fonte: Elaborado pelo autor.

Os quinze artigos obtidos a partir da fase final de pesquisa estão listados na Tabela 2.

Tabela 2 – Artigos selecionados

N°	Título	Autores	Ano	Origem
1	How Good Are Simplified Models for Protein Structure Prediction?	Swakkhar Shatabda, M. A. Hakim Newton, Mahmood A. Rashid, Duc Nghia Pham, and Abdul Sattar	2014	Advances in Bioinformatics
2	Prediction of Carbohydrate-Binding Proteins from Sequences Using Support Vector Machines	Seizi Someya, Masanori Kakuta, Mizuki Morita, et al.	2010	Advances in Bioinformatics
3	Predicting protein-ATP binding sites from primary sequence through fusing bi-profile sampling of multi-view features	Ya-Nan Zhang, Dong-Jun Yu, Shu-Sen Li, Yong-Xian Fan, Yan Huang and Hong-Bin Shen	2012	BMC Bioinformatics
4	Prediction of protein-protein interactions between viruses and human by an SVM model	Guangyu Cui, Chao Fang and Kyungsook Ha	2012	BMC Bioinformatics
5	Predicting RNA-binding sites of proteins using support vector machines and evolutionary information	Cheng-Wei Cheng, Emily Chia-Yu Su, Jenn-Kang Hwang, Ting-Yi Sung andWen-Lian Hsu	2008	BMC Bioinformatics
6	*Ab Initio* prediction of mycobacteriophages protein structure and function	Chiraag D Kapadia, Claire A Rinehart	2013	BMC Bioinformatics
7	Protein Function Prediction using Text-based Features extracted from the Biomedical Literature: The CAFA Challenge	Andrew Wong, Hagit Shatkay	2013	BMC Bioinformatics

8	Prediction of enzyme function by combining sequence similarity and protein interactions	Jordi Espadaler, Narayanan Eswar, Enrique Querol, Francesc X Avilés, Andrej Sali, Marc A Marti-Renom, Baldomero Oliva	2008	BMC Bioinformatics
9	Improving protein function prediction methods with integrated literature data	Aaron P Gabow, Sonia M Leach, William A Baumgartner, Lawrence E Hunter, Debra S Goldberg	2008	BMC Bioinformatics
10	On the encoding of proteins for disordered regions prediction	Becker J, Maes F, Wehenkel L	2013	PubMed
11	Protein functional class prediction using global encoding of amino acid sequence	Li X, Liao B, Shu Y, Zeng Q, Luo J	2009	PubMed
12	Using genetic algorithms to select most predictive protein features	Kernytsky A, Rost B.	2009	PubMed
13	Computational protein function prediction: are we making progress?	Godzik A, Jambon M, Friedberg I.	2007	PubMed
14	Subcellular localization prediction with new protein encoding schemes	Oğul H, Mumcuoğu EU.	2007	PubMed
15	Improvement of protein secondary structure prediction using binary word encoding	Kawabata T, Doi J.	1998	PubMed

Fonte: Elaborado pelo autor.

3.3 Análise da revisão sistemática da literatura

Diante da pergunta Q1 “Qual a informação utilizada para predição de função de proteínas? Primária? Secundária? Terciária? Combinação delas? Características físico-químicas?”, foi identificado dentre as estruturas da proteína, a secundária é mais utilizada na codificação. Porém, também existem autores que utilizam a combinação entre duas estruturas ou uma estrutura com características físico-químicas como, por exemplo, a solubilidade.

Para responder a Q2 “Qual a codificação utilizada na informação dos aminoácidos?”, temos vários tipos de codificações. Entre as codificações uma muito usada é a Matriz de Pontuação por Posição Específica (PSSM), onde o valor do aminoácido está relacionado à sua posição. Existem também variações da PSSM, cujo intervalo de pontuação é modificado e após o resultado passar por um processo de nivelamento considerando a proximidade ao elemento central. Outra codificação utilizada é a de contagem da quantidade de vezes que cada aminoácido existe na estrutura primária da proteína, criando um vetor de mesmo tamanho para todas as proteínas.

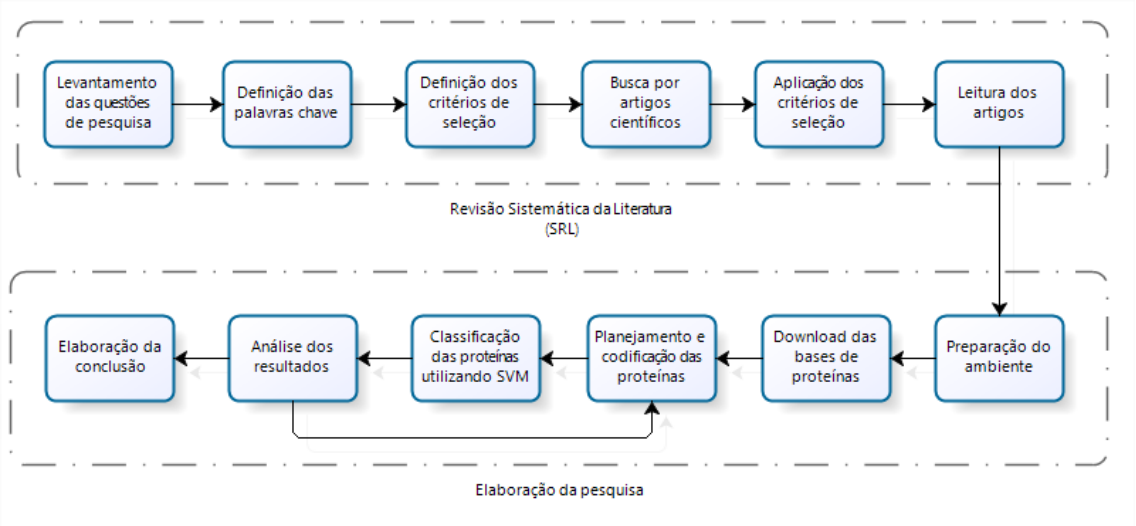
Com relação à pergunta Q3, “Qual a precisão média na predição de função de proteínas obtida através da codificação?”, é inviável dar uma resposta com valores, pois foi identificado que os resultados variam bastante diante da quantidade de variáveis existentes no processo, como os algoritmos classificados, a utilização ou não de normalização de dados entre outros.

Diante da pergunta Q4 “Qual a influência da codificação no desempenho do classificador?” foi observado que a codificação influencia diretamente no desempenho do classificador, pois é na codificação onde as informações e características das proteínas escolhidos para serem classificados estão, e devem estar de maneira que o classificador consiga distinguir as proteínas para um resultado com maior assertividade.

4 ELABORAÇÃO DA PESQUISA

A Figura 4 ilustra a metodologia usada para pesquisa e as subseções seguintes descrevem cada uma destas etapas.

Figura 4 – Metodologia Proposta



Fonte: Elaborado pelo autor.

4.1 Revisão Sistemática da Literatura

A revisão sistemática da literatura foi realizada conforme detalhado na seção anterior.

4.2 Preparação do ambiente

Para execução dos testes foi criado um ambiente de desenvolvimento utilizando a ferramenta Eclipse, onde foram feitas as codificações das proteínas através da linguagem de programação Java. Também foi configurada a ferramenta WEKA (disponível no endereço *http://www.cs.waikato.ac.nz/ml/weka/)* que contém o algoritmo classificador SVM, utilizado nos testes. Essas ferramentas foram utilizadas em uma máquina virtual rodando o sistema operacional Arch Linux.

4.3 *Download* das bases de proteínas

As proteínas foram obtidas de bancos públicos de proteínas, o Protein Data Bank (PDB) e o STING DB (disponíveis respectivamente nos endereços *http://www.rcsb.org/pdb/* e *http://www.cbi.cnptia.embrapa.br*).

O PDB foi criado em 1971, no *Brookhaven National Laboratory* sobre a liderança de Walter Hamilton. É o único repositório mundial com estruturas 3D de vários tipos de moléculas biológicas, inclusive as proteínas. O acesso é feito gratuitamente e sua base é atualizada semanalmente (BERMANT ET AL., 2000).

O STING DB é um banco de dados criado pela empresa Embrapa (Empresa Brasileira de Pesquisa Agropecuária) com base no PDB. É sincronizado semanalmente com o PDB e seus parâmetros são imediatamente calculados após a atualização. O acesso é feito gratuitamente (OLIVEIRA ET AL., 2007).

Figura 5 – Exemplo de arquivo de proteína

	A	B	C	D	E	F	G	H	I	J
1	PDB NAME = 1A6F									
2	CHAIN = A									
3										
4	OneLetterNam	Name()	Number()	PDBNumber(	3DEntropyCA	3DEntropyCA	3DEntropyCA	3DEntropyCA	3DEntropyCA	3DEntropyCA
5	A	ALA	1	0002	0,293333322	0,289999992	0,146666661	0,086666666	0,086666666	0,236000016
6	H	HIS	2	0003	0,393333346	0,49000001	0,25	0,173333332	0,186666667	0,340000004
7	L	LEU	3	0004	0,476666689	0,523333371	0,266666681	0,236666679	0,24333334	0,400000006
8	K	LYS	4	0005	0,373333335	0,563333333	0,289999992	0,283333331	0,25666666	0,476000011
9	K	LYS	5	0006	0,49000001	0,583333313	0,299999982	0,293333322	0,216666654	0,480000019
10	R	ARG	6	0007	0,526666641	0,639999986	0,336666673	0,316666663	0,25333333	0,467999995
11	N	ASN	7	0008	0,50666666	0,606666625	0,353333324	0,343333334	0,276666671	0,433999956
12	R	ARG	8	0009	0,406666666	0,513333321	0,306666672	0,306666672	0,289999992	0,436000019
13	L	LEU	9	0010	0,299999982	0,449999958	0,266666681	0,236666679	0,216666654	0,387999982
14	K	LYS	10	0011	0,356666654	0,453333288	0,233333349	0,206666663	0,196666673	0,371999979
15	K	LYS	11	0012	0,423333317	0,553333342	0,289999992	0,233333334	0,196666673	0,456
16	N	ASN	12	0013	0,540000021	0,623333395	0,326666683	0,286666662	0,230000004	0,456
17	E	GLU	13	0014	0,520000041	0,603333354	0,316666692	0,326666683	0,25999999	0,417999983
18	D	ASP	14	0015	0,376666665	0,50666666	0,299999982	0,370000005	0,289999992	0,430000007
19	F	PHE	15	0016	0,313333333	0,466666698	0,280000001	0,386666656	0,313333333	0,458000004
20	Q	GLN	16	0017	0,426666647	0,513333321	0,303333342	0,373333335	0,320000023	0,384000003
21	K	LYS	17	0018	0,483333319	0,573333323	0,356666684	0,413333327	0,359999985	0,385999978
22	V	VAL	18	0019	0,446666688	0,569999993	0,353333324	0,433333308	0,353333324	0,480000019
23	F	PHE	19	0020	0,413333327	0,593333304	0,366666675	0,423333317	0,333333343	0,518000007

Fonte: Elaborado pelo autor.

Depois de realizado o *download* dos arquivos das proteínas (Figura 5), foi feita a exclusão dos arquivos cujas cadeias eram iguais, e a separação dos arquivos por classes de proteínas, conforme apresentado na Tabela 3.

Tabela 3 – Arquivos de proteínas por classe

Classe	Número de proteínas
Hidrolases	161
Isomerases	57
Liases	60
Ligases	18
Oxidoredutases	76
Transferases	120

Fonte: Elaborado pelo autor.

4.4 Codificação das proteínas

No contexto de predição de função de proteínas, a codificação é muito importante, já que características diferentes fornecem desempenhos diferentes dos classificadores. O objetivo das codificações é gerar um arquivo com informações das proteínas que o SVM consiga processar e distinguir por classe gerando um resultado com alta assertividade. As codificações realizadas geram um arquivo com extensão ARFF, conforme Figura 6, com a quantidade de valores iguais para todos os atributos.

Figura 6 – Exemplo de arquivo ARFF

```
@relation FrequenciaAbsolutaAminoacidos
@attribute Amino-A real
@attribute Amino-C real
@attribute Amino-D real
@attribute Amino-E real
@attribute Amino-F real
@attribute Amino-G real
@attribute Amino-H real
@attribute Amino-I real
@attribute Amino-K real
@attribute Amino-L real
@attribute Amino-M real
@attribute Amino-N real
@attribute Amino-P real
@attribute Amino-Q real
@attribute Amino-R real
@attribute Amino-S real
@attribute Amino-T real
@attribute Amino-V real
@attribute Amino-W real
@attribute Amino-Y real
@attribute 'classe' {'Hidrolases','Isomerases','Liases','Ligases','Oxidoredutases','Transferases'}
@data
21,2,12,13,4,7,2,9,18,10,3,7,4,4,10,8,5,6,0,14,Hidrolases
40,2,16,19,7,27,10,23,16,32,21,9,21,12,20,24,15,28,1,4,Hidrolases
12,0,6,11,3,9,4,5,21,11,4,5,6,5,5,3,8,8,1,7,Hidrolases
6,0,4,9,5,3,3,6,17,13,1,6,2,6,11,7,4,6,0,4,Hidrolases
21,11,23,31,12,31,13,18,21,28,9,19,17,11,19,26,10,31,3,9,Hidrolases
33,4,18,12,4,20,11,18,6,18,3,14,8,9,6,28,16,13,9,14,Hidrolases
32,2,18,13,10,20,7,14,9,17,7,14,12,18,5,32,27,16,5,13,Hidrolases
15,0,6,13,2,12,6,11,8,12,3,1,4,8,5,9,6,10,0,4,Hidrolases
20,2,8,17,6,19,8,21,12,13,3,7,15,8,8,10,11,15,2,5,Hidrolases
```

Fonte: Elaborado pelo autor.

4.4.1 Codificação de frequência absoluta

A codificação de frequência absoluta foi feita através da contagem da quantidade de cada um dos tipos de aminoácidos existentes na cadeia da proteína, sendo o valor zero para os aminoácidos não existentes na proteína.

Pegando como exemplo uma proteína fictícia cuja cadeia de aminoácidos seja: **RAIAYRIAALRKKN**, os aminoácidos e seus respectivos valores seriam: R=3, A=4, I=2, Y=1, L=1, K=2, N=1 e para os demais aminoácidos não existentes na proteína o valor seria zero.

4.4.2 Codificação de frequência relativa

A codificação de frequência relativa foi feita através da proporção entre a quantidade de cada tipo de aminoácido e o número total de aminoácidos existentes em toda cadeia da proteína.

Pegando como exemplo a mesma proteína fictícia cuja cadeia de aminoácidos seja: **RAIAYRIAALRKKN**, os aminoácidos e seus respectivos valores seriam:

- R = 3/14 = 0,2142857142857143;
- A = 4/14 = 0,2857142857142857;
- I = 2/14 = 0,1428571428571429;
- Y = 1/14 = 0,0714285714285714,
- L = 1/14 = 0,0714285714285714,
- K = 2/14 = 0,1428571428571429,
- N = 1/14 = 0,0714285714285714;

E para os demais aminoácidos não existentes na proteína o valor seria zero.

4.4.3 Codificação de frequência absoluta e relativa

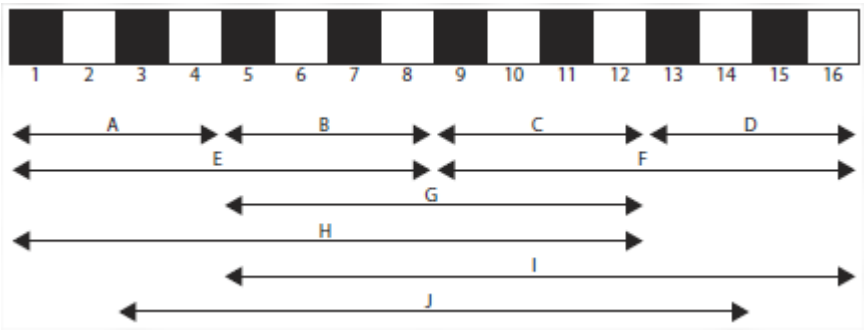
A codificação de frequência absoluta e relativa tem o objetivo de atingir o melhor resultado através da junção das duas codificações anteriores, frequência absoluta e relativa, para que possa ser aproveitado o melhor de cada uma.

A codificação é feita juntando o valor da quantidade de aminoácido e o valor da proporção da quantidade daquele aminoácido em relação a quantidade total de aminoácidos existentes na proteína.

4.4.4 Codificação de frequência absoluta e relativa por região

A codificação dos aminoácidos por frequência absoluta e relativa por região baseia na codificação acima, porém aplicada cada uma das dez regiões conforme divisão sugerida por FREITAS e SILLA (2011). A primeira divisão realizada divide a proteína em quatro regiões iguais (A, B, C, D), cada uma delas com 25% da proteína. A segunda divisão realizada divide a proteína em duas regiões (E, F), cada uma delas com 50% da proteína. A terceira divisão pega somente a região do meio da proteína (G) que compreende a extensão de 25% até 75% da proteína. As demais divisões pegam 75% da extensão da proteína, sendo que a quarta divisão pega a região (H) a partir do início da proteína, a quinta divisão pega a região (I) a partir dos 25% da extensão da proteína e a sexta divisão pega a região (J) do meio da proteína. A aposta desta divisão é conseguir um maior número de características de cada uma das regiões, as quais podem ter maior relevância na predição.

Figura 7 – Divisão da proteína em regiões



Fonte: FREITAS e SILLA, 2011.

4.4.5 Codificação de frequência absoluta, relativa e hidrofobicidade.

Nesta codificação, além da codificação de frequência absoluta e relativa acima detalhada, foi adicionada a característica físico-química de hidrofobicidade (FREITAS e SILLA). Ao invés de usar valores fixos para cada aminoácido, foram extraídas do STING DB as características de hidrofobicidade (*HydroKD(), HydroKDC(), HydroKDI(), HydroR(), HydroRC(), HydroRI*()) e realizada uma codificação para cada uma delas e também com todas juntas. A codificação foi realizada somando os valores de cada um dos aminoácidos para a característica em questão.

4.4.6 Codificação da média dos valores de cada uma das características

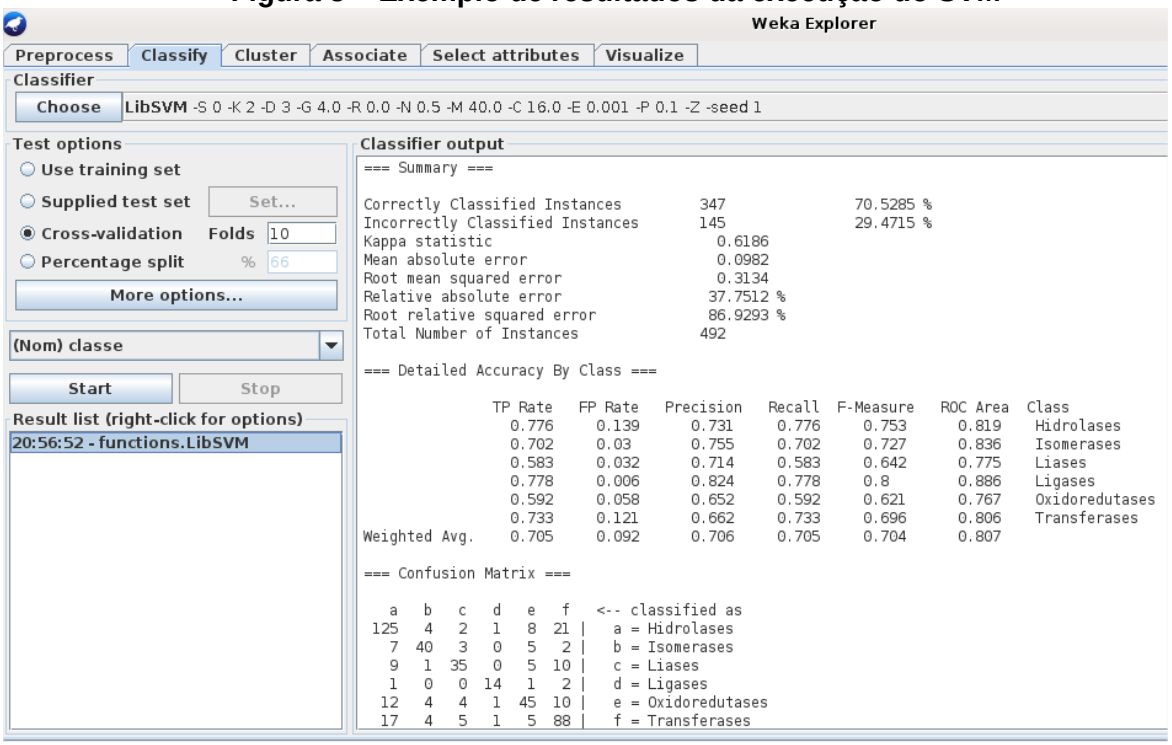
Esta codificação foi feita utilizando a média dos valores de cada uma das características da proteína, ou seja, para determinada característica tem-se o somatório dos valores de cada um dos aminoácidos dividido pela quantidade total de aminoácidos existentes na proteína. As características que possuíam valores do tipo *char* foram descartadas no momento da codificação por não ser possível realizar um somatório delas.

4.5 Classificação das proteínas utilizando SVM

Para classificação das proteínas foi utilizado o classificador SVM da ferramenta WEKA.

Após a execução, a ferramenta WEKA exibe os resultados da classificação de acordo com a Figura 8.

Figura 8 – Exemplo de resultados da execução do SVM



Fonte: Elaborado pelo autor.

4.6 Medidas de desempenho

O resultado das codificações realizadas foi avaliado através da precisão, sensibilidade e *f-measure*, obtidos da matriz de confusão gerada pelo classificador SVM.

Precisão é a taxa de instâncias corretamente classificadas como pertencentes à classe em questão, dentre todas as que foram classificadas na classe em questão, conforme a Equação 1.

$$Precisão = \frac{VP}{VP + FP} \quad (1)$$

Sensibilidade é a taxa de instâncias corretamente classificadas como pertencentes à classe em questão, dentre todas as que realmente são da classe em questão, como descrito na Equação 2.

$$Sensibilidade = \frac{VP}{VP + FN} \quad (2)$$

E a métrica *f-measure* (Equação 3) é uma média ponderada de precisão e sensibilidade.

$$F_1 = 2 \cdot \frac{Precisão \cdot Sensibilidade}{Precisão + Sensibilidade} \quad (3)$$

Sendo, Verdadeiro Positivo (VP) o número de instâncias classificadas corretamente em sua classe; Falso Negativo (FN) a quantidade de instâncias da classe em questão classificadas como de outra classe; Falso Positivo (FP) número de instâncias de outras classes classificadas como da classe em questão; Verdadeiro Negativo (VN) número instâncias corretamente classificadas como não pertencentes à classe em questão.

5 RESULTADOS E DISCUSSÕES

Antes da execução de cada teste foi utilizado o *grid search* da ferramenta *LibSVM*, disponível no endereço *http://www.csie.ntu.edu.tw/~cjlin/libsvm/.* O *grid search* foi executado para obter uma otimização dos parâmetros *gamma*, *cost* e *normalize* que consiste na busca exaustiva em subconjunto de valores manualmente especificado de um espaço de hiperparâmetros de um algoritmo de aprendizado.

A primeira codificação testada foi frequência absoluta dos aminoácidos. O SVM foi executado com os parâmetros *gamma = 4.0*, *cost = 16.0* e *normalize = true*. Foi obtido o resultado de **70,5285%** de acerto nas classificações. Para esta codificação a classe Ligase teve uma boa precisão e as classes Oxidoredutase e Transferase ficaram com a precisão baixa. A sensibilidade foi baixa para as classes Oxidoredutase e Liase, para as demais foi dentro do esperado. Na Tabela 4 podem ser vistos com detalhes os resultados obtidos.

Tabela 4 – Resultados da codificação por frequência absoluta

Codificação da frequência absoluta de aminoácidos			
Classe	Precisão	Sensibilidade	F-measure
Hidrolases	0,731	0,776	0,753
Isomerases	0,755	0,702	0,727
Liases	0,714	0,583	0,642
Ligases	0,824	0,778	0,8
Oxidoredutases	0,652	0,592	0,621
Transferases	0,662	0,733	0,696
Média	0,706	0,705	0,704

Fonte: Dados da pesquisa.

A segunda codificação testada foi frequência relativa dos aminoácidos. O SVM foi executado com os parâmetros *gamma = 4.0*, *cost = 8.0* e *normalize = true*. Foi obtido o resultado de **67,4797%** de acerto nas classificações. Para esta codificação a precisão foi baixa com exceção das classes Isomerase e Ligase, a sensibilidade só foi boa para a classe Hidrolase. A *f-measure* foi regular, com destaque positivo para as classes Hidrolase e Ligase, e destaque negativo para a classe Liase. Na Tabela 5 podem ser vistos com detalhes os resultados obtidos.

Tabela 5 – Resultados da codificação por frequência relativa

Codificação da frequência relativa de aminoácidos			
Classe	Precisão	Sensibilidade	F-measure
Hidrolases	0,672	0,764	0,715
Isomerases	0,761	0,614	0,68
Liases	0,64	0,533	0,582
Ligases	0,917	0,611	0,733
Oxidoredutases	0,667	0,684	0,675
Transferases	0,642	0,658	0,65
Média	0,679	0,675	0,673

Fonte: Elaborado pelo autor.

A terceira codificação testada foi frequência absoluta e relativa dos aminoácidos. O SVM foi executado com os parâmetros *gamma = 2.0*, *cost = 8.0* e *normalize = true*. Foi obtido o resultado de **70,935%** de acerto nas classificações, ficando melhor que as codificações com os valores individuais de frequência. A classe Ligase apresentou uma precisão acima da média e a classe Hidrolase uma sensibilidade acima da média. O resultado da sensibilidade da classe Liase ficou bem abaixo da média. Na Tabela 6 podem ser vistos com detalhes os resultados obtidos.

Tabela 6 – Resultados da codificação por frequência absoluta e relativa

Codificação da frequência absoluta e relativa de aminoácidos			
Classe	Precisão	Sensibilidade	F-measure
Hidrolases	0,717	0,82	0,765
Isomerases	0,766	0,632	0,692
Liases	0,739	0,567	0,642
Ligases	0,824	0,778	0,8
Oxidoredutases	0,681	0,618	0,648
Transferases	0,667	0,717	0,691
Média	0,712	0,709	0,707

Fonte: Dados da pesquisa.

A quarta codificação testada foi frequência absoluta e relativa por região. O SVM foi executado com os parâmetros *gamma = 0.25*, *cost = 4.0* e *normalize = true*. Foi obtido o resultado de **68,9024%** de acerto nas classificações. O destaque desta codificação foi a classe Ligase que apresentou uma precisão perfeita e a classe Liase que apresentou uma sensibilidade ruim. A *f-measure* das classes ficou bem regular, variando menos de 0,150 do melhor para o pior resultado. Na Tabela 7 podem ser vistos com detalhes os resultados obtidos.

Tabela 7 – Resultados da codificação por frequência absoluta e relativa por região

Codificação da frequência absoluta e relativa de aminoácidos por região da proteína			
Classe	Precisão	Sensibilidade	F-measure
Hidrolases	0,64	0,839	0,726
Isomerases	0,854	0,614	0,714
Liases	0,789	0,5	0,612
Ligases	1	0,611	0,759
Oxidoredutases	0,687	0,605	0,643
Transferases	0,661	0,683	0,672
Média	0,708	0,689	0,686

Fonte: Dados da pesquisa.

A quinta codificação testada, foi frequência absoluta e relativa juntamente com as seis características de hidrofobicidade. O SVM foi executado com os parâmetros *gamma = 2.0*, *cost = 4.0* e *normalize = true*. Foi obtido o resultado de **70,3252%** de acerto nas classificações. Vale ressaltar que também foram executados testes com as frequências e cada uma das características de hidrofobicidade individualmente, porém o resultado não teve alterações. Este resultado foi bem próximo do resultado anteriormente apresentado, porém com uma melhora na maioria das sensibilidades o que resultou em uma melhor *f-measure* também. Outra pequena diferença negativa foi nos valores da precisão. Na Tabela 8 podem ser vistos com detalhes os resultados obtidos com esta codificação.

Tabela 8 – Resultados da codificação por frequência absoluta + hidrofobicidade

Codificação da frequência absoluta e relativa de aminoácidos + características de hidrofobicidade			
Classe	Precisão	Sensibilidade	F-measure
Hidrolases	0,705	0,801	0,75
Isomerases	0,725	0,649	0,685
Liases	0,8	0,533	0,64
Ligases	0,867	0,722	0,788
Oxidoredutases	0,676	0,632	0,653
Transferases	0,659	0,725	0,69
Média	0,709	0,703	0,701

Fonte: Dados da pesquisa.

A sexta codificação testada foi a média dos valores de cada uma das características. O SVM foi executado com os parâmetros *gamma = 0.03125*, *cost = 512.0* e *normalize = true*. Foi obtido o resultado de **57,4775%** de acerto nas classificações. Nesta codificação a classe Hidrolase teve resultados regulares para precisão, sensibilidade e *f-measure*, as demais classes tiveram resultados ruins. Na Tabela 9 os resultados obtidos com esta codificação.

Tabela 9 – Resultados da codificação de média dos valores de cada uma das características

Codificação da média dos valores de cada uma das características			
Classe	Precisão	Sensibilidade	F-measure
Hidrolases	0,642	0,638	0,64
Isomerases	0,597	0,552	0,574
Liases	0,58	0,563	0,571
Ligases	0,524	0,5	0,512
Oxidoredutases	0,485	0,495	0,49
Transferases	0,556	0,585	0,57
Média	0,576	0,575	0,575

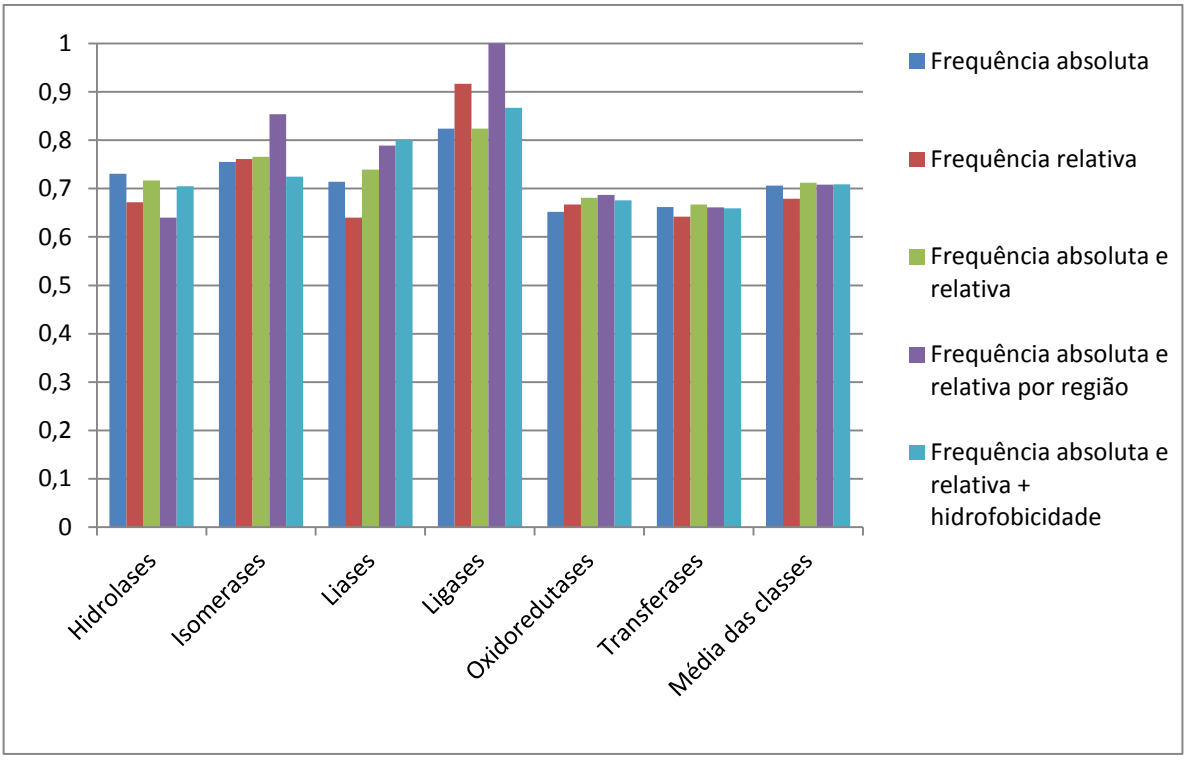
Fonte: Dados da pesquisa.

Apesar da grande quantidade de informações que esta codificação fornece ao SVM, assim como a codificação de frequência absoluta e relativa por região da proteína, o resultado foi inferior aos resultados das demais codificações, o que nos faz concluir que o importante não é quantidade de dados que é fornecida ao classificador, e sim a relevância que os dados têm na predição e a diversidade dos dados de classes diferentes.

Visto que esta codificação apresentou um resultado bem inferior comparado aos demais, a mesma não foi inserida nos gráficos comparativos.

Pela Figura 9 podemos verificar que a precisão varia para cada classe de acordo com a codificação utilizada. Na codificação de frequência absoluta e relativa por região de seis classes, três delas obtiveram a melhor precisão que as demais codificações, sendo ainda que a classe Ligases teve uma precisão de 100% com a codificação.

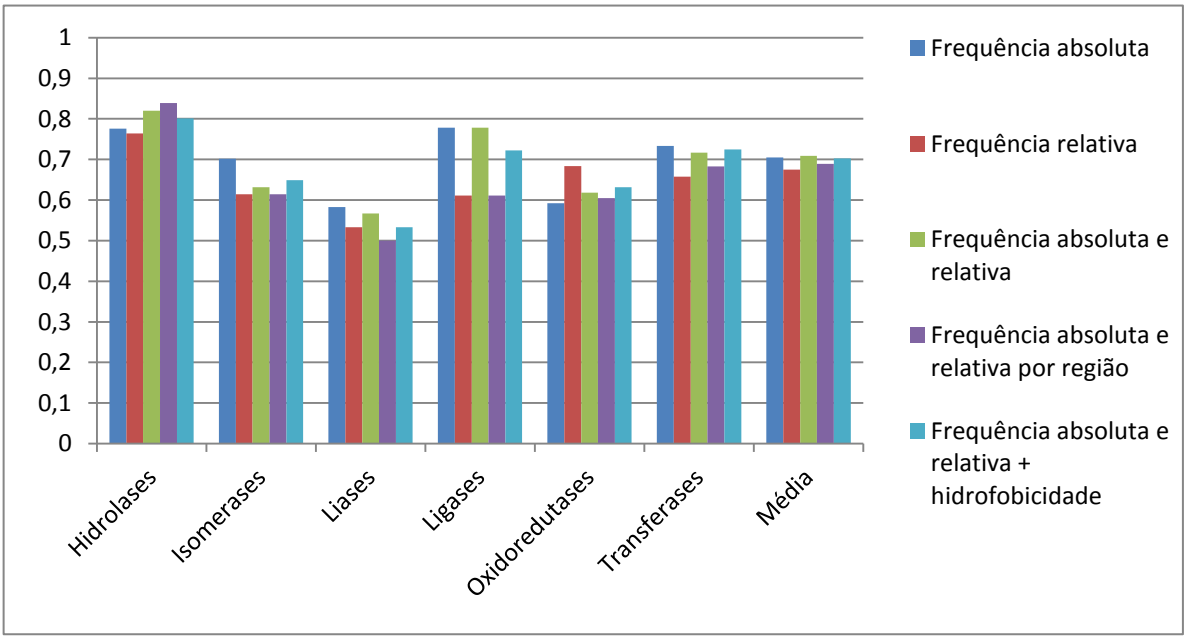
Figura 9 – Comparativa da precisão entre codificações



Fonte: Dados da pesquisa.

Para a sensibilidade, Figura 10, a codificação que teve melhor resultado foi a de frequência absoluta, sendo que de seis classes quatro tiveram melhor resultado com esta codificação. A classe que mais impactou a sensibilidade média foi a Liase, visto que em nenhuma das codificações o resultado superou 60%.

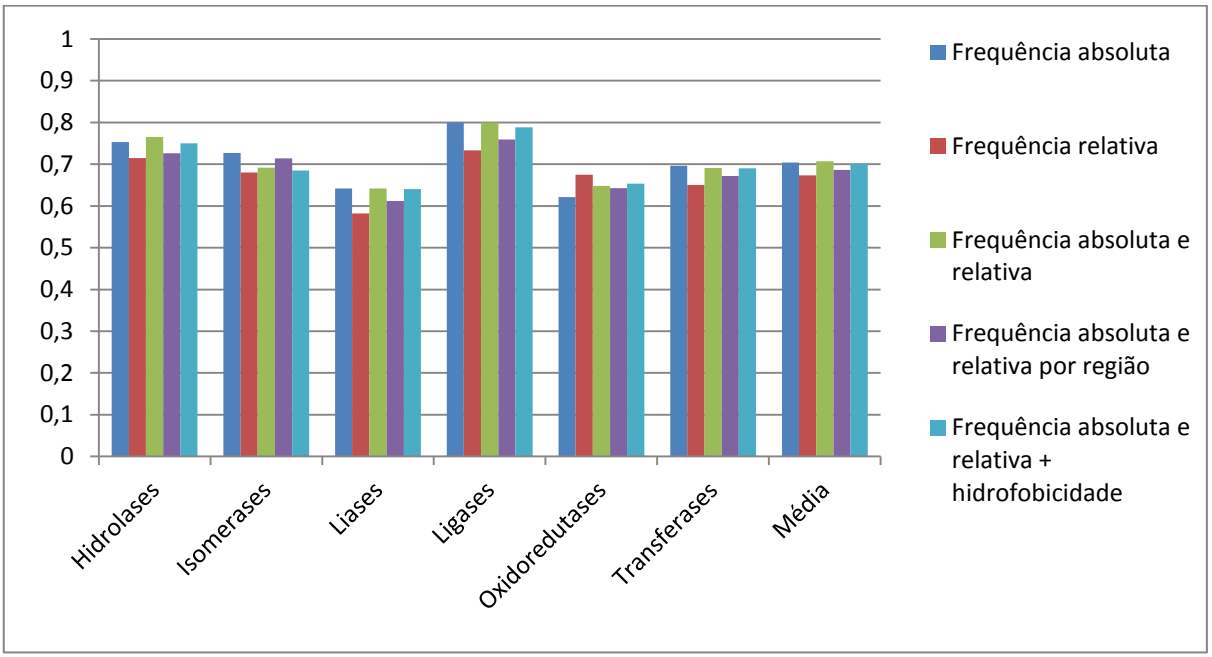
Figura 10 – Comparativa da sensibilidade entre codificações



Fonte: Dados da pesquisa.

Na média geral, o resultado da *f-measure,* Figura 11, ficou bem balanceado entre as codificações. Em quatro das seis classes usadas, os resultados da codificação de frequência absoluta e codificação da frequência absoluta e relativa foram superiores às demais codificações.

Figura 11 – Comparativa da *f-measure* entre codificações



Fonte: Dados da pesquisa.

6 CONSIDERAÇÕES FINAIS

Neste trabalho foi proposta e implementada uma codificação de proteínas que, submetida ao classificador SVM, capaz de predizer a função de uma proteína com um nível alto de assertividade. As codificações foram feitas com base nos aminoácidos e as características de hidrofobicidade das características extraídas dos bancos de dados PDB e STING DB.

Foram feitas seis codificações usando e combinando informações variadas das proteínas com objetivo de melhorar a precisão do classificador. Os resultados obtidos foram satisfatórios, porém ainda não suficientes para predizer a função exata das proteínas.

Acredita-se que com um estudo mais específico das proteínas com objeto de identificar informações mais relevantes na determinação da função, seja possível encontrar resultados mais precisos de predição.

REFERÊNCIAS

Borges-Osório M. R; Robinson W. M. **Genética Humana.** 3ª Edição, Porto Alegre – RS, 2013.

Boser, B.E; Guyon, I; Vapnik, V.N. **A training algorithm for optimal margin classifiers.** In Proceedings of the Fifth Annual Worlaghop of Computational Learning Theory, 5, 144-152, Pittsburgh, ACM, 1992.

Cheng C; Su E. C; Hwang J; Sung T; Hsu H. **Predicting RNA-binding sites of proteins using support vector machines and evolutionary information**. BMC Bioinformatics, 2008.

Cortes C; Vapnik V. **Support-Vector Networks.** AT&T Labs-Research, USA, 1995.
Floreano, D; Mattiussi, C; **Bio-Inspired Artificial Intelligence. Massachusetts Institute of Technology**. London, England, 2008.

Costa L. M. M; Zeilhofer P; Rodrigues W. S. **Avaliação do Classificador SVM (Support Vector Machine) no Mapeamento de Queimadas no Pantanal Mato-Grossense.** III Simpósio Brasileiro de Ciências Geodésicas e Tecnologias da Geoinformação, Recife - PE, Julho de 2010.

H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. **The Protein Data Bank**. UC San Diego, 2000. URL: http://www.rcsb.org/pdb. Acesso em: 26/05/2014.

Kapadia, C. D; Rinehart, C. A. ***Ab Initio* prediction of mycobacteriophages protein structure and function.** *BMC Bioinformatics* 2013, 14(Suppl 17):A10

Lehninger, A. L. Princípios de bioquímica. 3. Ed. São Paulo: Sarvier, 2002. 975 p.

Marsland S. **Machine Learning, An Algorithmic Perspective**. Massey University. Palmerston North, New Zealand, 2009.

Muppirala U. K; Honavar V. G; Dobbs D. **Predicting RNA-Protein Interactions Using Only Sequence Information**. BMC Bioinformatics, 2011.

REVISTA FOOD INGREDIENTS BRASIL. **Enzimas: Natureza e Ação nos Alimentos**. Edição Nº 16. São Paulo, SP, 2011.

S.R.M. Oliveira, G.V. Almeida, K.R.R. Souza, D.N. Rodrigues, P.R. Kuser-Falcão, M.E.B. Yamagishi, E.H. Santos, F.D. Vieira, J.G. Jardine and G. Neshich. **Sting_rdb: a relational database of structural parameters for protein analysis with support for data warehousing and data mining.** Genetics and Molecular Research, Embrapa Informática Agropecuária, Campinas, SP, Brasil, 2007.

Saeys Y; Inza I; Larrañaga P. **A review of feature selection techniques in bioinformatics.** Bioinformatics, 2007.

Silla Jr. C. N; Freitas A. A. **Selecting different protein representations and classification algorithms in hierarchical protein function prediction.** School of Computing and Centre for Biomedical Informatics.University of Kent, Canterbury, Kent, CT2 7NF, UK, 2011.

Someya, S; Kakuta, M; Morita, M; et al. **Prediction of Carbohydrate-Binding Proteins from Sequences Using Support Vector Machines**, Advances in Bioinformatics, vol. 2010, Article ID 289301, 9 pages, 2010. doi:10.1155/2010/289301

Sousa, B. F. S; Teixeira, A. S; Silva, F. A. T. F. **Classificação de bioma caatinga usando Support Vector Machines (SVM)**. Simpósio Brasileiro de Sensoriamento Remoto. Natal, Brasil, 2009.

SUN, Dengdi; HU, Maolin. **Determining protein function by protein-protein interaction network.** In: The 1st International Conference on Bioinformatics and Biomedical Engineering. [S.l.]: IEEE, 2007. p. 33–36.

Swakkhar Shatabda, M. A. Hakim Newton, Mahmood A. Rashid, Duc Nghia Pham, and Abdul Sattar. **How Good Are Simplified Models for Protein Structure Prediction?**, Advances in Bioinformatics, vol. 2014, Article ID 867179, 9 pages, 2014. doi:10.1155/2014/867179

Uzunian, A; Birner, E. **Biologia, volume único**, 3ª ed. Editora HARBRA, 2008.