Caracterização da Pesquisa: Predição de Estrutura 3D de Proteína

Nilcimar Neitzel Will[1]

[1] Prog. de Pós-graduação em Computação Aplicada – PPGCA
Universidade do Estado de Santa Catarina (UDESC) - Joinville, SC - Brasil

nil_cc@yahoo.com.br

Resumo. *Uma pesquisa ocorre quando há um problema a ser solucionado e possíveis soluções a serem analisadas. Ela deve ser realizada de forma a seguir um procedimento sistemático, com passos bem definidos, e para isto é necessário a caracterização da pesquisa. Identificar características da pesquisa quanto ao seu objetivo, nível de maturidade e procedimento de coleta de dados são algumas classificações importantes a serem identificadas para a definição da metodologia de pesquisa. Em vista disto, neste trabalho pretende-se caracterizar a pesquisa sobre Predição de Estrutura Terciária de Proteína quanto as possíveis classificações de uma pesquisa.*

1. Introdução

A previsão de estrutura tridimensional de proteína (PSP) é um dos problemas fundamentais da Bioinformática Estrutural, que ainda está em aberto. As proteínas são macromoléculas formadas por sequências de aminoácidos, sendo responsáveis por regular diversas atividades complexas nos seres vivos, o que inclui quase todos os processos biológicos fundamentais necessários a vida [School 2019]. Estas funções exercidas pelas proteínas estão diretamente relacionadas a sua forma tridimensional, pois é ao assumir a sua estrutura terciária que ela desenvolve a sua funcionalidade, sendo possível prever ou analisar a função que a mesma exerce no organismo [Lopes e Venske 2015]. Desta forma, conhecer sua estrutura 3D implica em também conhecer a sua função, o que é fundamental no desenvolvimento de soluções capazes de estimular, restringir ou suspender sua ação biológica [Gonçalves 2011].

Prever a estrutura 3D de uma proteína é um problema que envolve diversas variáveis, começando pelo fato de que existem 20 tipos de aminoácidos conhecidos, que podem formar milhares de combinações diferentes de proteínas, e cada proteína possui uma sequência única [Lopes e Venske 2015]. Em vista disso, pode-se dizer que o PSP pode ser considerado, de acordo com a teoria da complexidade computacional, como um problema NP-completo, devido a sua alta dimensionalidade e complexidade do espaço de busca conformacional [Berger e Leighton 1998]. Com isso, o PSP tem se mostrado um problema bastante desafiador, cujo objetivo principal seria minimizar a energia potencial da proteína, onde o espaço conformacional cresce a medida que aumenta o número de aminoácidos nas sequências de proteínas.

Identificado o problema, sua complexidade e objetivo, as próximas seções são dedicadas a categorizar a pesquisa quanto ao seu objetivo geral, nível de maturidade, raciocínio lógico, procedimento de coleta de dados, bem como a influência, natureza e forma das variáveis, e também o tipo de ciência ao qual esta pesquisa se encaixa. Tudo

isso se faz necessário para definir a metodologia de pesquisa mais adequada a este trabalho.

2. Caracterização da Pesquisa

Com relação ao objetivo geral da pesquisa, no momento atual é possível classificar esta pesquisa como Descritiva, pois conforme Gil (2010) já se passou pela etapa Exploratória, onde foi realizado um levantamento bibliográfico e já se adquiriu certo conhecimento geral sobre o problema. Nesse momento já se tem o entendimento dos parâmetros e algumas hipóteses que se pretende analisar. Por exemplo, já sabemos que agregando informação do problema do PSP ao algoritmo de otimização, pode-se reduzir o espaço de busca. O que se pretende nesse momento, é descrevê-las e identificar a associação entre estas variáveis (ditas como informação do problema), como elas podem reduzir de fato o espaço de busca e influenciar no resultado final.

Ainda com relação ao objetivo geral da pesquisa, pode-se afirmar que se trata de uma pesquisa com hipótese. Como já foi mencionado acima e segundo Volpato (2012), este tipo de pesquisa testa a relação entre duas ou mais variáveis. Neste caso, a relação que queremos identificar é como a utilização de informação do problema (por exemplo estrutura secundária ou fragmentos de proteína) está associada a melhoria na função de energia da proteína, não sendo esta uma relação de interferência, pois não é o uso de informação do problema que irá garantir a menor energia da proteína. Em vista disto, pode-se afirmar que se trata de uma pesquisa Descritiva com hipótese, onde se quer analisar a relação de associação entre as variáveis.

2.1. Nível de Maturidade

Segundo Wazlawick (2008), existem 5 graus de pesquisa quanto ao nível de maturidade. Destas, o autor define o nível 4 como sendo o nível mais maduro da pesquisa, pois usam testes padronizados, bancos de dados reconhecidos e métricas que possibilitam a reprodução dos testes. Com uma pesquisa do nível 4, pode-se dizer que foi realizado algo reconhecidamente melhor, com destaque para a palavra reconhecidamente, uma vez que é possível reproduzir os testes e comparar os resultados com o estado-da-arte.

Desta forma, esta pesquisa se encaixa no nível 4, pois para se realizar os experimentos com as proteínas, utiliza-se os mesmos bancos de dados de proteínas da literatura, que são bancos de dados reconhecidos e já consolidados, como por exemplo o PDB[1] (Protein Data Bank). Além disso, a cada 2 anos, ocorre o CASP (Critical Assessment of Structure Prediction), que é uma competição mundial de testes para o PSP, ao qual se pode ter acesso ao estado-da-arte e as proteínas utilizadas para teste, bem como as métricas consideradas. Também é comum obter este tipo de informação na literatura, pois já é padrão que os autores informem as proteínas testadas, as métricas utilizadas, bem como outras características relevantes aos testes. Desta forma, é possível comparar os resultados do meu experimento com os resultados apresentados pelos estados-da-arte ou por outros autores na literatura, sendo possível reconhecer qual é o melhor segundo aquele conjunto de testes e valores medidos.

[1] `https://www.rcsb.org/`

2.2. Raciocínio Lógico

Segundo Marconi e Lakatos (2005) tanto o pensamento indutivo quanto o dedutivo partem de um conjunto de constatações, premissas para inferir uma verdade. Porem, no pensamento dedutivo, se as premissas forem verdadeiras levam a uma conclusão verdadeira. Já no pensamento indutivo, mesmo as premissas sendo verdadeiras, o melhor que se pode chegar é a uma conclusão provavelmente verdadeira.

Desta forma, pode-se classificar esta pesquisa dentro do pensamento indutivo, pois se tomarmos como premissa que utilizando informação do problema é possível reduzir o espaço de busca e com isso reduzir a energia da proteína, mesmo que isto seja verdade para uma determinada proteína, não é possível afirmar que seja verdade para todas as proteínas existentes.

2.3. Procedimento de Coleta de Dados

Considerando que a pesquisa se encontra na etapa Descritiva, ainda se faz necessário algum procedimento de coleta de dados bibliográficos, desta vez de forma mais específica, pois já se tem o conhecimento geral do problema e das variáveis envolvidas. Ao mesmo tempo que já se inicia o procedimento de coleta Experimental [Gil 2010], com a construção do artefato com base nas variáveis já identificadas. Então, pode-se dizer que se trata de uma pesquisa Experimental, pois ao final deve-se apresentar os resultados de um experimento controlado. Porém, no momento atual, encontra-se em transição da coleta de dados bibliográfica e iniciando a coleta de dados experimental.

2.4. Influência das Variáveis

Segundo Cervo et al. (2007), é possível identificar 3 tipos de variáveis, sendo elas independentes, dependentes e intervenientes. Nesta pesquisa, estas variáveis são caracterizadas da seguinte forma:

- Independentes: onde encontram-se as variáveis relacionadas a informação do problema, como por exemplo: estrutura secundária e o tamanho dos fragmentos de proteína; e as variáveis relacionadas ao algoritmo de otimização, como por exemplo: tamanho da população, número de gerações, probabilidade de cruzamento e mutação.
- Dependentes: é o valor que deseja-se obter ao final do processo de otimização, que é o valor da energia da proteína.
- Intervenientes: são as variáveis que não se pode controlar, mas que influenciam o valor da energia da proteína. Neste caso, pode-se considerar o espaço de busca e a representação computacional da proteína.

Para reduzir o impacto do espaço de busca, é possível utilizar as variáveis independentes relacionadas a informação do problema. Com relação a representação computacional, pode-se usar uma representação *all-atom* que considera todos os átomos da proteína o que pode tornar o procedimento muito lento, ou pode-se utilizar uma representação centróide, que considera apenas os ângulos principais dos aminoácidos, reduzindo bastante o impacto de uma representação completa.

Desta forma, se espera aplicar informação do problema com um algoritmo de otimização, para obter o menor valor de energia da proteína, considerando um espaço de busca reduzido e uma representação computacional centróide.

2.5. Natureza e Forma das Variáveis

Quanto a abordagem da pesquisa, se trata de uma pesquisa Quantitativa, porque o resultado esperado é fundamental para a pesquisa, onde o objetivo é minimizar a energia da proteína, e chegar num valor aceitável para esta energia é o que valida todo o processo. Da mesma forma, os valores estatísticos são necessários para a comparação com os demais trabalhos da literatura e estado-da-arte.

Igualmente, as variáveis independentes e dependentes são quantitativas, e podem ser classificadas quanto a sua forma, segundo Barbetta (2006):

- Estrutura secundária: contínua, pois representa a probabilidade de um determinado aminoácido possuir tal conformação de estrutura secundária;
- Tamanho de fragmento de proteína: cardinal, podendo assumir os valores 3 ou 9;
- Tamanho da população: inteiro, podendo assumir qualquer valor inteiro maior que zero;
- Número de gerações: inteiro, podendo assumir qualquer valor inteiro maior que zero;
- Cruzamento: contínuo, sendo a probabilidade de ocorrência de cruzamento entre 2 indivíduos;
- Mutação: contínuo, sendo a probabilidade de ocorrência de mutação em um novo indivíduo.
- Energia: contínuo, representa o valor da energia potencial da proteína.

2.6. Tipo de Ciência

Em Wazlawick (2010) o autor faz uma reflexão quanto a Computação e os tipos de Ciências, identificando em quais subclassificações da Ciência as diferentes subáreas da Computação se encaixam. Conforme essa discussão, pode-se encaixar esta pesquisa dentro das Ciências Exatas e Inexatas, mais especificamente como uma ciência inexata, pela utilização de algoritmo evolutivo, que da mesma forma que os algoritmos genéticos, não produzem resultados exatos e podem ser imprevisíveis.

3. Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Referências

Barbetta, P. A. (2006). *Estatística Aplicada às Ciências Sociais*. Editora da UFSC. 6a ed.

Berger, B. e Leighton, T. (1998). Protein folding in the hydrophobic-hydrophilic (hp) is np-complete. Em *Proceedings of the Second Annual International Conference on Computational Molecular Biology*, RECOMB '98, páginas 30–39, New York, NY, USA. ACM.

Cervo, A. L., Bervian, P. A. e Da Silva, R. (2007). *Metodologia Científica*. Pearson Prentice Hall, 6a ed.

Gil, A. C. (2010). *Como Elaborar Projetos de Pesquisa*. Ed. São Paulo:Atlas. 5a ed.

Gonçalves, W. W. (2011). Um estudo da aplicação de algoritmos genéticos na predição da estrutura 3-d aproximada de proteínas. Monografia, Universidade Federal do Rio Grande do Sul.

Lopes, J. N. e Venske, S. M. (2015). Predição da estrutura de proteínas utilizando algoritmo evolutivo adaptativo. Em Bastos Filho, C. J. A., Pozo, A. R. e Lopes, H. S., editores, *Anais do 12 Congresso Brasileiro de Inteligência Computacional*, páginas 1–6, Curitiba, PR. ABRICOM.

Marconi, M. A. e Lakatos, E. M. (2005). *Fundamentos da Metodologia Científica*. Atlas. 5a ed.

School, H. M. (2019). *New deep-learning approach predicts protein structure from amino acid sequence*.

Volpato, G. L. (2012). *Tipos Lógicos de Pesquisa. Aula 20*. Disponível em: www.youtube.com/watch?v=XoTQo7fUf0s.

Wazlawick, R. S. (2008). *Metodologia de Pesquisa para Ciência da Computação*. Elsevier.

Wazlawick, R. S. (2010). Uma reflexão sobre a pesquisa em ciência da computação à luz da classificação das ciências e do método científico. *Revista de Sistemas de Informação da FSMA*, páginas 3–10.