Paralelização de um algoritmo de comparação sintática de genomas (BioDiff)

Vinicius Schacht, Marcelo Trindade Rebonatto, Marcelo Cezar pinto

Universidade de Passo Fundo
viniciuss *at* lci.upf.br, rebonatto *at* upf.br, mcpinto *at* upf.br

INTRODUÇÃO

Este trabalho apresenta uma análise da possibilidade de paralelização de um algoritmo para comparação sintática de seqüências genômicas, chamado BioDiff, elaborado pelo Professor Marcelo Cezar Pinto [PIN 02]. Este trabalho também aborda conceitos gerais de biotecnologia e biologia molecular, como DNA, RNA, cromossomos, genes, monômeros, polímeros e a estrutura formadora de proteínas e de cadeias de DNA [MEI 94]. Levando como base teórica estes conceitos, foram analisados aspectos relacionados com a dissertação de Pinto e o algoritmo desenvolvido por ele chamado de BioDiff utilizando a linguagem de programação C. Este algoritmo usa como bases teóricas principais a Teoria da Informação, o Casamento de Padrões (String Matching) e a Teoria dos Grafos, além dos conceitos biológicos já citados.

METODOLOGIA

O algoritmo seqüencial proposto por Pinto, foi analisado a fim de se entender seu funcionamento básico, desde funções até variáveis. Após esta análise, o algoritmo foi executado em um ambiente seqüencial com o intuito de se observar o funcionamento prático e os resultados obtidos. A seguir, fez-se uma avaliação do algoritmo em blocos, que foram definidos na análise inicial, para que pontos de cronometragem fossem definidos para posterior medição. Com essa medição é possível estabelecer os pontos a serem paralelizados. A paralelização está sendo realizada através da biblioteca MPI (*Mesage Passing Interface*) [PAC 97]. A paralelização está sendo realizada no cluster Colorado que pertence ao grupo de pesquisa Compadi da UPF utilizando a linguagem de programação C, já que está foi utilizada na codificação do algoritmo original.

RESULTADOS E DISCUSSÕES

Os resultados finais ainda não foram obtidos devido ao fato de o presente trabalho estar ainda em desenvolvimento. Mas alguns fatores podem ser considerados, como o fato de que dois pontos do algoritmo foram escolhidos para uma paralelização, pois apresentam laços de repetição que podem ser separados em domínios. No estágio atual algumas estruturas de dados e variáveis globais foram replicadas em todos os processos e a matriz principal, que consumia muita memória e acabava inviabilizando a execução do algoritmo com determinados genomas em máquinas comuns, foi dividida entre os

processos. No momento o algoritmo está sendo adaptado para que os resultados sejam correspondentes ao do original.

CONCLUSÃO

O algoritmo pode ser paralelizado, conseguindo talvez alcançar níveis de *speed-up* superescalar [KUM 94], diminuindo consideravelmente seu tempo de execução e agilizando a comparação das seqüências de DNA.

Referências

[KUM 94] KUMAR, V. et al. **Introduction to parallel computing:** design and analysis of algorithms. Califórnia: The Benjamin/Cummings, 1994.

[MEI 94] MEIDANIS, J.; SETÚBAL, J. C. **Uma introdução à Biologia Computacional**. In: IX ESCOLA DE COMPUTAÇÃO, 9, 1994, Recife. Trabalhos apresentados… Recife: UFPE, 1994, 132 p.

[PAC 97] PACHECO, P. S. **Parallel Programming with MPI**. San Fransisco: Morgan Kaufmann, 1997. 418 p.

[PIN 02] PINTO, M. C. **Um algoritmo para comparação sintática de genomas baseado na Complexidade Condicional de Kolmogorov**. Dissertação de Mestrado – Instituto de Computação, Unicamp, Campinas, 2002, 64p.

[MPI 94] MPI FORUM. **The MPI message passing interface standard**. Knoxville: University of Tennessee, 1994.

[STE 99] STERLING, T. L. et al. **How to build a beowulf**: a guide to the implementation and application of PC clusters. Cambridge: MIT, 1999. 239p.