V Simpósio de Iniciação Científica 2022
Ciência, Tecnologia e Inovação no Brasil

Implementação de novos *templates* e de uma nova matriz de substituição para atualização do GASS-WEB

Murillo Ventura Mendonça[1] (IC), Sandro Carvalho Izidoro (PQ)[1]
[1] *Universidade Federal de Itajubá.*

Palavras-chave: Algoritmo genético. Paralelismo. Proteínas. Sítios ativos.

Introdução

A ferramenta GASS (*Genetic Active Site Search*) foi modelada para utilizar informações estruturais de um sítio ativo *template* na busca de sítios ativos similares em outras proteínas. 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 para a comunidade científica na forma de um servidor (GASS-WEB).

Desde a implementação original do GASS em 2015, diversas proteínas tiveram suas estruturas elucidadas e anotadas no PDB, assim como novas anotações de sítios ativos foram adicionadas no *Catalytic Site Atlas* (CSA), principal base de dados para o funcionamento do GASS. Especialmente em 2017/2018, o CSA foi expandido com diversas anotações de mecanismos catalíticos, se tornando o *Mechanism and Catalytic Site Atlas* (M-CSA).

Além das atualizações dos sítios catalíticos, via M-CSA, uma atualização dos sítios de ligação foi realizada utilizando dados do BioLiP(YANG; ROY; ZHANG, 2012). O BioLiP é um banco de dados com curadoria semi-manual para interações ligante-proteína biologicamente relevantes, e tem atualização constante (última atualização: Abril 2022).

Estas atualizações e melhorias são propostas como objetivos deste trabalho, juntamente com uma atualização da interface e das funcionalidades das páginas do GASS-WEB.

Resultados e discussão

A Figura 1 apresenta a tela de busca de sítios catalíticos do GASS 2.0. No ***card*** *Step* 1, o usuário informa a proteína alvo da busca, podendo entrar com o código PDB ou mesmo um código UniProt, onde neste último caso, a estrutura da proteína será obtida da base de dados do AlphaFold (ALQURAISHI, 2019). No ***card*** *Step* 2 o usuário deve informar, com base no *EC Number*, a função da proteína e o número de resíduos no template. Ao fornecer esses dados o usuário está refinando a busca que o AG irá efetuar. E por fim, no ***card*** *Step* 3, o usuário escolhe o átomo utilizado como referência no template (carbono alfa ou o último átomo pesado da cadeia lateral) e se o GASS irá considerar mutações conservativas ou não, podendo utilizar a matriz de substituições do próprio GASS ou entrar com suas próprias mutações.

Para o GASS-Metal, o padrão de *cards* também foi utilizado. A diferença nas informações solicitadas estão com relação aos *templates* de sítios metálicos. Neste *card* (*Step* 2), é solicitado que o usuário informe o tipo de íon metálico, suas interações (se é apenas um único íon ou mais) e o número de resíduos do sítio.

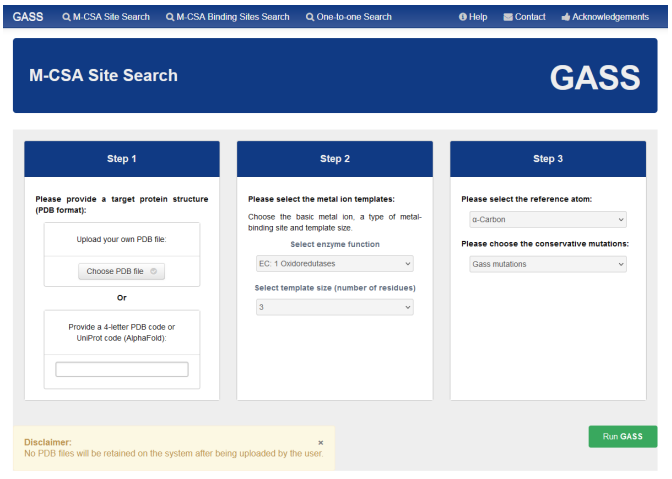

Figura 1 – *Cards* da tela de busca de sítios catalíticos do GASS 2.0.

A partir do M-CSA atualizado foi possível extrair dados de 741 novas proteínas, formando assim 820 novos *templates*. Cada um destes *templates* foi calculado tanto

usando o carbono alfa de cada resíduo como referência, quanto usando seu último átomo mais pesado (LHA).

Uma versão paralela do GASS, desenvolvida utilizando *threads* em C++ foi planejada e implementada com o objetivo de reduzir a penalidade de tempo de execução do método ao se adotar a busca por quadrantes antes de uma busca global. Esta implementação foi considerada bem sucedida já que obteve uma redução 33.98% no tempo de execução médio da ferramenta.

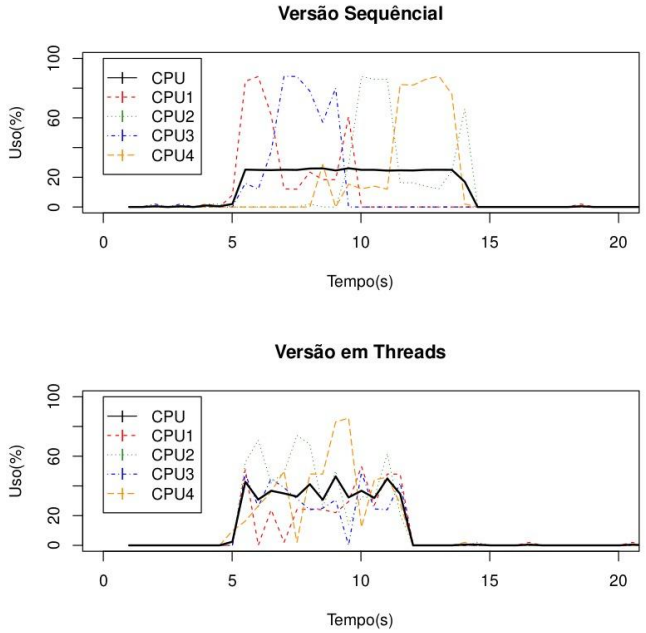

Figura 2 – Comparação da execução do AG sequencial e do AG paralelo.

Como visto na Figura 2, percebe-se que durante a execução da versão sequencial, o uso do processador a cada fatia de tempo está focado em apenas um dos núcleos do processador, com 79.3% de uso neste, representando cerca de 25% de uso total do processador para este processo. Este comportamento já era esperado, uma vez que apenas um processo referente ao GASS está em execução. Já na versão paralela em *threads*, percebe-se um perfil mais distribuído de carga de trabalho entre os núcleos, todos com porcentagem de uso significantes, tendo agora cerca de 40% de uso total do processador para esta versão.

Uma vez que as atualizações do GASS-WEB estava sendo feitas em paralelo com a implementação do GASS-Metal, o discente também participou do trabalho e do artigo do GASS-Metal (PAIVA et al., 2022a). O artigo foi publicado na revista *Briefings in Bioinformatics* (2022), que tem fator de impacto 13.994.

Além do artigo na *Briefings in Bioinformatics*, o discente também participou da publicação de outro artigo, agora na revista *Computers in Biology and Medicine* (PAIVA et al., 2022b), fator de impacto 6.698, onde foi responsável pela escrita das seções *Interactions at atomic/residue level* e *Databases*.

Conclusões

Ao final desta pesquisa, seus objetivos principais propostos foram alcançados, como a atualização dos *templates* do GASS utilizando a nova versão do CSA (M-CSA) e a reestruturação da página GASS-WEB. Embora não fosse prevista a implementação do modo paralelo para o GASS, esta também se mostrou bem sucedida e foi uma grande fonte de aprendizado.

Realizar esta pesquisa desencadeou diversas oportunidades para o discente e com elas experiências enriquecedoras, como participar da escrita de um artigo para uma grande revista, participar de projetos com outras instituições de ensino superior. Tanto a própria pesquisa, quanto às experiências proporcionadas por ela, fomentaram grande desejo de aprendizado e pesquisas futuras dentro desta área no discente, sendo de grande impacto na sua graduação e formação como engenheiro.

Agradecimento

O presente trabalho foi realizado com apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), sob forma de financiamento da pesquisa. Agradecimentos especiais também são reservados ao professor orientador Sandro Carvalho Izidoro pela disponibilidade e orientação constante durante toda a execução do projeto.

Referências

ALQURAISHI, M. AlphaFold at CASP13. Bioinformatics, Oxford University Press, v. 35, n. 22, p. 4862–4865, 2019.

IZIDORO, S. C.; MELO-MINARDI, R. C. de; PAPPA, G. L. GASS: identifying enzyme active sites with genetic algorithms. Bioinformatics, v. 31, n. 6, p. 864–870, 11 2014. ISSN 1367-4803. Disponível em: <https://doi.org/10.1093/bioinformatics/btu746>.

PAIVA, V. A. et al. GASS-Metal: identifying metal-binding sites on protein structures using genetic algorithms. Briefings in Bioinformatics, 05 2022. ISSN 1477-4054. Bbac178. Disponível em: <https://doi.org/10.1093/bib/bbac178>.

PAIVA, V. de A. et al. Protein structural bioinformatics: An overview. Computers in Biology and Medicine, p. 105695, 2022. ISSN 0010-4825. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0010482522004784>.

YANG, J.; ROY, A.; ZHANG, Y. BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions. Nucleic Acids Research, v. 41, n. D1, p. D1096–D1103, 10 2012. ISSN 0305-1048. Disponível em: <https://doi.org/10.1093/nar/gks966>.