Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, v. 8, n. 1, 2021.

Trabalho apresentado no XL CNMAC, Evento Virtual - Co-organizado pela Universidade do Mato Grosso do Sul (UFMS).

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Desenvolvimento da MDRPCA para Verificar Similaridades de Proteínas

Otaviano M. Monteiro [1]
Sandro R. Dias [2]
Thiago S. Rodrigues [3]
CEFET-MG, Belo Horizonte, Minas Gerais

Resumo. Este trabalho mostra o desenvolvimento e os resultados de uma metodologia baseada em Matriz de Distâncias Reduzida através de um Ponto entre os Carbonos Alfas (MDRPCA). Esta metodologia foi desenvolvida para representar as coordenadas atômicas de interações de proteínas, podendo ser utilizada com o *K-Means Clustering* para agrupar as interações proteicas. Os resultados obtidos foram comparados com MDC, MDRCCA, RID e aCSM. A MDRPCA apresentou eficácia e desempenho computacional satisfatórios.

Palavras-chave. MDRPCA, Interações de Proteínas, *K-Means Clustering*

1 Introdução

As pontes dissulfeto são estruturas covalentes. Elas são formadas pelos átomos de enxofre das cisteínas, que após oxidarem tornam-se cistinas. Essas ligações mantêm a estabilidade conformacional (estrutura tridimensional) de uma proteína [5].

O estudo das proteínas pode trazer melhorias para estas macromoléculas, através de propostas de mutações de um aminoácido específico ou de um conjunto de aminoácidos. Estas propostas de mutações podem aumentar ou diminuir a estabilidade, bem como a flexibilidade da proteína [2].

As informações das estruturas tridimensionais de diversas proteínas estão armazenadas em bancos de dados biológicos, através de arquivos de texto, como o Protein Data Bank (PDB) [1]. Dentre estas informações estão as interações de pontes dissulfeto.

Dentre os *softwares* que trabalham com arquivos de texto no formato do PDB, está o LSQKAB [11]. Este *software* tem o seu funcionamento baseado no algoritmo de Kabsch [6], através da minimização do desvio quadrado médio da raiz (RMSD) de duas estruturas proteicas. O LSQKAB também retorna as distâncias entre cada par de átomos comparados.

Existem outros *softwares* que trabalham com proteínas no formato do PDB, um exemplo é a MDC [7]. Esta metodologia permite a representação de interações de proteínas e o agrupamento com o *K-Means Clustering*. Para avaliar a precisão da MDC foram realizados experimentos com uma base de dados de pontes dissulfeto, extraídos do PDB pela ferramenta RID [3]. Os resultados da MDC foram comprados com a aCSM [10] e com o estratégia de busca da ferramenta RID [3]. A MDC obteve resultados satisfatórios no cenário comparado, alcançando uma maior precisão.

A forma de trabalho da MDC torna necessário o cálculo da distância Euclidiana entre todos os átomos de uma interação proteica [7]. Visando um ganho de processamento mas sem perder

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

precisão considerável, foi desenvolvida uma Matriz de Distâncias Reduzida cujos Centroides são os Carbonos Alfas (MDRCCA) [8]. A MDRCCA realiza apenas os cálculos de distância Euclidiana dos átomos considerados centroides (carbono alfa) em relação aos demais átomos não centroides. Deste modo, a MDRCCA trabalha com aproximadamente 30,30% do volume de informações da MDC. A MDRCCA apresentou um desempenho computacional satisfatório, mas a precisão foi inferior à MDC e RID.

Com o objetivo de melhorar a performance computacional da MDRCCA, bem como a sua precisão, desenvolvemos neste trabalho a Matriz de Distâncias Reduzida através de um Ponto entre os Carbonos Alfa (MDRPCA). A MDRPCA realiza os cálculos da distância Euclidiana de um ponto entre os carbonos alfas para os demais átomos da ponte dissulfeto. Esta metodologia trabalha com apenas 18,18% do volume de dados da MDC, obtendo assim a melhor performance computacional das metodologias comparadas, além de obter precisão superior ao aCSM, MDRCCA e RID no cenário comparado.

2 Objetivos

O objetivo geral deste trabalho consiste em desenvolver a MDRPCA para aperfeiçoar a representação e o agrupamento de interações de proteínas. Espera-se que esta nova metodologia obtenha performance computacional e precisão satisfatórias.

3 Metodologia

Os experimentos foram realizados com a mesma base de dados utilizada em [7], contendo 16.383 arquivos de interações de pontes dissulfeto. Cada interação contém os dois lados interagentes que formam a ponte dissulfeto. Cada lado contém a cadeia principal, um átomo do resíduo anterior à cadeia principal e um átomo do resíduo posterior à cadeia principal. Como exemplo, uma das interações da proteína "1EJG" é ilustrada pela Figura 1.

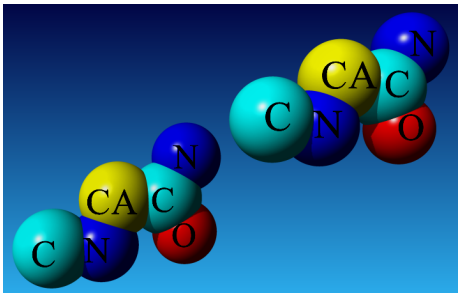
```
CRYST1   40.824   18.498   22.371  90.00  90.47  90.00 P 1 21 1      2
SCALE1      0.024495  0.000000  0.000201        0.00000
SCALE2      0.000000  0.054060  0.000000        0.00000
SCALE3      0.000000  0.000000  0.044702        0.00000
ATOM    1  C   THR A   2      14.172  10.757   7.221  1.00  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 1: Exemplo de um arquivo de interação da proteína "1EJG", obtido pela RID

De acordo com a Figura 1, a primeira linha do arquivo refere-se à célula cristalográfica. Os

operadores para a transformação de coordenadas são indicados pelas linhas 2, 3 e 4. Abaixo dessas linhas, inicia-se a seção de coordenadas. Nesta seção, a primeira coluna é referente ao tipo de registro desta linha. A segunda indica o número de série do átomo. A terceira mostra o símbolo atômico. 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 do local. A última coluna mostra o símbolo do elemento alinhado à direita [1] [7]. A Figura 2 ilustra a visualização tridimensional deste arquivo.

Figura 2: Visualização tridimensional da interação "1ejg_CYS-3-A_CYS-40-A_mc6.pdb"



A Figura 2 ilustra os dois lados interagentes que formam o arquivo de interação. O carbono mais à esquerda (C) é um átomo do resíduo anterior à cadeia principal do lado esquerdo. Os próximos átomos deste lado formam esta cadeia principal. São eles: nitrogênio (N), carbono alfa (CA), carbono (C) e oxigênio (O). O último nitrogênio (N) do lado esquerdo é um átomo do resíduo posterior à cadeia principal. O lado direito do arquivo de interação também contém 6 átomos, que seguem o mesmo padrão do lado esquerdo.

Os resultados foram comparados com a MDC e a MDRCCA, por terem obtidos resultados satisfatórios em [7] e [8]. Também foram realizadas comparações com a aCSM por ser uma das versões da [9], que é considerada o estado da arte na geração de padrões de grafos biológicos e com a ferramenta RID, por ser um *software* especializado na análise de interações de proteínas.

Os experimentos consistiram em realizar agrupamentos para cada técnica de representação de proteínas, com o *K-Means Clustering*, definindo 500, 750 e 1.000 grupos. A acurácia dos grupos formados foi verificada através de sobreposições atômicas feitas com o LSQKAB. Estas sobreposições ocorreram entre todos os registros que estiveram no mesmo grupo. Cada registro foi comparado duas vezes com todas as outras pontes dissulfeto de seu *cluster*. Na primeira comparação, um destes registros permaneceu "fixo"e o outro foi alinhado ao mesmo. Na segunda comparação, ocorreu o oposto. O limite definido foi de 0.5 Åpara cada par de aminoácido sobreposto.

4 Desenvolvimento

O fluxograma da MDRPCA é ilustrado pela Figura 3.

Conforme a Figura 3, para cada arquivo de interação contido na base de dados, calcula-se primeiro o ponto entre os dois carbonos alfas deste arquivo. Em seguida, calcula-se a distância Euclidiana de todos os demais átomos para este ponto. A Figura 4 ilustra alguns *clusters* formados por esta metodologia. Na ilustração da Figura 3, o vetor começou na posição 1, por ser a forma que o Matlab trabalha com vetores [4].

De acordo com a Figura 4, ao comparar alguns registros dos *clusters* 1 e 55, pode-se observar uma maior diferença nos valores das duas primeiras posições de cada grupo.

DOI: 10.5540/03.2021.008.01.0394

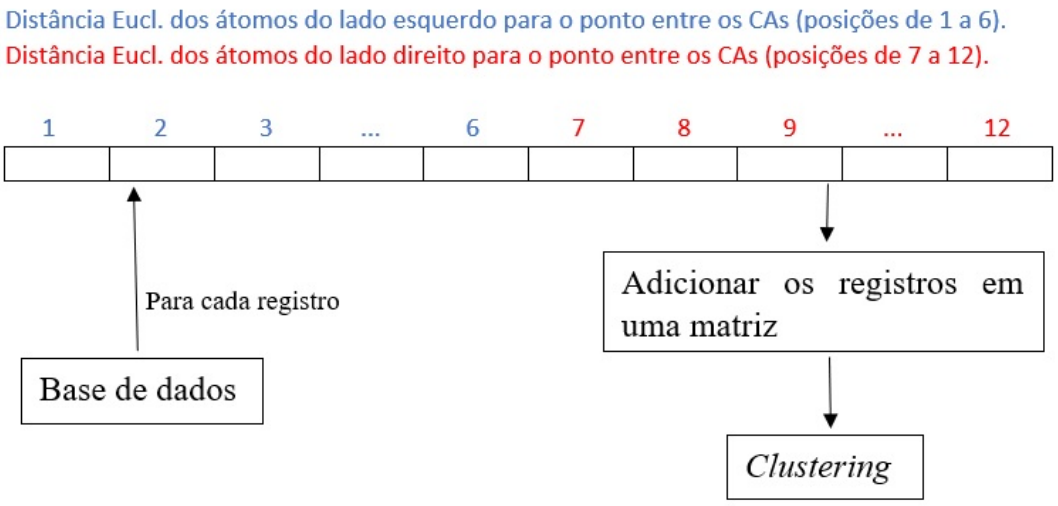


Figura 3: Fluxograma da MDRPCA

Registros contidos no *cluster* 1

Id	1	2	…	5	6	7	…	11	12
6.135	4,441	3,619	…	4,705	5,600	4,242	…	4,702	4,594
6.136	4,447	3,630	…	4,694	5,633	4,236	…	4,733	4,595
11.290	4,428	3,797	…	4,397	5,428	5,595	…	4,634	4,835
11.291	4,412	3,789	…	4,411	5,422	4,577	…	4,641	4,838

Registros contidos no *cluster* 55

Id	1	2	…	5	6	7	…	11	12
11.071	2,966	2,190	…	4,750	5,027	4,034	…	4,355	4,385
11.075	3,027	2,222	…	4,728	5,018	4,043	…	4,320	4,424
11.077	3,073	2,299	…	4,739	5,001	4,093	…	4,281	4,453
1.393	3,147	2,135	…	4,586	5,140	4,311	…	4,074	4,980

Figura 4: Exemplos de alguns *clusters*

5 Resultados

Os resultados obtidos ao definir 1.000 *clusters* estão contidos na Tabela 1. A primeira coluna indica as faixas de aproveitamento. À partir da segunda coluna estão as quantidades de registros que ficaram em *clusters* com estas respectivas taxas de aproveitamento. Estes valores referem-se à média, arredondando para o número inteiro mais próximo, considerando 31 execuções.

De acordo com a Tabela 1, a MDC e a RID apresentaram mais registros na faixa superior a 90%. No entanto, considerando a faixa de sobreposições aceitas entre 70% e 100%, a MDRPCA obteve uma quantidade próxima à estas duas metodologias. A MDRCCA e a aCSM obtiveram menos registros nestas melhores faixas de acertos. A Figura 5 ilustra graficamente a Tabela 1 e a Figura 6 ilustra o *boxplot* considerando as porcentagens totais de sobreposições aceitas.

Conforme a Figura 5, a MDRPCA obteve a maior parte dos seus registros em *clusters* com

DOI: 10.5540/03.2021.008.01.0394

Tabela 1: Quantidade de registros em *clusters* conforme a faixa de aproveitamento

Aproveitamento	MDRPCA	MDRCCA	MDC	RID	aCSM
0% até 10%	286	318	281	347	1.583
10% até 20%	856	968	763	735	2.125
20% até 30%	1.021	1.105	962	1.013	1.114
30% até 40%	1.175	1.162	1.084	1.148	1.272
40% até 50%	976	947	890	1.003	1.007
50% até 60%	934	1.009	904	1.026	745
60% até 70%	1.101	1.110	1.070	842	788
70% até 80%	1.396	1.579	969	892	959
80% até 90%	1.942	2.088	1.382	1.482	1.703
90% até 100%	6.696	6.097	8.078	7.894	5.087

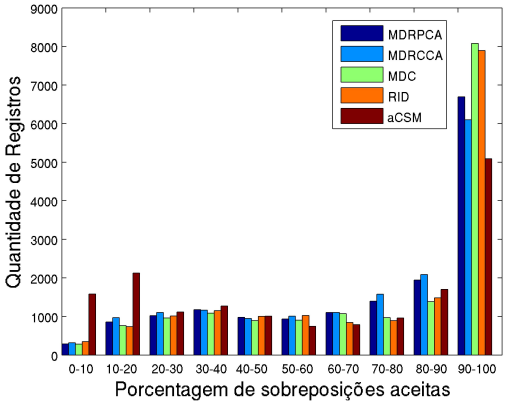


Figura 5: Aproveitamento por registros

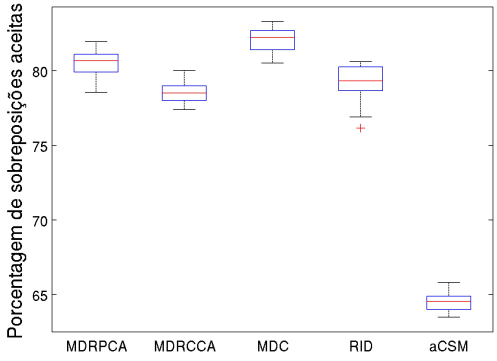


Figura 6: *Boxplot* com o aproveitamento total

aproveitamentos entre 70% e 100%. De acordo com a Figura 6, a MDC apresentou uma maior porcentagem de sobreposições aceitas, ao considerar as sobreposições realizadas em todos os *clusters*, de cada execução. Os resultados da MDRPCA e da RID ficaram próximos da MDC, tornando necessário o teste de Tukey para analisar se ocorreram diferenças significativas no aproveitamento total destas técnicas. A Figura 7 ilustra o teste de Tukey, definindo o nível de confiança com 99.

Conforme a Figura 7, a MDC obteve a maior porcentagem de sobreposições satisfatórias, obtendo vantagem significativa para as demais metodologias. A MDRPCA superou as outras 3 metodologias neste experimento. Os resultados destas análises ao considerar 500 e 750 grupos tiveram comportamentos similares destes experimentos com 1.000 *clusters*, mas as porcentagens de sobreposições aceitas foram menores.

A próxima análise é sobre a performance. A Tabela 2 ilustra o tempo médio de execução para construir os dados a serem agrupados, tempo para realizar os agrupamentos, tempo total e o ápice da porcentagem de consumo da memória RAM. Os dados de tempo nesta Tabela estão no formato de minutos:segundos. Os valores foram arredondados para o segundo mais próximo.

Conforme a Tabela 2, a MDRPCA foi a metodologia mais rápida na etapa de construção dos dados e no tempo total. Esta metodologia possibilitou a construção dos dados 9 vezes mais rapidamente do que a MDC. Além de obter um baixo consumo de memória RAM.

Após calcular o ponto entre o carbonos alfas, a MDRPCA tem a complexidade de $O(n)$ para

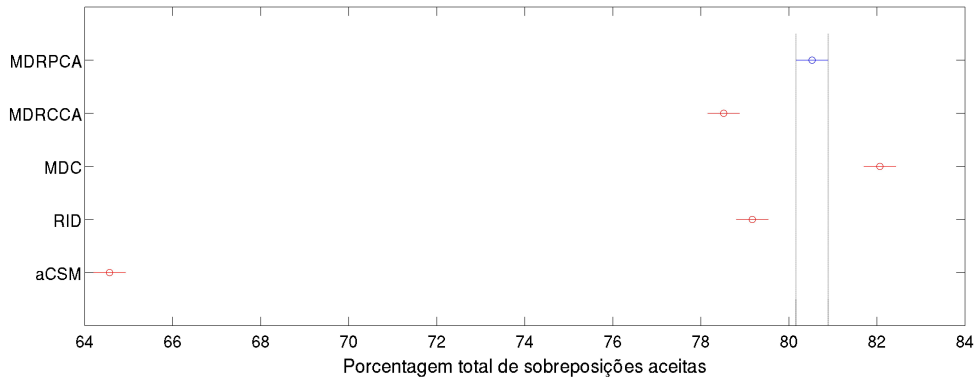


Figura 7: Teste de Tukey conforme a porcentagem sobreposições satisfatórias totais.

Tabela 2: Tempo para realização dos experimentos e consumo de memória RAM

Técnica	*Clusters*	Tempo const. dados	Tempo agrup.	Tempo total	% RAM
MDRPCA	500	00:01	00:26	00:27	6,7%
MDRCCA	500	00:02	00:36	00:38	7,0%
MDC	500	00:09	00:36	00:45	7,0%
RID	500	02:05	00:19	02:24	6,0%
aCSM	500	03:13	02:45	05:58	7,3%
MDRPCA	750	00:01	00:47	00:48	7,5%
MDRCCA	750	00:02	00:55	00:57	7,8%
MDC	750	00:09	00:55	00:54	7,8%
RID	750	02:05	00:31	02:36	6,8%
aCSM	750	03:13	03:10	06:23	8,3%
MDRPCA	1.000	00:01	01:12	01:13	8,3%
MDRCCA	1.000	00:02	01:34	01:36	8,6%
MDC	1.000	00:09	01:30	01:39	8,6%
RID	1.000	02:05	00:56	03:01	7,6%
aCSM	1.000	03:13	04:15	07:28	9,0%

transformar um arquivo de interação de “n” átomos em um vetor. A MDC tem a complexidade de $O((n^2 - n)/2))$ para realizar o mesmo procedimento.

6 Conclusões

Os resultados da MDRPCA sugerem que esta metodologia apresenta precisão e performance satisfatórias. A precisão da MDRPCA foi superior à MDRCCA, RID e aCSM. No entanto, a MDC apresentou precisão superior à MDRPCA.

A maior contribuição da MDRPCA é sua performance computacional. Esta metodologia foi consideravelmente mais rápida do que as demais técnicas avaliadas, além de ter um baixo consumo de memória RAM.

Os resultados obtidos por esta técnica sugerem utilizá-la em novos experimentos com outras bases de dados de interações de proteínas, principalmente com bases maiores, explorando o seu

DOI: 10.5540/03.2021.008.01.0394

desempenho computacional.

7 Agradecimentos

Agradecemos ao CEFET-MG pela estrutura disponibilizada e pela bolsa de estudos ofertada no início desta pesquisa. Também agradecemos à CAPES pela atual bolsa de doutorado.

Referências

[1] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. and Bourne, P. E. *The Protein Data Bank*, *Nucleic Acids Research*, 28:235-242, 2000. DOI: 10.1093/nar/28.1.235.

[2] Dias, S. R. *Residue interaction database: proposição de mutações sítio dirigidas com base em interações observadas em proteínas de estrutura tridimensional conhecida*, Tese de Doutorado, UFMG, 2012.

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

[4] Eddins, S. and Shure, L. *Matrix Indexing in MATLAB*. Mathworks, 2001. Disponível em: <https://www.mathworks.com/company/newsletters/articles/matrix-indexing-in-matlab.html/>. Acesso em: 29 de abr. 2021.

[5] Hunter, L. *Artificial intelligence and molecular biology, volume 445*. American Association for Artificial Intelligence, Califórnia, 1993.

[6] 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*, 32:922-923, 1976. DOI:10.1107/S0567739476001873.

[7] Monteiro, O. M., Dias, S. R. e Rodrigues, T. S. Desenvolvimento de uma Metodologia Baseada em Matriz de Distâncias para a Verificação de Similaridades de Proteínas, *Proceeding Series of the Brazilian Society of Computational and Applied Mathematics*, volume 7, 2020. DOI: 10.5540/03.2020.007.01.0369.

[8] Monteiro, O. M., Dias, S. R and Rodrigues, T. S. *Reduced Distance Matrix to Verify the Similarity Between Protein Structures*, *Brazilian Archives of Biology and Technology*, 2021.

[9] Pires, D. E. V., Minardi, R. C. M., Santos, M. A., Silveira, C. H., Santoro, M. M. and Meira, W. *Cutoff Scanning Matrix (CSM): structural classification and function prediction by protein inter-residue distance patterns. BMC genomics*, 12.4.1-11, 2011. DOI:10.1186/1471-2164-12-S4-S12.

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

[11] Winn, M. D., Ballard, C. C., Cowtan, K. D., Dodson, E. J., Emsley, P., Evans, P. R., Keegan, R. M., Krissinel, E. B., Leslie, A. G. W., McCoy, A., McNicholas, S. J., et al. *Overview of the CCP4 suite and current developments. Acta Crystallographica*, 67:235–242, 2011. DOI: 10.1107/S0907444910045749.

