Resenha Crítica: Uma Proposta de Algoritmo Memético Baseado em Conhecimento para o Problema de Predição de Estruturas 3-D de Proteínas

Correa, L. d. L. Uma proposta de algoritmo memético baseado em conhecimento para o problema de predição de estruturas 3-D de proteínas. Dissertação(Mestrado em Ciência da Computação). Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação. Porto Alegre. 2017.

A obra de 2017, intitulada por "Uma Proposta de Algoritmo Memético Baseado em Conhecimento para o Problema de Predição de Estrutura 3-D de Proteínas" se trata de uma dissertação de mestrado na área de Predição de Estrutura de Proteínas (PSP), cujo foco do autor está na exploração e na diversidade do modelo aplicado ao problema, através da implementação de duas metaheurísticas e a incorporação de informação de estruturas secundárias já conhecidas. No texto é apresentado desde os conceitos mais básicos envolvendo proteínas até a estruturação da proposta e os resultados alcançados.

O trabalho está dividido em 7 capítulos principais e um capítulo final dedicado às publicações e produção técnica do autor. Todos os capítulos seguem a mesma estrutura lógica, onde o autor explica claramente os termos relacionados ao capítulo em questão, bem como definições e aplicações, seguido da justificativa para a abordagem proposta, sempre deixando claro a relação do que está sendo visto com a proposta final. Além disso, todos os capítulos apresentam uma seção de “Resumo do Capítulo”, alinhando os conceitos abordados com o objetivo do capítulo.

No capítulo 1 o autor faz uma introdução ao problema do PSP e a área de Bioinformática Estrutural, relaciona o PSP a um problema NP-difícil, cita os principais métodos computacionais utilizados atualmente e termina justificando o uso de metaheurísticas. Já no capítulo 2 é apresentada a fundamentação biológica das proteínas e suas estruturas secundárias (ES), conceitos utilizados para definir a função de avaliação que será utilizada na proposta, composta por 3 valores: um valor referente a função de energia, um valor *SASA* (*Solvent Accessible Surface Area*) referente ao grau de exposição da proteína ao solvente e um termo de reforço de ES. É interessante notar que os capítulos 1 e 2 são destinados a esclarecer conceitos básicos relacionados ao problema, ao mesmo tempo que utiliza estas explicações para justificar as escolhas feitas na proposta. A linguagem e organização do texto é clara de forma a facilitar a compreensão dos termos apesar da complexidade, também prepara o leitor para a compreensão dos próximos capítulos.

Dando continuidade, o autor passa para o capítulo 3, onde inicia apresentando o CASP (*Critical Assessment of Protein Structure Prediction*), uma competição mundial de testes dos métodos para PSP, e deixa claro sua utilização como base de dados para determinar o estado da arte, sendo eles os métodos Rosetta e Quark. Em seguida apresenta os trabalhos relacionados seguindo uma ordem cronológica por metaheurística, de forma a justificar o uso de algoritmos meméticos, começando pelos algoritmos genéticos, uso de APL (Lista de Probabilidades Angulares), estratégias multiobjetivo, otimização multimodal e termina com os algoritmos meméticos. Aqui o autor não faz menção ao uso de alguma metodologia específica para a revisão de literatura, nem se foi feita alguma revisão de literatura, apenas cita alguns trabalhos relacionados conforme as metaheurísticas vão sendo apresentadas. Após justificar o uso de algoritmos meméticos o autor segue para a apresentação do modelo da proposta.

No capítulo 4, o autor começa apresentando a técnica APL e suas 7 variações, esta técnica é utilizada para identificar as preferências conformacionais dos aminoácidos através de sua ES, ele dedica certo esforço em apresentar tantos detalhes, o que se justifica apenas posteriormente no capítulo 5 quando

faz a análise dos resultados e comparação entres os 7 métodos de APL. Em seguida ele apresenta a estrutura do modelo, que é dividido em 2 etapas. A etapa 1 visa incluir diversidade nos modelos estruturais, e é divida em 3 sub etapas: Inicialização das soluções a partir da técnica APL; Filtragem das soluções iniciais a partir das funções RG (Raio de Giro) e *SASA*; Agrupamento das soluções através da técnica de clusterização hierárquica aglomerativa. A etapa 2 consiste na otimização das soluções da etapa 1, através do algoritmo memético. Nesta etapa o autor define diversos parâmetros para o algoritmo de forma a tentar aumentar o poder exploratório do método, como por exemplo a estruturação da população em árvore, e o uso de um algoritmo Colônia Artificial de Abelhas como técnica de busca local, sendo aplicado nos nodos da árvore. Nota-se que se trata de uma proposta bastante complexa, com etapas e sub-etapas, exigindo certa concentração do leitor e talvez algumas releituras para compreender as peças deste quebra-cabeça. O autor tenta abranger em sua proposta diversos pontos em aberto do problema, enriquecendo o trabalho e agregando em conhecimentos diversos para o leitor. Porém a maior contribuição e inovação do trabalho está na etapa 1, a inicialização de uma população mais numerosa (com 10.000 indivíduos) e não aleatória como de costume foi a grande sacada do autor para trazer diversidade ao modelo, gerando grande expectativa para os resultados dos testes.

Os demais capítulos o autor dedica a análise dos resultados e conclusões, ganhando destaque as análises da etapa 1, onde realiza testes com 100.000 soluções e 8 proteínas de diferentes tamanhos e formas estruturais. Nesta etapa o autor surpreende o leitor com diversas constatações inesperadas, como por exemplo de que não há diferença nos tipos de APL com relação a qualidade das soluções geradas, também não foi possível constatar a relação função de energia x RMSD, ou seja, não é possível dizer que um menor valor de função de energia represente um melhor valor de RMSD, o que traz uma série de implicações, levantando a necessidade por novos estudos em relação a funções de energia para o PSP. Também constatou-se que o processo de filtragem pode descartar soluções ruins e manter soluções boas, o que pode ser considerado como redução do espaço de busca, ponto positivo para o trabalho.

Já para a etapa 2 foi necessário reduzir para 10.000 soluções iniciais devido a capacidade computacional, porém o autor peca em não explicitar as configurações computacionais utilizadas para os testes. O autor compara os resultados com os métodos Rosetta e Quark, constatando que o método proposto foi capaz de atingir valores comparáveis aos estados da arte. Com estes resultados o autor conseguiu comprovar sua teoria inicial de que soluções iniciais melhores aplicadas a problemas difíceis podem gerar resultados melhores. Mas apesar das descobertas e melhorias, alguns enigmas ainda persistem, como a dificuldade na identificação de estruturas do tipo folha beta, evidenciado nas proteínas que tiveram os piores resultados. Para concluir, vale destacar a utilidade do trabalho em questão, que trouxe termos relevantes e atuais que não só agregam no conhecimento, como também despertam para a busca de novas soluções.

A obra foi escrita por Leonardo de Lima Corrêa, sob orientação do professor Dr. Márcio Dorn. Corrêa atualmente é bolsista de doutorado do CNPq no Programa de Pós-Graduação em Computação do Instituto de Informática da Universidade Federal do Rio Grande do Sul. Já possui diversos trabalhos publicados na área de Bioinformática e Inteligência Artificial, sendo que alguns estão descritos no capítulo 8 da dissertação em questão.

Nilcimar Neitzel Will
Mestranda do Programa de Pós-Graduação em Computação Aplicada,
Universidade do Estado de Santa Catarina - UDESC.