1.

Introdução

Biologia Computacional e Bioinformática são termos que definem uma área científica e tecnológica, cujo desenvolvimento permite prover métodos avançados no sentido de gerenciar e extrair informações úteis, a partir de seqüências de DNA, RNA e proteínas [1]. Dessa maneira, a Bioinformática tem, como um dos seus objetivos, prover métodos computacionais para copiar e interpretar dados adquiridos em diversos projetos de seqüenciamento de genomas, assim como de novas tecnologias em Biologia Molecular [2]. Para isso, a Bioinformática utiliza métodos de várias áreas da Ciência da Computação, tais como algoritmos, linguagens formais, redes neurais, bases de dados, mineração de dados, entre outras [1].

A análise de similaridades entre seqüências, obtidas a partir dos vários projetos de seqüenciamento, é normalmente utilizada na descoberta de funções biológicas. Para efetuar essa análise utiliza-se normalmente o *Basic Local Alignment Search Tool*, ou simplesmente BLAST, que se constitui em um conjunto de aplicações capaz de encontrar similaridades estatisticamente significativas entre seqüências a partir da avaliação do seu alinhamento. BLAST utiliza heurísticas que ajudam a acelerar o processo de análise, tornando a sua execução bastante eficiente em relação a outros algoritmos utilizados com a mesma finalidade [3].

BLAST é um software sofisticado, que se tornou um dos mais importantes na área de Bioinformática. Ele é utilizado com a finalidade de encontrar alinhamentos significativos entre um conjunto de seqüências de consulta e aquelas existentes em uma base de dados de seqüências biológicas [3].

Apesar de rápido [3], BLAST foi originalmente projetado para utilização em equipamentos de pequeno porte, com o objetivo de analisar pequenas seqüências de consulta e bases de dados [9]. Em função disso, apresenta limitações quando

trata de grandes bases de dados, como aquelas hoje existentes, e que crescem de forma exponencial em função de um volume cada vez maior de dados oriundos de projetos de seqüenciamento de genomas [3].

No entanto, as características de processamento intensivo, necessária à comparação de seqüências biológicas, e paralelização simples conduziram a vários estudos no sentido de executar o BLAST em ambientes de alta performance como *clusters*, e mais recentemente *grids* [18, 26, 28, 29], inclusive àquele que resultou no desenvolvimento do balaBLAST, na PUC-Rio [19].

O presente trabalho possui dois objetivos, a saber:

a) Complementar a avaliação da ferramenta balaBLAST, no que diz respeito ao aspecto de balanceamento de carga;

b) Efetuar uma avaliação do balaBLAST frente a outras ferramentas BLAST voltadas ao ambiente distribuído, considerando não somente o tempo de execução, mas também outros custos envolvidos, como é o caso do processamento de leituras e gravações em disco durante a sua execução.

A principal motivação diz respeito às pesquisas que lidam com a bioinformática e bancos de dados. Nesse sentido pretende-se:

a) Expandir as avaliações além de uma simples verificação do tempo de execução;

b) Medir não somente a eficiência obtida no processo de balanceamento de carga, mas também verificar a sua eficácia;

c) Implementar ferramentas que auxiliem as avaliações; e

d) Para fins de comparação, foi desenvolvida uma nova ferramenta voltada para o ambiente de grid.

O trabalho está organizado como se segue: no Capítulo 2 são apresentados os fatores e critérios a considerar na avaliação e comparação de ferramentas de software, pertinentes à análise do desempenho; no Capítulo 3 são relacionados os trabalhos pertinentes ao tema, incluindo alguns tópicos sobre a ferramenta BLAST e implementações do BLAST em ambientes distribuídos; no Capítulo 4, é relatado como foi efetuada a avaliação do balaBLAST quanto ao aspecto de

eficácia do balanceamento de carga; no Capítulo 5, é exposto como foi efetuada a avaliação e comparação de desempenho de ferramentas BLAST, incluindo o balaBLAST, em ambiente distribuído; no Capítulo 6, é apresentada uma pequena conclusão.