https://eventos.utfpr.edu.br//sicite/sicite2019

Estudo do caminho de dobramento de proteínas globulares usando dinâmica molecular com modelo granuloso

Study of the folding pathway of globular proteins using molecular dynamics with the coarse-grained model

Bruna Araújo Pinheiro
www.bruna.a.p@gmail.com
Universidade Tecnológica Federal do Paraná, Curitiba, Paraná, Brasil

César Manuel Vargas Benítez
cesarvargasb@gmail.com
Universidade Tecnológica Federal do Paraná, Curitiba, Paraná, Brasil

Leandro Takeshi Hattori
leandrotakeshihattori@gmail.com
Universidade Tecnológica Federal do Paraná, Curitiba, Paraná, Brasil

Lucas Destefani Fabri
lucasfabri@alunos.utfpr.edu.br
Universidade Tecnológica Federal do Paraná, Curitiba, Paraná, Brasil

RESUMO

Para as proteínas estarem em sua configuração funcional, suas estruturas precisam passar por um processo denominado dobramento proteico. No qual, independentemente de sua conformação inicial, atinge em uma estrutura final similar, a estrutura nativa. A má formação das estruturas pode causar doenças degenerativas como Alzheimer e alguns tipos de câncer. Na literatura existem diversos trabalhos que exploram a predição da estrutura nativa a partir das informações das sequências dos aminoácidos. Entretanto, as pesquisas sobre o caminho ou trajetória de dobramento ainda são escassos. Assim, este trabalho propõe a criação de conjunto de dados de trajetórias de dobramento de proteínas utilizando o método de Dinâmica Molecular (DM), a fim de auxiliar futuros estudos sobre este processo. Também foi proposta a paralelização do DM utilizando a arquitetura *Graphics Processing Unit* (GPU). Neste estudo foram selecionadas quatro proteínas globulares, 13FIBO, 2GB1, 1PLC e 5NAZ, com 13, 56, 99 e 229 aminoácidos respectivamente, sendo que a posição espaço-temporal de cada aminoácido, a energia potencial e os raios de giração foram armazenados. Os dados das trajetórias produzidos por DM indicaram ser coerentes com os paradigmas de convergência, formando configurações estruturais similares ao final das simulações.

PALAVRAS-CHAVE: Dinâmica molecular. Paralelismo. Dobramento de proteínas.

ABSTRACT

For proteins to be in their functional configuration, their structures need to go through a process called protein folding. In which, regardless of its initial conformation, it reaches into a similar final structure, called a native structure. Misfolded proteins can cause degenerative diseases such as Alzheimer's and some cancers. There are several works in the literature that explore the prediction of native structure from information of the amino acid sequences. However, research on the folding pathways is still scarce. Thus, this work proposes the creation of a protein folding trajectory dataset using the Molecular Dynamics (DM) method, in order to assist future studies on this process. DM parallelization was also proposed using the Graphics Processing Unit (GPU) architecture. In this study, we selected four globular proteins, 13FIBO, 2GB1, 1PLC and 5NAZ, with 13, 56, 99 and 229 amino acids, respectively. Where, the spacetime position of each amino acid, potential energy, and the radii of gyration has been stored. The trajectory data produced by DM indicated to be coherent with the convergence paradigms, forming similar structures at the end of the simulations.

KEYWORDS: *Molecular dynamics. Parallelism. Protein folding.*

Recebido: 19 ago. 2019.

Aprovado: 01 out. 2019.

INTRODUÇÃO

Quando as proteínas estão dobradas (em sua forma nativa) elas possuem uma configuração instável por conta da sua alta energia, então elas se dobram para atingir uma configuração mais estável, cuja energia é menor que a energia na forma nativa.

Uma das abordagens computacionais mais conhecidas para simular o dobramento das proteínas é o método da Dinâmica Molecular (DM). Esta abordagem calcula as forças newtonianas que agem sobre cada aminoácido para simular a trajetória de dobramento. Estes cálculos são realizados a cada iteração do algoritmo para simular as trajetórias dos aminoácidos. Este processo é realizado iterativamente, simulando o processo de dobramento até atingir uma configuração estável.

Por conta do grande número de cálculos do método de DM, é necessário um alto poder computacional para executar este método. Entretanto, dado ao desacoplamento dos cálculos de energia, DM é uma abordagem que pode ser altamente paralelizável. Atualmente, a arquitetura em GPU tem ganhado popularidade, principalmente após a disponibilização de *Application Programming Interface* (APIs) com o *Compute Unified Device Architecture* (CUDA). Outros fatores incluem o custo financeiro e o grande número de núcleos de processamento que podem ser utilizados para cálculos massivamente em paralelo. Dado a estes fatores, o uso de DM utilizando a arquitetura em GPU pode ser promissor.

MATERIAL E MÉTODOS

O algoritmo de DM utilizado neste trabalho foi baseado no trabalho de BENÍTEZ, (2010). Este algoritmo foi paralelizado em GPU utilizando a API CUDA utilizando a abordagem mestre-escravo, onde o mestre (processo em CPU) gerencia as tarefas para os escravos (*threads* da GPU). O programa também foi adaptado para executar simulações de proteínas com cadeias superiores à 200 aminoácidos.

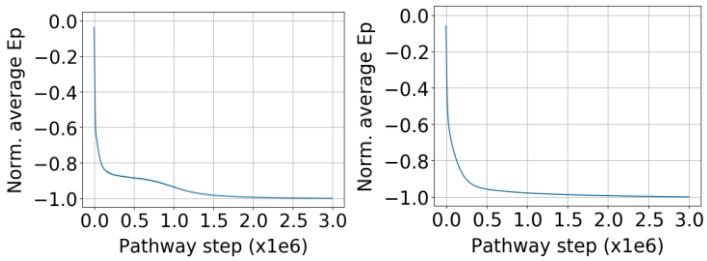
A cada execução do programa foi gerada um arquivo de trajetória de dobramento, contendo informações da estrutura, raio de giração, energia e o tempo de processamento. Para este estudo foram exploradas quatro sequências: 13FIBO, 2GB1, 1PLC e 5NAZ, com 13, 56, 99, e 229 aminoácidos, respectivamente.

Os datasets das quadro proteínas são compostas de 3500 trajetórias de dobramento no total. Após a geração dos *datasets,* foram gerados gráficos da energia potencial de cada *dataset*, bem como mapas de calor que apresentam graficamente a diferença entre as configurações iniciais (estruturas aleatoriamente geradas como auxílio do algoritmo M*ersenne Twister*) e finais (após a estrutura passar pelo algoritmo de dobramento) de cada proteína. Além de representações gráficas da configuração da proteína em cada estado do caminho de dobramento, bem como vídeos que representam um dos 1000 possíveis caminhos de dobramento.

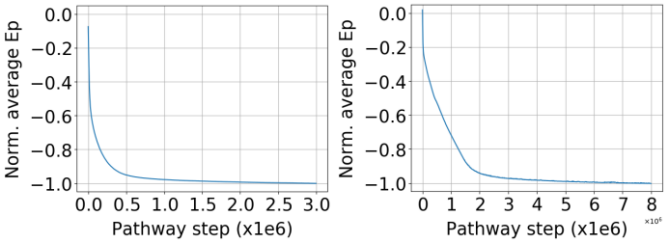
RESULTADOS E DISCUSSÃO

Este experimento tem o objetivo de apresentar o comportamento médio da energia de cada proteína durante a trajetória de dobramento dos *datasets*. Os resultados obtidos são apresentados na Figura 1. Foi possível observar o decaimento exponencial da energia ao longo do tempo da energia, estabilizando ao final da simulação. Entretanto, foi observado que para proteínas maiores a estabilização da energia tende a levar um maior número de iterações.

Figura 1 – Gráficos da energia potencial média de cada proteína



(a) 13FIBO (b) 2GB1

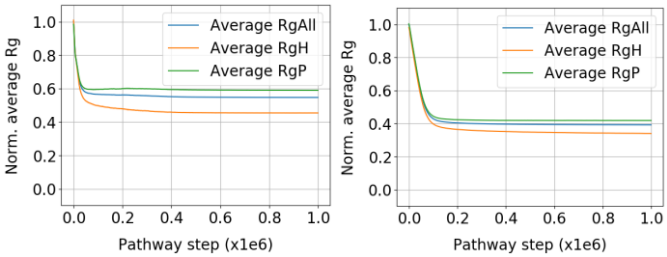


(c) 1PLC (d) 5NAZ

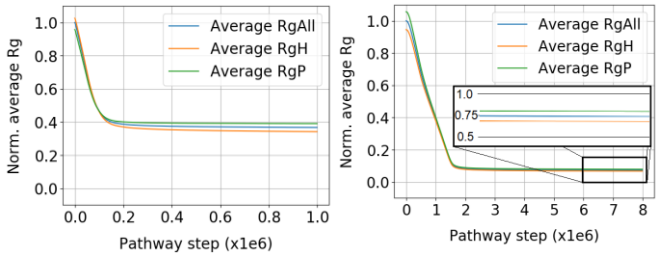
Fonte: produção do próprio autor.

Assim no experimento anterior, na Figura 2 é apresentado a média dos raios de giração ao longo das trajetórias de dobramento. Neste trabalho utilizamos os raios de giração dos elementos hidrofóbicos, polares e de todos (RgH, RgP e RgAll). Nos gráficos apresentados pela figura é possível observar que os valores dos raios de giração vão decaindo exponencialmente ao longo do tempo, assim como a energia potencial. Também foi possível observar que os RgH tem menor compacidade que os RgP, indicando uma formação do núcleo hidrofóbico, assim como em proteínas globulares.

Figura 2 - Gráficos do raio de giração médio de cada proteína



(a) 13FIBO (b) 2GB1

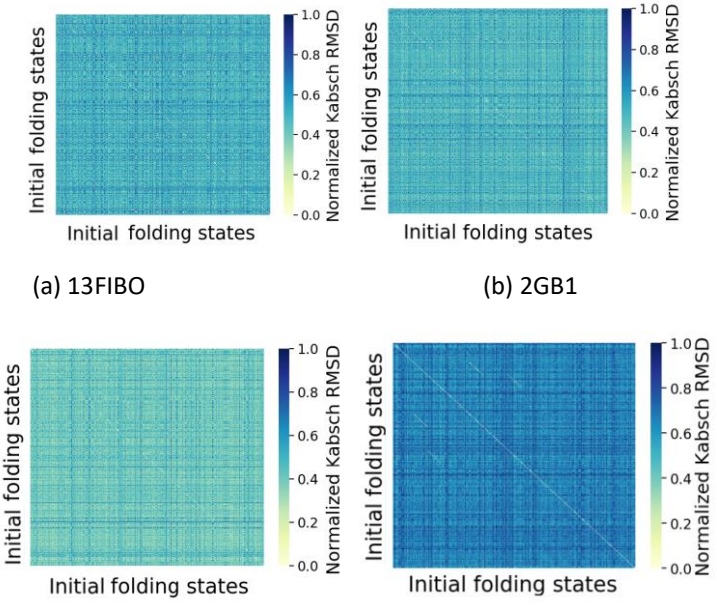


(c) 1PLC (d) 5NAZ

Fonte: produção do próprio autor.

A fim de verificar a semelhança entre as estruturas iniciais e finais das trajetórias das proteínas dos *datasets*, foram comparadas entre si usando o algoritmo Kabsch. E, através da matriz resultante desse algoritmo, foram produzidos os mapas de calor apresentados na Figura 3. Nessa figura podemos observar o mapa de calor que representa a similaridade entre as estruturas iniciais da proteína através de uma escala Kabsch de cores normalizada (entre 0 e 1). Nessa escala quanto mais próximo de 1, representado pela cor azul escuro, mais diferentes são as estruturas e quanto mais próximo de 0, representado pela cor amarelo, mais semelhantes são as estruturas.

Figura 3 – Mapa de calor das estruturas iniciais de cada proteína



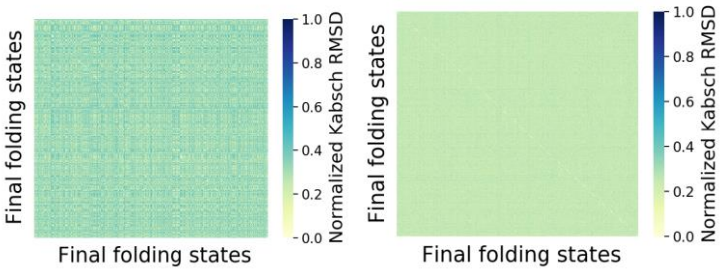
(a) 13FIBO (b) 2GB1

(c) 1PLC (d) 5NAZ

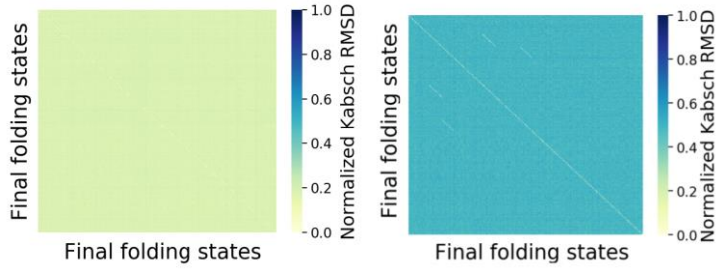
Fonte: produção do próprio autor.

Seguindo o mesmo método de produção dos mapas de calor das estruturas iniciais foram feitos mapas de calor comparando as estruturas finais, que são apresentados pela Figura 4. Comparando os mapas de calor é possível perceber que após passar pelo algoritmo de dobramento as estruturas ficam mais semelhantes entre si, o que indica que elas convergem para uma mesma configuração.

Figura 4 – Mapa de calor das estruturas finais de cada proteína



(a) 13FIBO (b) 2GB1

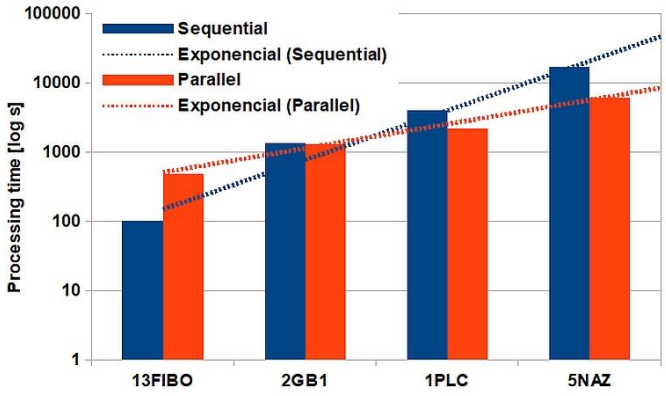


(c) 1PLC (d) 5NAZ

Fonte: produção do próprio autor.

Durante a execução do programa foram coletados os tempos de execução do algoritmo em CPU e em GPU para as quatro proteínas, a fim de mostrar a diferença computacional entre as duas arquiteturas em escala logarítmica, apresentado na Figura 6. Os resultados mostraram que o tempo de processamento em GPU é mais rápido que em CPU para proteínas globulares com mais de 56 aminoácidos (2GB1), superando o tempo de gerenciamento das *threads*.

Figura 6 - Comparação dos tempos de processamento entre CPU e GPU



Fonte: produção do próprio autor.

CONCLUSÃO

Tendo em vista os resultados obtidos, pode-se perceber que as diferentes estruturas iniciais de uma proteína convergem para uma estrutura similar. Além disso, a forma final da proteína tende a ter uma energia potencial final mais baixa que a inicial, como previsto na literatura. Comparando os gráficos da energia

potencial da proteína 5NAZ com as depois é possível observar que proteínas com um número maior de aminoácidos levam mais iterações do algoritmo para se estabilizar. Também pode-se perceber que o tempo de execução do algoritmo em CPU é menor para proteínas com menos de 56 aminoácidos, mostrando que para proteínas com mais de 56 aminoácidos é preferível a paralelização em GPU.

AGRADECIMENTOS

Agradecimentos à UTFPR e ao CNPQ pelas oportunidades e auxílio financeiro. Agradecimentos especiais aos professores César Manuel Vargas Benítez e Heitor Silvério Lopes e aos estudantes Leandro Takeshi Hattori e Lucas Destefani Fabri por todo o apoio e orientação durante o projeto.

REFERÊNCIAS

BENÍTEZ, César Manuel Vargas. **Contributions to the Study of the Protein Folding Problemusing Bioinspired Computation and Molecular Dynamics**. 2015. Tese (Doutorado em Engenharia de Computação) - Universidade Tecnológica Federal do Paraná, [*S. l.*], 2015. Disponível em: http://paginapessoal.utfpr.edu.br/cesarbenitez/trabalhos-realizados/CPGEI_Tese_112_2015.pdf/view. Acesso em: 31 jul. 2019.

BENÍTEZ, César Manuel Vargas. **Contributions to the Study of the Protein Folding Problemusing Bioinspired Computation and Molecular Dynamics.** 2010. Dissertação (Mestrado em Engenharia de Computação) - Universidade Tecnológica Federal do Paraná, [*S. l.*], 2010. Disponível em: http://paginapessoal.utfpr.edu.br/cesarbenitez/trabalhos-realizados/CPGEI_Dissertacao_532_2010.pdf/view. Acesso em: 31 jul. 2019.

KABSCH, W. A discussion of the solution of the best rotation to relate two sets of vectors. **Acta Crytallographica**, A34, p.827-828, 1978.