DESENVOLVIMENTO DE PIPELINE PARA DETECÇÃO DE MODELOS ATÔMICOS DE MACROMOLÉCULAS EM IMAGENS DE MICROSCOPIA ELETRÔNICA

LABORATÓRIO DE BIOLOGIA COMPUTACIONAL – LNBio – CNPEM

Orientador/Pesquisador responsável: Dr. Helder Veras Ribeiro Filho

Coorientador: M.e. João Victor da Silva Guerra

Pesquisador colaborador: Dr. Paulo Sérgio Lopes de Oliveira

Introdução

Esse projeto objetiva o desenvolvimento de um pipeline computacional automatizado para detecção de estruturas atômicas de macromoléculas de interesse biológico em imagens de microscopia eletrônica (ME) de contraste negativo. A coleta de imagens em contraste negativo permite a rápida caracterização estrutural da amostra e sua qualidade, sendo comumente utilizada como etapa prévia a determinação em alta resolução de estruturas macromoleculares por crio-microscopia eletrônica. Entretanto, essas amostras podem ser bastante complexas e heterogêneas, o que dificulta a interpretação e caracterização estrutural da macromolécula em estudo. Portanto, a aplicação do pipeline permitirá ao usuário interpretar as imagens obtidas no microscópio eletrônico utilizando modelos atômicos estruturais, determinados experimentalmente ou obtidos por modelagem molecular, que representam a macromolécula de interesse e suas possíveis conformações. Para isso, o pipeline percorrerá um caminho inverso ao *single particle analysis*, partindo do modelo atômico e utilizando projeções 2D desse modelo para realizar buscas nas micrografias. Como caso de estudo, o pipeline será aplicado em micrografias de amostras de uma proteína do SARS-CoV-2, e para detecção, serão utilizados modelos estruturais dessa proteína oriundos de simulações de dinâmica molecular *coarse-grained*. Em especial, o projeto avaliará a importância desse modelo de simulação para interpretação das imagens de ME. Com a implementação do pipeline, é esperado uma melhoria na interpretação das imagens coletadas, facilitando a tomada de decisão por parte do usuário e economizando tempo e recursos. Com a combinação de modelos estruturais e imagens de ME, os usuários poderão produzir interpretações estruturais e gerar ou validar suas hipóteses. O pipeline pode ainda ser futuramente expandido a outras técnicas de imagem, como tomografia crio-eletrônica (Ortiz et al., 2006).

Estado da arte

Os avanços na técnica de ME juntamente com o desenvolvimento de novos algoritmos voltados ao processamento de imagens obtidas por ME estão proporcionando um exponencial crescimento na determinação de novas estruturas macromoleculares de interesse biológico, como proteínas, ácidos nucleicos e partículas virais. Nesse contexto, a técnica de crio-microscopia eletrônica (Cryo-EM) nos últimos anos vem permitindo a determinação de estruturas cada vez mais complexas (Ribeiro-Filho et al., 2020) e em resolução atômica (Callaway, 2020). Para determinação da estrutura tridimensional das macromoléculas, as imagens (micrografias) obtidas pelo microscópio eletrônico são submetidas a um conjunto de técnicas de processamento de imagem denominado *single particle analysis* (SPA) que permite a construção de um mapa de densidades que pode ser visualizado na forma de um volume 3D (Afanasyev et al., 2017). Esse mapa é a referência para a construção do modelo atômico da macromolécula. Entretanto, um grande desafio ainda enfrentado para determinação de estruturas por Cryo-EM é a heterogeneidade da amostra analisada, tanto do ponto de vista de flexibilidade da macromolécula alvo quanto da capacidade desta de transitar entre diferentes estados oligoméricos.

Como uma etapa prévia a coleta de imagens por Cryo-EM, a técnica de contraste negativo *(negative staining* em inglês*)* é aplicada para caracterizar rapidamente a amostra analisada e diagnosticar sua qualidade. Essa técnica permite a visualização em baixa resolução da macromolécula que compõe a amostra analisada e possibilita a avaliação de tamanho e formato da macromolécula, grau de oligomerização e heterogeneidade conformacional. Um exemplo da importância dessa técnica, foi a recente caracterização de uma das proteínas do SARS-CoV-2 pelo LNBio e LNNano (resultados em fase de publicação). Nesse caso, a proteína do SARS-CoV-2 mostrou-se bastante heterogênea, com a formação de diferentes estados oligoméricos e diferentes conformações, dificultando sua interpretação estrutural.

Para colaborar na interpretação estrutural de macromoléculas nesses casos complexos, métodos computacionais podem fornecer informações preditivas relevantes em um intervalo curto de tempo. Para este fim, técnicas de modelagem por homologia juntamente com *docking* molecular podem predizer estruturas de macromoléculas compostas por vários domínios, assim como simulações de dinâmica molecular podem amostrar o espaço conformacional de macromoléculas flexíveis e complexas. Diante do mesmo objetivo, torna-se extremamente vantajosa a integração de métodos de modelagem computacional e imagens de ME para superar desafios e problemas de biologia estrutural de forma eficiente.

Objetivos

- Implementar pipeline de detecção automatizada de modelos atômicos 3D de macromoléculas biológicas em imagens de contraste negativo utilizando rotinas disponíveis à comunidade científica;
- Comparar e aprimorar acurácia e desempenho das rotinas implementadas no pipeline;
- Avaliar a utilização de simulações de dinâmica molecular com modelo CG como forma de amostragem conformacional para detecção e interpretação de partículas nas imagens de contraste negativo.

Metodologia

O pipeline compreenderá três tarefas principais (Figura 1) e será desenvolvido na linguagem de programação *Python* (https://www.python.org/). Em uma primeira etapa do projeto, utilizaremos rotinas de programas disponíveis à comunidade científica para construção do pipeline e avaliação de acurácia e desempenho. Em uma segunda etapa, implementaremos e/ou otimizaremos os algoritmos buscando melhorias na acurácia e desempenho do pipeline.

Para utilizar o pipeline, o usuário submeterá um conjunto de imagens de ME da amostra da sua macromolécula alvo. No projeto, como caso de estudo serão utilizadas imagens obtidas no LNNano de contraste negativo de uma proteína relacionada a um coronavírus purificada no LNBio. Juntamente com as imagens, o usuário submeterá um arquivo de metadados que conterá o tamanho do pixel de cada micrografia e o conjunto de modelos 3D atômicos de referência da macromolécula alvo no formato PDB. Essas estruturas podem ser oriundas do *Protein Data Bank* (PDB), podem ser modelos obtidos por homologia, ou quadros de simulações de dinâmica molecular. No presente projeto, como caso de estudo esses modelos serão quadros de simulações de dinâmica molecular com modelo CG. As simulações CG da proteína em estudo serão produzidas com o programa CafeMol (Kenzaki et al., 2011) e uma rotina disponível no CafeMol realizará a conversão de modelo CG para modelo atomístico.

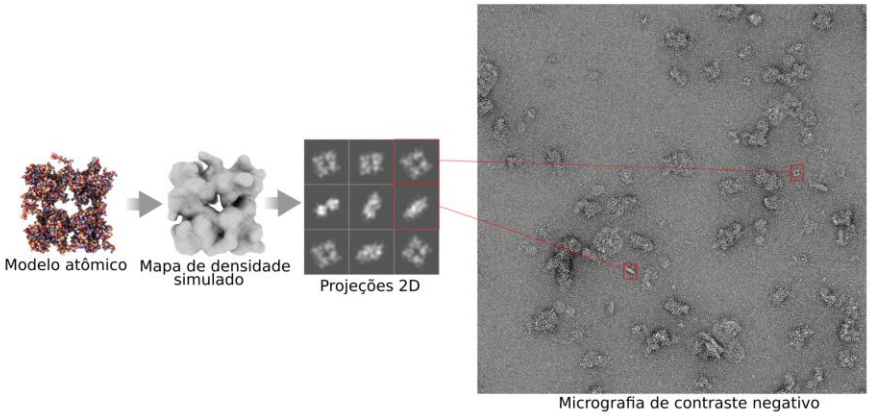

Figura 1: Fluxograma simplificado do funcionamento do pipeline.

Implementação do pipeline

A primeira tarefa do pipeline corresponderá a construção de mapas de densidade simulados a partir de modelos atômicos da proteína de estudo submetidos pelo usuário (Figura 1). Para cada um dos modelos (no caso de estudo, quadros da simulação) no formato PDB, serão construídos os mapas de densidade simulados no formato MRC. Na primeira etapa do projeto, utilizaremos a rotina *pdb2mrc* do programa EMAN2 (Tang et al., 2007). Essa rotina, produz densidades Gaussianas 3D para cada átomo da proteína e então insere estas densidades Gaussianas de todos os átomos em uma grade 3D para produzir um mapa de densidade simulado para toda a proteína. Esse método é amplamente utilizado (Igaev et al., 2019; Kim et al., 2019) na área de processamento de imagens de ME, porém teve sua origem na construção de mapas de densidade eletrônica simulados a partir de estruturas atômicas resolvidas por difração de raio-X. Quando utilizado para simular mapas em baixa ou intermediária resolução (8-20 Å), como no caso do presente projeto que envolve imagens de contraste negativo (resolução máxima aproximada de 20 Å), esse método produz uma aproximação aceitável. Entretanto, no contexto do Cryo-EM e estruturas de alta resolução, esse método possui limitações por não considerar características atômicas próprias como carga e ligações (Mostosi et al., 2019).

A segunda tarefa do pipeline utilizará os mapas de densidade simulados construídos na tarefa anterior e projetará esses volumes 3D em imagens 2D com determinada amostragem angular (Figura 1). Tais projeções corresponderiam as possíveis orientações da proteína no campo de visão do microscópio. Essa tarefa será implementada utilizando a função *Angular_project_library* do programa xmipp (de la Rosa-Trevín et al., 2013) e para cada mapa de densidade será construída uma coleção de possíveis projeções. Nessa tarefa, dependendo da amostragem angular escolhida e da resolução do mapa de origem, pode-se obter um conjunto grande projeções com projeções redundantes. Portanto, essa etapa poderá ser otimizada, conforme descrito na seção seguinte. Ainda nessa tarefa, o tamanho do pixel das imagens projetadas será ajustado para ser compatível ao das micrografias de contraste negativo.

Na terceira tarefa, as micrografias de contraste negativo serão examinadas para detecção das projeções da proteína alvo, representando suas diferentes orientações no plano, ou seja, essa etapa pode ser tratada como um problema de *template matching* (Figura 1). Em uma primeira etapa do projeto, a busca será realizada por correlação cruzada, métrica originalmente abordada em (Saxton & Frank, 1976) e implementada no programa Imagic-4D (van Heel et al., 2012). O desempenho do procedimento de busca será avaliado, uma vez que a correlação cruzada é dependente da rotação da projeção, o que tornará a etapa computacionalmente custosa.

Como resultado da terceira tarefa, as partículas mais similares com as projeções serão identificadas nas micrografias originais. O pipeline permitirá visualização comparativa, lado a

lado, das partículas detectadas e suas projeções similares. Além disso, será calculada estatística das partículas detectadas, o que pode colaborar, por exemplo, na identificação prévia de possível viés na orientação das partículas na grade utilizada na ME.

Otimização do pipeline

Após a implementação do pipeline, o presente projeto propõe-se a otimizar as rotinas das três tarefas descritas anteriormente. Para a primeira tarefa, a partir dos conhecimentos obtidos da rotina *pdb2mrc*, propõe-se avaliar sua utilização no contexto de alta resolução. Para a tarefa seguinte, busca-se a redução do espaço amostral de projeções, a partir da aplicação de técnicas de classificação, a fim de selecionar de forma não-supervisionada as projeções únicas. Para a última tarefa, diferentes algoritmos de *template matching* e aprendizado de máquina disponíveis na literatura serão testados; principalmente, os que possam se beneficiar de programação paralela tanto em CPU quanto GPU. Todas as otimizações implementadas ao pipeline buscarão o aumento da acurácia e desempenho.

Referências Bibliográficas

Afanasyev, P., Seer-Linnemayr, C., Ravelli, R. B. G., Matadeen, R., de Carlo, S., Alewijnse, B., Portugal, R. v., Pannu, N. S., Schatz, M., & van Heel, M. (2017). Single-particle cryo-EM using alignment by classification (ABC): The structure of Lumbricus terrestris haemoglobin. *IUCrJ*, *4*(5), 678–694. https://doi.org/10.1107/S2052252517010922

Callaway, E. (2020). "It opens up a whole new universe": Revolutionary microscopy technique sees individual atoms for first time. In *Nature* (Vol. 582, Issue 7811, pp. 156–157). NLM (Medline). https://doi.org/10.1038/d41586-020-01658-1

de la Rosa-Trevín, J. M., Otón, J., Marabini, R., Zaldívar, A., Vargas, J., Carazo, J. M., & Sorzano, C. O. S. (2013). Xmipp 3.0: An improved software suite for image processing in electron microscopy. *Journal of Structural Biology*, *184*(2), 321–328. https://doi.org/10.1016/j.jsb.2013.09.015

Igaev, M., Kutzner, C., Bock, L. v., Vaiana, A. C., & Grubmüller, H. (2019). Automated cryo-EM structure refinement using correlation-driven molecular dynamics. *ELife*, *8*. https://doi.org/10.7554/eLife.43542

Kenzaki, H., Koga, N., Hori, N., Kanada, R., Li, W., Okazaki, K. I., Yao, X. Q., & Takada, S. (2011). CafeMol: A coarse-grained biomolecular simulator for simulating proteins at work. *Journal of Chemical Theory and Computation*, *7*(6), 1979–1989. https://doi.org/10.1021/ct2001045

Kim, D. N., Moriarty, N. W., Kirmizialtin, S., Afonine, P. v., Poon, B., Sobolev, O. v., Adams, P. D., & Sanbonmatsu, K. (2019). Cryo_fit: Democratization of flexible fitting for cryo-EM. *Journal of Structural Biology*, *208*(1), 1–6. https://doi.org/10.1016/j.jsb.2019.05.012

Mostosi, P., Schindelin, H., Kollmannsberger, P., & Thorn, A. (2019). Automatic annotation of Cryo-EM maps with the convolutional neural network Haruspex. In *bioRxiv* (p. 644476). bioRxiv. https://doi.org/10.1101/644476

Ortiz, J. O., Förster, F., Kürner, J., Linaroudis, A. A., & Baumeister, W. (2006). Mapping 70S ribosomes in intact cells by cryoelectron tomography and pattern recognition. *Journal of Structural Biology*, *156*(2), 334–341. https://doi.org/10.1016/j.jsb.2006.04.014

Ribeiro-Filho, H. V., Coimbra, L. D., Cassago, A., Rocha, R. P. F., Padilha, A. C., da Silva Guerra, J. V., Leme, L., Trivella, D. B. B., Portugal, R. v., Lopes-de-Oliveira, P. S., & Marques, R. E. (2020). Cryo-EM structure of the mature and infective Mayaro virus at 4.4 Å resolution reveals new features of arthritogenic alphaviruses. In *bioRxiv* (p. 2020.11.06.371773). bioRxiv. https://doi.org/10.1101/2020.11.06.371773

Saxton, W. O., & Frank, J. (1976). Motif detection in quantum noise-limited electron micrographs by cross-correlation. *Ultramicroscopy*, *2*(C), 219–227. https://doi.org/10.1016/S0304-3991(76)91385-1

Tang, G., Peng, L., Baldwin, P. R., Mann, D. S., Jiang, W., Rees, I., & Ludtke, S. J. (2007). EMAN2: An extensible image processing suite for electron microscopy. *Journal of Structural Biology*, *157*(1), 38–46. https://doi.org/10.1016/j.jsb.2006.05.009

van Heel, M., Portugal, R., Rohou, A., Linnemayr, C., Bebeacua, C., Schmidt, R., Grant, T., & Schatz, M. (2012). *Four-dimensional cryo-electron microscopy at quasi-atomic resolution:* IMAGIC 4D (pp. 624–628). American Cancer Society. https://doi.org/10.1107/97809553602060000875