CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL

DESENVOLVIMENTO DE UMA METODOLOGIA BASEADA EM MATRIZ DE DISTÂNCIAS PARA A VERIFICAÇÃO DE SIMILARIDADES DE PROTEÍNAS

OTAVIANO MARTINS MONTEIRO

BELO HORIZONTE
AGOSTO DE 2017

OTAVIANO MARTINS MONTEIRO

DESENVOLVIMENTO DE UMA METODOLOGIA BASEADA EM MATRIZ DE DISTÂNCIAS PARA A VERIFICAÇÃO DE SIMILARIDADES DE PROTEÍNAS

Dissertação apresentada ao Programa de Pós-graduação em Modelagem Matemática e Computacional do Centro Federal de Educação Tecnológica de Minas Gerais, como requisito parcial para a obtenção do título de Mestre em Modelagem Matemática e Computacional.

Área de concentração: Modelagem Matemática e Computacional

Linha de pesquisa: Métodos Matemáticos Aplicados

Orientador: Thiago de Souza Rodrigues
Coorientador: Sandro Renato Dias

CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL
BELO HORIZONTE
AGOSTO DE 2017

Monteiro, Otaviano Martins
M775d Desenvolvimento de uma metodologia baseada em matriz de distâncias para a verificação de similaridades de proteínas / Otaviano Martins Monteiro. – 2017.
xi, 52 f. : il.: tabs, grafs.

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Modelagem Matemática e Computacional.
Orientador: Thiago de Souza Rodrigues.
Coorientador: Sandro Renato Dias.
Dissertação (mestrado) – Centro Federal de Educação Tecnológica de Minas Gerais.

1. Proteínas – Modelagem matemática – Teses. 2. Algorítimo de Kabsch – Teses. 3. Projeto Colaborativo Computacional em Cristalografia Macromolecular. 4. Matrizes (Matemática) – Teses. 5. Aglomeração – Teses. I. Rodrigues, Thiago de Souza. II. Dias, Sandro Renato. III Centro Federal de Educação Tecnológica de Minas Gerais. VI. Título.

CDD 512.9434

SERVIÇO PÚBLICO FEDERAL
MINISTÉRIO DA EDUCAÇÃO
CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS
COORDENAÇÃO DO CURSO DE MESTRADO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL

"DESENVOLVIMENTO DE UMA METODOLOGIA BASEADA EM MATRIZ DE DISTÂNCIAS NA VERIFICAÇÃO DE SIMILARIDADES DE PROTEÍNAS"

Dissertação de Mestrado apresentada por **Otaviano Martins Monteiro**, em 25 de agosto de 2017, ao Programa de Pós-Graduação em Modelagem Matemática e Computacional do CEFET-MG, e aprovada pela banca examinadora constituída pelos professores:

Prof. Dr. Thiago de Souza Rodrigues (Orientador)
Centro Federal de Educação Tecnológica de Minas Gerais

Prof. Dr. Sandro Renato Dias (Coorientador)
Centro Federal de Educação Tecnológica de Minas Gerais

Prof. Dr. Anton Semenchenko
Centro Universitário Newton Paiva

Prof. Dr. Arísio Mendes Lacerda
Centro Federal de Educação Tecnológica de Minas Gerais

Visto e permitida a impressão,

Prof. Dr. José Geraldo Peixoto de Faria
Coordenador do Programa de Pós-Graduação em
Modelagem Matemática e Computacional

Dedico este trabalho à minha mãe Creuzoni Martins Monteiro, e ao meu já falecido pai Geraldo Monteiro da Silva.

Agradecimentos

Agradeço ao meu orientador Thiago de Souza Rodrigues e ao meu coorientador Sandro Renato Dias pelos ensinamentos passados e pela dedicação às orientações.

À minha mãe Creuzoni, juntamente com seu esposo Adelmo, meu irmão Cristiano e minha noiva Ana Flávia pelo apoio e paciência no decorrer do curso.

Aos meus colegas de estudo e demais professores pelos ensinamentos passados.

Ao CEFET-MG pelo suporte oferecido e à CAPES pela bolsa de auxílio.

À Deus, por tornar tudo isso possível.

Resumo

As proteínas são macromoléculas presentes em todos os seres vivos e desempenham funções importantes, tais como manutenção de órgãos e tecidos, diferenciação celular, transporte, entre outras diversas finalidades. Várias proteínas, possuem a sua estrutura tridimensional resolvida e armazenada em bancos de dados biológicos, como o *Protein Data Bank* (PDB). Existem diversos *softwares* que trabalham com informações extraídas do PDB. Algumas dessas ferramentas, tem como função a verificação de similaridades entre estruturas proteicas. Um exemplo é o LSQKAB, que pertence ao pacote CCP4, e tem o seu funcionamento baseado no algoritmo de Kabsch, uma técnica frequentemente utilizada na área de bioinformática que possibilita a comparação entre duas estruturas de proteínas. Este trabalho tem como objetivo desenvolver uma metodologia baseada em uma matriz de distâncias, que verifique a similaridade de trechos de proteínas. A sua precisão deve ser a mesma do LSQKAB e ainda deve ser possibilitado o agrupamento de resíduos de acordo com as similaridades de seus valores de distâncias atômicas. Os experimentos foram realizados com arquivos de interações de resíduos de uma mesma proteína e os resultados foram comparados com o *Cutoff Scanning Matrix* (CSM) e com a técnica do arquivo de referência, utilizada na busca de pares interagentes pela ferramenta RID. Foi possível formar diversos grupos com arquivos similares através de cada técnica avaliada. A metodologia apresentada obteve resultados interessantes diante das outras técnicas comparadas, obtendo uma maior precisão.

Palavras-chave: Proteínas. LSQKAB. Matriz de distâncias. Agrupamentos.

Abstract

Proteins are macromolecules present in all living beings and perform important functions, such as maintenance of organs and tissues, cell differentiation, transportation, and other several tasks. Many proteins have their three-dimensional structure resolved and stored in biological databases, such as Protein Data Bank (PDB). There are various softwares working with informations extracted from PDB. Some of these tools, has the function of checking similarities between protein structures. An example is LSQKAB, which belongs to CCP4 package and has its operation based on Kabsch algorithm, a technique frequently used in bioinformatics that allows the comparison between two structures of proteins. This master thesis aims to develop a metodology based on a distances matrix that verifies the similarity of protein residues. Its precision must be the same as that of LSQKAB and it should still be possible to group residues according to the similarities of their values of atomic distances. The experiments were performed with files of residue interactions of the same protein and the results were compared with the Cutoff Scanning Matrix (CSM) and with the reference file technique, used in the search of interacting pairs by the RID tool. It was possible to form several groups with similar files through each technique evaluated. The presented methodology obtained interesting results in comparison to other techniques, obtaining a greater precision.

Keywords: Proteins. LSQKAB. Distance Matrix. Clustering.

Lista de Figuras

Lista de Algoritmos

Lista de Abreviaturas e Siglas

aCSM	*atomic Cutoff Scanning Matrix*
CSM	*Cutoff Scanning Matrix*
PDB	*Protein Data Bank*
RID	*Residue Interaction Database*
RSMD	*Root Mean Square Deviation*
SVD	*Singular Value Decomposition*

Sumário

1 Introdução

Conforme citado por Nelson, Lehninger e Cox (2008), as proteínas são as macromoléculas mais abundantes nas células vivas, ocorrendo em todas as células e em diversas variedades. Elas desempenham inúmeras tarefas para os seres vivos. Constituídas de um conjunto de aminoácidos, fornecem materiais para construção e manutenção de órgãos e tecidos, diferenciação celular, transporte, entre outras diversas finalidades.

As proteínas tem sido amplamente estudadas por diversos pesquisadores. Várias delas possuem a sua estrutura tridimensional resolvida e armazenada em bancos de dados biológicos, como o *Protein Data Bank* (PDB). De acordo com RSCB (2016a), neste banco de dados, são armazenados em arquivos texto, informações obtidas e registradas por pesquisadores, tais como coordenadas atômicas, ligações, distâncias, composição, função e outras particularidades importantes das proteínas.

De acordo com Dias (2012), o suporte fornecido pelo PDB possibilita diversos tipos de pesquisas na área de bioinformática. Conhecer as estruturas proteicas, bem como encontrar similaridades entre trechos de proteínas, pode resultar em propostas de mutações entre resíduos, visando melhorias em suas estruturas.

Existem diversos *softwares* que trabalham com informações de proteínas extraídas do PDB. Algumas dessas ferramentas, tem como função a verificação de similaridades de trechos de proteínas. Um exemplo é o LSQKAB, que pertence ao pacote CCP4. De acordo com CCP4 (2016), este *software* é baseado no algoritmo de Kabsch, um método que calcula a matriz de rotação ideal minimizando o desvio dos mínimos quadrados. O LSQKAB permite a realização de sobreposições atômicas de um único átomo ou de um trecho específico de átomos pertencentes a uma proteína no formato do PDB. O resultado da sobreposição pode ser escrito em um arquivo denominado de “deltas”, que contém uma lista com as diferenças de distâncias atômicas, entre todos os átomos comparados.

Além do LSQKAB, outros *softwares* que também utilizam o algoritmo de Kabsch são frequentemente utilizados na área de bioinformática, principalmente por cristalógrafos (físicos que resolvem estruturas das proteínas). Estes programas de computador, inclusive o LSQKAB, trabalham em três etapas para encontrar a matriz de rotação ideal entre dois trechos de proteínas. A primeira delas consiste na translação das coordenadas, a segunda é o cálculo da matriz de covariância e finalmente, é realizado o cálculo da matriz de rotação ideal. Na comparação entre poucos arquivos, o tempo de execução do LSQKAB é satisfatório. Entretanto, para comparar uma base de dados inteira, este processo torna-se demorado.

Uma das maneiras de utilizar o LSQKAB para comparar uma base de dados, é

como desenvolvido em Dias (2012). Inicialmente, foi escolhido um arquivo de referência. Em seguida, foi utilizado o LSQKAB para sobrepor todos os demais registros contra esta referência, gerando um arquivo de deltas com as diferenças das distâncias atômicas. Dias (2012) ainda calculou um valor de *score* para cada arquivo de delta gerado, assim, otimizando a busca.

Existem *softwares* que trabalham com arquivos do PDB, utilizando outras metodologias. Um exemplo é o *Cutoff Scanning Matrix* (CSM). Como citado por Pires et al. (2011), esta ferramenta gera assinaturas para grafos biológicos, com base no padrão de distância dos átomos de uma estrutura proteica. As assinaturas geradas são utilizadas em tarefas de classificação.

1.1 Justificativa

O LSQKAB é um *software* baseado no algoritmo de Kabsch, que retorna uma matriz de rotação ideal, com base no desvio dos mínimos quadrados entre dois conjuntos de proteínas. Este resultado é obtido através de uma série de cálculos computacionais. Este *software* é ideal para comparar poucos arquivos de uma forma precisa. Entretanto, este procedimento é demorado se a comparação for entre um arquivo e uma base de dados inteira, como a de 16.383 registros, utilizada nos experimentos.

O LSQKAB tem como entrada dois arquivos de texto com informações de proteínas, no formato do PDB. Ele gera como saída, arquivos de texto em diferentes formatos, um deles é o de “deltas”, que contém as diferenças de distâncias entre cada par de átomo comparado. Também pode-se gerar uma tabela de distâncias RMS entre pares de átomos para a cadeia principal e lateral de cada resíduo selecionado. Ainda é possível gerar um arquivo no formato do PDB, com as coordenadas que foram rotacionadas pela estrutura movimentada para o ajuste dos centróides. A constante leitura e escrita em disco, tornam ainda mais lenta a execução deste *software* para comparar uma base de dados grande.

Apesar de não ser indicado para a comparação de arquivos de uma base de dados inteira, o LSQKAB é utilizado por vários cristalógrafos e também por diversos *softwares*. Um exemplo é a ferramenta *Resídue Interaction Database* (RID), apresentada em Dias (2012). Neste *software*, o LSQKAB foi utilizado para encontrar pares de registros similares, com base nos arquivos de deltas resultantes das sobreposições atômicas de um arquivo escolhido como referência, contra os demais registros da base de dados.

Dias (2012) otimizou a busca criando um valor de *score* a partir do cálculo da distância euclidiana com as informações de distâncias em cada arquivo de delta gerado. Este procedimento diminuiu o tempo gasto nas comparações, mas a precisão não ficou ótima, pois arquivos com valores de distâncias atômicas diferentes podem obter o mesmo valor de *score*.

Visando o aumento da precisão na comparação entre estes arquivos, foram avaliados outros algoritmos como o Cutoff Scanning Matrix(CSM), apresentado em Pires et al. (2011). Este *software* gera assinaturas para grafos biológicos com base nos padrões de distâncias. Deste modo, foi decidido desenvolver uma metodologia baseada em matriz de distâncias, inspirada em uma das etapas do CSM.

A justificativa deste trabalho, consiste em desenvolver uma nova metodologia baseada em uma matriz de distâncias, para aperfeiçoar a busca de pares similares feita pela ferramenta RID, permitindo ainda o agrupamento dos registros conforme os seus valores de distâncias atômicas.

1.2 Motivação

Ao encontrar trechos similares de proteínas, existe a possibilidade de analisar e propor mutações entre os mesmos, o que pode trazer melhorias para as estruturas dessas macromoléculas.

De acordo com Dias (2012), estes estudos beneficiam diversos segmentos, tais como a saúde, no desenvolvimento de fármacos e vacinas. Na indústria, aprimorando enzimas digestivas que são utilizadas em vários processos. No meio ambiente, alterando as enzimas para a degradação de contaminantes, entre outras inúmeras áreas.

1.3 Objetivos

O objetivo geral deste trabalho é desenvolver uma metodologia baseada em uma matriz de distâncias, que possibilite o agrupamento de arquivos de interações extraídas do PDB.

Os objetivos específicos são:

1. Formar grupos com os arquivos de interações extraídas do PDB, conforme suas similaridades de distâncias atômicas;
2. Manter a precisão do LSQKAB;
3. Comparar a precisão desta nova técnica com o CSM e também com a utilização de um arquivo de referência, como foi feito na ferramenta RID;
4. A metodologia proposta deve realizar as comparações e possibilitar os agrupamentos em um tempo viável.

1.4 Organização do trabalho

O trabalho está organizado da seguinte maneira: o capítulo 2 contém os trabalhos relacionados com este projeto de pesquisa. O capítulo 3 mostra a fundamentação teórica,

explicando os principais itens para o bom entendimento deste trabalho, como as proteínas, os arquivos de interação, sobreposições atômicas, agrupamentos e grafos. O capítulo 4 exibe a metodologia de pesquisa, explicando o delineamento da pesquisa, a coleta e o tratamento dos dados e o cronograma. O item 5 contém o desenvolvimento do trabalho, bem como a execução dos experimentos, comparação com as outras metodologias e a análise dos resultados. O capítulo 6 mostra a conclusão, os trabalhos futuros e as considerações finais.

2 Trabalhos Relacionados

Kabsch (1976) desenvolveu o cálculo da matriz de rotação ideal entre duas estruturas de proteínas, que minimiza o RMSD (Root Mean Squared Derivation). No trabalho original, foram utilizados multiplicadores de Lagrange para provar esta solução.

Malischewski, Schumann e Hoffmann (2016) explica as principais partes do cálculo citado acima, através de técnicas mais simples de álgebra linear. Com a explicação voltada para a implementação desta solução em forma de algoritmo.

Conforme Malischewski, Schumann e Hoffmann (2016), a primeira etapa é a tradução, onde o centróide de uma estrutura é movimentado para o início do outro conjunto de dados, para que os centróides de ambas as estruturas se alinhem. Em seguida, é calculado a matriz de covariância, com a equação 1, onde P e Q são os conjuntos de dados:

$$A = P^T Q \tag{1}$$

Em seguida, na equação 2, é calculado o Singular Value Decomposition (SVD), da matriz de covariância obtida anteriormente. É necessário que a matriz “A” tenha “m” linhas e “n” colunas, com “m” > “n”.

$$VSW^T = A \tag{2}$$

Na equação 2, a matriz “A” pode ser escrita pela multiplicação das matrizes “V”, “S” e transposta da matriz “W”. Onde “V” possui “m” linhas e “m” colunas, “S” contém “m” linhas e “n” colunas e “W” possui “n” linhas e “n” colunas.

No próximo passo, através da equação 3, é verificado se a matriz de rotação necessita de ajustes. As matrizes “W” e “V” obtidas anteriormente, são utilizadas nesta etapa.

$$d = sign(det(WV^T)) \tag{3}$$

Por fim, a matriz de rotação ideal é calculada pela equação 4, com o valor "d"obtido no passo anterior sendo utilizado na fórmula.

$$U = W \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & d \end{bmatrix} V^T \tag{4}$$

Dentre os trabalhos que se destinam a encontrar similaridades entre trechos de proteínas, pode-se destacar os *softwares* desenvolvidos a partir deste algoritmo. Um exemplo é a ferramenta LSQKAB, que pertence ao pacote *Collaborative Computational Project No*. 4 (CCP4) e realiza sobreposições atômicas seguindo os passos mencionados acima.

De acordo com CCP4 (2016), através da ferramenta LSQKAB é possível realizar sobreposições atômicas de um único átomo ou de um trecho específico de átomos pertencentes a uma proteína que estão em arquivos texto, no formato do PDB. O subconjunto para as comparações pode ser escolhido com base no número de resíduos, cadeia lateral, principal ou ainda pela posição do átomo no arquivo.

Esta ferramenta possibilita a saída através de um arquivo denominado de “delta”, que contém as diferenças de distâncias para cada par de átomo comparado. Também pode-se gerar uma tabela de distâncias RMS entre pares de átomos para a cadeia principal e lateral de cada resíduo selecionado. E ainda é possível gerar na forma de um arquivo do PDB, as coordenadas que foram rotacionadas pela estrutura movimentada para o ajuste dos centróides.

Dentre os trabalhos que utilizaram o LSQKAB, pode-se citar, Dias (2012), que desenvolveu uma ferramenta conhecida como Residue Interaction Database (RID). Esta aplicação consiste em um algoritmo e uma nova base de dados que propõem mutações entre pares de resíduos de proteínas com estrutura tridimensional conhecida, visando aumentar a estabilidade da proteína, focando na conformação da cadeia principal. As mutações sugeridas mantêm o enovelamento da proteína alvo, para que ela continue com a sua função. Neste trabalho citado, o LSQKAB foi utilizado para sobrepor todos os arquivos de interações de proteínas contidos na base de dados do *software* RID contra um arquivo de referência. Esta referência foi escolhida de acordo com a resolução indicada no arquivo PDB.

De acordo com RSCB (2016b), a resolução é uma medida da qualidade dos dados coletados do cristal da proteína ou ácido nucléico. Ela é a medida do nível de detalhamento tanto para o experimento de difração quanto para a medição no mapa de densidade eletrônica. Neste último, uma estrutura com resolução alta (até 1 angstrom) permite uma melhor localização atômica.

Após sobrepor todos os arquivos da base de dados contra o arquivo de referência escolhido, Dias (2012) otimizou a busca criando um valor de *score* a partir da distância euclidiana de cada arquivo delta gerado. Este procedimento ocorreu para diminuir o custo da busca por pares similares que o LSQKAB gastaria para comparar todos os arquivos da base de dados. Os resultados obtidos foram bons, mas não ótimos.

Além do LSQKAB, existem outros *softwares* que comparam as estruturas proteicas baseando em valores de RMSD. De acordo com Kufareva e Abagyan (2012), esta é a forma

mais comum de comparação destas estruturas.

Um outro *software* que segue esta métrica de comparação é o *Superpose*. Conforme citado em Maiti et al. (2004), o *Superpose* é um servidor que realiza a sobreposição de duas ou mais estruturas de proteínas por vez. Este *software* permite o alinhamento em sequência, alinhamentos por estruturas, estatísticas do RMSD obtido, além de imagens interativas das estruturas sobrepostas.

Algumas outras técnicas, apresentam diferentes adaptações na verificação por RMSD. Um exemplo é o Zemla (2003), onde foi apresentado uma nova ferramenta para encontrar similaridades em estruturas tridimensionais de proteínas. Foi verificado além do RMSD de todos os carbonos alfas correspondentes, a similaridade entre regiões locais das estruturas comparadas.

Neste trabalho, além de comparar as estruturas proteicas, temos como objetivo agrupá-las, conforme as similaridades dos valores de distâncias. Na área de bioinformática, alguns autores desenvolveram trabalhos utilizando técnicas de agrupamentos. Um exemplo é Singh e Thornton (1991), que apresentaram um atlas de 400 interações possíveis referentes a cada um dos 20 tipos de aminoácidos de algumas estruturas de proteínas. Estas interações foram medidas através das distâncias e ângulos dos resíduos interagentes.

Kawaji, Takenaka e Matsuda (2004) também utilizou uma técnica de *clustering*, mas adicionada de uma estrutura baseada em grafos. O objetivo dos autores foi encontrar relacionamentos distantes entre proteínas, onde apenas uma região de suas sequências sejam similares. Como resultado, foi desenvolvido um algoritmo de agrupamento eficiente em tempo e espaço ocupado.

Pires et al. (2011) apresentou um novo modelo para geração de assinaturas para grafos biológicos, denominada *Cutoff Scanning Matrix* (CSM). Esta metodologia apresentada foi utilizada satisfatoriamente em diversos cenários, como anotação automática proteica, predição de ligantes e atividade de pequenas moléculas.

De acordo com Pires et al. (2011), a metodologia do CSM consiste em gerar vetores de atributos, que representam padrões de distâncias entre nós de um grafo, que são usados em tarefas de classificação. A geração da matriz CSM é dada do seguinte modo: para cada proteína do conjunto de dados, é gerado um vetor de atributos. Inicialmente é calculado a distância Euclidiana entre todos os pares dos centróides que representam os resíduos (Carbono Alfa ou Beta). Estes valores são inseridos em uma matriz de distâncias. Foi definido um intervalo de distâncias (*cutoffs*) a ser considerado (de 0,0 até 30,0 angstrons) e um passo (0,2 angstrom). As distâncias geradas, contidas nestes intervalos são avaliadas e a frequência de pares de resíduos dentro do intervalo definido são computadas. Deste modo, foi gerado para cada proteína, um vetor com 151 valores. Unidos, esses vetores compõem a matriz CSM, onde cada linha da matriz representa uma proteína, e cada coluna

indica a frequência de pares de resíduos a uma distância menor ou igual a cada passo do valor de corte. Ainda foi utilizado o *Singular Value Decomposition* (SVD), para reduzir a dimensionalidade e os ruídos dos dados, assim diminuindo o gasto com processamento.

Após a apresentação do CSM, foram desenvolvidos novos trabalhos baseados neste algoritmo, como em Pires et al. (2013) e Pires e Ascher (2016). No primeiro trabalho, foi desenvolvido o *atomic Cutoff Scanning Matrix* (aCSM), que é uma versão atômica do CSM e suporta assinatura de *pockets* baseada em grafos. O outro trabalho, trata-se do CSM-lig, um servidor web para avaliação e comparação de proteínas e pequenas moléculas, onde a modelagem foi definida pelo CSM. Os átomos foram representados como nós e suas interações como arestas. Foram extraídos padrões de distâncias entre os componentes.

3 Fundamentação Teórica

Nos seguintes tópicos são descritos os principais itens para o bom entendimento do trabalho.

3.1 Proteínas

De acordo com Nelson, Lehninger e Cox (2008), as proteínas são as macromoléculas mais abundantes nas células vivas, ocorrendo em todas as células e em diversas variedades. Elas desempenham inúmeras tarefas para os seres vivos, fornecendo materiais para construção e manutenção de órgãos e tecidos, diferenciação celular, transporte, entre outras diversas finalidades.

As proteínas são formadas por um conjunto de 20 aminoácidos distintos, ligados em sequências lineares. Os aminoácidos são compostos por uma cadeia principal de átomos (nitrogênio, carbono alfa, carbono e oxigênio), que efetuam a ligação peptídica com a cadeia principal de outro aminoácido. Eles ainda possuem uma cadeia lateral ligada ao carbono alfa, iniciando no carbono beta, exceto para a Glicina que possui apenas um átomo de hidrogênio em sua cadeia lateral.

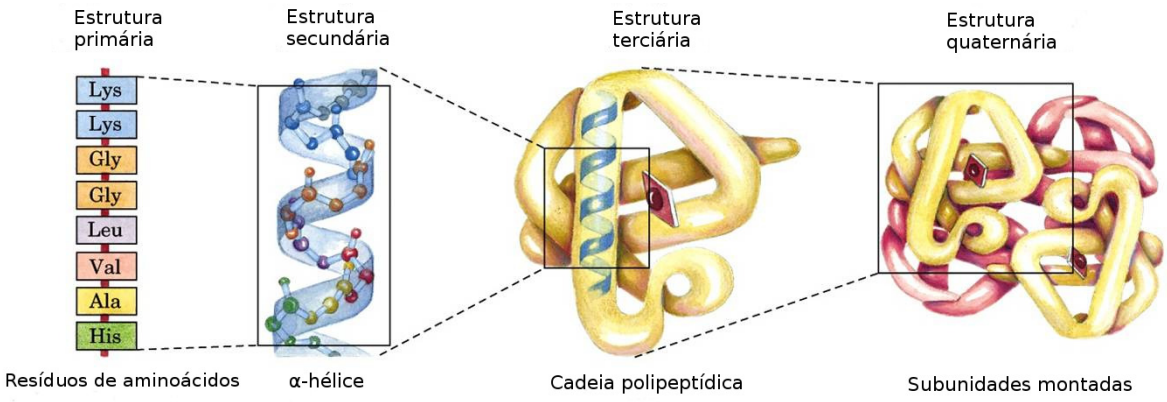
Nelson, Lehninger e Cox (2008), afirmam que existem quatro níveis de estrutura protéica, sendo elas: primária, secundária, terciária e quaternária. Elas são ilustradas na Figura 1.

A estrutura primária é formada por uma sequência de aminoácidos unidos entre si por ligações peptídicas. A estrutura secundária é dada pela forma em que os aminoácidos se interagem, formando ligações intermoleculares. Com estas interações, é criada uma conformação espacial de polipeptídeo. A estrutura terciária consiste na conformação tridimensional completa de uma cadeia polipeptídica. A estrutura quaternária é referente a localização espacial de múltiplas cadeias polipeptídicas, associadas de uma maneira estável.

3.2 Arquivos de Interações Extraídas do PDB

O *Protein Data Bank* (PDB) é um banco de dados biológico que fornece suporte para os estudos das proteínas. De acordo com RSCB (2016a) , o PDB armazena coordenadas atômicas, ligações, distâncias, composição, função e outras particularidades importantes das proteínas. Ele mantém em sua estrutura, arquivos no formato texto, contendo várias

Figura 1 – Níveis de estrutura de uma proteína



Fonte: Nelson, Lehninger e Cox (2008)

informações obtidas e registradas por pesquisadores, que permitem a visualização da estrutura da proteína pelas suas coordenadas atômicas.

A Figura 2 ilustra a estrutura tridimensional completa de um receptor de acetilcolina contido no PDB. Este polipeptídeo foi representado no PDB pelo arquivo "PDB 1PEN"[1].

A partir do exemplo de estrutura tridimensional completa ilustrada pela Figura 2, podemos destacar as interações dissulfeto. As pontes dissulfeto são ligações covalentes formadas pela interação entre os átomos de enxofre das cisteínas, que após serem oxidadas se tornam cistinas (átomos de cor verde) [2]. A Figura 3 mostra detalhadamente a estrutura de uma ponte dissulfeto.

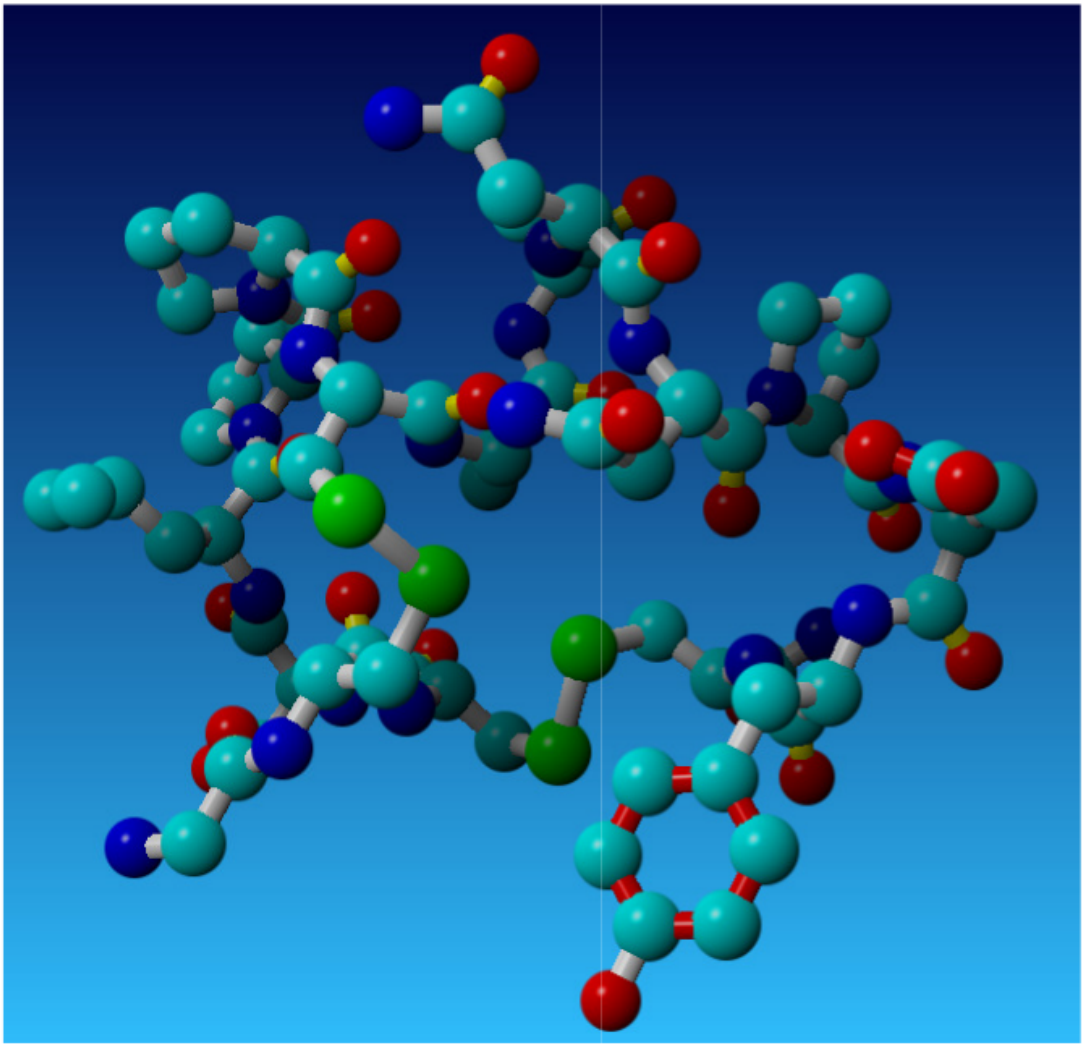
Através da Figura 3, podem ser visualizados os átomos de Enxofre Gama (SG) das cisteínas que se ligam através dos átomos de Carbono Beta (CB), formando uma ponte dissulfeto. Com esta ligação, as cadeias de Nitrogênio, Carbono Alfa, Carbono e Oxigênio de cada lado são estabilizadas. Esta cadeia é chamada de cadeia principal, por conter a estrutura principal do aminoácido que se liga a outro.

Neste trabalho, os experimentos foram realizados com interações de pontes dissulfeto, extraídas do PDB no formato de arquivos de texto pela ferramenta RID. Foram extraídas a cadeia principal dos resíduos interagentes completa e apenas o último átomo do resíduo anterior e o primeiro átomo do resíduo posterior. A cadeia lateral foi ignorada, pois a manutenção da cadeia principal implica na manutenção da função da proteína, no caso de uma mutação. A Figura 4 mostra a estrutura dos arquivos que foram extraídos. Como exemplo, foi ilustrado o arquivo “1ejg_CYS-3-A_CYS-40-A_mc6.pdb”.

[1] PDB ID 1PEN. Hu, S.H., Gehrmann, J., Guddat, L.W., Alewood, P.F., Craik, D.J., Martin, J.L. The 1.1. Å crystal structure of the neuronal acetylcholine receptor antagonist, alpha-conotoxin alpha conotoxin PnIA from Conus pennaceus, PubMed ed 8740364, Structure. 1996 Apr 15;4(4):417-23

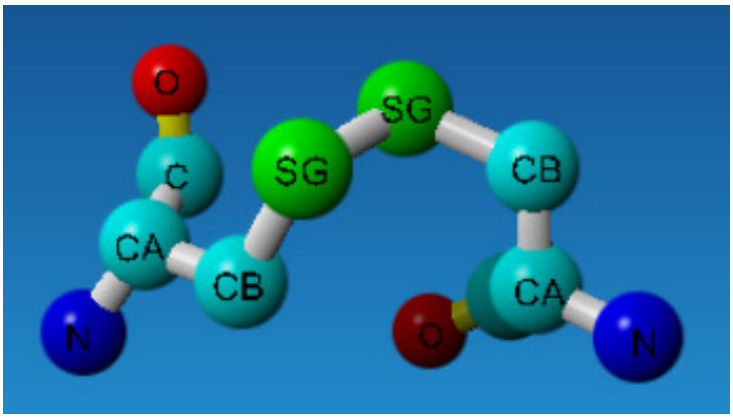
[2] Em um processo de oxidação, as cisteínas perdem elétrons e se transformam em cistinas.

Figura 2 – Ilustração do arquivo de interação



Fonte: Dias (2012)

Figura 3 – Ilustração do arquivo de interação



Fonte: Dias (2012)

Figura 4 – Exemplo de um Arquivo de Interação

```
CRYST1   40.824   18.498   22.371  90.00  90.47  90.00 P 1 21 1      2
SCALE1      0.024495  0.000000  0.000201        0.00000
SCALE2      0.000000  0.054060  0.000000        0.00000
SCALE3      0.000000  0.000000  0.044702        0.00000
ATOM    1  C   THR A   2      14.172  10.757   7.221  1.00  2.52         C
ATOM    2  N   CYS A   3      13.438  11.192   8.264  1.00  2.26         N
ATOM    3  CA  CYS A   3      13.607  10.675   9.600  1.00  2.06         C
ATOM    4  C   CYS A   3      12.210  10.413  10.164  1.00  1.94         C
ATOM    5  O   CYS A   3      11.324  11.246   9.996  1.00  2.78         O
ATOM    6  N   CYS A   4      12.015   9.277  10.850  1.00  1.74         N
ATOM    7  C   THR A  39      20.535  13.026  11.330  1.00  4.95         C
ATOM    8  N   CYS A  40      19.748  11.982  11.477  1.00  4.08         N
ATOM    9  CA  CYS A  40      18.460  12.120  12.139  1.00  3.64         C
ATOM   10  C   CYS A  40      18.660  12.202  13.669  1.00  3.83         C
ATOM   11  O   CYS A  40      19.515  11.523  14.227  1.00  4.96         O
ATOM   12  N   PRO A  41      17.847  13.009  14.329  1.00  4.47         N
END
```

A primeira linha do arquivo é referente a descrição da célula cristalográfica utilizada para extrair diversas informações da estrutura atômica, sendo identificada por CRYST1. As linhas 2, 3 e 4, indicados por SCALEn mostram os operadores para a transformação de coordenadas ortogonais em coordenadas da célula cristalográfica. Abaixo dessas 4 primeiras linhas, é iniciada a seção de coordenadas. Esta seção é uma das mais importantes do arquivo. A partir de informações contidas em PDB (2010), foi possível construir a tabela 1 para explicar detalhadamente os itens desta seção.

Como pode ser observado pela tabela da seção de coordenadas, a primeira coluna refere-se ao tipo de registro contido naquela linha. A segunda coluna indica o número de série do átomo. A terceira mostra o símbolo do átomo, sendo que carbono (C), nitrogênio (N), carbono alfa (CA) e oxigênio (O). As próximas três colunas referem respectivamente ao nome do resíduo, cadeia e número de série. As colunas 7, 8 e 9 indicam as coordenadas dos átomos em angstroms quanto aos eixos x, y e z. As duas colunas seguintes indicam respectivamente a probabilidade do átomo estar naquela localização e a medida de confidencialidade da localização do átomo. A última coluna mostra o símbolo do elemento que está alinhado à direita.

Tabela 1 – Seção de Coordenadas de um arquivo de Interação

Colunas	Tipo de Dados	Campo	Definição
1 - 6	Nome do registro	ATOM	
7 - 11	Inteiro	serial	Número de série do átomo
13 - 16	String	name	Nome do átomo
17	Caracter	altLoc	Indicador de localização alternativa
18 - 20	Nome do resíduo	resName	Nome do resíduo
22	Caracter	chainId	Identificação da cadeia
23 - 26	Inteiro	resSeq	Número de sequência do resíduo
27	Caracter	iCode	Código da inserção de resíduos
31 - 38	Real	x	Coordenada ortogonal para x em angstrons
39 - 46	Real	y	Coordenada ortogonal para y em angstrons
47 - 54	Real	z	Coordenada ortogonal para z em angstrons
55 - 60	Real	occupancy	Probabilidade do átomo estar naquele local
61 - 66	Real	tempFactor	Fator de Temperatura
77 - 78	String	element	Símbolo do elemento alinhado à direita
79 - 80	String	charge	Carga

3.3 Sobreposições Atômicas de Trechos de Proteínas

De acordo com Lodish et al. (2014), a estrutura tridimensional de uma proteína está ligada com a sua função, ou seja, qualquer alteração em sua estrutura nativa pode alterar o seu funcionamento. Mitchell et al. (2006) complementa que substituições de aminoácidos das proteínas, através de mutações também podem impactar na função da proteína. Se os aminoácidos trocados forem semelhantes, ou seja, possuírem tamanho, carga e outras características iguais, os impactos da mutação podem não ser altos, mas caso a substituição seja realizada por um outro aminoácido com estrutura significativamente diferente, esta mudança poderá resultar em ganho ou perda de função da proteína.

A sobreposição atômica, que é realizada por *softwares* como o LSQKAB, permite verificar a similaridade entre trechos de proteínas. Nestas comparações, os valores das distâncias atômicas são importantes para identificar similaridades entre os resíduos. Esta verificação pode indicar possíveis mutações entre as estruturas comparadas.

Conforme é citado por Kufareva e Abagyan (2012), dentre as métricas utilizadas para medir a similaridade entre proteínas, a mais utilizada é o *Root Mean Square Deviation* (RMSD). As finalidades mais comuns para o seu uso são: cálculos entre átomos de carbono alfa contidos nos dois trechos de proteínas; Cálculos entre todos os átomos de um subconjunto específico de resíduos; Cálculos entre todos os átomos de uma pequena molécula ou

ligante. A sua fórmula é dada por:

$$RMSD = \sqrt{\frac{1}{n}\sum_{i=1}^{n} d_i^2} \tag{5}$$

Como pode ser visto pela equação 5, através da fórmula do RMSD, é calculado o somatório das diferenças de distância entre cada par de átomos. O valor do somatório é dividido pela quantidade de átomos comparados. Por fim, é calculado a raiz quadrada. Com o valor do RMSD, é obtido o quanto uma estrutura de proteína se desviou da outra.

De acordo com Tsai (2003), um valor de RMSD de 0,0 até 3,0 angstrons indica uma forte similaridade estrutural. É comum utilizar um baixo valor de RMSD como limiar (0,5 angstrom, por exemplo).

Neste trabalho, as sobreposições consideradas como satisfatórias, foram as quais o maior valor das 12 distâncias atômicas do arquivo de delta fosse igual ou inferior à 0,5 angstrom. A escolha deste valor se deu por ser uma distância baixa. Deste modo, é garantido que o valor de RMSD será igual ou inferior a 0,5 angstrom e ainda é evitado a ocorrência de algum átomo mal sobreposto em comparações que a maioria dos outros átomos possuísse uma boa sobreposição.

3.4 Grafos

Conforme é citado por Gross e Yellen (2004), os grafos consistem em representações abstratas de entidades (indicados por vértices) e seus relacionamentos (arestas), que podem ter uma ou mais direções.

A utilização desta estrutura de dados é frequentemente utilizada na área de computação, e através dela pode-se representar sistemas no mundo real, tais como circuitos elétricos, moléculas orgânicas, estradas entre outros.

De acordo com Pires et al. (2011), o Cutoff Scanning Matrix (CSM) é um modelo para geração de assinaturas para grafos biológicos. O CSM gera vetores de atributos que representam padrões de distâncias entre nós de um grafo, que são então usados como evidência em tarefas de classificação.

3.5 Matriz de Distâncias

Conforme citado por Pinho (2017), uma matriz é um tipo de dado usado para representar uma certa quantidade de variáveis do mesmo tipo. Estes valores são acessíveis por um único nome e são armazenados contiguamente na memória. Cada variável pode ser

acessada individualmente pelo uso de índices. A matriz pode ser unidimensional (também chamada de vetor), ou pode ter mais dimensões.

Neste trabalho, foi utilizado uma matriz unidimensional para armazenar o resultado do cálculo da distância euclidiana entre todos os átomos de um mesmo arquivo de interação. As informações de cada registro foram inseridas em um vetor de 66 posições. Posteriormente, foi formada uma matriz com 16.383 linhas e 66 colunas (cada linha referente a um registro), que foi agrupada pela técnica de *clustering* K-Means.

3.6 Agrupamentos

De acordo com Quilici-Gonzalez e Zampirolli (2015), a tarefa de *clustering* é muito utilizada em mineração de dados. Ela consiste em segmentar um grupo de registros em conjuntos menores, separando-os por atributos iguais ou características similares.

Conforme citado por MathWorks (2017), o objetivo da análise de *clusters* consiste em encontrar padrões ocultos ou grupos em uma base de dados. Os algoritmos de *clustering* agrupam os dados de um modo que os registros dentro de um certo grupo, possuem uma maior similaridade entre si, se comparados com os dados de qualquer outro grupo.

MathWorks (2017) ainda afirma que a análise de *cluster* é um método de aprendizagem sem supervisão e uma tarefa importante na análise de dados exploratórios. Dentre as principais medidas de similaridades utilizadas para formar os grupos estão a distância euclidiana e a distância probabilística. Algumas das técnicas de *clustring* mais comumente utilizadas são:

- *Hierarchical Clustering*: Constrói uma árvore de *clusters*. Esta árvore não é um conjunto único de *clusters*, trata-se de uma hierarquia multinível. Os *clusters* de um nível são unidos como agrupamentos no próximo nível.
- *K-Means Clustering*: Particiona os dados em “k” grupos distintos, baseados na distância para o centróide de cada *cluster*. O centróide é o ponto para o qual a soma das distâncias de todos os elementos desse cluster é a menor possível. O *k-means* utiliza um algoritmo iterativo que minimiza a soma das distâncias de cada item ao centróide do seu grupo. Ele move objetos entre *clusters*, até que a soma não possa ser mais diminuída.
- *Gaussian mixture models*: Modela os *clusters* como uma mistura de componentes de densidade normal multivariada.
- *Self-organizing maps*: Utiliza redes neurais que aprendem a topologia e a distribuição dos dados. Neste processo, os dados são agrupados por similaridade.

De um modo geral, os estudos que envolvem a manipulação de estruturas de proteínas trabalham com uma grande quantidade de registros. Dividir as estruturas protéicas

em grupos menores, de acordo com as suas semelhanças de distâncias atômicas, é uma tarefa de extrema importância.

Neste trabalho, foram feitos experimentos utilizando o algoritmo de *clustering k-means* para agrupar os valores em cada técnica avaliada. Além da precisão desta técnica, um outro motivo desta escolha é por ela permitir o uso da distância euclidiana para formação dos grupos, uma vez que todas as metodologias avaliadas utilizaram este cálculo em diferentes etapas.

4 Metodologia

Este capítulo aborda todos os passos que foram seguidos para realizar os experimentos. No item 4.1, são apresentados os *hardwares* e os *softwares* utilizados. A seção 4.2 contém as primeiras análises que foram realizadas. A seção 4.3 explica a metodologia CSM. O item 4.4 apresenta a metodologia desenvolvida neste trabalho. A seção 4.5 mostra a metodologia do arquivo de referência, utilizada pela ferramenta RID. O item 4.6 explica como os experimentos foram desenvolvidos. A seção 4.7 mostra como os dados foram coletados e tratados.

4.1 Hardwares e Softwares Utilizados

Os experimentos foram desenvolvidos em um notebook Intel (R) Core (TM) i3, com 4 GB RAM e sistema operacional Ubuntu 16. Os *softwares* utilizados foram o Matlab R2014a; o LSQKAB na versão 6.5; o *software Cutoff Scanning Matrix*(CSM), desenvolvido por Pires et al. (2011); a técnica de busca utilizando um arquivo de referência, desenvolvido por Dias (2012), através da ferramenta RID; além de scripts em Shell Script, executados via prompt de comando para manipular e organizar arquivos de texto.

4.2 Primeiras Análises

A primeira etapa do trabalho consistiu em entender a estrutura das proteínas e conceitos de sobreposições atômicas.

Em seguida, foram avaliadas as metodologias e ferramentas utilizadas para representação e comparação destas estruturas. Concluímos que o algoritmo de Kabsch é uma técnica frequentemente utilizada nesta área, principalmente por cristalógrafos (físicos que resolvem estrutura de proteínas). Existem ainda diversos *softwares* baseados neste algoritmo. Um exemplo é o LSQKAB, que é utilizado para verificar as similaridades entre arquivos de proteínas no formato do PDB, fornecendo um fácil uso e saídas detalhadas.

Após definir a metodologia e escolher o LSQKAB como um "*software* de referência", foi definido o critério para avaliar quando duas estruturas de proteínas são similares. A regra definida foi: todos os pares de átomos comparados devem ter no máximo 0,5 angstrom de diferença de distância atômica. Deste modo, seria garantido que o valor de RMSD da estrutura ficasse em até 0,5 angstrom e evitaria a ocorrência de átomos mal sobrepostos.

Ao realizar os primeiros experimentos, foi visto que o uso do LSQKAB é demorado na comparação entre vários arquivos de proteínas. Devido a este problema, foram avaliados

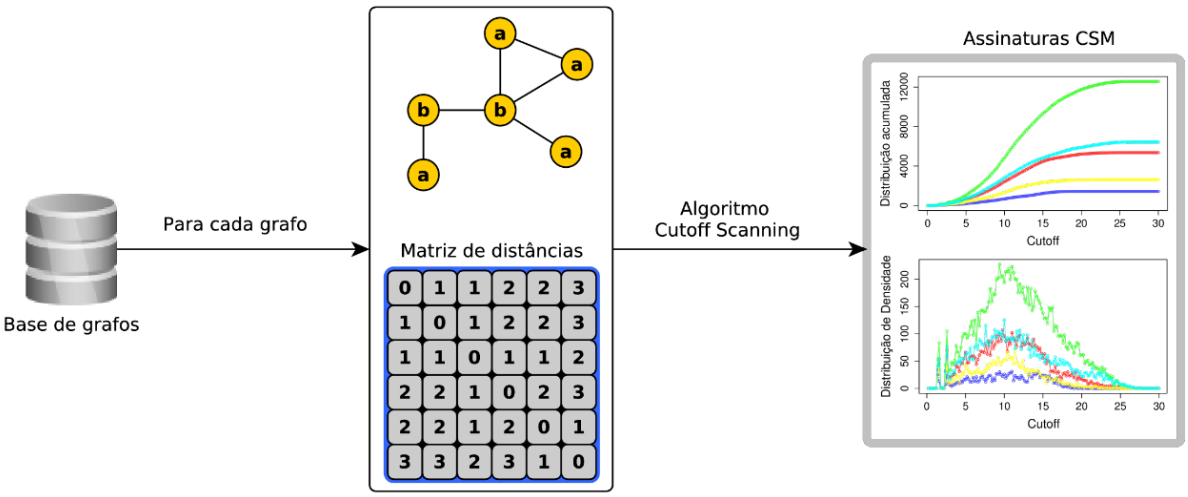
outros algoritmos, como o Cutoff Scanning Matrix (CSM), que gera assinatura para grafos biológicos com base nos padrões de distâncias. Deste modo, decidiu-se desenvolver uma metodologia baseada em uma matriz de distâncias, inspirada em uma das etapas do algoritmo CSM.

4.3 CSM

O *Software Cutoff Scanning Matrix* (CSM) foi apresentado em Pires et al. (2011). Ele gera assinaturas para grafos biológicos. Sua metodologia consiste em gerar vetores de atributos, que representam padrões de distâncias entre nós de um grafo, que são usados em tarefas de classificação.

De acordo com Pires et al. (2011), para construir a matriz CSM é necessário gerar um vetor de atributos para cada arquivo. Inicialmente é calculada a distância euclidiana entre todos os pares dos centróides escolhidos. Estes valores são inseridos em uma matriz de distâncias. Através do limite de corte previamente definido pelo usuário, com os parâmetros de distância mínima, distância máxima e o passo a ser incrementado, as distâncias contidas nestes intervalos são avaliadas pelo algoritmo CSM e a frequência de pares de resíduos dentro deste intervalo são computadas. Deste modo, é gerada a matriz CSM. A Figura 5 ilustra as etapas necessárias para gerar a matriz CSM.

Figura 5 – Fluxo para geração da matriz CSM



Fonte: Pires (2012)

Conforme a Figura 5, cada arquivo da base de dados refere-se a uma proteína que é representada em forma de grafo. Foi gerada uma matriz de distâncias para representar cada arquivo. A leitura desta matriz é feita do seguinte modo: o valor contido na primeira linha e segunda coluna, equivale à distância euclidiana do primeiro átomo de um arquivo para o segundo átomo desta mesma proteína. Deste modo, o valor da segunda linha e

primeira coluna será o mesmo, pois trata-se da distância euclidiana entre estes mesmos átomos. Assim, a matriz gerada é simétrica e a diagonal principal contém apenas valores "0". Após a geração da matriz, o algoritmo CSM obtém a frequência dos átomos escolhidos como centróides e constrói a matriz CSM, aonde cada linha refere-se a uma proteína e cada coluna é referente a frequência de cada átomo para uma determinada distância.

Pires et al. (2011) realizou diferentes experimentos tendo como centróides os carbonos alfas e betas. Em seu trabalho, a avaliação com os carbonos alfas apresentou um melhor resultado. Entretanto, neste trabalho, o objetivo não é apenas avaliar as distâncias entre os resíduos por um centróide específico, mas a análise deve-se levar em consideração todos os átomos contidos no arquivo de interação. Assim, os experimentos realizados, não tiveram apenas um tipo específico de átomo como centróide, foram computadas as frequências de distâncias entre todos os tipos de átomos. Como consequência, a matriz CSM gerada, ficou similar ao resultado do algoritmo aCSM (uma versão do algoritmo CSM baseada em todos os átomos).

Para trabalhar com todos os átomos como centróides, foi verificado que a maior distância euclidiana entre qualquer tipo de aminoácido foi de 11,22 angstrons. Este valor foi utilizado como o parâmetro de "distância máxima".

Após descobrir a maior distância entre os átomos, foi definido que cada matriz CSM gerada teria 151 colunas, como nos experimentos feitos por Pires et al. (2011). Deste modo, o valor do passo foi 0,074.

Após a definição dos parâmetros, a matriz de distâncias explicada anteriormente para representar cada arquivo, é avaliada pelo CSM e a frequência dos centróides é computada. A principal etapa do CSM é explicada pelo algoritmo 1.

Algoritmo 1: Algorítmo para o Cálculo de Assinaturas CSM

```
Entrada: BaseDeGrafos, TipoRotulo, DistMin, DistMax, DistPasso
Saida: CSM;
Function CSM
for all grafo i E BaseDeGrafos do
    j ← 0;
    distMatriz ← calculaDistancias(grafo)
    for dist ← DistMin ; to DistMax; passo DistPasso do
        for all tipo E (TipoRotulo) do
            CSM[i][j] ← obtemFrequencia(distMatriz, dist, tipo);
            j++;
        end
    end
end
return CSM;
```

Como pode ser visto pelo alogritmo 1, o cálculo do CSM tem como entrada as

proteínas que serão estudadas, o tipo do átomo, que neste caso foi desprezado, pois foram considerados todos os átomos, a distância mínima e máxima a ser avaliada, além do passo a ser incrementado. Dentro do primeiro *for*, é calculada a matriz de distância. Dentro dos outros dois laços *for*, a matriz de distâncias é utilizada para obter a frequência dos átomos para gerar a matriz CSM, conforme os valores de distância mínima, distância máxima e valor do passo.

A Figura 6, ilustra a saída do algoritmo CSM, ao comparar a frequência de distância considerando todos os átomos. Esta saída consiste na própria matriz CSM escrita em arquivo texto. Esta figura foi modificada para que as principais partes do arquivo gerado fosse explicado.

Figura 6 – Exemplo de um Arquivo CSM

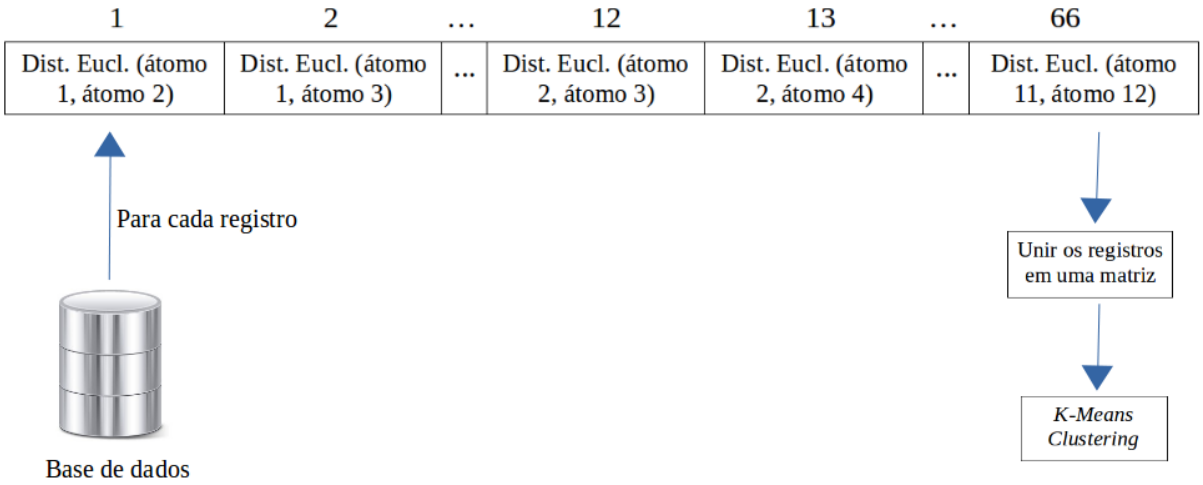
```
11bg_CYS-26-B_CYS-84-B_mc6.pdb,66 66 66,...,65 65 64 64 64 62 62 61,...
11bg_CYS-40-A_CYS-95-A_mc6.pdb,66 66 66,...,64 63 63 63 61 61 60 59,...
11bg_CYS-58-B_CYS-110-B_mc6.pdb,66 66 66,...,61 61 61 61 60 60 59 59,...
.
.
.
```

Conforme a Figura 6, a primeira informação contida em cada linha refere-se ao nome do arquivo de interação. Em seguida, cada valor é referente à frequência de átomos entre uma certa distância do arquivo. É importante ressaltar que a saída do algoritmo CSM é dada da maior frequência para a menor. Como os experimentos foram realizados considerando todos os átomos, pode-se notar que os primeiros valores dos três arquivos do exemplo é 66. Ou seja, todos os 66 possíveis valores de distâncias entre átomos estão em até 11,22 angstrons (valor da distância máxima). A coluna seguinte, indica a frequência dos átomos que estão a uma distância máxima de 11,22 subtraída pelo valor do passo (0,074), e assim sucessivamente. Os valores das distâncias que foram ilustrados correspondem às 3 primeiras informações de cada arquivo e em seguida dos dados contidos entre a posição 28 e 35 de cada arquivo. É possível observar que esta metodologia encontrou diferenças de padrões de distâncias entre os 3 arquivos utilizados como exemplo nesta figura. A matriz CSM gerada foi agrupada pela técnica *K-Means clustering*.

4.4 Metodologia Baseada em Matriz de Distâncias

A principal contribuição deste trabalho, deve-se ao desenvolvimento de uma metodologia para comparação de similaridades entre arquivos de interações extraídas do PDB. A metodologia proposta foi inspirada em uma das etapas do CSM, referente à construção da matriz de distâncias que é utilizada para obter as frequências dos átomos. A Figura 7 ilustra as etapas da metodologia proposta.

Figura 7 – Fluxo da metodologia proposta



Como pode ser observado pela Figura 7, para cada arquivo de interação extraída do PDB pela ferramenta RID, o mesmo foi representado por um vetor de 66 posições. O vetor se iniciou pela posição 1, pois ao contrário da maioria das outras linguagens de programação em que os vetores iniciam por 0, no Matlab o vetor tem o seu início pela posição 1. Cada posição é referente ao cálculo da distância euclidiana, para as coordenadas “x”, “y” e “z” em relação a cada um dos outros 11 átomos contidos em um mesmo registro. Posteriormente, estes vetores foram unidos em uma matriz.

Após a construção da matriz, a mesma foi agrupada pela técnica *K-means Clustering*, no *software* Matlab. O principal motivo da escolha desta técnica se deu pela utilização da distância euclidiana para formação dos grupos, através do parâmetro *’sqeuclidean’*. Cada agrupamento possui um centróide, que é o ponto para o qual a soma das distâncias de todos os elementos desse *cluster* é a menor possível.

Representar os arquivos desta maneira é interessante e viável para o problema em questão. Todos os 16.383 arquivos de interações avaliados possuem exatamente 12 átomos. Cada arquivo contém a cadeia principal dos resíduos interagentes completa e apenas o último átomo do resíduo anterior e o primeiro átomo do resíduo posterior. A cadeia lateral foi ignorada, com o intuito de manter a cadeia principal, visando a manutenção da função da proteína, no caso de uma mutação.

Devido ao fato dos arquivos estudados possuírem o mesmo tamanho, foi possível unir todos os vetores construídos em uma matriz de 16.383 linhas e 66 colunas. Cada linha representa um arquivo de interação e cada coluna contém os respectivos valores de cada registro.

4.5 Metodologia do Arquivo de Referência

Esta metodologia é baseada na busca de arquivos similares da ferramenta RID, desenvolvida por Dias (2012). A etapa inicial consistiu na escolha de um arquivo de interação para ser utilizado como referência. O critério utilizado foi a sua resolução indicada no arquivo PDB.

De acordo com RSCB (2016b), a resolução é uma medida da qualidade dos dados coletados do cristal da proteína ou ácido nucléico. Ela é a medida do nível de detalhamento tanto para o experimento de difração quanto para a medição no mapa de densidade eletrônica. Neste último, uma estrutura com resolução alta (até 1 angstrom) permite uma melhor localização atômica. Ao avaliar a resolução dos arquivos do PDB, foi verificado que o arquivo "pdb1ejg.ent" possui um alto valor de resolução. Desse modo, a referência escolhida foi uma interação desta proteína, o arquivo "1ejg_CYS-3-A_CYS-40-A_mc6.pdb".

Após esta escolha, foi utilizado o *software* LSQKAB, para realizar sobreposições atômicas de todos os outros 16.382 arquivos de interação contra a referência escolhida. Como resultado, foram gerados 16.382 arquivos deltas. Cada arquivo obtido contém uma lista com as 12 diferenças de distâncias entre cada par de átomo dos arquivos sobrepostos. A Figura 8 mostra a estrutura do "1ejg_CYS-3-A_CYS-40-A_mc6.pdb-1cbn_CYS-3-A_CYS-40-A_mc6.pdb.deltas", um dos arquivos gerados.

Figura 8 – Exemplo de um Arquivo Delta

```
0.015       2C   A              2C   A
0.017       3N   A              3N   A
0.033       3CA  A              3CA  A
0.031       3C   A              3C   A
0.032       3O   A              3O   A
0.023       4N   A              4N   A
0.031      39C   A             39C   A
0.013      40N   A             40N   A
0.029      40CA  A             40CA  A
0.031      40C   A             40C   A
0.047      40O   A             40O   A
0.026      41N   A             41N   A
```

O nome do arquivo delta é dado pela união dos nomes dos registros que o formou, separados por "-". Neste exemplo, os arquivos formadores foram o "1ejg_CYS-3-A_CYS-40-A_mc6.pdb" (arquivo utilizado como referência) e o "1cbn_CYS-3-A_CYS-40-A_mc6.pdb". A primeira coluna refere-se ao valor das diferenças das distâncias atômicas entre estes arquivos. As próximas duas colunas referem-se às informações do primeiro arquivo e as

duas últimas ao segundo arquivo. Estas informações são: número do resíduo da proteína, identificação do átomo e cadeia.

A informação mais importante de cada arquivo delta é a sua primeira coluna. Ela representa a diferença de distâncias entre os átomos de cada registro em relação aos mesmos átomos do arquivo de referência. Estes valores foram importados pelo Matlab e agrupados pela técnica *k-means Clustering*.

4.6 Realização dos Experimentos

Os experimentos foram realizados com 16.383 arquivos de interações de pontes dissulfeto, extraídas do PDB pela ferramenta RID.

A precisão do *software* apresentado neste trabalho foi comparada com a performance do CSM, cuja implementação de Pires et al. (2011), encontra-se disponível em Pires (2014). E também foi comparado com o método de busca utilizado na ferramenta RID.

A avaliação de cada uma das três metodologias foi feita por agrupamentos através do *k-means clustering*, definindo 500, 750 e 1.000 grupos. A precisão dos *clusters* formados foi verificada usando os resultados das sobreposições feitas com o LSQKAB. Foram realizadas sobreposições atômicas entre todos os registros que ficaram no mesmo grupo. Cada arquivo foi comparado duas vezes com todos os outros arquivos de seu grupo. Na primeira comparação, um destes registros se mantinha “fixo” e o outro se ajustava ao mesmo. Na segunda comparação, ocorria o inverso entre estes arquivos. Foi gerado um arquivo de deltas para cada comparação.

Cada registro gerado contém as informações das diferenças de distâncias entre cada átomo sobreposto. O valor de tolerância definido foi de no máximo 0,5 angstrom para cada par de aminoácido sobreposto. Deste modo, é garantido o valor de RMSD de no máximo 0,5 angstrom, o que garante uma forte similaridade entre os arquivos comparados, como citado em Tsai (2003).

Também verificou-se em todas as técnicas avaliadas, a média, a variância e o desvio padrão das distâncias dos registros em relação ao centróide de seu grupo. Ainda foi avaliado o tempo gasto por cada metodologia para construção dos registros e realização do agrupamento. Além do consumo de memória de cada metodologia para agrupar os registros.

4.7 Repetição dos Experimentos

O algoritmo *K-means Clustering* apresenta alguns passos aleatórios na formação dos primeiros grupos e escolha dos centróides iniciais. Deste modo, cada experimento

contou com 1.000 interações. Mesmo definindo um valor alto de interações, alguns registros ficaram em grupos diferentes de uma execução para a outra.

Deste modo, calculou-se quantas vezes seria necessário executar os experimentos de modo a garantir um alto nível de confiança e um baixo intervalo de confiança com base no cálculo do tamanho da amostra.

O cálculo do tamanho da amostra determina o número de observações ou repetições necessárias para estimar a variabilidade de um fenômeno em uma amostra estatística. Como os agrupamentos foram realizados considerando todos os 16.383 registros da base de dados, foi-se utilizada a seguinte fórmula:

$$n' = \frac{n}{1 + \frac{z^2 \hat{p}(1-\hat{p})}{e^2 N}} \tag{6}$$

Na equação 6, ”n” é o tamanho da amostra, “z” refere-se ao valor do *z-score*. Este valor está diretamente ligado com o nível de confiança. Para garantir uma boa precisão, o nível de confiança foi definido com 99% e consequentemente, o *z-score* foi de 2,58. O tamanho da população é representado por “N”, sendo 16.383. A proporção da população foi dada por “$\hat{p}''$ e teve o valor de 99%. O erro foi representado por “e”. O resultado é expresso por n'.

Ao realizar este cálculo com os parâmetros citados acima, foi constatado que 20 execuções de cada experimento são suficientes para obter um nível de confiança de 99%, com um intervalo de confiança de 6%.

4.8 Coleta e tratamento de dados

Para facilitar a manipulação dos 16.383 arquivos de interações, o nome de cada arquivo e as informações de suas respectivas coordenadas atômicas foram importadas para um único arquivo texto, através de comandos em Shell Script. A importação teve como referência o guia PDB (2010), onde é descrita toda a estrutura de um arquivo do PDB. Após a importação para um arquivo texto, foi verificado que nenhuma dentre as informações a serem utilizadas estavam incompletas.

Para realizar os experimentos com a nova metodologia, os dados foram importados para o Matlab. Em seguida, representados na forma de grafos e posteriormente agrupados pelo *K-Means Clustering*.

Para realizar os experimentos com o CSM, estes dados foram importados pelo código-fonte do CSM, implementado na linguagem Perl. A matriz CSM gerada foi importada para o Matlab e em seguida, agrupada pelo *K-Means Clustering*.

Os experimentos utilizando um arquivo como referência, teve como primeiro passo,

avaliar a resolução dos arquivos do PDB. Esta avaliação foi feita através de comandos em ShellScript. Após a escolha do arquivo com melhor resolução, foi utilizado o LSQKAB para sobrepor todos os registros contra esta referência e assim foram gerados arquivos de deltas. Foram realizados comandos em ShellScript para organizar estes registros gerados em um único arquivo texto, assim facilitando a importação dos dados para o Matlab. Em seguida, os dados importados foram agrupados pelo *K-Means Clustering*.

Em ambas as técnicas avaliadas, foram utilizados comandos em Shell Script para organizar os arquivos de interação conforme a formação dos grupos. O LSQKAB foi utilizado para gerar sobreposições atômicas entre todos os arquivos de um mesmo grupo, e assim verificar a precisão dos *clusters* formados por cada técnica. Foram utilizados comandos em Shell Script para ler os arquivos deltas gerados e verificar quantas sobreposições realizadas foram satisfatórias.

A avaliação da média, variância e desvio padrão das distâncias dos registros em relação ao centróide de seu grupo, e ainda a geração dos gráficos foi feita através da ide Matlab. A avaliação do tempo gasto consistiu em realizar as etapas de cada metodologia por 5 vezes e calcular a média do tempo de execução. A avaliação do uso de memória RAM se deu pelo monitoramento do momento de maior consumo por cada metodologia.

5 Análise e Discussão dos Resultados

Este capítulo apresenta as implementação da metodologia apresentada e os resultados obtidos da sua comparação com as outras metodologias. A seção 5.1 mostra a representação pela matriz de distâncias e o seu agrupamento. A seção 5.2 contém a comparação entre os métodos, sendo dividida em cinco subseções. A subseção 5.2.1 mostra a taxa de aproveitamento pelos *clusters* obtidos em cada metodologia. A subseção 5.2.2 contém a análise da quantidade de registros que estiveram em grupos com diferentes taxas de acertos. A subseção 5.2.3 contém uma análise dos dados através da regressão Linear, avaliando a relação entre a taxa de acerto dos grupos e a média de distância dos registros para o centróide de seu grupo. O item 5.2.4 analisa a média de distância em relação à variância. No item 5.2.5 é feita a análise da média de distância em relação ao desvio padrão. A subseção 5.3 mostra um comparativo do tempo de execução de cada metodologia. O item 5.4 mostra o uso de memória RAM para realizar os agrupamentos.

5.1 Agrupamentos Através da Matriz de Distâncias

A primeira etapa do desenvolvimento desta metodologia, consistiu em organizar os dados dos arquivos de interações a serem estudados. Deste modo, a identificação de cada arquivo, bem como os seus respectivos valores de coordenadas atômicas, foram organizadas em um único arquivo texto, através de comandos em Shell Script. Em seguida, estes dados foram importados pelo *software* desenvolvido, na IDE Matlab.

Após a importação dos dados, os mesmos foram convertidos para uma matriz de distâncias. Foi calculada a distância euclidiana entre as coordenadas atômicas “x”, “y” e “z” de todos os 12 átomos de um mesmo arquivo. Deste modo, foi obtido um vetor de 66 posições referente a cada arquivo. Ou seja, a primeira posição é referente a distância euclidiana entre o primeiro e o segundo átomo, a segunda posição contém o valor do cálculo do primeiro átomo com o terceiro, e assim sucessivamente até ser obtido o cálculo entre todos os átomos.

Como exemplo do funcionamento desta metodologia, o arquivo de interação “1ejg_CYS-3-A_CYS-40-A_mc6.pdb”, ilustrado pela Figura 4 no capítulo de fundamentação teórica, teve a saída ilustrada pela Figura 9:

A Figura 9 ilustra o vetor construído. A posição 1 indica a distância euclidiana das coordenadas “x”, “y” e “z” do primeiro átomo do arquivo “1ejg_CYS-3-A_CYS-40-A_mc6.pdb” (14.172, 10.757 e 7.221 respectivamente) para o segundo átomo deste mesmo arquivo (13.438, 11.192 e 8.264 respectivamente). As outras 65 posições são referentes ao cálculo

Figura 9 – Vetor formado pela matriz de distâncias a partir do arquivo “1ejg_CYS-3-A_CYS-40-A_mc6.pdb”

```
Pos. 1: 1.3475   Pos. 2: 2.4465   Pos. 3: 3.5537   Pos. 4: 4.0064
Pos. 5: 4.4736   Pos. 6: 7.9070   Pos. 7: 7.1208   Pos. 8: 6.6657
Pos. 9: 7.9879   Pos. 10: 8.8441  Pos. 11: 8.3127  Pos. 12: 1.4425
Pos. 13: 2.3927  Pos. 14: 2.7334  Pos. 15: 3.5185  Pos. 16: 7.9455
Pos. 17: 7.1249  Pos. 18: 6.4107  Pos. 19: 7.5831  Pos. 20: 8.5204
Pos. 21: 7.7152  Pos. 22: 1.5292  Pos. 23: 2.3864  Pos. 24: 2.4600
Pos. 25: 7.5178  Pos. 26: 6.5531  Pos. 27: 5.6645  Pos. 28: 6.6649
Pos. 29: 7.5520  Pos. 30: 6.7667  Pos. 31: 1.2276  Pos. 32: 1.3413
Pos. 33: 8.8030  Pos. 34: 7.8107  Pos. 35: 6.7733  Pos. 36: 7.5557
Pos. 37: 8.4323  Pos. 38: 7.4741  Pos. 39: 2.2547  Pos. 40: 9.4758
Pos. 41: 8.5848  Pos. 42: 7.5019  Pos. 43: 8.2596  Pos. 44: 9.2234
Pos. 45: 8.0270  Pos. 46: 9.3207  Pos. 47: 8.2164  Pos. 48: 7.1612
Pos. 49: 7.7884  Pos. 50: 8.5264  Pos. 51: 7.7488  Pos. 52: 1.3156
Pos. 53: 2.4044  Pos. 54: 3.1089  Pos. 55: 3.4194  Pos. 56: 4.0274
Pos. 57: 1.4547  Pos. 58: 2.4570  Pos. 59: 2.7978  Pos. 60: 3.5780
Pos. 61: 1.5452  Pos. 62: 2.4144  Pos. 63: 2.4418  Pos. 64: 1.2261
Pos. 65: 1.3221  Pos. 66: 2.2363
```

da distância euclidiana entre todos os outros átomos do mesmo arquivo.

Posteriormente, estes vetores foram unidos e formaram uma matriz de 16.383 linhas e 66 colunas. Cada linha é referente a um arquivo e cada coluna refere-se ao cálculo da distância euclidiana entre os átomos deste registro.

Ao término da construção desta matriz baseada na distância euclidiana, a mesma foi agrupada pela técnica *k-means Clustering*. O agrupamento foi realizado dividindo os registros em 500, 750 e 1.000 *clusters*.

Do mesmo modo, as saídas geradas pelo CSM e pela técnica do arquivo de referência, também foram agrupadas pelo *k-means Clustering*. O item a seguir, mostra o resultado das comparações entre as 3 técnicas, seguindo os critérios mencionados no capítulo de metodologia.

5.2 Comparação Entre as Metodologias

As comparações entre as metodologias, após as sobreposições feitas pelo LSQKAB, ocorreram de diferentes maneiras. Foi verificado qual técnica formou mais grupos com boas taxas de acertos; Qual técnica contou com uma maior quantidade de registros dentro dos melhores grupos; E ainda, foi verificado a média de distância entre os elementos, foi

calculada a regressão linear, variância e desvio padrão, para analisar os grupos que foram gerados por ambas as técnicas.

5.2.1 Comparação pela Taxa de Aproveitamento dos Clusters

Esta primeira avaliação, verificou a taxa de aproveitamento obtida em cada grupo, após as sobreposições realizadas com o LSQKAB. Como a técnica *K-Means Clustering* utiliza de alguns passos aleatórios para encontrar o melhor centróide e mover registros entre grupos, os experimentos foram repetidos por 20 vezes. Foi calculado a média dos resultados obtidos por cada técnica para uma melhor análise dos resultados. Conforme o cálculo do tamanho da amostra, a quantidade de vezes que os experimentos foram repetidos possibilitam um nível de confiança de 99% e um intervalo de confiança de 6%. A tabela 2, indica os resultados obtidos pela média de *clusters* por taxa de aproveitamento.

Tabela 2 – Média de Clusters por Taxa de Aproveitamento

Técnica		Mat. Dist.			Referência			CSM		
	Nº de clusters	500	750	1.000	500	750	1.000	500	750	1.000
Taxa de Acerto em cada Cluster	0% até 10%	29	40	56	25	31	59	92	107	115
	10% até 20%	67	77	79	57	74	70	84	113	150
	20% até 30%	60	61	77	73	68	75	52	78	81
	30% até 40%	51	90	97	45	76	96	37	63	84
	40% até 50%	36	53	69	43	63	64	25	40	64
	50% até 60%	42	47	59	55	52	67	26	44	52
	60% até 70%	39	49	67	25	50	54	29	27	48
	70% até 80%	30	51	57	33	44	54	31	50	56
	80% até 90%	52	62	70	41	57	73	52	74	84
	90% até 100%	94	220	369	103	235	388	72	154	266

A tabela de aproveitamento por *clusters* contém na primeira coluna as faixas de acertos consideradas. Em seguida, é exibido quantos grupos estiveram nestas faixas de acertos da pela Matriz de Distâncias, para os experimentos com 500, 750 e 1.000 grupos. Do mesmo modo, as próximas colunas mostram os resultados pela técnica utilizando um arquivo como referência e pelo CSM, respectivamente. Estes valores foram baseados na média obtida após várias execuções.

Como pode ser visto na tabela, para os experimentos com 500 grupos, a metodologia que utilizou um arquivo como referência apresentou mais grupos com a taxa de acerto superior a 90%, com 103 *clusters*. Destes, 31 tiveram 100% de acertos. A técnica da matriz de distâncias, obteve 94 grupos com o aproveitamento superior à 90%, sendo que 31

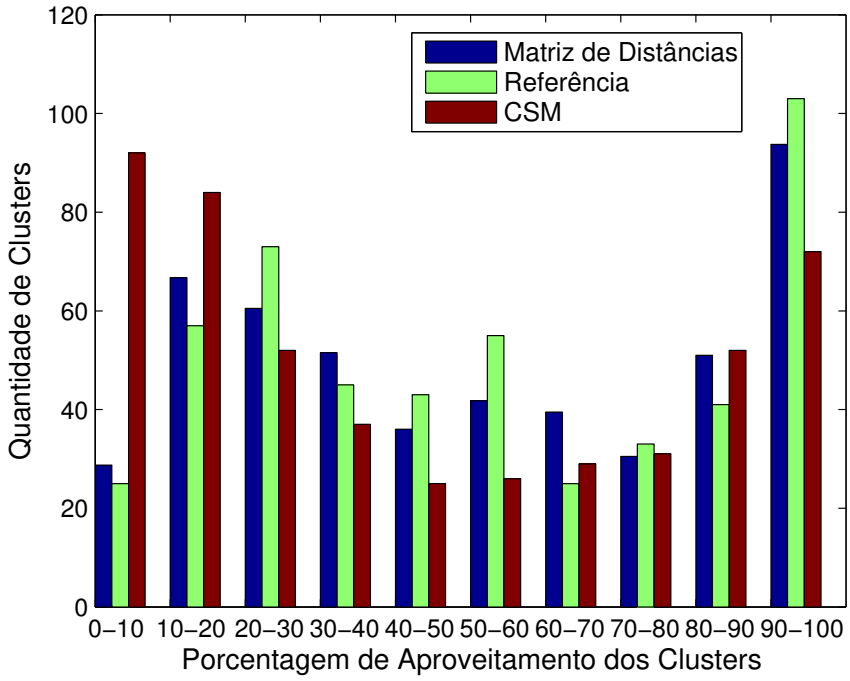
tiveram 100%. A metodologia CSM conseguiu 72 grupos com o mais de 90% de acertos e 32 destes possuíram 100% de acertos.

Avaliando a taxa de acerto entre 80% e 100%, ainda considerando 500 *clusters*, a representação pela matriz de distâncias obteve uma maior quantidade de grupos, foram 145. A técnica do arquivo de referência contou com 144 agrupamentos e o CSM obteve 124.

Verificando o outro extremo na análise por 500 grupos, a metodologia CSM obteve mais agrupamentos nas duas piores taxas de acertos. Foram 92 *clusters* com o aproveitamento inferior à 10% e 84 grupos com a taxa entre 10% e 20%. A técnica da matriz de distâncias obteve dentro destas faixas 29 e 67 grupos respectivamente. A técnica do arquivo de referência obteve 25 e 57 para as duas piores faixas, respectivamente.

A Figura 10 ilustra mais claramente os resultados de cada metodologia, considerando 500 grupos.

Figura 10 – Análise dos 500 Grupos Formados em Relação à porcentagem de Acertos



A Figura 10 contém no eixo “x” as faixas de aproveitamento. No eixo “y” é dado a quantidade de grupos que ficaram entre cada taxa de acerto. Os resultados obtidos pela metodologia apresentada, são representados pela cor azul escuro. A técnica do arquivo de referência é representada pela cor verde claro. A cor vermelho grená, representa a técnica CSM.

Como pode ser observado, a técnica do arquivo de referência e a representação proposta, obtiveram os melhores resultados. Principalmente na quantidade de grupos com a taxa de acerto superior à 90%. E ainda, estas metodologias obtiveram menos grupos nas

duas piores faixas de acertos do que o CSM. A técnica CSM obteve muitos grupos com pouco aproveitamento, apresentando até mais agrupamentos nas piores faixas de acertos do que nas melhores.

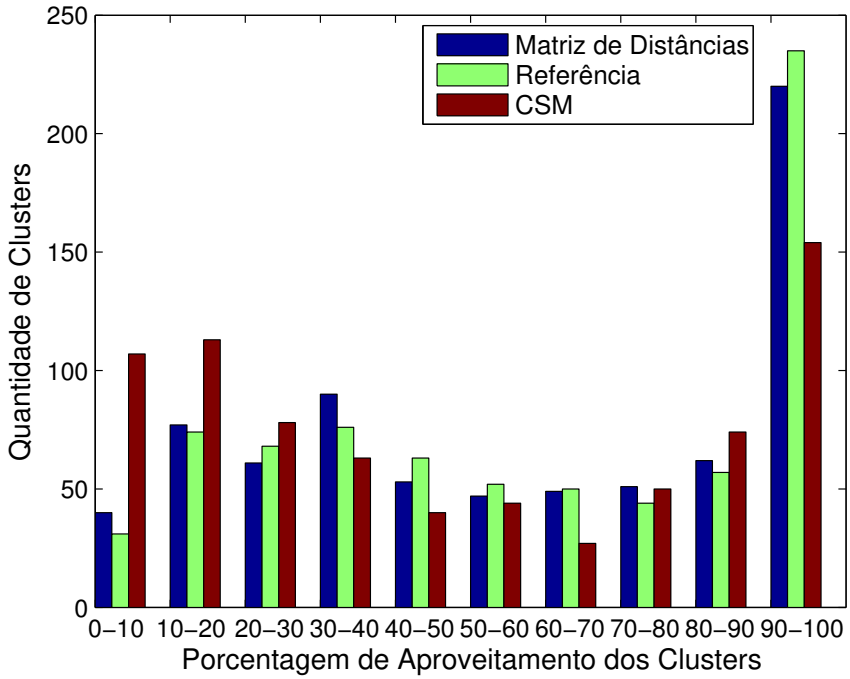
Analisando os resultados por 750 grupos da tabela 2, novamente a representação considerando um arquivo de referência conseguiu obter mais grupos na faixa de acertos acima de 90%. Foram 235 grupos, sendo que 133 tiveram 100% de acertos. A técnica da matriz de distâncias formou 220 agrupamentos com mais de 90% de acertos, sendo 115 com 100% de satisfação. O CSM formou 154 grupos com mais de 90% de acertos, destes 105 tiveram 100% de aproveitamento.

Avaliando a taxa de aproveitamento entre 80% e 100%, a técnica do arquivo de referência conseguiu 292 grupos. A matriz de distâncias e o CSM conseguiram 282 e 228 respectivamente.

Avaliando o outro extremo, ainda considerando 750 grupos, novamente o CSM apresentou os piores resultados, 230 grupos com até 20% de acertos. Entretanto, o CSM aumentou consideravelmente a quantidade de grupos satisfatórios.

Através da Figura 11 é possível observar mais claramente a análise por 750 grupos.

Figura 11 – Análise dos 750 Grupos Formados em Relação à porcentagem de Acertos



Como pode ser observado pela Figura 11, ao definir mais grupos, todas as três técnicas conseguiram melhorar consideravelmente os seus resultados. A quantidade de grupos com mais de 90% de acertos mais que dobrou para todas as técnicas em relação aos experimentos com 500 grupos.

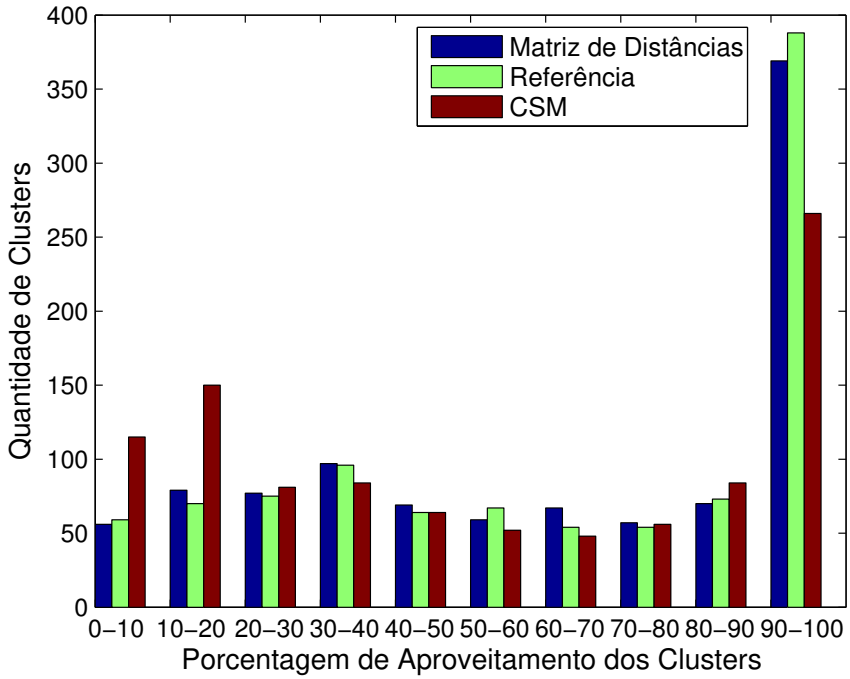
Analisando os resultados por 1.000 grupos da tabela 2, novamente a representação considerando um arquivo de referência conseguiu obter mais grupos na faixa de acertos acima de 90%. Foram 388 grupos, sendo que 258 tiveram 100% de acertos. A técnica da matriz de distâncias obteve 369 grupos com mais de 90% de satisfação, sendo 226 com 100% de acertos. O CSM obteve 266 grupos com mais de 90% de precisão, destes 187 com 100%.

Avaliando os grupos que tiveram entre 80% e 100% de todas as técnicas, temos a técnica do arquivo de referência com 461 grupos, a metodologia da matriz de distâncias com 439 agrupamentos e o CSM com 350.

O CSM ainda continuou sendo a técnica que teve mais grupos nas piores faixas de acerto. Entretanto, com 1.000 grupos formados, a quantidade de agrupamentos satisfatórios foi superior.

A representação gráfica da análise por 1.000 *clusters* pode ser vista pela Figura 12

Figura 12 – Análise dos 1000 Grupos Formados em Relação à porcentagem de Acertos



Como pode ser observado pela Figura 12, todas as técnicas aumentaram consideravelmente a quantidade de grupos com mais de 90% de acertos. Em todas as técnicas, o aumento desta faixa de acerto foi superior do que as outras faixas.

5.2.2 Comparação pela Quantidade de Arquivos nos Clusters

O objetivo desta análise é verificar quantos registros estiveram em grupos que obtiveram determinadas taxas de acertos. Deste modo, é possível concluir qual das metodologias

avaliadas formou mais grupos com uma maior quantidade de arquivos similares. Do mesmo modo que a análise anterior, este experimento também foi repetido por 20 vezes, garantindo um nível de confiança de 99% e um intervalo de confiança de 6%. A média obtida pelos experimentos foi utilizada para realizar uma melhor análise.

A tabela de quantidade de arquivos por aproveitamento dos *clusters*, mostra na sua primeira coluna as faixas de porcentagem definidas. A próxima divisão é referente à técnica da matriz de distâncias, contendo a quantidade de registros que estiveram em grupos com a taxa de acerto indicada pela primeira coluna, considerando 500, 750 e 1.000 grupos. Seguindo este mesmo padrão, a terceira e a quarta divisão contém os valores obtidos pela metodologia do arquivo de referência e pelo CSM, respectivamente.

Tabela 3 – Quantidade de Arquivos por Taxa de Aproveitamento dos Clusters

Técnica		Mat. Dist.			Referência			CSM		
	Nº de Clusters	500	750	1.000	500	750	1.000	500	750	1.000
Taxa de Acerto em cada Cluster	0% até 10%	464	333	276	548	278	357	2.695	2.041	1.573
	10% até 20%	1.456	1.099	760	1.257	1.117	728	2.424	2.071	2.144
	20% até 30%	1.422	1.009	959	1.658	1.060	1.026	1.550	1.499	1.110
	30% até 40%	1.072	1.318	1.067	1.172	1.302	1.146	1.105	1.303	1.283
	40% até 50%	1.024	925	889	1.224	1.290	993	705	744	1.010
	50% até 60%	1.389	997	909	1.807	1.059	1.020	877	868	749
	60% até 70%	1.325	1.028	1.074	717	1.012	835	943	707	776
	70% até 80%	1.119	1.302	968	1.154	1.121	897	1.128	1.198	938
	80% até 90%	2.282	1.597	1436	1.820	1.604	1.480	2.235	2.067	1.724
	90% até 100%	4.830	6.775	8.045	5.025	6.539	7.890	2.721	3.885	5.076

Como pode ser observado pela tabela 3, nos experimentos considerando 500 *clusters*, a metodologia que utiliza um arquivo como referência obteve 5.025 registros em grupos que obtiveram mais de 90% de acertos, sendo 989 em grupos com 100% de satisfação. Ainda foram contabilizados 1.820 arquivos na faixa de acerto entre 80% e 90%. Considerando todos os registros que estiveram em agrupamentos com a taxa de acerto entre 70% e 100% foram 7.999 arquivos.

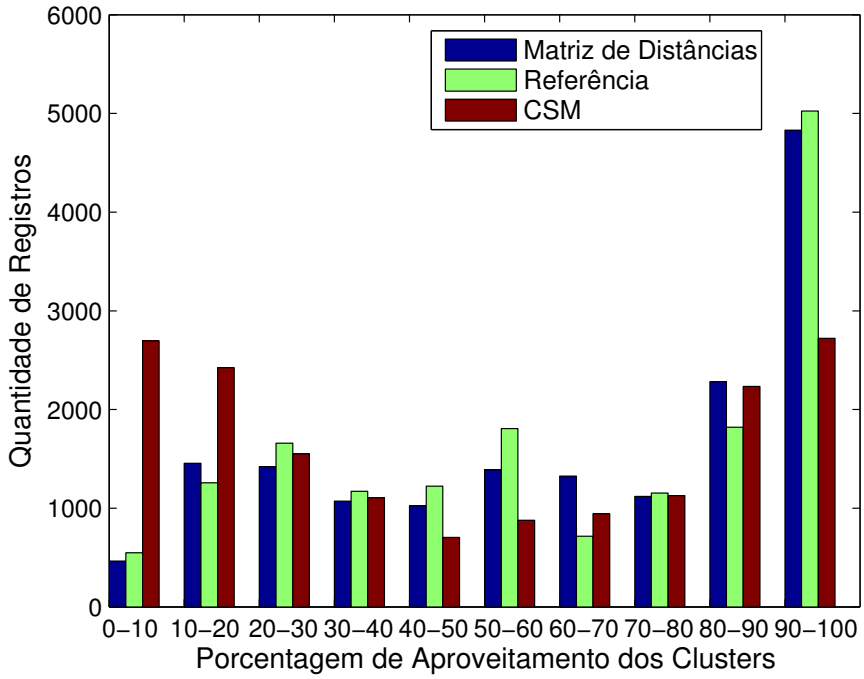
Ainda avaliando os experimentos considerando 500 grupos, a metodologia da matriz de distâncias obteve em média 4.830 registros em grupos com mais de 90% de acertos, tendo a média de 854 arquivos em grupos com 100% de acertos. Ainda foram registrados em média 2.282 registros na faixa de 80% a 90%. Considerando os arquivos existentes em grupos entre 70% e 100% de acertos, são 8.230 arquivos de interação. Avaliando o outro extremo, foram 3.340 arquivos em grupos com menos de 30% de acertos.

Através do CSM, foram encontrados 6.669 arquivos nos grupos com até 30% de acertos. Por outro lado. 6.084 elementos estiveram nos agrupamentos com o índice de

acerto superior a 70%, com 2.721 registros em grupos com mais de 90% e ainda 817 elementos em *clusters* com 100% de aproveitamento.

A Figura 13 ilustra através de um gráfico a comparação entre estas metodologias, ao considerar 500 *clusters*.

Figura 13 – Análise dos Registros em Relação à Porcentagem de Acertos dos 500 Grupos



A Figura 13 contém no eixo “x” as faixas de aproveitamento. No eixo “y” é dado a quantidade de registros que ficaram entre cada taxa de acerto. Os resultados obtidos pela metodologia apresentada, são representados pela cor azul escuro. A técnica que utiliza o arquivo de referência é ilustrada pela cor verde. A técnica CSM é representada pela cor vermelho grená.

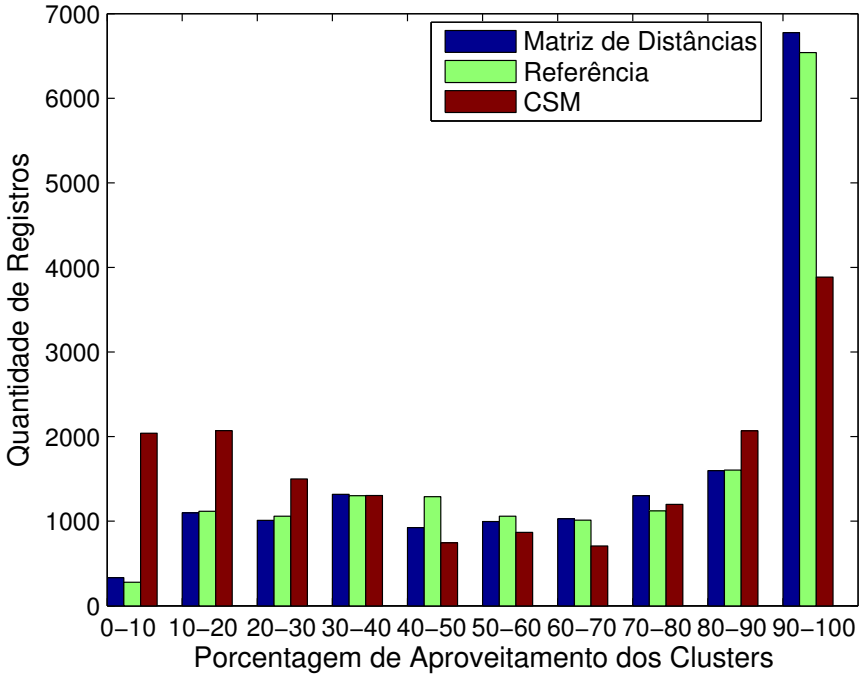
Como pode ser observado pela tabela 3 e pela Figura 13, os agrupamentos realizados pela matriz de distâncias e pela metodologia do arquivo de referência, além de obterem um melhor resultado no experimento anterior, obtiveram também vantagens nesta avaliação em relação ao CSM. A principal diferença se deu na faixa de acertos entre 90% e 100%. Foram em média 4.830 arquivos para a metodologia desenvolvida, 5.025 para a metodologia que utiliza o arquivo de referência e apenas 2.721 registros para o CSM. Além disso, o CSM obteve 5.119 registros em grupos que tiveram até 20% de aproveitamento.

A metodologia que utiliza o arquivo de referência conseguiu formar mais grupos com a porcentagem de acertos superior a 90%. Entretanto, de um modo geral, a representação pela matriz de distâncias obteve um maior aproveitamento entre todas as sobreposições realizadas. Foram 75% de acertos pela técnica da matriz de distâncias, contra 73% de acertos pela metodologia do arquivo de referência e 55% pelo CSM.

Os resultados contidos na tabela 3, indicam que a técnica da matriz de distâncias obteve nos experimentos de 750 grupos em média 6.775 registros em *clusters* com a taxa de acerto superior à 90%. Destes registros, 2.616 estiveram em grupos com 100% de acertos. A metodologia que utiliza um arquivo como referência teve 6.539 arquivos em grupos com mais de 90% de acertos, sendo 2.860 em grupos com 100% de precisão. O CSM obteve em média 3.885 registros em grupos com mais de 90% de acertos e 3.001 registros em *clusters* com 100%.

A representação pela matriz de distâncias, novamente obteve uma média geral de acertos superior às demais metodologias, obtendo 79,36%. A representação pelo arquivo de referência obteve 75,71% e o CSM 60,70%. As taxas de acertos com 750 agrupamentos foram superiores aos experimentos com 500 grupos. A Figura 14 mostra graficamente estes resultados.

Figura 14 – Análise dos Registros em Relação à Porcentagem de Acertos dos 750 Grupos

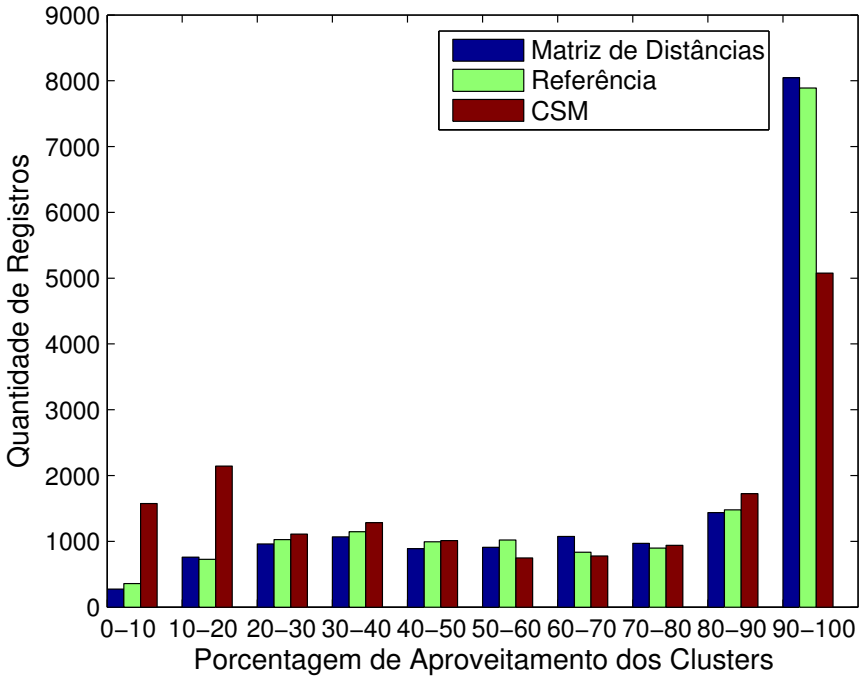


Como pode ser observado pela Figura 14, o aumento de 500 grupos para 750, possibilitou para todas as metodologias avaliadas uma diminuição na quantidade de registros nas três piores faixas de aproveitamento (entre 0% e 30%). Todas as três metodologias obtiveram também um aumento significativo de registros em grupos com a precisão superior à 90%.

A tabela 3 indica que a metodologia da matriz de referência obteve mais registros em agrupamentos com mais de 90% de acertos. Foram 8.045 arquivos, destes 4.130 registros estiveram em agrupamentos com 100%. A técnica do arquivo de referência obteve 7.890 arquivos em agrupamentos com a precisão superior à 90% e 4.510 em *clusters*

com 100% de acertos. O algoritmo CSM obteve 5.076 registros em grupos com a taxa de acerto superior à 90%, sendo 3.001 arquivos em agrupamentos com 100% de satisfação. A Figura 15 ilustra graficamente os resultados considerando 1000 grupos.

Figura 15 – Análise dos Registros em Relação à Porcentagem de Acertos dos 1000 Grupos



Como pode ser observado pela Figura 15, novamente com o aumento da quantidade de grupos, em todas as metodologias avaliadas os registros ficaram mais concentrados em agrupamentos com grandes taxas de acertos. Nestes experimentos, a técnica da matriz de distâncias e a representação pelo arquivo de referência, conseguiram obter resultados satisfatórios. A metodologia CSM teve uma forte melhora na medida em que a quantidade de agrupamentos definidos aumentou. Entretanto, os seus resultados ainda foram inferiores às outras técnicas comparadas.

5.2.3 Média e Regressão Linear das Distâncias dos Registros para os centróides de seus grupos

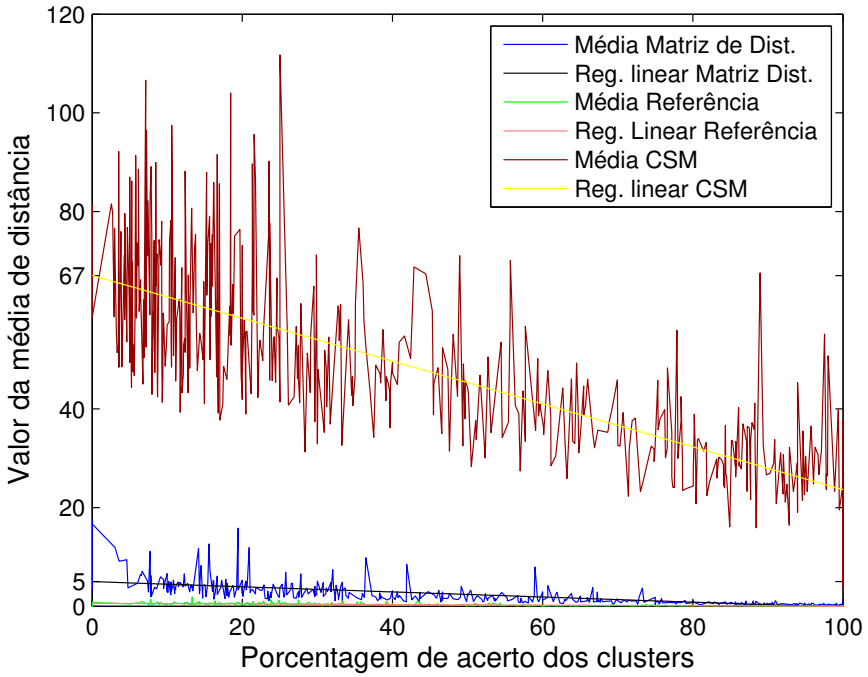
Para entender melhor a formação dos agrupamentos, foi avaliado a média da distância de cada registro em relação ao centro do seu grupo. Estes experimentos foram realizados para as três metodologias, considerando 500, 750 e 1.000 *clusters*.

Dentre todas as execuções de cada técnica para uma certa quantidade de *clusters*, foi escolhida a que obteve uma melhor taxa de acerto para representar a sua metodologia neste experimento. Por exemplo, dentre todas as execuções da técnica da matriz de distâncias, com 500 grupos, a que apresentou uma maior taxa de acerto foi escolhida para

representar a sua metodologia. O mesmo se repetiu para as outras técnicas e para as outras quantidades de grupos definidos.

A primeira destas análises, consistiu em calcular a média de distância entre todos os elementos de cada *cluster* para o seu respectivo centróide. O objetivo desta análise é verificar a relação entre a taxa de aproveitamento dos grupos (eixo x), para o valor médio da distância dos registros para o centro de cada agrupamento(eixo y). A figura Figura 16 ilustra o resultado para o experimento com 500 *clusters*.

Figura 16 – Média de Distância para o Centróide, com todas as Metodologias, em 500 grupos



A Figura 16 tem no eixo “y” a média de distância dos registros para o centro do grupo. O eixo “x”, mostra a respectiva porcentagem de acertos destes agrupamentos. A linha de cor azul ilustra a média das distâncias pela metodologia da Matriz de distâncias e a linha preta indica a regressão linear desta técnica. A linha de cor verde, mostra a média de valores para os agrupamentos utilizando um arquivo de referência e a linha de cor laranja é a sua regressão linear. Por fim, a linha de cor vermelha grená, indica a média de distâncias pelo CSM e a linha amarela mostra a sua regressão linear.

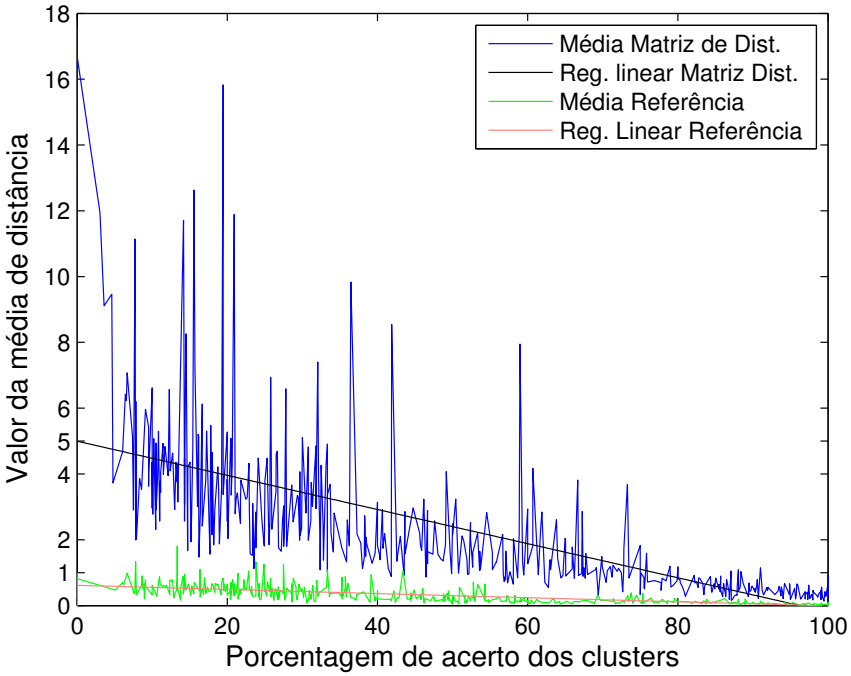
Conforme a Figura 16, nos resultados de todas as técnicas, os *clusters* que tiveram maiores taxas de acertos, possuíram um menor valor de diferença de distância para o centro de seu grupo. A linha de regressão linear, mostra o aumento da distância dos registros em relação aos seus respectivos centróides, na medida em que a taxa de acerto sofre queda.

É importante destacar que pela técnica do arquivo de referência, a linha de regressão linear teve 0,6 em seu fim. Nas técnicas da matriz de distâncias e CSM, as suas respectivas

linhas de regressão linear tiveram 5 e 67.

Como a metodologia CSM apresentou valores altos de médias de distâncias, e assim, ofuscou os detalhes das outas técnicas avaliadas, foi gerado um novo gráfico. A Figura 17 mostra apenas os resultados dos agrupamentos pela matriz de distâncias e pelo arquivo de referência.

Figura 17 – Média de Distância para o Centróide, desconsiderando o CSM, em 500 grupos

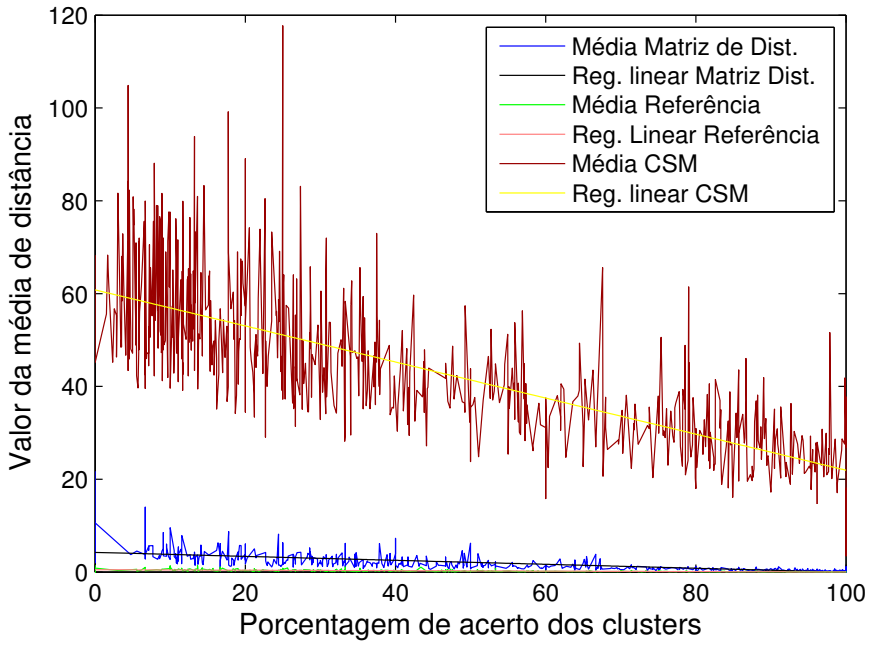


Pela Figura 17, é possível verificar de uma forma mais clara os resultados da técnica da matriz de distâncias (linha azul) e do arquivo de referência (linha verde). Na medida em que os agrupamentos destas técnicas tinham piores taxas de acertos (eixo x), a média de distância para os centróides teve um aumento (eixo y). A metodologia que utilizou um arquivo como referência, apresentou um aumento de distância muito leve, mesmo para os grupos em que a taxa de acerto foi muito baixa, atingindo no 0,6 no fim da linha de regressão linear. Este acontecimento se deu por esta ser a única das técnicas avaliadas que realiza sobreposições antes do agrupamento. Deste modo, os registros já são agrupados de uma maneira mais próxima. Entretanto, esta proximidade das distâncias indica que arquivos próximos entre si, não serão necessariamente similares pelo LSQKAB. A explicação para este acontecimento é que um arquivo muito diferente da referência, não será necessariamente idêntico a um outro arquivo também diferente. Ou seja, cada átomo de cada registro, tem as coordenadas “x”,“y” e “z”. Ao comparar estas coordenadas com o arquivo de referência, será dado um único valor para expressar a diferença entre os átomos, em relação a essas três coordenadas. Deste modo, a representação com um arquivo de referência conseguiu deixar os arquivos próximos em seus *clusters*, mas esta proximidade

não necessariamente indica que estes registros são realmente similares. De um modo geral, considerando as sobreposições aceitas pelo LSQKAB, para os experimentos com 500 grupos, a taxa de acertos da metodologia do arquivo de referência foi de 73,90% e a matriz de distâncias obteve 76,70% (mesmo os registros ficando mais distantes ente si).

A análise da distância dos registros para o centróide de seu grupo, para os experimentos com 750 *clusters* é ilustrada pela Figura 18

Figura 18 – Média de distância para o centróide, para todas as metodologias, com 750 grupos



De acordo com a Figura 18, a média da distância dos registros para os centróides, seguiram o mesmo padrão, mas ficaram menores. A técnica CSM (linha vermelha) continuou com os registros mais dispersos nos agrupamentos do que as outras técnicas. Entretanto, pode-se notar pela linha de regressão linear que ao considerar 750 grupos, a distância dos registros para os centróides teve uma queda significativa. O mesmo aconteceu para as outras duas metodologias, como pode ser visto mais claramente pela Figura 19.

Assim como ilustrado pela Figura 19, os registros de ambas as técnicas ficaram mais próximos dos centros de seus grupos. Na técnica da matriz de distâncias (linha azul), a linha de regressão linear (linha preta) atingiu 4,2 no seu fim. A linha da regressão linear da técnica do arquivo de referência (linha vermelha), obteve no seu fim o valor de 0,5. Comparando esta análise com a avaliação geral destas execuções pelo LSQKAB, nos experimentos com 750 grupos, esta execução da matriz de distâncias obteve 80,42% de acertos e a representação pelo arquivo de referência obteve 76,86%.

A Figura 20 ilustra a análise da média de distância dos registros para os seus respectivos centróides, nos experimentos de 1.000 grupos.

Figura 19 – Média de distância para o Centróide, desconsiderando o CSM, com 750 grupos

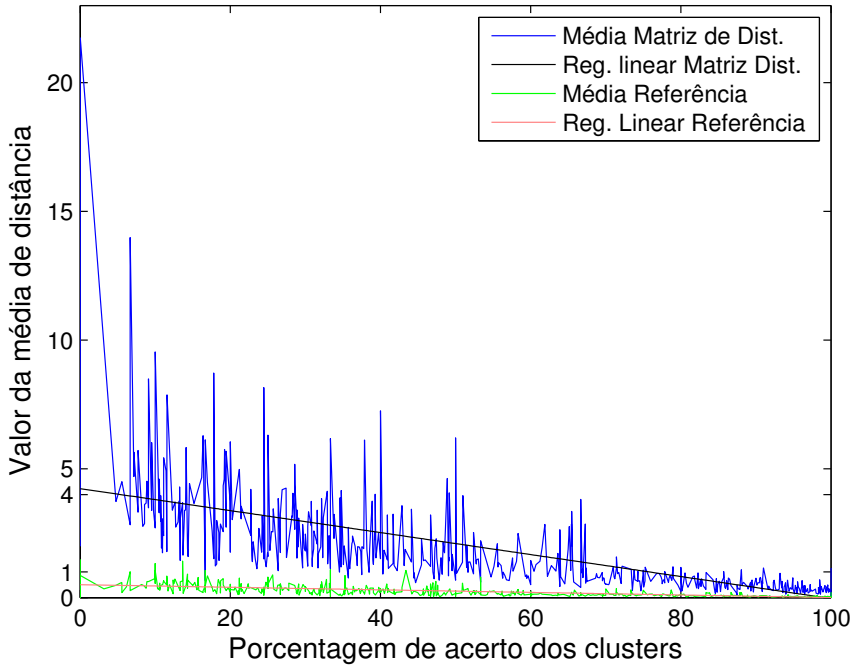
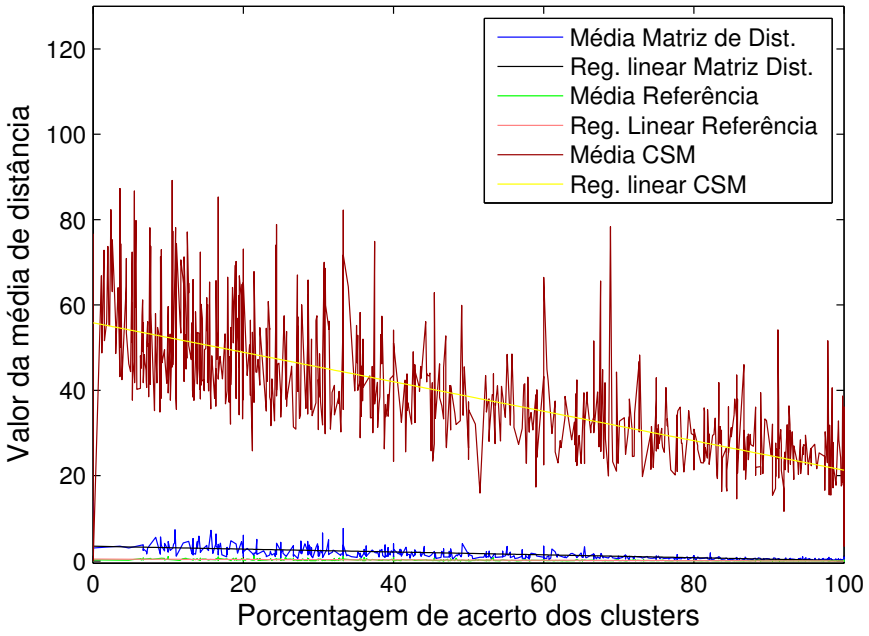


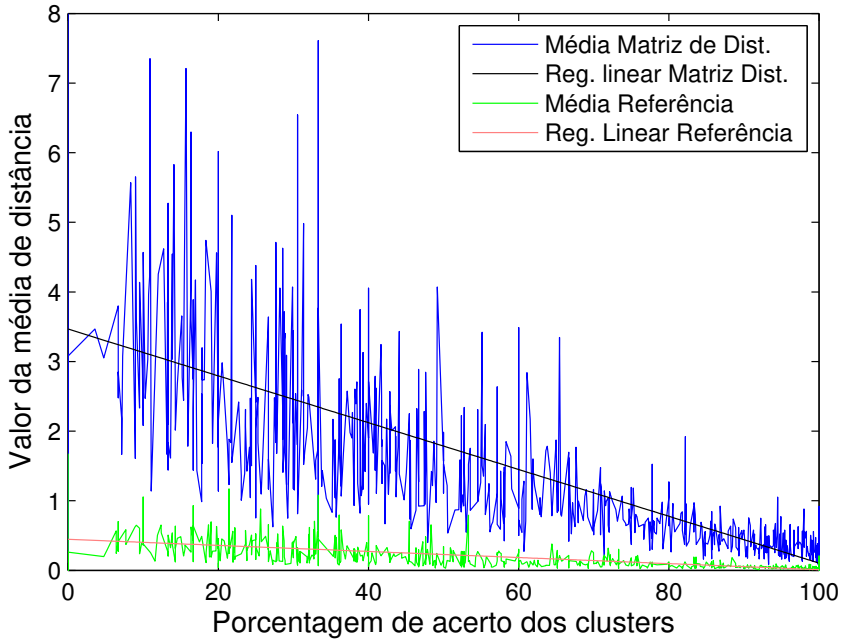
Figura 20 – Média de distância para o centróide, para todas as metodologias, com 1.000 grupos



Conforme ilustrado pela Figura 20, o melhor experimento do CSM com 1.000 grupos, teve a sua linha de regressão linear mais próximos em relação aos seus experimentos anteriores. Desta vez, a sua linha de regressão linear atingiu em seu fim o valor 55. A verificação desta execução pelo LSQKAB atingiu 65,9% de acertos. Entretanto, estes

resultados ainda ficaram abaixo dos obtidos pelas outras técnicas. A Figura 21 ilustra com maiores detalhes os resultados das outras duas técnicas comparadas.

Figura 21 – Média de Distância para o Centróide sem considerar o CSM (1.000 grupos)



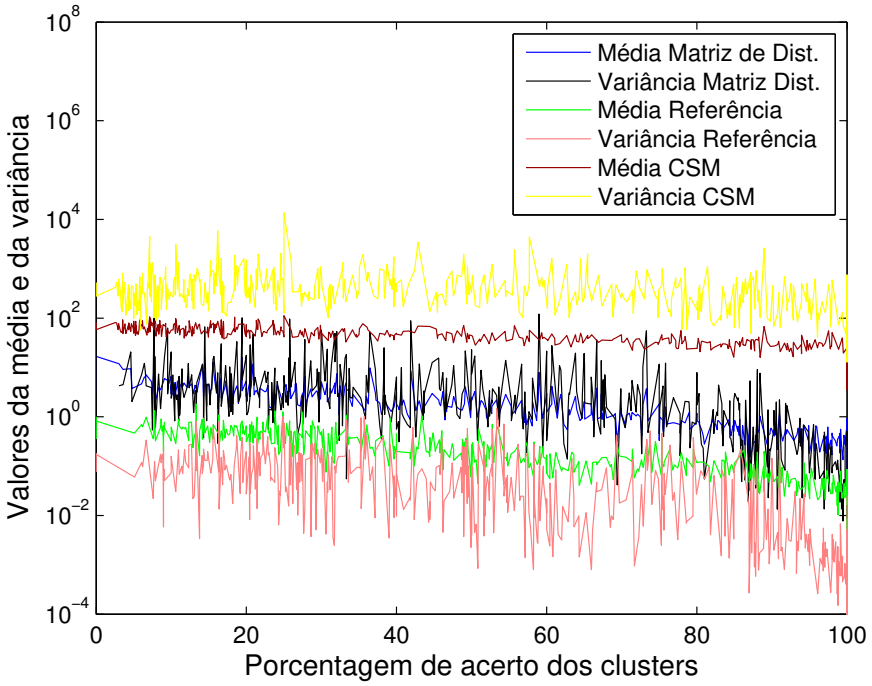
Conforme a Figura 21, nos experimentos com 1.000 grupos, estas metodologias obtiveram os seus registros ainda mais próximos. A linha de regressão linear da técnica da matriz de distâncias (linha preta), obteve em seu fim 3,4. A linha de regressão linear da técnica do arquivo de referência obteve em seu fim 0,44. Comparando estes valores com a taxa de acerto pelo LSQKAB, cada técnica obteve respectivamente 83,10% e 81,2%.

5.2.4 Variância das Distâncias dos Registros para os Centróides de seus grupos

As próximas análises consistem na comparação entre a média de distâncias para os respectivos centróides de cada grupo e a variância. Novamente, estes experimentos foram feitos tendo como base a melhor execução de cada técnica para os experimentos com 500, 750 e 1.000 grupos. A Figura 22 ilustra os resultados para os experimentos com 500 grupos.

A Figura 22 contém no eixo “y” os valores da média e da variância em escala logarítmica. Este tipo de escala foi utilizada devido ao CSM apresentar valores de variância muito altos e ofuscar os resultados das outras técnicas. No eixo “x”, estão as porcentagens de acerto. Para esta análise, foram definidas a linha de cor azul para ilustrar a média das distâncias pela metodologia da Matriz de distâncias e a linha preta indica a sua respectiva variância. A linha de cor verde, mostra a média de valores para os agrupamentos utilizando

Figura 22 – Média e variância da Distância para o Centróide, para todas as metodologias, com 500 grupos



um arquivo de referência e a linha de cor laranja é a variância desta metodologia. Por fim, a linha de cor vermelho grená, indica a média de distâncias pelo CSM e a linha amarela mostra a sua respectiva variância.

Como pode ser visto pela Figura 22, o valor da variância pela metodologia com um arquivo de referência foi muito baixo, inclusive ficando menor do que a sua média de distâncias e sendo a menor dentre as técnicas avaliadas. Os valores da variância da metodologia da matriz de distâncias, esteve baixa para a maioria dos grupos que tiveram o aproveitamento próximo a 100%. Entretanto, o valor aumentou para a maioria dos grupos. A técnica CSM apresentou valores de variância muito altos, chegando a atingir 10^4 em alguns grupos.

A Figura 23 indica a análise da variância para os experimentos com 750 grupos.

Como pode ser observado pela Figura 23, todas as técnicas tiveram um comportamento similar ao experimento com 500 *clusters*. Entretanto, como o aumento da quantidade de grupos, o valor da variância foi levemente menor em cada técnica.

A Figura 24 mostra a análise da variância considerando 1.000 grupos.

A Figura 24 indica que o comportamento de ambas as técnicas em relação à variância foi similar também ao definir 1.000 grupos. A variância da matriz de distâncias foi baixa, mas não tão excelente quanto a técnica do arquivo de referência. O fato dos registros serem sobrepostos antes do agrupamento ajudou na obtenção deste resultado. Em todos

Figura 23 – Média e variância da Distância para o Centróide, com todas as técnicas, com 750 grupos

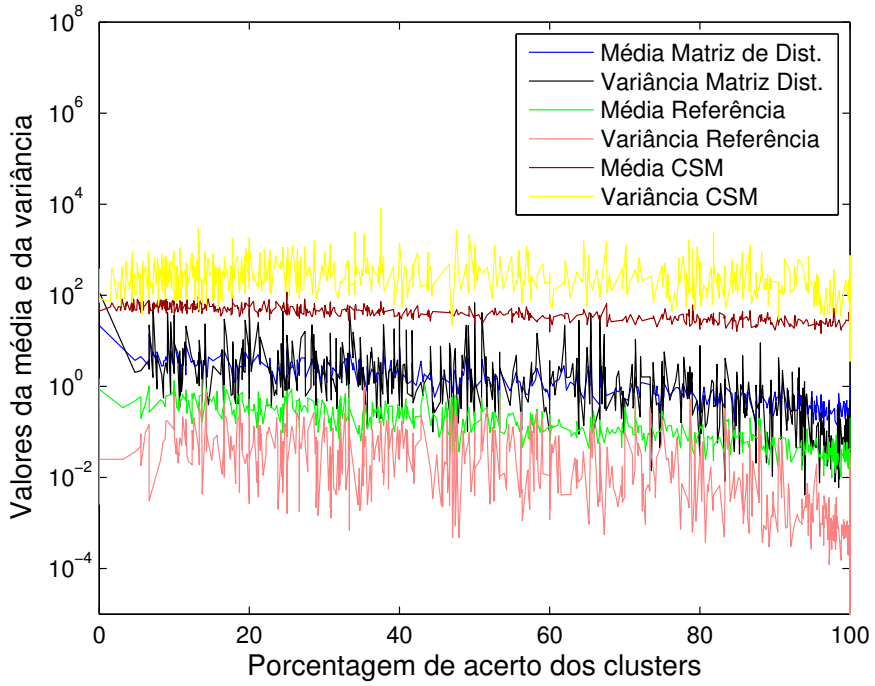
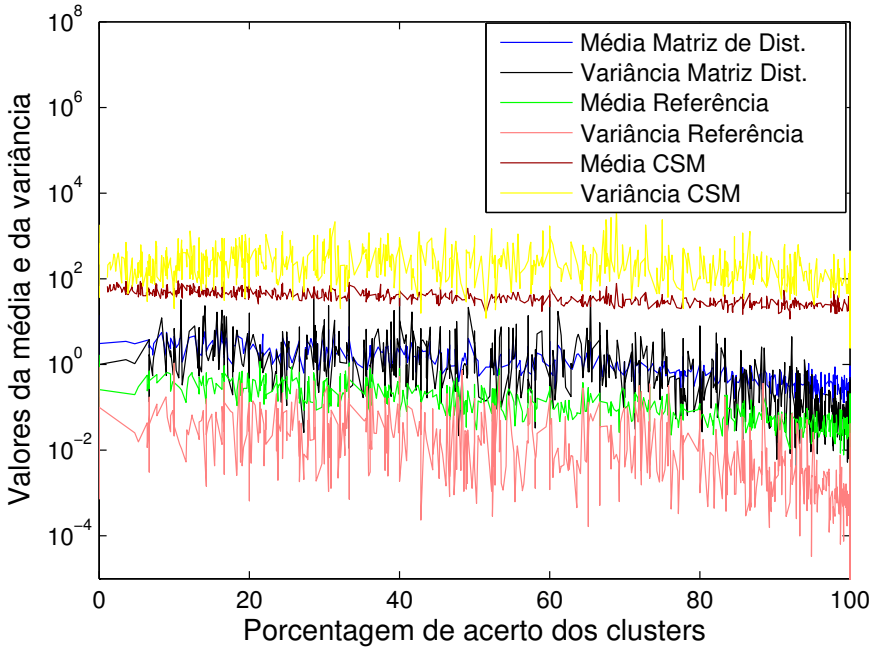


Figura 24 – Média e variância da Distância para o centróide, com todas as técnicas, com 1.000 grupos



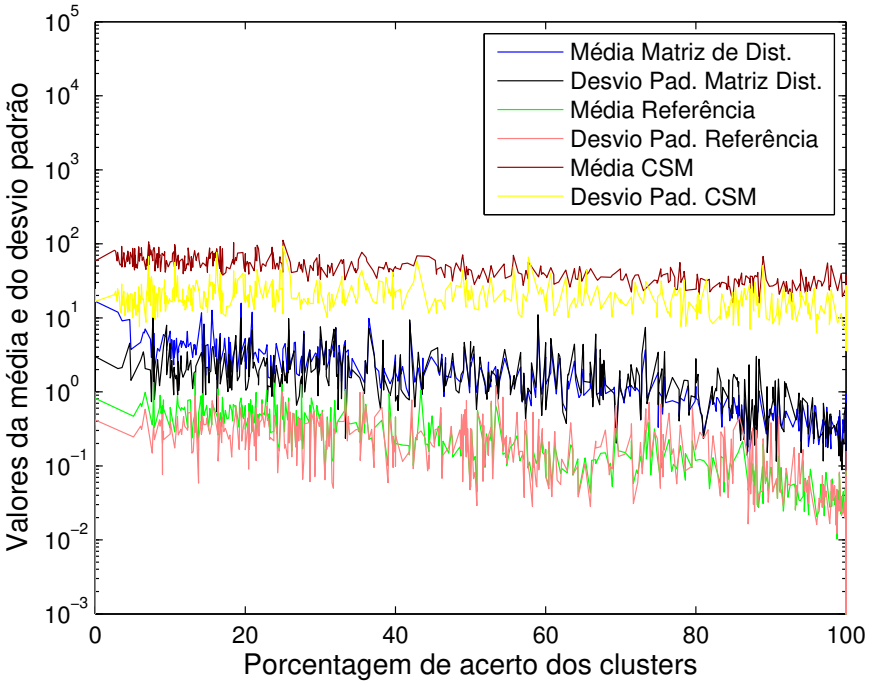
os cenários, o CSM apresentou uma variância muito maior do que as outras técnicas.

5.2.5 Desvio Padrão das Distâncias dos Registros para os Centróides de seus grupos

A medida pelo desvio padrão possui uma vantagem em relação à variância, pois é permitido uma interpretação mais direta da variação do conjunto de dados. Do mesmo modo que nos experimentos anteriores, cada metodologia teve o seu desvio padrão avaliado com base na execução que obteve o melhor resultado em cada experimento com 500, 750 e 1.000 grupos.

A Figura 25 faz a análise do desvio padrão em relação à média de distâncias dos registros para os seus centróides, considerando 500 grupos.

Figura 25 – Desvio Padrão da Média de Distância para o centróide, considerando todas as técnicas, em 500 grupos



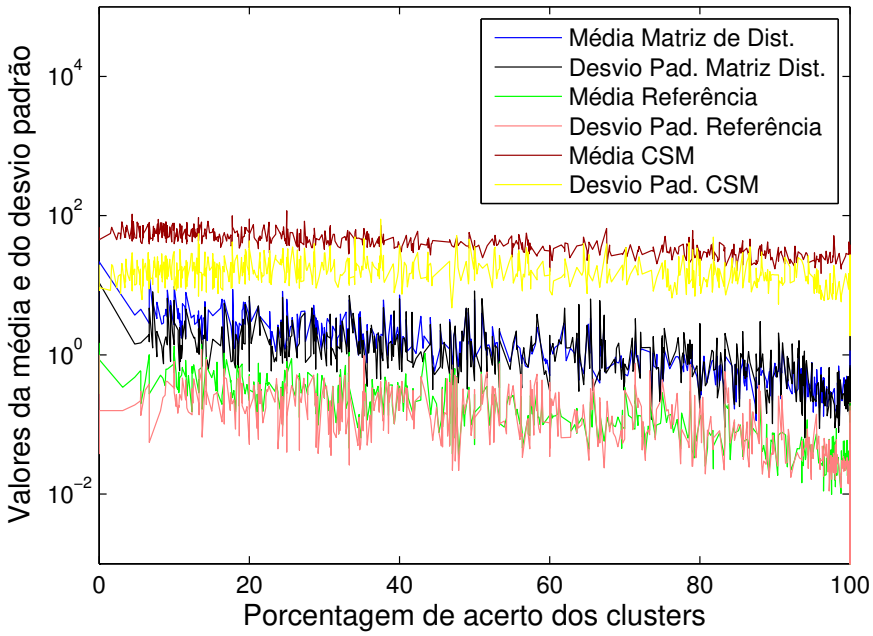
A Figura 25 contém no eixo “y”, os valores da média e do desvio padrão. Estes valores foram representados em escala logarítmica pois o CSM ficou fora de escala. O eixo “x” contém a taxa de acerto dos *clusters*. Para esta análise, foram definidas a linha de cor azul para ilustrar a média das distâncias pela metodologia da Matriz de distâncias e a linha preta indica o seu desvio padrão. A linha de cor verde, mostra a média de valores para os agrupamentos utilizando um arquivo de referência e a linha de cor laranja é o desvio padrão desta metodologia. A linha de cor vermelho grená, indica a média de distâncias pelo CSM e a linha amarela mostra o seu desvio padrão.

Como pode ser observado pela Figura 25, o desvio padrão do CSM foi bem menor do que o valor da média de distâncias. A Figura 25 indica que a metodologia da matriz de

distâncias teve o desvio padrão (linha preta) um pouco menor do que a média de distâncias (linha azul). A metodologia do arquivo de referência teve os valores da média de distância (linha verde) e do desvio padrão (linha laranja) muito baixos, sendo que este último também foi menor do que a média das distâncias.

Os resultados obtidos na análise com 750 grupos podem ser visualizados pela Figura 26

Figura 26 – Desvio Padrão da Média de Distância para o centróide, com todas as metodologias, em 750 grupos

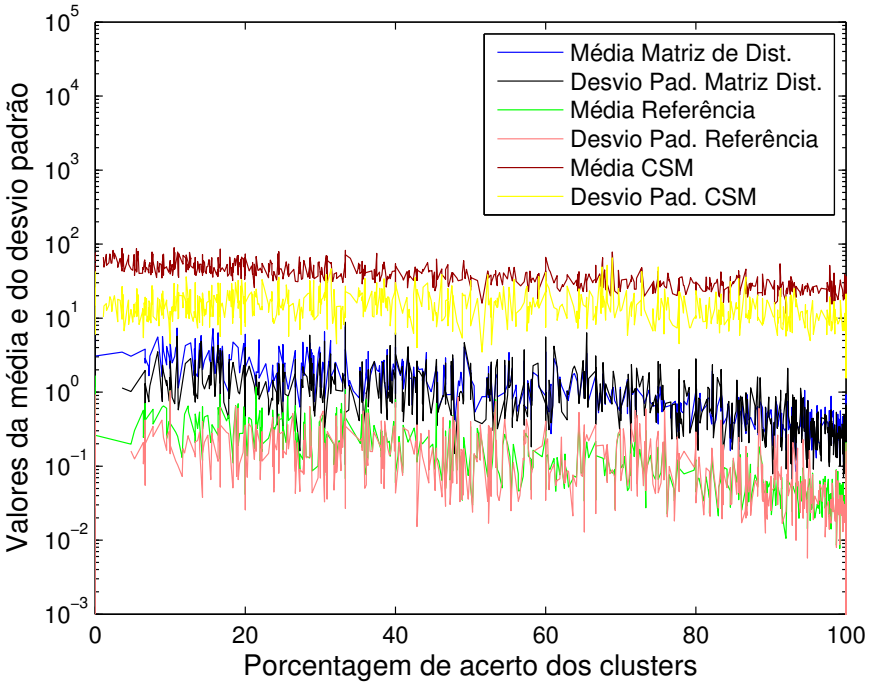


De acordo com a Figura 26, o comportamento de ambas as técnicas foi similar ao experimento com 500 *clusters*. Entretanto, em todas as metodologias, o desvio padrão ficou menor devido ao aumento da quantidade de grupos definidos.

Os experimentos com 1.000 grupos podem ser vistos pela Figura 27.

De acordo com a Figura 27, novamente, ambas as técnicas tiveram o mesmo comportamento. Os valores da média e do desvio padrão tiveram uma nova queda devido ao aumento da quantidade de grupos. A metodologia apresentada teve o desvio padrão baixo, principalmente nos grupos que tiveram o aproveitamento próximo à 100%. A metodologia que apresentou os melhores valores de desvio padrão foi a técnica do arquivo de referência, que teve um desvio padrão baixo até para os grupos com pequenas taxas de acertos. Por outro lado, o CSM apresentou os maiores valores de desvio padrão.

Figura 27 – Desvio Padrão da Média de Distância para o centróide, com todas as metodologias, em 1.000 grupos



5.3 Análise de Complexidade dos Algoritmos

A metodologia proposta tem o tempo de complexidade para transformar um arquivo de interação na forma de vetor em $O(n^2/2)$. Onde “n” é a quantidade de átomos que o arquivo possui.

O CSM tem o seu tempo de complexidade em $O(n^2)$ para transformar um registro no formato de grafo (onde “n” é o número de átomos que o arquivo possui). Esta metodologia também possui a complexidade $O(p * n^2)$ para obter o padrão de distâncias de um registro da matriz de grafos e inseri-lo na matriz CSM. Neste último caso, “p” é a quantidade de passos definidos para obter a frequência dos átomos.

O método de busca da ferramenta RID utiliza o LSQKAB para realizar sobreposições atômicas. O LSQKAB trabalha conforme o algoritmo de Kabsch. De acordo com (SHIBUYA, 2010), o algoritmo de Kabsch compara duas estruturas tridimensionais em um tempo $O(n * m)$, onde “n” e “m” referem-se ao tamanho das estruturas comparadas.

Conforme citado por (NAZEER; SEBASTIAN, 2009), o tempo de complexidade do algoritmo *K-Means Clustering* é de $O(n^2)$ para definir os primeiros centróides, onde “n” é a quantidade de registros. De acordo com (CHRISTOPHER; PRABHAKAR; HINRICH, 2008), para realizar os próximos passos, o seu tempo de complexidade é aproximadamente O(n*k*i*d). Onde “n” é a quantidade de registros a serem agrupados, “k” o número de *clusters*, “i” é a quantidade de iterações e “d” é o número de atributos de cada registro.

5.4 Avaliação do Tempo de Execução

O tempo de execução foi comparado em um notebook Intel (R) Core (TM) i3, com 4 GB RAM e sistema operacional Ubuntu 16. Foi avaliado o tempo gasto pelas três metodologias para construir os dados a serem agrupados, o tempo gasto para realizar o agrupamento e o tempo total. Os experimentos foram repetidos por 5 vezes e foi calculada a média do tempo de execução. A tabela 4 ilustra os resultados obtidos.

Tabela 4 – Tempo de Execução

Técnica	Mat. Dist.			Referência			CSM		
Nº de clusters	500	750	1.000	500	750	1.000	500	750	1.000
Construção dos dados	0:46	0:46	0:46	4:40	4:40	4:40	4:00	4:00	4:00
Agrupamento	3:14	5:10	6:26	1:05	1:48	2:38	9:24	11:40	12:25
Tempo total	4:00	5:56	7:12	5:45	6:28	7:20	13:24	15:40	16:25

Como pode ser observado pela tabela 4, a técnica do arquivo de referência demorou mais tempo para construir os dados. Foram gastos 4 minutos e 40 segundos para gerar os arquivos de deltas entre cada arquivo de interação com o registro escolhido como referência. É importante ressaltar que caso o usuário ainda não tenha escolhido o arquivo de referência, e queira escolher o de melhor resolução, como feito em Dias (2012), seria necessário ainda executar um outro *script* para avaliar a resolução de cada arquivo. O CSM demorou em média 4 minutos para gerar o padrão de distâncias para todos os registros avaliados. A técnica da matriz de distâncias demorou 46 segundos para construir os dados a serem agrupados, sendo bem mais rápida do que as demais metodologias.

Na etapa de construção dos dados, a técnica do arquivo de referência teve um procedimento mais demorado. Ainda foram necessários 8,7 *mega bytes* em disco rígido para armazenar os 16.382 arquivos de deltas que foram gerados.

Por outro lado, como pode ser visto pela tabela de tempo de execução, a técnica do arquivo de referência apresentou um melhor desempenho nos agrupamentos. Um dos motivos é que esta técnica agrupou com 12 valores para cada registro. A técnica da matriz de distâncias agrupou com 66 valores para cada registro e o CSM, 151 para cada arquivo. O outro motivo da vantagem da técnica do arquivo de referência, deve-se a proximidade dos registros em relação ao seu centróide, fazendo com que não fosse necessário a realização das 1.000 interações. Por sua vez, o CSM apresentou uma grande demora para realizar a etapa de agrupamento. Como os registros ficaram distantes de seus centróides, foi necessário um número maior interações por parte do *k-means*.

Avaliando o tempo total, em todos os casos, a utilização técnica da matriz de distancias tornou-se mais rápida. Entretanto, em todas as técnicas, o tempo gasto com o agrupamento aumentou na medida em que a quantidade definida de *clusters* também

aumentou. Como a técnica do arquivo de referência apresentou vantagens no tempo de agrupamento, o tempo total gasto por esta metodologia nos experimentos com 1.000 grupos foi próximo da técnica da matriz de distâncias.

5.5 Avaliação do Uso de Memória RAM

O uso da memória foi comparado no mesmo notebook dos experimentos anteriores. A ide utilizada foi o Matlab e o sistema operacional foi o Ubuntu 16. Como cada metodologia agrupou números diferentes de informações para representar os registros, foi avaliado o momento de maior consumo de memória RAM pelas pelas três metodologias durante os agrupamentos. A tabela 5 ilustra os resultados obtidos.

Tabela 5 – Uso de Memória

Técnica	Mat. Dist.			Referência			CSM		
Nº de clusters	500	750	1.000	500	750	1.000	500	750	1.000
% de uso de memória RAM	22	23,5	25	20,6	22,3	24	21	22,6	24,5

Como pode ser observado pela tabela 5, embora a técnica do arquivo de referência tenha consumido uma quantidade menor de memória RAM, a diferença não foi alta em relação às outras metodologias. Também é importante notar que em todas as metodologias, ao aumentar a quantidade de grupos, ocorreu também um aumento do uso de memória para realizar os agrupamentos.

6 Conclusão

Os resultados obtidos através dos experimentos foram satisfatórios. Foi possível encontrar diversos grupos de estruturas tridimensionais de proteínas, conforme as similaridades de suas distâncias atômicas.

A metodologia da matriz de distâncias mostrou uma maior precisão na porcentagem de acerto total se comparada com as outras técnicas avaliadas. Embora o intervalo de confiança não permita afirmar a sua superioridade, a metodologia apresentada obteve mais acertos entre todas as sobreposições realizadas pelo LSQKAB, em todos os experimentos. Obtendo ainda uma maior quantidade de registros em grupos com mais de 90% de acertos para os experimentos com 750 e 1.000 grupos. Uma outra vantagem desta metodologia é referente ao tempo de execução, pois ela não necessita de realizar sobreposições atômicas com um arquivo de referência e a matriz a ser agrupada é construída facilmente. Deste modo, esta metodologia demonstrou ser suficiente para seu uso com eficiência no agrupamento de itens similares dentre os arquivos de interações.

A representação por um arquivo de referência, obteve ótimos resultados na quantidade de registros em grupos com as melhores taxas de aproveitamento. Apesar de depender de um maior tempo para construir os valores a serem agrupados, os *clusters* formados por esta metodologia, deixaram os registros muito perto dos centróides de seus grupos, e consequentemente próximos entre si. Com esta proximidade dos registros, esta metodologia apresentou também ótimos resultados na avaliação da variância e do desvio padrão. Além disso, como esta metodologia depende de apenas 12 valores por registro para realizar os agrupamentos, parte do tempo demorado na geração dos arquivos de deltas foi compensada na etapa de agrupamento. O consumo de memória RAM por esta metodologia foi levemente menor do que as outras.

A técnica CSM tornou-se interessante por permitir as comparações diretamente pelos arquivos de interações, não sendo necessário escolher um arquivo como referência e realizar sobreposições dos demais registros da base de dados contra o arquivo escolhido. Entretanto, os seus resultados ficaram abaixo das demais metodologias comparadas, tanto na proximidade entre os registros de um mesmo grupo, quanto na taxa de acerto pelo LSQKAB.

6.1 Trabalhos Futuros

Apesar dos resultados serem satisfatórios, ainda existem novos experimentos a serem realizados, tais como:

1. Realizar experimentos com outras bases de dados: Verificar os resultados em uma base de dados de aproximadamente 100.000 registros, e com arquivos de interações diferentes. É esperado que a metodologia apresentada obtenha uma precisão similar ou superior à obtida com esta base de dados. A utilização de mais arquivos formará grupos com os registros mais próximos entre si.
2. Incluir novas metodologias e realizar comparações: Pesquisar por outros algoritmos e comparar a sua precisão com as metodologias avaliadas neste trabalho.
3. Adaptação da metodologia proposta para a ferramenta RID: Incluir a metodologia aqui proposta na ferramenta RID para substituição da estratégia usada. Visando a melhoria da identificação de itens similares na busca por candidatos à mutação.

Referências

CCP4. **LSQKAB (CCP4: Supported Program)**. 2016. Disponível em: <http://www.ccp4.ac.uk/html/lsqkab.html>. Acesso em: 23 de janeiro de 2016. Citado 2 vezes nas páginas 1 e 6.

CHRISTOPHER, D. M.; PRABHAKAR, R.; HINRICH, S. Introduction to information retrieval. **An Introduction To Information Retrieval**, v. 151, p. 177, 2008. Citado na página 45.

DIAS, S. R. **Residue interaction database: proposição de mutações sítio dirigidas com base em interações observadas em proteínas de estrutura tridimensional conhecida**. Tese (Doutorado) — Universidade Federal de Minas Gerais - UFMG, 2012. Disponível em: <http://www.bibliotecadigital.ufmg.br/dspace/bitstream/handle/1843/BUOS-9GQKBD/tese_sandrord_vfinal9_ok.pdf?sequence=1>. Acesso em: 22 de novembro de 2015. Citado 8 vezes nas páginas 1, 2, 3, 6, 11, 17, 22 e 46.

GROSS, J. L.; YELLEN, J. **Handbook of graph theory**. [S.l.]: CRC press, 2004. Citado na página 14.

KABSCH, W. A solution for the best rotation to relate two sets of vectors. **Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography**, International Union of Crystallography, v. 32, n. 5, p. 922–923, 1976. Citado na página 5.

KAWAJI, H.; TAKENAKA, Y.; MATSUDA, H. Graph-based clustering for finding distant relationships in a large set of protein sequences. **Bioinformatics**, Oxford Univ Press, v. 20, n. 2, p. 243–252, 2004. Citado na página 7.

KUFAREVA, I.; ABAGYAN, R. Methods of protein structure comparison. **Homology Modeling: Methods and Protocols**, Springer, p. 231–257, 2012. Citado 2 vezes nas páginas 6 e 13.

LODISH, H. et al. **Biologia celular e molecular**. [S.l.]: Artmed Editora, 2014. Citado na página 13.

MAITI, R. et al. Superpose: a simple server for sophisticated structural superposition. **Nucleic acids research**, Oxford Univ Press, v. 32, n. suppl 2, p. W590–W594, 2004. Citado na página 7.

MALISCHEWSKI, S.; SCHUMANN, H.; HOFFMANN, D. **kabsch Algorithm**. 2016. Disponível em: <http://biomolecularstructures.readthedocs.io/en/latest/kabsch/>. Acesso em: 13 de setembro de 2015. Citado na página 5.

MATHWORKS. **Machine learning method for finding and visualizing natural groupings and patterns in data**. 2017. Disponível em: <https://www.mathworks.com/discovery/cluster-analysis.html>. Acesso em: 20 de junho de 2017. Citado na página 15.

MITCHELL, R. N. et al. Robbins e cotran-fundamentos de patologia. **São Paulo**, 2006. Citado na página 13.

NAZEER, K. A.; SEBASTIAN, M. Improving the accuracy and efficiency of the k-means clustering algorithm. In: **Proceedings of the World congress on Engineering**. [S.l.: s.n.], 2009. v. 1, p. 1–3. Citado na página 45.

NELSON, D. L.; LEHNINGER, A. L.; COX, M. M. **Lehninger principles of biochemistry**. [S.l.]: Macmillan, 2008. Citado 3 vezes nas páginas 1, 9 e 10.

PDB. **Coordinate Section**. 2010. Disponível em: <http://www.wwpdb.org/documentation/file-format-content/format33/sect9.html>. Acesso em: 05 de fevereiro de 2016. Citado 2 vezes nas páginas 12 e 24.

PINHO, M. S. **Matriz e Vetores**. 2017. Disponível em: <http://www.inf.pucrs.br/~pinho/LaproI/Vetores/Vetores.htm>. Acesso em: 20 de maio de 2017. Citado na página 14.

PIRES, D. E.; ASCHER, D. B. Csm-lig: a web server for assessing and comparing protein–small molecule affinities. **Nucleic acids research**, Oxford University Press, v. 44, n. W1, p. W557–W561, 2016. Citado na página 8.

PIRES, D. E. et al. Cutoff scanning matrix (csm): structural classification and function prediction by protein inter-residue distance patterns. **BMC genomics**, BioMed Central, v. 12, n. 4, p. S12, 2011. Citado 8 vezes nas páginas 2, 3, 7, 14, 17, 18, 19 e 23.

PIRES, D. E. et al. acsm: noise-free graph-based signatures to large-scale receptor-based ligand prediction. **Bioinformatics**, Oxford University Press, v. 29, n. 7, p. 855–861, 2013. Citado na página 8.

PIRES, D. E. V. **CSM: Uma Assinatura para Grafos Biológicos Baseada em Padrões de Distâncias**. Tese (Doutorado) — Universidade Federal de Minas Gerais - UFMG, 2012. Disponível em: <http://homepages.dcc.ufmg.br/~dpires/publications/thesis.pdf>. Acesso em: 22 de abril de 2016. Citado na página 18.

PIRES, D. E. V. **Cutoff Scanning Matrix (CSM): funciton prediction and fold recognition by protein inter-residue distance patterns**. 2014. Disponível em: <http://homepages.dcc.ufmg.br/~dpires/csm/index.html>. Acesso em: 23 de janeiro de 2017. Citado na página 23.

QUILICI-GONZALEZ, J. A.; ZAMPIROLLI, F. de A. **Sistemas Inteligentes e Mineração de Dados**. [S.l.]: Francisco de Assis Zampirolli, 2015. Citado na página 15.

RSCB. **About the PDB Archive and the RCSB PDB**. 2016. Disponível em: <http://www.rcsb.org/pdb/static.do?p=general_information/about_pdb/index.html>. Acesso em: 26 de janeiro de 2016. Citado 2 vezes nas páginas 1 e 9.

RSCB. **About the PDB Archive and the RCSB PDB**. 2016. Disponível em: <http://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/resolution>. Acesso em: 05 de junho de 2016. Citado 2 vezes nas páginas 6 e 22.

SHIBUYA, T. Searching protein 3-d structures in linear time. **Journal of Computational Biology**, Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA, v. 17, n. 3, p. 203–219, 2010. Citado na página 45.

SINGH, J.; THORNTON, J. M. **Atlas of protein side-chain interactions**. [S.l.]: IRL Press at Oxford University Press, 1991. Citado na página 7.

TSAI, C. S. **An introduction to computational biochemistry**. [S.l.]: John Wiley & Sons, 2003. Citado 2 vezes nas páginas 14 e 23.

ZEMLA, A. Lga: a method for finding 3d similarities in protein structures. **Nucleic acids research**, Oxford Univ Press, v. 31, n. 13, p. 3370–3374, 2003. Citado na página 7.