IMPLEMENTAÇÃO DE NOVOS TEMPLATES E DE UMA NOVA MATRIZ DE SUBSTITUIÇÃO PARA ATUALIZAÇÃO DO GASS-WEB

João V. S. Ribeiro [1] (IC), Sandro C. Izidoro (PQ) [1]
[1] *Universidade Federal de Itajubá - Campus Theodomiro Carneiro Santiago.*

Palavras-chave: Bioinformática. Predição de função de proteínas. Sítios catalíticos. Sítios de ligação.

Introdução

O Genetic Active Site Search (GASS) (IZIDORO; MELO-MINARDI; PAPPA, 2015) foi modelado para utilizar informações estruturais de um sítio ativo *template* (modelo) na busca de proteínas com sítios ativos similares. O método pode encontrar sítios ativos com resíduos em cadeias diferentes e é capaz de lidar com mutações conservativas, além de não impor quaisquer restrições quanto ao número de resíduos no sítio ativo e a distância entre eles.

Posteriormente, o GASS foi disponibilizado na forma de um servidor web (GASS-WEB[1]) (MORAES et al., 2017) quando, além de utilizar *templates* baseados em informações do Catalytic Site Atlas (CSA) (PORTER; BARTLETT; THORNTON, 2004) e dados do NCBI-VAST (GIBRAT; MADEJ; BRYANT, 1996; PANCHENKO; MADEJ,2004), apresentou uma novidade em relação ao original: busca por sítios de ligação.

O GASS-WEB utiliza de *templates* de sítios ativos para encontrar proteínas com sítios ativos similares. Supondo uma dada proteína com um sítio ativo formado por três resíduos de aminoácidos: HIS, 57, E; ASP, 102, E; SER, 195, E (nome do resíduo de aminoácido, posição na sequência e cadeia). O GASS-WEB utiliza essa informação para encontrar um sítio ativo similar em uma outra proteína alvo. Mutações conservativas podem aparecer (e.g., serina por alanina), e GASS-WEB pode lidar com essa situação utilizando uma matriz de substituição que lhe informa quando pode trocar um resíduo de aminoácido por outro para efetuar a busca. Essa matriz de substituição é montada com base nas mutações conservativas observadas nas proteínas homólogas à proteína usada como template.

Este projeto pretende utilizar a nova versão do CSA (RIBEIRO et al., 2017), agora M-CSA, e também a nova versão do Biolip (base de dados de sítios de ligação - *binding*) (ZHANG et al., 2023) para atualizar os *templates* e a matriz de substituição do GASS-WEB. Além disso, outras melhorias no servidor serão implementadas, como a otimização da exibição dos resultados.

Para a execução do projeto serão necessárias novas implementações em Python para filtrar e atualizar os novos templates, gerar uma nova matriz de substituição, bem como otimizar as páginas de resultados do servidor.

Metodologia

A atualização dos *templates* do GASS-WEB será feita conforme as novas versões do M-CSA, BioLip e PDB (BERMAN, 2000). Durante a atualização dos *templates*, serão gerados dados utilizando informações do Carbono Alfa (CA) e Last Heavy Atom (LHA) de cada resíduo de aminoácido do *template*. As mutações conservativas encontradas nos sítios ativos das proteínas homólogas serão adicionadas na matriz de substituição que o GASS-WEB utiliza nas buscas por sítios similares. A escolha dessas mutações será baseada na matriz Blosum62 (HENIKOFF; HENIKOFF, 1992). A Figura 1 ilustra o procedimento de atualização dos *templates* e da matriz de substituição, para a geração de uma base de dados que será utilizada pelo GASS-WEB..

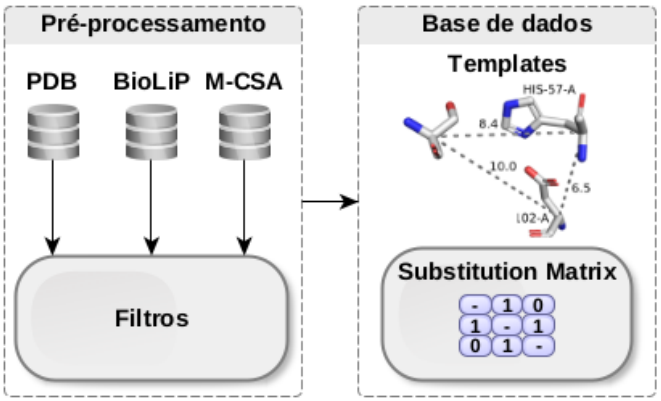
Inicialmente, procedemos com a coleta dos nomes das proteínas LIT presentes no M-CSA, bem como de seus homólogos, que são proteínas compartilhando características semelhantes com aquelas de interesse. As proteínas LIT são aquelas que possuem ampla notoriedade na literatura científica, sendo alvo de numerosos estudos e análises, conferindo-lhes, portanto, uma maior confiabilidade.

Posteriormente, realizamos a interseção entre as bases de dados do BioLip e as proteínas previamente obtidas do M-CSA. Isso resultou na geração dos templates por

[1] https://gass.unifei.edu.br/

meio de uma pesquisa na extensa base de dados redundante do BioLip, visando identificar os ligantes de interesse. Adicionalmente, promovemos a criação das mutações conservativas, comparando as ligações presentes nas proteínas do M-CSA com aquelas encontradas em suas contrapartes homólogas. Nesse processo, empregamos a matriz Blosum62 (Matriz de Substituição de Blocos de Aminoácidos) para validar as mutações. Essa etapa culminou na elaboração da matriz de substituição, a qual foi utilizada para a busca por sítios catalíticos e de ligação.

Figura 1 – Procedimentos de conferência e atualização dos templates e da matriz de substituição.



Cada busca no GASS-WEB reporta uma população final de respostas, uma vez que é baseado em um algoritmo genético. Essa característica resulta em um problema no GASS, no qual resultados redundantes podem ser gerados. Para solucionar essa questão, adicionamos uma rotina ao algoritmo, a qual é executada após a geração dos resultados. Essa rotina tem como objetivo eliminar permutações que possam prejudicar a busca, otimizando, assim, o algoritmo e aumentando as chances de que o resultado desejado esteja entre os 10 primeiros resultados na população final.

Todos os procedimentos serão executados utilizando as linguagens C/C++ e Python.

Resultados e discussão

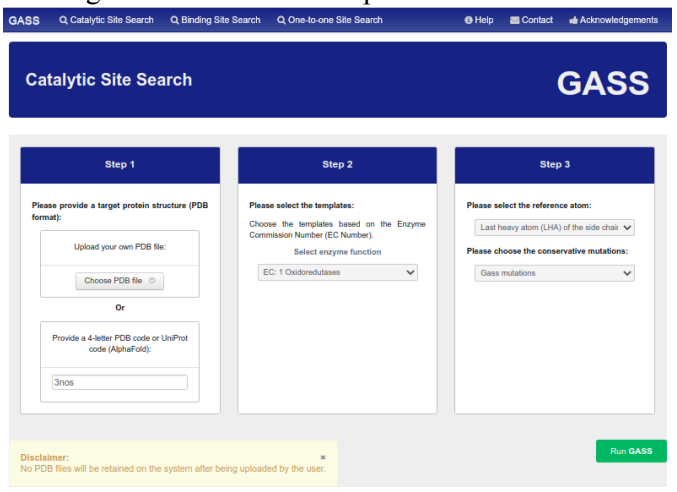
A Tabela 1 mostra a quantidade de *templates* antes e depois das operações de atualização. Quanto aos *templates* de sítios catalíticos, a nova versão tem menos *templates* e menos mutações. Isso acontece pois o M-CSA atualizou sua base de dados, retirando muitos *templates* redundantes. Em relação aos *templates* de sítio de ligação, o número aumentou pois o BioLip adicionou muitas estruturas homólogas na sua nova versão. As mutações conservativas tiveram seu número diminuído pois nessa versão o critério de corte utilizado com base na matriz Blosum62 foi 0 (zero).

Tabela 1 - Número de templates e mutações conservativas do GASS-WEB e do GASS-WEB 2.0.

	GASS-WEB	GASS-WEB 2.0
Templates de sítios catalíticos	1691	863
Mutações conservativas	1520	514
Templates de sítios de ligação	1158	2734
Mutações conservativas	567	180

Para a nova versão do GASS, um novo servidor web foi montado e configurado[2]. A Figura 2 mostra a tela de busca de sítios catalíticos.

Figura 2 – Tela da busca por sítios catalíticos.

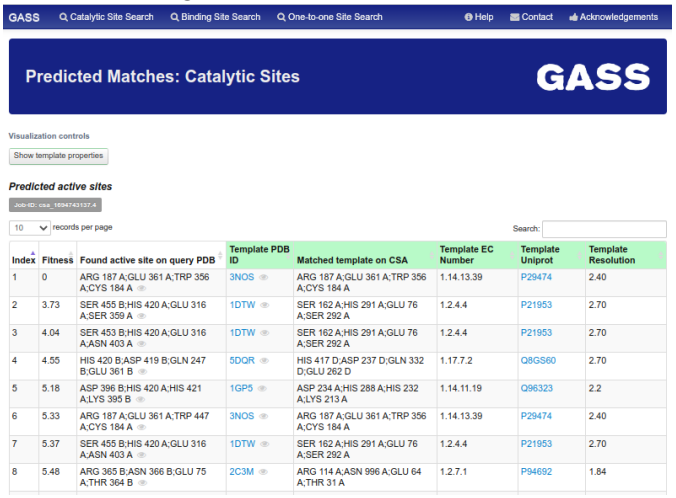


A Figura 3 mostra a tela de resultados para uma busca por sítios catalíticos Os resultados agora estão sem as duplicações de sítios em cadeias diferentes. Isso faz com que mais resultados diferentes e significativos apareçam logo nas primeiras posições da população final. Além disso, o GASS-WEB agora conta com uma nova página de visualização dos resultados na estrutura proteica. A Figura 5 mostra uma estrutura proteica com os resíduos

[2] https://gass2.unifei.edu.br/

de aminoácidos encontrados (destacados em vermelho) e uma lista de resultados na parte inferior da janela. Dessa forma, o usuário do GASS-WEB pode escolher e visualizar todos os resultados para a busca efetuada.

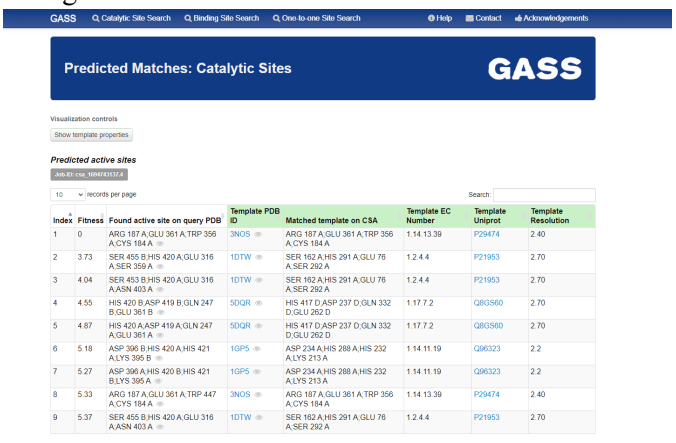
Figura 3 – Tela de resultados.



Index	Fitness	Found active site on query PDB	Template PDB ID	Matched template on CSA	Template EC Number	Template Uniprot	Template Resolution
1	0	ARG 187 A;GLU 361 A;TRP 356 A;CYS 184 A	3NOS	ARG 187 A;GLU 361 A;TRP 356 A;CYS 184 A	1.14.13.39	P29474	2.40
2	3.73	SER 455 B;HIS 420 A;GLU 316 A;SER 359 A	1DTW	SER 162 A;HIS 291 A;GLU 76 A;SER 292 A	1.2.4.4	P21953	2.70
3	4.04	SER 453 B;HIS 420 A;GLU 316 A;ASN 403 A	1DTW	SER 162 A;HIS 291 A;GLU 76 A;SER 292 A	1.2.4.4	P21953	2.70
4	4.55	HIS 420 B;ASP 419 B;GLN 247 B;GLU 361 B	5DQR	HIS 417 D;ASP 237 D;GLN 332 D;GLU 262 D	1.17.7.2	Q8GS60	2.70
5	5.18	ASP 396 B;HIS 420 A;HIS 421 A;LYS 395 B	1GP5	ASP 234 A;HIS 288 A;HIS 232 A;LYS 213 A	1.14.11.19	Q96323	2.2
6	5.33	ARG 187 A;GLU 361 A;TRP 447 A;CYS 184 A	3NOS	ARG 187 A;GLU 361 A;TRP 356 A;CYS 184 A	1.14.13.39	P29474	2.40
7	5.37	SER 455 B;HIS 420 A;GLU 316 A;ASN 403 A	1DTW	SER 162 A;HIS 291 A;GLU 76 A;SER 292 A	1.2.4.4	P21953	2.70
8	5.48	ARG 365 B;ASN 366 B;GLU 75 A;THR 364 B	2C3M	ARG 114 A;ASN 996 A;GLU 64 A;THR 31 A	1.2.7.1	P94692	1.84

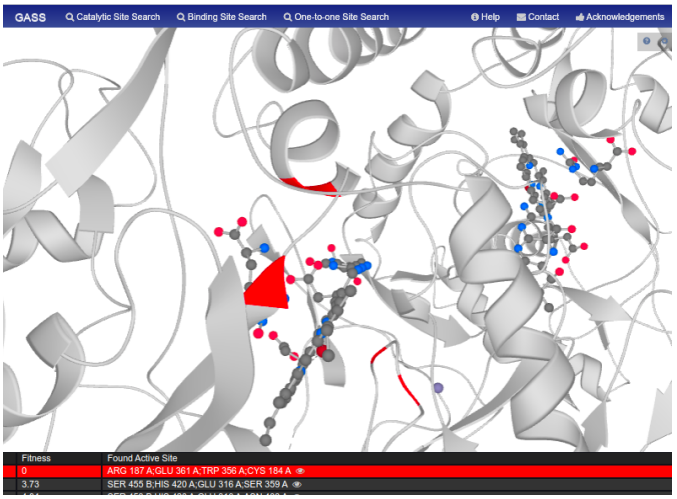
Ao comparar a figura 4 com a figura 3 podemos ver que o template 5DQR estava com o mesmo sítio porém em cadeias diferentes, portanto o algoritmo que trata os resultados após a execução do GASS retirou a linha com o maior fitness.

Figura 4 – Tela de resultados antes do tratamento.



Index	Fitness	Found active site on query PDB	Template PDB ID	Matched template on CSA	Template EC Number	Template Uniprot	Template Resolution
1	0	ARG 187 A;GLU 361 A;TRP 356 A;CYS 184 A	3NOS	ARG 187 A;GLU 361 A;TRP 356 A;CYS 184 A	1.14.13.39	P29474	2.40
2	3.73	SER 455 B;HIS 420 A;GLU 316 A;SER 359 A	1DTW	SER 162 A;HIS 291 A;GLU 76 A;SER 292 A	1.2.4.4	P21953	2.70
3	4.04	SER 453 B;HIS 420 A;GLU 316 A;ASN 403 A	1DTW	SER 162 A;HIS 291 A;GLU 76 A;SER 292 A	1.2.4.4	P21953	2.70
4	4.55	HIS 420 B;ASP 419 B;GLN 247 B;GLU 361 B	5DQR	HIS 417 D;ASP 237 D;GLN 332 D;GLU 262 D	1.17.7.2	Q8GS60	2.70
5	4.87	HIS 420 A;ASP 419 A;GLN 247 A;GLU 361 A	5DQR	HIS 417 D;ASP 237 D;GLN 332 D;GLU 262 D	1.17.7.2	Q8GS60	2.70
6	5.18	ASP 396 B;HIS 420 A;HIS 421 A;LYS 395 B	1GP5	ASP 234 A;HIS 288 A;HIS 232 A;LYS 213 A	1.14.11.19	Q96323	2.2
7	5.27	ASP 396 A;HIS 420 B;HIS 421 B;LYS 395 A	1GP5	ASP 234 A;HIS 288 A;HIS 232 A;LYS 213 A	1.14.11.19	Q96323	2.2
8	5.33	ARG 187 A;GLU 361 A;TRP 447 A;CYS 184 A	3NOS	ARG 187 A;GLU 361 A;TRP 356 A;CYS 184 A	1.14.13.39	P29474	2.40
9	5.37	SER 455 B;HIS 420 A;GLU 316 A;ASN 403 A	1DTW	SER 162 A;HIS 291 A;GLU 76 A;SER 292 A	1.2.4.4	P21953	2.70

Figura 5 – Visualização dos resultados na estrutura da proteína.



Conclusões

O projeto de atualização do Genetic Active Site Search (GASS-WEB) representa um avanço significativo na pesquisa de identificação de sítios ativos e de ligação em proteínas. Ao incorporar as novas versões do M-CSA e do BioLip, bem como otimizar o servidor web, a equipe demonstrou um compromisso em manter o GASS-WEB relevante e eficaz para a comunidade científica.

A metodologia empregada na atualização dos templates e na geração da matriz de substituição, baseada em informações do Carbono Alfa (CA) e Last Heavy Atom (LHA) de resíduos de aminoácidos, juntamente com a validação de mutações conservativas usando a matriz Blosum62, demonstra um rigor científico na seleção e atualização dos dados. Isso garante a precisão e a confiabilidade das buscas por sítios ativos e de ligação em proteínas.

A implementação de uma rotina para eliminar resultados redundantes, melhorando a eficiência do algoritmo genético, é uma adição valiosa que aprimora a experiência do usuário e reduz a permutação de resíduos nos resultados obtidos.

As melhorias na interface do usuário, com a eliminação de duplicações de sítios em cadeias diferentes e a introdução de uma nova página de visualização de resultados na estrutura proteica, tornam o GASS-WEB mais acessível e útil para os pesquisadores que desejam explorar os resultados de suas buscas de forma mais eficiente.

No geral, as melhorias na metodologia, na base de dados

e na interface do usuário garantem que o GASS-WEB continue a desempenhar um papel importante na pesquisa biomolecular e na descoberta de alvos terapêuticos.

Agradecimentos

Agradecemos à Universidade Federal de Itajubá (Unifei) pelo apoio institucional crucial ao longo deste trabalho de pesquisa. A Unifei proporcionou recursos e um ambiente acadêmico propício à realização deste estudo.

Também estendemos nossos agradecimentos ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo generoso financiamento que tornou possível a execução deste projeto. Seu suporte financeiro foi fundamental para o sucesso deste trabalho.

Por fim, gostaríamos de expressar nossa apreciação a todas as pessoas que, de alguma forma, contribuíram para o desenvolvimento deste projeto.

Referências

BERMAN, H. M. The Protein Data Bank. **Nucleic Acids Research**, v. 28, n. 1, p. 235–242, 1 jan. 2000.

GIBRAT, J.-F.; MADEJ, T.; BRYANT, S. H. Surprising similarities in structure comparison. **Current Opinion in Structural Biology**, v. 6, n. 3, p. 377-385, 1996. ISSN 0959-440X.

HENIKOFF, S.; HENIKOFF, J. G. Amino acid substitution matrices from protein blocks. **Proceedings of the National Academy of Sciences of the United States of America**, v. 89, n. 22, p. 10915–10919, 15 nov. 1992.

IZIDORO, S. C.; MELO-MINARDI, R. C. de; PAPPA, G. L. GASS: identifying enzyme active sites with genetic algorithms. **Bioinformatics**, Oxford University Press, v. 31, n. 6, p. 864 - 870, 2015.

MORAES, J. et al. GASS-WEB: a web server for identifying enzyme active sites based on genetic algorithms.**Nucleic Acids Research**, v. 45, n. W1, p. W315-W319, 04, 2017. ISSN 0305-1048.

PORTER, C. T.; BARTLETT, G. J.; THORNTON, J. M. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. **Nucleic Acids Research**, v. 32, n. suppl1, 2004.

RIBEIRO, A. et al. Mechanism and Catalytic Site Atlas (M-CSA): a database of enzyme reaction mechanisms and active sites. **Nucleic Acids Research**, v. 46, n. D1, p. D618-D623,11 2017. ISSN 0305-1048.

ZHANG C.; ZHANG X.; FREDDOLINO, P. L., AND ZHANG, Y. BioLiP2: an updated structure database for biologically relevant ligand-protein interactions, **Nucleic Acids Research**, gkad630 (2023).