"Anotação de seqüências e expansão do sistema EGene/CoEd"

Ricardo Yamamoto Abe

16 de novembro de 2006

Outline

Processamento em bioinformática

Na área de bioinformática, realiza-se freqüentemente o processamento de seqüências de bases nitrogenadas (Adenina, Citosina, Guanina, Timina e Uracil) ou dos aminoácidos formados por essas bases.

Processamento em bioinformática

Exemplo de dado de entrada (formato FASTA)

```
>Ac-A01-1705
CAGGCCCTTCCCAACTAAGGTAAACCCTAAGCCCTAAACCCTAAACCCTA
AACCCTAAACCCTAAACCCTCAAACGTGTCGATGGATGGGAGTGAACTGG
CGCACTAATGAGTAGAAAATTTGCAGACATTCTCCGCTCCCTGCGACGTA
AACGGCAGCATAATGGCGGACAGCTCCAAAATAGAAGAAGGGCTTCAACC
```

Processamento em bioinformática

O processamento envolve várias tarefas computacionais interconectadas, cada um com um protocolo de entrada e saída de dados distinto. Usualmente, é criado um script, chamado *pipeline* que executa as tarefas em ordem determinada pelo usuário – cada tarefa é denominada como componente.

Processamento em bioinformática

Com o processamento das seqüências é possível obter algumas informações, chamadas evidências, a partir das quais um biólogo pode gerar anotações, ou seja, vincular partes de uma seqüência a funções reguladoras, componentes celulares, entre outros.

Processamento em bioinformática

Entretanto, a criação de cada script torna-se onerosa no sentido de que é necessário considerar que cada par de programas que troca dados necessita de um processamento adicional para que a saída de um seja compatível com a entrada do outro.

EGene/CoEd

Nesse cenário, existe o EGene, um sistema de geração de pipelines que torna mais fácil a implementação dos mesmos. Existe ainda o CoEd, uma ferramenta gráfica que facilita a criação dos arquivos de configuração utilizados pelo EGene.

Tela do CoEd

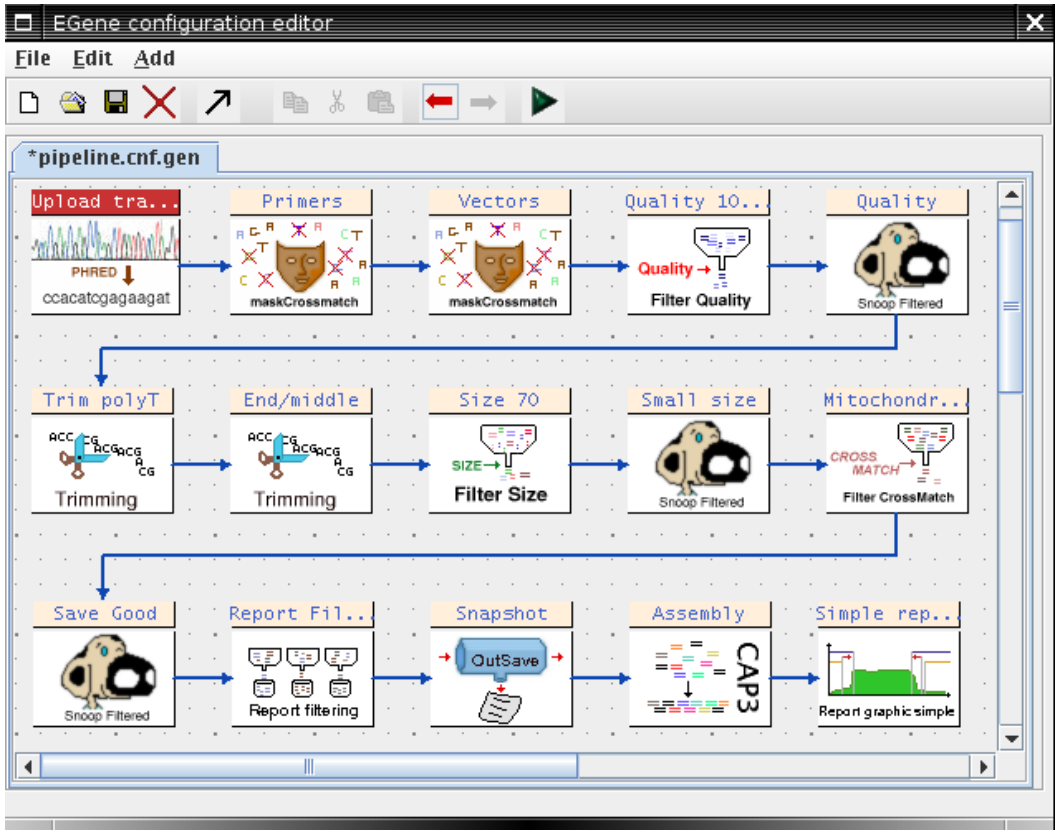

Figura: *Um pipeline gerado pelo CoEd.*

Objetivos do trabalho

O trabalho realizado durante a iniciação científica envolveu duas partes: a expansão do sistema EGene/Coed e geração de evidências.

O sistema original

A versão original do EGene permite uma única arquitetura de interligação de componentes, denominada *pipeline*. Num *pipeline*, os dados seguem um único fluxo, sendo processados seqüencialmente por cada componente.

O sistema original

Figura: *Arquitetura disponível no EGene 1.0.*

Alterações do sistema

Durante a iniciação científica, foram criadas duas novas estruturas para o processamento: *forks* e seletores.

Forks

Num *fork*, um dado pode ser enviado para vários componentes diferentes, gerando novas possibilidades de fluxo.

Forks

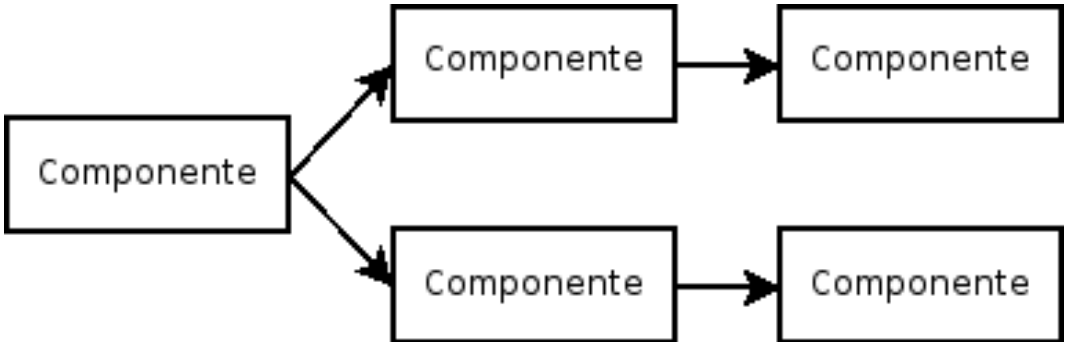

Figura: *Exemplo de arquitetura utilizando forks.*

Seletores

Com seletores, é possível encapsular um componente e fazer com que uma seqüência só seja processada pelo mesmo se uma dada condição for satisfeita.

Seletores

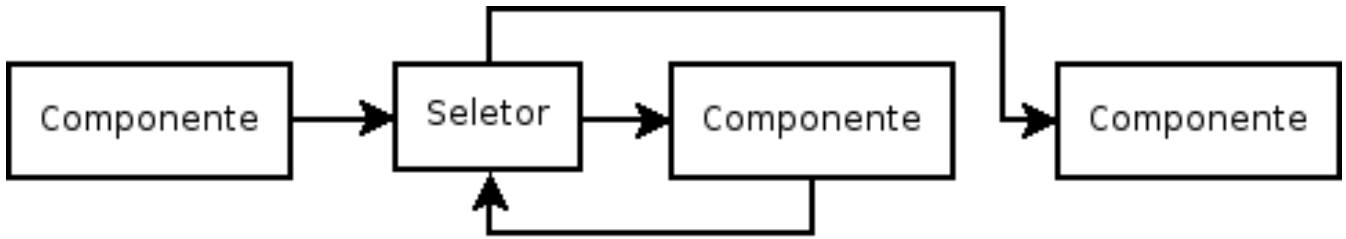

Figura: *Exemplo de arquitetura utilizando um seletor.*

Forks e Seletores

Com essas alterações, houve um ganho de desempenho de processamento. Com os seletores, é possível ignorar a execução de um ou mais componentes sobre uma dada seqüência. Com os *forks* temos ainda mais paralelismo do que o oferecido pelos *pipes* encontrados nos sistemas UNIX.

Forks e Seletores

Com forks e seletores, o EGene é capaz de implementar o processanto de programas em praticamente toda arquitetura possível em grafos direcionados acíclicos.

Arquitetura não implementável

Uma única estrutura não é implementável com as alterações: dois ou mais componentes enviando dados para um outro. Entretanto, essa limitação é aceitável no contexto de processamentode seqüências, dado que ela não é utilizada usualmente.

Arquitetura não implementável

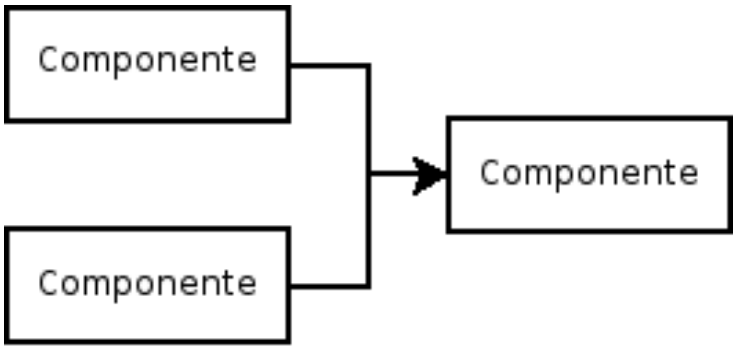

Figura: *Exemplo de arquitetura não implementável no novo EGene/CoEd.*

Geração de evidências

Os componentes do EGene original permitem basicamente realizar o pré-processamento das seqüências, eliminando contaminantes e sub-seqüências de baixa qualidade.

Geração de evidências

Em nosso grupo de bioinformática, foram desenvolvidos componentes para geração de evidências e componentes de anotação automática no formato *feature table* de submissão de seqüências anotadas.

Geração de evidências

Tomando por base uma modelagem das bases de dados do sistema original, foi feita uma ampliação da mesma de modo a permitir a inclusão dos dados de evidências. Foram identificados 4 tipos de evidências: similaridade, multi-intervalo, estatística e gráfico.

FIM