PUC

ISSN 0103-9741

Monografias em Ciência da Computação
n° 28/09

Protein World Database: Geração do Esquema Lógico e Processo de ETL

Cristian Tristão
Sergio Lifschitz

Departamento de Informática

PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO DE JANEIRO
RUA MARQUÊS DE SÃO VICENTE, 225 - CEP 22451-900
RIO DE JANEIRO - BRASIL

Monografias em Ciência da Computação, No. 28/09 ISSN: 0103-9741
Editor: Prof. Carlos José Pereira de Lucena Outubro, 2009

Protein World Database: Geração do Esquema Lógico e Processo de ETL

Cristian Tristão, Sergio Lifschitz

{ctristao, sergio}@inf.puc-rio.br

Abstract. The automated methods development of DNA sequencing on a large scale, coupled with the technologies development for high performance computing and more efficient algorithms, has provided generation of a large amount of data, and enabled the scientific community to study the genomes structure, organization and evolution. Today, the main challenge is the biological data organization, storage and availability, without compromising the biological systems interpretation and understanding, and their interactions. This report describes the logical schema used to store the data from the Protein World Database project and integrate data from various public sources, e.g. RefSeq, Swissprot, NCBI Taxonomy, Pfam, KEGG and GO, as well as the difficulties encountered in the data Extraction, Transformation, and Loading process (ETL). This scheme allows obtaining information relevant to the similarity analysis between proteins, basic process in functional annotation and new proteins discovery.

Keywords: Logical Schema, Biological Databases, Data Integration, Sequence Alignment, Protein World DB.

Resumo. O desenvolvimento de métodos automáticos de seqüenciamento de DNA em larga escala, aliado ao desenvolvimento de tecnologias de computação de alto desempenho e de algoritmos mais eficientes, tem proporcionado a geração de uma grande quantidade de dados, e permitido à comunidade científica o estudo da estrutura, organização e evolução de genomas. Hoje, o principal desafio é a organização, armazenamento e disponibilização desses dados biológicos, sem que haja comprometimento da interpretação e compreensão dos sistemas biológicos e suas interações. Este relatório descreve o esquema lógico utilizado para armazenar os dados provenientes do projeto Protein World Database e integrar dados de diferentes fontes públicas, e.g. Refseq, Swissprot, Taxonomy NCBI, Pfam, KEGG e GO, bem como as dificuldades encontradas no processo de Extração, Transformação, e Carga (ETL) dos dados. Este esquema possibilita a obtenção de informações relevantes para a análise de similaridade entre proteínas, processo fundamental na anotação funcional e descoberta de novas proteínas.

Palavras-chave: Esquema Lógico, Base de Dados Biológicos, Integração de Dados, Alinhamento de Seqüências, comparação de proteínas, Protein World DB.

Responsável por publicações:

Rosane Teles Lins Castilho
Assessoria de Biblioteca, Documentação e Informação
PUC-Rio Departamento de Informática
Rua Marquês de São Vicente, 225 - Gávea
22451-900 Rio de Janeiro RJ Brasil
Tel. +55 21 3527-1516 Fax: +55 21 3527-1530
E-mail: bib-di@inf.puc-rio.br

1. Introdução

O avanço tecnológico apresentado nesses últimos anos proporcionou um aumento no poder computacional e no desenvolvimento de novas e mais eficientes técnicas de processamento e pesquisa. Na área da biologia, a bioinformática acelerou o processo de estudo do material genético de inúmeros organismos, levando a crer, por exemplo, que a análise sistemática de todo o conteúdo genético de um organismo tem o potencial de levar à compreensão integral da genética, da bioquímica, da fisiologia e da patogênese dos microrganismos, sendo capaz de concretizar-se através do estudo comparativo dos genomas ou de regiões sintênicas de duas ou mais espécies, subespécies ou cepas.

Um exemplo de estudo computacional do material genético é o Projeto Comparação de Genomas [1], projeto este executado em parceria com o *World Community Grid* [4], onde as seqüências codificantes para proteínas de todos os genomas completos depositados no RefSeq (versão 21) [3] foram comparadas entre si, utilizando-se para tal o programa SSEARCH [5], uma implementação gratuita e aberta do algoritmo de *Smith-Waterman* [2]. O resultado do programa SSEARCH foi um conjunto de dados de similaridade, os quais permitem a obtenção de critérios para a anotação de genes e famílias gênicas, correção e reanotação de genes, classificação funcional de proteínas, assim como diversos estudos onde a organização e evolução dos genomas são abordadas.

Por serem de extrema importância para o avanço científico, e úteis em vários aspectos, os dados de similaridade obtidos devem ser disponibilizados integralmente para a comunidade científica. No entanto, mesmo levando em consideração o avanço tecnológico alcançado, por ser uma área de pesquisa em constante avanço e descobertas, essa não é uma tarefa simples de ser implementada por dois motivos: (I) grande volume de dados gerado, e (II) divergência entre conceitos e relacionamentos.

O grande volume de dados gerado é um fator que pode inviabilizar o armazenamento, manipulação e disponibilização dos dados. O Projeto Comparação de Genomas resultou em um volume de dados superior a 900 GB. Estes dados correspondem a informações de similaridade entre as proteínas pertencentes aos genomas utilizados.

Atualmente, na literatura, não existe um consenso referente aos conceitos biológicos e seus relacionamentos, impossibilitando a geração de um esquema de dados genético amplo o bastante para representar diferentes domínios de pesquisa, ou específico o bastante para representar novas áreas de pesquisas com conceitos próprios. Dados biológicos são extensos e diversos, englobando vários domínios de conhecimento como biologia molecular e celular, genética, biologia estrutural, farmacologia e fisiologia. Cada domínio contempla tipos de entidades sobrepostas e complementares, e com suas próprias terminologias e necessidades de dados. Além disso, a variedade de procedimentos experimentais e analíticos resulta em dados relacionados, porém não idênticos. Um exemplo de conflito entre conceitos é apresentado por Macedo [19]:

- Termos têm o mesmo nome, mas diferentes semânticas: por exemplo, a palavra colônia é usada em zoologia para significar um grupo de animais da mesma

espécie que vive junto e depende um do outro. Em microbiologia, colônia é um grupo de microorganismos desenvolvidos a partir de uma única célula.

- Termos têm diferentes nomes com a mesma semântica: por exemplo, um esquema de dados usa o termo gene para nomear o conceito gene, enquanto outro esquema usa a palavra "Gene Humano" para definir o mesmo conceito de gene.

Nota-se que existem diferentes iniciativas de pesquisas focadas na modelagem de dados para bioinformática (e.g. [6, 7]). Contudo, nenhuma das soluções propostas nesses projetos se aplica a esta pesquisa em questão. Levando em consideração o problema da divergência entre conceitos e relacionamentos envolvidos no projeto de comparação de proteínas e dificuldade de reutilizar algum esquema de dados genéticos presente na literatura, foi desenvolvido um esquema conceitual de dados genéticos para representar e contextualizar os conceitos e relacionamentos envolvidos no alinhamento entre proteínas de genomas completos, conforme descrito em [25].

Após o desenvolvimento do esquema conceitual [25], o passo seguinte foi a elaboração, construção e carga do esquema lógico para a realização de pesquisas e análises sobre os dados de similaridade obtidos no Projeto Comparação de Genomas. Somado a esses dados, encontram-se dados de diferentes fontes públicas, e.g. Refseq, Swissprot, Taxonomy NCBI, Pfam, KEGG e GO (*Gene Ontology*), que ajudam na análise e identificação de novas proteínas. Este relatório apresenta a geração do esquema lógico proposto, mostrando as principais decisões e alternativas de modelagem, bem como as dificuldades e cuidados tomados no processo de ETL (*Extraction, Transformation, and Loading*) dos dados provenientes de fontes públicas, uma vez que, somados, esses dados ultrapassam a marca de 1 TB.

O restante desse documento está organizado da seguinte forma: uma visão geral do projeto *Protein World DB* descrevendo sua origem, objetivos, etapas e estado atual de desenvolvimento é apresentado no Capítulo 2. No Capítulo 3 é descrito o processo de mapeamento do esquema conceitual [25] para o esquema lógico, apresentando as regras de transformação e algumas decisões tomadas durante o processo. O processo de ETL das tabelas existentes no esquema lógico, a partir dos dados de comparação de similaridade e bases de dados externas, é descrito no Capítulo 4. Por fim, no Capítulo 5, são expostas as conclusões obtidas com o desenvolvimento deste trabalho, bem como algumas possibilidades de trabalhos futuros.

2. Protein World DB

O Projeto Comparação de Genomas é um projeto da equipe de Bioinformática do Laboratório de Genômica Funcional e Bioinformática do Instituto Oswaldo Cruz (IOC), FIOCRUZ, com o objetivo de calcular o grau de similaridade entre seqüências, comparando o conteúdo de proteínas de genomas completamente seqüenciados de centenas de organismos, incluindo os seres humanos e várias outras espécies com grande importância na indústria, medicina, comércio, ou pesquisa. Para a realização do processo de comparação de proteínas foram utilizados os recursos de computação distribuída fornecidos pelo *World Community Grid*, os quais utilizam recursos ociosos de computadores, possibilitando a distribuição de tarefeas e otimização do

processamento. A informação genômica, obtida no processo de comparação, pode ser usada para melhorar a qualidade e interpretação de dados biológicos, assim como a compreensão dos sistemas biológicos e das interações ambientais. Esta informação pode desempenhar um papel crítico no desenvolvimento de melhores medicamentos e vacinas, bem como métodos de diagnóstico.

As seqüências utilizadas para as comparações no Projeto Comparação de Genomas foram seqüências de aminoácidos obtidas de duas bases de dados distintas: o *Reference Sequence* (RefSeq) do NCBI, um conjunto de seqüências (que podem ser genômicas, mRNAs, RNAs, ou de proteínas) não redundantes e bem anotadas, e o swissProt, uma base de dados de proteínas anotadas. Elas possuem organização e propostas diferentes com distintas formas de acesso.

Na primeira fase do projeto, utilizando o *World Community Grid*, foi realizada a comparação do tipo "todos contra todos" através do algoritmo de *Smith-Waterman* [2] em mais de 2,8 milhões de seqüências de proteínas de aproximadamente 3.774 organismos, incluindo vírus, e dentre estes organismos, mais de 400 possuem a seqüência completa do genoma decifrada. A maioria dessas seqüências protéicas é originada a partir da análise computacional de genomas por parte de muitos grupos de pesquisa desde os anos sessenta. Elas são depositadas em bancos de dados públicos, juntamente com a anotação funcional, que na sua maioria são preditas computacionalmente. Para a análise comparativa de genomas, as seqüências foram agrupadas em blocos contendo 2.000 seqüências cada, e mais que 1 milhão de comparações do tipo bloco-a-bloco foram feitas. A partir de 20 de dezembro de 2006, 4 milhões de comparações foram realizadas. A conclusão dessa fase se deu em 31 de março de 2007.

Para a segunda fase do projeto, o conjunto inicial de dados foi atualizado com dados genômicos mais recentes (RefSeq), sendo adicionadas 393.999 novas seqüências protéicas. Além disso, um novo conjunto curado de dados de referência (SwissProt), representando mais 254.609 seqüências, foi incluído nas comparações. Esta segunda fase foi concluída em 14 de maio de 2007

Finalmente, um conjunto de dados experimental representando mais de 3 milhões de seqüências potencialmente codificantes foi adicionado na tentativa de identificação de novas seqüências codificantes. Este conjunto de dados foi derivado de *Open Reading Frames* (ORFs) com mais de 300 pb (par-base) para as quais uma predição clássica de seu potencial codificante não alcançou resultados positivos. Apenas as ORFs integralmente contidas em regiões descritas como sendo não-codificantes foram incluídas; qualquer tipo de sobreposição com uma seqüência codificante anotada como tal resultou na exclusão da referida ORF do conjunto de dados analisados. Em termos de tamanho, os dados gerados a partir do Projeto Comparação de Genomas somam, compactados, aproximadamente 300Gb; expandidos, estes dados ocupam quase 900Gb de espaço em disco. Esta fase final do projeto terminou em 21 de julho de 2007.

Ao todo, a tarefa de comparação de proteínas levou aproximadamente 5 meses de processamento no *World Community Grid*. Um exemplo de uma linha do resultado do programa SSEARCH pode ser visto na Figura 1. Somente a linha contendo os valores é armazenada. A linha superior contém os descritores dos valores: query gi (identificador da seqüência query), subject gi (identificador da seqüência subject); SW

score (pontuação Smith-Waterman), bit score, e-value, % identity (percentual de identidade), alignment length, (tamanho do alinhamento), query start (início da seqüência query), query end (final da seqüência query), subject start (início da seqüência subject), subject end (final da seqüência subject), query gaps (gaps na seqüência query), subject gaps (gaps na seqüência subject).

```
query gi, subject gi, SW score, bit score, e-value, % identity, alignment length, query start, query end, subject start, subject end, query gaps, subject gaps

67523787,67540134,2166,488.8,2.6e-138,0.336,1320,35,1275,67,1367,79,19
```

Figura 1. Resultado da execução do algoritmo de Smith-Waterman

Um dos problemas a ser resolvido no contexto desse projeto diz respeito ao sistema de banco de dados que fará o armazenamento dos dados e dará suporte às pesquisas. A colaboração do grupo de pesquisa em Bioinformática do Departamento de Informática da PUC-Rio se deu nesse sentido: desenvolver um banco de dados para armazenamento dos dados resultantes do processo de comparação de proteínas, integrar dados de diferentes fontes públicas, e disponibilizar o acesso dos mesmos a toda comunidade científica.

Como solução e resultado da parceria entre Fiocruz e PUC-Rio, foi desenvolvido o banco de dados biológico *Protein World DB*, com o patrocínio da IBM e *World Community Grid*. Ele é o primeiro produto do Projeto Comparação de Genomas. Algumas de suas funcionalidades incluem a recuperação de identificadores, anotação, termos de ontologias e domínios de proteínas. Também pode-se realizar pesquisas de similaridade utilizando a ferramemta BLAST. Outras características incluem a seleção de genes únicos e aglomerados de proteínas (pré-processados e armazenados em banco de dados).

Para o desenvolvimento do *Protein World DB*, várias questões foram e estão sendo levadas em consideração. Em primeiro lugar a questão de persistência dos dados, que já são da ordem de um terabyte. Sabe-se que não é suficiente apenas comprar mais dispositivos de armazenamento com maior capacidade para que se resolva a questão de acesso e busca eficientes. Os sistemas gerenciadores de bancos de dados (SGBD) ajudam no caso de bancos de dados ditos convencionais. Porém, essa questão ainda é um problema em aberto para a biologia. Não existe comprovação de que dados de projetos genômicos podem ser armazenados e tratados convenientemente pelos SGBDs mais populares, que usam o modelo relacional de dados, como é o caso aqui.

3. Projeto Lógico

Após a definição do esquema conceitual do *Protein World DB*, conforme descrito em [25], realizou-se o mapeamento e geração do esquema lógico. Normalmente surgem dúvidas quanto à diferença entre modelo conceitual e lógico. Resumidamente, a modelagem conceitual é o mais alto nível de abstração, onde busca-se identificar todas as entidades e os relacionamentos existentes entre elas, sem limitações ou aplicação de tecnologia específica. O diagrama de Entidade e Relacionamento (ER) é o mais usado. O esquema lógico, por sua vez, já leva em conta algumas limitações e implementa alguns recursos como adequação de padrão e nomenclatura. Também deve-se pensar

em chaves primárias e estrangeiras, levando em conta as entidades e relacionamento criados no esquema conceitual. É uma preparação para o esquema físico (esquema gerado para um SGBD específico).

No processo de transição do esquema conceitual para o esquema lógico, foram utilizadas algumas regras básicas de mapeamento e, quando necessário, algumas decisões de modelagem foram tomadas. De acordo com uma dessas regras, todas as entidades presentes no esquema conceitual tornam-se tabelas no esquema lógico com seus respectivos atributos. Deste modo, as entidades PROTEIN, ORF_T, GENOMIC SEQUENCE, CDS, GENE, TAXONOMY, RANK, DOMAIN, GENE_ONTOLOGY e ENZYME foram mapeadas para as tabelas protein, orf_t, genomicsequence, cds, gene, taxonomy, tax_rank, domain, gene_ontology e enzyme respectivamente.

Levando em consideração a aplicação dessa regra, o restante deste capítulo apresenta o mapeamento dos relacionamentos presentes no esquema conceitual para o lógico, descrevendo as outras regras empregadas, e algumas das particularidades e dificuldades encontradas durante o processo. Por fim, apresentamos o resultado desta transformação como um todo, e detalhamos os atributos de cada tabela resultante.

3.1. Hits

O resultado da primeira etapa do Projeto de Comparação de Genomas foi o conjunto de arquivos contendo dados de similaridade (hits) entre proteínas e ORFs traduzidas (tORFs). Como a relação de similaridade entre proteínas e tORFs é do tipo "muitos-para-muitos", foi preciso criar uma tabela intermediária para cada relação, tendo como chave primária a composição das chaves primárias das outras duas tabelas envolvidas no relacionamento. As tabelas criadas foram: (a) hits_oo, relação entre tORFs, (b) hits_pp, relação entre proteínas, e (c) hits_op, relação entre tORFs e proteínas. A Figura 2 ilustra esse processo de mapeamento.

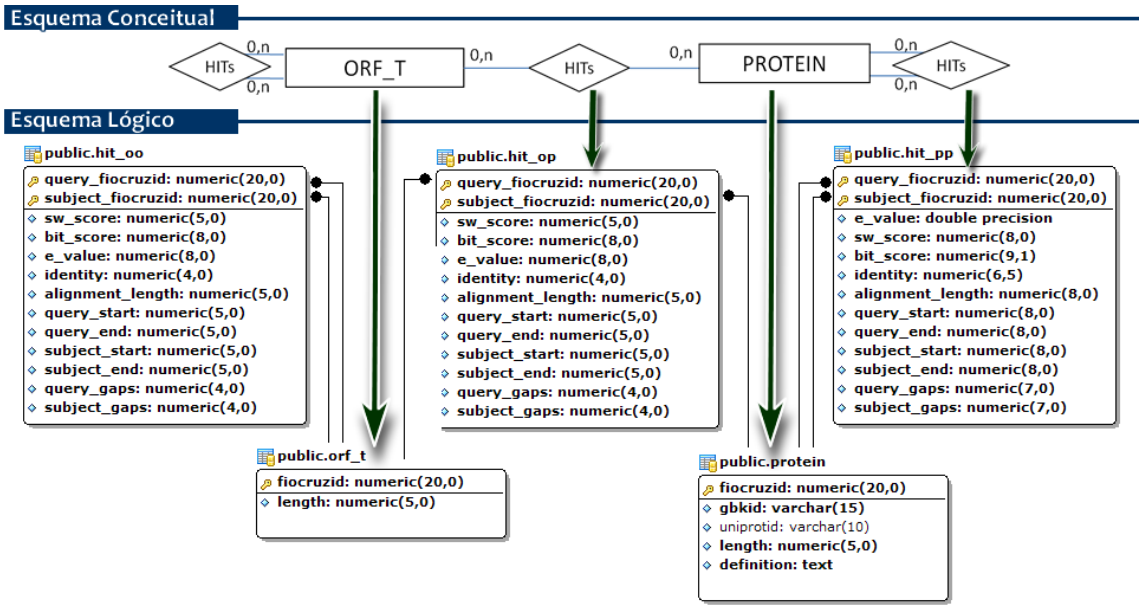

Figura 2. Mapeamento da similaridade (hits) entre proteínas e tORFs

As tabelas intermediárias (hits) devem receber os atributos que identificam a relação entre as entidades. Neste caso, cada tabela possui os atributos que identificam o alinhamento entre cada proteína/tORF, que são: query_fiocruzid (identificador da seqüência query), subject_fiocruzid (identificador da seqüência subject); sw_score (pontuação Smith-Waterman), bit_score, e_value, identity (percentual de identidade), alignment, (tamanho do alinhamento), query_start (início da seqüência query), query_end (final da seqüência query), subject_start (início da seqüência subject), subject_end (final da seqüência subject), query_gaps (gaps na seqüência query), subject_gaps (gaps na seqüência subject).

3.2. Taxonomia

De acordo com [8], taxonomia é uma excelente forma de organização no estudo de sistemas biológicos. Herança, homologia por descendência comum e à conservação de seqüência e estrutura na determinação da função são todas as idéias centrais em biologia que estão diretamente relacionados com a história evolutiva de um grupo de organismos. Para o mapeamento da taxonomia, os relacionamentos devem ser analisados em separado. A Figura 3 ilustra o resultado da aplicação das regras de transformação explicitadas abaixo:

- Relacionamento entre taxonomias: é um tipo de auto-relacionamento "um-para-muitos". Com ele representamos a linhagem taxonômica dos indivíduos, onde cada nodo apresenta um link (referência) para o seu nível taxonômico superior. Dessa forma a tabela *taxonomy* recebe um novo atributo denominado "*father*" (*foreing key* para *taxonomy*). A cardinalidade "0.1" da extremidade filha indica que essa referência pode ser nula, caso usado na representação de nodos do tipo raiz, os quais não possuem nodo pai.

- Relacionamento entre taxonomia e *rank*: esse é um tipo de relacionamento "um-para-muitos". Por isso a tabela taxonomia recebe o atributo de referência ao *rank* relacionado, denominado tax_rank_id. O fato da cardinalidade ser "1.1" indica que a taxonomia deve possuir um e somente um *rank*, ou seja, não aceita valor nulo.

- Relacionamento entre taxonomia e proteína/tORF: da mesma forma que o anterior, esse é um tipo de relacionamento "um-para-muitos" de cardinalidade "1.n". Por isso tanto a tabela protein quanto a tabela orf_t recebem o atributo de referência taxonomy_id. Isso indica que proteínas e torfs devem possuir uma taxonomia, e por sua vez, uma linhagem taxonômica.

Uma característica muito importante é que o atributo synonymous é do tipo composto. Por isso houve a necessidade de retirar esse atributo da tabela taxonomy e de criar uma tabela denominada "synonymous" que possui os diferentes nomes de uma mesmo indivíduo. Ela também deve possuir o taxonomy_id (*foreing key*) para referenciar o indivíduo relacionado.

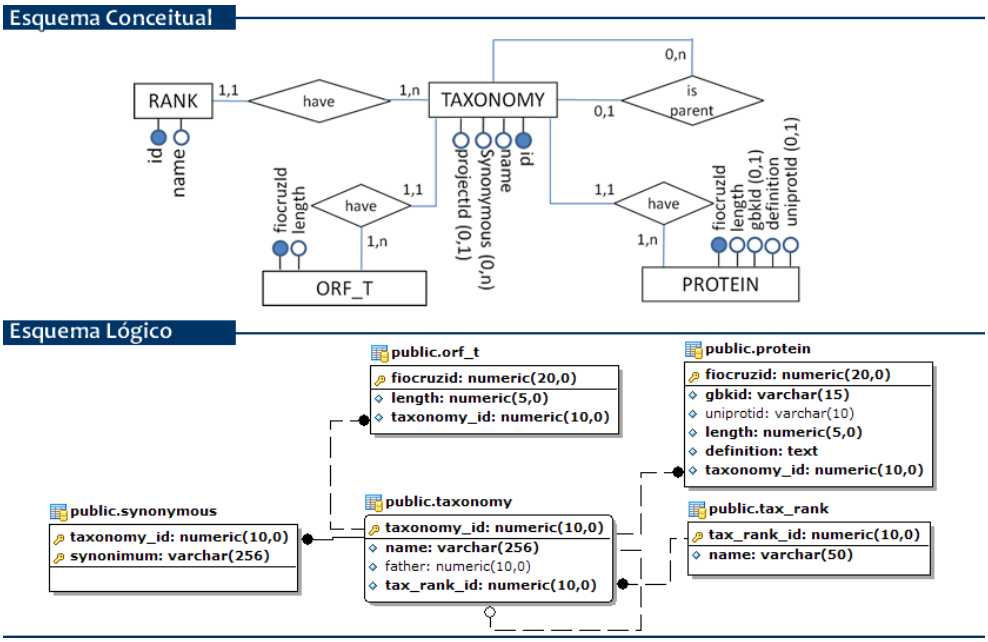

Figura 3. Mapeamento da taxonomia

3.3. Dogma Central da Biologia Molecular

O dogma central da biologia afirma que a informação genética codificada em DNA é transcrito em blocos transportáveis individuais, composto de RNA mensageiro (mRNA). Cada bloco de mRNA contém o programa para a síntese de uma proteína particular (ou número pequeno de proteínas) . Este trio crítico de macromoléculas – DNA → RNA → proteínas - está presente em todas as células [24]. Conforme descrito em [25], a parte do esquema conceitual que representa o dogma central da biologia molecular foi simplificado e adequado às necessidades de pesquisa e disponibilidade de dados.

A seguir são apresentadas as regras de transformação utilizadas no mapeamento dos relacionamentos envolvidos na parte destinada ao dogma central da biologia molecular, conforme ilustrado pela Figura 4:

- Relacionamento entre torfs e seqüência genômica: de acordo com o esquema conceitual, uma seqüência genômica pode conter zero ou mais orfs, e uma orf é proveniente de uma e somente uma seqüência genômica. Desta forma, esse é um tipo de relacionamento "um-para-muitos". Uma alternativa para de mapeamento para o modelo lógico seria introduzir na tabela torf uma referência para a seqüência genômica que a origina, além dos atributos existentes nesse relacionamento. Contudo, foi decidido criar uma tabela denominada de orf_region para armazenar os atributos deste relacionamento por motivos organizacionais e pelo fato de representarem a região da seqüência genômica onde a orf é obtida. Por fim, foram acrescentados na tabela orf_region os atributos de referência à torf, fiocruzid (*primary key*), e seqüência genômica, gbkid (*foreing key*).

- Relacionamentos de CDS: como CDS possui uma relação do tipo “um-para-muitos” com seqüência genômica e gene, cada relação é convertida na inserção de atributos de referência, gbkid e geneid respectivamente, à “tabela” CDS. Por sua vez, CDS possui um relacionamento do tipo "um-para-um" com proteína. Para este tipo de relacionamento existe mais de uma forma de mapeamento. No entanto, optou-se por referenciar a tabela proteína em CDS pelo atributo fiocruzid.

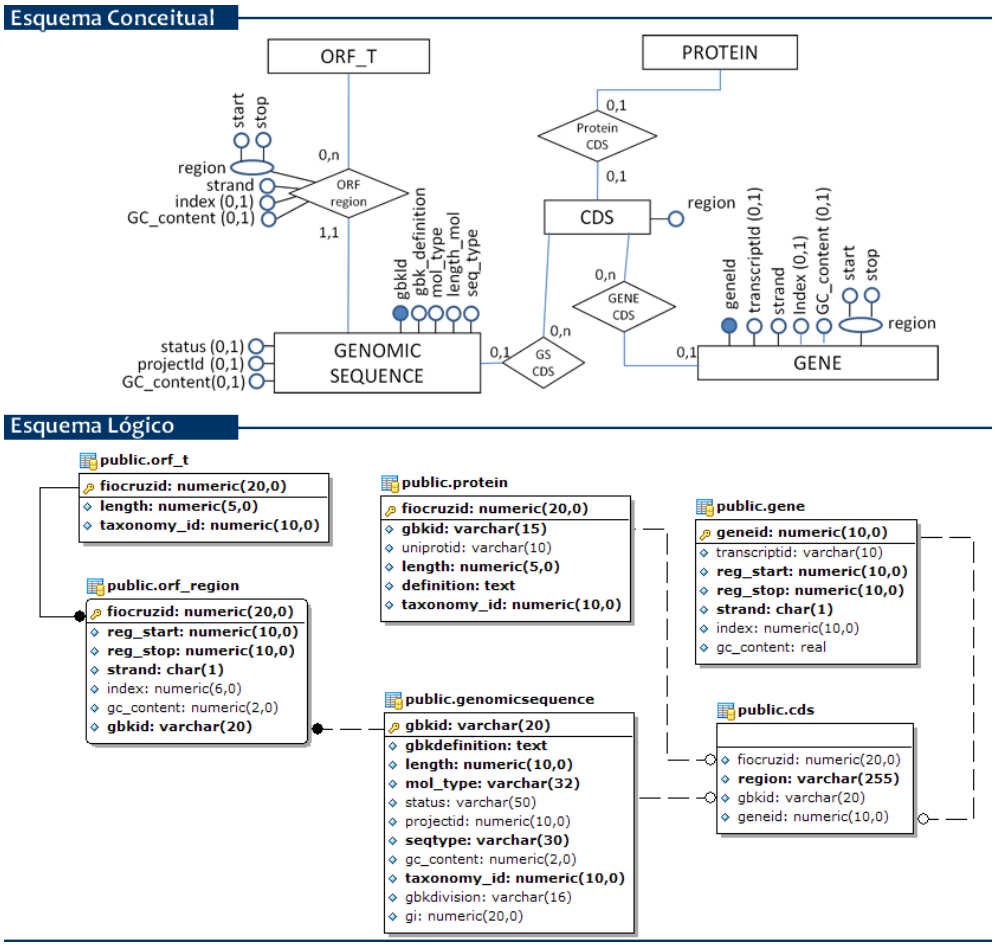

Figura 4. Mapeamento do dogma central

Neste ponto do mapeamento houve uma segunda mudança no esquema lógico em relação ao conceitual. Por motivos de praticidade e desempenho, foi acrescentado na tabela genomicsequence uma referência para taxonomia (taxonomy_id). Assim, essa informação, que é muito utilizada em consultas referentes à seqüência genômica, é obtida diretamente sem a necessidade de percorrer outras tabelas.

3.4. Anotações

De acordo com [27], anotação é processo de atribuição das funções biológicas preditas e características estruturais aos dados brutos, e.g. à seqüência primária da proteína. Vale ressaltar que a predição das funções celulares (estruturais, enzimas, transportadores, sinalizadores, etc.) é em sua maioria hipotética. A maior parte dessas possíveis funções

foi atribuída por análises *in silico* e somente uma pequena fração das proteínas preditas teve suas funções confirmadas por experimentos laboratoriais.

Com base nessas informações, conclui-se que o relacionamento de proteínas com suas anotações é do tipo "muitos-para-muitos". Sendo assim, o mapeamento é realizado criando-se uma tabela intermediária para cada relacionamento, os quais terão como chave primária a composição das chaves primárias das outras duas entidades/tabelas envolvidas, podendo ainda acrescentar os atributos que identificam essa relação.

Como resultado do mapeamento da relação da entidade *protein* entre as entidades *domain*, *enzyme* e *gene ontology*, foram criadas as tabelas domainannotation, enzyneannotation e goannotation respectivamente. Por fim, foram mapeados os auto-relacionamentos de enzyme e gene ontology, os quais podem ser visualizados graficamente pela Figura 5:

- Enzyme: por ser um auto-relacionamento do tipo "um-para-muitos" foi aplicada a mesma regra empregada em taxonomia. A tabela *enzyme* recebe um novo atributo denominado "*father*" (*foreing key* para *enzyme*), podendo ser nulo, conforme cardinalidade presente no esquema conceitual ("0.1").

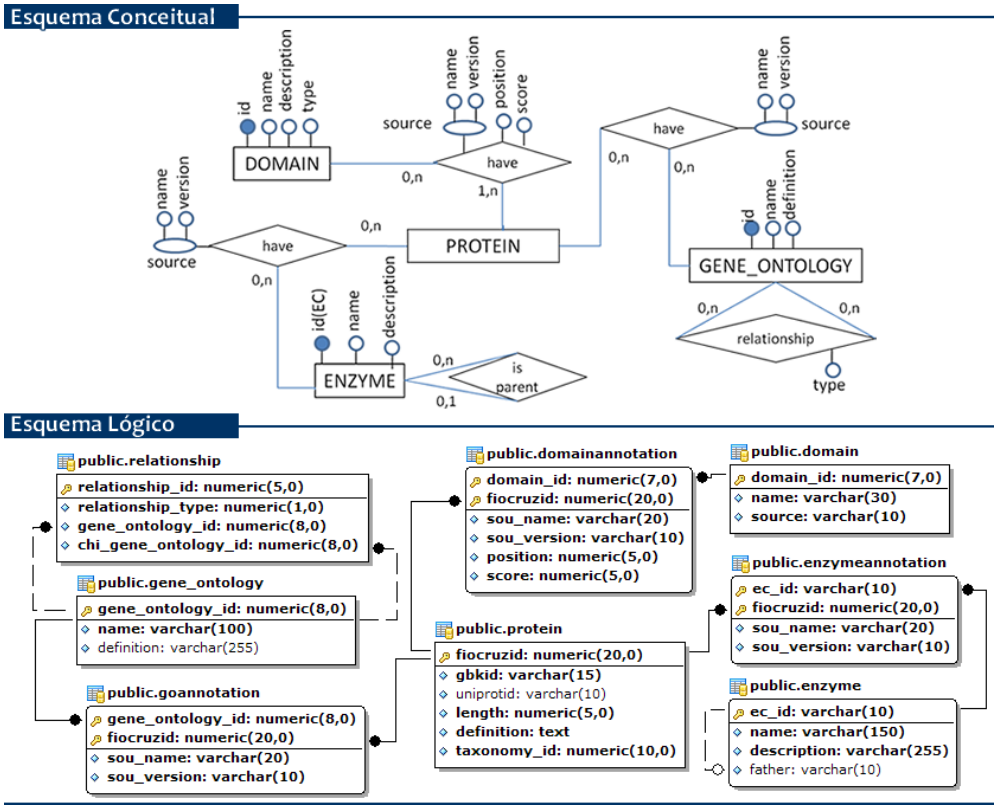

Figura 5. Mapeamento das anotações

- Gene ontology: o auto-relacionamento de gene ontology é do tipo "muitos-para-muitos". Por esse motivo foi criada a tabela relationship. Nela foram acrescentadas as referências para a tabela *ontology*, o atributo tipo de relacionamento (relationship_type) e um atributo seqüencial denominado

relationship_id (*primary key*). A necessidade da criação desse atributo para identificação de chave primária, e a não utilização dos atributos de referência à ontologia como chave, deve-se ao fato da existência de diferentes tipos de relacionamentos para uma mesma relação.

3.5. Visão Geral

As Figura 6 e Figura 7 ilustram respectivamente o esquema conceitual original, elaborado e descrito em [25], e o esquema lógico resultante do mapeamento utilizando o conjunto de regras de transformação descritos nesse capítulo.

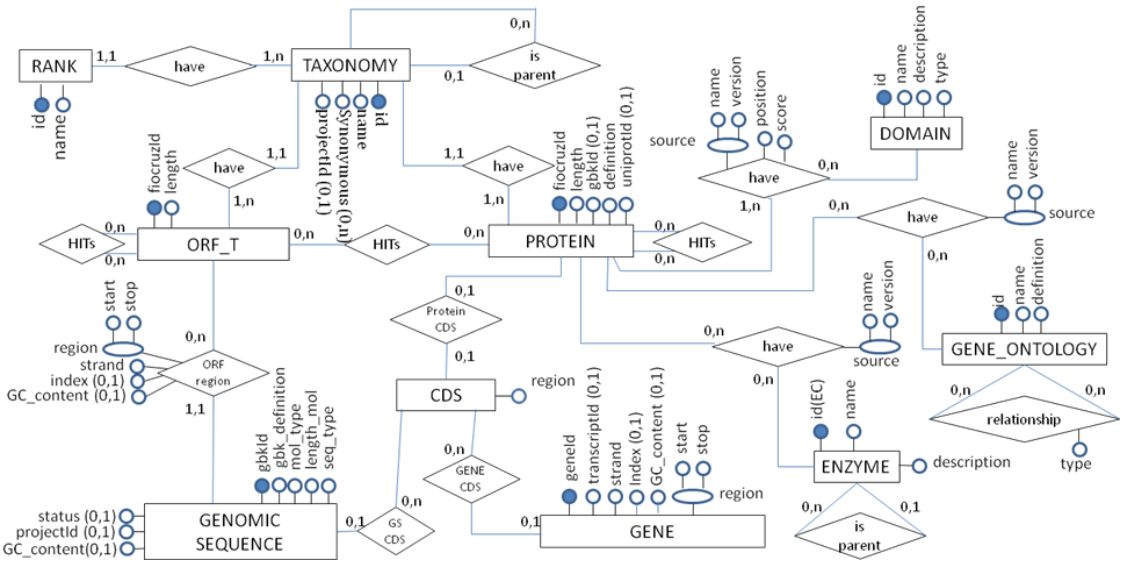

Figura 6. Protein World Database: esquema conceitual

Outra característica muito importante na elaboração do esquema lógico está relacionada ao tipo e tamanho definidos para cada atributo. Pesquisas na literatura e, principalmente a expertise do grupo de pesquisa, foram fundamentais na concepção do esquema. A seguir são descritos os principais atributos que necessitaram de uma pesquisa mais profunda para sua definição:

- Identificador de proteína/tORF (fiocruzid): por não haver uma padronização, cada projeto de pesquisa ou banco de dados público define e utiliza um identificador próprio para proteína. Para o Projeto Comparação de Genoma não foi diferente. O processo de comparação envolveu proteínas de diferentes bases, e conseqüentemente diferentes identificadores, e seqüências de aminoácidos não catalogadas, as tORFS, que receberam um identificador próprio. Todos esses identificadores são numéricos e nenhum ultrapassa a casa das 10^{20} posições. Sendo assim, nosso identificador de proteínas/tORFs, e demais referências, foram definidos como numérico de 20 posições.

- Atributos de tamanho e posição: todos os atributos que se referem ao tamanho ou posição em seqüências de aminoácidos foram definidos com tamanho 5, pois de acordo com estatísticas do release 57.8 do UniProtKB/Swiss-Prot, datado em 22/09/2009, a maior seqüência de aminoácidos presente na base é a da proteína

TITIN_MOUSE (A2ASS6) que possui a quantidade de 35213 aa (aminoácidos), enquanto que atributos que se referem ao tamanho ou posição em seqüências de nucleotídeos foram definidos com tamanho 10, pois, de acordo com dados do Projeto Genoma Humano (HGP), coordenado pelo *U.S. Department of Energy and the National Institutes of Health* [26], o comprimento do cromossoma humano varia de 51 milhões a 245 milhoes de bp (base-par), e o gene *dystrophin* é o maior gene encontrado na natureza medindo 2.4 Mb, segundo dados do Entrez Gene NCBI [10].

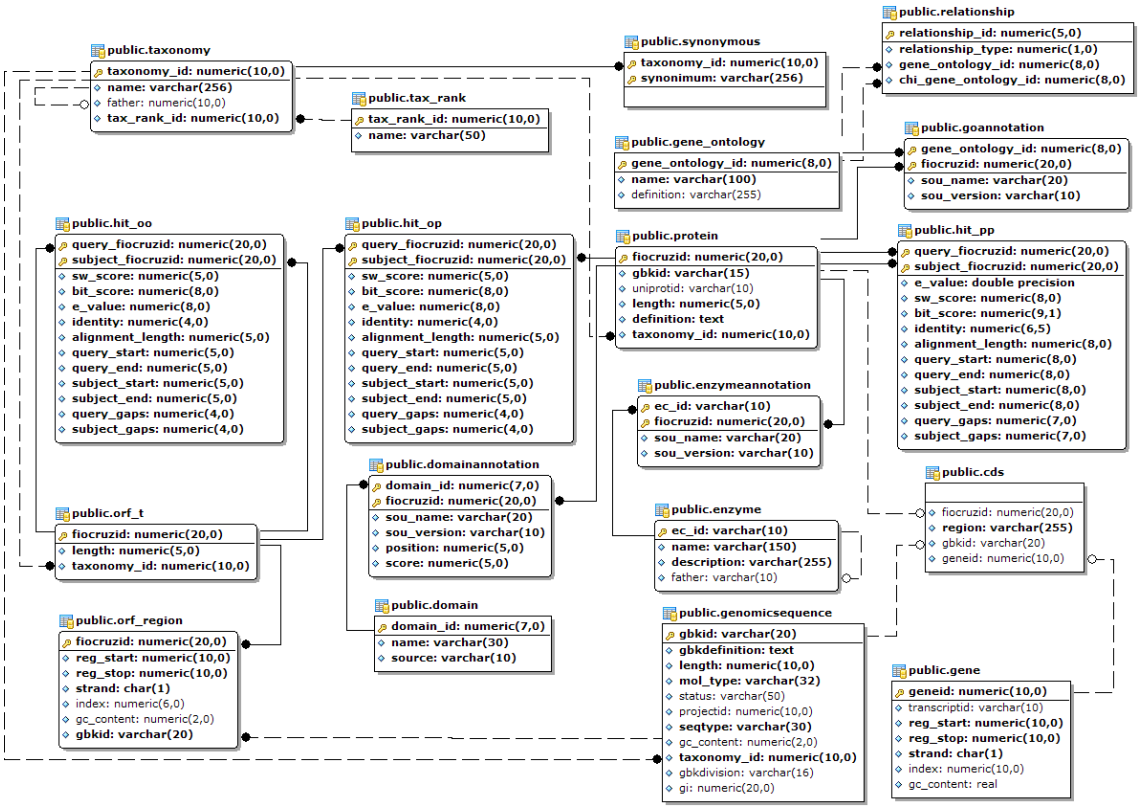

Figura 7. Protein World Database: esquema lógico

4. Implantacao do BD

Após a geração do esquema lógico, descrito no capítulo anterior, o passo seguinte foi a implantação do *Protein World DB*, e o processo de ETL (*Extraction, Transformation, and Loading*). Para a atividade de desenvolvimento do esquema lógico utilizou-se o Sistema de Gerência de Banco de Dados (SBGD) PostgreSQL [12] versão 8.4 por ser um software livre, de código aberto, e amplamente difundido no meio acadêmico e de pesquisa.

ETL é um processo fundamental na tarefa de integração de base de dados. Esse trabalho de integração de informação possibilita a construção de um repositório de referência para a comunidade de anotadores através da utilização dos índices de similaridade obtidos no Projeto Comparação de Genomas, juntamente com uma

nomenclatura padronizada de genes e seus produtos. O processo de ETL é dividido em 3 fases, conforme ilustrado pela Figura 8: extração, transformação e carga.

A primeira fase do processo de ETL envolve a extração de dados de diferentes origens, que na maioria das vezes usam diferentes formatos e organização dos dados. Neste projeto, a primeira fase envolveu, além dos arquivos resultantes do processo de comparação entre proteínas, a extração de dados de fontes externas, tais como: (a) Refseq [3], (b) Swissprot [9], (c) Taxonomy NCBI [11], que constituem o núcleo da informação, e (d) Pfam [13] - domínios de proteínas -, (e) KEGG [14] - vias metabólicas e (f) Gene Ontology [15] - vocabulário controlado -, que abrangem informações referentes às anotações das proteínas.

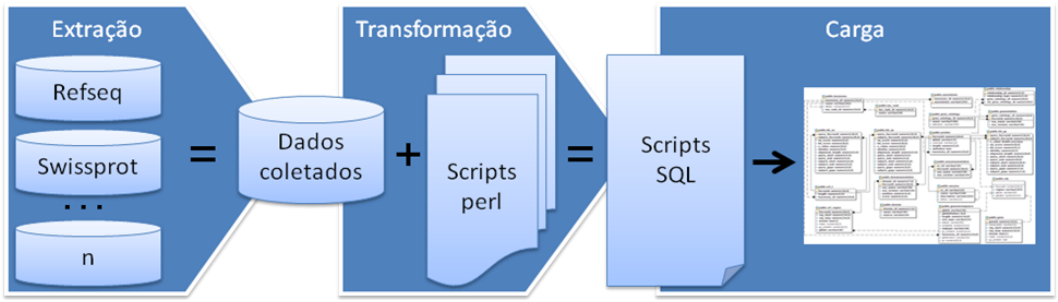

Figura 8: O processo de ETL

A fase de transformação envolve a aplicação de uma série de regras e/ou funções sobre os arquivos extraídos para que estejam de acordo com o formato e organização dos arquivos necessários para a terceira fase, a tarefa de carga. Para esta fase do processo foram utilizados scripts baseados na linguagem de programação Perl [16], os quais resultaram em scripts SQL. A tarefa de carga, que é a última fase do processo de ETL, envolveu a execução dos scripts SQL para inserção dos dados formatados.

A seguir, o processo de ETL para cada uma das fontes externas de dados é descrito.

4.1. Resultado do Programa SSEARCH

Conforme já mencionado, como etapa inicial do Projeto Comparação de Genoma, foi realizada uma tarefa de comparação de proteínas, utilizando a infra-estrutura de *grid* do *World Community Grid*. Como resultados foram gerados arquivos de similaridade (Figura 1) entre as proteínas envolvidas nesse processo (aproximadamente 900 GB de dados).

Estes arquivos de similaridade gerados devem ser armazenados integralmente nas tabelas hit_oo, hit_op e hit_pp. A única diferença é que hit_oo armazena apenas dados de comparação entre ORFs traduzidas, hit_pp entre proteínas, e hit_op entre ORFs traduzidas e proteínas. No arquivo resultante do processo de comparações, para diferenciar a origem das proteínas, o identificador da seqüência (*gi*) foi representado da seguinte forma:

- Origem RefSeq: representado pelo *Genbank Identification* (GI), como por exemplo 31543974.

- Origem SwissProt: representado por um número seqüencial, começando por 150000 ou 900000 mais 8 casas decimais, como por exemplo 900000000000002.

- Origem tORF: representado por um número seqüencial simples de 8 casas decimais, como por exemplo 00000041.

O único trabalho nesse processo de carga é a identificação das partes envolvidas na comparação para decidir em qual tabela a linha resultante deve ser inserida, conforme ilustrado pela Figura 9. A principal dificuldade foi realmente a quantidade de linhas de comparações existentes, tornando o processamento de carga uma tarefa demorada (aproximadamente 2 semanas).

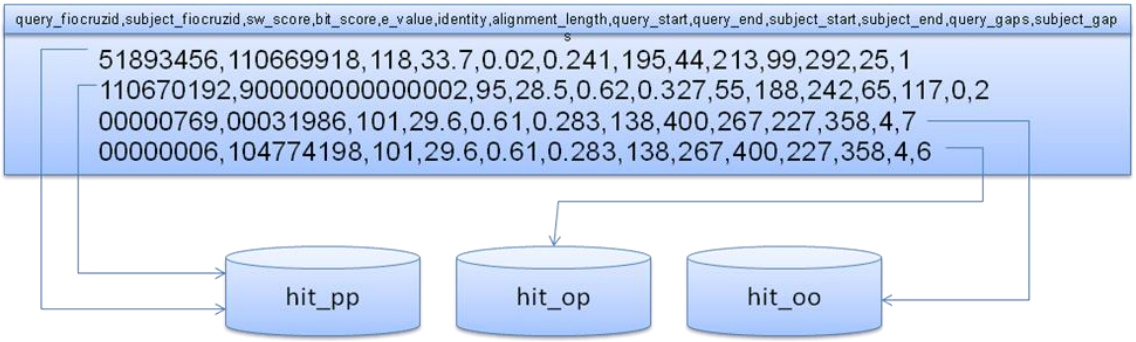

Figura 9. Processo de carga das tabelas de hits

4.2. Refseq

O NCBI *Reference Sequence* (RefSeq) [3] é uma rica coleção não-redundante de anotações de DNA, RNA e proteínas de vários táxons[1]. A coleção inclui seqüências de plasmídeos, organelas, vírus, archaea, bactérias e eucariotos. Cada seqüência RefSeq representa uma única molécula natural de um organismo. O objetivo é fornecer um conjunto abrangente e comum que representa a informação da seqüência de uma espécie. Convém notar, porém, que o RefSeq foi construído utilizando somente dados de bancos de dados públicos.

Para o acesso às informações mais atualizadas disponíveis na base de dados Refseq, optou-se pela integração da versão mais recente na época (versão 33). Isso significa que é representado um conjunto de seqüências de proteínas com dois subconjuntos: seqüências que foram comparadas no Projeto Comparação de Genomas (Refseq versão 21 e swissprot versão 51.5) e seqüências que não foram comparadas, por terem sido depositadas nessas bases de dados após o processo de comparação. A formação deste segundo subconjunto não traz qualquer tipo de problema.

Por sua vez, dentre as seqüências de aminoácidos comparadas no Projeto Comparação de Genomas, existem seqüências de proteínas de diferentes origens:

[1] Táxon (ou taxa no plural) é uma unidade taxonômica, essencialmente associada a um sistema de classificação. Táxons podem estar em qualquer nível de um sistema de classificação: um reino é um táxon, assim como um género é um táxon, assim também como uma espécie também é um táxon ou qualquer outra unidade de um sistema de classificação dos seres vivos.

- Seqüências de proteínas provenientes de projetos genômicos que possuem anotação de genes, mRNA e CDS na seqüência genômica (Refseq) – Figura 10;
- Seqüências de proteínas para as quais as únicas seqüências de nucleotídeos de origem são mRNA (Refseq) – Figura 11;
- Seqüências de proteínas obtidas diretamente do seqüenciamento da molécula de proteína, sem seqüência de nucleotídeo de origem (swissprot) – Figura 12;
- Seqüências de ORFs traduzidas (tORFs), obrigatoriamente geradas a partir de nucleotídeos, que possuem apenas as coordenadas de sua posição na seqüência genômica (apenas genomas bacterianos completos do refSeq versão 21).

```
LOCUS       NC_000085               6229 bp    DNA     linear   CON 10-JUL-2007
DEFINITION  Mus musculus chromosome 19, reference assembly (C57BL/6J).
ACCESSION   NC_000085 REGION: 45225205..45231433
VERSION     NC_000085.5  GI:149323268
DBLINK      Project:169
                                          ...
FEATURES             Location/Qualifiers
     source          1..6229
                     /organism="Mus musculus"
                     /mol_type="genomic DNA"
                     /strain="C57BL/6J"
                     /db_xref="taxon:10090"
                     /chromosome="19"
     gene            1..6229
                     /gene="Tlx1"
                     /note="Derived by automated computational analysis using
                     gene prediction method: BestRefseq. Supporting evidence
                     includes similarity to: 1 mRNA"
                     /db_xref="GeneID:21908"
                     /db_xref="MGI:98769"
     mRNA            join(1..778,2802..3003,5180..6229)
                     /gene="Tlx1"
                     /product="T-cell leukemia, homeobox 1"
                     /exception="unclassified transcription discrepancy"
                     /note="Derived by automated computational analysis using
                     gene prediction method: BestRefseq. Supporting evidence
                     includes similarity to: 1 mRNA"
                     /transcript_id="NM_021901.2"
                     /db_xref="GI:31543873"
                     /db_xref="GeneID:21908"
                     /db_xref="MGI:98769"
     CDS             join(202..778,2802..3003,5180..5402)
                     /gene="Tlx1"
                                          ...
                     /codon_start=1
                     /product="T-cell leukemia, homeobox 1"
                     /protein_id="NP_068701.1"
                     /db_xref="GI:11276073"
                     /db_xref="CCDS:CCDS29858.1"
                     /db_xref="GeneID:21908"
                     /db_xref="MGI:98769"
                                          ...
```

Figura 10. RefSeq de seqüência genômica contendo anotação de genes, mRNA e CDS

```
LOCUS       XP_001891547            211 aa            linear   INV 01-APR-2008
DEFINITION  AT hook motif family protein [Brugia malayi].
ACCESSION   XP_001891547
VERSION     XP_001891547.1  GI:170570967
DBSOURCE    REFSEQ: accession XM_001891512.1
                                          ...
FEATURES             Location/Qualifiers
     source          1..211
                     /organism="Brugia malayi"
                     /db_xref="taxon:6279"
```

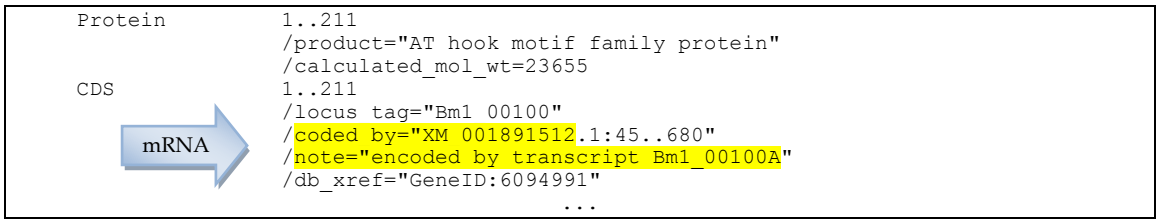

Figura 11. RefSeq de proteína originada de um mRNA

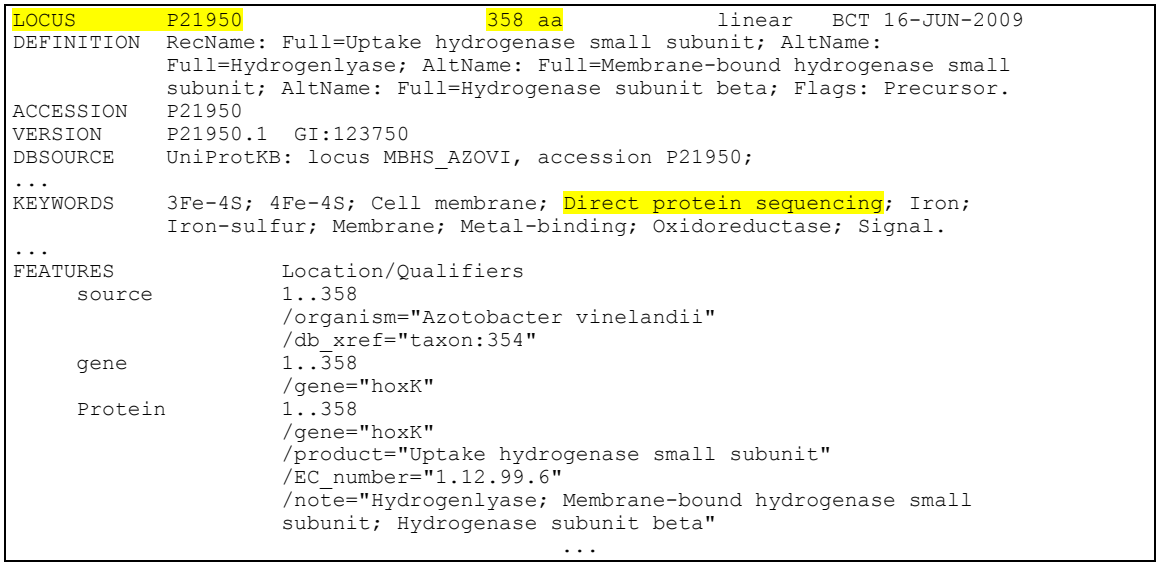

Figura 12. RefSeq de proteína obtida diretamente do seqüenciamento da molécula de proteína (swissprot)

Por causa dessas diferenças, o algoritmo responsável pela carga do núcleo, representado no esquema lógico, é obrigado a selecionar dados específicos a proteínas que possuem diferentes origens. Outra característica que deve-se levar em consideração é com relação ao tipo de arquivo RefSeq. O número de acesso RefSeq pode ser distinguido pelo formato do prefixo, 2 caracteres seguidos por pelo caractere *underscore* ('_'). Por exemplo:

- Proteínas são representadas por: AP_, NP_, XP_, YP_, ZP_
- Genomas são representados por: AC_, NC_, NG_, NT_, NW_, NZ_, NS_
- mRNAs são representados por: NM_, XM_
- RNAs são representados por: NR_, XR_

Dependendo do tipo de molécula, pode-se extrair diferentes informações. A Tabela 1 apresenta os identificadores que podem ser encontrados em um arquivo RefSeq. Alguns valores estão inseridos em strings e, neste caso, estão identificados pelos caracteres “xxx”.

Tabela 1: Mapeamento de dados - RefSeq (proteína/genoma) x atributos de tabelas

Identificador	Tipo arquivo	Atributo	Tabela
VERSION – GI:	proteína	fiocruzid	protein
ACCESSION. VERSION	proteína	gbkid	protein
--	proteína	uniprotid	protein
LOCUS – "accession" xxx aa	proteína	length	protein
DEFINITION	proteína	definition	protein
FEATURES – source – /db_xref="taxon:xxx"	proteína	taxonomy_id	protein
FEATURES – CDS – /db_xref="GI:xxx"	genoma	fiocruzid	cds
FEATURES – CDS	genoma	region	cds
ACCESSION	genoma	gbkid	cds
FEATURES – gene – /db_xref="GeneID:xxx"	genoma	geneid	cds
FEATURES – gene – /db_xref="GeneID:xxx"	genoma	geneid	gene
FEATURES – mRNA – /transcriptid="xxx"	genoma	transcriptid	gene
FEATURES – gene	genoma	reg_start	gene
FEATURES – gene	genoma	reg_stop	gene
FEATURES – gene (complement – Y ou N)	genoma	strand	gene
** Computado **	genoma	index	gene
** Computado **	genoma	gc_content	gene
ACCESSION. VERSION	genoma	gbkid	genomicsequence
DEFINITION	genoma	gbkdefinition	genomicsequence
LOCUS – "accession" xx bp	genoma	length	genomicsequence
FEATURES – source – /mol_type="xxx"	genoma	mol_type	genomicsequence
DEFINITION – "definition", xxx	genoma	status	genomicsequence
DBLINK – Project	genoma	projectid	genomicsequence
FEATURES – source – /"*tipoSeqüênciaGen.*"="xxx"	genoma	seqtype	genomicsequence
** Computado **	genoma	gc_content	genomicsequence
FEATURES – source – /db_xref="taxon:xxx"	genoma	taxonomy_id	genomicsequence
LOCUS – "accession" "length", "mol_type", "*topology*" xxx	genoma	gbkdivision	genomicsequence
VERSION – GI:	genoma	gi	genomicsequence

Uma questão muito importante explícita no esquema conceitual que deve ser considerada é o fato de que o atributo gbkid da tabela CDS referencia somente seqüências genômicas. Durante o processo de carga deve-se levar em consideração o prefixo da referência de acordo com o tipo de molécula (conforme identificado acima). Isso porque, o identificador "FEATURES – CDS – /coded_by=" sempre referencia um gbkid. No entanto, esse gbkid nem sempre refere-se a uma seqüência genômica. Ele também pode referenciar transcritos.

Proteínas que são geradas diretamente de um mRNA, no caso do seu mapeamento na seqüência genômica ser desconhecido, a informação CDS → GENE é obtida a partir do arquivo de proteínas.

4.3. Taxonomy NCBI

Para organizar os dados de seqüência de acordo com a classificação filogenética de organismos existente, o NCBI [11] mantém seu próprio banco de dados taxonômico, que contém os nomes de todos os organismos que estão representados no GenBank. Esta taxonomia é disponibilizada por meio de um "*dump*"

da base de dados de taxonomia do GenBank, e é composto pelos seguintes arquivos: nodes, names, division, gencod (*genetic codes*), delnodes (*deleted nodes*), merged (*merged nodes*), e citations. No entando apenas as tabelas nodes e names foram utilizadas no processo de ETL.

O arquivo "nodes" possui tuplas com informações referentes aos nodos da taxonomia. Cada tupla possui, dentre as inúmeras informações, os atributos tax_id (identificador do nodo na base taxonômica do GenBank), parent tax_id (identificador da base taxonômica do GenBank para o nodo pai), e rank (string informando o rank do nodo, e.g. superkingdom, kingdom, ...).

O arquivo "names" possui todos os possíveis nomes que os nodos podem apresentar. Os seus principais atributos são: tax_id (referência ao nodo que possui esse nome, name_txt (o nome propriamente dito), name class (tipo de nome, como por exemplo, synonym, equivalent name, scientific name, common name).

Inicialmente foi realizada a seleção dos ranks distintos e criação de um identificador único para cada um deles, com base nos dados presentes no atributo rank do arquivo "nodes". Como resultado, foram geradas as informações necessárias para a carga da tabela rank do nosso esquema lógico. A Tabela 2 apresenta os ranks com os respectivos identificadores.

Tabela 2: Ranks dos nodos taxonômicos

Identificador	rank
1	'no rank'
2	'superkingdom'
3	'genus'
4	'species'
5	'order'
6	'family'
7	'subspecies'
8	'subfamily'
9	'tribe'
10	'phylum'
11	'class'
12	'forma'
13	'suborder'
14	'superclass'
15	'varietas'

Identificador	rank
16	'subclass'
17	'kingdom'
18	'superfamily'
19	'infraorder'
20	'subphylum'
21	'infraclass'
22	'superorder'
23	'subgenus'
24	'parvorder'
25	'superphylum'
26	'species group'
27	'species subgroup'
28	'subtribe'
29	'subkingdom'

O passo seguinte foi percorrer o arquivo "names" selecionando, de cada tupla, apenas os dados referentes ao nome, identificador do nodo e tipo de nome. A partir dessa seleção foi gerado o arquivo de carga para a tabela synonymous que contém os possíveis nomes para cada nodo. Por fim, percorreu-se o arquivo "nodes" para gerar a carga da tabela taxonomy. Para cada tupla desse arquivo, foram selecionados: (I) o identificador do nodo (tax_id), (II) a referência para o nodo pai (parent tax_id), (III) o nome científico do nodo, que é obtido selecionando o nome de tipo "scientific name" do arquivo "names", e (IV) o identificador do rank, obtido com a comparação do atributo rank da tupla com o nome do rank da tabela rank gerada anteriormente. A Figura 13 ilustra todo esse processo.

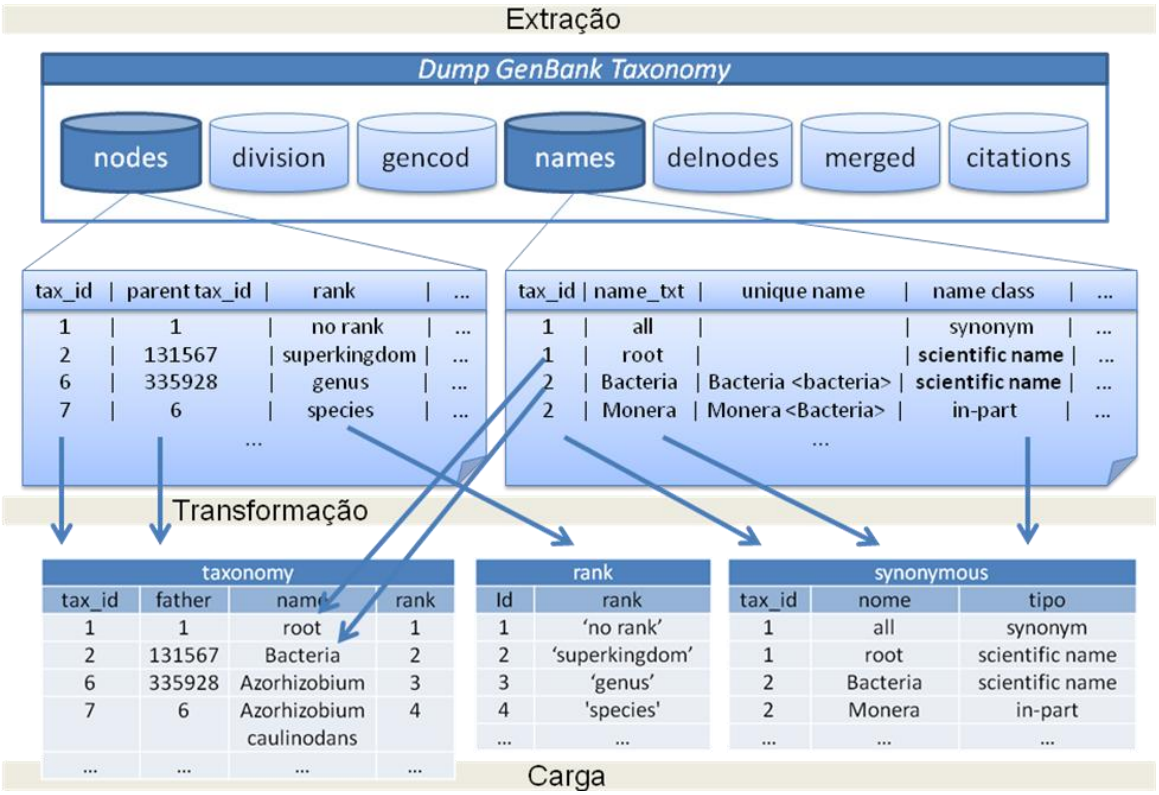

Figura 13: Processo de ETL da taxonomia

4.4. Arquivos FASTA de tORFs

Durante o mapeamento das tORFs, nas seqüências genômicas, foram gerados arquivos do tipo FASTA contendo as informações mais relevantes. A Figura 14 apresenta um trecho de um destes arquivos mostrando sua organização e as informações que podem ser obtidas:

- fiocruzid: identificador interno de uma tORF. É um número seqüencial simples de 8 casas decimais.
- gbkid: identificador do genoma que originou a respectiva tORF. O restante dos dados genômicos são obtidos pelo processo de ETL do arquivo RefSeq.
- ORF region: indica a região genômica onde foi extraída a ORF, posteriormente traduzida em tORF. Com base nessa informação obtém-se o tamanho da tORF e os atributos reg_start e reg_stop da tabela orf_region.
- gbk definition: definição do genoma da tORF.

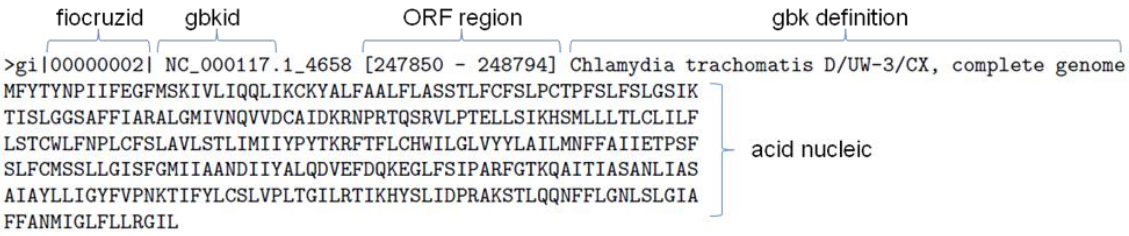

Figura 14: Exemplo de arquivo FASTA contendo informações de tORFs

4.5. Demais Fontes de Dados

Conforme já descrito, cada fonte de dados armazena diferentes dados com diferentes estruturas e formas de acesso. Desta forma, cada fonte de dados deve ser analisada e manipulada com muito cuidado, fazendo com que esse processo demande muito tempo.

Ainda restam analisar algumas fontes externas de dados para que a etapa de ETL seja concluída. A adição dessas fontes ao *Protein World DB* contribuirá com informações estruturais, de domínio e semântica às proteínas comparadas, dando-lhes maior cobertura para análises e apoio ao pesquisador para identificação e anotação de novas seqüências de proteínas. As fontes externas de dados que fornecerão este tipo de informação e que ainda estão sendo estudadas são:

- *Universal Protein Knowledgebase* (UniProtKB)/Swiss-Prot [9]: é um banco de dados de seqüências de proteínas curado que se esforça para fornecer (I) uma maior quantidade e melhor qualidade de anotação, e.g. descrição da função de uma proteína, a sua estrutura, domínios, modificações pós-translacionais, variantes, etc, (II) um nível mínimo de redundância e (III) um alto nível de integração com outras bases de dados. Da mesma forma que o RefSeq, está sendo efetuada uma análise sobre uma versão mais atual do *uniprot/sprot*. A vantagem desta base é que ela é disponibilizada no formato XML, formato este bastante difundido na área de banco de dados e suportado por muitos SGBDs comerciais.

- Pfam [13]: O banco de dados Pfam é uma grande coleção de famílias de proteínas, cada uma representada pelo alinhamento múltiplo de seqüências e modelos ocultos de Markov (HMMs). A principal contribuição dessa base são os domínios protéicos. Domínio é a designação para uma região funcional de uma proteína. Diferentes combinações de domínios dão origem a uma grande variedade de proteínas encontradas na natureza. A identificação dos domínios que ocorrem dentro das proteínas pode, portanto, fornecer informações sobre a sua função.

- *Kyoto Encyclopedia of Genes and Genomes* (KEGG) [14]: KEGG é um banco de dados de sistemas biológicos, constituído por um conjunto de dados, que podem ser de diferentes tipos: (I) estruturais, de genes e proteínas (KEGG genes), (II) químicos, de substâncias endógenas e exógenas (KEGG ligante), (III) diagramas, de ligação molecular, de interação e redes de reação (KEGG PATHWAY), e (IV) hierarquias e relações de vários objetos biológicos (KEGG

BRITE). Ele fornece uma base de conhecimento de referência para ligar genomas, tanto a sistemas biológicos quanto a ambientes, pelo processo de mapeamento PATHWAY e mapeamento BRITE. Para esta base, a principal contribuição são os mapeamentos PATHWAYs.

- *Gene Ontology* (GO) [15] – O projeto Gene Ontology é uma importante iniciativa na área de bioinformática, com o objetivo de padronizar a representação de genes e seus produtos. O projeto prevê um vocabulário controlado de termos para descrever as características dos produtos de genes e os seus dados anotados a partir dos membros do Consórcio GO, bem como ferramentas para acessar e processar estes dados.

Vale ressaltar que para as bases externas Pfam, KEGG e GO buscam-se apenas os identificadores de domínio, e.g. funcional, vias de interação molecular e ontologia, respectivamente. A relação entre as proteínas e esses domínios são oriundas das fontes RefSeq e UniProt/Swiss-Prot.

5. Conclusões e Trabalhos Futuros

A genômica comparativa é uma das áreas de pesquisa em franca expansão no momento, pois permite a obtenção de dados interessantes para a pesquisa gênica. Esses dados permitem o estudo de fatores relativos à evolução dos organismos, um 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. Devido ao seu grande potencial, o Projeto Comparação de Genomas foi o primeiro projeto da América Latina a ter sido aceito pelo *World Community Grid* para processamento. Todavia, estes dados de pouco valem se não forem devidamente armazenados e analisados.

Por lidar com técnicas de Bioinformática e Biologia Computacional, o Projeto Comparação de Genomas possui um grande potencial multidisciplinar, tornando possível a troca de experiências entre as diferentes áreas de pesquisa. Além disso, o *Protein World DB* torna-se uma importante fonte de dados para responder a diversas perguntas sobre a organização, estrutura e anotação de genes e genomas. O banco de dados integrador gerado será disponibilizado a toda comunidade científica, passando a constituir um repositório de grande valor para todos os pesquisadores envolvidos no tema.

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 disco para armazenamento com no mínimo 1TB disponível. Outra possibilidade seria dispor de estruturas de armazenamento de dados mais eficientes.

Futuramente, para a concepção de uma nova versão da base, pretendemos aplicar idéias já discutidas em [17], [18] e [19], 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. [20] e [21], 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 seqüências 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. Novamente aqui pretende-se utilizar técnicas e metodologias sugeridas em [22] e [23].

References

[1] ProteinWorld DB, http://bioinfo.pdtis.fiocruz.br/ProteinWorldDB/

[2] Smith T.F., Waterman M.S.: Comparison of Biosequences. Adv. Appl. Math. 2, 482 - 9 (1981)

[3] NCBI Reference Sequences, http://www.ncbi.nlm.nih.gov/RefSeq/

[4] World Community Grid, http://www.worldcommunitygrid.org/

[5] Pearson W.R.: SSearch. Genomics 11, 635--650 (1991)

[6] Pastor O.: Conceptual Modeling meets the Human Genome, Procs ER Conceptual Modeling, pp 1-11 (2008)

[7] Chen, J.Y. and Carlis J.V.: Genomic Data Modeling, Information Systems 28(4), pp 287-310 (2003)

[8] The NCBI Handbook: http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=handbook

[9] The UniProt Knowledgebase, http://www.uniprot.org/

[10] NCBI Entrez Gene, http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene

[11] NCBI Taxonomy Database, http://www.ncbi.nlm.nih.gov/Taxonomy/

[12] PostgreSQL, http://www.postgresql.org/

[13] The Pfam Protein Families Database, http://pfam.sanger.ac.uk/

[14] KEGG: Kyoto Encyclopedia of Genes and Genomes, http://www.genome.jp/kegg/

[15] The Gene Ontology, http://www.geneontology.org

[16] The Perl Directory at Perl.org, http://www.perl.org/

[17] Seibel, L.F.B, Macedo, J.A.F., Lemos, M., Lifschitz, S., Miranda, A.B., Alves, M., Degrave, W., A conceptual model for molecular biology information. II Workshop Brasileiro de Bioinformática (WOB), pp 47--56, 2003

[18] Macedo J.A.F de, Lifschitz S., Porto F., Picouet P., de Miranda A.B., Otto, T.D. Towards a Conceptual Modeling Language for Biological Domains, Brazilian Simposium in Bioinformatics (poster proceedings), pp 128-137, 2007.

[19] Macedo J.A.F de, Lifschitz S., Porto F., Picouet P.. Dealing with Some Conceptual Data Model Requirements for Biological Domains, International Symposium on Bioinformatics and Life Sciences Computing, pp 651-656, 2007.

[20] Lifschitz, S.,Mauro, R.C., An i/o device driver for bioinformatics tools: the case for blast. Genetics and Molecular Research (GMR), 4(1):563--570, 2005.

[21] Noronha, M.F. Implementacao e avaliação de desempenho de um driver para gerência de E/S em aplicações de bioinformática. Dissertação de Mestrado, Departamento de Informática da PUC-Rio (orientação: Sérgio Lifschitz), Setembro 2006

[22] Costa, R.L.C., Lifschitz, S. Database allocation strategies for parallel blast evaluation on clusters. Distributed and Parallel Databases 13(1):99--127, 2003.

[23] Sousa, D. Estratégias de fragmentação mista para bancos de dados de seqüências genômicas, Dissertação de Mestrado, Departamento de Informática da PUC-Rio (orientação: Sérgio Lifschitz), Julho 2007

[24] Lodish, H., Berk, A., Zipursky, L. S., Matsudaira, P., Baltimore, D., Darnell, J. Molecular Cell Biology , 6th Edition. W H Freeman & Co., 2007.

[25] TRISTÃO, C.; MIRANDA, A.B.; LIFSCHITZ, S. A conceptual data model involving protein sets from complete genomes: a biological point of view. Technical Report MCC 27/09, Departamento de Informática, PUC-Rio, 2009.

[26] U.S. Department of Energy and the National Institutes of Health. Human Genome Project Information, http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml

[27] NCBI Annotation Information - Information on NCBI genome annotation methods, http://www.ncbi.nlm.nih.gov/genome/guide/build.shtml