A ferramenta de apoio na tomada de decisão na anotação de non-coding RNAs

A tool to support decision making in the annotation of non-coding ANNs

Vinícius Gabriel Pereira
Viniciuspereira@alunos.utfpr.edu.br
Universidade Tecnológica Federal do Paraná, Cornélio Procópio, Paraná, Brasil

Alexandre Rossi Paschoal
alerpaschoal@gmail.com
Universidade Tecnológica Federal do Paraná, Cornélio Procópio, Paraná, Brasil

RESUMO

Atualmente, uma das grandes áreas de investigação da ciência é a análise de dados. A anotação de RNAs não-codificantes (ncRNAs) é uma etapa não trivial que muitas vezes envolve o uso de várias abordagens para dar confiabilidade nos resultados obtidos via métodos *in silico*.. A base para construção foi a ferramenta BEDTools, que é um poderoso conjunto de scripts para uma ampla variedade de tarefas de análise genômica e afins. Foi usado a biblioteca Tkinter da linguagem Python na criação da interface gráfica. O software desenvolvido permite as seguintes funcionalidades: (i) a entrada dos dados obtidos do relatório do programa Infernal; (ii) o comando ***Ordenação*** que vai organizar os grupos de resultados a partir da sua localização genômica; (iii) e o comando ***Merge*** que permite analisar quais resultados tem sobreposição ou não, gerando assim dois relatórios. Essa ferramenta proposta irá auxiliar os pesquisadores e os cientistas da área de bioinformática, biologia, genômica e afins para otimizar a análise de anotação de ncRNAs.

PALAVRAS-CHAVE: Ferramenta. Não codificante

. RNA.

ABSTRACT

Currently, one of the major areas of science research is data analysis. The annotation of non-coding RNAs (ncRNAs) is a non-trivial step that often involves the use of several approaches to give reliability in the results obtained by in silico methods. The basis for construction was the BEDTools tool, which is a powerful set of scripts for a wide variety of genomic and related analysis tasks. The Tkinter library of the Python language was used in the creation of the graphical interface. The developed software allows the following functionalities: (i) the input of the data obtained from the Infernal program report; (ii) the Sorting command that will organize the result groups from their genomic location; (iii) and the Merge command that allows you to analyze which results have overlapping or not, thus generating two reports. This proposed tool will assist researchers and scientists in the field of bioinformatics, biology, genomics and the like to optimize ncRNA annotation analysis**KEYWORDS:** Tools. Non-coding. RNA.

Recebido: 09 fev. 2016.
Aprovado: 12 mar. 2016.

INTRODUÇÃO

A bioinformática pode ser definida como uma área interdisciplinar que busca a aplicação de técnicas computacionais e de exatas no tratamento de dados biológicos [1]. Por exemplo, tem-se o desenvolvimento de programas computacionais que permitam reconhecer sequências de genes; predizer a configuração tridimensional de proteínas; identificar inibidores de enzimas; organizar e relacionar informação biológica; entre outros [1].

Os RNAs não-codificantes (ncRNAs) são ácidos ribonucleicos (RNAs) que é transcritos mas não traduziados em protéinas, ainda que a função de muitos deles não seja clara. A anotação de RNAs não-codificantes (ncRNAs) é uma etapa não trivial que muitas vezes envolve o uso de várias abordagens para dar confiabilidade nos resultados obtidos via métodos *in silico*. O programa Infernal (*Inference of RNA alignments*) [REF] é uma abordagem largamente utilizada para identificar ncRNAs via modelos de covariância por busca estrutural de ncRNAs. A saída, dentre vários formatos, é um resultado tabulado que necessita uma análise para sua investigação final. Principalmente se aplicado e escala genômica

Considerando que a análise pós-Infernal não é trivial e demanda conhecimentos de programação para tratamento e exploração desses resultados, este projeto desenvolveu um programa para atender esta demanda. A partir dos dados de entrada do resultado do Infernal, nosso programa modela e disponibiliza um relatório de modo fácil e amigável para o usuário fazer sua análise. Deste modo, a tarefa de anotação dos ncRNAs passa a ser possível para todos, inclusive os que não tem conhecimento de exatas/programação.

MATERIAIS E MÉTODOS

BEDTools: MANIPULAÇÃO DE INTERVALO GENÔMICO

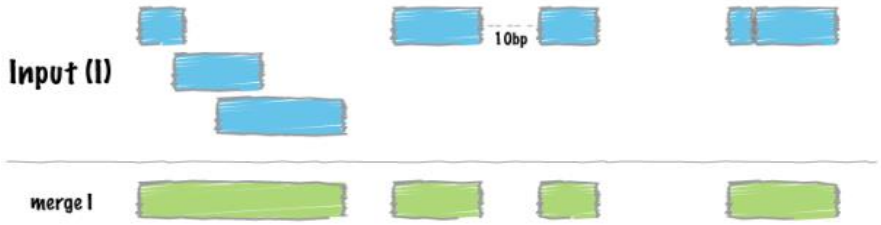
BEDTools é um poderoso conjunto de ferramentas para uma ampla variedade de tarefas de análise genômica ou aritmética de genoma. É uma ferramenta de linha de comando que permite realizar a interseção(*intersect*), combinação (merge), contagem *(count*), complemento*(complement*), e o embaralhamento(*shuffle*). Cada ferramenta individualmente realiza um trabalho relativamente simples, entretanto, se combinadas as múltiplas operações podem realizar análises bastantes sofisticadas (BEDTools, 2011).

No software desenvolvido foi utilizado o comando ***merge****, pois* combina todas as sobreposições em um arquivo de intervalo com um único recurso que abrange todos os recursos combinados (QUINLAN;KINDLON, 2017). Como ilustrado na Figura 1.

Para utilizar o comando merge é preciso que o arquivo de entrada esteja ordenado por cromossomo e depois pela posição inicial. Caso o arquivo não estiver neste formato o comando merge retornará um erro. Este pré-requisito permite

que o algoritmo de mesclagem trabalhe rapidamente sem utilização da memória. Para ordenar o arquivo é utilizado o comando ***sort*** no terminal UNIX (QUINLAN;KINDLON, 2017). Depois de se executar o comando merge é gerado um arquivo no formato BED, que foi desenvolvido pela UCSC para descrever anotações em um genoma. Ele é um formato texto, com colunas separadas por caracteres TAB que representam intervalos no genoma associados a anotações (VARUZZA, 2013).

Figura 1 - Comportamento padrão do comando merge.



Fonte: Autoria Própira

Entretanto para se alcançar tal resultado utilizando o método apresentado pôde ser muito demorado, pois é preciso ter uma noção profunda de programação e muita das vezes os pesquisadores da área de biologia não possuem tal conhecimento. Nesse contexto o objetivo deste trabalho é apresentar uma interface gráfica que visa otimizar e facilitar todo trabalho do comando *merge,* desde sua ordenação até o resultado obtido por seu comando, precisando ter apenas o programa BEDTools instalado na máquina. Deste modo, o não-especialista poderá fazer a investigação dos resultados da anotação via estrutura, bem como gerar relatórios e demais analises.

1.2 DESENVOLVIMENTO DO PROGRAMA

Para realizar este trabalho foi utilizado a ferramenta BEDTools, primeiramente é preciso instalá-la no terminal UNIX. A interface gráfica foi desenvolvida utilizando a biblioteca Tkinter da linguagem Python que acompanha a instalação padrão e permite desenvolver interfaces gráfica, ou seja, qualquer computador que tenha o interpretador Python instalado é capaz de criar interfaces gráficas. Um dos motivos de escolher o Tkinter é sua facilidade de uso e recursos disponíveis.

RESULTADOS E DISCUSSÃO

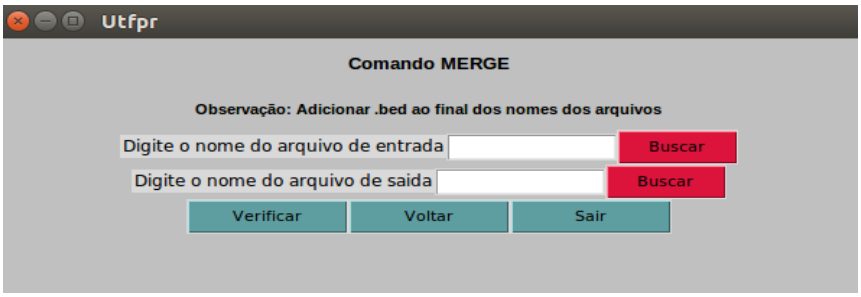
A interface gráfica possui dois comandos um é a ordenação do arquivo que está sendo inserido, o segundo comando executa o comando *merge* do BEDTools. Esses comandos são exemplificados na Figura 2. Cada comando leva a uma segunda tela, entretanto com funções diferentes, no botão *Converter para o formato Bed* vai abrir uma segunda janela, aonde é denominado o nome e a localização do arquivo de entrada e é definido também o nome e a localização onde você deseja salvar o arquivo de saída, esse comando irá ordenar as posições do menor para o maior da coluna 10 e 11 da posição do cromossomo utilizando a vertente da 10° coluna(se está coluna estiver com sinal " + " inverte os valores

daquela determinada linha da coluna 10 e 11, e se estiver com sinal " - " permanece a posição). No botão *Comando Merge* irá abrir uma segunda tela como ilustrado na figura 3 onde também deverá ser definido os nomes dos arquivos de entrada e saída, sua função já foi definida na Introdução. Essa figura também possui três botões: *Verificar*, *Voltar* e *Sair*. Ao pressionar o Botão *Verificar* depois de ter sido preenchido os campos de entrada e saída irá retornar o arquivo desejado. No botão *Voltar* você retorna a tela principal e no botão *Sair* é explicativo. Lembrando, que como já explanado na Introdução o comando Merge só deve ser executado depois que o arquivo for ordenado pelo comando *Converter para o formato Bed.* Se não ocasionará erro no Script.

Figura 2 - Página principal do Script

Fonte: Autoria Própira

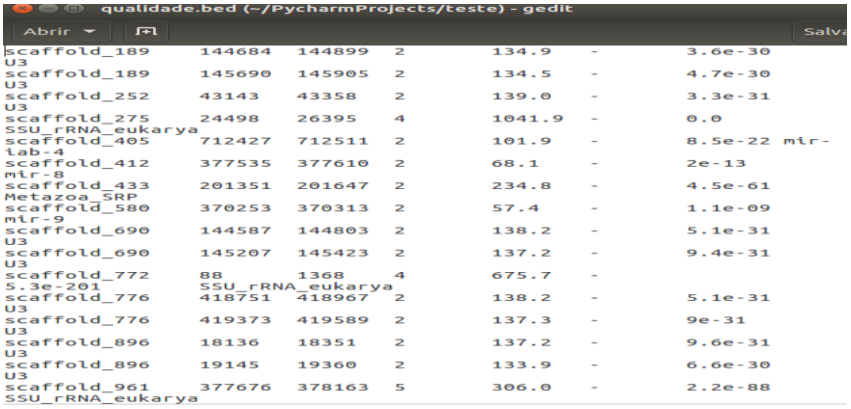
Figura 3 - Página do script para aplicar o comando Merge



Fonte: Autoria Própira

Para validação da interface gráfica, foi analisado o resultado obtido pela interface gráfica desenvolvida comparando-a com o resultado do teste de mesa feito sobre o ArquivoTeste.

Figura 4 - Sobreposições encontradas pela ferramenta



Fonte: Autoria Própira

Na figura 4 encontra-se o resultado do Comando merge, a primeira coluna se encontra o nome do cromossomo, na segunda e terceira coluna se encontra respectivamente o começo e o fim da sobreposição, na quarta coluna encontra-se o número de sobreposições(ou quantidade de fitas), por exemplo o número 2 significa que tem duas fitas, uma sobreposta a outra naquele intervalo descrito. O início do intervalo do cromossomo é descrito na coluna 2 e sua posição final na coluna 3. Entretanto o número 1 significa que não há sobreposição naquele intervalo. Prosseguindo na coluna 5 temos a probabilidade em porcentagem de haver aquela sobreposição, na coluna 6 defini a vertente, na coluna 7 está o p-Value (probabilidade de haver a sobreposição) em notação cientifica e na 8 é apresentado o nome da linha BED. A ordem das colunas do arquivo de entrada é diferente da ordem do arquivo de saída, pois têm como objetivo respeitar o padrão do formato BED estabelecido pelo BEDTools, pois alguns podem assumir colunas opcionais como estabelecido pelo formato. Na figura 4 é possível observar também que o arquivo retornado contém todas as sobreposições encontradas para cada característica.

Após feito o teste confrontado os testes o resultado foi satisfatório, pois todas as sobreposições encontradas no cromossomo pelo teste de mesa, também foram identificadas pela ferramenta como visto.

CONCLUSÕES

Com os resultados obtidos foi possível observar o potencial da ferramenta proposta e observar que seu objetivo foi alcançado. Deste modo, o usuário tem um programa *standalone* que pode ser executado em qualquer máquina sem restrição de uso de configuração. Essa ferramenta proposta irá auxiliar os pesquisadores e os cientistas da área de bioinformática, biologia, genômica e afins de forma que irá otimizar a análise de anotação de ncRNA.

REFERÊNCIAS

QUINLAN, A. Bedtools Tutorial. 2017. [Acessado em 05/10/2017]. Disponível em: <http://quinlanlab.org/tutorials/bedtools/bedtools.html>.

BEDTools: Merge. 2011. Disponível em: <http://bedtools.readthedocs.io/en/latest/content/tools/merge.html> [Acessado em 22/07/2018]. Citado na página 45.

QUINLAN, A.; KINDLON, N. Bedtools: a powerful toolset for genome arithmetic. 2017. [Acessado em 05/10/2017]. Disponível em: <http://quinlanlab.org/tutorials/bedtools/bedtools.html>.

VARUZZA, L. Introdução à análise de dados de sequenciadores de nova geração. 2013. [Acessado em 22/07/2018]. Disponivel em: <http://lvaruzza.com/files/apostila_bioinfo_2.1.pdf>.