

Departamento de Ciência da Computação – IME-USP
PIPELINE BIOINFORMÁTICO - UMA DESCRIÇÃO SUCINTA
(Rascunho)
Jesús P. Mena-Chalco
jmena@ime.usp.br

Neste documento, apresentamos uma pequena descrição do pipeline bioinformático criado, em Agosto do 2004, no Laboratório de Bioinformática do Departamento de Ciência da Computação do IME-USP, com parceria do Laboratório de Biotecnologia do Departamento de Ciências Biológicas da ESALQ-USP.

O desenvolvimento do pipeline foi realizado por Jesús P. Mena-Chalco e Henrique S. Alves, sob a direção dos professores Dr. Roberto Marcondes Cesar Jr. e Dra. Helaine Carrer.

Instâncias

Atualmente, o pipeline está configurado para a execução dos projetos de seqüenciamento e análise de ESTs correspondentes a os projetos:

- *Eucalyptus grandis*, disponível no site, <http://malariadb.ime.usp.br:8026/pipeline/>
- *Pantoea agglomerans*, disponível no site, <http://malariadb.ime.usp.br:8026/pantoea/>

1 Esquema geral

O pipeline bioinformático, representado na Figura 1 é um *Common Gateway Interface* (CGI) desenvolvido em JSP e Perl, que permite a recepção dos arquivos de cromatograma e os parâmetros de execução para o processo (*Phred*), eliminação dos vetores de clonagem e validação da qualidade dos *reads* (*Lucy*), montagem (*Phrap*), comparação por pares de nucleotídeos com outras seqüências de organismos registrados no NCBI (*Blastcl3*), e alinhamento dos contigs e singlets com uma seqüência de DNA dada como parâmetro (*MUMmer*).

1.1 Programas utilizados

- Phred version 0.020425.c
- Lucy 1.19p
- Phrap version 0.990329
- Blastcl3 2.2.9
- MUMmer 3.0
- Apache tomcat 4.1.30
- Perl 5.8.4

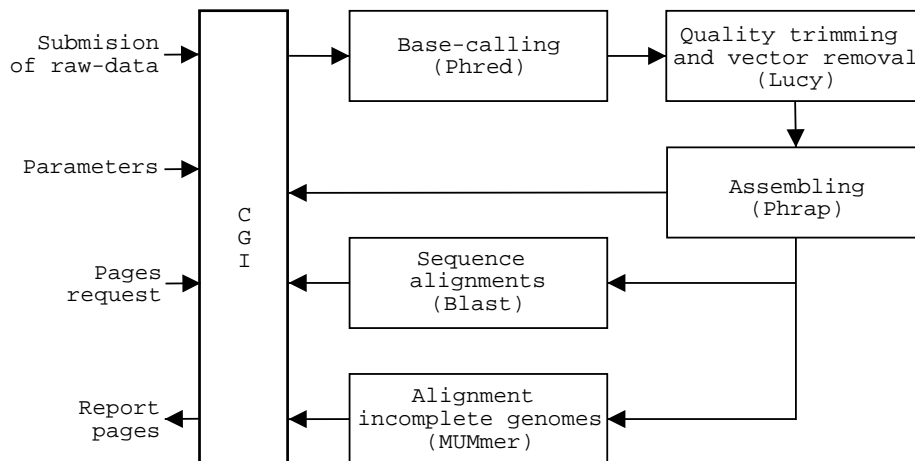


Figura 1: Esquema do Pipeline Bioinformático.

2 Implementação

O pipeline bioinformático de dados de genomas é organizado como um conjunto de etapas computacionalmente custosas:

- **Identificação das bases (base-calling).** As entradas para o pipeline são as leituras (*reads*) dos fragmentos produzidos pelo seqüenciador automático de DNA com informações analógicas que representam as bases lidas deste equipamento (*raw data*), chamados de arquivos de cromatograma, de um dos fragmentos próprios do método de seqüenciamento.

Para converter esses dados analógicos em fragmentos de bases, são submetidas diretamente a um programa de identificação das bases, denominado *base-caller*, o qual as identifica como *A*, *T*, *C* ou *G*, atribuindo um valor numérico de qualidade para cada um identificado, no caso contrario atribui o valor *N*. As bases de cada fragmento são aceitos como corretos.

O programa utilizado nesta etapa é o *Phred*, que pode gerar arquivos de saída em diferentes formatos, sendo FASTA o formato mais utilizado com informação da identificação e os valores de qualidade de cada base.

- **Validação de qualidade dos reads.** O programa utilizado nesta etapa é o *Lucy* que elimina os vetores do clonagem existentes nos fragmentos e os avalia. É considerado como valido um fragmento de qualquer tamanho cuja qualidade seja aceitável (promédio do erro máximo 0,025).
- **Montagem das seqüências.** O programa utilizado nesta etapa é o *Phrap* que consiste na geração genômica desde os fragmentos já digitalizados.
- **Comparações por pares de nucleotídeos.** O programa utilizado nesta etapa é o *Blastcl3* que compara cada contig com outras seqüências de organismos registrados no NCBI.
- **Alinhamento dos contigs e singlets.** O programa utilizado nesta etapa é o *MUMmer*.

3 Relatórios do pipeline

São apresentados relatórios correspondentes a:

- Reads.
- Singlets.
- Contigs.
- Alinhamentos.
- *Status*.

4 Configurações

- *Eucalyptus grandis*.
 - Vetores de clonagem: pCMVSPORT1, pGemT, SYNPU18CV, PUC19, pBS+, pBluescript II KS(+).
 - Comparações por pares de nucleotídeos (blastn nr): Seqüências de cloroplasto.
 - Seqüência para o alinhamento: *Nicotiana tabacum*.
- *Pantoea agglomerans*.
 - Vetores de clonagem: pCMVSPORT1, pGemT, SYNPU18CV, PUC19, pBS+, pBluescript II KS(+).
 - Comparações por pares de nucleotídeos (blastn nr): Seqüências de bactérias.
 - Seqüência para o alinhamento: *E. coli*.