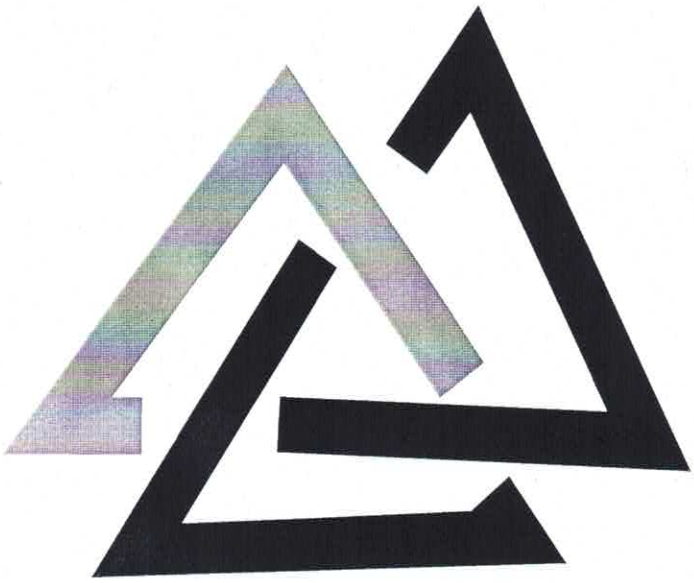
#66

Montagem de genomas completos utilizando MAQ

Bruno Zonovelli Da Silva, Carlos Cristiano Hasenclever Borges, Wagner Antonio Arbex

Resumo: O sequenciamento do DNA é um processo que determina a ordem dos nucleotídeos, em uma dada sequência, a partir de uma amostra biológica.Atualmente, as tecnologias que visam o sequenciamento do DNA sofreram grandes avanços e são capazes de gerar dados de milhões de pares de bases em uma única corrida. As plataformas de sequênciamento de nova geração (Next Generation Sequencing - NGSs) são fundamentadas no método de Sanger e estão sendo amplamente empregadas por serem procedimentos menos custosos e mais velozes do que os métodos clássicos de sequenciamento. A montagem do genoma a partir de sequências de DNA é uma tarefa exclusivamente computacional. Tendo seu início com a leitura dos arquivos originados das máquinas de sequenciamento, que após o tratamento correto, contêm as sequências de nucleotídeos e podem conter ou não as informações relativas a qualidade de sequenciamento, eles são conhecidos como FASTA, quando contém somente os nucleotídeos, e FASTQ quando contêm também a informação de qualidade. O processo de montagem de um genoma é divido em: montagem (validação e edição), scaffolding e o fechamento dos gaps ou espaços entre os contigs. A montagem de fragmentos de DNA consiste em construir uma sequência de nucleotídeos continua, construída a partir de um conjunto de fragmentos sobrepostos, essa sequência é conhecida como contig. Se o número de fragmentos for muita grande, a resolução do problema será como resolver um quebra-cabeça, que possui uma característica fundamental: a colocação das peças nos locais corretos. Por isso, uma das tarefas mais difíceis num projeto consiste na montagem dos fragmentos, principalmente quando se compara o tamanho dos mesmos com o do genoma completo. O processo de remontagem visa obter a sequência completa do DNA do genoma de um indivíduo, anteriormente sequenciado em plataformas de NGS ou não. Os arquivos contendo as sequências são armazenados em repositórios, de forma que o processo se inicia com a obtenção desses arquivos que são em geral do tipo FASTQ. O próximo passo é a definição da montagem que será utilizada como referência, em seguida as sequências são alinhadas com o genoma de referência escolhido, obtendo assim o genoma consenso, ou genoma alvo. Foram utilizados neste artigo, as sequencias do genoma de duas espécies distintas, uma animal e outra vegetal. O genoma principal é o de um animal da espécie bos taurus, raça Fleckvieh, que foi sequenciado utilizando NGS. Também foi remontado o genoma da Arabidopsis thaliana, escolhido, devido ao grande volume de informação disponível, e principalmente sequências de NGS. O software utilizado para a remontagem dos genoma foi o MAQ, que é um software de montagem e alinhamento de sequências, que utiliza a informação de qualidade para alinha-las, e trabalha principalmente com dados gerados pela plataforma Solexa. Porém, possui funções para tratar dados sequenciados na plataforma ABI SOLiD. O MAQ inicia o processo de montagem pelo alinhamento dos reads em relação ao genoma de referência, gerando em seguida os consensos. Na etapa de mapeamento ele executa o alinhamento, utilizando o algoritmo de Smith-Waterman, sem a presença de gap. Para DNA de fita única o alinhamento aceita de 2 a 3 mismatches e de 1 a 2 para fita dupla. Entretanto, esses valores podem ser alterados por meio de parâmetros definidos durante o mapeamento. Na etapa de montagem cada consenso tem um valor estatístico calculado. Esse valor é utilizado para maximizar a probabilidade posterior de cada posição do consenso. Além das funções principais o



SIMMEC EMMCOMP 2014

XI Simpósio de Mecânica Computacional
II Encontro Mineiro de Modelagem Computacional
Juiz de Fora, 28-30 de Maio de 2014, UFJF