Anotação de seqüências e expansão do sistema EGene/CoEd

Ricardo Yamamoto Abe Nº USP: 3670866
Orientador: Alan Mitchell Durham – IME/USP
Co-orientador: Arthur Gruber – ICB/USP
Universidade de São Paulo – Instituto de Matemática e Estatística
ricardoy@linux.ime.usp.br

Introdução

O processamento de seqüências em bioinformática 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). 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.

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.

Nesse cenário, existe o EGene [1], 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.

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

Expansão do sistema EGene/CoEd

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.

Cada componente do EGene é um programa feito em Perl que utiliza o SequenceObject para troca de dados. O SequenceObject mantém um padrão sobre formatação dos dados das seqüencias processadas. Ele utiliza as facilidades dos *pipes* do UNIX, de modo que num *pipeline* com vários programas ocorre um ganho com paralelismo de forma transparente, sem a necessidade de haver um controle explícito de concorrência, já que o próprio sistema operacional é encarregado disso.

Durante a iniciação científica, foram criadas duas novas estruturas para o processamento: *forks* e seletores. Num *fork*, um dado pode ser enviado para vários componentes diferentes, gerando novas possibilidades de fluxo. Com seletores, é possível encapsular um componente e fazer com que uma seqüência só seja processada pelo mesmo se uma dada condição for satisfeita. Foi necessário ainda atualizar a ferramenta CoEd para que ela fosse compatível com as novas funcionalidades.

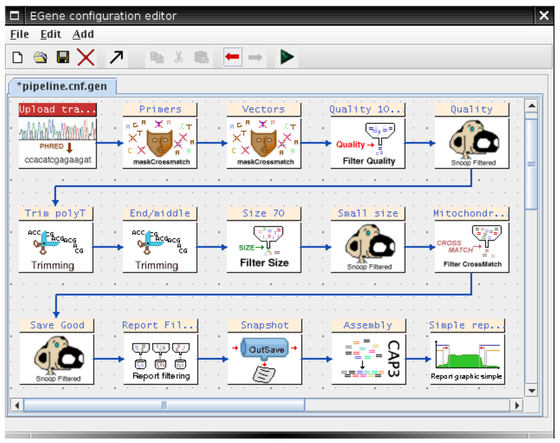

Figure 1: *Um pipeline gerado pelo CoEd.*

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

Além do aumento de arquiteturas ocorre também um ganho de desempenho. 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.

Geração de evidências

Os componentes do EGene original permitem basicamente realizar o pré-processamento das seqüências, eliminando contaminantes e subseqüências de baixa qualidade. Em nosso grupo de bioinformática, a aluna Milene Ferro desenvolveu componentes para geração de evidências e componentes de anotação automática [2] no formato *feature table* de submissão de seqüências anotadas.

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.

Os componentes para geração de evidência também foram incluídos no sistema CoEd, de modo que a utilização deles pode ser feita de maneira fácil e intuitiva.

Referências

[1] Durham, A.M., et al. EGene: a configurable pipeline generation system for automated sequence analysis. *Bioinformatics,* 21(12): 2812-2813, 2005.

[2] Ferro, M. Desenvolvimento de componentes de anotação de seqüências para o sistema EGene de geração de pipelines, 2006.