5
Conclusão e Trabalhos Futuros

Não existem dúvidas com relação à importância da genômica comparativa para o estudo de fatores relativos à evolução dos organismos, ou conhecimento mais aprofundado da genética, bioquímica e fisiologia dos mesmos, além de gerar dados com grande potencial de aplicações práticas, como, por exemplo, na identificação de genes marcadores e genes apropriados para o desenvolvimento de novas drogas.

O grande problema está em organizar estes dados, entender como eles se relacionam, e gerenciar e desvendar todo tipo de informação escondida nestes dados. Se por um lado a bioinformática acelerou o processo de geração de dados biológicos, por outro, ela originou novos desafios. Além da grande quantidade de dados gerados, há muitos tipos de dados e cada um tem sua própria comunidade e seus próprios repositórios. Além disso, existe o agravante de que esses dados biológicos são armazenados, na sua maioria, em arquivos no formato texto.

Uma contribuição desta Tese está relacionada neste sentido. Primeiramente abordamos a dificuldade de representar e organizar os conceitos envolvidos no domínio da biologia molecular. Neste domínio, a modelagem/padronização de conceitos são realizadas sobre informações que na verdade são muitas vezes suposições e/ou inferências sobre a forma como a vida realmente é.

O modelo de dados proposto, além de trazer um padrão para a representação dos dados da biologia molecular, traz uma abstração dos conceitos e, por utilizar a metodologia de Entidade e Relacionamento (ER), facilita o processo de análise e entendimento de suas relações.

Apesar de não ser uma unanimidade entre a comunidade científica, o uso de SGBDs para gerenciar dados biológicos não é uma novidade. Conforme apresentado no Capítulo 2, inúmeras propostas de pesquisa tem surgido como alternativas para armazenar ou melhorar o desempenho de consultas. No entanto, a grande maioria dos trabalhos relacionados apresentados não está preocupada com a manipulação e obtenção de informação biológica. A única

preocupação é armazenar os dados sem nenhum mecanismo específico de acesso.

A outra contribuição desta Tese está relacionada neste sentido. Não houve uma preocupação com o armazenamento físico de sequencias genômicas, mas sim em como representá-las e como obter e manusear a informação da forma mais simples, sem a necessidade de conhecimentos prévios da biologia molecular. Para tanto, apresentamos uma alternativa de solução para representar sequencias biológicas.

Com relação aos trabalhos futuros, podemos apresentar algumas alternativas. No que diz respeito ao grande volume de dados e a relação entre dados em disco e memória RAM disponível, fica claro que é necessário dispor de soluções eficientes para que não ocorram gargalos de processamento. Uma das possibilidades seria dispor de máquinas de alto poder de processamento e um grande disco para armazenamento. Outra alternativa seria dispor de estruturas de armazenamento de dados mais eficientes.

Além disso, podemos pensar em aplicar algumas ideias já discutidas e apresentadas por [Seibel et. al. 2003], [Macedo et. al. 2007a] e [Macedo et. al. 2007b], por exemplo. Já, para aproveitar ao máximo os recursos disponíveis, explorar as noções de drivers dedicados e escalonadores ad-hoc para aplicações de bioinformática, e.g. [Lifschitz and Mauro 2005] e [Noronha 2006], são estratégias que se mostram bastante interessantes.

Outra possibilidade interessante seria investigar o uso de bancos de dados distribuídos, tecnologia já consolidada para bases de dados convencionais, porém pouco explorada para bases de dados biológicos e sequências genômicas. No caso, o próprio SGBD distribuído lida com os dados não mais centralizados e gerencia a transparência do acesso, do ponto de vista dos usuários. Além disso, aproveita-se a maior disponibilidade de máquinas, logo maior poder de processamento, e otimiza-se cada acesso específico. Também pode-se aproveitar a distribuição dos dados e aplicar técnicas de paralelismo no processamento. Para esta alternativa podemos sugerir utilizar as técnicas e metodologias apresentadas por [Costa and Lifschitz 2003] e [Souza 2007].

Por fim, podemos elencar outros trabalhos de pesquisa relacionados à:

- Atualização da base de dados com novas proteínas e anotações;
- Extensão do modelo proposto e criação de novas consultas/funções;
- Extensão do banco para suportar comparações blast e índices sobre sequências;

- Criação de um tipo específico para representar sequências genômicas;
- Criação e utilização de visões materializadas para melhorar o desempenho de consultas; e
- Aplicar técnicas de *data mining* para obter algum tipo de padrão e extrair algum tipo de informação.