UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Biológicas
Programa Interunidades de Pós-Graduação em Bioinformática

Lucio Paccori Lima

ALGORITMOS PARA MINERAÇÃO DE DADOS UTILIZANDO REGRESSÃO LOGISTICA - APLICAÇÃO EM BIOINFORMÁTICA ESTRUTURAL

Belo Horizonte
2023

Lucio Paccori Lima

ALGORITMOS PARA MINERAÇÃO DE DADOS UTILIZANDO REGRESSÃO LOGISTICA - APLICAÇÃO EM BIOINFORMÁTICA ESTRUTURAL

Tese apresentada ao Programa de Pós-graduação em Bioinformática do Instituto de Ciências Biológicas da Universidade Federal de Minas Gerais, como requisito parcial para a obtenção do grau de Doutor em Bioinformática.

Orientador: Prof. Dr. Marcos Augusto dos Santos

Belo Horizonte
Setembro de 2023

043 Lima, Lucio Paccori.
Algoritmos para mineração de dados utilizando regressão logistica – aplicação em bioinformática estrutural [manuscrito] / Lucio Paccori Lima. – 2023.
83 f. : il. ; 29,5 cm.

Orientador: Prof. Dr. Marcos Augusto dos Santos.
Tese (doutorado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas. Programa Interunidades de Pós-Graduação em Bioinformática.

1. Bioinformática. 2. Modelos Logísticos. 3. Mineração de Dados. I. Santos, Marcos Augusto dos. II. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. III. Título.

CDU: 573:004

Ficha catalográfica elaborada pela bibliotecária Fabiane C M Reis – CRB 6 - 2680

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA INTERUNIDADES DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

ATA DE DEFESA DE TESE

LUCIO PACCORI LIMA

Às nove horas do dia **27 de setembro de 2023**, reuniu-se, em formato híbrido, presencial na Sala 2077 do ICEx e por videoconferência pela plataforma Zoom, a Comissão Examinadora de Tese indicada pelo Colegiado do Programa para julgar, em exame final, o trabalho do discente **Lucio Paccori Lima,** intitulado: **"ALGORITMOS PARA MINERAÇÃO DE DADOS UTILIZANDO REGRESSÃO LOGÍSTICA - APLICAÇÃO EM BIOINFORMÁTICA ESTRUTURAL"**, requisito para obtenção do grau de Doutor em **Bioinformática.** Abrindo a sessão, o Presidente da Comissão, **Dr. Marcos Augusto dos Santos**, após dar a conhecer aos presentes o teor das Normas Regulamentares do Trabalho Final, passou a palavra ao candidato, para apresentação de seu trabalho. Seguiu-se a arguição pelos Examinadores, com a respectiva defesa do candidato. Logo após, a Comissão se reuniu, sem a presença do candidato e do público, para julgamento e expedição de resultado final. Foram atribuídas as seguintes indicações:

Professor(a)/Pesquisador(a)	**Instituição**	**Indicação**
Dr. Marcos Augusto dos Santos - Orientador	UFMG	**Aprovado**
Dra. Letícia Pereira Pinto	UFMG	**Aprovado**
Dr. Frederico Ferreira Campos Filho	UFMG	**Aprovado**
Dr. João Arthur Ferreira Gadelha Campelo	BDMG	**Aprovado**
Dr. Bráulio Roberto Gonçalves Marinho Couto	Biobyte	**Aprovado**
Dr. José Miguel Ortega	UFMG	**Aprovado**

Pelas indicações, o candidato foi considerado: **Aprovado**

O resultado final foi comunicado publicamente ao candidato pelo Presidente da Comissão. Nada mais havendo a tratar, o Presidente encerrou a reunião e lavrou a presente ATA, que será assinada por todos os membros participantes da Comissão Examinadora.

Belo Horizonte, 27 de setembro de 2023.

Documento assinado eletronicamente por **Jose Miguel Ortega**, **Servidor(a)**, em 28/09/2023, às 14:35, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Frederico Ferreira Campos Filho**, **Professor do Magistério Superior**, em 28/09/2023, às 18:28, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Leticia Pereira Pinto**, **Professora Magistério Superior-Substituta**, em 28/09/2023, às 18:41, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Bráulio Roberto Gonçalves Marinho Couto**, **Usuário Externo**, em 29/09/2023, às 10:49, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **João Arthur Ferreira Gadelha Campelo**, **Usuário Externo**, em 01/10/2023, às 18:58, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Marcos Augusto dos Santos**, **Membro de comissão**, em 02/10/2023, às 18:09, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **2662616** e o código CRC **6F865E62**.

Referência: Processo nº 23072.258949/2023-53 SEI nº 2662616

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA INTERUNIDADES DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

FOLHA DE APROVAÇÃO

Lucio Paccori Lima

"ALGORITMOS PARA MINERAÇÃO DE DADOS UTILIZANDO REGRESSÃO LOGÍSTICA - APLICAÇÃO EM BIOINFORMÁTICA ESTRUTURAL"

Tese aprovada pela banca examinadora constituída pelos Professores:

Prof. Marcos Augusto dos Santos - Orientador
UFMG

Profa. Letícia Pereira Pinto
UFMG

Prof. Frederico Ferreira Campos Filho
UFMG

Prof. João Arthur Ferreira Gadelha Campelo
BDMG

Prof. Bráulio Roberto Gonçalves Marinho Couto
Biobyte

Prof. José Miguel Ortega
UFMG

Belo Horizonte, 27 de setembro de 2023.

Documento assinado eletronicamente por **Jose Miguel Ortega**, **Servidor(a)**, em 28/09/2023, às 14:35, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Frederico Ferreira Campos Filho**, **Professor do Magistério Superior**, em 28/09/2023, às 18:28, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Leticia Pereira Pinto**, **Professora Magistério Superior-Substituta**, em 28/09/2023, às 18:41, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Bráulio Roberto Gonçalves Marinho Couto**, **Usuário Externo**, em 29/09/2023, às 10:50, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **João Arthur Ferreira Gadelha Campelo**, **Usuário Externo**, em 01/10/2023, às 18:58, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Marcos Augusto dos Santos**, **Membro de comissão**, em 02/10/2023, às 18:10, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **2662620** e o código CRC **6CC238CC**.

Referência: Processo nº 23072.258949/2023-53 SEI nº 2662620

I dedicate this work to people who, with a lot of effort, study and humility, fight to achieve their goals.

Agradecimentos

Ao meu orientador, professor Marcos, pela confiança e compreensão.
Aos meus familiares e amigos, pela amizade e apoio.
A UFV pelo tempo cedido.
A UFMG.

"A maior tristeza de não haver vencido
é a vergonha de não ter lutado !"
(Ruy Barbosa)

Resumo

Vários problemas em bioinformática estão centrados na busca de relacionamentos entre componentes não detectáveis à primeira vista, semelhantes à recuperação de informação latente em máquinas de busca como Yahoo, Google etc. Técnicas de inteligência artificial usadas por essas máquinas transformam a enorme quantidade de dados em descoberta e conhecimento. Dentre os recursos disponíveis na área, a Regressão Logística ocupa um lugar pouco explorado, mas que tem muito a oferecer. Neste trabalho, mostramos como a Regressão Logística tradicional, com algumas das modificações sugeridas aqui, pode ser usada com sucesso em uma classe de problemas de bioinformática: aquela que envolve a estrutura tridimensional de proteínas. Esta aplicação se junta a outras de forma que a versatilidade da técnica pôde ser aferida, uma vez que ela já foi explorada em classificadores construídos a partir de *microarray*, no reaproveitamento de fármacos, em triagem virtual de ligantes, na busca de alvos drogáveis e em árvores filogenéticas.

Palavras-chave: bioinformática; regressão logística; mineração de dados.

Abstract

Several problems in bioinformatics are rooted in the search for relationships between components that are not detectable at first glance, similar to the retrieval of latent information in search engines such as Yahoo, Google, etc. Artificial intelligence techniques used by these machines transform enormous amounts of data into discovery and knowledge. Among the resources available in the area, Logistics Regression occupies a little explored place, but has a lot to offer. In this work, we show how traditional Logistic Regression, with some of the modifications suggested here, can be used successfully in a class of bioinformatics problems; namely, that involving the three-dimensional structure of proteins. This application joins others, where the versatility of the technique could be assessed. It has already been explored in classifiers built from microarrays, in the reuse of drugs, in virtual screening of ligands, in the search for druggable targets and in phylogenetic trees.

Palavras-chave: bioinformatics; logistic regression; data mining.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

Este trabalho aborda a aplicação de álgebra linear e estatística em um contexto desafiador para a mineração de dados [Dos Santos *et al.*, 2012], com ênfase em problemas com mais atributos do que entidades, que abundam em Bioinformática. Exemplo deste fato ocorrem em, por exemplo, em reaproveitamento de fármacos, triagem virtual baseada em ligante [Leite *et al.*, 2020] e busca de alvos drogáveis [Silvério-Machado *et al.*, 2015]. A aplicação dos métodos usuais de Inteligência Artificial [Coppin *et al.*, 2010] (redes neurais, *support vector machine* etc) nessa classe de problemas (onde se tem mais atributos que entidades) é desafiador. Os artefatos de mineração [Carvalho *et al.*, 2015] usuais não são conhecidos por serem adequados para escolher quais atributos (e seus valores) que mais contribuem para uma determinada entidade estar em um determinado grupo. Esta fase é conhecida como *feature selection* [Vidyavathi *et al.*, 2019]; é uma área em aberto na ciência da computação. Precede, por assim dizer, a própria técnica escolhida para resolver o problema. Assim, quando necessário, é feita, a priori, uma seleção dos atributos capazes de discernir as entidades de um grupo de interesse. Esta dificuldade, em geral, é resolvida limitando a natureza combinatória do problema, que é, sabidamente, de complexidade exponencial. Mas aqui, na verdade, não temos esta dificuldade; a própria seleção de atributos nos métodos que utilizamos é um subproduto da aplicação dos algoritmos aqui propostos. Em nenhum momento, procedemos à seleção prévia dos atributos.

Um método interessante, que tem como resultado colateral uma avaliação da importância dos atributos, é a Regressão Logística [Hosmer e Lemeshow *et al.*, 2013]; uma técnica bem conhecida e muito utilizada na área médica [Ahmad *et al.*, 2014], [Andriani & Chamidah, *et al.*, 2019]. Mas a sua utilização genérica como instrumento de classificação traz uma série de inconvenientes [Abreu *et al.*, 2019] e [Yi *et al.*, 2006] como overfitting ou underfitting. Possivelmente seja esta razão por ele não estar precisamente contemplado na taxonomia apresentada na Figura 1.1. Nem sempre nos problemas reais tem-se um domínio com equilíbrio adequado entre a quantidade (e variedade) dos membros do grupo que se deseja classificar e a quantidade (e variedade) da população total. Por exemplo, no

repositório de problemas de classificação Kaggle (`https://www.kaggle.com/datasets`) abundam problemas com 10^9 indivíduos e um subgrupo de interesse para classificação com não mais que 10^3 entidades. É o caso, por exemplo, de fraudes de cartões de crédito [Felipe *et al.*, 2012]; muitas transações lícitas contra um relativamente pequeno conjunto de transações ilícitas (conjunto desbalanceados). Nos problemas que escolhemos para exemplificar e testar os métodos, temos domínios que contam com grupos de 27000 entidades coexistindo com outros limitados a algumas dezenas (conjunto desbalanceados) [Barella *et al.*, 2015]. Estas e outras inconveniências são resolvidas nos algoritmos que estamos propondo e/ou adaptando.

Já existem aplicações baseadas nessas ferramentas, como em microarrays [Morais *et al.*, 2020], reaproveitamento de fármacos e triagem virtual de ligantes [Leite *et al.*, 2020], busca de alvos drogáveis [Silvério-Machado *et al.*, 2015], árvores filogenéticas [Santos *et al.*, 2011]. Todas elas estão no contexto onde o número de atributos é superior - em várias ordens de grandeza - ao número de entidades.

Dessa forma, exploramos, agora, essa metodologia em Bioinformática Estrutural [Silva *et al.*, 2023] e [do Nacimento *et al.*, 2020]. É uma das áreas fundamentais da ciência que lida com proteínas. Notadamente, são elas os alvos primários a serem modulados por drogas na busca do restabelecimento do reequilíbrio da vida que, eventualmente, tenha sido abalado por alguma situação adversa. Sua função é ditada pela sua estrutura tridimensional.

Para testar nossos algoritmos, elegemos problemas de classificação, usando conjuntos de proteínas bem conhecidos e facilmente disponíveis [Pires *et al.*, 2011]. Decidimos por esta aplicação simples, priorizando mostrar a viabilidade do nosso propósito ao lado de incentivar o leitor a buscar solucionar problemas mais sofisticados na área Bioinformática

Existem ótimos métodos para classificar e recuperar proteínas em suas respectivas super famílias, famílias e subgrupos diversos. A partir da descrição das proteínas por um *vector space model* (VSM) [Yi *et al.*, 2006], no qual cada entidade é representada por um conjunto de tamanho fixo de atributos numéricos, são aplicados algoritmos tradicionais, a saber, redes neurais, pesquisa em árvores, *support vector machine* e outros (Figura 1.1) [Vapkin et al., 1999], [Frank et al,. 2004], [Wang et al., 2005] e [Ma et al., 2014]. Estes algoritmos ão capazes de predizer com eficiência a que grupo uma dada proteína pertence. Ao que nos consta, em [Pires DE, de Melo-Minardi RC, dos Santos MA, Santoro MM, et al., 2012], são apresentados o melhores resultados, que permitem construir oráculos perfeitos (considerados por nós quando a média harmônica obtida de validações cruzadas são superiores à 0,95).

Entretanto, os artefatos de mineração citados não são, como já dito, adequados para escolher quais atributos (e seus valores) contribuem para uma determinada entidade estar em um certo grupo [Silva *et al.*, 2023].

A justificativa, e muito da motivação inicial para este trabalho, era encarar um fato

que ocorre na natureza. É sabido que as proteínas com sequências similares correspondem a estruturas tridimensionais também similares. Entretanto, existem casos em que estruturas similares não guardam similaridade quanto às sequências primárias [Antczak *et al.*, 2016], [Kolodny *et al.*, 2004] e [Leach *et al.*, 2001]. Esperávamos encontrar assinaturas que fossem invariantes quanto a este aspecto. Para o seu desenvolvimento, acabamos por ter a necessidade de buscar por oráculos perfeitos, que passou a ser o principal objetivo do trabalho.

É importante também comentar, brevemente, sobre a arquitetura que adotamos para este texto e algumas das escolhas feitas. Permeia o nosso trabalho, o intuito de divulgar as técnicas discutidas em um formato que permita aos pesquisadores interessados em usar inteligência artificial, encontrar aqui uma forma rápida, robusta e eficiente de fazê-lo. Para tanto, não utilizamos os mecanismos tradicionais da área e apresentamos uma abordagem que resgata processos de recuperação de informação, que, por diversos motivos, estão negligenciados na literatura. A abordagem, neste aspecto, é nova; consistente com isto, fazemos exigências modestas sobre o conhecimento matemático do leitor. Além disto, é de nosso interesse que partes deste texto estejam presentes em um livro que estamos planejando produzir.

Taxonomia de Mineração de Dados

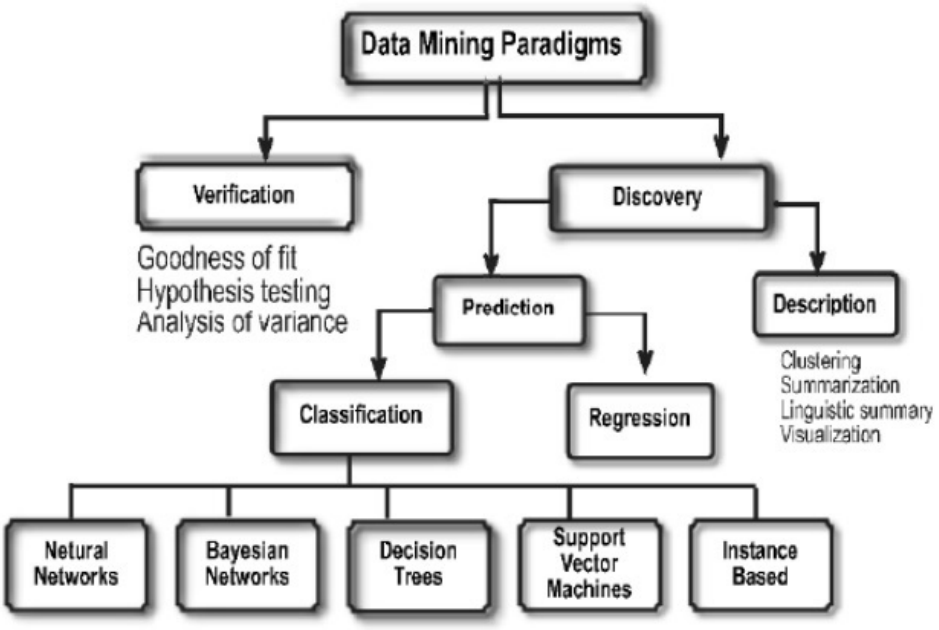


Figura 1.1. Taxonomia do processo de mineração de dados. Ilustra a inter-relação e agrupamento dos métodos. Cabe distinguir dois principais tipos: *Verification* (verifica a hipótese do usuário) e *Discovery* (encontra novas regras e padrões de forma autônoma). Adaptado de [Maimon, et al., 2005].

1.1 Objetivos

1.1.1 Objetivo Geral

Validar o uso de Regressão Logística para aplicações em Bioinformática Estrutural.

1.1.2 Objetivos Específicos

- Escolher parâmetros e bases de dados para avaliar os métodos e estratégias propostas.
- Explorar os **vsm's** conhecidos para o problema.
- Explorar e propor algoritmos eficientes para serem utilizados em Bioinformática estrutural.

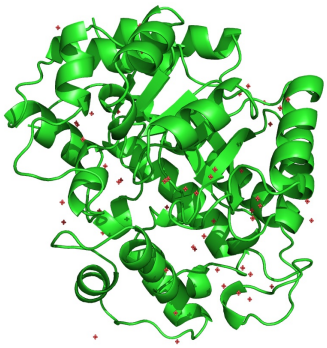
Capítulo 2

Metodologia

Os experimentos foram realizados em um microcomputador de processador Intel(R) Core(TM)i5 4200M CPU @ 1.60GHz, 8.00 GB de memória RAM. O sistema operacional instalado nesse microcomputador é o Windows 10 Professional 64 bits. Usando o ambiente de prototipagem **MatLab** versão R2019b (`https://www.mathworks.com`) com os *Toolboxes* de Bioinformática, Estatística e Otimização. As cópias utilizadas são as de avaliação, periodicamente renovadas.

2.1 Conjunto de dados

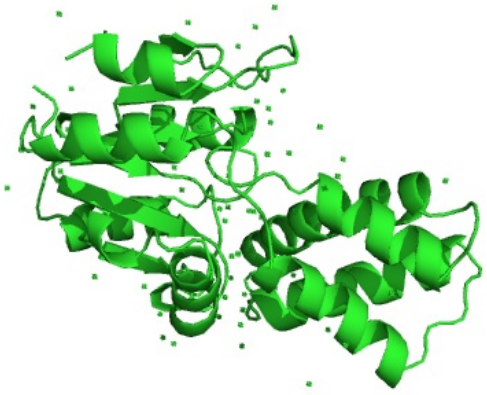
A primeira base de dados conta com o conjunto de superfamílias de enzimas, tidas como um padrão ouro, que utilizam mecanismos distintos para executar suas funções [Brown et al., 2006]. Neste conjunto, são consideradas seis superfamílias (*Amdohydrolase, Crotonase, Haloacid dehalogenase, Isoprenoind synthase type I e Vicinal oxygen chelate*) (ver Figura 2.1), compreendendo 47 famílias distribuídas em 896 diferentes cadeias.



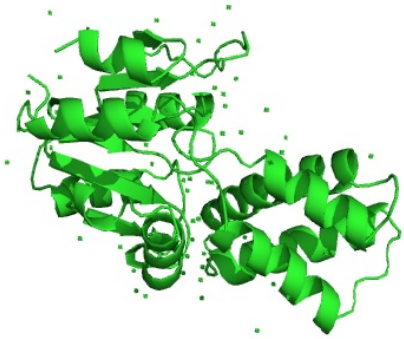
(a) Amidhydrolase: 2ADA

(b) Crotonase: 1WDM

(c) Enolase: 2ONE



(d) Haloacid Dehalogenese: 1JUD



(e) Isoprenoid Syn. Type I: 1HXG

(f) Vicinal Oxygen Chel.: 1CJX

Figura 2.1. Visualização de estruturas tridimencionais de proteínas(PDB): Amidhydrolase (2ADA:A), Cotronase (1WDM:A), Enolase (2ONE:A), Haloacid dehalogenase (1JUD:A), Isoprenoid synthase type I (1HXG:A) e Vicinal oxygen chelate (1CJX:A), observe-se claramente não são semelhantes nas estruturas tridimensinais pois são de superfamílias diferentes. Figuras geradas usando o PyMOL.

Mostra-se um exemplo de alinhamento estrutural entre duas entidades: uma *Amidhydrolase* (2ADA:A) e uma *Crotonase* (1WDM:A) do conjunto de dados *gold standard*, logo após, duas rodadas de alinhamentos estruturais e tentando reduzir o RMSD. Obtêm-se o RMSD final aproximado de 25. Lembre-se que valores altos de RMSD indicam uma sobreposição ruim entre estruturas pois as proteínas *Amidhydrolase* (2ADA:A) e uma *Crotonase* (1WDM:A), Figura 2.2 escolhidas como exemplo, são de grupos diferentes.

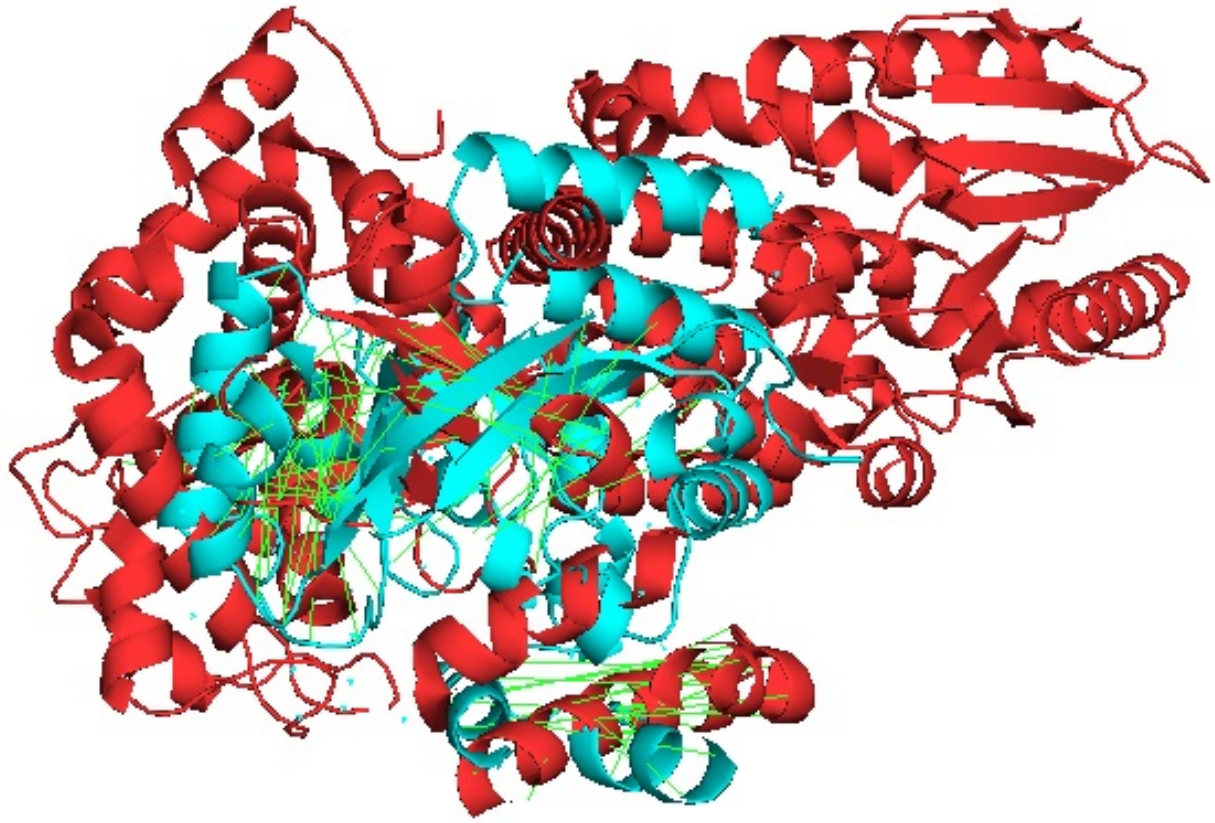


Figura 2.2. Visualização do alinhamento de sequências uma Amidhydrolase (2ADA:A) e uma Crotonase (1WDM:A), Figura gerada usando o PyMOL.

Os outros dataset que utilizamos têm características muito diferentes e são trabalhados com dados agrupados segundo uma classificação estrutural. Utilizamos um conjunto de quatro *data sets* não balanceados, a saber 6SSE, 5SSE, 4SSE, e 3SSE [Jain & Hirst, 2010], que são estruturas de pequenas proteínas.

Finalmente, o terceiro conjunto de dados congrega tipos e dados com variações tanto no tamanho quanto no número de entidades em categorias específicas. Escolhemos o SCOP (*Structural Classification Of Protein, versão 1.75*), de onde foram recuperados os identificadores no *Protein Data Bank* (**PDB**) (`\https://www.rcsb.org/`) e o SCOP que classifica as proteínas em classes, enovelamentos, superfamílias e famílias respectivamente [Berman et al., 2000].

2.2 Métodos Baseados em Álgebra Linear para Recuperação de Informação de Forma Inteligente

A Álgebra Linear fornece instrumentos amplamente utilizados na mineração de dados. As modernas máquinas de busca (Yahoo, Google etc) se esteiam em métodos e formas de representação que remetem à mineração de dados e aos problemas de topologia, comumente estudados como espaços normados na matemática, caracterizados por matrizes e vetores. Páginas na *web*, as quais correspondem às entidades, deixam de ser coleções semânticas de caracteres e passam a ser representadas por vetores de números reais; o conjunto desses vetores formam matrizes de números com propriedades de posto (*rank*), espectro (autovalores e autovetores), dentre outras. Na *web*, o **VSM** é um dicionário de radicais de uma língua - por exemplo, palavras como computador, computável e computação referem-se a um mesmo radical "comput". Uma modelagem que considera esses instrumentos sofisticados e escaláveis parte da definição do que será o **VSM** - é a ponte que liga uma realidade em questão aos problemas abstratos de Álgebra Linear. Com isto, situações de áreas distintas podem ser mapeados para o uso dos elementos básicos dessa área da matemática. Por exemplo, uma sequência de resíduos de aminoácidos de uma proteína pode ser representada como as frequências que uma dada janela de k resíduos aparece [Couto, B. R. G. M.; Santos, M. A. *et al.*, 2007]. No nosso problema, onde buscamos representar uma estrutura tridimensional de uma proteína, inicialmente mapeamos o **VSM** segundo uma ideia de [Pires DE, de Melo-Minardi RC, dos Santos MA, Santoro MM, *et al.*, 2011] que é o *cutoff scanning*. Posteriormente, mudamos esta abordagem; adotamos uma outra proposta neste trabalho, mais adequada aos nossos objetivos.

No que se segue, primeiro comentamos alguns aspectos que lançamos mãos daqueles instrumentos de Álgebra Linear que viabilizam a recuperação de informação em conjuntos de dados de grande porte. Depois mostramos os **VSM's** que utilizamos e, no final desta seção, apresentamos a Regressão Logística e as alterações com as quais esperamos que ela se transforme em um instrumento digno de constar na taxonomia da Figura 1.1.

2.2.1 Decomposição em valores singulares

A decomposição por valores singulares ou *singular value decomposition* (**svd**) [Eldén, 2006, 2007; Berry, 1995] é uma técnica da álgebra linear utilizada para reduzir a dimensionabilidade das entidades sem comprometer a sua essência. Com o **svd** é possível, em vez de trabalhar com uma matriz com o posto aproximado, simplesmente usar a combinação linear dos padrões presentes na matriz que aproximam o posto. Em geral, para o propósito de recuperar informação em problemas reais, não são necessários muitos padrões, implicando em uma representação mais econômica. Por exemplo, no caso da *web*, não obstante o número de radicais de uma língua ser elevado, o número de padrões, segundo

a referência citada, varia de 40 até 240. Reforçando o que queremos dizer, isto implica que uma entidade nesse contexto não tem mais que 240 atributos.

Mais formalmente, a decomposição por valores singulares é uma fatoração de uma matriz qualquer em três outras matrizes com propriedades importantes. Possui várias aplicações, tanto diretas, nas quais se aplicam os resultados extraídos de suas matrizes fatores, quanto como um passo em muitos algoritmos.

Definição 1 *Dado $A \in \mathbb{R}^{m \times n}$, não necessariamente de posto completo, a decomposição por valores singulares de A é uma fatoração tal que:*

- $A = USV^T$;
- $U \in \mathbb{R}^{m \times m}$, *são os vetores singulares à esquerda de* AA^T *e é ortogonal;*
- $V \in \mathbb{R}^{n \times n}$, *são os vetores singulares à direita de* A^TA *e é ortogonal;*
- $S \in \mathbb{R}^{m \times n}$, *é diagonal se* $m = n$, *caso contrário adiciona-se* $m - n$ *linhas de zeros em* S *e é formado por a raiz quadrada dos autovalores de* AA^T.

$$S = diag(\sigma_1, \sigma_2, .., \sigma_n)$$

$$\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_n \geq 0.$$

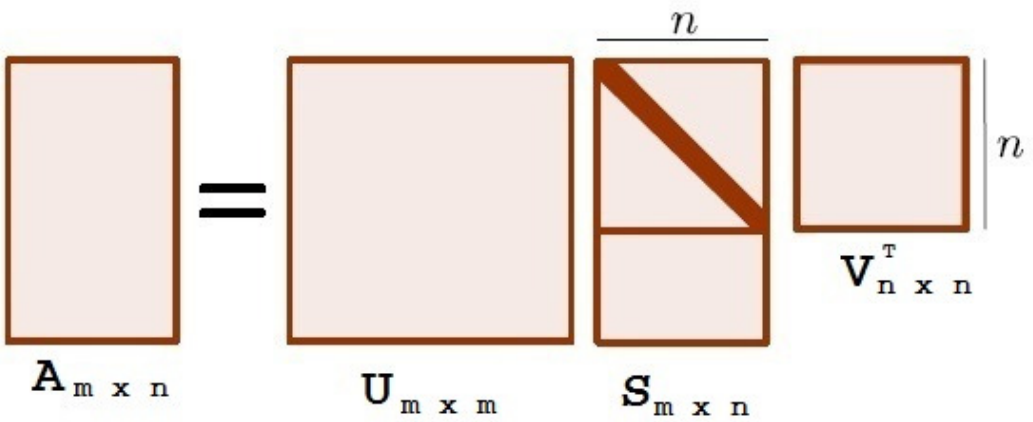


Figura 2.3. Decomposição em Valores Singulares de uma matriz A.

Os valores $(\sigma_1, \sigma_2, ...\sigma_n)$ são chamados de valores singulares de A. As colunas de U, $(u_1, u_2, ...u_m)$ são chamadas de vetores singulares à esquerda de A e as colunas da matriz V, $(v_1, v_2, ..., v_n)$ são os vetores singulares à direita de A. As colunas de U podem ser interpretadas como padrões das entidades em A com um peso correspondente σ_i; já V são os padrões das linhas, que têm o mesmo peso associado.

Uma redução de posto da matriz A para ser útil à recuperação de informação, consiste na escolha de um valor k, inferior ao posto da matriz, que implicará na representação

de A por uma outra matriz A_k cujo posto será precisamente k. Usando os fatores oriundos do *svd* sabe-se que a matriz A_k é a mais próxima de A segundo a norma de Frobennius. Assim, uma aproximação de A, para nós dependerá da escolha do número de valores singulares a serem utilizados e, consequentemente, do número de padrões que serão utilizados na representação de A_k.

Considerando este número como $k < p$, onde p é o posto de A, a fatoração de A pode ser vista de uma outra forma, conhecida como decomposição diálica:

$$A \approx A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T$$

Neste formato, onde cada termo da somatória é uma matriz, dado que $\sigma_i u_i v_i^T$ é um produto externo de vetores que resulta em uma matriz de posto 1, observa-se que os valores $\sigma_i \approx 0$ imputam o que pode ser considerado como ruído às entidades. Sua retirada tem o efeito de propiciar a formação de grupos, eventualmente não detectáveis na forma original de A [Eldén *et al.*, 2007]. Na prática, temos observado que, ao serem plotados os valores de σ_i em um gráfico, o início da cauda longa é um bom indicativo do valor de k a ser adotado. Entretanto este processo é mais elaborado em outros trabalhos, como por exemplo em [Santos, A R; Santos, M A ; Azevedo, V. *et al.*, 2011].

Quanto à redução de dimensionalidade, ela pode ser entendida quando escrevemos $A = USV^T = U(SV^T)$. Estes parênteses são especiais; eles indicam que cada coluna do produto $H = (SV^T)$ pode ser interpretada como a combinação linear dos padrões em U para gerar a correspondente coluna de A. Ao procedermos à redução do posto de A, temos que $H_k = S_k V_k^T$ é uma matriz de k linhas por n colunas. Se duas entidades são próximas enquanto vetores da matriz $A_k = U_k S_k V_k^T$, é porque as combinações lineares em H_k também são próximas, implicando que basta trabalhar com $H_K \in \mathbb{R}^{k \times m}$ em vez de $A_k \in \mathbb{R}^{m \times n}$, lembrando que $k < p < m$. A representação de uma entidade $q \in \mathbb{R}^m$, no espaço $\mathbb{R}^k$ gerado por U_k, é dada por uma projeção. Esta projeção q_k é a solução de um sistema de equações lineares $U_k q_k = q$, que, pelo método da minimização da somatória dos quadrados dos resíduos, é dada por $q_k = U_k^T q$.

Estamos utilizando o *svd* neste trabalho com três propósitos. O primeiro deles é o fato de que se trata de uma ferramenta que possibilita a visualização de conjuntos de dados no espaço tridimensional [Marcolino LC, Couto BRGM, dos Santos MA, 2010], o que nos permite ter uma avaliação inicial das eventuais dificuldades no processo de classificação. Outra aplicação vem do fato de que, nas melhorias que estamos propondo nos algoritmos, temos de fazer escolhas e recuperar um subconjunto de entidades que são mais próximas a uma dada consulta; usando as técnicas de máquina de busca, conseguimos melhores resultados. Finalmente, usamos diretamente os valores singulares como **VSM**, conforme visto a seguir.

2.2.2 Representação de entidades no contexto de estruturas de proteínas

Neste trabalho, estamos utilizando duas representações para as entidades. A primeira delas, *o cutoff scanning*, proposto por [Pires DE, de Melo-Minardi RC, dos Santos MA, Santoro MM, et al., 2012], deu origem ao melhor, ou um dos melhores classificadores para esta classe de problemas. A outra forma, *a spectral pattern use*, é uma ideia original que estamos validando.

Representação de entidades usando o *cutoff scanning* .
Neste **VSM**, a proteína é descrita a partir do número de átomos que existe em subintervalos de 0,2 Å(ångströms) até uma distância máxima de 30Å. As distâncias interatômicas estão sendo usadas como características estruturais. Cada linha representa o vetor de características dos padrões de distância entre os átomos do conjunto selecionado de uma das cadeias utilizadas. Foi usado por [Pires DE, de Melo-Minardi RC, dos Santos MA, Santoro MM, et al., 2011], em cujo trabalho são mostrados resultados excepcionais para esta classe de problemas de classificação.

Nova representação de entidades: *spectral pattern* .
Existe um fato interessante e pouco conhecido que exploramos neste trabalho. A matriz de distâncias D entre os átomos de uma proteína (ou parte deles) tem uma propriedade peculiar e, até onde fomos capazes de verificar, exclusiva; experimentos explorando este fato estão no seção 3.4. Ao plotarmos em um gráfico os valores singulares obtidos da decomposição desta matriz, observamos a presença de seis valores significativos; sempre a cauda longa se inicia a partir do sexto valor. Para ilustrar este fato, considere a proteína 1A4M, onde mostramos um histograma das distâncias (Figura 2.4) e o gráfico dos seus valores singulares (Figura 2.5).

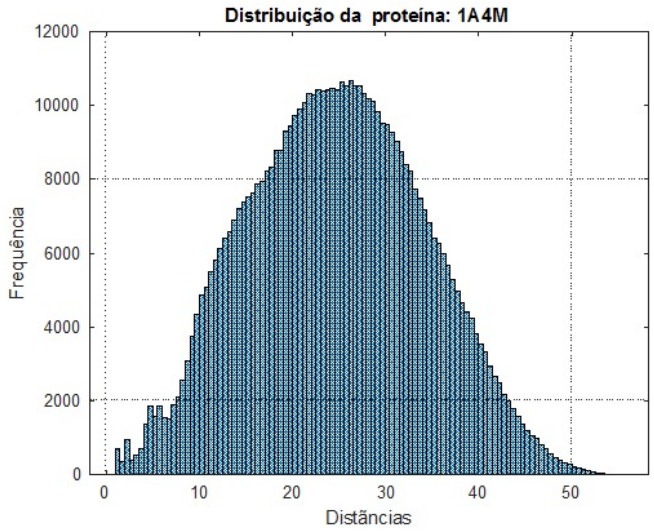


Figura 2.4. Histograma das 547581 distâncias interatômicas dos C_α da proteína 1A4M.pdb, pertencentes ao intervalo [1,31 55,17] Angstrons.

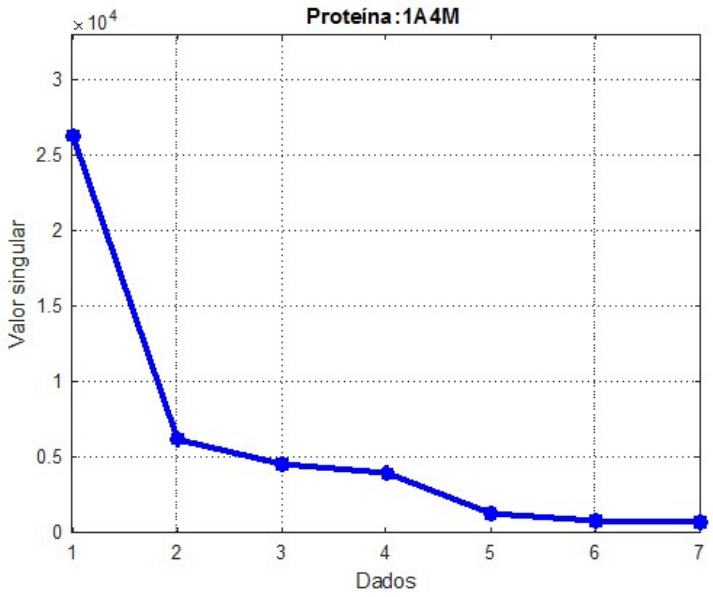


Figura 2.5. Valores Singulares σ_i associados à matriz de distâncias entre os C_α da proteína 1A4M. Dos 1047 valores, são mostrados somente os iniciais; os demais são valores muito mais baixos quando comparados com estes e foram omitidos.

Este fato, até onde fomos capazes de verificar, ocorre com todas as proteínas, não importando quais conjuntos de átomos foram considerados na construção da matriz de distâncias. Podem ser usados todos os átomos, ou somente um subgrupo deles (cadeia principal, C_α etc). Isto nos motivou a utilizar os próprios valores singulares como atributos do VSM. Na Figura 2.6, escolhemos ao acaso seis proteínas pertencentes a superfamílias distintas do conjunto de dados que testamos na subseção 3.4.2. Pode-se observar uma clara diferenciação no espectro dos valores singulares dessas proteínas.

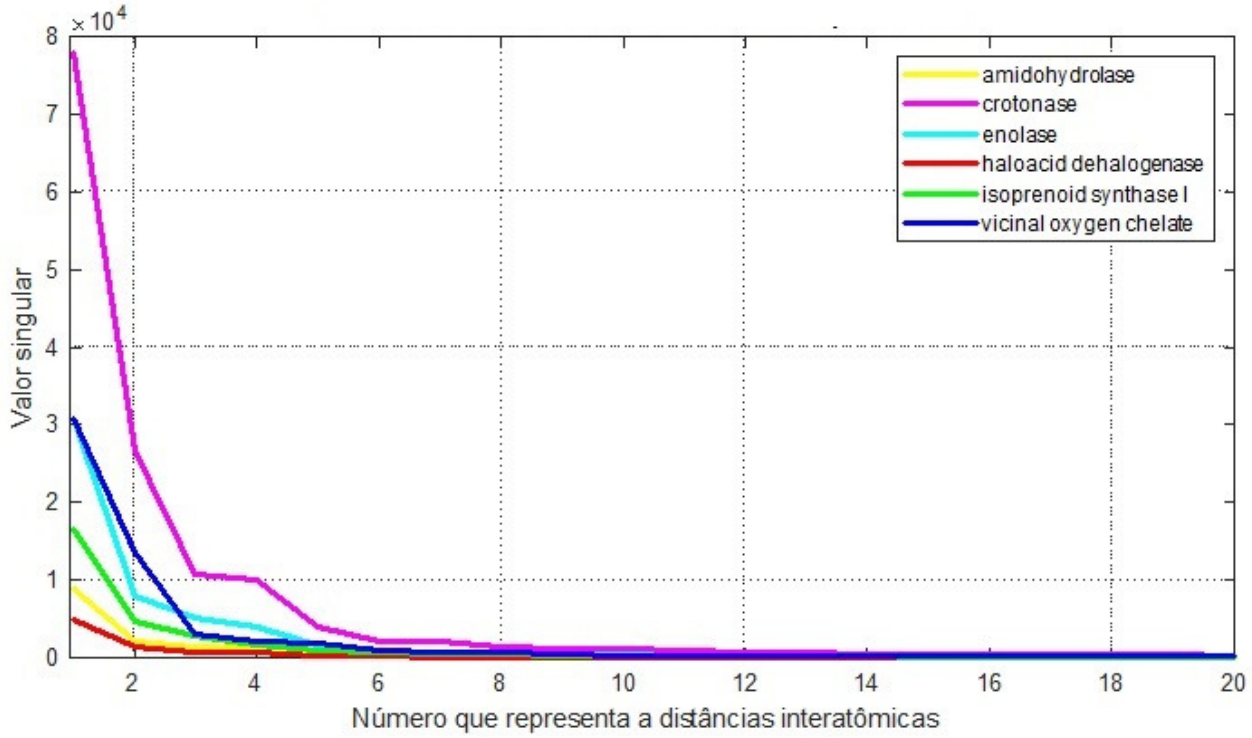


Figura 2.6. Primeiros valores singulares das matrizes de distâncias entre os átomos de proteínas de superfamílias distintas retiradas do conjunto de dados *golden standard*.

2.3 Regressão Logística

A Regressão Logística consiste em encontrar valores de $\alpha = (\alpha_1, \alpha_2, ..., \alpha_n)^T$ para ajustar a equação (2.1) para cada uma das entidades j. O valor da função $P_j(x)$, conhecida como *logit*, com $P_j(x) \in [0, 1]$ informa a probabilidade de uma dada entidade ser classificada como pertencente a um subconjunto específico.

$$P_j(x) = \frac{e^{\sum_{i=1}^{m} \alpha_i x_i}}{1 + e^{\sum_{i=1}^{m} \alpha_i x_i}} \tag{2.1}$$

Observa-se que quando $e^{\sum_{i=1}^{m} \alpha_i x_i}$ tende para zero, $P_j(x)$ também tende para zero. Por outro lado, se $e^{\sum_{i=1}^{m} \alpha_i x_i}$ tende para infinito, $P_j(x)$ aproxima-se da unidade (figura 2.7).

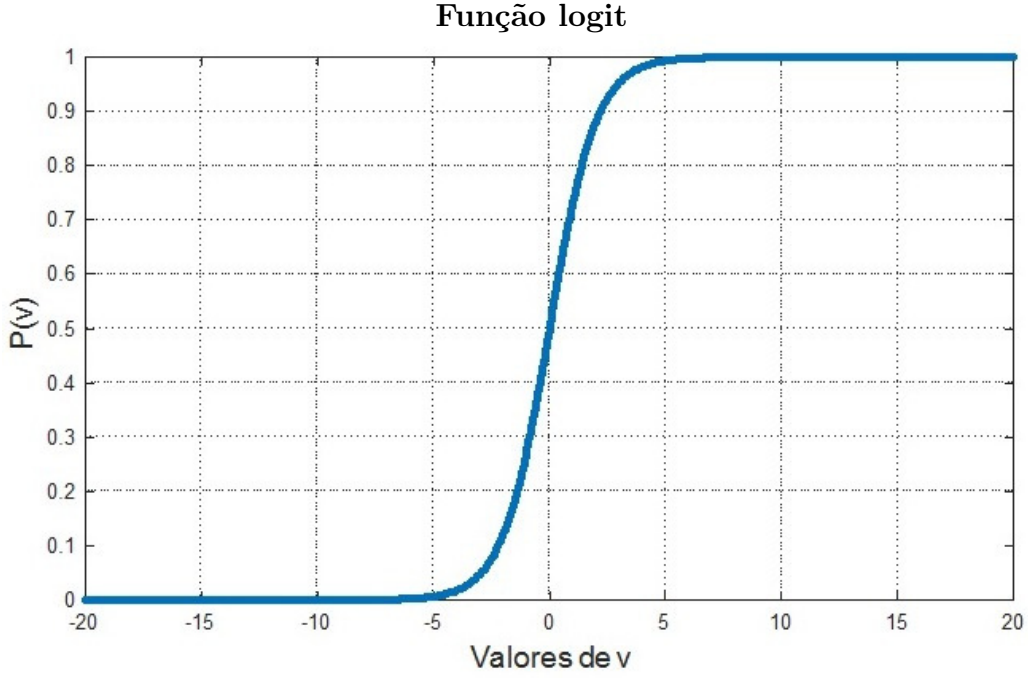


Figura 2.7. Gráfico da função logit $P(v) = \dfrac{e^v}{1+e^v}$, onde $v = \sum_{i=1}^{m} \alpha_i x_i$

Para proceder à determinação dos valores de α, faz-se uma transformação linear que remeterá o problema à solução de um problema de álgebra linear. Seja a chance $C_j(x)$ é definido por:

$$C_j(x) = \frac{P_j(x)}{1 - P_j(x)}. \tag{2.2}$$

A expressão da equação (2.2) usando (2.1), temos

$$C_j(x) = e^{\sum_{i=1}^{m} \alpha_i x_i} \tag{2.3}$$

Tomando-se o logaritmo na base e em ambos os lados de (2.3), obtemos um sistema de equações lineares para determinar α:

$$b_j = \sum_{i=1}^{n} \alpha_i x_i \tag{2.4}$$

onde $b_j = \ln(C_j)$, para j=1,2,...m.

Associamos ao conjunto de dados uma matriz $A = \{a_{i,j}\} \in \mathbb{R}^{m \times n}$. As linhas representam as entidades e as colunas estão associadas aos atributos. Assim, o valor de cada $a_{i,j}$ é o valor do atributo j na proteína i. É importante observar que isto é diferente da forma que tratamos anteriormente, quando discutimos a decomposição por valores singulares (esta matriz é a transposta da anterior).

Seja $b = (b_1, b_2, ..., b_m)^T$ o sistema de equações lineares (2.4) pode ser representado por:

$$A\alpha = b \tag{2.5}$$

A seguir discutimos formas de atribuir valores à α_i.

1. Abordagem clássica Para aplicar a Regressão Logística, tradicionalmente, supõe-se $m > n$ e os parâmetros α são computados resolvendo-se o seguinte problema de programação não linear irrestrito que minimiza a somatória dos quadrados dos resíduos associados à cada linha do sistema (2.5):

$$Minimize \;\; q(\alpha) = ||A\alpha - b||^2 \tag{2.6}$$

Se A^TA tem posto completo, a solução única de (2.6) é dada resolvendo o sistema de equações lineares de ordem n:

$$A^TA\alpha = A^Tb \tag{2.7}$$

2. Modelo quadrático para problemas com mais atributos que entidades Neste caso temos $m < n$ e a matriz A^TA não tem rank completo. Geralmente, para contornar essa dificuldade, descarta-se um grupo das variáveis (*feature selection*), mantendo apenas um subconjunto das variáveis originais. Usamos um termo estabilizador no modelo de Regressão Logística, encontrado nos trabalhos de [Morais R. F. et al. 2020], que permite a atribuição de valores para os parâmetros α que minimizam a soma dos quadrados dos resíduos $(A\alpha - b)$, adicionado aos quadrados de α. Assim, para encontrar uma solução a (2.5), resolvemos um problema otimização quadrática irrestrito dado por

$$Minimize \;\; q(\alpha) = ||\alpha||^2 + ||A\alpha - b||^2 \tag{2.8}$$

Como a função $q_q(\alpha)$ é convexa, o argumento α^* que minimíza (2.8) é dado pela derivação de $q_q(\alpha)$ em relação à α e igualando-se o resultado a zero. Isto resulta em um sistema de equações lineares

$$(I + A^TA)\alpha = A^Tb, \tag{2.9}$$

onde I é a matriz identidade da ordem n. Ou seja, a solução ótima para α em (2.8) é obtida pela solução de (2.9) e é única.

Então, dada uma proteína (*query*) $q = (q_1, q_2, ..., q_n)$ com n atributos, a probabilidade de q pertencer a uma classe associada ao sistema relacionado é dado por:

$$P(q) = \frac{e^{q\alpha}}{1 + e^{q\alpha}}. \tag{2.10}$$

Saliente-se que (2.9) é para o caso em que o número de incógnitas é maior que o número de equações, que geralmente ocorre com o método **Adhoc**, visto abaixo.

2.3.1 Método Adhoc

Adhoc é uma expressão latina cuja tradução literal é "para isto" ou "para esta finalidade". É usada para designar algo ou alguma coisa que foi formada ou usada para um propósito ou necessidade particular e imediata, sem planejamento prévio.

A metodologia descrita a seguir, aqui intitulada **Adhoc**, usa modelos construídos exclusivamente em resposta a uma única demanda por classificação. Dada uma consulta q, um modelo *adhoc*, específico para esta consulta, é construído a partir da escolha de k_0 entidades mais próximas à q tais que $P_j(x) = 0$, ao lado de k_1 entidades escolhidas dentre aquelas mais próximas à q tal que $P_j(x) = 1$.

Os valores de k_0 e k_1 são determinados experimentalmente. Nos nossos ensaios, constatamos que estes valores são baixos; na validação cruzada de 10 partições (é usado a validação cruzada para comparar os resultados do modelos propostos com os resultados do artigo [Pires *et al.*, 2011]), verificamos que o desempenho da classificação é superior quando atribuímos valores na ordem de algumas unidades a estes dois parâmetros. Isto tem como consequência a construção de matrizes de atributos $M \in \mathbb{R}^{m \times n}$ nas quais $m < n$, o que impede o uso da Regressão Logística tradicional. Resolvemos esta limitação usando a Regressão Logística modificada segundo explicado na seção anterior.

Outra dificuldade está na escolha das entidades k_0 e k_1 mais próximas, o que poderia impactar o tempo de resposta nos problemas de grande porte. Isto é resolvido organizando as entidades em árvores de pesquisa segundo os recursos usados em máquinas de busca (**svd**, clusterização etc). A matriz de atributos original A é particionada em duas outras, A_0 e A_1, segundo $P_j(x) = 0$ ou 1; com cada uma delas organizada enquanto elementos de uma máquina de busca. Assim, a recuperação das entidades mais próximas a uma consulta q fica extremamente eficiente e não impacta o tempo de processamento de forma perceptível.

Concluindo, o método **Adhoc** é um recurso para a Regressão Logística que permite a sua aplicação em cenários desbalanceados. O algoritmo 1 é dado a seguir.

Algoritmo 1: Método Adhoc

Entrada: q : consulta q, A1, A0
Saída: P: Probabilidade de q pertencer ao grupo 1
A1: Conjunto de entidades que pertencem a categoria 1;
A0: Conjunto de entidades que não pertencem a categoria 0;
M1 $\leftarrow$ Conjunto dos k_1 elementos mais próximos de q em A1;
M0 $\leftarrow$ Conjunto dos k_0 elementos mais próximos de q em A0;
M $\leftarrow$ [M1; M0];
$i_1 \leftarrow 1$, $\forall$ elemento de M1;
$i_0 \leftarrow 0$, $\forall$ elemento de M0;
i $\leftarrow [i_1; i_0]$;
P $\leftarrow$ Aplique a Regressão Logística modificada para M e i

Tabela 2.1. Regressão Logística para ambientes desbalanceados.

Capítulo 3

Resultados

Neste capítulo, apresentamos a experiência computacional usando três conjuntos de dados: um deles congrega entidades sob aspectos funcionais (*golden standard*); um outro é uma coleção de quatro conjuntos desbalanceados que usam a própria estrutura (SSE's); e, finalizando, usamos o *full* **SCOP**, uma coleção tradicional de estruturas. Para todos os ensaios, usamos cristais depositados no **PDB** para extrair as coordenadas dos átomos para os classificadores.

Os átomos considerados neste trabalho para a construção do **VSM** foram, majoritariamente, os da cadeia principal (ou parte dela), constituída seja pelos C_α, seja pelos átomos (C_α, C, N). Mas também, em alguns casos, utilizamos todos os átomos da proteína.

Planejamos os experimentos para inicialmente responder uma primeira pergunta: aquela que tange a adequação dos nossos algoritmos à classificação de estruturas de proteínas. Os métodos que usamos são interessantes para avaliar atributos e, em outras aplicações, como por exemplo, aquelas que utilizam *microarrays*, o efeito colateral que no caso é identificar marcadores biológicos de uma doença, tem resultado em contribuições interessantes [Abreu *et al.*, 2019], [Santos *et al.*, 2017], [Morais, Rodrigues, F., et al., 2020]. Assim, passamos a testar os métodos descritos na seção anterior para seguir a busca de assinaturas estruturais para atender ao nosso objetivo.

No que se segue, apresentamos os resultados da Regressão Logística em dois conjuntos: um que se apresenta muito bem balanceado (*golden standard*) e outro sem este aspecto (SSE). Com os resultados que alcançamos, buscamos identificar as assinaturas usando como **VSM** o *cutoff scanning* . Mas, não tivemos sucesso e passamos a trabalhar com um novo *VSM*, constituído por atributos da decomposição espectral da matriz de distâncias entre os átomos. Com esta nova forma de identificar as entidades, obtivemos resultados muito próximos ao que consideramos como oráculo perfeito (média harmônica na validação cruzada de 10 partições acima de 0,95).

É importante falar sobre as métricas que utilizamos para aferir o desempenho dos classificadores. Utilizando-se as métricas de sensibilidade (SEN $= \frac{VP}{(VP+FN)}$), especificidade

(ESP = $\frac{VN}{(VN+FP)}$), média harmônica entre sensibilidade e especificidade (MH=$\frac{2*SEN*ESP}{(SEN+ESP)}$) e área sob a curva (ROC).

Utiliza-se a validação cruzada neste trabalho para fins de comparação, e consiste no particionamento dos dados de entrada em n conjuntos, em que $n-1$ partes serão utilizadas para a construção do modelo. Denominamos tais partes como conjunto de treino. A parte restante é utilizada, então, para avaliar a adequação do modelo. Essa parte é denominada conjunto de teste. Esse procedimento é repetido n vezes, variando sistematicamente os conjuntos de treino e de teste, são calculadas as médias das métricas de qualidade dos classificadores para as n execuções. Neste caso, dizemos que executamos a validação em n-partições.

3.1 Validação da Regressão Logística como instrumento para classificação estrutural - caso 1: conjuntos balanceados

Neste item, estão os resultados com modelos construídos com *cutoff scanning* [Pires DE, de Melo-Minardi RC, dos Santos MA, Santoro MM, et al., 2011], no conjunto *golden standard*, cuja característica que mais nos chama a atenção é o bom balanceamento entre as entidades em todos os conjuntos. Na referência citada, os melhores resultados foram com a matriz de distâncias entre os carbonos alfa C_α da proteína em que o posto foi reduzido usando **svd**. Vários métodos tradicionais de mineração de dados foram utilizados.

Nossos melhores resultados foram alcançados usando a regressão tradicional com as coordenadas dos átomos da cadeia principal - ver Tabela 3.1. Usando a Regressão Logística com a estrategia **Adhoc**, o desempenho foi levemente inferior - ver Tabela 3.2. Resultados de outros experimentos usando C_α encontram-se nos anexos (Tabela C.1). Possivelmente, embora excepcionais, estes resultados possam ser melhorados usando como em Pires [Pires *et al.*, 2011], a decomposição por valores singulares. Neste primeiro momento, como nos interessa avaliar atributos para entender uma possível assinatura, evitamos este recurso, pois estaríamos perdendo um certo mapeamento dos atributos originais, dado que o **svd** trabalha em espaços projetados.

Cumpre observar que tivemos dificuldade para comparar diretamente os nossos resultados com aqueles obtidos em Pires [Pires *et al.*, 2011]. O desempenho dos métodos é aferido a partir de 10 *folders* aleatoriamente construídos. Não há, na referência citada, uma sugestão de *data set* para testes. Além disto, adotamos, por considerar mais adequado ao nosso problema, critérios de avaliação baseados na média harmônica e curva **roc**. Entretanto, acreditamos que os nossos resultados seguem semelhantes àqueles que nos servem de referência.

Resultados da Regressão Logística no conjunto *golden standard*

Superfamilia	Sensibilidade. (Média)	Especificidade. (Média)	Media Harmônica (Média)	Curva ROC (Média)
amidhydrolase	0,9756	0,9581	0,9664	0,9597
crotonase	1,0000	0,9890	0,9944	0,9984
enolase	0,9779	0,9931	0,9850	1,0000
haloacid dehalogenase	0,9571	0,9573	0,9551	0,9587
isoprenoid synthase type I	1,0000	0,9994	0,9997	1,0000
vicinal oxygen chelate	1,0000	0,9968	0,9984	1,0000

Tabela 3.1. Com a regressão tradicional alcançamos a média harmônica (média) 98,31% , sensibilidade média de 98,51% e uma especificidade média de 98,23%. As coordenadas dos átomos para a construção da matriz de distâncias foram os da cadeia principal.

Resultados do método Adhoc no conjunto *golden standard*

Superfamilia	Sensibilidade. (Média)	Especificidade. (Média)	Media Harmônica (Média)	Curva ROC (Média)
amidhydrolase	0,9699	0,9419	0,9553	0,9794
crotonase	0,9818	0,9643	0,9727	0,9836
enolase	0,9351	0,9194	0,9266	0,9584
haloacid dehalogenase	0,9737	0,9559	0,9643	0,9831
isoprenoid synthase type I	1,0000	0,9742	0,9866	0,9926
vicinal oxygen chelate	0,9533	0,9657	0,9587	0,9734

Tabela 3.2. Regressão Logística em *data sets* bem balanceados: o método Adhoc alcançou 96,07% de média harmônica (média), sensibilidade média de 96,90% e uma especificidade média de 95,36%, usando uma matriz de distâncias construída a partir das coordenadas dos átomos carbono alfa (C_α) da proteína.

3.2 Validação da Regressão Logística como instrumento para classificação estrutural - caso 2: conjuntos não balanceados

Este é um bom exemplo de como o algoritmo tradicional de Regressão Logística aplicado em *data sets* desbalanceados (onde um dos conjuntos tem um número alto de entidades e outro um número pequeno de entidades) não funciona; a média harmônica oscila em torno de 50% (estes resultados não estão mostrados aqui). Nos problemas a seguir, sempre aplicaremos o método **Adhoc**.

Utilizamos os quatro conjuntos de dados *SSE's* (base de pequenas proteínas *SSEs*) [Jain *et al.*, 2010], que são congregados por estruturas, cada um com mais de 50 superfamílias com várias classificações. Os resultados estão na Tabela 3.3. O valor médio entre as médias harmônicas foi de 94,32 %. Foram utilizados os átomos da cadeia principal, sinalizado nos resultados anteriores como a estratégia que pode alcançar os melhores resultados.

Em uma análise superficial, dado que não sabemos exatamente como os testes em Pires [Pires *et al.*, 2011] foram realizados, nos parece que os nossos resultados são ligeiramente inferiores. Embora os oráculos não sejam oráculos perfeitos na nossa acepção, o desempenho é muito bom e sinaliza a efetividade do método **Adhoc**.

Resultados do método Adhoc em bases desbalanceadas

Conjuntos	Sensibilidade. (Média)	Especificidade. (Média)	Media Harmônica (Média)	Curva ROC (Média)
3SSEs	0,9450	0,9556	0,9483	0,9659
4SSEs	0,9644	0,9664	0,9645	0,9824
5SSEs	0.8960	0,9451	0,9111	0,9148
6SSEs	0,9434	0,9599	0,9489	0,9711

Tabela 3.3. O desempenho em média dos conjuntos SSEs foi: 94,32% de média harmônica (média), sensibilidade média de 93.37% e uma especificidade média de 95,68%.

De acordo a Tabela 3.3, para bases desbalanceadas o método Adhoc resolve o problema de classificação. Para a base SCOP [Berman *et al.*, 2000] uma base desbalanceada, será analisado pelo método Ahdoc com um novo VSM *spectral pattern* na seção 3.4.2.

3.3 Busca de assinaturas estruturais usando o VSM *cutoff scanning* com o método Adhoc

Assim, dado que os resultados alcançados foram animadores, passamos à próxima etapa que é a definição da assinatura, escolhendo e procurando interpretar os atributos mais importantes na classificação. A Regressão Logística indica isto a partir dos valores de α; usam-se os valores de α_i mais positivos e os mais negativos. Por exemplo, no gráfico mostrado na figura (3.1), escolheríamos alguns atributos que estivessem topologicamente mais distantes do eixo das abscissa. Aqueles atributos cujos valores de α_i são mais próximos de zero não têm poder discriminatório.

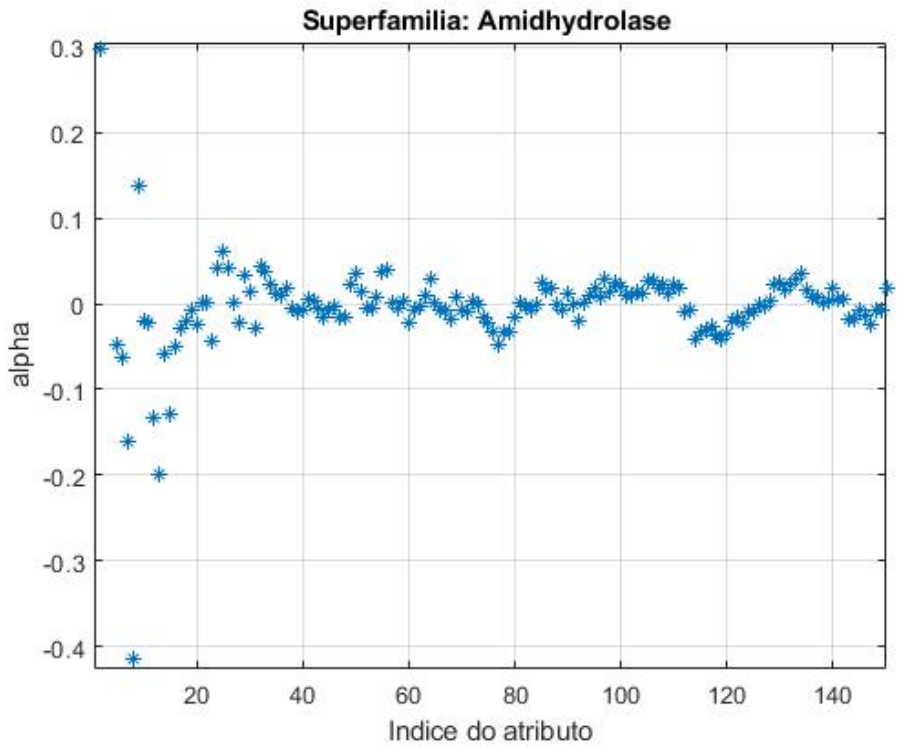


Figura 3.1. Valores dos pesos (α_i) obtidos em um modelo de Regressão Logística, para a superfamília Amidhydrolase do conjunto de dados "gold standard". Foram usados os átomos da cadeia principal.

Por exemplo, para o oráculo da superfamília *Amidhydrolase*, escolheríamos os atributos que são mostrados na Tabela 3.4. O próximo passo consiste em validar esta escolha de atributos. Caso os oráculos construídos somente com estes atributos sejam eficientes, tem-se uma prova de conceito, como utilizamos em outros problemas da área, quando buscávamos marcadores biológicos.

Alfa	Intervalos (ångströms)
Negativos	$(7,4\ ,\ \ 7,6)$
	$(12,0\ ,\ \ 12,2)$
	$(6,0\ ,\ \ 6,2)$
	$(5,6\ ,\ \ 5,8)$
	$(7,2\ ,\ \ 7,4)$
Positivos	$(5,8\ ,\ \ 6,0)$
	$(2,6\ ,\ \ 2,8)$
	$(1,2\ ,\ \ 1,4)$
	$(1,0\ ,\ \ 1,2)$
	$(1,6\ ,\ \ 1,8)$

Tabela 3.4. Escolha de atributos baseados nos parâmetros α_i da superfamília Amidhydrolase. Foram utilizados os átomos da cadeia principal.

Mas, infelizmente, um fato saltou às vistas quando passamos a analisar seleções de

atributos em cada um dos oráculos criados para cada consulta no método **Adhoc**. Não houve consenso: os atributos mais importantes variavam e não conseguimos algo que fosse robusto e estável. Isto nos levou a abandonar o *VSM* proposto por Pires [Pires *et al.*, 2011]. A seguir discutimos o novo *vector space model*, o *spectral pattern*.

3.4 Resultados com o VSM spectral pattern

Nesta seção apresentamos os experimentos computacionais utilizando os conjuntos SSE's e o SCOP (*full*). Como já havíamos decidido, continuamos utilizando os átomos da cadeia principal e, claro, dado o desbalanceamento (onde um dos conjuntos tem um número alto de entidades e o outro um número pequeno de entidades) destes *data sets*, usamos a Regressão Logística com o método **Adhoc**.

Antes de seguir com os testes, apresentamos 2 ensaios que podem ajudar a entender a especificidade das matrizes de distâncias construídas a partir das coordenadas dos átomos de uma proteína.

3.4.1 Preâmbulo: a exclusividade de uma matriz com as distâncias entre os átomos de uma proteína

Para ilustrar a especificidade dos valores singulares das matrizes D obtidas a partir de estruturas terciárias das proteínas, consideramos uma matriz simétrica D cujos valores foram gerados segundo distribuições uniforme e normal. Esta matriz tem o mesmo tamanho e valores no mesmo intervalo de uma matriz de distâncias reais (proteína 1A4M).

As figuras 3.2 e 3.3 tratam de uma matriz D construída segundo uma distribuição uniforme.

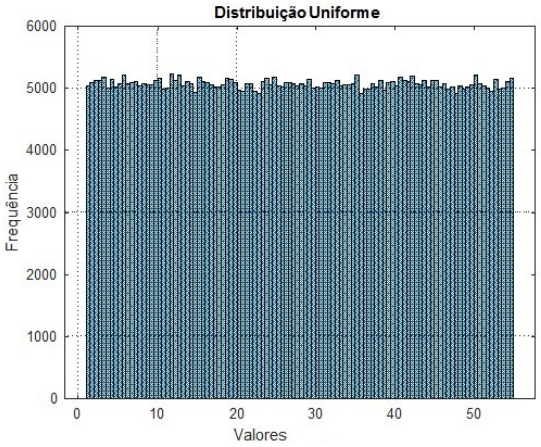


Figura 3.2. O Histograma dos 547581 valores aleatórios gerados a partir de uma distribuição uniforme no intervalo (1.31 , 55.17), simulando uma proteína real.

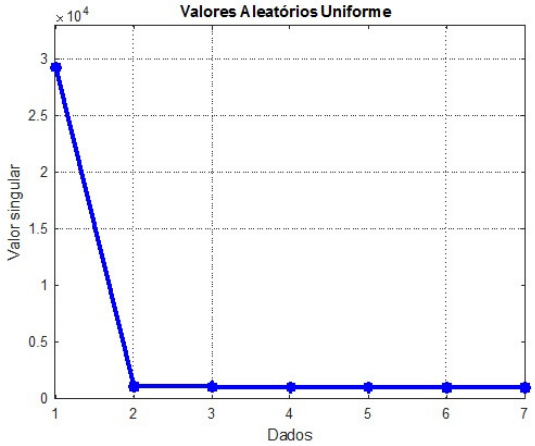


Figura 3.3. Valores singulares de uma matriz de números aleatórios uniformemente distribuídos. São mostrados os seis primeiros de um total de 1047; os valores omitidos correspondem a uma cauda longa tendendo para zero.

Já para uma matriz D obtida segundo uma distribuição normal, tem-se o resultado mostrado nas figuras 3.4 e 3.5.

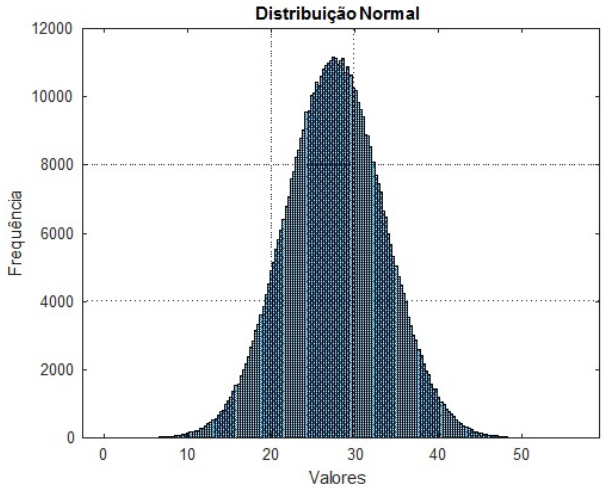


Figura 3.4. O Histograma dos 547581 números aleatórios gerados a partir de uma distribuição normal no intervalo (1.35 , 55.17) simulando a proteína real.

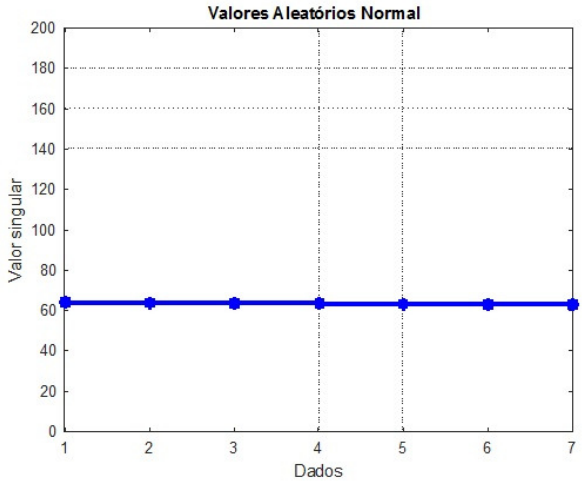


Figura 3.5. Valores singulares de uma matriz de números aleatórios segundo uma distribuição normal (normalizado por $z = \frac{x-\mu}{\sigma}$ os valores da figura (3.4)). São mostrados os seis primeiros de um total de 1047; os valores omitidos correspondem a um decréscimo constante tendendo para zero.

Formulamos a seguinte hipótese que ainda carece de confirmação e que procuraremos validar na sequência deste trabalho. Considere uma matriz de distâncias entre os átomos (ou parte deles) de uma proteína qualquer D_p cuja decomposição espectral (**svd**) é dada por (U_p, S_p, V_p). Se D é uma matriz de números aleatórios, no mesmo intervalo de D_p seguindo uma distribuição uniforme e com a decomposição dada por (U, S, V), então, tem-se que

$$D_p = US_pV^t.$$

Ou seja, os vetores singulares (padrões) de uma matriz de distâncias entre os átomos de uma proteína são os **mesmos** de uma matriz da mesma ordem, construída com números aleatórios e no mesmo intervalo das distâncias que existem na proteína, seguindo uma distribuição uniforme. O que altera no espectro da matriz de distâncias da proteína são somente os valores singulares (pesos) associados a cada padrão. Nos experimentos que fizemos até o presente momento, isto se confirmou, tanto no caso em que se considera todos os átomos da proteína, quanto no caso em que se considera somente uma parte deles (cadeia principal ou C_α).

Esta hipótese, caso seja confirmada experimentalmente, corrobora a importância do nosso *vector space model* que tem como atributos os valores do espectro da matriz de distâncias da proteína. Observe que isto pode ser um resultado marcante, podendo dar origem a uma série de ideias para serem aproveitadas em algoritmos da área de Bioinformática estrutural.

3.4.2 Resultados do VSM spectral pattern use

Usamos o Algoritmo 2 para obter o conjunto de entidades que será representado pelos valores singulares de cada matriz de distâncias. Dado um conjunto de arquivos no formato **PDB**, para cada um, extrai-se as coordenadas dos átomos da cadeia principal e depois computa-se a matriz de distâncias. Dessa matriz, procede-se à decomposição por valores singulares e extrai-se os seus valores singulares que representarão cada uma dessas estruturas que estão descritas no formato **PDB** .

Algoritmo 2: VSM: Spectral Pattern Use

Entrada: pdbs (Proteínas)
Saída: A: Matriz de entidades e atributos
para *pdb* $\in$ *(pdbs)* **faça**
  [x y z] $\leftarrow$ Coordenadas x, y e z dos átomos C_α, C, N retiradas do pdb;
  D $\leftarrow$ Distancia entre todos os átomos, dadas pelas coordenadas x, y e z;
  [U S V] $\leftarrow$ svd(D);
  s=diag(S);
  s $\leftarrow$ s(1:6): Escolha dos valores singulares mas significativos de S;
  A $\leftarrow$ [A; s];
fim

Tabela 3.5. Geração do *vector space model* para arquivos **PDB**.

A Tabela (3.6) apresenta o desempenho do método **Adhoc** para o segundo *data sets* [Jain *et al.*, 2010] contem entidades desbalanceadas. Observa-se que o desempenho do método, enquanto processo de classificação, os resultados de [Pires *et al.*, 2011] obtidos são tão bons quanto os nossos resultados.

Método Adhoc usando o *spectral pattern use*

Conjunto de dados	Sensibilidade. (Média)	Especificidade. (Média)	Media Harmônica (Média)	Curva ROC (Média)
3SSE	0,9661	0,9173	0,9391	0,9445
4SSE	0,9934	0,9221	0,9299	0,9161
5SSE	0,9713	0,9662	0,9684	0,9710
6SSE	0,9388	0,9326	0,9310	0,9564

Tabela 3.6. Para os conjuntos não balanceados *SSE's* usando o método Adhoc e o novo *VSM*, alcançamos uma média harmônica (média) 94,21%, sensibilidade média de 96,74% e uma especificidade média de 93,46%. Foram usados os átomos da cadeia principal para construir a matriz de distâncias.

A Tabela (3.7) apresenta o desempenho do método **Adhoc** para o terceiro *data sets*

SCOP [Berman *et al.*, 2000] que contem entidades desbalanceadas e com o novo *vector space model: Spectral Pattern* observa-se que o desempenho do método Adhoc, enquanto processo de classificação os nossos resultados são tão bons comparados aos de [Pires *et al.*, 2011].

Método Adhoc usando o *spectral pattern use*

Superfamília	Sensibilidade. (Média)	Especificidade. (Média)	Média Harmônica (Média)	Curva ROC (Média)
Globin-like*	0,9105	0,9062	0,9105	0,9567
Alpha-helical ferredoxin*	0,9169	0,9344	0,9255	0,9563
Chaperone J-domain*	0,9282	0,9538	0,9406	0,9685
Recombination endonuclease VII, C-terminal and dimerization domains*	0,9419	0,9534	0,9473	0,9690
Prefoldin*	0,8918	0,9199	0,9054	0,9365
HR1 repeat*	0,9282	0,9403	0,9339	0,9548
tRNA-binding arm*	0,9080	0,9313	0,9192	0,9431
Eukaryotic DNA topoisomerase I, dispensable insert domain *	0,9033	0,9229	0,9126	0,9432
C-terminal UvrC-binding domain of UvrB*	0,8888	0,9416	0,9135	0,9360
Epsilon subunit of F1F0-ATP synthase C-terminal domain*	0,9490	0,9622	0,9553	0,9703
Fe,Mn superoxide dismutase (SOD), N-terminal domain*	0,9204	0,9539	0,9366	0,9563
Sporulation inhibitor Sda*	0,9380	0,9660	0,9516	0,9605

Tabela 3.7. Para a base de dados SCOP usando o método Adhoc e o novo *VSM*, alcançamos uma média harmônica (média) 93,01%, sensibilidade média de 93,03% e uma especificidade média de 93,04% para Backbone. * formado por todas as superfamílias independente das Classes e Folds.

Entretanto, como nosso objetivo principal é buscar por assinaturas estruturais, resolvemos seguir em busca de oráculos perfeitos. Algumas trilhas apareceram para melhorar o método **Adhoc**. Na próxima seção, delineamos a mais promissora delas.

3.5 Novos métodos

3.5.1 Usando Programação Linear

Determinar uma direção para a dilatação de espaços de Shor é definir um hiperplano

$$H(x) = \{x \in \mathbb{R}^n : d'x = 0\}$$

Resolveremos o problema através de otimização e mais especificamente, pelo fato de o problema de viabilidade algébrica linear ser um problema de programação linear (PPL) clássico.

Consideremos $A = [A_0; A_1]$ uma partição de $A \in \mathbb{R}^{m \times n}$ e $A_0 \in \mathbb{R}^{m0 \times n}$, $A_1 \in \mathbb{R}^{m1 \times n}$; I_{m0}, I_{m1} matrizes identidade de ordem m_0 e m_1, $Z_0 \in \mathbb{R}^{m0 \times m1}$ e $Z_1 \in \mathbb{R}^{m1 \times m0}$ matrizes de zeros.

Consideremos $y = \{x, x_f, x_s\}$ com $x \in \mathbb{R}^n,\ x_f \in \mathbb{R}^{m0}$ e $x_s \in \mathbb{R}^{m0}$, as variáveis x_f e x_s são variáveis artificiais acrescidas às restrições relativas à A_0 e à A_1 no problema de viabilidade algébrica linear (Problema 2) a seguir.

Problema 1 (melhor direção)
$Minimize\ f(x) = c_1'x$
Subject to
$|x_i| \leq 1$ para $i = 1, 2...n$ {não é restrição; colocar como upper e lower bounds das variáveis: $x_i \leq 1$ e $-x_i \leq 1$}
onde $c_1 = [-sum(A_1) + sum(A_0)]'$
O operador $sum(A)$ produz um vetor linha com a soma das colunas de uma matriz A.

Problema 2 (Viabilidade Algébrica Linear)
$Minimize\ f(y) = c2'y$
Subject to
$By \leq 0$
$y(n+1 : n+m0+m1) \geq 0$

onde B= $\begin{bmatrix} A_0 & -I_0 & Z_0 \\ -A_1 & Z_1 & -I_1 \end{bmatrix}$, $c2 = [zeros(n,1); ones(m0,1); ones(m1,1)]$
Os problemas a seguir usam os resultados da viabilidade algébrica linear. No que se segue, x_f, x_s são variáveis artificiais que estão relacionados ao argumento y^* que minimiza a função objetivo do Problema 2.

O Problema 2, assinala valores para os pesos α no contexto em que se faz necessário ter um conjunto reduzido de atributos.

Problema 3 (Nova Regressão Logística)
$Minimize\ f(x) = \alpha' * \alpha$
Subject to
$A_0\alpha \leq b_f$
$-A_1\alpha \leq b_s$
$\alpha \in \mathbb{R}^n$

onde $b_f = y^*(n+1 : n+m0)$ e $b_s = y^*(n+m0+1+1 : end)$ do Problema 2.

Observa-se que este problema pode ficar irrestrito usando-se como penalidade os valores das variáveis duais dos PPL's. Com isto consegue-se resolver o problema simplesmente a partir de um sistema de equações lineares como é feito na Regressão Logística modificada.

Pode ser interessante associar o conceito de ambiguidade (seção 3.5.2) às entidades cujos valores de $x_f = y^*(n+1 : n+m0)$ e $x_s = y^*(n+m0+1+1 : end)$, obtidos no Problema 2, sejam maiores que zero e esto sera feito na seção 3.5.2.

O valor das variáveis duais podem ser utilizadas para penalizar as entidades ambíguas ao se buscar uma direção para dilatação. O ideal é que estas entidades alterem o mínimo possível, ficando confinadas em uma região bem definida no espaço após a dilatação.

Na discussão que se segue, abordaremos o problema mais referenciado na literatura da mineração de dados: o conjunto de dados de flores do tipo Iris [Unwin *et al.*, 2021] e [Fisher *et al.*, 1936] (figura 3.6). Ele foi a inspiração de muitos métodos que apareceram depois que este problema foi publicado. O domínio consta de 150 entidades, representando plantas da espécie Iris, descritas por 4 atributos reais (comprimento e largura de pétalas e sépalas). Estão categorizados em 3 classes: Iris Setosa, Iris *Versicolour* e Iris *Virginica*, com 50 exemplares cada.

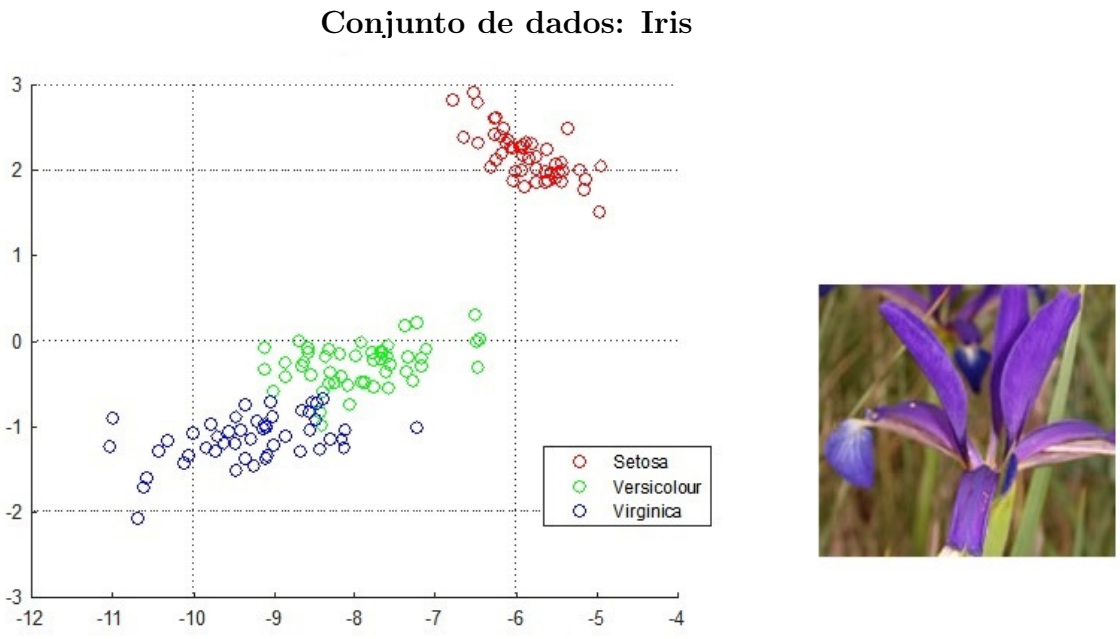


Figura 3.6. Visualização da Iris no espaço R^2 aplicando SVD.

O domínio é extremamente simples: uma das classes é separada das outras e é aquela em que os algoritmos se saem muito bem. A pior dentre as três para a classificação é a que está entre as outras duas - a Iris Versicolour. Os resultados da validação cruzada em 10 folderes estão apresentados na Tabela 3.8.

Resultados da Regressão Logística tradicional

Classes	Médias Harmônicas
Setosa	1,0000
Versicolour	0,7890
Virginica	0,9314

Tabela 3.8. O desempenho da base de dados Iris, aplicando Regressão Logística tradicional. As métrica de sensibilidade, especificidade e média harmônica são médias da validação cruzada de 10 folders.

A Tabela 3.9 apresenta o desempenho da Iris usando a logística tradicional e Programação Linear. Observa-se que o desempenho do método PPL, enquanto processo de classificação, foi muito melhor em comparação a logística tradicional.

Resultados da Iris

Classes	Médias Harmônicas (Tradicional)	Médias Harmônicas (Progamação Linear)
Setosa	1,0000	1,0000
Versicolour	0,7890	0,9800
Virginica	0,9314	0.9898

Tabela 3.9. O desempenho da base de dados Iris. Foi calculada a média harmônica (médias) da validação cruzada de 10 folders. 1ª coluna usando Regressão Logística tradicional e a 2ª coluna adotamos o método Adhoc usando programação Linear.

Observa-se que como a PPL funciona muito bem para o exemplo tradional da Iris, por isso ela foi aplicada em uma da nossas base de dados. A Tabela 3.10 apresenta o desempenho usando PPL em *data set golden standard*. Observa-se o desempenho de 98% em média harmonica do método PPL, enquanto processo de classificação o modelo é considerado oráculo perfeito.

Método Adhoc usando o programação linear

Conjunto de dados	Sensibilidade. (Média)	Especificidade. (Média)	Média Harmônica (Média)	Curva ROC (Média)
amidhydrolase	0,9807	0,9910	0,9857	0,9818
crotonase	0,9967	0,9891	0,9927	0,9983
enolase	0,9911	0,9644	0,9774	0,9688
haloacid dehalogenase	1,0000	0,9650	0,9820	0,9780
isoprenoid synthase type I	0,9400	0,9784	0,9562	0,9339
vicinal oxygen chelate	0,9927	0,9968	0,9947	0,9945
Média	0,9839	0,9808	0,9815	0,9760

Tabela 3.10. Desempenho na base de dados *gold standard* usando a programação linear: alcançamos uma média harmônica (média) 98%, sensibilidade média de 98% e uma especificidade média de 98%. Foram usados os átomos da cadeia C_α.

3.5.2 Ambiguidades

Tratamento de ambiguidades não é um tema visto com frequência na área de mineração de dados. Em geral, entidades ambíguas são tidas como um preço a pagar pelo processo de modelar a realidade. Esta modelagem é sempre uma simplificação, isto é, diminuir as entidades.

Entretanto, modelos que lidam com a biologia têm um tema presente no sentido em que as entidades que representam uma dada realidade nem sempre são perfeitamente dicotomizáveis. Este esforço em embutir a ambiguidade nos modelos é perfeitamente coerente com a área que estamos atuando. O método **Adhoc**, que discutimos, oferece uma oportunidade para avançar neste tema.

Explorando a separabilidade e assumindo que os atributos de entidades que se imiscuem com as outras geram ambiguidades, estamos propondo um novo algoritmo para tratar essa situação.

3.5.3 Resolvendo a ambiguidade: método Adhoc Extendido

A presente proposta leva em conta que a separação topológica entre as classes de entidades é fundamental para o desempenho dos algoritmos baseados em Regressão Logística. A separabilidade aqui é entendida como a existência de um hiperplano que separa das demais os membros da classe que se deseja classificar. Como exemplo, discutimos, na seção 3.5.1, o problema de classificação da Iris; a Iris Setosa congrega um conjunto de entidades perfeitamente separadas das demais. Isto não ocorre com as outras, impactando negativamente o desempenho da regressão Tabela 3.4. Assim, para lidar com problemas de separabilidade e inspirados no modelo **Adhoc**, sugerimos a construção de vizinhanças predefinidas como sugere a figura (3.7)

O domínio é previamente mapeado em vizinhanças B e $\bar{B}$, indicadas por bolas de duas cores (vermelha e azul na figura), dependendo da predominância de uma classe frente

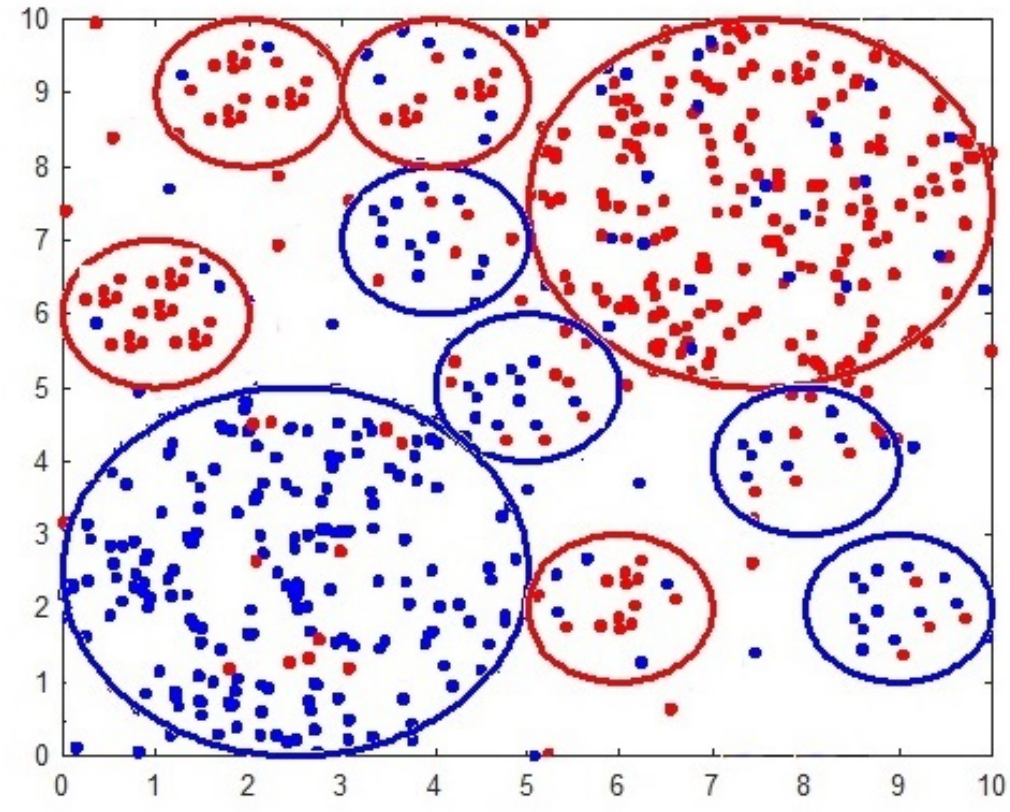


Figura 3.7. Visualização de um domínio com vizinhanças predefinidas.

à outra. Considera-se que nestas vizinhanças, no caso mais difícil, contenha somente os elementos de uma única classe.

Para montar os modelos, escolhem-se as duas vizinhanças B e $\bar{B}$ (duas bolas de cores diferentes) mais próximas à projeção da consulta no espaço em que estão representadas as entidades. Supondo que estas vizinhanças sejam o caso mais complexo, em que ambas contêm elementos dos dois grupos - membros da classe que se deseja classificar junto com os demais -, o desfecho da consulta é dado pela construção **Adhoc** de dois modelos; P e $\bar{P}$. O modelo P é construído com as entidades da bola B que são da classe que se deseja classificar com as entidades da bola $\bar{B}$ que não são desta classe. Observe que este modelo contém entidades perfeitamente separáveis. Já o modelo $\bar{P}$ é construído com as entidades da classe que se deseja classificar, mas em $\bar{B}$ junto com aqueles que estão em B e que não são da classe desejada. Seja $d = |p(q) - \bar{p}(q)|$, onde q é a consulta e $p(q)$ e $\bar{p}(q)$ são as probabilidades computadas pelos modelos P e $\bar{P}$. Se d for próximo de 1, q é considerado ambígua e o desfecho é dado por $\bar{p}(q)$; caso contrário, se d for próximo de zero, o desfecho é dado por $p(q)$.

Na discussão que se segue, aplicamos o modelo de ambíguos ao problema da Iris e chegamos aos seguintes resultados:

Resultados da Iris: Modelo de ambíguos

Classes	Médias Harmônicas (Tradicional)	Médias Harmônicas (Ambíguos)
Setosa	1,0000	0.9970
Versicolour	0,7890	0,9882
Virginica	0,9314	0.9310

Tabela 3.11. O desempenho da base de dados Iris. Foi calculada a média harmônica (médias) da validação cruzada de 10 folders. 1ª coluna usando Regressão Logística tradicional e a 2ª coluna adotamos o método Adhoc usando o primeiro modelo de ambíguos.

Na Tabela 3.11 claramente observa-se que o primeiro modelo de ambíguos é melhor que o modelo tradicional dicutidos no preâmbulo do inicio da seção 3.5.3 para classe da iris Versicolour, chegando a 98 % de média harmônica (média), com 88 % (44 entidades) da clase 1 e 96 % (96 entidades) da clase 0.

Na discussão que se segue, abordaremos o problema em conjuntos de dados relativos a amostras de vinho verde tinto. O objetivo é modelar a qualidade do vinho com base em testes físico-químicos [Cortez *et al.*, 2009]. Esse conjunto de dados também está disponível no repositório *machine learning repository* UCI. `https://archive.ics.uci.edu/ml/datasets/wine+quality`, O dataset do Wine-Quality comtem 4898 entidades e 11 atributos (propriedades Fisico-químicos).

Mostraremos o funcionamento do modelo de ambíguos para o *dataset* wine-quality. Foram feitos todos os testes de acordo a Figura 3.8. Mostraremos aqui para as qualidades 3-4 e 8-9 com 183 e 180 entidades respectivamente.

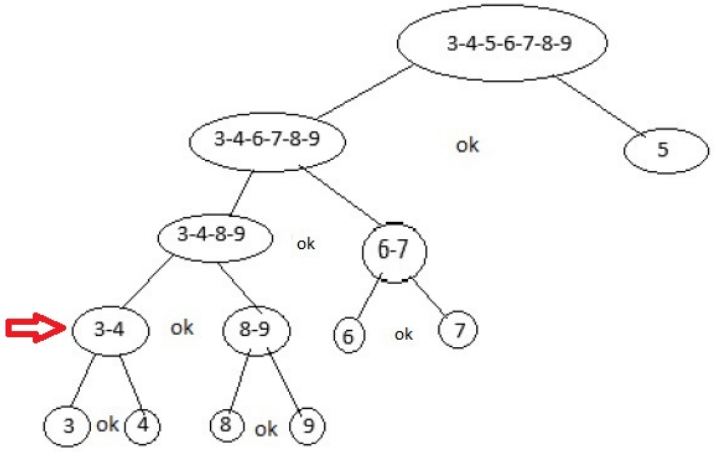


Figura 3.8. As qualidades 0, 1, 2 e 10 não tem atributos, logo, não constam na árvore. Foram feitos os testes para cada par de ramos da árvore. Por exemplo primeiramente foi feito para as qualidades 3-4-6-7-8-9 e 5, e assim sucessivamente.

A Figura abaixo (Figura 3.9) mostra as probabilidades de cada um das entidades. Observa-se que suas probabilidades melhoram aplicando-se a viabilidade algébrica em comparação a Regressão Logistica tradicional.

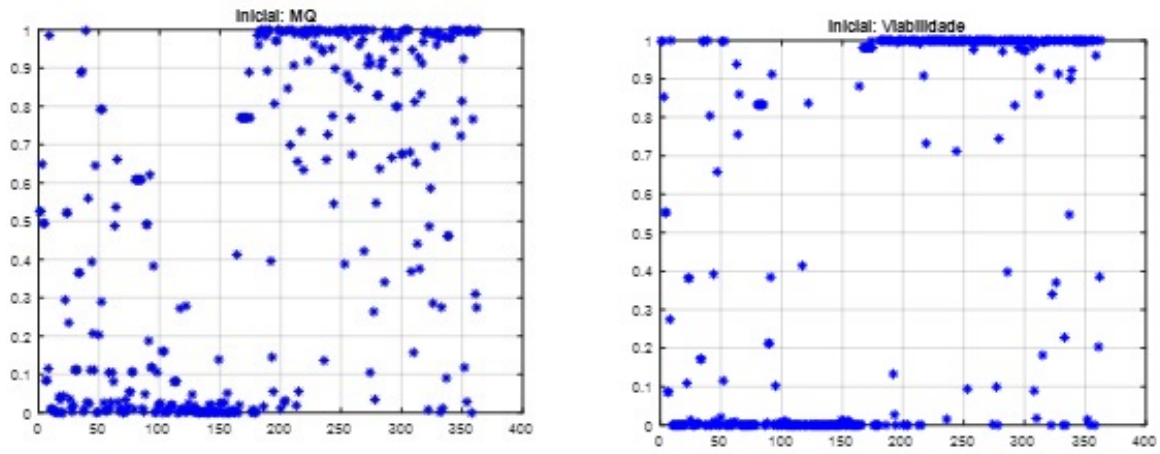


Figura 3.9. À esquerda temos o gráfico das probabilidades 0 ou 1, ou seja, pertence ou não ao grupo usando a Regressão Logística tradicional. No gráfico da direta foi aplicada a viabilidade algébrica para melhorar as probabilidade (desempenho). Observamos que temos 183 e 180 entidades 0 ou 1, respectivamente.

Aplicamos o primeiro modelo de ambíguos às entidades que foram aplicadas a viabilidade algébrica (seção 3.5.1) e separamos em dois grupos de ambíguos (44 entidades da classe zero e 31 entidades da classe 1) e não-ambíguos (136 entidades da classe zero e 148 entidades da classe um). Nos não-ambíguos ficaram as entidades que acertaram e no grupo dos ambíguos as que erraram, conforme a Figura 3.10.

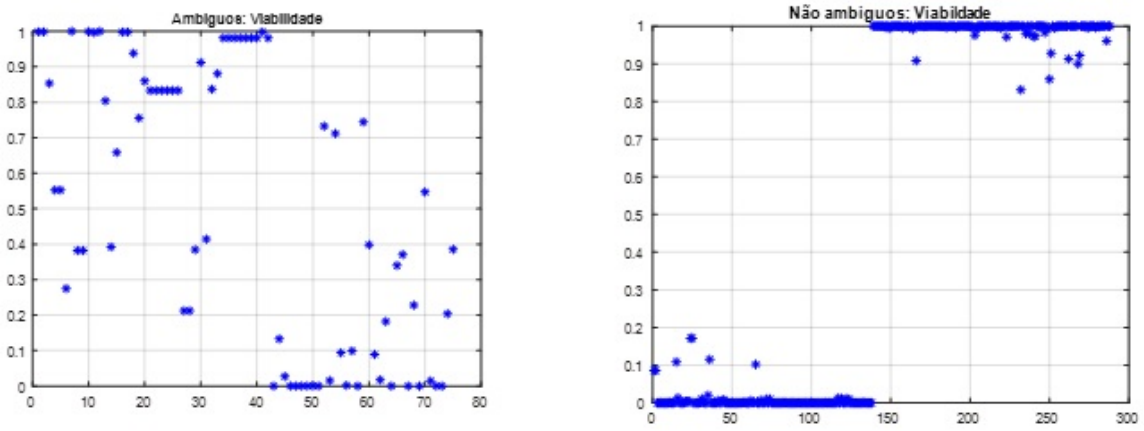


Figura 3.10. O gráfico mostra a separação dos dois grupos de ambíguos e não-ambíguos: foi feito um corte quando se acerta em um 80 %. O gráfico da esquerda mostra as probabilidades das entidades ambíguas e o da direta as probabilidades dos não-ambíguos.

Aplicamos a programação linear para a viabilidade algébrica (seção 3.5.1) as entidades ambiguas (44 entidades da classe zero e 31 entidades da classe 1) da Figura 3.10 e

obtemos Figura 3.11, logo aplicamos a Figura 3.11 um segundo modelo de ambiguidade ao grupo de entidades e separamos em dois grupos ambíguos (7 entidades da classe zero e 3 entidades da classe 1) e não-ambíguos (37 entidades da classe zero e 32 entidades da classe um). Nos não-ambíguos ficaram as entidades que acertaram e no grupo dos ambíguos as que erraram, conforme a Figura 3.12.

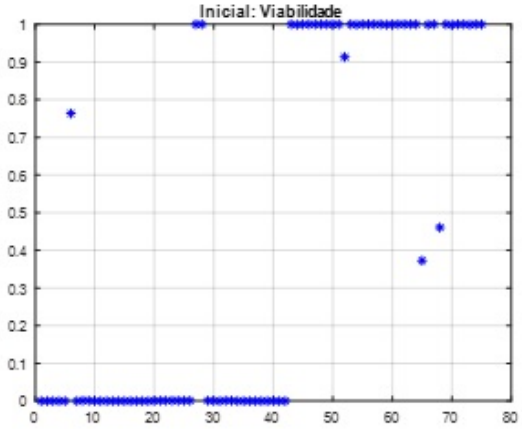


Figura 3.11. A visualização das probabilidades do *dataset* após a aplicação da programação linear.

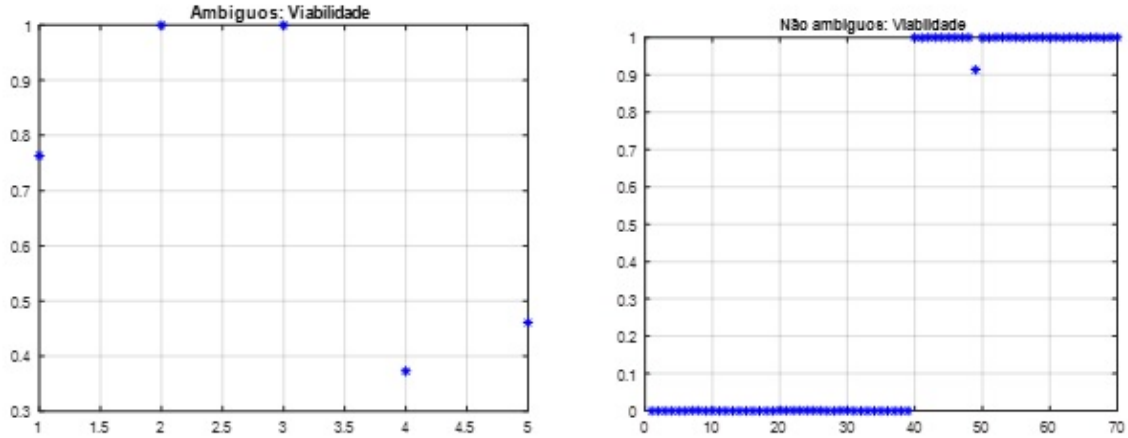


Figura 3.12. O gráfico mostra a separação dos dois grupos de ambíguos e não-ambíguos: foi feito um corte quando se acerta em um 80 %. O gráfico da esquerda mostra as probabilidades das entidades ambíguas e o da direta as probabilidades dos não-ambíguos. Percebe-se que o desempenho dos não-ambíguos é ótimo.

Portanto, aplicando dois modelos de ambíguos e eliminando entidades usando Programação Linear por meio da viabiliadade algébrica (seção 3.5.1) para obter uma separação com probabilidade proximo de um. Os resultados da proposta de ambiguidade é relevante para continuar trabalhando.

Capítulo 4

Conclusões

Embora as técnicas de inteligência artificial estejam em constante evolução, em seu núcleo permanecem alguns conceitos discutidos aqui, como *vector space model*, matrizes, padrões etc. Acreditamos que a Regressão Logística é uma abordagem prática e fácil de aplicar para aproveitar os mecanismos de recuperação de informação, presentes nas máquinas de busca como Yahoo e Google, em Bioinformática. Apresentamos ideas e métodos que foram aplicados na área de Bioinformática Estrutural, e que dão nova vida a esta metodologia. São gerais o suficiente para serem aplicadas em outros contextos e aplicações. Por exemplo, o método Ahdoc pode lidar com conjuntos de dados desbalanceados, não importando se a aplicação é em Bioinformática. Mesmo havendo interesse em utilizar algum outro recurso de Inteligência Artificial, os métodos apresentados são escaláveis e não demandam sistemas computacionais sofisticados, podendo ser utilizados como ferramenta de pré-processamento para seleção de atributos.

Neste trabalho, validamos o uso deste instrumento (a Regressão Logística), muito pouco utilizado na arte de recuperar informação em geral, na Bioinformática Estrutural. Escolhemos um problema simples (classificação de proteínas) para atingirmos os nossos objetivos. Tal fato foi atestado frente os resultados que obtivemos utilizando os mesmos conjuntos de dados que o melhor algoritmo de classificação publicado utilizou.

Um exemplo de aplicação quase que direta do que foi apresentado diz respeito às cavidades presentes nas proteínas e que são alvos de drogas aprovadas que as modulam. A partir das distâncias interatômicas de cada cavidade, pode-se associar um *VSM*, e, assim, mecanismos de busca latente podem ser desenvolvidos para reposicionar drogas. Isto, salvo o modelo utilizado para representar as entidades, é similar ao que foi usado com sucesso para obter novos alvos de drogas [Silvério-Machado *et al.*, 2015]. Outra fonte de problemas interessantes são os conjuntos de dados depositados no **NCBI** com experimentos com miRNA's. As técnicas aqui apresentadas têm aplicação quase que direta. Existem oportunidades para kits de diagnóstico e descoberta de marcadores de doenças.

As proteínas são artefatos da natureza que sustentam a vida como a conhecemos.

Suas estruturas tridimensionais lhes conferem a função que desempenham. Parece natural representá-las a partir das próprias coordenadas de seus átomos (ou partes deles). A matriz de distâncias interatômicas foi a forma que encontramos para fazermos isto com sucesso, dado que a representação usando o espectro da matriz mostrou-se robusta e consistente. No entanto, não nos escapa que o significado da decomposição, (matrizes de posto um) associada ao valor de cada um dos atributos, ainda não foi descoberto e carece de uma investigação profunda que no momento não foi possível realizar.

Concluímos, assim, que o novo **VSM** - *spectral pattern* com a presença de seis valores singulares (seção 3.4) obtidos da decomposição da matriz das distâncias interatômicas, O **VSM** mostrou-se eficaz, robusto e consistente, teve sucesso em diferentes cenários, portanto, é uma das contribuições importantes deste trabalho.

Os resultados da proposta de ambiguidade e a viabilidade algébrica são relevantes (seção 3.5.1 e 3.5.2). Após análise dos resultados, percebeu-se que algumas melhorias podem aprimorar e expandir a metodologia proposta. Entre elas está a procura de interpretação de atributos importantes na classificação de estruturas pela proposta de ambiguidades.

Referências Bibliográficas

Antczak M, Kasprzak M, Lukasiak P, Blazewicz J. *Structural alignment of protein descriptors - a combinatorial model.* BMC Bioinformatics. 2016 Sep 17;17:383.

Antony Unwin, Kim Kleinman, *The Iris Data Set: In Search of the Source of Virginica*, Significance, Volume 18, Issue 6, December 2021, Pages 26–29,

Abreu, Ana Paula (2019) *Prospeccão de Biomarcadores para Cancer de Mama Subtipo Luminal A em estágio inicial usando Álgebra Linear*. Mestrado em Bioinformática. Instituto de Ciências Biológicas - Universidade Federal de Minas Gerais.

Ahmad, W. M. A. W., Nawi, M. A. B. A., Aleng, N., Halim, N., Mamat, M., Hamzah, M., & Ali, Z. (2014). *Association of hypertension with risk factors using logistic regression*. Applied Mathematical Sciences, 8(52), 2563-2572.

Andriani, P., & Chamidah, N. (2019, August). *Modelling of Hypertension Risk Factors Using Logistic Regression to Prevent Hypertension in Indonesia*. In Journal of Physics: Conference Series 1306(1), 012027. IOP Publishing

Baeza-Yates, Ricardo; Ribeiro-Neto, Berthier. *Recuperação de Informação: conceitos e tecnologias das máquinas de busca*. (2nd. Ed.) Porto Alegre:Bookman, 2013.

Barella, Victor Hugo. *Técnicas para o problema de dados desbalanceados em classificação hierárquica.* 2015. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2015.

Berry, M. W.; Dumais, S.T. & O'Brien, G.W.(1995). *Using linear algebra for intelligent information retrieval*. SIAM reviem, 37(4): 573-595

Baldi P. and Brunak S. *Bioinformatics: the Machine Learning approach*. MIT Press, 2 edition, 2001.

Berman , H. M,; Westbrook, J.; Feng, Z.; Gilliland, G,; Bhat, T.; Weissig, H,; Shindyalov, I. N. & Bourne, P. E. (2000). *The protein data bank*. Nucleic acids research, 28(1): 235-242.

Brown SD1, Gerlt JA, Seffernick JL, Babbitt PC.(2006) *A gold standard set of mechanistically diverse enzyme superfamilies.* Genome Biology, 7(1):R8

Carvalho, D. R.; Dallagassa, M. R. .*Mineração de dados: aplicações, ferramentas, tipos de aprendizado e outros subtemas.* AtoZ: novas práticas em informação e conhecimento, v. 3, p. 82, 2015.

Chandra, Nagasuma; Anand, Praveen; Yeturu, Kalidas (dezembro de 2010). *Structural bioinformatics: Deriving biological insights from protein structures.* Interdisciplinary Sciences: Computational Life Sciences (em inglês). 2 (4): 347–366.

Coppin,B. *Inteligência artificial.* Rio de Janeiro: LTC, 2010;

Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, José Reis (2009), *Modeling wine preferences by data mining from physicochemical properties*, Decision Support Systems,Volume 47, Issue 4, Pages 547-553.

Couto, B. R. G. M.; Ladeira, A. P. ; Santos, M. A. (2007).*Application of latent semantic indexing (LSI) to evaluate the similarity of sets of sequences without multiples alignments character-by-character.* Genetics and Molecular Research , v. 6, p. 983-999.

Eldén, L.(2007) *Matrix methodos in Data Mining and Pattern Recognition (Fundamentals of Algoritms)* Society for industrial and Aplied Mathematics.

Felipe Júnior, José *Mineração de Dados para Detecção de Fraudes em Transações Eletrônicas* Dissertação (mestrado) — Universidade Federal de Minas Gerais. — Belo Horizonte, 2012

Fisher, R. A.(1936) *The use of multiple measurements in taxonomic problems* Annals of Eugenics, 7(2), 179–188.

Fort, G. and Lambert-Lacroix, S. (2005) *Classification using partial least squares with penalized logistic regression,* Bioinformatics, 21(7): 1104-1111.

Frank, E.; Hall, M.; Trigg, L.; Holmes, G., Witten, I. H. (2004). *Data mining in bioinformatics using weka. Bioinformatics,* 20(15):2479–2481.

Goldschmidt, R.; Passos, E. *Data Mining – um guia prático.* Rio de Janeiro: Elsevier, 2005.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X.. (2013) *Applied Logistic Regression (Vol. 398).* John Wiley & Sons.

Hunter Sarah, Jones Philip, Mitchell Alex, Apweiler Rolf, Attwood Teresa, Bateman Alex, Bernard Thomas, Binns David, Bork Peer, Burge Sarah, Castro Edouard, Coggill Penny, Corbett Matt, Das Ujjwal, Daugherty Louise, Duquenne Lauranne, Finn

Robert, Fraser Matthew, Gough Julian, Yong Siew-Yit. (2011). *InterPro in 2011: new developments in the family and domain prediction database* , Nucleic Acids Research, Volume 40, Issue 10, D306-D312.

Jain P1, Hirst JD. *Automatic structure classification of small proteins using random forest.* BMC Bioinformatics. 2010 Jul 1;11:364. doi: 10.1186, 1471-2105-11-364.

Kolodny R, Linial N. *Approximate protein structural alignment in polynomial time.* Proc Natl Acad Sci U S A. 2004 Aug 17;101(33):12201-12206.

Kloczkowski A., Jernigan R. L., Wu Z., Song G., Yang L., Kolinski A., Pokarowski P.(2009)*Distance matrix-based approach to protein structure prediction.* J Struct Funct Genomics. 2009 Mar;10(1):67-81.

Kolodny R, Linial N. *Approximate protein structural alignment in polynomial time.* Proc Natl Acad Sci U S A. 2004

Landwehr, N. Hall, M. & Frank, E (2005). *Logistic model trees.* Machine Learning, 59(1-2): 161-2015.

Leach A. R. (2001) *Molecular Modelling: Principles and Applications.* Prentice Hall; 2nd ed. edição (2001)

Leite, C. F. V. et al.*Milk-Way algorithm for ligand-based virtual screening: CDK2 case study. Trends in Developmental Biology*, v. 13, 2020.

Lorena A. C., De Carvalho, André C. P. L. F. (2007) *Uma introdução às support vector machines,* Revista de Informática Teórica Aplicada, vol. 14, no. 2, pp. 43-67.

Leandro S. Marcolino, Bráulio R. G. M. Couto, Marcos A. dos Santos. (2010) *Genome visualization in space* . In Procceedings of IWPACCBB, Springer Berlin Heidelberg pp. 225-232.

Ma, C.; Zhang, H. H., Wang, X. (2014). *Machine learning for big data analytics in plants.* Trends in plant science, 19(12):798–808.

Morais R.F.; Ortega, J. M. ; Azevedo, V. A.C. ; Dos Santos, M. A., et al., (2020) *Analysis of the microarray gene expression for breast cancer progression after the application modified logistic regression* GENE, v. 726, p. 144-168

do Nascimento Júnior, Francisco. *Classificação de Proteínas usando Máquinas de Aprendizagem e Descoberta de Padrões.* Dissertações de Mestrado - Ciência da Computação. Universidade Federal de Pernambuco 2014

Pires, D. E. V. ; Minardi, R. C. M. ; Santos, M. ; Da Silveira, C. H. ; Santoro, M. M. ; Meira Junior, W. (2011) *Cutoff Scanning Matrix (CSM): funciton prediction and fold recognition by protein inter-residue distance patterns* . BMC Genomics, v. 12 Suppl 4, p. S12.

Perfetti R, Ricci E(2006). *Analog neural network for support vector machine learning.* IEEE Trans Neural Netw. 17(4):1085-1091.

Richardson, Alice. (2011). *Logistic Regression: A Self-Learning Text, Third Edition by David G. Kleinbaum, Mitchel Klein.* International Statistical Review. 79. 296-296.

Silvério-Machado R., Couto B. R., Dos Santos M. A. (2015).*Retrieval of Enterobacteriaceae drug targets using singular value decomposition.* Bioinformatics 31, 1267?1273. 10.1093/bioinformatics/btu792 - DOI - PubMed

Santos, Alysson dos (2017) *Early Breast Cancer Detection Using Logistic Regression Models.* Mestrado em Bioinformáica. Instituto de Ciências Biológicas - Universidade Federal de Minas Gerais.

Santos, Anderson R; Santos, Marcos A ; Baumbach, Jan ; McCulloch, John A ; Oliveira, Guilherme C ; Silva, Artur ; Miyoshi, Anderson ; Azevedo, Vasco . *A singular value decomposition approach for improved taxonomic classification of biological sequences.* BMC GENOMICS, v. 12, p. S11, 2011.

dos Santos, Eduardo Campos (2012) *Mineração de dados usando álgebra linear para a predição de alvos drogáveis.* Doutorando em Bioinformática. Instituto de Ciências Biológicas - Universidade Federal de Minas Gerais.

Silva, AC; Nunes, BRP; Xavier, J. *De onde vêm as proteínas?* BIOINFO. ISSN: 2764-8273. Vol. 3. p.16 (2023). doi: 10.51780/bioinfo-03-16

Vidyavathi, B. M. (2019). *A new approach to feature selection for data mining.* Computational Intelligence Research, 7(3).

Wang, J. T.; Zaki, M. J.; Toivonen, H. T., Shasha, D. (2005). *Introduction to data mining in bioinformatics.* Em Data Mining in Bioinformatics, pp. 3–8. Springer.

Yi Liu, Yuan F. Zheng ,FS-SFS: *A novel feature selection method for support vector machines.* Journal of Pattern Recognition,39(7) pp 1333-1345,2006.

.

Apêndice A

Artigo 1: Algoritmos para Classificação Estrutural de Proteínas

1. Autores: Lima, Paccori Lucio; dos Santos, Marcos Augusto
2. **DOI**: 10.55905/revconv.16n.11-255
3. INSS: 1988-7833; Qualis CAPES: A4
4. Volume:16; No:11
5. Ano: 2023
6. URL: https://doi.org/10.55905/revconv.16n.11-006
7. Publicado em: 30/11/2023.

Algoritmos para Classificação Estrutural de Proteínas

Lúcio Paccori Lima[1] and Dr. Marcos Augusto dos Santos[2]

[1] *UFMG, Belo Horizonte/MG, Brazil*
[2] *UFMG, DCC, Belo Horizonte /MG, Brasil*

Abstract

Neste trabalho, estruturas de proteínas são mapeadas em um *vector space model* para sua manipulação em ambientes que remetem às modernas maquinas de busca (Yahoo, Google etc). Instrumentos da álgebra linear junto com a regressão logística com modificações esteiam um novo algoritmo, aqui denominado **Adhoc**, para a busca de assinaturas de estruturas tridimensionais. Este método transcende a aplicação aqui discutida; pode ser utilizado de forma geral em outros problemas de recuperação de informação. Para a validação, testamos as ideias para classificar proteínas em seus respectivos grupos e famílias, a partir da estrutura já determinada. Os resultados apresentados são animadores e outros experimentos estão sendo providenciados.

Keywords: Bioinformática, Regressão Logística, SVM

1 INTRODUÇÃO

Este trabalho contempla o uso de algoritmos de classificação para mineração de dados de estruturas proteicas e suas assinaturas. Uma assinatura, na acepção aqui considerada, é uma marca que comprova um conjunto único de atributos para identificar uma entidade (proteína ou parte da proteína) como membro de um conjunto previamente escolhido.

Existem ótimos métodos para classificar e recuperar proteínas em suas respectivas super famílias, famílias e subgrupos diversos. A partir da descrição das proteínas por um *vector space model* (VSM), onde cada entidade é representada por um conjunto de tamanho fixo de atributos numéricos, são aplicados algoritmos tradicionais, a saber, redes neurais, pesquisa em árvores, *vector support machine* e outros [Vapkin et al., 1999], [Frank et al,. 2004], [Wang et al., 2005] e [Ma et al., 2014]. São capazes de predizer com eficiência a que grupo uma dada proteína pertence. Ao que nos consta, em [Pires DE, de Melo-Minardi

Contato: Lúcio, paccori@uol.com.br

RC, dos Santos MA, Santoro MM, et al., 2012] tem-se os melhores resultados, onde são apresentados resultados que permitem construir oráculos perfeitos (considerados por nós quando a média harmônica obtida de validações cruzadas são superiores à 0,95).

Entretanto, os artefatos de mineração citados no paragrafo acima não são, sabidamente, adequados para escolher quais os atributos (e seus valores) que contribuem para um determinado indivíduo estar em um certo grupo. Este processo de escolha de atributos (*feature selection*) configura uma área em aberto na ciência da computação.

Um método interessante que tem como resultado colateral uma avaliação de tais atributos, é a regressão logística [Kleinbaum et al., 2002], [Mesquita, 2014]; uma técnica bem conhecida e muito utilizada na área médica [Fort, et al., 2005], [Hosmer et al., 2000]. Mas a sua utilização genérica como instrumento de classificação traz uma série de inconvenientes; talvez seja esta razão por ele não estar precisamente contemplado na taxonomia do processo de mineração de dados. Nem sempre nos problemas reais tem-se um domínio com equilíbrio adequado entre a quantidade (e variedade) dos membros do grupo que se deseja classificar e a quantidade (e variedade) da população total. Nos próximos trabalhos, estaremos usando superfamílias com 27000 entidades ao lado de outras com apenas algumas dezenas. Estas e outras inconveniências são resolvidas nos algoritmos que estamos propondo e/ou adaptando.

A justificativa e muito da motivação inicial para este trabalho, era encarar um fato que ocorre na natureza. É sabido que às proteínas com sequências similares, correspondem estruturas tridimensionais também similares [Leach, 2001]. Entretanto, existem casos em que estruturas similares não guardam similaridade quanto às sequências primárias. Esperamos encontrar assinaturas que sejam invariantes quanto a este aspecto.

No desenvolvimento do trabalho, acabamos por ter a necessidade de buscar por oráculos perfeitos. Este é um dos principais aspectos tratados neste trabalho.

2 METODOLOGIA

2.1 Conjunto de dados

Os testes foram realizadas no conjunto de superfamílias de enzimas, tidas como um padrão ouro, que utilizam mecanismos distintos para executar suas funções [Brown et al., 2006]. Neste conjunto são consideradas seis superfamílias (amdohydrolase, crotonase, haloacid dehalogenase, isoprenoind synthase type I e vicinal oxygen chelate), compreendendo 47 famílias distribuídas em 896 diferentes cadeias.

2.2 Decomposição por valor singular

A decomposição por valores singulares ou *singular value decomposition* (**svd**) [Eldén, 2006, 2007; Berry, 1995] é uma técnica da álgebra linear utilizada para reduzir a dimensionabilidade das entidades sem comprometer a sua essência. Com o **svd** é possível, ao invés de trabalhar com uma matriz com o posto aproximado, simplesmente usar a combinação linear dos padrões presentes na matriz que aproximam o posto. Em geral, para o propósito de recuperar informação em problemas reais, não são necessários muitos padrões, implicando em uma representação mais econômica.

Mais formalmente, a decomposição por valores singulares é uma fatoração de uma matriz qualquer em três outras matrizes com propriedades importantes. Possui várias

aplicações, tanto diretas, nas quais se aplicam os resultados extraídos de suas matrizes fatores, quanto como um passo em muitos algoritmos.

Definição 1. *Dado $A \in \mathbb{R}^{m \times n}$, não necessariamente de posto completo, a decomposição por valores singulares de A é uma fatoração tal que:*

- $A = USV^T$;
- $U \in \mathbb{R}^{m \times m}$, *são os autovetores de* AA^T *e é ortogonal;*
- $V \in \mathbb{R}^{n \times n}$, *são os autovetores de* A^TA *e é ortogonal;*
- $S \in \mathbb{R}^{m \times n}$, *é diagonal se* $m = n$, *caso contrário adiciona-se* $m - n$ *linhas de zeros em* S *e é formado por a raiz quadrada dos autovalores de* AA^T.

$$S = diag(\sigma_1, \sigma_2, .., \sigma_n)$$

$$\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_n \geq 0.$$

Os valores $(\sigma_1, \sigma_2, ...\sigma_n)$ são chamados de valores singulares de A. As colunas de U, $(u_1, u_2, ...u_m)$ são chamadas de vetores singulares à esquerda de A e as colunas da matriz V, $(v_1, v_2, ..., v_n)$ são os vetores singulares à direita de A. As colunas de U podem ser interpretadas como padrões das entidade em A com um peso correspondente σ_i; já V são os padrões das linhas, que têm o mesmo peso associado.

Estamos utilizando o *svd* neste trabalho com três propósitos. Primeiro é uma ferramenta que possibilita a visualização de conjuntos de dados no espaço tridimensional [Marcolino LC, Couto BRGM, dos Santos MA, 2010], o que nos permite ter uma avaliação inicial das eventuais dificuldades no processo de classificação. Outra aplicação vem do fato de que nas melhorias que estamos propondo nos algoritmos, temos de fazer escolhas e recuperar um subconjunto de entidades que são mais próximas a uma dada consulta; usando as técnicas de máquina de busca, conseguimos melhores resultados. Finalmente, usamos diretamente os valores singulares como **VSM**, como será usado futuramente.

2.2.1 Representação de entidades no contexto de estruturas de proteínas

Neste trabalho estamos utilizando a representação de entidades usando *o cutoff scanning* que foi proposto por [Pires DE, de Melo-Minardi RC, dos Santos MA, Santoro MM, et al., (2012)]; deu origem ao melhor, ou um dos melhores classificadores para esta classe de problemas.

Representação de entidades usando o *cutoff scanning* .

Neste **VSM**, a proteína é descrita a partir do número de átomos que existe em subintervalos de 0,2 ångströms até uma distância máxima de 30Å. Foi usado por [Pires DE, de Melo-Minardi RC, dos Santos MA, Santoro MM, et al., 2011]; onde são mostrados resultados excepcionais para esta classe de problemas de classificação.

2.3 Regressão Logística

A regressão logística consiste em encontrar valores de $\alpha = (\alpha_1, \alpha_2, ..., \alpha_n)^T$ para ajustar a equação (1) para cada uma das entidades j. O valor da função, conhecida como *logit*, $P_j(x)$, com $P_j(x) \in [0, 1]$ informa a probabilidade da uma dada entidade ser classificada como pertencente a um subconjunto específico.

$$P_j(x) = \frac{e^{\sum_{i=1}^{m} \alpha_i x_i}}{1 + e^{\sum_{i=1}^{m} \alpha_i x_i}} \tag{1}$$

Observa-se que quando $e^{\sum_{i=1}^{m} \alpha_i x_i}$ tende para zero, $P_j(x)$ também tende para zero. Por outro lado, se $e^{\sum_{i=1}^{m} \alpha_i x_i}$ tende para infinito, $P_j(x)$ aproxima-se da unidade.

Para proceder a determinação dos valores de α, faz-se uma transformação linear que remeterá o problema à solução de um problema de álgebra linear. Seja a chance $C_j(x)$ é definido por:

$$C_j(x) = \frac{P_j(x)}{1 - P_j(x)}. \tag{2}$$

A expressão da equação (2) usando (1), temos

$$C_j(x) = e^{\sum_{i=1}^{m} \alpha_i x_i} \tag{3}$$

Tomando-se o logaritmo em ambos os lados de (3), obtemos um sistema de equações lineares para determinar α:

$$b_j = \sum_{i=1}^{n} \alpha_i x_i \tag{4}$$

onde $b_j = \log(C_j)$, para j=1,2,...m.

Associamos ao conjunto de dados uma matriz $A = \{a_{i,j}\} \in \mathbb{R}^{m \times n}$. As linhas representam as entidades e às colunas estão associados os atributos. Assim, o valor de cada $a_{i,j}$ é o valor do atributo j na proteína i. Observar que isto é diferente da forma que tratamos anteriormente, quando discutimos a decomposição por valores singulares (esta matriz é a transposta dessa anterior).

Seja $b = (b_1, b_2, ..., b_m)^T$; o sistema de equações lineares (4) pode ser representado por:

$$A\alpha = b \tag{5}$$

Em (5), cumpre observar que quando o número de equações é inferior ao número de incógnitas, a solução do sistema, fornecida pelo método da minimização da somatória dos quadrados dos resíduos, é indeterminado (infinitas de soluções). Em outras palavras, a abordagem clássica em álgebra linear é minimizar $||A\alpha - b||^2$, o que requer o posto completo de $A^T A$ - uma propriedade não é esperada das matrizes A nos cenários em que estamos trabalhando. Geralmente, para contornar essa dificuldade descarta-se um grupo das variáveis, mantendo apenas um subconjunto das variáveis originais. Este procedimento é chamado de *feature selection* em mineração de dados - uma área de pesquisa em aberto. Nós usamos um termo estabilizador no modelo de regressão logística, encontrado nos trabalhos de [Morais, Rodrigues, F., et al., 2020], que permite a atribuição de valores para os parâmetros α que minimizam a soma dos quadrados dos resíduos $(A\alpha - b)$,

adicionado aos quadrados de α. Assim, para encontrar uma solução a (5), resolvemos um problema otimização quadrática irrestrito dado por

$$Minimize f(\alpha) = ||\alpha||^2 + ||A\alpha - b||^2 \tag{6}$$

Como $f(\alpha)$ é função convexa, o argumento α^* que minimiza (6) é dado pela derivação de $f(\alpha)$ em α e igualando resultado a zero, que resulta em um sistema de equações lineares

$$(I + A^T A)\alpha = A^T b, \tag{7}$$

onde I é a matriz identidade da ordem n.

A solução ótima para α de (6) é obtida pela solução de (7) e é única. Então, dada uma proteína (*query*) $q = (q_1, q_2, ..., q_n)$ com n atributos, a probabilidade de q pertencer a uma classe associada ao sistema relacionado é dado por:

$$P(q) = \frac{e^{q\alpha}}{1 + e^{q\alpha}}. \tag{8}$$

Saliente-se que (7) é para o caso em que o número de incógnitas é maior que o número de equações, que geralmente ocorre com o método **Adhoc**, visto a seguir.

2.3.1 Método Adhoc

Adhoc é uma expressão latina cuja tradução literal é "para isto" ou "para esta finalidade". É usada para designar algo ou alguma coisa que foi formada ou usada para um propósito ou necessidade particular e imediata, sem planejamento prévio.

A metodologia descrita a seguir, aqui intitulada **Adhoc**, usa modelos construídos exclusivamente em resposta a uma única demanda por classificação. Dada uma consulta q, um modelo **Adhoc**, específico para esta consulta, é construído a partir da escolha de k_0 entidades mais próximas à q tais que $P_j(x) = 0$, ao lado de k_1 entidades escolhidas dentre aquelas mais próximas à q tal que $P_j(x) = 1$.

Os valores de k_0 e k_1 são determinados experimentalmente. Nos nossos ensaios constatamos que estes valores são baixos; na validação cruzada verificamos que o desempenho da classificação é superior quando atribuímos valores na ordem de algumas unidades a estes dois parâmetros. Isto tem como consequência a construção de matrizes de atributos $M \in \mathbb{R}^{m \times n}$ nas quais $m < n$ o que impede o uso da regressão logística tradicional. Resolvemos esta limitação usando a regressão logística modificada segundo explicado na seção anterior.

Outra dificuldade está na escolha das entidades $k_{0/1}$ mais próximas, o que poderia impactar o tempo de resposta nos problemas de grande porte. Isto é resolvido organizando as entidades em árvores de pesquisa segundo os recursos usados em máquinas de busca (**svd**, clusterização etc). A matriz de atributos original A é particionada em duas outras, A_0 e A_1, segundo $P_j(x) = 1/0$; com cada uma delas organizada enquanto elementos de uma máquina de busca. Assim, a recuperação das entidades mais próximas a uma consulta q fica extremamente eficiente e não impacta o tempo de processamento de forma perceptível.

Concluindo, o método **Adhoc** é um recurso para a regressão logística que permite a sua aplicação em diferentes cenários. A tabela 1 apresenta o algoritmo utilizado para o método **Adhoc.**

Algoritmo 1: Método Adhoc

Entrada: q : consulta q, A1, A0
Saída: P: Probabilidade de q pertencer ao grupo 1
A1: Conjunto de entidades que pertencem a categoria 1;
A0: Conjunto de entidades que não pertencem a categoria 0;
M1 ← Conjunto dos k_1 elementos mais próximos de q em A1;
M0 ← Conjunto dos k_0 elementos mais próximos de q em A0;
M ← [M1; M0];
$i_1 \leftarrow 1$, $\forall$ elemento de M1;
$i_0 \leftarrow 0$, $\forall$ elemento de M0;
i ← $[i_1; i_0]$;
P ← Aplique a regressão logística modificada para M e i

Tabela 1: Regressão logística para ambientes desbalanceados

3 RESULTADOS E DISCUSSÃO

Nesta seção apresentamos a experiência computacional usando o conjuntos de dados que congrega entidades sob aspectos funcionais (*golden standard*). Para todos os ensaios usamos cristais depositados no **pdb** para extrair as coordenadas dos átomos para os classificadores.

Os átomos considerados neste trabalho para a construção do **vsm** foram os da cadeia principal (ou parte dela), constituída, seja pelos C_α, seja pelos átomos (C_α, C, N). Mas também, em alguns casos, utilizamos todos os átomos da proteína. Entretanto, não houve grandes alterações nos resultados (não exibidos no texto).

Planejamos os experimentos para inicialmente responder uma primeira pergunta; aquela que tange a adequação dos nossos algoritmos à classificação de estruturas de proteínas. Os métodos que usamos são interessantes para avaliar atributos e, em outras aplicações, como por exemplo, aquelas que utilizam *microarrays*, o efeito colateral que no caso é identificar marcadores biológicos de uma doença, tem resultado em contribuições interessantes [Abreu A.P. (2019)], [Santos A. dos, (2017)], [Morais, Rodrigues, F., et al., 2020]. Assim, passamos a testar os métodos descritos na seção anterior para seguir à busca de assinaturas estruturais para atender ao nosso objetivo.

No que se segue, apresentamos os resultados da regressão logística em um conjunto, um que se apresenta muito bem balanceado (*golden standard*). Com os resultados que alcançamos, buscamos identificar as assinaturas usando como **vsm** o *cutoff scanning* , e obtivemos resultados muito próximos ao que consideramos como oráculo perfeito (média harmônica na validação cruzada acima de 0,95).

3.1 Validação da regressão logística como instrumento para classificação estrutural

Neste item estão os resultados com modelos construídos com *cutoff scanning* [Pires DE, de Melo-Minardi RC, dos Santos MA, Santoro MM, et al., 2011], no conjunto *golden standard* cuja característica que mais nos chama a atenção é o bom balanceamento entre as entidades em todos os conjuntos. Na referência citada, os melhores resultados foram com a matriz de distâncias entre os carbonos alfa C_α da proteína em que o posto foi reduzido usando **svd**. Vários métodos tradicionais de mineração de dados foram utilizados.

Nossos melhores resultados foram alcançados usando a regressão tradicional com as

coordenadas dos átomos da cadeia principal - ver tabela 2. Usando a regressão logística com a estrategia **Adhoc**, o desempenho foi levemente inferior - ver tabela 3. Resultados de outros experimentos usando C_α encontram-se nos anexos (tabela 5). Possivelmente, embora excepcionais, estes resultados possam ser melhorados usando como em Pires (ref. citada), a decomposição por valores singulares. Neste primeiro momento, como nos interessa avaliar atributos para entender uma possível assinatura, evitamos este recurso pois estaríamos perdendo um certo mapeamento dos atributos originais, dado que o **svd** trabalha em espaços projetados. Entretanto, tão logo seja possível, faremos isto.

Cumpre observar que tivemos dificuldade para comparar diretamente os nossos resultados com aqueles obtidos em Pires (referência citada). O desempenho dos métodos é aferido a partir de 20 *folders* aleatoriamente construídos. Não há na referência citada uma sugestão de *data set* para testes. Além disto, adotamos, por considerar mais adequado ao nosso problema, critérios de avaliação baseados na média harmônica e curva **ROC** . Entretanto, acreditamos que os nossos resultados seguem semelhantes àqueles que nos servem de referência.

Resultados da regressão logística no conjunto *golden standard*

Superfamilia	Sensibilidade. (Média)	Especificidade. (Média)	Media Harmônica (Média)	Curva ROC (Média)
amidhydrolase	0,9756	0,9581	0,9664	0,9597
crotonase	1,0000	0,9890	0,9944	0,9984
enolase	0,9779	0,9931	0,9850	1,0000
haloacid dehalogenase	0,9571	0,9573	0,9551	0,9587
isoprenoid synthase type I	1,0000	0,9994	0,9997	1,0000
vicinal oxygen chelate	1,0000	0,9968	0,9984	1,0000
Média	0,9851	0,9823	0,9831	0,9861

Tabela 2: Com a regressão tradicional alcançamos a média harmônica (média) 98% , sensibilidade média de 98% e uma especificidade média de 98%. As coordenadas dos átomos para a construção da matriz de distâncias foram os da cadeia principal

Resultados do método Adhoc no conjunto *golden standard*

Superfamilia	Sensibilidade. (Média)	Especificidade. (Média)	Media Harmônica (Média)	Curva ROC (Média)
amidhydrolase	0,9699	0,9419	0,9553	0,9794
crotonase	0,9818	0,9643	0,9727	0,9836
enolase	0,9351	0,9194	0,9266	0,9584
haloacid dehalogenase	0,9737	0,9559	0,9643	0,9831
isoprenoid synthase type I	1,0000	0,9742	0,9866	0,9926
vicinal oxygen chelate	0,9533	0,9657	0,9587	0,9734
Média	0,9690	0,9535	0,9607	0.9784

Tabela 3: Regressão logística em *data sets* bem balanceados: o método **Adhoc** alcançou 96% de média harmônica (média), sensibilidade média de 96% e uma especificidade média de 95%, usando uma matriz de distâncias construída a partir das coordenadas dos átomos carbono alfa (C_α) da proteína.

3.2 ***Busca de assinaturas estruturais usando o VSM*** *cutoff scanning* ***com o método Adhoc***

Asim, dado que os resultados alcançados foram animadores, passamos à próxima etapa que é a definição da assinatura, escolhendo e procurando interpretar os atributos mais importantes na classificação. A regressão logística indica isto a partir dos valores de α; usam-se os valores de α_i mais positivos e os mais negativos. Por exemplo, no gráfico mostrado na figura (1), escolheríamos alguns atributos que estivessem topologicamente mais distantes do eixo das abcissas. Aqueles atributos cujos valores de α_i são próximos de zero, não têm poder discriminatório.

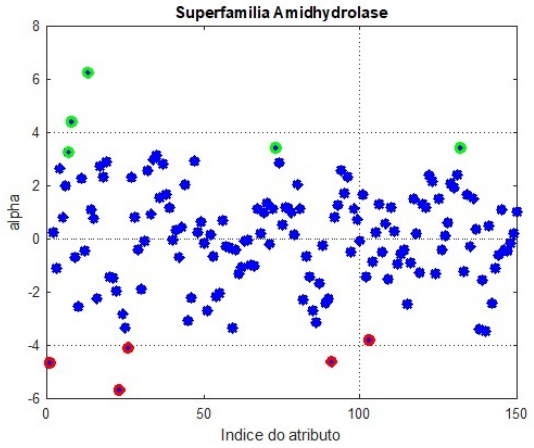


Fig. 1: Valores dos pesos (α_i) obtidos em um modelo de regressão logística, para a superfamília Amidhydrolase do conjunto de dados *gold standard* . Foram usados os átomos do carbono alfa (C_α).

Por exemplo, para o oráculo da superfamília Amidhydrolase, escolheríamos os atributos que são mostrados na tabela (4). O próximo passo consiste em validar esta escolha de atributos. Caso os oráculos construídos somente com estes atributos sejam eficientes, tem-se uma prova de conceito, como utilizamos em outros problemas da área, quando buscávamos marcadores biológicos. [Leite, et al. 2020]e [Morais R. F. et al. 2020].

Alfa	Atributo	Intervalos (ångströms)	Alfa	Atributo	Intervalos (ångströms)
Positivos	7	$(2,2\ ,\ 2,4)$	Negativos	6	$(2,0\ ,\ 2,2)$
	8	$(2,4\ ,\ 2,6)$		3	$(1,4\ ,\ 1,6)$
	4	$(1,6\ ,\ 1,8)$		2	$(1,2\ ,\ 1,4)$
	1	$(1,0\ ,\ 1,2)$		18	$(4,4\ ,\ 4,6)$
	5	$(1,8\ ,\ 2,0)$		17	$(4,2\ ,\ 4,6)$

Tabela 4: Escolha de atributos baseados nos parâmetros α_i da superfamília Amidhydrolase. Foram utilizados os átomos da cadeia inteira

4 CONCLUSSÃO

O objetivo principal deste trabalho é a busca por assinaturas de estruturas tridimensionais de proteínas que passa, necessariamente, por processos que atestem a sua pertinência.

Dentre os caminhos para alcançar este objetivo, escolhemos aqueles que remetem à mineração de dados. Imaginávamos um caminho mais fácil.

Um trabalho [Morais, R. F. et al. 2020] do grupo de pesquisa ao qual pertenço a publicado há alguns anos atrás e obteve um resultado importante. A partir de um modelo para descrever as estruturas (*cutoff scanning*) e usando as técnica usuais de mineração de dados, Pires [Pires, et al., 2012]) conseguiu classificadores que se situam entre os de melhor desempenho para esta classe de problemas. Entretanto, em função dos métodos utilizados, nada foi especulado quanto ao papel dos atributos e sua importância em cada classe em que ele foi aplicado. Imaginávamos que poderíamos utilizar nossos métodos e assim descortinar o papel dos atributos.

Os nossos métodos mostraram-se eficientes para a esta classe de problemas. Como efeito colateral, consideramos que com o método **Adhoc** conseguimos situar a regressão logística como uma alternativa eficaz aos métodos usuais de mineração de dados. Estendemos a possibilidade de usar a regressão logística em cenários desbalanceados em geral. Os resultados transcendem esta aplicação deste nosso trabalho em bioinformática e tem aplicação geral em na ciência da computação. Resolve a grande falha da regressão logística na escalabilidade da sua aplicação nos cenários reais.

Além da eficiência, cumpre chamar a atenção dos recursos computacionais que o método demanda: nada além de um microcomputador. Não são necessários servidoras e nem instalações sofisticadas para usar a metodologia.

Um efeito colateral deste nosso trabalho tem sido a melhoria dos algoritmos de classificação; não é exatamente nosso objetivo primário mas, precisamos desses métodos para mostrar a efetividade das nossas escolhas. Nossos métodos se traduzem também como uma contribuição para área de mineração de dados em ciência da computação. Gostaria de salientar que em nenhum momento este foi o nosso foco. O Laboratório de Bioinformática e Sistemas (**LBS**) ao qual estamos ligados, tem como um dos projetos o reposicionamento de fármacos. Nesta trilha, sempre aparece a necessidade de criar um domínio com as cavidades nas quais os fármacos modulam as proteínas. A dificuldade em mapear estas entidades sempre esteve presente e este nosso trabalho pode vir a ser uma resposta a este problema.

Referências

[1] Abreu, A. P. (2019) *Prospecção de Biomarcadores para Câncer de Mama Subtipo Luminal A em estágio inicial usando Álgebra Linear*. Mestrado em Bioinformáica. Instituto de Ciências Biológicas - Universidade Federal de Minas Gerais.

[2] Berry, M. W.; Dumais, S.T. & O'Brien, G.W.(1995). *Using linear algebra for intelligent information retrieval*. SIAM reviem, 37(4): 573-595.

[3] Brown SD1, Gerlt JA, Seffernick JL, Babbitt PC.(2006) *A gold standard set of mechanistically diverse enzyme superfamilies*. Genome Biology, 7(1):R8

[4] Fort, G. and Lambert-Lacroix, S. (2005) *Classification using partial least squares with penalized logistic regression,* Bioinformatics, 21(7): 1104-1111.

[5] Hosmer D. W.; Lemeshow , S. (2000) *Applied Logistic Regression* . H 2ad ed. New York: John Wiley & Sons.

[6] Jain P1, Hirst JD. *Automatic structure classification of small proteins using random forest.* BMC Bioinformatics. 2010 Jul 1;11:364. doi: 10.1186, 1471-2105-11-364.

[7] Kleinbaum, D. G.; Klein, M. (2002) *Logistic Regression, a Self-Learning Text,* 3ª edição, Londres-ENG, Springer, 701p.

[8] Kloczkowski A., Jernigan R. L., Wu Z., Song G., Yang L., Kolinski A., Pokarowski P.(2009)*Distance matrix-based approach to protein structure prediction.* J Struct Funct Genomics. 2009 Mar;10(1):67-81.

[9] Landwehr, N. Hall, M. & Frank, E (2005). *Logistic model trees.* Machine Learning, 59(1-2): 161-2015.

[10] Leach A. R. (2001) *Molecular Modelling: Principles and Applications.* Prentice Hall; 2nd ed. edição (2001)

[11] Leite, C. F. V. et al.*Milk-Way algorithm for ligand-based virtual screening: CDK2 case study. Trends in Developmental Biology*, v. 13, 2020.

[12] Lorena A. C., De Carvalho, André C. P. L. F. (2007) *Uma introdução às support vector machines,* Revista de Informática Teórica Aplicada, vol. 14, no. 2, pp. 43?67.

[13] Ma, C.; Zhang, H. H., Wang, X. (2014). *Machine learning for big data analytics in plants.* Trends in plant science, 19(12):798–808.

[14] Leandro S. Marcolino, Bráulio R. G. M. Couto, Marcos A. dos Santos. (2010) *Genome visualization in space* . In Procceedings of IWPACCBB, Springer Berlin Heidelberg pp. 225-232.

[15] Mesquita, P. S. B. (2014) *Um Modelo de regressão logística para avaliação dos programas de Pós-graduação no Brasil* . Dissertação (Mestrado) Universidade Estadual do Norte Fluminense, Campos dos Goytacazes.

[16] Morais R.F.; Ortega, J. M. ; Azevedo, V. A.C. ; Dos Santos, M. A., et al., (2020) *Analysis of the microarray gene expression for breast cancer progression after the application modified logistic regression* GENE, v. 726, p. 144168

[17] Pires, D. E. V. ; Minardi, R. C. M. ; Santos, M. ; Da Silveira, C. H. ; Santoro, M. M. ; Meira Junior, W. (2011) *Cutoff Scanning Matrix (CSM): funciton prediction and fold recognition by protein inter-residue distance patterns* . BMC Genomics, v. 12 Suppl 4, p. S12.

[18] Santos, Anderson R; Santos, Marcos A ; Baumbach, Jan ; McCulloch, John A ; Oliveira, Guilherme C ; Silva, Artur ; Miyoshi, Anderson ; Azevedo, Vasco. (2011).*A singular value decomposition approach for improved taxonomic classification of biological sequences.* BMC Genomics, v. 12, p. S11.

[19] Santos, Alysson dos (2017) *Early Breast Cancer Detection Using Logistic Regression Models.* Mestrado em Bioinformáica. Instituto de Ciências Biológicas - Universidade Federal de Minas Gerais.

[20] Wang, J. T.; Zaki, M. J.; Toivonen, H. T., Shasha, D. (2005). *Introduction to data mining in bioinformatics.* Em Data Mining in Bioinformatics, pp. 3–8. Springer.

5 Anexo

5.1 Regressão Logistica usando o vetor space model com o cutoff scanning

Resultados do regressão logística no conjunto *golden standard*

Superfamilia	Sensibilidade. (Média)	Especificidade. (Média)	Media Harmônica (Média)	Curva ROC (Média)
amidhydrolase	0,8881	0,8727	0,8778	0,9266
crotonase	1,0000	0,9455	0,9477	1,0000
enolase	0,9094	0,9402	0,9196	0,9500
haloacid dehalogenase	0,9125	0,9017	0,8969	0,8633
isoprenoid synthase type I	1,0000	0,9908	0,9453	1,0000
vicinal oxygen chelate	1,0000	0,9687	0,9836	0,9919

Tabela 5: Com a regressão tradicional alcançamos uma média harmônica (média) 92,85%, sensibilidade média de 95,17% e uma especificidade média de 93,16% As coordenadas dos átomos para a construção da matriz de distâncias foram os do carbono Alfa (C_α)

Resultados do regressão logística no conjunto *golden standard*

Superfamilia	Sensibilidade. (Média)	Especificidade. (Média)	Media Harmônica (Média)	Curva ROC (Média)
amidhydrolase	0,9703	0,9645	0,9666	0,9668
crotonase	0,9812	0,9878	0,9840	0,9225
enolase	0,9775	0,9746	0,9759	0,9996
haloacid dehalogenase	0,9262	0,9743	0,9468	0,9853
isoprenoid synthase type I	1,0000	1,0000	1,0000	1,0000
vicinal oxygen chelate	1,0000	0,9826	0,9911	0,9905

Tabela 6: Com regressão tradicional alcançamos uma média harmônica (média) 97,61%, sensibilidade média de 98,59% e uma especificidade média de 98,39%, As coordenadas dos átomos para a construção da matriz de distâncias foram usados todos átomos da cadeia

Resultados do método Adhoc no conjunto *golden standard*

Superfamilia	Sensibilidade. (Média)	Especificidade. (Média)	Media Harmônica. (Média)	Curva ROC (Média)
amidhydrolase	0,9626	0,9477	0,9548	0,9769
crotonase	0,9784	0,9566	0,9670	0,9828
enolase	0,9222	0,9267	0,9237	0,9570
haloacid dehalogenase	0,9677	0,9537	0,9598	0,9762
isoprenoid synthase type I	1,0000	0,9856	0,9927	0,9958
vicinal oxygen chelate	0,9600	0,9589	0,9584	0,9757

Tabela 7: Regressão logística em *data sets* balanceados: o método **Adhoc** alcançou 95,94% de média harmônica (média), sensibilidade média de 96,52% e uma especificidade média de 95,49%, usando uma matriz de distâncias construída a partir das coordenadas dos átomos da cadeia principal da proteína

Resultados do método Adhoc no conjunto *golden standard*

Superfamilia	Sensibilidade. (Média)	Especificidade. (Média)	Media Harmônica (Média)	Curva ROC (Média)
amidhydrolase	0,9725	0,9730	0,9726	0,9856
crotonase	0,9895	0,9705	0,9797	0,9870
enolase	0,9595	0,9537	0,9563	0,9761
haloacid dehalogenase	1,0000	0,9733	0,9863	0,9940
isoprenoid synthase type I	0,9866	0,9763	0,9808	0,9847
vicinal oxygen chelate	0,9833	0,9800	0,9825	0,9896

Tabela 8: Regressão logística em *data sets* balanceados: o método **Adhoc** alcançou 97,64% de média harmônica (média), sensibilidade média de 98,22% e uma especificidade média de 97,11%, usando uma matriz de distâncias construída a partir das coordenadas dos todos átomos da um das cadeias

Apêndice B

Artigo 2: On the bridging the gap to use search engine techniques in bioinformatics

1. Autores: **Lima, PL**; Santos, MA
2. Ano: 2023
3. Artigo em processo de revisão

Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Subject Section

On the bridging the gap to use search engine techniques in bioinformatics

Lucio Paccori Lima [1,*], Co-Authors, Marcos A. dos Santos [2]

[1]Institute of Biological Sciences and Department of Computer Science, Federal University of Minas Gerais, Belo Horizonte, MG 31270-901, Brazil and [2]Department of Computer Science, Federal University of Minas Gerais, Belo Horizonte, MG 31270-901, Brazil.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Several problems in bioinformatics are rooted in the search for relationships between components not prima facia exposed. They share some similarities with the latent information retrieval process in modern search engines. Machines as Yahoo, Google and others, use artificial intelligence techniques to transform huge amount of data in discovering and knowledge. In their core stable applied mathematics, statistics and computer science. The logistic regression occupies a rough explored place, even so it has a lot to offer. It is robust, consistent, elegant and an easy to understand methodology. Meantime, serious limitations in certain data sets prevent or limit its use. It happens when the number of attributes of an entity is much greater than the domain itself, which is present in classification problems with microarrays, repurposing of approved drugs, virtual ligand screennig, snips, etc. Furthermore, an unbalanced domain may also be present alongside other drawbacks such as the very topology of entities in space.
Results: We show how applied mathematics and computer science can be used for classification, information retrieval and data mining in bioinformatics. One of the consequences, which we have successfully taken advantage of, is the assessment of the importance of attributes. It is a result that presents itself collaterally and that refers to an open area in data mining. In addition, new algorithms, built from the neighborhood to a specific query, are presented. As an exercise of the new proposals, we developed an application in the area of structural bioinformatics. It adds to the others in which we have successfully explored this relationship. Algebraic tools that make search engines viable, along with logistic regression modifications, underpin the new algorithms.
Contact: paccori@uol.com.br
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

This work maps out algorithms to use logistic regression, mathematical programming and linear algebra for data mining and information retrieval in bioinformatics. We already have a range of applications based in these tools, such as in microarrays (Morais-Rodrigues *et al.*, 2020), repurposing of drugs and virtual screening of ligands (Leite *et al.*, 2020), search for druggable targets (Silvério-Machado R, 2014), phylogenetic trees (Santos *et al.*, 2011). All these applications start from a vector space model (**vsm**) from which they access an overarching set of applied mathematical results that makes search engines feasible. We now explore this methodology in structural bioinformatics. We have aimed to make modest demands on the mathematical expertise of the reader and use a simple application that prioritizes to show the feasibility of the technique and inspire and stalk other applications. In this opportunity, we added new strategies capable of making the application of logistic regression more robust and efficient. The algorithms, although contextualized here in this specific application, are added to the previous processes and are general enough to be used in other artificial intelligence problems.

There are great methods to recover proteins in their respective superfamilies, families and diverse subgroups in structural bioinformatics. From the description of proteins by a **vsm**, where each entity is represented by a fixed size set of numerical attributes, traditional algorithms are applied, namely; neural networks, tree search, vector support machine and others (Perfetti and Ricci, 2006), (Landwehr *et al.*, 2005), (Wang *et al.*, 2005)

and (Ma *et al.*, 2014). They are able to efficiently predict which group a given protein belongs to. As far as we know, (Pires *et al.*, 2011) has one of the best results. They allow the construction of perfect oracles (considered by us when the harmonic mean obtained from cross-validation is greater than 0.90). This classification problem, which is already well solved in the literature, will serve to validate the methodology. Other more sophisticated applications may come from the template presented here (see the conclusions).

Also, the justification and much of the motivation for this work is to face a fact that happens in nature. It is known that proteins with similar primary sequences correspond to tertiary structures that are also similar (Leach, 1996), (Antczak *et al.*, 2016) e (Kolodny and Linial, 2004). However, there are cases in which similar three-dimensional structures do not keep similarity to the primary sequences. We hope to be contributing to the search for signatures that are invariant in this regard.

The aforementioned mining artifacts are not known to be suitable for choosing which attributes (and their values) contribute most to a given individual being in a certain group. An interesting method that has, as a collateral result, an evaluation of such attributes, is logistic regression (Richardson, 2011) ; a well-known and widely used technique in the medical field (Fort and Lambert-Lacroix, 2004), (Hosmer and Lemeshow, 2000). But its generic use as a classification instrument brings a series of inconveniences. In real problems, there is not always a domain with an adequate balance between the quantity (and variety) of the members of the group to be classified and the quantity (and variety) of the total population (domain). For example, in our case, we have superfamilies with 27000 entities alongside others with only a few dozen. These and other inconveniences are resolved by the algorithms we are proposing and/or adapting.

2 Material and Methods

2.1 Data Sets

We selected three well known data sets to discuss our algorithms in structural bioinformatics. Initially, we used a set of enzyme superfamilies, considered a gold standard, which have different mechanisms to perform their functions (Brown *et al.*, 2006). In this set, six superfamilies are considered (amdohydrolase, crotonase, haloacid dehalogenase, isoprenoind synthase type I and vicinal oxygen chelate), comprising 47 families distributed in 896 different chains. Next, we work with data grouped according to a structural classification. It contains a set of 4 unbalanced data sets, namely 6SSE, 5SSE, 4SSE and 3SSE (Jain and Hirst, 2010). Finally, we used all the structures present in SCOP (*Structural Classification of Proteins, version 1.75*), from which the identifiers for the *Protein Data Bank* (**PDB**) were retrieved (`\https://www.rcsb.org/`), which classifies proteins into classes, folds, superfamilies and families (Berman *et al.*, 2002).

2.2 Singular Value Decomposition

The singular value decomposition (**svd**) is a linear algebra technique used to assign a space with a reduced dimension to represent the set of entities without compromising their essence (Elden, 2019; Berry *et al.*, 1995). With **svd**, instead of working with a matrix of approximate rank of the original matrix, we can use another one, where entities are represented by a linear combination of the most important patterns present in the matrix. If two entities are similar, they also have linear combination of such patterns that are similar. In general, for the purpose of retrieving information in real problems, not many patterns are needed, implying a more economical and consistent representation. Formally, the singular value decomposition is a factorization of any matrix into three other matrices with important properties. It has several applications, both direct, in which the results extracted from its factor matrices are applied, and as a step in many algorithms.

Definition 1. *Given $A \in \mathbb{R}^{m\times n}$, not necessarily of full rank, the singular value decomposition of A is a factorization such that:*

- $A = USV^T$;
- $U \in \mathbb{R}^{m\times m}$, *are the eigenvectors of AA^T and is orthogonal;*
- $V \in \mathbb{R}^{n\times n}$, *are the eigenvectors of A^TA and is orthogonal;*
- $S \in \mathbb{R}^{m\times n}$, *is diagonal if $m = n$, otherwise add $m - n$ rows of zeros in S and is formed by the square root of the eigenvalues of AA^T.*

$$S = diag(\sigma_1, \sigma_2, .., \sigma_n)$$

$$\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_n \geq 0.$$

The values $(\sigma_1, \sigma_2, ...\sigma_n)$ are called singular values of A. The columns of U, $(u_1, u_2, ...u_m)$ are called left singular vectors of A and the columns of the matrix V, $(v_1, v_2, ..., v_n)$ are the singular vectors to the right of A. U columns can be interpreted as patterns of entities in A with a corresponding weight σ_i; V are the patterns of the lines, which have the same associated weight. Note that since $A = U(SV^T)$ and U are patterns ($UU^T = I$), each column j of the matrix SV^T indicates the combination of patterns to obtain the entity j. To this observation we can add a different way of looking at **svd**, known as the dialic decomposition of A:

$$A = \sum_{i=1}^{p} \sigma_i u_i v_i^T,$$

which corresponds to a rank one sum of matrices (p is a rank of A). It is important to note that a value of σ_i close to zero implies a contribution almost like noise to the relationship between entities and can be a way of masking similarities that may exist. By discarding these contributions to work with an approximation of rank k of A; $A_k \approx A$ where $A_k = \sum_i \sigma_i u_i v_i^T$, $i = 1, 2, \ldots, k$; the formation of clusters of entities with latent relationships are encouraged. But it is not necessary to use A_k that has the same dimensions of$A$. The matrix $S_k(U_k)^T$ has k lines, instead of m, where k represents the rank adopted and U_k (V_k) is the k first columns (lines) of U (V_k) is used to represent A. The value of k is obtained experimentally. Its search starts from the point where the long tail, usually existing in the plot of σ values, begins. When the matrix has the same number of rows and columns, we have a particular case. The singular value decomposition is better known as spectral decomposition, where the weights σ_i are the eigenvalues of the matrix.

We are employing **svd** to improve the data mining process in three ways. First, as a tool to help visualize data sets in three-dimensional space (Marcolino *et al.*, 2010), which allows us to have an initial assessment of possible difficulties in the classification process with regard to the topology of the groups of entities. Another application comes from the improvements that we are proposing in the algorithms. As we retrieve subsets of entities that are the closest to a given query, we get the best results applying the techniques for building search engines (Elden, 2019). Finally, we directly use the singular values (eigenvalues) like a **vsm** to represent the structures.

2.3 Representation of the protein structures

In Artificial Intelligence, the representation of domain entities (**vsm**) is fundamental. The adopted **vsm** is expected to be robust, in the sense that different entities do not occupy the same point in space. Furthermore, it has to be consistent. The distance between two entities in the domain, measured by an appropriate metric, must reflect reality. We use two representations for the three-dimensional structures of proteins. The first one,

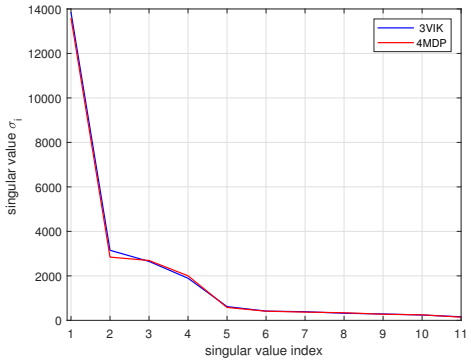


Fig. 1: *Spectral pattern* of the proteins $3VIK$ and $4MDP$ that have high structural similarity. Its residues have only 35 percent of identity, but the cosine between them is ≈ 1.

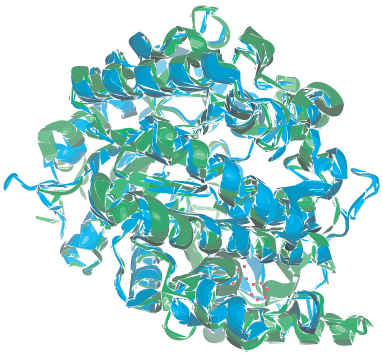


Fig. 2: Structural alignment of proteins $3VIK$ and $4MDP$ (PyMOL) that share only 35 percent of residues in common.

cutoff scanning, the protein is described from the number of atoms in subintervals from 0.2 ångströms up to a maximum distance of 30 ångströms (Pires *et al.*, 2011). The *spectral pattern* representation is based on an interesting and unexplored fact.The matrix of interatomic euclidean distances D of a protein (or part of them) has a peculiar property. When plotting the singular values obtained from the decomposition of this matrix, we observe the presence of six significant values. Always the long tail starts from the sixth value. Occurs in all proteins, no matter which are the atoms used (main chain, alpha carbons, etc). Thus, to represent each protein, we adopt the set of singular values (σ) derived from the decomposition of its distance matrix D. Each singular value (σ_i) is one attribute in the **vsm** to represent the structure of each protein. Behind this idea is the fact that, given two proteins $p1$ and $p2$ with similar three-dimensional structures, their interatomic distance matrices D_{p1} and D_{p2} are such that $\parallel D_{p1} - D_{p2} \parallel_{\mathbf{F}} \leq \delta$, where $\delta > 0$ is a small enough value. Therefore, they use the rank 1 matrices from the dialic decomposition in a similar way. As an example of a possible application, suppose a hypothetical protein is known for which a model is available. It is then possible to determine which protein deposited in the **PDB** has a similar three-dimensional structure comparing only their *spectral pattern* **vsm**. As an example, consider the two proteins $3VIK$ (491 residues) and $4MDP$ (481 residues). From the alignment we see that they have 35% of identity. Their spectral *spectral pattern* are displayed in figure **1**. The structural alignment between them (figure **2**) is consistent with the high similarity of the two representations.

2.4 Logistic Regression

Logistic Regression consists of finding values of $\alpha = (\alpha_1, \alpha_2, ..., \alpha_n)^T$ to fit

$$P_j(x) = \frac{e^{\sum_{i=1}^{m} \alpha_i x_i}}{1 + e^{\sum_{i=1}^{m} \alpha_i x_i}} \tag{1}$$

for each of the j entity. The value of the function (1), $P_j(x)$, with $P_j(x) \in (0, 1)$ informs the "probability" of a given entity being classified as belonging to a specific subset. Note that when $e^{\sum_{i=1}^{m} \alpha_i x_i}$ tends to zero, $P_j(x)$ also tends to zero. On the other hand, if $e^{\sum_{i=1}^{m} \alpha_i x_i}$ approaches infinity, $P_j(x)$ approaches to 1.

In order to proceed with the determination of the values of α, a linear transformation is carried out which will refer the problem to the solution of a linear algebra problem. Let the chance $C_j(x)$ be defined by:

$$C_j(x) = \frac{P_j(x)}{1 - P_j(x)}. \tag{2}$$

Substituting (1) into (2), we get

$$C_j(x) = e^{\sum_{i=1}^{m} \alpha_i x_i} \tag{3}$$

Taking the logarithm on both sides of (3), we obtain a system of linear equations for determining α:

$$b_j = \sum_{i=1}^{n} \alpha_i x_i \tag{4}$$

where $b_j = \log(C_j), \quad \text{for } j = 1, 2, \ldots, m.$

Thus, we associate a matrix $A = \{a_{i,j}\} \in \mathbb{R}^{\mathrm{m} \times \mathrm{n}}$ to the data set. Now, rows represent entities and the columns of A are associated with attributes. The $a_{i,j}$ is the value of the j attribute in the i structure. Note that this is different from the way we treated it earlier when we discussed the singular value decomposition (this matrix is the transpose of the previous one). The value of m is the number of entities and n the number of attributes.

Let $b = (b_1, b_2, ..., b_m)^T$; the system of linear equations (4) can be represented by:

$$A\alpha = b. \tag{5}$$

In order to assign values to the vector α when the number of entities exceeds the number of attributes ($m > n$), the classical approach associates to (5) an unrestricted nonlinear programming problem that minimizes the sum of the squares of the residuals associated with each line of the system:

$$Minimize \;\; q(\alpha) = ||A\alpha - b||^2 \tag{6}$$

If $A^T A$ has rank n, the unique solution of (6) is given by solving the system of linear equations of order n:

$$A^T A\alpha = A^T b. \tag{7}$$

In problems where $m < n$, which normally occur in the myriad of classification problems in bioinformatics (see references in the introduction), the quadratic model (6) does not apply because the normal matrix $A^T A$ does not have full rank. Generally, to get around this difficulty, a group of variables is discarded (*feature selection*), keeping only a subset of the original variables. We use a stabilizing term in the logistic regression model,

found in the works of (Morais-Rodrigues *et al.*, 2020), which allow the assignment of values for the parameters α by minimizing the sum of squares of the residuals ($A\alpha - b$), added to the squares of α. Thus, to find a solution to (5), we solve an unrestricted quadratic optimization problem given by

$$Minimize\ q_q(\alpha) = ||\alpha||^2 + ||A\alpha - b||^2. \quad (8)$$

As the function $q_q(\alpha)$ is convex, the argument α^* that solves (8) is given by the derivation of $q_q(\alpha)$ with respect to α and setting the result to zero. This results in a system of linear equations

$$(I + A^T A)\alpha = A^T b, \quad (9)$$

where I is an identity matrix of order n. That is, the optimal solution for $q_q(\alpha)$ in (8) is obtained by the solution of (9) and it is unique.

Some authors include a variable α_0 to compute $P_i(x)$ in (1). It is not associated with any attribute and can be interpreted as a regularization of the opening of the closed interval $[0, 1]$ (note that at the extremes of the interval, the logarithm of chance associated with the entity i is not defined). This does not significantly alter the classifier results and, for clarity, has not been included in the text.

Other mathematical programming models can be formulated to assign values to α. In particular, instead of formulating equalities, we can think of inequalities and work with other objective functions to circumvent, for example, ambiguities present in the considered domain. As this is not a recurring case in bioinformatics problems, we refrained from presenting these models in this work.

Finally, once given a structure represented by $q = (q_1, q_2, ..., q_n)$, the "probability" that q belongs to a class associated with the related system is given by

$$P(q) = \frac{e^{q\alpha}}{1 + e^{q\alpha}}. \quad (10)$$

2.5 Adhoc Method

Adhoc is from a Latin expression whose literal translation is "for this" or "for this purpos". It is used to designate something that was formed or used for a particular and immediate purpose or need, without prior planning. The **Adhoc** method uses models built exclusively in response to a single demand for classification. Given a query q, a model **Adhoc**, specific to this query, is built from the choice of k_0 entities closest to q such that $P_j(x) = 0$, alongside of k_1 entities chosen among those closest to q such that $P_j(x) = 1$. In summary, the **Adhoc** method is a resource that explores a neighborhood concept in logistic regression, allowing its application in different scenarios.

We didn't apply a sophisticated method to assign values to k_0 and k_1; they were easily determined experimentally. In our tests we found that these values are very low. In the cross-validation, we verified that for the good performance of the classification, it is enough to adopt values in the order of a few units to these two parameters. Not uncommon, sometimes this results in matrices $M \in \mathbb{R}^{(k_0+k_1)\times n}$ in with $(k_0 + k_1) < n$ which prevents the use of traditional logistic regression. We address this limitation using modified logistic regression as explained in the previous section.

Another aspect is the choice of closest $k_{(0/1)}$ entities, which could impact the response time in large problems. This is solved by organizing the entities in search trees according to resources used in search engines as **svd**, clustering, etc. Thus, the matrix A is partitioned into two others, A_0 and A_1, according to $P_j(x) = 1/0$. Each of them has its entities organized like elements of a search engine. In this way, retrieving the closest entities to a query q is efficient and does not noticeably impact processing time.

3 Results

The experiments were designed to validate logistic regression in the context of the proposed structural bioinformatics application. So, first we checked how it behaves against the techniques that were used in (Pires *et al.*, 2011). The same dataset and the same **vsm** were used. We also tested the new way of representing the structure of proteins - the *spectral pattern*. To handle unbalanced data sets, the **Adhoc** method was applied.

We adopted the harmonic mean for the evaluation of the classifiers. This metric relates sensitivity and specificity and, together with the **ROC** curve, it gives a precise indication of the behavior of the classifiers. The performance was measured in 20 randomly constructed *folders*. The procedure was repeated to exhaustion in almost all cases and the presented results remained without expressive alterations.

The atoms considered were those of the main chain (or part of it), constituted either by C_α, or by atoms (C_α, C, N). But also, in some cases, we use all of the atoms in the protein. However, as there were no major changes in the results, we decided to present, in most cases, the results considering only C_α.

Once the **PDB** records of the classifier domain have been downloaded, the **vsm** for each protein can be determined, constituting the columns of an A matrix. Thus, the decomposition by singular values is applied to the matrix of entities ($A = USV^T$) allowing to display their singular values (matrix S diagonal). In the *golden standard* proteins represented by **vsm** *cutoff scanning*, the spectrum (not shown) suggests the presence of six main patterns. Each entity can be represented by six values, which are the linear combination of the six most important patterns of A, given by $S_6(V_6)^T$. This is done to display the domain of the data set in three dimensions (see figure 3). From this figure it is apprehended that the domain in which the proteins are represented, does not seem to have any specificity that can be used in the classification process. This fact occurs in practically all the problems discussed in this work, regardless of the **vsm** used.

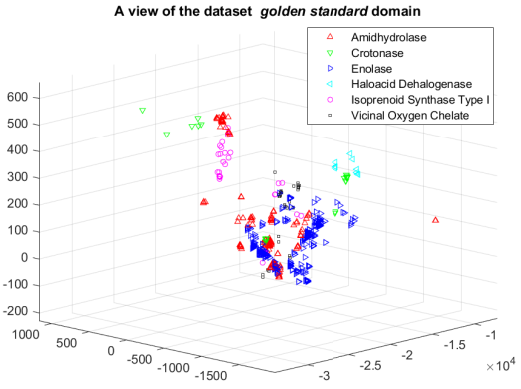


Fig. 3: Projection of the representation of protein structures described by **vsm** *cutoff scanning*. After the singular values decomposition, each protein started to be represented by 6 attributes from the original 150. The plotting of these proteins in the R^3 space was done according to (Marcolino *et al.*, 2010).

Below we present the results obtained with the 3 data sets. In the tables that follow, we refrained from placing a specific column with the results obtained in (Pires *et al.*, 2011), since there is no specific set for tests. Another observation is that in the cited work there is an improvement in

performance using a reduced rank matrix $A_k \approx A$ instead of A, which has the values of the attributes modified. As our other objective refers to the collection of evidences for some possible attribute selection (and their values) that may characterize structural signatures, we avoid any alteration in **vsm**.

Superfamilia	Sensibilidade. (Média)	Especificidade. (Média)	Media Harmônica (Média)	Curva ROC (Média)
amidhydrolase	0,9756	0,9581	0,9664	0,9597
crotonase	1,0000	0,9890	0,9944	0,9984
enolase	0,9779	0,9931	0,9850	1,0000
haloacid dehalogenase	0,9571	0,9573	0,9551	0,9587
isoprenoid synthase type I	1,0000	0,9994	0,9997	1,0000
vicinal oxygen chelate	1,0000	0,9968	0,9984	1,0000
Média	0,9851	0,9823	0,9831	0,9861

(a) *cutoff scanning*

Superfamilia	Sensibilidade. (Média)	Especificidade. (Média)	Media Harmônica (Média)	Curva ROC (Média)
amidhydrolase	0,8925	0,7608	0,8145	0,8943
crotonase	0,7626	0,9061	0,8228	0,8424
enolase	0,9245	0,8874	0,9044	0,9466
haloacid dehalogenase	0,3745	0,9970	0,5353	0,6833
isoprenoid synthase type I	0,8242	0,9880	0,8900	0,9075
vicinal oxygen chelate	0,8417	0,8495	0,8429	0,8699
Média	0,7700	0,8687	0,8017	0,8573

(b) *spectral pattern*

Table 1. Performance of the classifiers constructed using Logistic Regression in the *golden standard* dataset. In **1a** the best results were obtained using all atoms coordinates of the protein backbone. In table **1b**, the best results were obtained using only the C_α coordinates.

The *gold standard* dataset is well balanced in the sense that each category has roughly the same amount of entities. Traditionally applied logistic regression performs similarly to the cited reference when using the same **vsm** (see table **1a**). With the new representation (*spectral pattern*), the results indicate a difficulty in obtaining perfect oracles (see table **1b**). Perhaps its performance can be improved by adding others dimensions to *spectral pattern*, such as the number of residuals and/or some interpro annotation (Hunter *et al.*, 2012). When the **Adhoc** was applied, we got perfect oracles in both representations (see tables **2a** and **2b**). In particular, note the excellent result obtained with the use of **vsm** *spectral patern*.

Superfamilia	Sensibilidade. (Média)	Especificidade. (Média)	Media Harmônica (Média)	Curva ROC (Média)
amidhydrolase	0,9699	0,9419	0,9553	0,9794
crotonase	0,9818	0,9643	0,9727	0,9836
enolase	0,9351	0,9194	0,9266	0,9584
haloacid dehalogenase	0,9737	0,9559	0,9643	0,9831
isoprenoid synthase type I	1,0000	0,9742	0,9866	0,9926
vicinal oxygen chelate	0,9533	0,9657	0,9587	0,9734
Média	0,9690	0,9535	0,9607	0,9784

(a) *Cutoff scaning*

Superfamilia	Sensibilidade. (Média)	Especificidade. (Média)	Media Harmônica (Média)	Curva ROC (Média)
amidhydrolase	0,9983	0,9991	0,9986	0,9978
crotonase	0,9969	0,9983	0,9975	0,9965
enolase	0,9972	0,9981	0,9972	0,9981
haloacid dehalogenase	0,9961	0,9931	0,9994	0,9960
isoprenoid synthase type I	1,0000	0,9997	0,9998	0,9997
vicinal oxygen chelate	1,0000	0,9990	0,9998	0,9991
Média	0,9981	0,9979	0,9987	0,9979

(b) *spectral pattern*

Table 2. Performance of the classifiers constructed using **Adhoc** method in the *golden standard* dataset. All of then use C_α coordinates to build the interatomic distance matrices for each protein.

Conjuntos	Sensibilidade. (Média)	Especificidade. (Média)	Media Harmônica (Média)	Curva ROC (Média)
3SSEs	0,9450	0,9556	0,9483	0,9659
4SSEs	0,9644	0,9664	0,9645	0,9824
5SSEs	0.8960	0,9451	0,9111	0,9148
6SSEs	0,9434	0,9599	0,9489	0,9711
Média	0,9372	0,9568	0,9432	0,9586

(a) *Cutoff scaning*

Conjunto de dados	Sensibilidade. (Média)	Especificidade. (Média)	Media Harmônica (Média)	Curva ROC (Média)
Full Scop	0,9303	0,9304	0,9301	0,9322
3SSE	0,9661	0,9173	0,9391	0,9445
4SSE	0,9934	0,9221	0,9299	0,9161
5SSE	0,9713	0,9662	0,9684	0,9710
6SSE	0,9388	0,9326	0,9310	0,9564
Média	0,9600	0,9338	0,9385	0,9440

(b) *spectral pattern*

Table 3. Results of the **Adhoc** method in unbalanced datasets. We achieve perfect oracles in all tested instances.

The two other datasets are unbalanced. SCOP covers a large set of proteins that give rise to various types of oracles. We haven't done exhaustive testing using *cutoff scanning* **vsm** where the best performance seems to be with all the atoms of the protein. It would demand disproportionate resources to those we used in the other case (*spectral pattern*), where we used only the atoms of the main chain. However, everything seems to indicate that the results would be very similar. We chose to summarize the results with this dataset in a single line in the table in **3b**. So, it can be seen that the use of the **Adhoc** strategy resulted in perfect oracles in both representations of the structure. Here too, it is worth noting that, given the variety of types and sizes of proteins used, we believe that performance could be improved by introducing new dimensions in the structure representation.

The results corroborate the usefulness of logistic regression in the proposed application, since perfect oracles were obtained. The performance of the classifiers also indicates the robustness and consistency of the representation of the structures by the singular values of the interatomic distance matrices. In figure **4** the values of the **vsm** *spectral pattern* of the proteins of each of the superfamilies in the dataset *golden standard* are displayed. It is possible to see subgroups and possible archetypal curves to represent the various groupings.

To select the attributes that can characterize a structure signature, we follow a simple procedure. We discard those attributes whose values of α_i are close to zero (which do not have discriminatory power). Of course, this process can be more sophisticated. However, the simple act of choosing attributes whose α_i values are topologically farthest from the x axis has worked adequately in other applications. These attributes are taken from the A matrix and a cross-validation provides a proof of concept for this procedure.

For example, consider the profiles of a set of patients with luminal A early breast cancer alongside healthy others whose miRNA counts are deposited in the **NCBI** (`https://www.ncbi.nlm.nih.gov/`) under the identifier $GSE46335$. The patient profile is easily retrieved and the A matrix can be assembled. The independent terms b_i associated with patients are set to $log(chance(x) = 1)$ or $log(chance(x) = 0)$, according to the profile (has or does not have cancer) of each patient. Thus, solving the system of linear equations (9) and displaying the values of α we select those attributes such that $|\alpha_i| \geq \pm 1.2 \times 10^{-3}$ (see figure 5). It is not a case of cross-validation, given the reduced number of patients. But, in this simple example, the importance of the technique can be assessed.

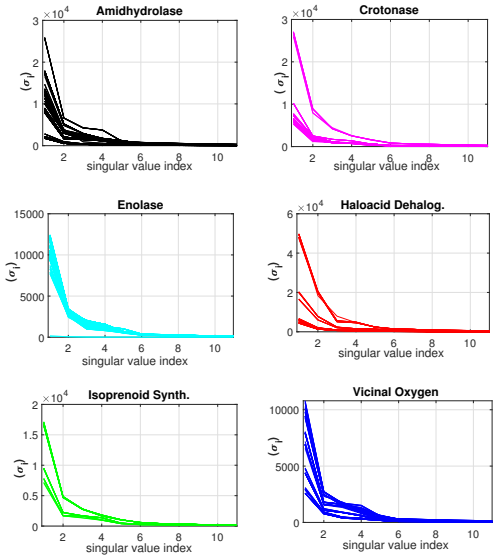


Fig. 4: Set of the 11 highest singular values (σ_i) of the interatomic distance matrices of all proteins in the *gold standard* dataset.

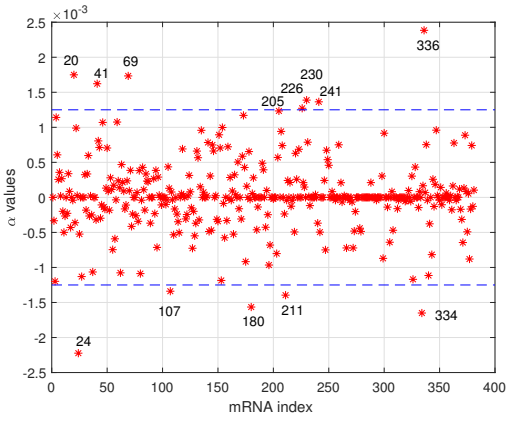


Fig. 5: Values of weights α associated with miRNAs from dataset $GSE46355$. Note that the positive α_i weights contribute to the occurrence of the disease and the negative are protective factors (possibly). The adopted cut-off value was $\pm 1.2 \times 10^{-3}$. Thus, the attributes (points) selected in the figure are: hsamiR121 (20), hsamiR1393p (41), hsamiR182 (69) , hsamiR411 (205), hsamiR4835p (226), hsamiR4863p (230) , hsamiR494 (241), hsamiR6285p (336), hsamiR126 (24), hsamiR216 (107), hsamiR3625p (180), hsamiR429 (211) and hsamiR625 (334).

From the methodology we are using, it is possible to obtain a set of possible biological markers to be investigated. In (Morais-Rodrigues *et al.*, 2020) there are studies with microarrays of gene expressions of different types of breast cancer.

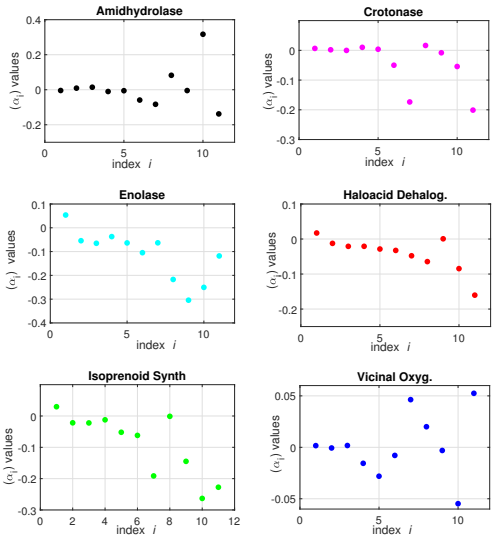


Fig. 6: Scatter plots to help choose the most significant attributes to classify each superfamily of the *golden standard* dataset. The weights (α_i), computed in classifiers, are associated to the attributes.

When applying the same reasoning with *cutoff scanning*, we were not successful, because the attributes of this **vsm** changed when applying the **Adhoc** method. Furthermore, it was not clear to us the biological significance associated with a set of selected attributes. The results regarding the stability aspect of the attribute were quite different with the *spectral pattern*. In the figure (6) a group of attributes that are distant from the x axis in the cross-validation can be selected. Such attributes of the *spectral pattern* are weights associated with the rank 1 matrices that are added to recompose the matrix of interatomic distances in the dialic decomposition. We do not have an explanation of the meaning of each of these rank 1 matrices. Perhaps they are associated with the existing force fields between the protein atoms.

4 Conclusions

Some problems in bioinformatics resembles the art of retrieving information on current search engines.They are the ones that deal with huge datasets in which their components keep latent information among themselves. Although artificial intelligence techniques are constantly evolving, at their core remain some concepts discussed here, such as vector space models, matrices, patterns, etc. We believe that logistic regression is a practical and easy to apply approach to take advantage of the search engine tools in bioinformatics. Our proposed improvements breathe new life into this methodology and are general enough to be used in other contexts and applications. Even if there is interest in using some other resource to retrieve information, the presented methods can be used in large scale problems, do not demand sophisticated computer systems and can be used as a pre-processing tool to select attributes.

As an example of an almost direct application, concerns to the cavities present in proteins and which are targets of approved drugs that modulate them. By associating a **vsm** to each cavity, mechanisms can be developed to reposition drugs from latent searches, similar to the way that was

successfully used to obtain new drug targets (Silvério-Machado R, 2014). Another source of interesting problems are the datasets deposited at the NCBI with experiments with miRNA's. The techniques presented here have almost direct application. There are opportunities for diagnostic kits and disease marker discovery.

Proteins are wonderful artifacts of nature designed to sustain life as we know it. Their three dimensional structures give them the function they perform. It seems natural to represent them using the coordinates of their atoms (or parts of them). The interatomic distance matrix was the way we found to do this. The representation using the matrix spectrum is robust and consistent. However, it has not escaped us that the meaning of the decomposition (matrices of rank one) associated with the value of each of the attributes has yet to be discovered.

References

Antczak, M. *et al.* (2016). Structural alignment of protein descriptors – a combinatorial model. *BMC Bioinformatics*, **17**, 383.

Berman, H. M. *et al.* (2002). The Protein Data Bank. *Acta Crystallographica Section D*, **58**(6 Part 1), 899–907.

Berry, M. W. *et al.* (1995). Using linear algebra for intelligent information retrieval. *SIAM Review*, **37**(4), 573–595.

Brown, S. *et al.* (2006). A gold standard set of mechanistically diverse enzyme superfamilies. *Genome biology*, **7**, R8.

Elden, L. (2019). *Matrix Methods in Data Mining and Pattern Recognition*. SIAM-Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2nd edition.

Fort, G. and Lambert-Lacroix, S. (2004). Classification using partial least squares with penalized logistic regression. *Bioinformatics*, **21**(7), 1104–1111.

Hosmer, D. W. and Lemeshow, S. (2000). *Applied logistic regression*. H 2ed New York: John Wiley and Sons.

Hunter, S. *et al.* (2012). Interpro in 2011: New developments in the family and domain prediction database. *Nucleic acids research*, **40**(1), D306–D312.

Jain, P. and Hirst, J. D. (2010). Automatic structure classification of small proteins using random forest. *BMC Bioinform.*, **11**, 364.

Kolodny, R. and Linial, N. (2004). Approximate protein structural alignment in polynomial time. *Proceedings of the National Academy of Sciences*, **101**(33), 12201–12206.

Landwehr, N. *et al.* (2005). Logistic model trees. *Mach. Learn.*, **59**(1-2), 161–205.

Leach, A. (1996). *Molecular Modelling: Principles and Applications*. Pearson education. Longman.

Leite, C. *et al.* (2020). Milk-way algorithm for ligand-based virtual screening: Cdk2 case study. *Trends in Developmental Biology*, **13**, 20.

Ma, C. *et al.* (2014). Machine learning for big data analytics in plants. *Trends in Plant Science*, **19**(12), 798–808.

Marcolino, L. S. *et al.* (2010). Genome visualization in space. In M. P. Rocha, F. F. Riverola, H. Shatkay, and J. M. Corchado, editors, *Advances in Bioinformatics*, pages 225–232, Berlin, Heidelberg. Springer Berlin Heidelberg.

Morais-Rodrigues, F. *et al.* (2020). Analysis of the microarray gene expression for breast cancer progression after the application modified logistic regression. *Gene*, **726**, 144–168.

Perfetti, R. and Ricci, E. (2006). Analog neural network for support vector machine learning. *IEEE Transactions on Neural Networks*, **17**(4), 1085–1091.

Pires, D. *et al.* (2011). Cutoff scanning matrix (csm): Structural classification and function prediction by protein inter-residue distance patterns. *BMC genomics*, **12 Suppl 4**, S12.

Richardson, A. (2011). Logistic regression: A self-learning text, third edition by david g. kleinbaum, mitchel klein. *International Statistical Review*, **79**, 296–296.

Santos, A. R. *et al.* (2011). A singular value decomposition approach for improved taxonomic classification of biological sequences. *BMC genomics*, **12 Suppl 4**, S11.

Silvério-Machado R, Couto BR, D. S. M. (2014). Retrieval of enterobacteriaceae drug targets using singular value decomposition. *Bioinformatics. 2015 Apr 15;31(8):1267-73*, **31**.

Wang, J. T. L. *et al.* (2005). *Introduction to Data Mining in Bioinformatics*, pages 3–8. Springer London, London.

Apêndice C

Regressão Logística usando o vetor space model com o cutoff scanning

C.1 Resultados usando a Logística Tradicional para diferentes átomos

Resultados do Regressão Logística no conjunto *golden standard*

Superfamília	Sensibilidade. (Média)	Especificidade. (Média)	Média Harmônica (Média)	Curva ROC (Média)
amidhydrolase	0,8881	0,8727	0,8778	0,9266
crotonase	1,0000	0,9455	0,9477	1,0000
enolase	0,9094	0,9402	0,9196	0,9500
haloacid dehalogenase	0,9125	0,9017	0,8969	0,8633
isoprenoid synthase type I	1,0000	0,9908	0,9453	1,0000
vicinal oxygen chelate	1,0000	0,9687	0,9836	0,9919

Tabela C.1. Com a regressão tradicional alcançamos uma média harmônica (média) 92,85%, sensibilidade média de 95,17% e uma especificidade média de 93,16%. As coordenadas dos átomos para a construção da matriz de distâncias foram os do carbono Alfa (C_α)

Resultados do Regressão Logística no conjunto *golden standard*

Superfamília	Sensibilidade. (Média)	Especificidade. (Média)	Média Harmônica (Média)	Curva ROC (Média)
amidhydrolase	0,9756	0,9581	0,9664	0,9597
crotonase	1,0000	0,9890	0,9944	0,9984
enolase	0,9779	0,9931	0,9850	1,0000
haloacid dehalogenase	0,9571	0,9573	0,9551	0,9587
isoprenoid synthase type I	1,0000	0,9994	0,9997	1,0000
vicinal oxygen chelate	1,0000	0,9968	0,9984	1,0000

Tabela C.2. Com regressão tradicional alcançamos uma média harmônica (média) 98,31%, sensibilidade média de 98,51% e uma especificidade média de 98,23%. As coordenadas dos átomos para a construção da matriz de distâncias foram as da cadeia principal

Resultados do Regressão Logística no conjunto *golden standard*

Superfamília	Sensibilidade. (Média)	Especificidade. (Média)	Média Harmônica (Média)	Curva ROC (Média)
amidhydrolase	0,9703	0,9645	0,9666	0,9668
crotonase	0,9812	0,9878	0,9840	0,9225
enolase	0,9775	0,9746	0,9759	0,9996
haloacid dehalogenase	0,9262	0,9743	0,9468	0,9853
isoprenoid synthase type I	1,0000	1,0000	1,0000	1,0000
vicinal oxygen chelate	1,0000	0,9826	0,9911	0,9905

Tabela C.3. Com regressão tradicional alcançamos uma média harmônica (média) 97,61%, sensibilidade média de 98,59% e uma especificidade média de 98,39%, As coordenadas dos átomos para a construção da matriz de distâncias: foram usados todos átomos da cadeia

C.2 Resultados usando a Logística Modelo Adhoc para diferentes centróides

Resultados do método Adhoc no conjunto *golden standard*

Superfamília	Sensibilidade. (Média)	Especificidade. (Média)	Média Harmônica (Média)	Curva ROC (Média)
amidhydrolase	0,9699	0,9419	0,9553	0,9794
crotonase	0,9818	0,9643	0,9727	0,9836
enolase	0,9351	0,9194	0,9266	0,9584
haloacid dehalogenase	0,9737	0,9559	0,9643	0,9831
isoprenoid synthase type I	1,0000	0,9742	0,9866	0,9926
vicinal oxygen chelate	0,9533	0,9657	0,9587	0,9734

Tabela C.4. Regressão Logística em *data sets* balanceados: o método Adhoc alcançou 96,07% de média harmônica (média), sensibilidade média de 96,90% e uma especificidade média de 95,36%, usando uma matriz de distâncias construída a partir das coordenadas dos átomos carbono alfa (C_α) da proteína

Resultados do método Adhoc no conjunto *golden standard*

Superfamília	Sensibilidade. (Média)	Especificidade. (Média)	Média Harmônica. (Média)	Curva ROC (Média)
amidhydrolase	0,9626	0,9477	0,9548	0,9769
crotonase	0,9784	0,9566	0,9670	0,9828
enolase	0,9222	0,9267	0,9237	0,9570
haloacid dehalogenase	0,9677	0,9537	0,9598	0,9762
isoprenoid synthase type I	1,0000	0,9856	0,9927	0,9958
vicinal oxygen chelate	0,9600	0,9589	0,9584	0,9757

Tabela C.5. Regressão Logística em *data sets* balanceados: o método Adhoc alcançou 95,94% de média harmônica (média), sensibilidade média de 96,52% e uma especificidade média de 95,49%, usando uma matriz de distâncias construída a partir das coordenadas dos átomos da cadeia principal da proteína

Resultados do método Adhoc no conjunto *golden standard*

Superfamília	Sensibilidade. (Média)	Especificidade. (Média)	Média Harmônica (Média)	Curva ROC (Média)
amidhydrolase	0,9725	0,9730	0,9726	0,9856
crotonase	0,9895	0,9705	0,9797	0,9870
enolase	0,9595	0,9537	0,9563	0,9761
haloacid dehalogenase	1,0000	0,9733	0,9863	0,9940
isoprenoid synthase type I	0,9866	0,9763	0,9808	0,9847
vicinal oxygen chelate	0,9833	0,9800	0,9825	0,9896

Tabela C.6. Regressão Logística em *data sets* balanceados: o método Adhoc alcançou 97,64% de média harmônica (média), sensibilidade média de 98,22% e uma especificidade média de 97,11%, usando uma matriz de distâncias construída a partir das coordenadas de todos átomos de uma das cadeias