

Montagem "de novo" de genomas: uso de Decomposição em Valores Singulares para eliminação de repetições e polimorfismo de base única

Bráulio Roberto Gonçalves Marinho Couto (Autor), Lucas Felipe Silva (Co-Autor)

Na montagem "de novo" de genomas, somente dados obtidos dos sequenciadores de DNA são conhecidos, sem qualquer informação extra de espécies similares. A montagem "de novo" é um problema NP-Difícil, computacionalmente intratável. Neste trabalho, técnicas baseadas em Decomposição em Valores Singulares (SVD) são usadas na montagem "de novo" de genomas. O objetivo é apresentar os resultados do uso da técnica na primeira etapa de montagem "de novo" genomas a partir de fragmentos de DNA ou reads: a eliminação de repetições e polimorfismo de base única (NPS). Um dos principais problemas da montagem "de novo" de genoma utilizando metodologias clássicas é a necessidade de verificar se a montagem está ocorrendo de forma esperada sem a inserção de reads repetidos ou com NPS (erros de sequenciamento) na montagem de contigs. Nessas técnicas clássicas não é possível a leitura completa de todos os reads, pois demanda de grande capacidade computacional, portanto, as mesmas conseguem verificar somente as pontas dos reads e realizam uma estimativa para tentar minimizar ao máximo esse problema. Para solucionar esse problema, inicialmente, cada fragmento de sequência de DNA é codificado como um vetor de frequência de hexapletos de nucleotídeos (A, C, T, G). Ou seja, cada um dos reads da base de dados é transformado num vetor de tamanho 4^6 , considerando todas as 4.096 combinações possíveis de nucleotídeos seis-a-seis. Toda a base de dados é transformada então numa matriz M de dimensão $4.096 \times n$, onde n é o número de reads. A matriz original é decomposta pela SVD e, para a eliminação de repetições, faz-se comparações par-a-par usando cosseno como medida de similaridade entre vetores de reads. Aqueles reads com cosseno acima de um ponto de corte (por exemplo, 0,95) são considerados repetições, sendo eliminados da base de dados. A técnica foi avaliada numa base com 196.784 reads (<http://metagenomics.anl.gov/>), cujos resultados foram comparados com os métodos clássicos.

Instituição de Ensino: Centro Universitário de Belo Horizonte