Banco de dados de genótipos da Rede genômica animal

Vinícius Fernandes Dias[1]
Roberto Hiroshi Higa[2]

O Banco de Dados de Genótipos (BDG) é um banco cumulativo (só é permitido incluir e acessar dados, eles nunca são removidos) que compreende o armazenamento organizado e padronizado de conjuntos de dados resultantes de procedimentos de genotipagem em larga escala, além de alguns resultados de análises básicas cujo resultado é utilizado por diversas análises subsequentes. Genericamente, seu objetivo é prover suporte para sistemas computacionais que implementem a (semi-) automação de processos de análises, vinculados a programas de melhoramento animal que utilizem dados de genotipagem.

Os dois processos encarregados de garantir a interface com o BDG têm por objetivo disciplinar a forma como esses acessos são realizados. No caso do processo “Padronizar e armazenar”, seu objetivo é garantir que, juntamente com os conjuntos de dados, sejam incluídos no BDG metadados que permitam a posterior recuperação destes conjuntos de dados. Já o processo “Recuperar e padronizar” tem por objetivo garantir que os conjuntos de dados recuperados do BDG sejam formatados em um formato padrão conhecido, facilitando assim sua utilização em procedimentos de análise subsequentes.

Tendo em vista a granularidade considerada nas manipulações de dados comumente utilizadas ao realizar essas análises, optou-se por manipular de forma agregada as informações de genótipo de um indivíduo e o painel de marcadores utilizados para genotipagem de um conjunto de amostras. Para isso, utiliza-se o conceito de campos do tipo *Binary Large Object* (BLOB)

[1] Universidade Estadual de Campinas - vinicius.dias@colaborador.embrapa.br
[2] Embrapa Informática Agropecuária - roberto.higa@embrapa.br

de sistemas de bancos de dados relacionais. Com este tipo de modelagem evita-se uma granularidade excessiva dos dados e uma consequente superpopulação de registros em algumas tabelas, o que dificulta a realização de consulta sobre esses dados. A desvantagem é o armazenamento de dados redundantes, ficando as aplicações que acessam esses dados, responsáveis manipulá-los de forma consistente.

Para analisar a viabilidade desta abordagem para as dimensões esperadas dos conjuntos de dados a serem armazenados no BDG, estão sendo realizados testes para analisar a viabilidade do modelo. Utilizando o banco *Postgresql* (POSTGRESQL, 2013) foi criada uma tabela de genótipos com campos numéricos (como id do genótipo), texto (descrição) e um campo do tipo BLOB, que armazena um arquivo compactado contendo as informações de genótipo. Cada arquivo possui 5 colunas e 700000 linhas de dados gerados aleatoriamente, apenas para testes de tempo.

Foram feitos scripts na linguagem Python para manipulação do banco de dados utilizando a bibilioteca *psycopg2* (PSYCOG, 2013), e mediu-se o tempo de inserção e seleção de dados no banco. O tempo para inserir inclui a compactação do arquivo texto de 700000 linhas e a carga no banco e o tempo de consulta consiste em fazer a busca de todas as tuplas presentes e descompactar o binário presente no campo BLOB da Tabela 1.

Tabela 1. Tempo de inserção e seleção de dados no banco.

Inserção		Seleção	
Itens inseridos	Tempo	Itens selecionados	Tempo
10	20s	1000	4m27s
1000	34m50s	5000	29m32s
5000	168m00s	15000	83m20s
10000	333m00s		

É possível observar que o tempo de inserção e seleção crescem de maneira linear com o tamanho da entrada. Entretanto, no momento, mais testes estão sendo realizados com o intuito de verificar o comportamento do banco para maiores volumes de dados e confirmar a tendência observada.

Referências

POSTGRESQL. **Postgresql**. 2013. Disponível em: <http://www.postgresql.org/>. Acesso em: 30 set. 2013.

PSYCOPG. **Psycopg – PostgreSQL database adapter for python**. 2013. Disponível em: <http://pythonhosted.org/psycopg2/>. Acesso em: 30 set. 2013.