UMA PLATAFORMA COMPUTACIONAL PARA ANÁLISE DE EXPRESSÃO GÉNICA

Diogo André Rocha Teixeira

Dissertação realizada sob a orientação do *Prof. Rui Camacho* e co-orientação de *Nuno Fonseca*
na *Faculdade de Engenharia da Universidade do Porto*

1. Contexto

A biologia molecular é um ramo da biologia que estuda as atividades biológicas dos seres vivos, ao nível molecular. As bases para esta área de estudo foram criadas no início da década de 1930, embora apenas tenha emergido na sua forma mais moderna na década de 1960, com a descoberta da estrutura do DNA. Entre os processos estudados por este ramo da biologia está a expressão génica. A expressão génica é o processo através do qual moléculas de DNA são transformadas em produtos genéticos úteis, tipicamente proteínas, que são essenciais para os organismos vivos. Este conhecimento não é apenas importante em áreas como biologia molecular ou evolutiva, mas tem aplicações cruciais em áreas como medicina. Um exemplo de uma destas aplicações é a utilização de análise de expressão génica no diagnóstico e tratamento de pacientes com cancro [1].

Com o advento das técnicas de *Next Generation Sequencing* (NGS) os investigadores têm à sua disposição grandes quantidades de dados de sequenciação, cuja produção é mais barata e rápida. Estes dados podem ser usados para obter informação relevante sobre a expressão génica de organismos. Mas, à medida que o custo da sequenciação de genomas é reduzido, o custo do processamento dessa informação aumenta. Técnicas NGS costumam produzir *reads*[1] mais curtas quando comparadas com aquelas produzidas por técnicas anteriores, apresentando um problema mais desafiante, de ponto de vista computacional [2].

2. Problem de Domínio

Apesar dos grandes avanços nas últimas décadas, a biologia molecular é ainda uma área recente e, como tal, ainda existem muitas incógnitas e conhecimento parcial. Alguns dos mecanismos reguladores da expressão génica são ainda desconhecidos. Um destes mecanismos regula a velocidade de transcrição de RNA. Um dos objetivos desta dissertação é preceber de que forma é que as sequências finais do exões de um gene afetam a velocidade com que estes são transcritos. O outro objetivo é perceber de que forma a manipulação das *RNA-binding proteins* (RBP) pode ser usada para melhor perceber a expressão génica de um organismo. Estas são tarefas complexas que podem ser decompostas nos três principais problemas que vão ser endereçados nesta dissertação.

Alinhamento de *reads* contra um genoma de referência e análise da expressão diferencial entre amostras de diferentes indivíduos. Este é o problema mais complexo tratado nesta dissertação. Serão usados dados obtidos através de uma técnica de sequenciação donominada *RNA Sequencing*.

Enriquecimento de genes e análise de RBPs. Esta parte do trabalho tem como objetivo recolher informação relevante sobre os genes em estudo, de forma a ajudar os biologistas a melhor perceber a função desses genes. Informação sobre RBPs é particularmente importante para manipular genes e muito útil para melhor perceber a expressão génica.

Análise dos dados produzidos, usando técnicas de *data mining*, mais especificamente técnicas de *clustering*. Estas técnicas serão aplicadas num esforço de dar aos biologistas mais informação relevante sobre expressão génica, descobrindo possíveis relações implícitas nessa informação.

Resolver estes problemas requer o uso de ferramentas computacionais. Como tal, o desenvolvimento de um sistema informático (ou vários sistemas) para resolver estes problemas surge como um objetivo secundário da dissertação.

3. Motivação e Objetivos

A análise de expressão génica é essencial para a biologia molecular moderna. Entre muitas das possíveis aplicações desta informação, podemos destacar: melhor classificação e diagnóstico de doenças; avaliação a reação de células a um tratamento específico.

Embora existam hoje ferramentas computacionais poderosas para resolver inúmeros problemas de biologia, muitas dessas ferramentas exigem um conjunto muito específico de competências técnicas e tem uma curva de aprendizagem íngreme. Assim, a motivação mais importante por trás desta dissertação é criar ferramentas mais fáceis de utilizar. Isto significa desenvolver um sistema simples, para que qualquer utilizador possa aprender a operá-lo num curto espaço de tempo, com pouco esforço; O sistema deve também ser suficientemente avançado para atender às necessidades do utilizador.

Outro problema é a dispersão da informação e a tarefa repetitiva e prolongada de compilar essa informação. Frequentemente é necessário juntar manu-

[1] Uma *read* é um fragmento de um genoma/transcriptoma, obtido através de técnicas de sequenciação.

almente informação proveniente de um grande número de plataformas, que usam formatos e notações inconsistentes. O segundo objectivo é, portanto, desenvolver um sistema que seja capaz de realizar esta tarefa pelo utilizador, tornando o processo mais rápido e simples.

4. Projeto

O projecto revolve em torno do desenvolvimento do protótipo de um sistema informático capaz de resolver os problemas acima mencionados. Devido à complexidade deste sistema, o seu desenvolvimento seguiu uma organização modular. O sistema foi assim dividido em três modulos.

O *pipeline* de análise de expressão diferencial é responsável por alinhar *reads* contra um genoma de referência, comparando depois os resultados para diferentes amostras. O *pipeline* é baseado na ferramenta iRAP. As funcionalidades do *pipeline* são ainda reforçadas com ferramentas para configuração automática de experiências e consolidação de resultados de expressão diferencial.

O fluxo de trabalho de análise de RBPs agrega informações sobre RBPs de várias plataformas biológicas (Ensembl, NCBI, UniProt, etc.), organizando essa informação de maneiras que são úteis para os investigadores. Além disso, esta informação é agrupada usando técnicas de *data mining*, a fim de revelar os grupos de genes e RBPs que podem ter relevância biológica.

A plataforma web é responsável por armazenar e gerir dados genéticos, coordenar a interação entre os outros componentes do sistema e fornecer uma interface web para interação com os utilizadores.

5. Caso de Estudo

Dois casos de estudo foram realizados a fim de avaliar a qualidade do sistema desenvolvido.

O primeiro caso de estudo focou as melhorias que foram desenvolvidos para o *pipeline* de análise de expressão génica. Foi baseado numa experiência passada (experiência ArrayExpress *E-GEOD-48829*) que estudou a bactéria *E. coli*. O objetivo deste caso de estudo foi verificar se a ferramenta desenvolvida poderia ajudar a melhorar a confiança dos investigadores nos resultados de expressão diferencial, combinando os melhores resultados de várias ferramentas. Os resultados foram comparados com os resultados brutos e os resultados filtrados por *p-value*[2]. Conclui-se que a ferramenta reduziu significativamente o número de genes nos resultados, aumentando a confiança nesses resultados, dando ao mesmo tempo um conjunto de genes menos extenso e portanto mais fácil de analisar em experiências futuras.

O segundo caso de estudo foi realizado em colaboração com especialistas do Instituto de Biologia Molecular e Celular (IBMC). O conjunto de dados estudado era composto por vinte e três genes da família *RhoGTPase*, do genoma de *Rattus norvegicus* (rato norueguês). Este caso de estudo teve três objetivos principais: avaliar a utilidade geral da ferramenta de análise de RBPs; comparar a solução desenvolvida com as já existentes; e avaliar o impacto do desempenho dos computadores no desempenho global da ferramenta. Os resultados obtidos foram verificados em termos da sua integridade, correção e relevância do ponto de vista da biologia. A validação biológica foi realizada por especialistas IBMC. Concluiu-se que com a ferramenta desenvolvida é possível obter os mesmos resultados que um especialista obteria, numa fracção do tempo e disponibilizando mais informação útil.

6. Conclusões

Os nossos objetivos, do ponto de vista do estudo do problema e esboço de uma solução, foram totalmente cumpridos. A solução proposta corresponde a todas as nossas expectativas. No entanto a implementação do sistema de análise e alinhamento de dados RNA-Seq não foi totalmente concluída, devido a limitações de tempo. Como tal, o nosso objectivo de criar e testar um protótipo do sistema completo não foi atingido.

7. Trabalho Futuro

A continuação do trabalho proposto passaria por terminar a implementação e integração do *pipeline* de análise de dados de RNA-Seq. Isso permitiria à nossa solução funcionar conforme foi projetada, integrando todo o processo de análise. Além disso, seria interessante estudar as ferramentas desenvolvidas em termos de desempenho, quando usadas com grandes volumes de informação e pedidos. Embora as ferramentas tenham sido desenvolvidas tendo em consideração o seu desempenho, o seu funcionamento em larga escala necessita de outro tipo de infra-estrutura, que não foi considerada nesta dissertação.

Referências

[1] Lajos Pusztai, Mark Ayers, James Stec, e Gabriel N Hortobágyi. Clinical Application of cDNA Microarrays in Oncology. *The Oncologist*, 8(3):252–258, Janeiro 2003.

[2] Jochen B W Wolf. Principles of transcriptome analysis and gene expression quantification: an RNA-seq tutorial. *Molecular ecology resources*, 13(4):559–72, Julho 2013.

[3] Steven Goodman. A dirty dozen: Twelve p-value misconceptions. *Semin Hematol*, 45:135–140.

[2] O *p-value* é usado para aferir a significância estatística de resultados [3].