PROCESSAMENTO PARALELO DE BIOSSEQÜÊNCIAS

Aluno: Guilherme Lemos Mazie
Orientador: Sérgio Lifschitz

Introdução

Foram feitas implementações de diferentes estratégias para processamento paralelo de biosseqüências. A partir de tais execuções, foram obtidos relatórios contendo informações sobre a similaridade de um dado conjunto de seqüências de entrada com um banco de dados, ou melhor, uma base de seqüências biológicas. Para o processamento dos dados foi utilizado um cluster de 32 máquinas do Departamento de Informática da PUC-Rio.

Além disso, foi feita a tradução de um programa de formatação de bases para utilização com estratégias de processamento paralelo desenvolvidas na PUC-Rio, além do desenvolvimento de um programa alternativo com a mesma finalidade.

Por fim, também foi renovado e atualizado todo o site do Laboratório de Bioinformática.

Objetivos

Comparar os desempenhos de diferentes estratégias de processamento paralelo de seqüências, mostrando as diferenças entre eles e sua maior eficiência em relação aos métodos que utilizam processamento seqüencial.

Traduzir o programa de formatação de bases de dados biológicos, previamente escrito em PERL, para a linguagem C e desenvolver um programa alternativo, utilizando uma técnica diferente de formatação.

Tornar disponíveis as estratégias implementadas na PUC-Rio através da renovação do site do Laboratório de Bioinformática.

Metodologia

Utilizando o processamento paralelo, consegue-se um considerável ganho de tempo na comparação das seqüências e, por isso, estratégias dessa natureza são criadas com crescente freqüência. Este método se utiliza do fato de que um conjunto de computadores conectados a um servidor de controle processa comparações de forma mais eficiente do que um único computador, mesmo que muito mais moderno do que os do cluster (conjunto de computadores).

Foram implementadas e testadas cinco estratégias: mpiBLAST[1], Replicada, Fragmentada, Sob Demanda e Corretiva, sendo as quatro últimas[2] desenvolvidas no Laboratório de BioInformática da PUC-Rio.

A estratégia Replicada consiste em replicar todo o banco de dados em todas as máquinas do cluster, de forma que todo o conteúdo da base encontra-se em cada nó. Em seguida, cada seqüência do arquivo de entrada é enviado para um nó. Quando um nó termina a sua comparação, ele recebe outra seqüência. Esse método é relativamente pouco eficiente pois, apesar da troca de informações entre o servidor e os nós ser pequena, o custo de pré-processamento é muito alto pois o tempo que se leva para replicar a base é grande.

A estratégia Fragmentada Pura consiste em dividir o banco de dados em tamanhos relativamente iguais e distribuir cada fragmento para um nó do cluster. Em seguida, a

seqüência de entrada é comparada com cada fragmento em cada nó separadamente e as respostas são uniformizadas pelo servidor.

A estratégia Fragmentada Corretiva consiste numa estratégia mais complexa. Nela, o banco é fragmentado e cada nó recebe os fragmentos mas trabalha com um fragmento principal. A seqüência de entrada é enviada a todas as máquinas do cluster e a comparação é feita com cada fragmento principal. Caso um nó se torne ocioso, ou seja, tenha acabado todas as suas comparações com o fragmento principal, ele passa a trabalhar com um fragmento secundário, reduzindo o trabalho de uma outra máquina que ainda esteja em processamento.

A estratégia Fragmentada Sob Demanda, assim como a Corretiva, recebe a base de dados fragmentada e cada nó possui um fragmento principal. As seqüências do arquivo de entrada são enviadas uma a uma para cada nó e comparadas com seu fragmento principal, gerando uma resposta que é enviada para o servidor. Caso uma máquina se torne ociosa, ela recebe uma seqüência ainda não trabalhada por um nó em operação e realiza a comparação com um fragmento secundário correspondente ao principal da máquina ainda em operação.

Essas estratégias foram testadas com diferentes arquivos de entradas contendo seqüências biológicas reais e algumas criadas e comparadas com bases reais, especificamente as bases “ecoli”, nt e nr. As duas primeiras são bases de nucleotídeos, enquanto a segunda é uma base de proteínas.

Em relação à tradução do programa de formatação, ele consiste em um programa que lê um arquivo de entrada com as bases e as divide em n fragmentos pré-determinados pelo usuário de tamanhos mais homogêneos possíveis. A outra estratégia desenvolvida varia em relação aos critérios de divisão: ao invés de dividir por tamanhos homogêneos, divide-se por quantidade de seqüências e, como elas variam de tamanho, conseqüentemente os tamanhos não serão parecidos.

O site do LabBio foi renovado e atualizado de forma a colocar-se à disposição de todos as estratégias desenvolvidas.

Conclusões

Notadamente, as estratégias desenvolvidas no Laboratório de Bioinformática obtiveram desempenho superior ao mpiBLAST, na grande parte dos testes. As estratégias Fragmentada Pura e Replicada foram as que obtiveram piores desempenhos, nessa ordem, entre as quatro desenvolvidas na PUC-Rio. O desempenho da estratégia Fragmentada Pura é sempre o desempenho do nó mais lento, ou melhor, que demorou mais tempo para efetuar suas comparações. A estratégia Replicada, por sua vez, tem um tempo de pré-processamento muito alto, além de não tratar o caso da ociosidade de uma máquina também.

Por outro lado, as estratégias Fragmentada Sob Demanda e Fragmentada Corretiva obtiveram tempos muito bons, muito inferiores aos apresentados pelo mpiBLAST. Ambas evitam que um computador se torne ocioso enquanto outro ainda está em processamento, ou seja, o paralelismo está sempre acontecendo, reduzindo o tempo da melhor forma possível.

Referências

1 – http://www.mpiblast.org

2 – DE SOUSA, Daniel Xavier. **Estratégias de Balanceamento de Carga para Avaliação Paralela do BLAST com Bases de Dados Replicadas e Fragmentos Primários**. 1ª edição: Rio de Janeiro, 2007.