Trabalho apresentado no XXXIX CNMAC, Uberlândia - MG, 2019.

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Desenvolvimento de uma Metodologia Baseada em Matriz de Distâncias para a Verificação de Similaridades de Proteínas

Otaviano Martins Monteiro [1]
Sandro Renato Dias [2]
Thiago de Souza Rodrigues [3]
Centro Federal de Educação Tecnológica de Minas Gerais

Resumo Várias proteínas possuem a sua estrutura tridimensional armazenada no *Protein Data Bank* (PDB). Dentre os *softwares* que trabalham com informações extraídas do PDB está o LSQKAB. Este *software* calcula a diferença de distâncias entre pares de átomos de duas estruturas proteicas. O seu uso torna-se lento ao comparar diversos registros. Neste trabalho, desenvolvemos uma metodologia baseada em matriz de distâncias, visando o agrupamento de registros, conforme as similaridades dos valores de distâncias atômicas. Avaliamos a precisão desta metodologia, utilizando o LSQKAB como referência, bem como avaliamos o tempo de execução e o consumo de memória RAM. Os experimentos foram realizados com arquivos de interações de resíduos de uma mesma proteína e os resultados foram comparados com o *atomic Cutoff Scanning Matrix* (aCSM) e com a técnica de busca da ferramenta *Residue Interaction Database* (RID). A metodologia apresentada obteve resultados interessantes diante das outras técnicas, obtendo por exemplo, uma maior precisão.

Palavras-chave. Agrupamentos, LSQKAB, Matriz de Distâncias, Proteínas

1 Introdução

As proteínas são as macromoléculas mais abundantes nas células vivas. Conforme [6], várias proteínas têm sua estrutura tridimensional resolvida e armazenada no banco de dados biológico *Protein Data Bank* (PDB), através de arquivos de texto.

O LSQKAB [1] é um *software* que trabalha com informações extraídas do PDB. O seu funcionamento é baseado no algoritmo de Kabsch [3], através da minimização do desvio quadrado médio da raiz (RMSD), de duas estruturas. Esta ferramenta escreve em um arquivo chamado de "deltas", as diferenças de distâncias entre os átomos comparados [1].

Devido ao tempo de execução, não é indicado usar o LSQKAB para comparar todos os registros de uma base de dados. Deste modo, a ferramenta *Residue Interaction Database* (RID), descrita em [2], utiliza o LSQKAB para sobrepor todos os registros da base de dados contra um registro de referência, gerando um arquivo delta para cada arquivo sobreposto.

[1] otavianomartins@hotmail.com
[2] sandrord@cefetmg.br
[3] tsouza@decom.cefetmg.br

O *Cutoff Scanning Matrix* (CSM), apresentado em [4] e suas variações, como o aCSM, descrito em [5], geram assinaturas para grafos biológicos. Estas assinaturas são geradas com base no padrão de distâncias dos átomos de uma estrutura proteica, extraída do PDB.

Neste trabalho, desenvolveu-se uma metodologia baseada em matriz de distâncias, inspirada no CSM. Esta metodologia possibilita agrupar interações do PDB com base nas distâncias atômicas, de uma forma eficiente e com um tempo hábil. Os experimentos foram desenvolvidos através de interações de pontes dissulfeto, ilustradas pela figura 1.

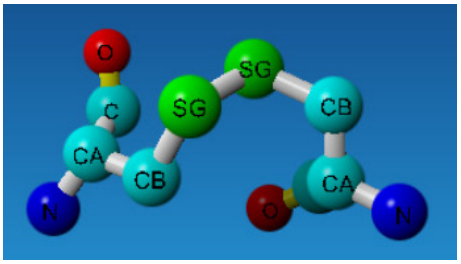

Figura 1: Formação de uma ponte dissulfeto, obtida pelos átomos de Enxofre Gama (SG), que se ligam através dos átomos de Carbono Beta (CB)
[2]

Através da figura 1, 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 (N), Carbono Alfa (CA), Carbono (C) e Oxigênio (O) de cada lado são estabilizadas.

As estruturas avaliadas neste trabalho foram extraídas do PDB pela ferramenta RID [2]. Cada registro contém a cadeia principal dos resíduos interagentes completa e apenas o último átomo do resíduo anterior e o primeiro átomo do resíduo posterior. A figura 2 mostra a estrutura de uma das interações da proteína "1EJG" (sua identificação no PDB).

A primeira linha do arquivo é referente à célula cristalográfica. As linhas 2, 3 e 4 indicam os operadores para a transformação de coordenadas. Abaixo dessas linhas, é iniciada a seção de coordenadas. Nesta seção, a primeira coluna refere-se ao tipo de registro desta linha. A segunda indica o número de série do átomo. A terceira mostra o símbolo do átomo. As próximas três colunas referem-se respectivamente ao nome do resíduo, cadeia e número de série. As colunas 7, 8 e 9 indicam as coordenadas dos átomos em Å 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. A última coluna mostra o símbolo do elemento que está alinhado à direita.

2 Objetivos

O objetivo geral é desenvolver uma metodologia baseada em matriz de distâncias, que possibilite o agrupamento de arquivos de interações. Os objetivos específicos são: formar grupos conforme as similaridades de distâncias atômicas dos registros, em um tempo hábil; Manter a precisão do LSQKAB; Comparar os resultados com o aCSM e com a RID.

DOI: 10.5540/03.2020.007.01.0369

```
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
```

Figura 2: Exemplo de um arquivo de interação da proteína "1EJG", obtido pela RID

3 Metodologia

Foram realizados experimentos com 16.383 arquivos de interações de pontes dissulfeto. Os resultados foram comparados com o aCSM, por ser o estado da arte na geração de padrões de grafos biológicos. Também foram feitas comparações com a técnica de busca da ferramenta RID, por ser um *software* especializado na análise de interações de proteínas.

A avaliação das metodologias foi feita por agrupamentos através do *K-Means Clustering*, definindo 500, 750 e 1.000 grupos. A precisão dos *clusters* foi verificada usando os resultados das sobreposições feitas com o LSQKAB, entre todos os registros que ficaram no mesmo grupo. Foi gerado um arquivo de deltas para cada comparação. Cada arquivo de deltas contém as diferenças de distâncias para cada par de átomo sobreposto. A tolerância definida foi de no máximo 0.5 Å por par. Deste modo, é garantido o RMSD de no máximo 0.5 Å, indicando uma forte similaridade entre os arquivos, conforme [8].

Também verificou-se, em todas as técnicas avaliadas, a média e o desvio padrão das distâncias dos registros em relação ao centróide de seu grupo. Ainda foi avaliado o tempo gasto por cada metodologia e o consumo de memória RAM.

4 Desenvolvimento

A metodologia da Matriz de Distâncias é ilustrada pela figura 3.

Cada uma das posições da figura 3, são referentes ao cálculo da distância euclidiana, para as coordenadas x, y e z, em relação a cada um dos outros 11 átomos do mesmo registro. Posteriormente, estes vetores foram unidos em uma matriz. Devido ao fato dos arquivos estudados possuírem o mesmo tamanho, a matriz construída teve 16.383 linhas

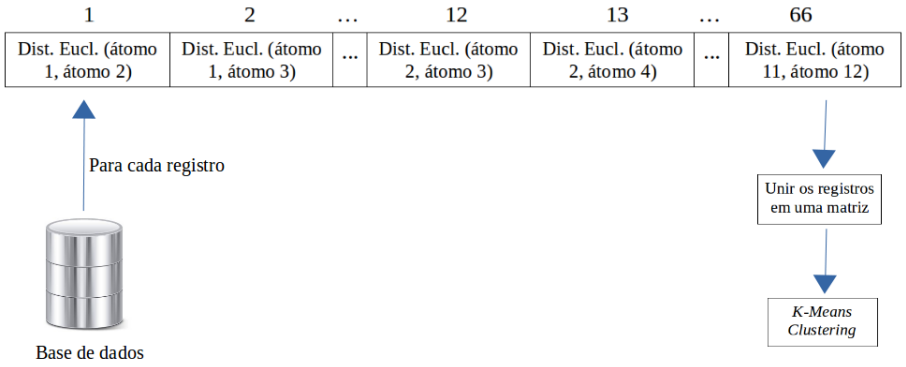

Figura 3: Fluxo da metodologia baseada em matriz de distâncias

e 66 colunas. Cada linha representa um arquivo de interação e cada coluna contém os respectivos valores de cada registro. Em seguida, foi realizado o agrupamento da matriz. A figura 4 ilustra os *clusters* 1 e 251 após o agrupamento.

Alguns registros do *cluster* 1

Id	1	2	...	32	33	34	…	65	66
9.243	1,3359	2,4545	...	1,3325	6,8214	5,5619	...	1,3331	2,2545
9.244	1,3367	2,4565	...	1,3339	6,8516	5,5894	...	1,3340	2,2576
9.248	1,3371	2,4592	...	1,3351	6,8803	5,6126	...	1,3344	2,2609

Alguns registros do *cluster* 251

Id	1	2	...	32	33	34	…	65	66
9.243	1,3398	2,4548	...	1,3232	4,9544	4,5800	...	1,4415	2,3050
557	1,3335	2,4375	...	1,3265	4,9650	4,7786	...	1,3344	2,2571
8.515	1,3281	2,4422	...	1,3270	4,9247	4,8332	...	1,3306	2,2619

Figura 4: Exemplos de alguns *clusters*

Conforme a figura 4, a formação dos grupos ocorreu por pequenas diferenças de valores. Os arquivos destes *clusters* diferenciaram-se com maior intensidade nas posições 33 e 34.

5 Resultados

Os resultados obtidos neste trabalho, foram divididos nas subseções abaixo.

5.1 Taxa de aproveitamento por *Clusters* e Registros

A tabela 1 indica quantos registros estiveram em grupos que obtiveram determinadas taxas de acertos. A primeira coluna contém as faixas de acertos. Em seguida, é exibido quantos grupos estiveram nestas faixas de acertos para as metodologias comparadas, nos experimentos com 500, 750 e 1.000 grupos. Estes valores foram baseados na média obtida após 30 execuções e foram arredondados para o número inteiro mais próximo.

DOI: 10.5540/03.2020.007.01.0369

Tabela 1: Quantidade média de registros por taxa de aproveitamento dos *clusters*

Técnica		Matriz de Distâncias			RID			aCSM		
	Nº de clusters	500	750	1.000	500	750	1.000	500	750	1.000
Taxa de Acertos	0% até 10%	442	345	281	564	297	347	2.673	2.031	1.583
	10% até 20%	1.473	1.069	763	1.246	1.104	735	2.436	2.082	2.125
	20% até 30%	1.421	1.033	962	1.648	1.054	1.013	1.565	1.492	1.114
	30% até 40%	1.114	1.299	1.084	1.186	1.296	1.148	1.093	1.297	1.272
	40% até 50%	1.051	912	890	1.205	1.285	1003	705	735	1.007
	50% até 60%	1.389	1.023	904	1.798	1.053	1.026	861	848	745
	60% até 70%	1.311	1.017	1.070	729	1.008	842	956	720	788
	70% até 80%	1.058	1.319	969	1.147	1.128	892	1.134	1.202	959
	80% até 90%	2.209	1.582	1.382	1.810	1.589	1.482	2.217	2.059	1.703
	90% até 100%	4.915	6.784	8.078	5.049	6.568	7.894	2.743	3.897	5.087

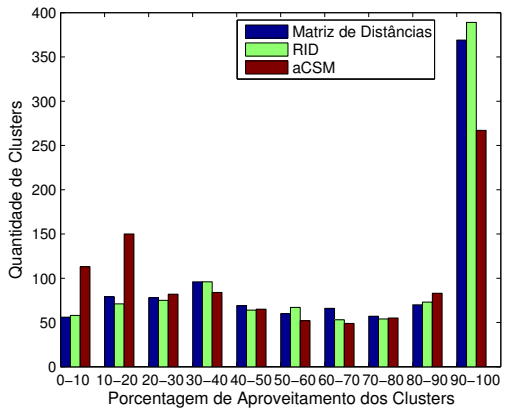

Figura 5: Aproveitamento por *clusters*

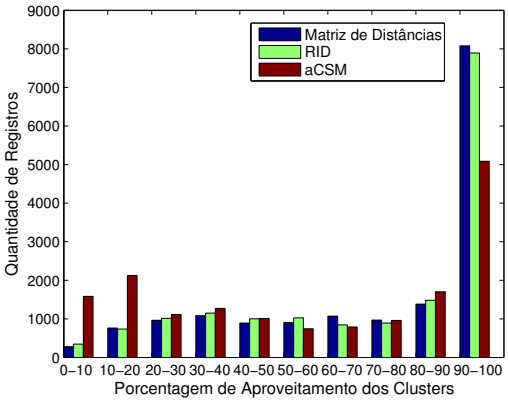

Figura 6: Aproveitamento por registros

Conforme indicado pela tabela 1, a metodologia da Matriz de Distâncias obteve, em todos os experimentos, mais registros em grupos que tiveram mais de 80% de sobreposições satisfatórias. Nos experimentos com 1.000 grupos, quase 9.500 registros estiveram nesta faixa. Ainda tendo, em média, 8.078 registros em grupos com taxas de acertos superiores à 90%. A ferramenta RID também obteve bons resultados, entretanto, um pouco inferiores aos da metodologia apresentada. O aCSM apresentou resultados abaixo das duas técnicas.

As figuras 5 e 6 ilustram os experimentos com 1.000 *clusters*. A figura à esquerda ilustra a quantidade de grupos por taxas de acertos. A figura à direita ilustra a tabela 1.

Conforme a figura 5, todas as técnicas avaliadas obtiveram uma maior quantidade de *clusters* na faixa de acerto superior à 90%. De acordo com a figura 6, a metodologia apresentada foi a que obteve mais registros na faixa de acertos superior a 90%.

A metodologia da matriz de distâncias obteve ainda a melhor taxa de acerto geral.

DOI: 10.5540/03.2020.007.01.0369

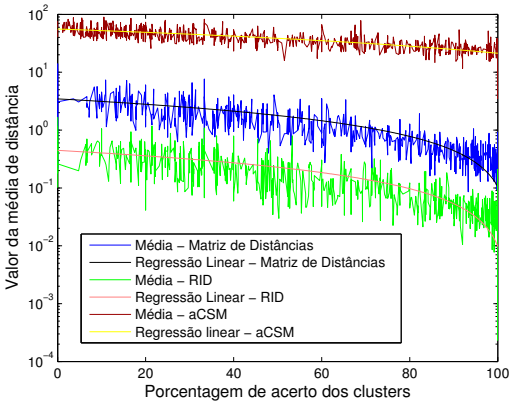

Figura 7: Média e regressão linear

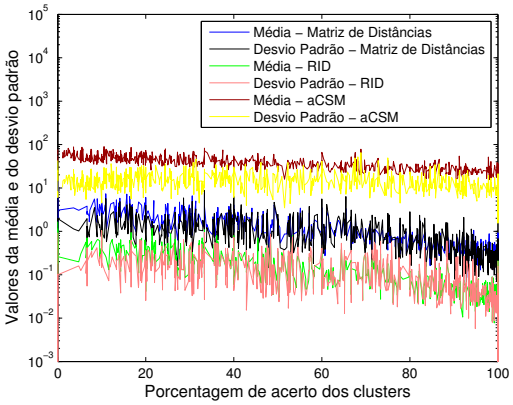

Figura 8: Média e desvio padrão

5.2 Média, Regressão Linear e Desvio Padrão

A média e regressão linear (figura 7), indicam a relação entre a taxa de aproveitamento dos grupos (eixo x), para o valor médio da distância dos registros em relação ao centróide de seu *cluster* (eixo y). O desvio padrão é ilustrado pela figura 8.

Conforme a figura 7, a regressão linear indica que em todas as técnicas avaliadas, os *clusters* que tiveram as maiores taxas de acertos geralmente possuíram registros próximos dos centróides. A metodologia apresentada e, principalmente, a RID, obtiveram baixos valores de distâncias. O mesmo comportamento ocorreu com o desvio padrão (figura 8).

5.3 Complexidade dos Algoritmos, Tempo de Execução e Memória RAM

A metodologia baseada em matriz de distâncias tem a complexidade de $O(n^2/2)$ para transformar um arquivo de interação de "n" átomos em um vetor. O aCSM tem a complexidade de $O(n^2)$ para transformar um registro no formato de grafo. $O(p * n^2)$ para obter o padrão de distâncias de um registro, sendo "p" a quantidade de passos. A RID utiliza o LSQKAB para realizar sobreposições. De acordo com [7], a complexidade é $O(n * m)$.

Conforme a tabela 2, a metodologia apresentada foi a mais rápida na construção dos dados e obteve o menor tempo total. A RID consumiu menos memória RAM.

Tabela 2: Tempo de execução e uso de memória RAM

Técnica	Matriz de Distâncias			RID			aCSM		
Nº de clusters	500	750	1.000	500	750	1.000	500	750	1.000
Tempo const. os dados	0:46	0:46	0:46	4:40	4:40	4:40	4:00	4:00	4:00
Tempo para agrupar	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
% de RAM consumida	22	23,5	25	20,6	22,3	24	21	22,6	24,5

DOI: 10.5540/03.2020.007.01.0369

6 Conclusões

Os resultados foram satisfatórios. A metodologia da matriz de distâncias mostrou uma maior precisão na formação dos grupos, além de um baixo tempo de execução e facilidade na construção da matriz. Deste modo, esta metodologia demonstrou ser eficiente na verificação e agrupamento de arquivos de interações, conforme as distâncias atômicas.

Agradecimentos

Agradecemos ao Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG) pelo suporte e bolsa de doutorado. Agradecemos também à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela bolsa durante o mestrado.

Referências

[1] CCP4. *LSQKAB (CCP4: Supported Program)*. CCP4, Oxford, 2019.

[2] S. R. Dias, R. C. Garrat and R. A. P. Nagem. The Use of a Residue-Residue interaction database for a Enggineering of Mutants Enzymes. In ENAPEBI 2012 - Ciencia sem Fronteiras - Encontro de Pesquisa em Bioquimica e Imunologia, *UFMG.*, 2012.

[3] W. Kabsch. A solution for the best rotation to relate two sets of vectors. In Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography, *International Union of Crystallography*, A32:922-923, 1976. DOI: 10.1107/S0567739476001873

[4] D. E. V. Pires, R. C. M. Minardi, M. A. Santos, C. H. Silveira, M. M. Santoro and W. Meira. Cutoff Scanning Matrix (CSM): structural classification and function prediction by protein inter-residue distance patterns. In BMC genomics, *BioMed Central.*, 12.4:S12, 2011. DOI:10.1186/1471-2164-12-S4-S12

[5] D. E. V. Pires, R. C. M. Minardi, C. H. Silveira, F. F. Campos and W. Meira. aCSM: noise-free graph-based signatures to large-scale receptor-based ligand prediction. In Bioinformatics, *Oxford University Press.*, 29:855-861, 2013. DOI:10.1093/bioinformatics/btt058

[6] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne. The Protein Data Bank. In Nucleic Acids Research, 28.1:235-242, 2000. DOI: https://doi.org/10.1093/nar/28.1.235.

[7] T. Shibuya. Searching protein 3-D structures in linear time, In Journal of Computational Biology, *Mary Ann Liebert.*, 17.3:203-219, 2010. DOI: 10.1089/cmb.2009.0148.

[8] C. S. Tsai. An introduction to computational biochemistry. *Wiley Online Library.* 2003. DOI: 10.1002/0471223840.

