Montagem de Regiões Gênicas

Aluno: Pedro Ivo Gomes de Faria
Orientador: Prof. Dr. Alan Durham
Trabalho de Formatura Supervisionado - Instituto de Matemática e Estatística - USP

Introdução

A montagem de sequências refere-se ao alinhamento e fusão de fragmentos (os fragmentos fundidos denominam-se *contigs*) vindos de uma molécula de DNA maior para poder reconstruir a sequência original. Isto é necessário pois a tecnologia atual de sequenciamento de DNA não consegue lidar com cromossomos inteiros, mas apenas com pequenos fragmentos (chamados de *reads*) de tamanho entre 20 e 1000 pares de bases.

Além da grande quantidade de dados gerada pelos ditos sequenciadores da "próxima geração", outros problemas incluem a presença de erros nos *reads* e a existência de sequências quase idênticas (conhecidas como repetições), que podem dificultar a montagem (gerando *contigs* que não existem na molécula original, chamados de quimeras).

Objetivos

O objetivo deste trabalho é a implementação de um *pipeline* para **montagem de regiões gênicas**. Para tentar evitar as dificuldades causadas pelas repetições, a ferramenta desenvolvida tentará apenas obter os genes (e suas regiões adjacentes) de interesse do usuário (mais precisamente, tentará montar apenas os *reads* que tenham um mínimo de similaridade com as sequências de interesse). Idealmente, iremos obter também os elementos cis-regulatórios (regiões do DNA que regulam a expressão de genes localizados na mesma molécula) dos genes em questão (um deles - a região promotora - está presente um pouco antes do início do gene, na região 5'). O usuário pode indicar quais são as regiões gênicas de interesse fornecendo os seguintes tipos de sequências:

- proteínas;
- ESTs (*expressed sequence tags*);
- *clusters* "*full length*" de ESTs.

Abordagem - estratégia de sequenciamento

Embora o *pipeline* possa ser executado para montar quaisquer tipos de fragmentos de DNA, neste projeto foram utilizadas sequências oriundas de um sequenciamento BAC a BAC, cuja abordagem facilita montagens mais precisas do que as feitas a partir do sequenciamento *shotgun*.

Isso ocorre pois em vez de gerar fragmentos de DNA de forma aleatória para todo o genoma (como ocorre no sequenciamento *shotgun*), o sequenciamento BAC a BAC divide-o em partes menores, determina de que forma essa partes são mapeadas no genoma (ou seja, "ordena" as partes) e apenas então gera fragmentos de DNA aleatórios para cada uma dessas partes. Esse processo é ilustrado pela figura a seguir:

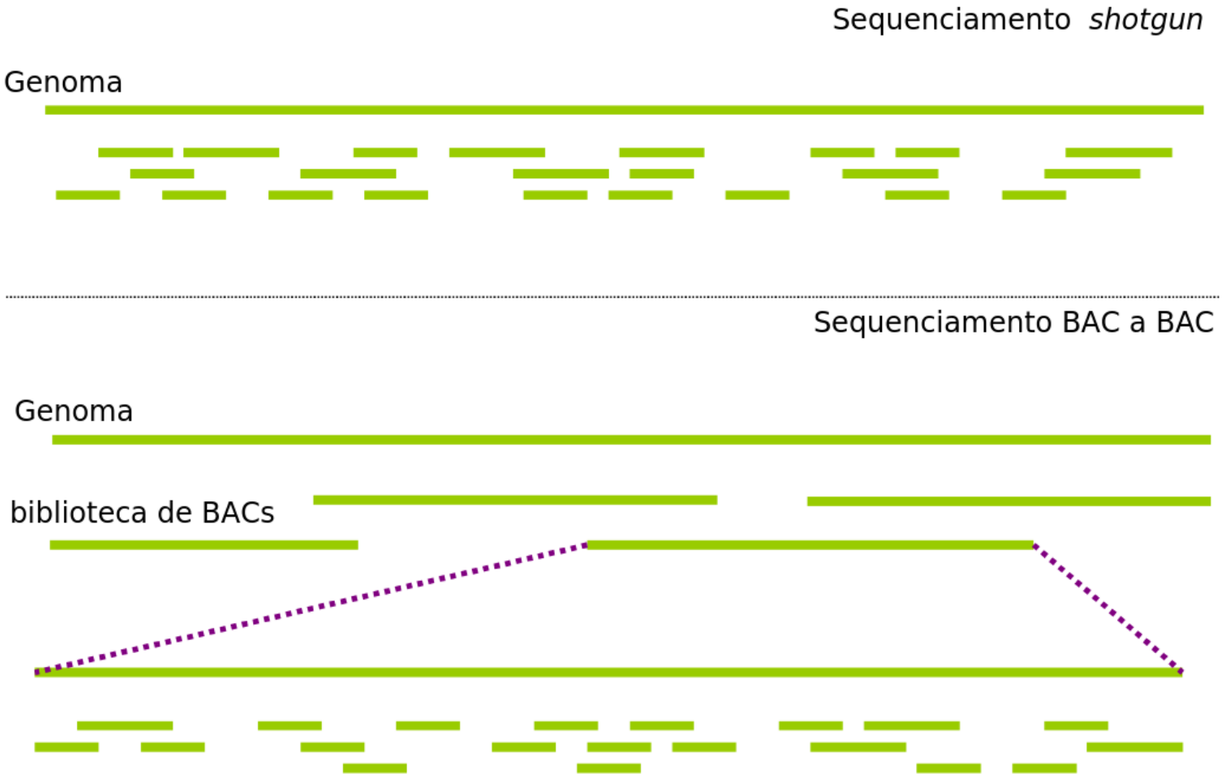

Figura: Diferenças entre os sequenciamentos *shotgun* e BAC a BAC

Abordagem - estratégia de montagem

Em vez de utilizar apenas o conjunto de fragmentos para a montagem (ab initio), o *pipeline* utiliza sequências de referência (que supostamente são similares às sequências de DNA a serem montadas) para auxiliar a montagem (montagem comparativa). Tais abordagens são ilustradas a seguir:

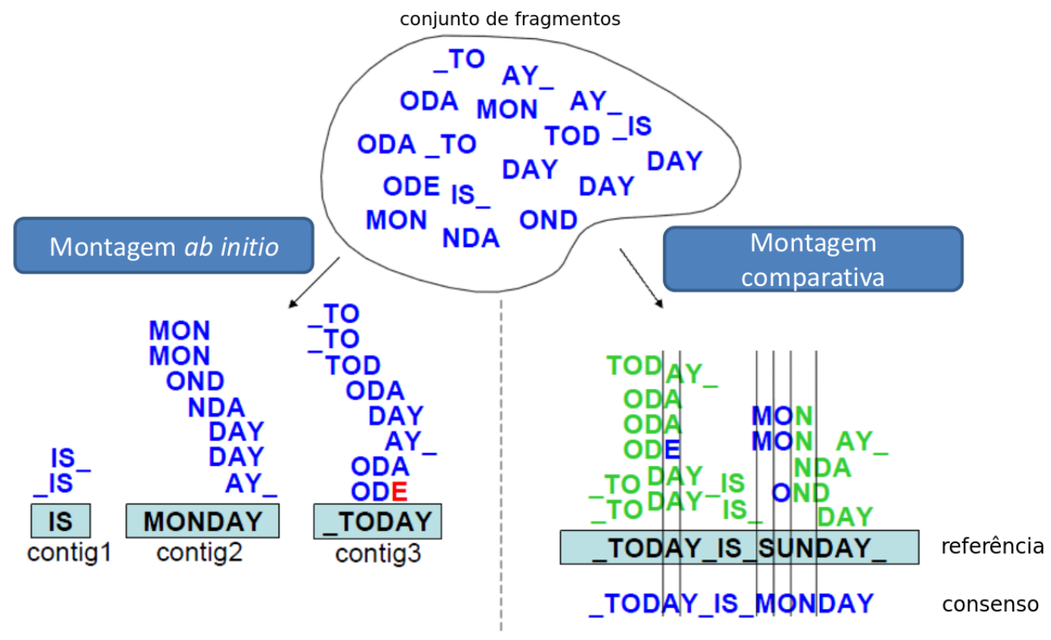

Figura: Diferenças entre as abordagens de montagem ab initio e comparativa

Resultados

O *pipeline* (implementado em Perl) foi usado para montar sequências de DNA de cana-de-açúcar (*Saccharum officinarum*), utilizando como referência as sequências de proteínas do sorgo (*Sorghum bicolor*), que é a planta de cultivo evolutivamente mais próxima da cana-de-açúcar. Para o BAC SHCRBa_218_D04, os resultados foram:

nome do contig	nome da proteína	tamanho da região 5' (pb)	tamanho da região 3' (pb)	identidade na proteína	cobertura no BAC	identidade no BAC
C_1	P_1	708	X	0,99	0,95	1
C_2	P_2	515	X	0,90	1	0,99
C_3	P_2	X	X	0,97	0,99	1
C_4	P_2	X	266	0,92	0,98	0,99
C_5	P_3	20	X	0,96	1	1
C_6	P_4	282	X	0,93	0,94	1
C_7	P_5	740	X	0,96	0,97	0,99
C_8	P_5	X	855	0,96	0,93	0,99
C_9	P_6	940	X	0,94	0,93	0,99
C_{10}	P_6	X	605	0,96	1	0,99
C_{11}	P_7	618	450	0,96	0,91	0,99

Conclusões

Para os 6 BACs que puderam ser montados e validados (pois a sequência supostamente correta do BAC era conhecida):

- aproximadamente 70% das regiões gênicas puderam ser estendidas em algum sentido (a 5' do início de tradução ou a 3' do fim da tradução);
- de modo geral, não houve ocorrência de quimeras (os contigs foram mapeados com aproximadamente 96% de cobertura e 99% de identidade na sequência supostamente correta do BAC).

Logo, o *pipeline* poderia ser utilizado como uma forma razoavelmente confiável (embora limitada) de montar regiões gênicas, com alguma chance de conseguir estender a montagem até a região promotora do gene.

Principais Referências

[1] SETUBAL, J.; MEIDANIS, J. *Introduction to computational molecular biology*. 1ª ed. Boston: PWS, 1997.
[2] ALBERTS, B. et al. Biologia molecular da célula. 5ª ed. Porto Alegre: Artmed, 2010.
[3] GRIVET, L.; ARRUDA, P. *Sugarcane genomics: depicting the complex genome of an important tropical crop*. *Current Opinion in Plant Biology* 5:122–127, 2001.