Alinhamento de Seqüências de DNA em Aglomerados de Computadores

Daniela Saccol Peranconi*, Gerson G. H. Cavalheiro

Programa Interdisciplinar de Pós-Graduação em Computação Aplicada
Universidade do Vale do Rio dos Sinos
Av. Unisinos, 950 - CEP 93022-000 - São Leopoldo, RS - Brasil
{danielap, gersonc}@exatas.unisinos.br

Palavras-chave

Alinhamento, seqüências biológicas, aglomerados de computadores, mensagens ativas, *multithreading*.

Alinhamento de Seqüências Biológicas

O algoritmo de Smith-Waterman [SMI 81], baseado em programação dinâmica, é considerado o mais sensitivo para alinhamento de seqüências biológicas (DNA ou proteínas), apresentando os melhores resultados. No entanto, as grandes necessidades computacionais deste algoritmo, no que se refere a utilização de memória e tempo de processamento, limitam sua utilização. O algoritmo manipula uma matriz de dados, de tamanho $(n+1)\times(m+1)$, onde cada célula contém a similaridade entre um elemento de uma seqüência X, de tamanho n, e um elemento de uma seqüência Y, de tamanho m. A matriz é preenchida de cima para baixo e da esquerda para a direita, em um processo de inundação, com o elemento (i, j) necessitando de três valores previamente calculados: $(i\text{-}1, j)$, $(i\text{-}1, j\text{-}1)$ e $(i, j\text{-}1)$.

Para a obtenção de uma implementação concorrente eficiente para este algoritmo, é necessário considerar as dependências de dados do método de programação dinâmica. Além disso, é preciso definir a granularidade adequada ao problema, de forma a evitar um número muito grande de tarefas concorrentes para que o custo necessário às sincronizações entre elas não sobreponha o potencial ganho de uma execução paralela. Uma solução ao problema de granularidade é dividir a matriz de similaridades em blocos retangulares de elementos, aumentando a granularidade do cálculo e diminuindo o número de tarefas criadas [MAR 01].

Suporte à Execução de Aplicações em Aglomerados

O objetivo deste trabalho é desenvolver uma ferramenta de comunicação em aglomerados de computadores que explore a capacidade de execução concorrente intra e entre-nós destas arquiteturas. Esta ferramenta oferece um modelo de comunicação eficiente para o processamento de alto desempenho, com um grau de abstração elevado para ativação remota de cálculos. A implementação está sendo realizada com uso de

*Bolsista PROPUP/Capes

multiprogramação leve (*multithreading*) e *sockets*. A materialização dos resultados se apresentará sob a forma de uma biblioteca de comunicação, cujo desenvolvimento foi dividido em duas partes: (*1*) implementação do mecanismo para comunicação entre os nodos do aglomerado e (*2*) integração com o ambiente Anahy [CAV 03].

Para a primeira parte da implementação está sendo utilizado o modelo de Mensagens Ativas [EIC 92, ROL 04], que demonstraram bons resultados de desempenho em experimentos preliminares [DAL 04]. O módulo de mensagens ativas serve como base para a implementação de serviços para migração de tarefas (requisição de trabalho, envio de trabalho e envio de dados) e trocas de dados (requisição de dados e retorno de dados), possibilitando a exploração da concorrência entre-nós.

A segunda parte da implementação visa tornar Anahy, que atualmente disponibiliza a exploração da concorrência intra-nó, operacional para aglomerados de computadores. Para tanto, é necessário tornar o controle da lista de tarefas, a criação e a sincronização de tarefas e o escalonamento de Anahy, atualmente aplicáveis ao contexto local a um nó, igualmente aplicáveis ao contexto distribuído. Quando da integração do mecanismo com Anahy, a presença de uma *thread* especializada em comunicação (*daemon* de comunicação) em cada nó, permitirá que os processadores virtuais de Anahy dediquem-se exclusivamente à execução de operações com cálculo útil, sobrepondo o tempo gasto em comunicações com computação [VAL 90].

A validação da ferramenta proposta será obtida através da implementação de uma aplicação para alinhamento de seqüências biológicas, utilizando o algoritmo apresentado.

Referências

[CAV 03] CAVALHEIRO, G. G. H. et al. Uma Biblioteca de Processos Leves para a Implementação de Aplicações Altamente Paralelas. In **Anais do IV Workshop de Sistemas Computacionais de Alto Desempenho**, São Paulo - Brasil, 2003.

[DAL 04] DALL'AGNOL, E. C., VILLA REAL, L. C., PERANCONI, D. S., CARDOZO JR., M. A., CAVALHEIRO, G. G. H. Construção de um Mecanismo de Comunicação para Ambientes de Processamento de Alto Desempenho. In **Anais do V Workshop de Sistemas Computacionais de Alto Desempenho**, Foz do Iguaçu, Brasil, Out. 2004 (a aparecer).

[EIC 92] EICKEN, T. von et al. Active Messages: a Mechanism for Integrated Communication and Computation. In **Proceedings the 19th Annual International Symposium on Computer Architecture, ACM SIGARCH**, v.20, n.5, p.256–266, May 1992.

[MAR 01] MARTINS, W. S. et al. A Multithreaded Parallel Implementation of a Dynamic Programming Algorithm for Sequence Comparison. In **Proceedings of the Pacific Symposium of Biocomputing**, p.311–322, Jan. 2001.

[ROL 04] ROLOFF, E. et al. Variações de Mensagens Ativas para Aglomerados de Computadores. In **Anais da 4^a Escola Regional de Alto Desempenho**, p.289–292, Pelotas, Jan. 2004.

[SMI 81] SMITH, T.; WATERMAN, M. Identification of Common Molecular Subsequences. **Journal of Molecular Biology**, v.147, p.195–197, 1981.

[VAL 90] VALIANT, L. G. A Bridging Model for Parallel Computation. **Communications of the ACM**, v.33, n.8, p.103–111, Aug. 1990.