UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

MIDB: UM MODELO DE INTEGRAÇÃO DE DADOS BIOLÓGICOS

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação, para obtenção do título de mestre em Ciência da Computação.

Orientação: Prof. Dr. Ricardo Rodrigues Ciferri

CAROLINE BEATRIZ PERLIN

São Carlos - SP
Fevereiro/2012

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

P451mm

Perlin, Caroline Beatriz.
MIDB : um modelo de integração de dados biológicos / Caroline Beatriz Perlin. -- São Carlos : UFSCar, 2012.
103 f.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2012.

1. Banco de dados. 2. Bioinformática. 3. Modelo de integração de dados. 4. Integração de esquemas. 5. Integração de instâncias. I. Título.

CDD: 005.74 (20ª)

Universidade Federal de São Carlos

Centro de Ciências Exatas e de Tecnologia
Programa de Pós-Graduação em Ciência da Computação

"MIDB: Um Modelo de Integração de Dados Biológicos"

CAROLINE BEATRIZ PERLIN

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação

Membros da Banca:

Ricardo Rodrigues Ciferri

Prof. Dr. Ricardo Rodrigues Ciferri
(Orientador - DC/UFSCar)

Marilde Santos

Profa. Dra. Marilde Terezinha Prado Santos
(DC/UFSCar)

Prof. Dr. Wilson Araújo da Silva Júnior
(USP/RP)

São Carlos
Fevereiro/2012

Agradecimentos

Em primeiro lugar, agradeço a Deus pela força e perseverança em concluir este trabalho, e por todas as etapas vencidas.

Aos meus pais, Rosa e Luiz, e aos meus irmãos Carlos e Christian, por me incentivarem no estudo e pelo apoio incondicional.

Ao Hiro pelo amor, força e companheirismo.

Aos meus amigos e colegas Eduardo Ortiz, Karina Rocha, Marina Soubhia, Chrislac, Rafael ABF, Felipe Callori, Sandro Amaral, Fernanda e Daniel Oshiro, Rebeca e Luis Gustavo Ywata, Tati e Andy e Mateus Carrijo, Taciana Lemos, Ricardo Almeida, Douglas Fukuhara, Pablo Matos, Juliana Duque, Renata Tsuruda, Rafael Durelli, Thiago Siqueira, Vanessa Maia, Walter Coelho, Marcus Teixeira, Bruno Tomazela, Felipe Louza, Arthur Nascimento, André Silva, Anderson Hieda, e tantos outros, pela amizade nos bons e maus momentos e pelo apoio.

Ao meu orientador Ricardo Ciferri pelo incentivo, competência, dedicação e paciência e por todo o aprendizado que me proporcionou.

À professora Cristina Ciferri por ter acompanhado o meu trabalho, sempre trazendo críticas construtivas.

À professora Sandra Abib pela amizade e carinho.

A todos os que, de maneira direta ou indireta, contribuíram para a conclusão deste trabalho.

Resumo

Na bioinformática, existe um imenso volume de dados sendo produzidos, os quais estão relacionados a sequências de nucleotídeos e aminoácidos que se encontram, em quase a sua totalidade, armazenados em Bancos de Dados Biológicos (BDBs). Para uma determinada sequência existem algumas classificações de informação: dados genômicos, dados evolutivos, dados estruturais, dentre outros. Existem BDBs que armazenam somente uma ou algumas dessas classificações. Tais BDBs estão hospedados em diferentes *sites* e servidores, com sistemas gerenciadores de banco de dados distintos e com uso de diferentes modelos de dados, além de terem instâncias e esquemas com heterogeneidade semântica. Dentro desse contexto, o objetivo deste projeto de mestrado é propor um *Modelo de Integração de Dados Biológicos*, com novas técnicas de integração de esquemas e integração de instâncias. O modelo de integração proposto possui um mecanismo especial de integração de esquemas, e outro mecanismo que realiza a integração de instâncias de dados (com um dicionário acoplado) permitindo resolução de conflitos nos valores dos atributos; e um Algoritmo de Clusterização é utilizado, com o objetivo de realizar o agrupamento de entidades similares. Além disso, o especialista de domínio participa do gerenciamento desses agrupamentos. Esse modelo foi validado por meio de um estudo de caso com ênfase na integração de esquemas e integração de instâncias com dados de sequências de nucleotídeos de genes de organismos do gênero *Actinomyces*, provenientes de quatro diferentes fontes de dados. Como resultado, obteve-se que aproximadamente 97,91% dos atributos foram categorizados corretamente na integração de esquemas e a integração de instâncias conseguiu identificar que aproximadamente 50% dos clusters gerados precisam de tratamento do especialista, evitando erros de resolução de entidades. Além disso, algumas contribuições são apresentadas, como por exemplo a Categorização de Atributos, o Algoritmo de Clusterização, as funções de distância propostas e o modelo MIDB em si.

Palavras-chave: Bioinformática, Bancos de Dados Biológicos, Integração de Dados Biológicos, Modelo de Integração de Dados, Integração de Esquemas, Integração de Instâncias.

Abstract

In bioinformatics, there is a huge volume of data related to biomolecules and to nucleotide and amino acid sequences that reside (in almost their totality) in several Biological Data Bases (BDBs). For a specific sequence, there are some informational classifications: genomic data, evolution-data, structural data, and others. Some BDBs store just one or some of these classifications. Those BDBs are hosted in different sites and servers, with several data base management systems with different data models. Besides, instances and schema might have semantic heterogeneity. In such scenario, the objective of this project is to propose a biological data integration model, that adopts new schema integration and instance integration techniques. The proposed integration model has a special mechanism of schema integration and another mechanism that performs the instance integration (with support of a dictionary) allowing conflict resolution in the attribute values; and a Clustering Algorithm is used in order to cluster similar entities. Besides, a domain specialist participates managing those clusters. The proposed model was validated through a study case focusing on schema and instance integration about nucleotide sequence data from organisms of *Actinomyces* gender, captured from four different data sources. The result is that about 97.91% of the attributes were correctly categorized in the schema integration, and the instance integration was able to identify that about 50% of the clusters created need support from a specialist, avoiding errors on the instance resolution. Besides, some contributions are presented, as the Attributes Categorization, the Clustering Algorithm, the distance functions proposed and the proposed model itself.

Keywords: Bioinformatics, Biological Databases, Biological Database Integration, Data Integration Model, Schema Integration, Instance Integration.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ALGORITMOS

LISTA DE LISTAGENS

LISTA DE ABREVIATURAS E SIGLAS

3D – Tridimensional

API – *Application Programming Interface*

BD – Banco de Dados

BDB – Banco de Dados Biológicos

BLAST – *Basic Local Alignment Search Tool*

DDBJ – *DNA Data Bank of Japan*

DNA – *Deoxyribo Nucleic Acid* (ácido desoxirribonucleico)

EMBL – *European Molecular Biology Laboratory*

ER – Entidade-Relacionamento (do modelo conceitual ER)

GAV – *Global As a View*

GBD – *Genome Data Base*

LAV – *Local As a View*

PDB – *Protein Data Bank*

PBDe – *Protein Data Bank in Europe*

RE – Resolução de Entidades

RNA – *Ribo Nucleic Acid* (ácido ribonucleico)

SCOP – *Structure Classification of Proteins*

SGBD – Sistema Gerenciador de Banco de Dados

SQL – *Structured Query Language*

VEGA – *VErtebrate Genome Annotation*

XML – *eXtensible Markup Language*

W3C – *World Wide Web Consortium*

SUMÁRIO

Capítulo 1

INTRODUÇÃO

Este capítulo apresenta o contexto, a motivação e os desafios que deram origem ao desenvolvimento deste trabalho de mestrado. Os principais objetivos são discutidos e algumas das contribuições almejadas são apresentadas, finalizando com a descrição da organização da dissertação.

1.1 Considerações Iniciais

Devido à sua aplicabilidade em projetos genoma, proteoma e meta-genoma, a bioinformática tem se destacado atualmente como uma área de pesquisa bastante relevante (GUSNANTO et al., 2012; AMES et al., 2012). No I *Workshop* Brasileiro de Bioinformática (2002), a bioinformática foi definida como uma nova área do conhecimento científico baseada na união entre a biologia e a ciência da computação, a qual emprega ferramentas matemáticas, computacionais, estatísticas e de outras áreas relacionadas na solução de problemas da biologia.

No contexto desta pesquisa em nível de mestrado, a bioinformática é definida como um campo da ciência que engloba métodos da biologia, da matemática, da estatística e principalmente da ciência da computação para resolver questões biológicas relevantes a partir de sequências de nucleotídeos, de aminoácidos e de informações relacionadas (SILVA JUNIOR, 2006).

Um dos fatos que impulsionou a expansão da bioinformática foi o projeto genoma. Esse projeto produziu grande quantidade de dados que precisavam ser armazenados, anotados e consultados. Uma forma de armazenamento para esses

dados do domínio da bioinformática é por meio do uso de Bancos de Dados Biológicos (BDBs).

Um BDB pode armazenar dados sobre sequências de nucleotídeos, sequências de aminoácidos, promotores, genes e anotações relacionadas, proteínas e suas estruturas tridimensionais, bem como outros produtos biológicos. Os BDBs, em sua maioria, são de domínio público, possibilitando que os cientistas tenham acesso à informação derivada de outros laboratórios e, além disso, possam trocar e compartilhar informações e sequências genéticas (NAKANO, 2005; YANAGA, 2006). Nos BDBs, as sequências são constituídas por *strings* que representam nucleotídeos compostos pelas bases nitrogenadas adenina (A), guanina (G), citosina (C), timina (T) e uracila (U). As três primeiras bases são comuns ao DNA e RNA. A timina é uma base exclusiva do DNA, a qual é substituída pela uracila quando ocorre a transcrição do DNA para o RNA, sendo a uracila uma base exclusiva do RNA. Nos BDBs também são usadas as *strings* que representam os 20 aminoácidos que compõem as proteínas (i.e., {A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}).

Em geral, os BDBs são construídos independentemente e possuem dados redundantes. Por exemplo, diferentes BDBs podem armazenar sequências de nucleotídeos diferentes para um mesmo organismo, indicando inconsistência nos dados ou indicando diferentes versões de um sequenciamento genômico. Por outro lado, BDBs podem ser complementares. Por exemplo, uma sequência de um certo organismo pode ser exclusiva de um BDB, enquanto outra sequência de nucleotídeos de um outro organismo pode ser encontrada apenas em um outro BDB. Em particular, a carga dos dados é efetuada por programas específicos em cada BDB, sem nenhuma preocupação quanto a consistência dos dados armazenados nos diversos BDBs. Em raros casos, tais como ocorre com os BDBs *Genbank* (NCBI) (GENBANK, 2010) e DNA *Data Bank of Japan* (DDBJ) (DNA Data Bank of Japan, 2010), há uma replicação controlada das sequências biológicas armazenadas no final de cada período (i.e., um horário pré-estabelecido para a troca e a sincronização de sequências). No entanto, biólogos e bioinformatas frequentemente necessitam analisar conjuntamente dados de vários BDBs e com isso se deparam com o problema de integração dos dados.

1.2 Motivação e Desafios

Este projeto de pesquisa em nível de mestrado desenvolve um novo modelo de integração de dados. O intuito deste modelo é permitir a integração de diferentes fontes de dados, ou seja, os BDBs, resolvendo os seguintes principais problemas provenientes da integração: integração de esquemas, identificação de agrupamentos de entidades similares (resolução de entidades) e resolução de conflitos de valores nos atributos de entidades.

O modelo proposto por este projeto de pesquisa poderá ser utilizado para integrar dados biológicos de sequências de nucleotídeos e por conseguinte auxiliar biólogos e demais pesquisadores da área de saúde (e.g. biólogos, médicos, pesquisadores) a mapear tais sequências e suas características, mesmo que estejam armazenadas em BDBs heterogêneos. O problema de integração de BDBs torna-se cada vez mais importante com o aumento do volume de dados sequenciados. Atualmente, o volume de dados biológicos cresce de forma exponencial e a quantidade de BDBs também apresenta um crescimento acelerado (NCBI-GENBANK, 2011). Portanto, é essencial prover mecanismos eficientes para a integração dos dados dos diversos BDBs.

O modelo de integração proposto possui um mecanismo especial de integração de esquemas, e outro mecanismo que realiza a integração de instâncias de dados (com um dicionário acoplado) permitindo resolução de conflitos nos valores dos atributos; e um Algoritmo de Clusterização é utilizado, com o objetivo de realizar o agrupamento de entidades similares. Além disso, o especialista de domínio participa do gerenciamento desses agrupamentos. Esses, portanto, são os principais diferenciais da proposta do modelo de integração.

1.3 Objetivo

Este trabalho de pesquisa tem como objetivo propor um modelo de integração de dados biológicos. Esse modelo será apoiado no uso de novas técnicas de integração de esquemas e integração de instâncias, além de contar com o auxílio do

especialista de domínio para tomada de decisão em determinados casos. Esta pesquisa enfoca o problema de integração dos dados segundo as seguintes perspectivas: (1) integração de esquemas com a identificação de relacionamentos entre os atributos de diferentes esquemas, (2) identificação de agrupamentos de entidades similares, e (3). integração de instâncias em BDBs com resolução de conflitos de valores (i.e., valores armazenados nas tuplas ou registros). O modelo proposto será validado por meio de estudo de caso visando-se analisar a eficiência do processo de integração.

As contribuições advindas deste trabalho de mestrado são: uma nova técnica de integração de esquemas (denominada Categorização de Atributos), uma nova forma de integrar instâncias (Algoritmo de Clusterização), as distâncias de Edição de Caracteres e de Edição de Palavras criadas para comparação de *strings* em uma clusterização, e o modelo de integração de dados biológicos em si.

O modelo proposto é aplicável na biologia sistêmica (por exemplo, em bancos de dados de tumores). Para um determinado gene, em conjunto com informações clínicas é possível identificar os genes expressos e as mutações utilizando o modelo proposto para realizar comparações entre genes.

1.4 Organização da Dissertação

O conteúdo desta dissertação está estruturado da seguinte maneira:

- Os Capítulos 2 e 3 apresentam a fundamentação teórica que será usada no desenvolvimento deste trabalho, sendo que o Capítulo 2 descreve conceitos sobre Bancos de Dados Biológicos e o Capítulo 3 descreve conceitos relacionados à Integração de Bancos de Dados, com uma seção que foca na Integração de Dados Biológicos.
- O Capítulo 4 descreve os trabalhos correlatos a este projeto.
- O Capítulo 5 detalha o modelo MIDB de integração proposto por este trabalho de mestrado.
- O Capítulo 6 realiza estudos de caso como forma de avaliação do modelo MIDB.

- O Capítulo 7 apresenta a conclusão, destacando as contribuições alcançadas e avaliando o quão adaptável é o modelo proposto, além de apresentar trabalhos futuros.
- Por fim, são apresentadas as referências consultadas.

Capítulo 2

BANCOS DE DADOS BIOLÓGICOS

Este capítulo descreve os conceitos e as características básicas de Bancos de Dados Biológicos, os tipos de dados biológicos armazenados nesses BDBs e cita os principais BDBs de acesso público para cada um dos tipos de dados. Além disso, é apresentado o resultado para uma pesquisa em três dos principais BDBs, demonstrando o diferente formato de dados armazenado em cada BDB para sequências de genes.

2.1 Considerações Iniciais

A bioinformática é uma área que adquiriu crescimento considerável nos últimos anos, principalmente devido à sua aplicabilidade nos projetos genoma e proteoma. Assim, as pesquisas da área de bioinformática produzem grande volume de dados biológicos que precisam ser armazenadas em bancos de dados próprios, denominados mais especificamente de Bancos de Dados Biológicos (BDBs).

2.2 Tipos de dados biológicos

Dentre os dados biológicos que os BDBs podem armazenar incluem-se sequências de nucleotídeos e sequências de aminoácidos, ácidos nucleicos, genomas, proteomas, meta-genomas, anotações de genes, estrutura tridimensional de proteínas, dados bibliográficos, enfim, dados referentes aos seres vivos.

Em seguida, exemplificamos alguns desses tipos de dado biológico que podem ser armazenado em um BDB:

- Sequência de nucleotídeos e aminoácidos: As sequências são constituídas por *strings* que representam nucleotídeos compostos pelas bases nitrogenadas citosina (C), timina (T), guanina (G), adenina (A) e uracila (U), ou por *strings* que representam os 20 aminoácidos que compõem as proteínas (i.e., {A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}). As bases adenina, guanina e citosina estão presentes tanto no DNA quanto no RNA. Contudo, a base timina é exclusiva do DNA, e a uracila é exclusiva do RNA. *Keratin type II cytoskeletal 2 epidermal* é uma proteína cuja sequência de aminoácidos pode ser observada na Figura 1.

```
  10         20         30         40         50         60
MSCQISCKSR GRGGGGGGFR GFSSGSAVVS GGSRRSTSSF SCLSRHGGGG GGFGGGGFGS

        70         80         90        100        110        120
RSLVGLGGTK SISISVAGGG GGFGAAGGFG GRGGGFGGGS SFGGGSGFSG GGFGGGGFGG

       130        140        150        160        170        180
GRFGGFGGPG GVGGLGGPGG FGPGGYPGGI HEVSVNQSLL QPLNVKVDPE IQNVKAQERE

       190        200        210        220        230        240
QIKTLNNKFA SFIDKVRFLE QQNQVLQTKW ELLQQMNVGT RPINLEPIFQ GYIDSLKRYL

       250        260        270        280        290        300
DGLTAERTSQ NSELNNMQDL VEDYKKKYED EINKRTAAEN DFVTLKKDVD NAYMIKVELQ

       310        320        330        340        350        360
SKVDLLNQEI EFLKVLYDAE ISQIHQSVTD TNVILSMDNS RNLDLDSIIA EVKAQYEEIA

       370        380        390        400        410        420
QRSKEEAEAL YHSKYEELQV TVGRHGDSLK EIKIEISELN RVIQRLQGEI AHVKKQCKNV

       430        440        450        460        470        480
QDAIADAEQR GEHALKDARN KLNDLEEALQ QAKEDLARLL RDYQELMNVK LALDVEIATY

       490        500        510        520        530        540
RKLLEGEECR MSGDLSSNVT VSVTSSTISS NVASKAAFGG SGGRGSSSGG GYSSGSSSYG

       550        560        570        580        590        600
SGGRQSGSRG GSGGGGSISG GGYGSGGGSG GRYGSGGGSK GGSISGGGYG SGGGKHSSGG

       610        620        630
 GSRGGSSSGG GYGSGGGGSS SVKGSSGEAF GSSVTFSFR
```

Figura 1 Sequência de aminoácidos que forma a proteína *Keratin type II cytoskeletal 2 epidermal*, adaptada de Uniprot (2010).

- Ácido nucleico: São compostos químicos formados por ácido fosfórico, uma pentose, e base nitrogenada, conforme exibido na Figura 2. São macromoléculas que contêm a informação genética da célula. Existem dois ácidos nucleidos:

- Ácido desoxirribonucleico (DNA): A pentose que o forma é a desoxirribose e pode ter as seguintes bases nitrogenadas: A, T, G, C.
- Ácido ribonucleico (RNA): A pentose que o forma é a ribose e pode ter as seguintes bases nitrogenadas: A, U, G, C.

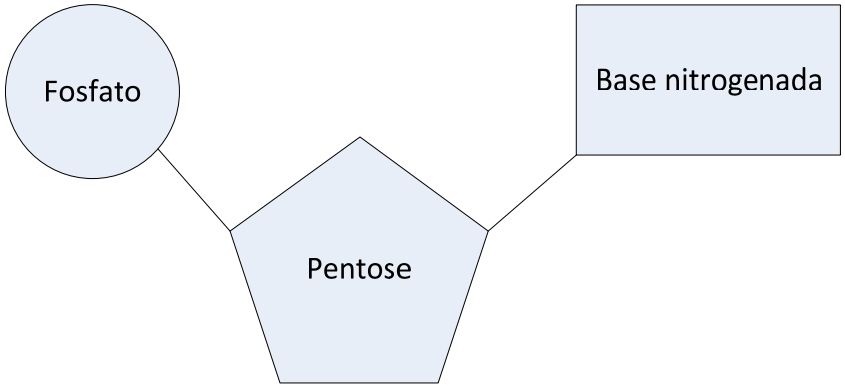

Figura 2 Esquema do nucleotídeo, com seus componentes fosfato, pentose e base nitrogenada.

- Genoma: Conjunto de genes de uma espécie, onde um gene é uma sequência de nucleotídeos que contém informação genética. O Projeto Genoma (LANDER et al., 2001), cujo objetivo foi o sequenciamento e a análise do genoma humano, é um dos projetos de sequenciamento de genoma mais conhecidos. A Figura 3 mostra a relação entre cromossomo, gene e DNA. Sumarizando, os organismos vivos de uma determinada espécie possuem *n* cromossomos, os quais são formados por moléculas de DNA. O DNA é formado por uma sequência de nucleotídeos. Em algumas partes dessas sequências podem-se identificar informações genéticas, caracterizando os genes. Assim, genoma é o conjunto das informações genéticas presentes em todos os cromossomos.

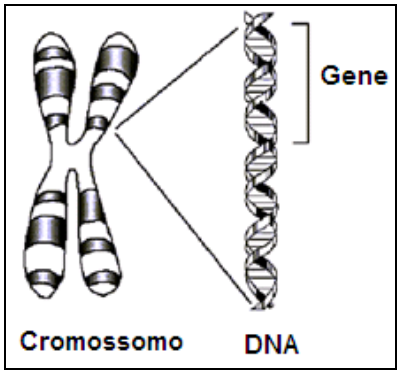

Figura 3 Esquema mostrando a relação entre cromossomo, gene e DNA, adaptada de Predictive (2003).

- Proteoma: Conjunto de proteínas que participam de processos químicos e biológicos numa célula, tecido ou organismo de uma espécie. Após a conclusão do Projeto Genoma, um dos grandes desafios é o Projeto Proteoma, que tem como objetivo o mapeamento do proteoma da espécie humana. Exemplo de proteoma: proteoma da bactéria *Escherichia coli*, o qual pode ser encontrado no Banco de Dados de Genoma e Proteoma da *E.coli* (GenProtEC, 2010).
- Anotação de gene: Os genes são anotados com o intuito de incluir informações importantes para a sua análise e interpretação. A anotação genômica pode ser em nível de nucleotídeo, em nível de proteína ou em nível de processo (TEIXEIRA, 2008). De acordo com o ambiente onde a anotação é realizada, ela pode ser classificada em manual (realizada por um anotador humano), automática (conduzida por ferramentas computacionais) ou importada (a anotação foi extraída de uma outra fonte de dados, geralmente um BDB público). A Figura 4 mostra a anotação do gene BRD2 da espécie *Canis familiaris* (cachorro).

Vega Gene: OTTCANG00000000028 (BRD2) [Region in detail]
Vega gene OTTCANG00000000028 (annotated by Havana) has 6 transcripts: OTTCANT00000000040, OTTCANT00000000043, OTTCANT00000000039, OTTCANT00000000041, OTTCANT00000000038, OTTCANT00000000042
Description: bromodomain containing 2

Source: v37; **Species**: Canis familiaris; Gene; **Feature type**: Gene; Canis familiaris;

Figura 4 Exemplo de anotação de gene da espécie *Canis familiaris* (cachorro), adaptada de Vega (2010).

- Estrutura de proteínas: alguns BDBs armazenam estruturas tridimensionais (3D) que exibem a representação geométrica das proteínas no espaço. Na Figura 5 pode-se visualizar a estrutura da proteína *serum albumina* dos seres humanos, a qual também provê informações sobre ligações químicas e detalhes experimentais, como exibido na Figura 6.

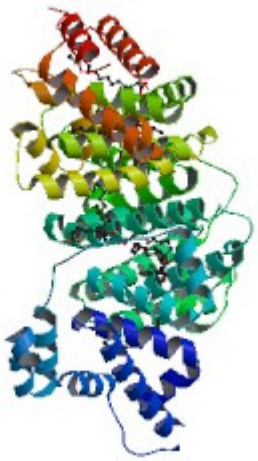

Figura 5 Estrutura da proteína *serum albumina* humana, adaptada de wwPDB (2010).

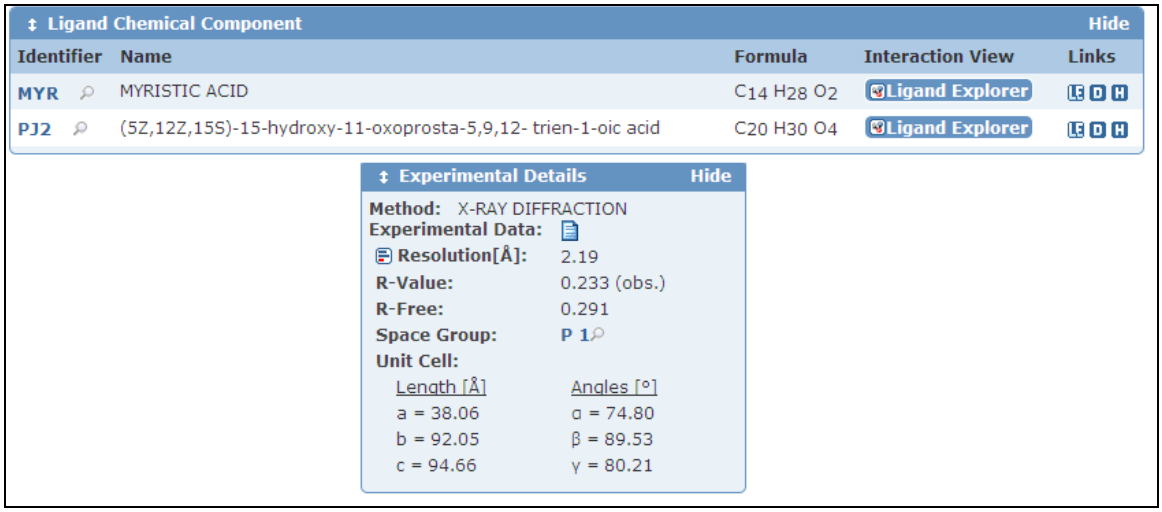

Figura 6 Informações adicionais sobre a proteína *serum albumina* humana: ligações químicas e detalhes experimentais, adaptada de wwPDB (2010).

- Dados bibliográficos: São quaisquer publicações, citações, resumos e artigos completos disponíveis para consulta. O PubMed (NATIONAL, 1997) é um dos BDBs mais conhecidos que armazena dados bibliográficos. Utilizando a expressão "*soya cancer*", é possível recuperar referências que relacionam a soja com algum tipo de câncer, como pode ser observado em um trecho da consulta, exibido na Figura 7.

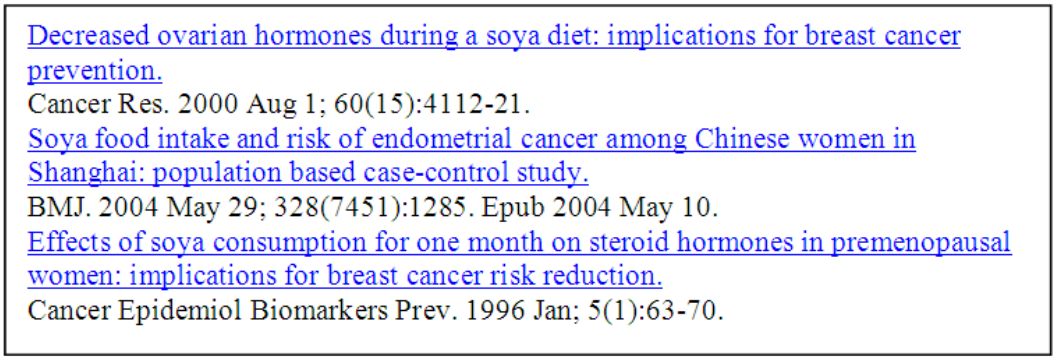

Figura 7 Trecho do resultado da busca por "*soya cancer*" no PubMed.

A maioria dos BDBs especializam-se em armazenar um ou mais dos dados citados anteriormente. Em seguida, são citados BDBs de acesso público para cada um desses tipos de dados.

BDBs de sequência de nucleotídeos

- Genbank, disponível em http://www.ncbi.nlm.nih.gov/Genbank/.
- DNA Data Bank of Japan (DDBJ), disponível em http://www.ddbj.nig.ac.jp/Welcome-e.html
- EMBL Nucleotide Sequence Database, disponível em http://www.ebi.ac.uk/embl/index.html

BDBs de sequência de aminoácidos (proteínas)

- UniProt, disponível em http://www.uniprot.org/.
- PROSITE, disponível em http://www.expasy.org/prosite/.
- PRINTS, disponível em http://www.bioinf.manchester.ac.uk/dbbrowser/PRINTS/index.php.
- Pedant 3 database, disponível em http://pedant.gsf.de .

BDBs de genomas

- CAMERA Marine Microbial Ecology, disponível em http://camera.calit2.net/index.php/ .
- Maize GBD, disponível em http://www.maizegdb.org/ .
- Ensembl, disponível em http://www.ensembl.org/index.html
- Vertebrate Genome Annotation (VEGA) database, disponível em http://vega.sanger.ac.uk/index.html

BDBs de proteomas

- GenProtEC, *E.coli* Genome and proteome database, disponível em http://genprotec.mbl.edu/.

BDBs de anotação de genes

- VEGA - Vertebrate Genome Annotation database, disponível em http://vega.sanger.ac.uk/index.html.

BDBs sobre a estrutura tridimensional de proteínas

- Protein Data Bank (PDB), disponível em http://www.wwpdb.org/ .

- CATH Protein Structure Classification, disponível em http://www.cathdb.info/ .
- SCOP Structure Classification of Proteins Database, disponível em http://scop.mrc-lmb.cam.ac.uk/scop/.
- Protein Data Bank in Europe (PBDe), disponível em http://www.ebi.ac.uk/pdbe/

BDBs bibliográficos

- Pubmed, disponível em http://www.ncbi.nlm.nih.gov/pubmed/ .

Mesmo sendo de uma mesma categoria, o formato dos dados armazenados por cada BDB difere entre si. Por exemplo, na categoria de BDBs de sequência de nucleotídeos, tem-se o Genbank, DDBJ, EMBL e o Oralgen. Fazendo uma pesquisa sobre sequência de nucleotídeos do organismo *Actinomyces naeslundii* nos três primeiros BDBs e sobre *Actinobacillus actinomycetemcomitans* no último BDB, a estrutura dos mesmos é bem diferenciada, como pode ser observado na Listagem 1, Listagem 2, Listagem 3 e Listagem 4.

Listagem 1 Pesquisa *Actinomyces naeslundii* no Genbank.

```
LOCUS       AF048778                315 bp    DNA     linear   BCT 09-FEB-1999
DEFINITION  Actinomyces naeslundii urease gamma-subunit (ureA) gene, complete
            cds.
ACCESSION   AF048778
VERSION     AF048778.1  GI:4249596
KEYWORDS    .
SOURCE      Actinomyces naeslundii
  ORGANISM  Actinomyces naeslundii
            Bacteria; Actinobacteria; Actinobacteridae; Actinomycetales;
            Actinomycineae; Actinomycetaceae; Actinomyces.
REFERENCE   1  (bases 1 to 315)
  AUTHORS   Morou-Bermudez,E. and Burne,R.A.
  TITLE     Genetic and physiologic characterization of urease of Actinomyces
            naeslundii
  JOURNAL   Infect. Immun. 67 (2), 504-512 (1999)
   PUBMED   9916052
REFERENCE   2  (bases 1 to 315)
  AUTHORS   Morou-Bermudez,E. and Burne,R.A.
  TITLE     Direct Submission
  JOURNAL   Submitted (17-FEB-1998) Center for Oral Biology, Univ. of
            Rochester, 601 Elmwood Ave, Rochester, NY 14642, USA
FEATURES             Location/Qualifiers
     source          1..315
                     /organism="Actinomyces naeslundii"
                     /mol_type="genomic DNA"
                     /strain="WVU45"
                     /db_xref="taxon:1655"
     gene            1..315
                     /gene="ureA"
     CDS             1..315
```

```
                     /gene="ureA"
                     /codon_start=1
                     /transl_table=11
                     /product="urease gamma-subunit"
                     /protein_id="AAD13723.1"
                     /db_xref="GI:4249597"
                     /translation="MHLTPREQEKLLIVVAADLARRRKDRGIRLNHPEAVAYITAEIL
                     EGAREGRTVTDLMAYGTTLLTYDDVMEGVPEMIRAVQVEATFPDGTKLVSVHDPIRRR
                     LP"
ORIGIN
        1 atgcacctca ccccgcgtga gcaggagaag ctactcatcg tcgtcgccgc agacctggca
       61 cgaagacgca aggacagagg gatccggctc aaccaccccg aggcagtcgc ctacatcacc
      121 gctgagatcc tcgagggagc ccgagagggg cgcacggtca cggatctcat ggcctacggg
      181 accaccctgc tgacctacga cgacgtcatg gagggagtcc ccgagatgat ccgcgcggtc
      241 caggtggagg cgaccttccc cgacggaacc aagctcgtct ccgtccacga cccgattcgc
      301 aggaggctgc catga
//
```

Listagem 2 Pesquisa *Actinomyces naeslundii* no DDBJ.

```
-----------------------------------
Number = [ EU621004 ]
-----------------------------------

LOCUS       EU621004                552 bp    DNA     linear   BCT 31-MAR-2009
DEFINITION  Actinomyces naeslundii strain CCUG 34725 citrate synthase I (gltA)
            gene, partial cds.
ACCESSION   EU621004
VERSION     EU621004.1
KEYWORDS    .
SOURCE      Actinomyces naeslundii
  ORGANISM  Actinomyces naeslundii
            Bacteria; Actinobacteria; Actinobacteridae; Actinomycetales;
            Actinomycineae; Actinomycetaceae; Actinomyces.
REFERENCE   1  (bases 1 to 552)
  AUTHORS   Henssge,U., Do,T., Radford,D.R., Gilbert,S.C., Clark,D. and
            Beighton,D.
  TITLE     Emended description of Actinomyces naeslundii and descriptions of
            Actinomyces oris sp. nov. and Actinomyces johnsonii sp. nov.,
            previously identified as Actinomyces naeslundii genospecies 1, 2
            and WVA 963
  JOURNAL   Int. J. Syst. Evol. Microbiol. 59 (PT 3), 509-516 (2009)
   PUBMED   19244431
REFERENCE   2  (bases 1 to 552)
  AUTHORS   Henssge,U., Do,T., Radford,D., Gilbert,S., Clark,D. and Beighton,D.
  TITLE     Direct Submission
  JOURNAL   Submitted (02-APR-2008) Infection Research Group, Dental Institute
            King's College London, Floor 17, Microbiology, Tower Wing, London
            SE1 9RT, United Kingdom
FEATURES             Location/Qualifiers
     source          1..552
                     /organism="Actinomyces naeslundii"
                     /mol_type="genomic DNA"
                     /strain="CCUG 34725"
                     /db_xref="taxon:1655"
     gene            <1..>552
                     /gene="gltA"
     CDS             <1..>552
                     /gene="gltA"
                     /codon_start=1
                     /transl_table=11
                     /product="citrate synthase I"
                     /protein_id="ACF21239.1"
                     /db_xref="GI:193794054"
                     /translation="GLPLLYPDPQRSYVEDFIRLTFGMPYQSYEIDPAVVRALDMLLI
                     LHADHEQNCSTSTVRLVGSADANMYASVAAGVGALSGPLHGGANEAVLRMLDTIQSSG
                     MSTAEFVRKVKDKEDGVRLMGFGHRVYKNYDPRAAIVKETAHDVLTRLGSDDGDRKLE
```

```
                     IAMELEETALRDEYFVSRSLYPNV"
BASE COUNT           91 a          203 c          176 g           82 t
ORIGIN
        1 ggcctgccgc tgctctaccc cgacccgcag cgctcctacg tcgaggactt catccgcctg
       61 accttcggca tgccctacca gtcctacgag atcgacccgg ccgtggtgcg ggccctggac
      121 atgctcctca tcctgcacgc cgaccacgag cagaactgct cgacctccac cgtgcgcctc
      181 gtgggctcgg ccgacgccaa catgtacgcc tccgtggccg cgggcgtggg tgccctgtcc
      241 ggcccgctgc acggcggcgc gaacgaggcc gtcctgcgga tgctggacac gatccagagc
      301 tcggggatga gcacggccga gttcgtccgc aaggtcaagg acaaggagga cggcgtccgg
      361 ctcatgggct tcggccaccg ggtctacaag aactatgacc cgcgcgccgc gatcgtcaag
      421 gagaccgccc acgacgtcct gacccgtctg ggctccgacg acggcgaccg caagctcgag
      481 atcgccatgg agctggagga gacggcgctg cgcgacgagt acttcgtctc gcgcagcctg
      541 tacccgaacg tc
//
```

Listagem 3 Pesquisa *Actinomyces naeslundii* no EMBL.

```
ID   EU621004; SV 1; linear; genomic DNA; STD; PRO; 552 BP.
XX
AC   EU621004;
XX
DT   31-MAR-2009 (Rel. 100, Created)
DT   31-MAR-2009 (Rel. 100, Last updated, Version 1)
XX
DE   Actinomyces naeslundii strain CCUG 34725 citrate synthase I (gltA) gene,
DE   partial cds.
XX
KW   .
XX
OS   Actinomyces naeslundii
OC   Bacteria; Actinobacteria; Actinobacteridae; Actinomycetales;
OC   Actinomycineae; Actinomycetaceae; Actinomyces.
XX
RN   [1]
RP   1-552
RX   DOI; 10.1099/ijs.0.000950-0.
RX   PUBMED; 19244431.
RA   Henssge U., Do T., Radford D.R., Gilbert S.C., Clark D., Beighton D.;
RT   "Emended description of Actinomyces naeslundii and descriptions of
RT   Actinomyces oris sp. nov. and Actinomyces johnsonii sp. nov., previously
RT   identified as Actinomyces naeslundii genospecies 1, 2 and WVA 963";
RL   Int. J. Syst. Evol. Microbiol. 59(Pt 3):509-516(2009).
XX
RN   [2]
RP   1-552
RA   Henssge U., Do T., Radford D., Gilbert S., Clark D., Beighton D.;
RT   ;
RL   Submitted (02-APR-2008) to the INSDC.
RL   Infection Research Group, Dental Institute King's College London, Floor 17,
RL   Microbiology, Tower Wing, London SE1 9RT, United Kingdom
XX
DR   StrainInfo; 127122; 0.
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..552
FT                   /organism="Actinomyces naeslundii"
FT                   /strain="CCUG 34725"
FT                   /mol_type="genomic DNA"
FT                   /db_xref="taxon:1655"
FT   gene            <1..>552
FT                   /gene="gltA"
FT   CDS             <1..>552
FT                   /codon_start=1
FT                   /transl_table=11
FT                   /gene="gltA"
FT                   /product="citrate synthase I"
```

```
FT                   /db_xref="GOA:C1IM42"
FT                   /db_xref="InterPro:IPR002020"
FT                   /db_xref="InterPro:IPR016141"
FT                   /db_xref="InterPro:IPR016142"
FT                   /db_xref="InterPro:IPR016143"
FT                   /db_xref="InterPro:IPR019810"
FT                   /db_xref="UniProtKB/TrEMBL:C1IM42"
FT                   /protein_id="ACF21239.1"
FT                   /translation="GLPLLYPDPQRSYVEDFIRLTFGMPYQSYEIDPAVVRALDMLLIL
FT                   HADHEQNCSTSTVRLVGSADANMYASVAAGVGALSGPLHGGANEAVLRMLDTIQSSGMS
FT                   TAEFVRKVKDKEDGVRLMGFGHRVYKNYDPRAAIVKETAHDVLTRLGSDDGDRKLEIAM
FT                   ELEETALRDEYFVSRSLYPNV"
XX
SQ   Sequence 552 BP; 91 A; 203 C; 176 G; 82 T; 0 other;
     ggcctgccgc tgctctaccc cgacccgcag cgctcctacg tcgaggactt catccgcctg        60
     accttcggca tgccctacca gtcctacgag atcgacccgg ccgtggtgcg ggccctggac       120
     atgctcctca tcctgcacgc cgaccacgag cagaactgct cgacctccac cgtgcgcctc       180
     gtgggctcgg ccgacgccaa catgtacgcc tccgtggccg cgggcgtggg tgccctgtcc       240
     ggcccgctgc acggcggcgc gaacgaggcc gtcctgcgga tgctggacac gatccagagc       300
     tcggggatga gcacggccga gttcgtccgc aaggtcaagg acaaggagga cggcgtccgg       360
     ctcatgggct tcggccaccg ggtctacaag aactatgacc cgcgcgccgc gatcgtcaag       420
     gagaccgccc acgacgtcct gacccgtctg ggctccgacg acggcgaccg caagctcgag       480
     atcgccatgg agctggagga gacggcgctg cgcgacgagt acttcgtctc gcgcagcctg       540
     tacccgaacg tc                                                           552
//
```

Listagem 4 Pesquisa *Actinobacillus actinomycetemcomitans* no Oralgen.

Record 1 of 1 from the Actinobacillus actinomycetemcomitans database

Gene ID:AA01943

Genbank Locus Tag:

DNA Molecule Name:

Genbank ID:

Gene Name:

Definition:

conserved hypothetical protein

Cellular Location:

Cytoplasm [Evidence]

Gene Start:

1313148

Gene Stop:

1312990

Gene Length:

159

Molecular Weight*:

6029

pI*:

Net Charge*:

EC:

Functional Class:

Unknown

Gene Ontology:

Pathway: pathway table

Comment:

Human Oral Microbiome Database:

View in HOMD Genome Viewer

Blast Summary: PSI-Blast Search

This sequence is similar to 10954430, an unknown from Actinobacillus actinomycetemcomitans.

Top Blast Hits: Updated monthly

Click here to view the entire PsiBlast results.

gi|10954430|ref|NP_067568.1| hypothetical protein pVT745_p2... 72 8e-12

Extract AA Sequences | Multiple Alignment | Reset

InterPro Summary: InterProScan

No hits reported.

COGS Summary: COGS Search

No hits to the COGs database.

Blocks Summary: Blocks Search

ProDom Summary: Protein Domain Search

Residues 1 to 53 match (1e-13) PD:PD253686 which is described as PLASMID

Paralogs: Local Blast Search

No paralogous sequences are found to this sequence.

AA01943 has no significant similarity (blastp p-value < 1e-3) to any other gene in this genome.

Pfam Summary: Pfam Search

Top PDB Hits:

Gene Protein Sequence:

MSITTQETYRKEGYLPQKVTLSEEVVFFTKIDNGEIKVKSNDEVLANLKK
IIG

Gene Nucleotide Sequence: Sequence Viewer

ATGAGCATAACAACACAAGAAACCTATCGTAAGGAAGGCTATTTGCCGCA
AAAAGTGACACTTTCAGAAGAAGTTGTATTTTTCACTAAAATTGACAATG
GCGAAATTAAAGTAAAATCCAATGATGAGGTCTTAGCCAACCTGAAAAAG
ATTATTGGT

2.3 Formato de Dados de Sequência nos BDBs

Nas listagens 1 a 4 foram apresentados dados de sequência extraídos de alguns principais BDBs: Genbank, DDBJ, EMBL e Oralgen. A Tabela 1 apresenta uma comparação entre os atributos de sequência para cada um desses BDBs. O Genbank e o DDBJ possuem a maioria dos atributos de sequência em comum, exceto o atributo “comentário”, que não aparece nas sequências armazenadas no DDBJ.

O EMBL possui vários atributos similares aos atributos do Genbank e do DDBJ, porém com nomes diferentes, como pode ser verificado na tabela. Por exemplo, o campo *locus* do Genbank e DDBJ possui a nomenclatura de *ID* no EMBL.

Em contrapartida, os atributos da Tabela 1 a partir do atributo *Gene id* são exclusivos do Oralgen e podem ser utilizados para apresentar informação mais completa sobre uma determinada pesquisa.

Tabela 1 Comparação entre os atributos da sequência dos BDBs.

Atributos da sequência	Bancos de Dados Biológicos			
	Genbank	**DDBJ**	**EMBL**	**Oralgen**
Locus	Locus	Locus	ID	Genbank locus tag
Definição	Definition	Definition	DE	Definition
Accesso	Accession	Accession	AC	-
Versão	Version	Version	-	-
Palavras-chave	Keywords	Keywords	KW	-
Fonte	Source	Source	OS + OC	-
Referência	Reference	Reference	RN	-
Comentário	Comment	-	-	Comment
Características	Features	Features	FHFT	Features/gene/gene = gene name Features/gene = gene start + gene stop Feature/CDS/translation *
Origem	Origin	Origin	SQ	Gene nucleotide sequences
Gene id	-	-	-	Gene id
Localização celular	-	-	-	Cellular location
Tamanho do gene	-	-	-	Gene length
Peso molecular	-	-	-	Molecular weight
PL	-	-	-	PL
Net charge	-	-	-	Net charge
EC	-	-	-	EC

Classe funcional	-	-	-	Functional class
Gene Ontology	-	-	-	Gene ontology
Vias metabólicas	-	-	-	Pathway
Evidência secundária	-	-	-	Secondary evidence
Human oral microbiome database	-	-	-	Human oral microbiome database
Sumário no blast	-	-	-	Blast summary
Top blast hits	-	-	-	Top blast hits
Sumário no Inter Pro	-	-	-	Inter Pro summary
COGS summary	-	-	-	COGS summary
Blocks summary	-	-	-	Blocks summary
ProDom summary	-	-	-	ProDom summary
Paralogs	-	-	-	Paralogs
Pfam summary	-	-	-	Pfam summary
Características estruturais	-	-	-	Structural features
Top PDB hits	-	-	-	Top PDB hits

* abrangência parcial

Além da diferença dos termos, também existe a diferença nos formatos de dados mostrados pelos BDBs escolhidos. Por exemplo, a sequência exibida pelo DDBJ é diferente da sequência do EMBL, como pode ser observado nas listagens 2 e 3 deste capítulo.

2.4 Considerações Finais

Neste capítulo foram apresentados os conceitos sobre Bancos de Dados Biológicos, tipos de dados biológicos e foram exibidas consultas em alguns BDBs. Analisando tais consultas, foi possível identificar diferenças entre o formato dos dados armazenados em três principais BDBs públicos e o Oralgen.

Capítulo 3

INTEGRAÇÃO DE BANCOS DE DADOS

Neste capítulo são descritos os principais conceitos sobre Integração de Bancos de Dados, citando os níveis afetados pela integração de dados, arquitetura de integração, teoria sobre Integração de Esquemas e Integração de Instâncias, algoritmos para Resolução de Entidades e fatores que devem ser considerados na integração de dados biológicos.

3.1 Considerações Iniciais

A expansão do uso dos bancos de dados e a evolução de algumas aplicações de negócios trouxeram a necessidade de integrar bancos de dados de um mesmo domínio (LIU e ÖZSU, 2009). Por exemplo, diferentes departamentos de uma universidade podem desenvolver bancos de dados independentes que podem ter informação em comum tal como pesquisador. Um passo futuro é integrar esses dados, sendo que o BD de um departamento 1 pode ter um conjunto de atributos A, B e C, enquanto o BD do departamento 2 pode ter outro conjunto de atributos C, D e E, ou seja, alguns atributos são comuns, enquanto outros são específicos de cada banco de dados. O intuito da integração desses dados é consolidar os dados de ambas as fontes de dados, que pode ser efetuado em um terceiro banco de dados, com os atributos A, B, C, D e E. Essa “consolidação” é realizada com o uso de técnicas e métodos da área de Integração de Banco de Dados.

Além disso, o dinamismo dos negócios requer adequação das aplicações e bancos de dados utilizados para prover infra-estrutura, além de um ambiente

propício para tomada de decisões em cenários competitivos. A Integração de Banco de Dados vem suprir essa necessidade.

A integração de dados afeta três níveis diferentes: nível de sistemas, nível lógico e desafios sociais (LING e ÖZSU, 2009).

No **nível de sistemas**, tem-se que a *Structured Query Language* (SQL) é uma linguagem padrão de consultas a Sistemas Gerenciadores de Bancos de Dados (SGBDs), porém cada um deles implementa o SQL de uma forma. Além disso, mesmo que a forma de armazenamento mais comum dos dados seja o SGBD relacional, existem diversas outras formas de armazenamento de dados, por exemplo: arquivos textos, documentos semi-estruturados (*eXtensible Markup Language* – XML), bancos de dados orientados a objeto, bancos de dados hierárquicos, entre outros.

Embora limitando o domínio de aplicação, no **nível lógico** surgem diferenças entre os esquemas a serem integrados, como as citadas abaixo:

- Nomes de tabelas e atributos;
- Organização dos dados (tabular ou hierárquica, por exemplo);
- Cobertura do domínio em nível de detalhes;
- Diferenças na representação de um objeto do mundo real no banco de dados.

Dentre os **desafios sociais** na integração de dados, os principais são em relação a encontrar um determinado dado no sistema, e a dificuldade de cooperação das pessoas nesse processo.

Segundo Ziegler e Dittrich (2004), há duas razões principais para integrar dados. Primeiro, para um conjunto de sistemas de informação existentes, uma visão integrada pode ser criada para facilitar o acesso aos dados e o seu reuso por meio de um único ponto de acesso. Em segundo lugar, dados de diferentes sistemas de informação complementares são combinados para gerar um banco de dados mais abrangente para satisfazer uma necessidade específica.

Ziegler e Dittrich (2004) afirmam ainda que para a integração, os dados devem ser representados utilizando os mesmos princípios de abstração. A tarefa de integração de dados inclui a detecção e resolução de conflitos de esquema e dados relacionados a estrutura e semântica.

3.2 Conceitos de Integração de Dados

O objetivo da integração de dados é criar uma visão integrada a fim de tornar mais fácil o acesso aos dados e permitir o seu reuso por meio de um único ponto de acesso. Nesse sentido, dados complementares de diferentes sistemas de informação são combinados para gerar um banco de dados mais abrangente.

A Figura 8 exibe a arquitetura de um sistema de integração de dados: de um lado temos várias fontes de dados que desejam-se integrar por meio de um mapeamento para gerar um único banco de dados destino.

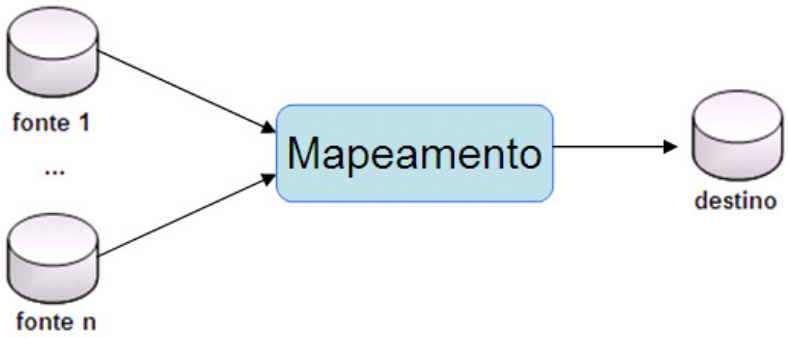

Figura 8 Arquitetura de um sistema de integração de dados.

A integração de dados pode ser feita em nível de esquema e/ou em nível de instância. A primeira ocorre quando deseja-se integrar fontes de dados estruturalmente, enquanto a última existe quando deseja-se integrar os valores dos atributos.

Para a ***integração em nível de esquema***, existem três possíveis arquiteturas de integração: integração usando esquema mediado, integração baseada em *data warehouse* e integração por meio de banco de dados federados.

A Figura 9 exibe a **arquitetura de integração usando esquema mediado**. As fontes de dados são consultadas por pequenos programas, chamados *wrappers* (encapsuladores), que também têm como função retornar uma resposta adequada para o esquema mediado, e consequentemente ao usuário. Os *wrappers* são basicamente “tradutores”, que traduzem a consulta do usuário para as fontes de dados heterogêneas, e também realizam a tradução da resposta fornecida por tais fontes para a visualização do usuário.

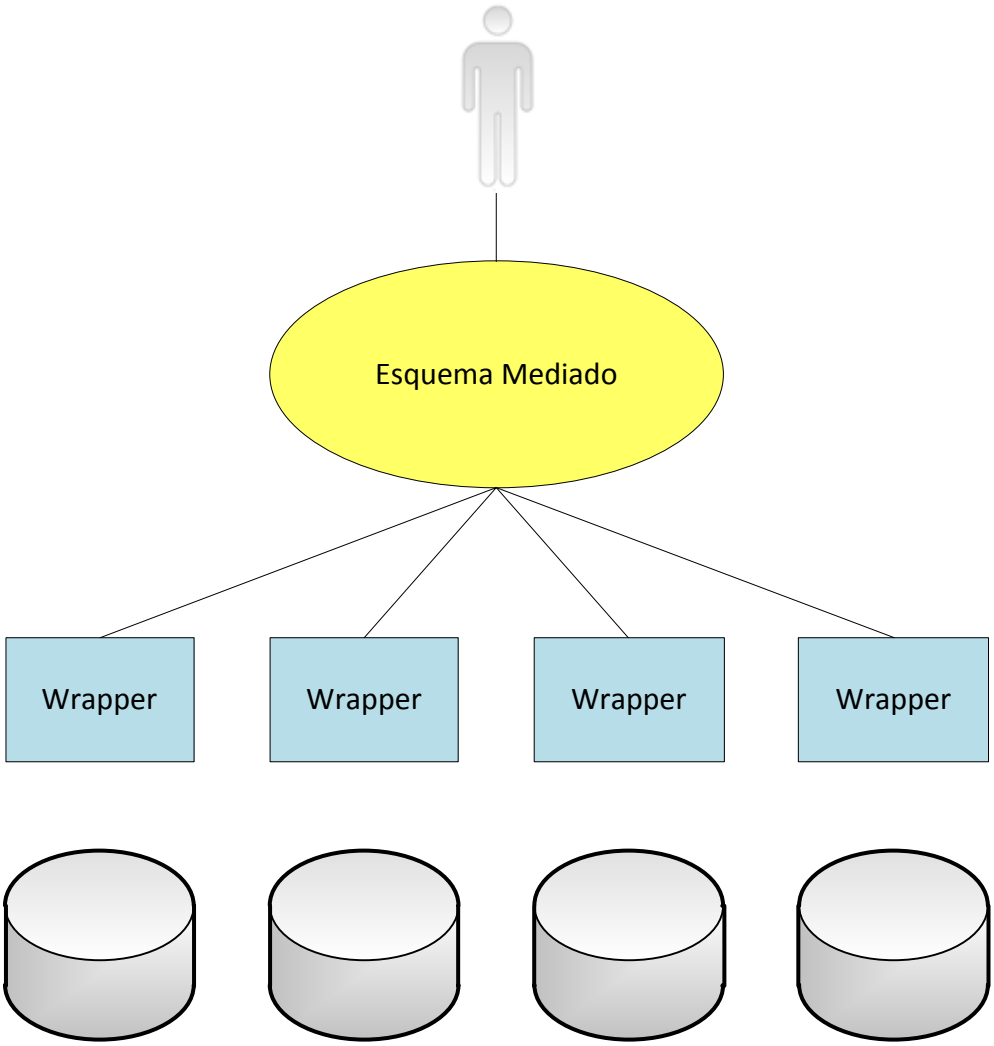

Figura 9 Arquitetura de integração usando esquema mediado.

O esquema mediado é um esquema lógico que possui apenas informações necessárias para a aplicação. Ele não contém todas as informações de todas as fontes de dados, mantendo apenas informações relevantes para a aplicação. Descrições das fontes permitem o mapeamento entre os dados das fontes e o esquema global onde o usuário realiza suas consultas. Na arquitetura apresentada na Figura 9, o usuário tem a impressão de estar interagindo com apenas um sistema de informação.

Existem duas linguagens que realizam o mapeamento do esquema nessa arquitetura:

- *Global As a View* (GAV): Descreve o esquema mediado como um conjunto de definições sobre as relações fonte.

- *Local As a View* (LAV): As fontes de dados são descritas como visões sobre o esquema mediado.

A linguagem GAV facilita o processamento de consultas conceitualmente, pois existe uma correspondência direta entre o esquema mediado e as relações dos dados da fonte. Por outro lado, a linguagem LAV é mais fácil de estender e dar manutenção, pois independe das fontes de dados. Com a LAV, é mais fácil de descrever restrições e utilizar o menor número de fontes para uma consulta, diminuindo os custos envolvidos nessa operação (LIU e ÖZSU, 2009).

Na **arquitetura de integração baseada em *data warehouse***, dados são extraídos das várias fontes e integrados em um *data warehouse*, como apresentado na Figura 10. As principais vantagens dessa arquitetura são a disponibilidade dos dados e o desempenho das consultas.

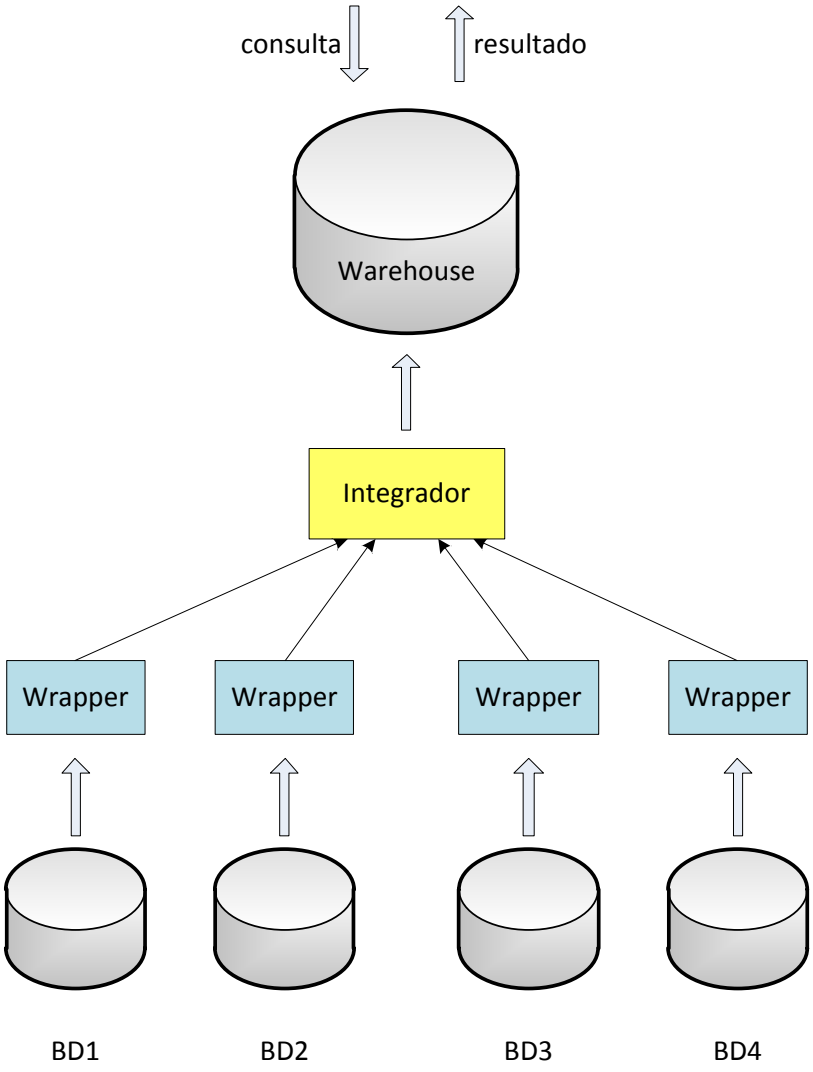

Figura 10 Arquitetura de integração baseada em *data warehouse*, adaptada de Medcraft (2003).

Na **arquitetura de integração por meio de bancos de dados federados** são definidos um modelo de dados e uma linguagem de consulta globais. Como mostrado na Figura 11, a consulta é disparada no SGBD global, sobre o esquema global. O usuário que interage com o esquema global não precisa conhecer detalhes dos bancos de dados locais. É realizada uma conversão do esquema global para os esquemas de dados locais.

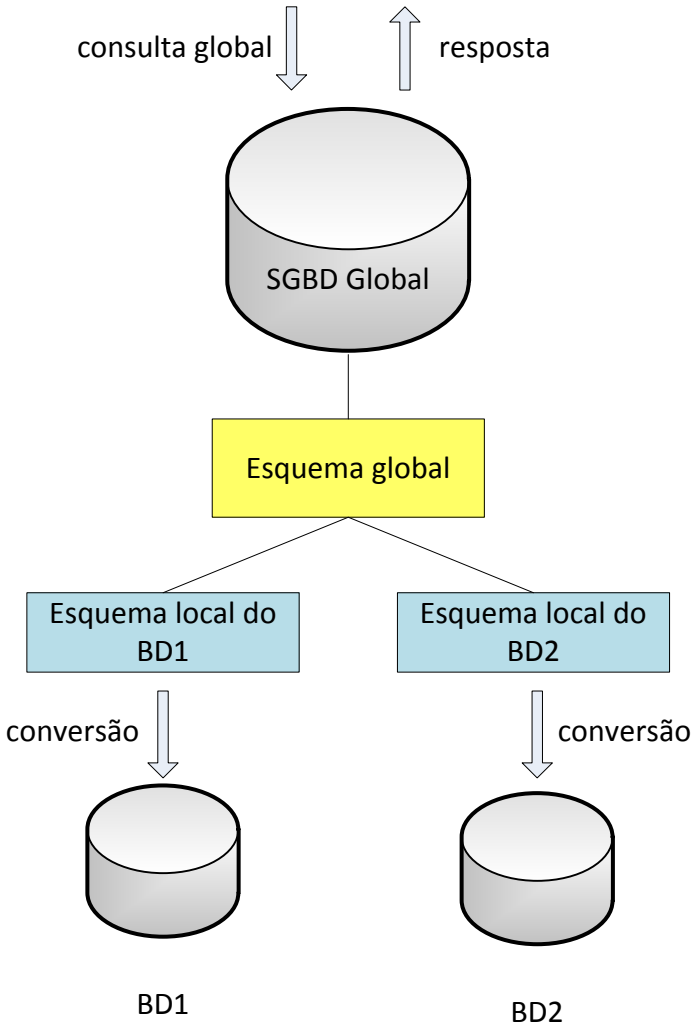

Figura 11 Arquitetura de integração por meio de bancos de dados federados, adaptada de Medcraft (2003).

A ***integração de instâncias*** é a integração dos dados em nível de tuplas, no caso de um banco de dados relacional. Esse último tipo de integração também é conhecido como integração de entidades, Resolução de Entidades (RE), *deduplication* e *merge-purge* (BENJELLOUN et al., 2009).

Menestrina, Benjelloun e Garcia-Molina (2006) afirmam que a resolução de entidades identifica os registros que se referem (ou são similares) a uma entidade e realiza a fusão (*merge*) desses registros. Um registro que sofreu fusão torna-se um "composto" dos registros fonte.

Assim, a RE baseia-se em encontrar registros que são similares e podem ser fundidos. A fusão dos registros dá origem a um novo registro, que por sua vez também é comparado aos demais para identificação de similaridade.

Por exemplo, suponha três entidades e_1, e_2 e e_3 denotadas da seguinte forma, que também estão representadas na Figura 12.

e_1: Definição, Gene

e_2: Gene, Tam

e_3: Def, Gene

E1	Definição	Gene
	Fly Eyeless	ACT...

E2	Gene	Tam
	ACT...	35

E3	Def	Gene
	Fly ey	ACT...

Figura 12 Exemplo de entidades a serem integradas.

É possível fundir e_1 e e_2 com base no atributo *Gene*. Assim será criada uma nova entidade, representada na Figura 13.

E12	Gene	Definição	Tam
	ACT...	Fly eyeless	35

Figura 13 Integração das entidades E1 e E2.

A entidade e_{12} pode ser fundida à entidade e_3, pois existe uma correspondência dos atributos *Gene* e *Definição* comuns a ambas as entidades, e *Fly ey* é a abreviação de *Fly eyeless*. Assim, uma nova entidade será criada, conforme mostra a Figura 14.

E123	Gene	Definição	Tam
	ACT...	Fly eyeless	35

Figura 14 Integração das entidades E12 e E3.

Formalizando, a resolução de entidades genérica (MENESTRINA, BENJELLOUN e GARCIA-MOLINA, 2006) é baseada nas funções *match* e *merge*:

- A função *match*, denotada por M(*r*, *s*) retorna o valor verdadeiro caso *r* e *s* representem a mesma entidade, representada por $r \approx s$.
- A função *merge* cria um novo registro a partir de dois registros similares. A representação dessa operação entre dois registros *r* e *s* é denotada por <*r*,*s* >.

A definição da resolução de entidades genérica, segundo Menestrina, Benjelloun e Garcia-Molina (2006) é que dado um conjunto de registros R, o resultado da resolução de entidade RE(R) é o menor conjunto S tal que:

1. $R \subseteq S$
2. Para quaisquer registros $r_1, r_2 \in S$, se $r_1 \approx r_2$ então $< r_1, r_2> \in S$

Assim, o conjunto S deve ter tanto os registros originais quanto os registros derivados dos *merges*.

Benjelloun et al. (2009) apresentam o algoritmo de força bruta para gerar o conjunto S = RE(R), conforme mostrado na Figura 15.

```
1: entrada: um conjunto I de registros
2: saída: um conjunto I' de registros, I' = RE(I)
3: I' ← I ; N ← ∅
4: repita
5:     I' ←I' ∪ N; N ←∅
6:     para todos pares de registros (r, r') em I'
7:         se r ≈ r' então
8:             merged ← < r, r' >
9:           se merged ∉ I' então
10:                inclua merged em N
11:          fim-se
12:        fim-se
13:    fim-para
14: até N = ∅
15: para todos pares de registros (r, r') em I ' onde r ≠ r'
16:     se r' ≤ r então
17:         Remova r' de I'
18:     fim-se
19: fim-para
```

Figura 15 Algoritmo de força bruta para gerar RE(R), adaptado de Benjelloun et al. (2009).

O algoritmo de força bruta executa várias comparações desnecessárias, que poderiam ser evitadas caso o histórico das combinações fosse armazenado. Para

eliminar essas comparações, Benjelloun et al. (2009) propõem um novo algoritmo, G-Swoosh, apresentado na Figura 16.

```
1: entrada: um conjunto I de registros
2: saída: um conjunto I' de registros, I' = RE(I)
3: I' ← ∅
4: enquanto I ≠ ∅
5:     r ← um registro de I
6:     Remova r de I
7:     para todos registros r' em I' U {r}
8:         se r ≈ r' (resp. r' ≈ r) então
9:             merged ← < r, r' > (resp. < r', r >)
10:            se merged ∉ I U I' U {r} então
11:                inclua merged em I
12:            fim-se
13:        fim-se
14:    fim-para
15:    Inclua r em I'
16: fim-para
17: Remova os registros dominados de I' (veja linhas 15 a 18 do algoritmo de força bruta)
18: retorne I'
```

Figura 16 Algoritmo G-Swoosh, adaptado de Benjelloun et al. (2009).

Além desses algoritmos, existem vários outros que realizam a resolução de entidades: On et al. (2006), Bhattacharya e Getoor (2007); Chen, Kalashnikov e Mehrotra (2005).

3.3 Integração de Dados Biológicos

Devido à particularidade dos dados biológicos, para integrá-los, é necessário considerar, além dos aspectos descritos na Seção 3.2, as premissas descritas abaixo:

- A quantidade de dados biológicos. O aumento do sequenciamento de dados biológicos gerou, consequentemente, um crescimento na quantidade de dados armazenados pelos BDBs. Por exemplo, em 2006 o Genbank possuía mais de 61 milhões de sequências (BENSON et al., 2006), e em 2011 ele possui mais de 146 milhões sequências reportadas (NCBI-GENBANK, 2011). Ou seja, em cinco anos, o número de sequêcias armazenadas no Genbank aumentou 2,4 vezes. A taxa rápida de crescimento de dados pode levar a

mudanças no esquema do banco de dados como uma tentativa de lidar com essa grande quantidade de dados. Porém, mudar o esquema de um banco de dados altera a forma como é feito o armazenamento dos dados e a consulta dos mesmos é realizada, complicando o esforço de integração.

- Os BDBs não cresceram somente em tamanho, mas também em quantidade. Em 2005, o número de banco de dados biológicos existentes era de 719 (GALPERIN, 2005), enquanto que em 2011 esse número passou para 1330 BDBs (GALPERIN e COCHRANE, 2011). A diferença foi de 1,8 a mais de BDBs em seis anos. Alguns desses BDBs são locais e tem foco em um determinado organismo de um grupo de pesquisa local.
- O acesso aos dados pode ser realizado com diversas técnicas. Em bancos de dados privados, geralmente a técnica utilizada é a linguagem de consulta SQL, enquanto em bancos de dados públicos geralmente existe uma interface de consulta, com diferentes recursos disponíveis para o usuário. A interface de consulta costuma ser um formulário *web*, com vários parâmetros, alguns deles estão preenchidos com o valor *default*, facilitando assim a compreensão do usuário. A interface pode continuar a mesma, ainda que existam mudanças no esquema do banco de dados. Para os bancos de dados públicos existe ainda a opção de realizar consultas via *e-mail*. Mesmo com tantas facilidades, ainda existem campos cuja nomenclatura é difícil de entender pelos usuários, dificultando a interpretação dos dados de entrada e o resultado de uma consulta. Quando o objetivo é integrar vários BDBs, esse problema é agravado, pois eles podem existir em cada BDB individualmente.
- Há uma grande quantidade de formatos e tipos de dados. Como não há uma padronização, um mesmo dado pode ser representado de uma certa forma em um BDB A e em um formato diverso no BDB B. Considerando o número de BDBs existentes supramencionado, é possível ter uma ideia de que podem existir muitos formatos e tipos criados, principalmente para BDBs autônomos. A representação dos dados é um ponto que deve ser considerado no momento da integração de dados, para identificação de mesma entidade. Quando existem muitos dados de formatos e tipos heterogêneos, o esforço de integração dos mesmos é maior.
- Além da heterogeneidade de tipos e formatos de dados, podem ocorrer também heterogeneidade de nomes (heterogeneidade sintática), de valores,

de estrutura do banco de dados e heterogeneidade semântica. Essa última refere-se ao significado dos dados entre os BDBs, e é mais difícil de ser detectada, pois existem situações em que dois termos diferentes referem-se à mesma entidade (sinônimos) e existe o caso em que um mesmo termo pode ter significados diferentes para dois ou mais BDBs distintos.

- A qualidade dos dados pode ter ruídos devido a erros provenientes de diferentes fatores, como técnicas de laboratório que geram os dados biológicos ou erro humano. Por exemplo, Wesche, Gaffney e Keightley (2004), em seu trabalho que usa o genoma do camundongo como referência, estimam que a taxa de erro de um único nucleotídeo para a codificação do DNA é de 0,1%. É uma taxa relativamente baixa, mas ainda acontece, pois os equipamentos de sequenciamento trabalham com probabilidades. Em dados mais antigos, a taxa de erro é maior devido a equipamentos mais antigos. Ruídos também podem ocorrer em dados de *strings* devido a erros de digitação humanos. O uso de distância de edição pode auxiliar a mapear similaridade de literais, como uma tentaitva de minimizar esse problema.

3.4 Distância de Edição

Para calcular a similaridade entre duas *strings* de caracteres se faz necessário o uso de uma métrica que compare esses caracteres e retorne um valor, indicando a similaridade entre os dois literais. Existem vários algoritmos (e várias métricas) de similaridade possíveis. O nome dado a esse tipo de métrica é distância de edição.

A distância de edição é importante na comparação entre valores de instâncias, no momento da integração das mesmas. Ela pode ser utilizada para indicar similaridade entre literais, fornecendo um resultado que pode ser analisado para indicar se os literais referem-se ao mesmo objeto do mundo real.

Duas distâncias de edição bem conhecidas são a Distância de Hamming e a Distância de Levenshtein.

A Distância de Hamming (HAMMING, 1950) realiza a contagem do número de posições cujos símbolos sejam diferentes.

A Distância de Levenshtein (LEVENSHTEIN, 1966) por sua vez calcula o número mínimo de edições necessárias para transformar um literal em outro, usando as operações de adição de caracter, remoção ou substituição de caracter.

Por exemplo, calculando a distância de edição de Hamming (DH) e de Levenshtein (DL) para os seguintes literais, tem-se:

1. *Actinomyces naeslundii* strain E1-20 DNA gyrase subunit A (gyrA) gene, partial cds.
2. *Actinomyces naeslundii* strain R24330 DNA-directed RNA polymerase beta-subunit (rpoB) gene, partial cds.

DH(1,2) = 48

DL(1,2) = 34

3.5 Considerações Finais

Neste capítulo foram apresentados os conceitos sobre integração de dados, a teoria sobre integração de esquemas e integração de instâncias. Baseando-se nos conceitos de integração de dados e nos dados biológicos mencionados no Capítulo 2, foi possível fazer algumas considerações sobre a integração de dados biológicos. O próximo capítulo apresenta os trabalhos correlatos ao MIDB.

Capítulo 4

TRABALHOS CORRELATOS

Neste capítulo são descritos os trabalhos envolvendo Integração de Bancos de Dados Biológicos, alguns que realizam integração de esquemas e outros que tratam de integração de instâncias. Os principais trabalhos correlatos a este projeto são detalhados e comparados. Ao final, será apresentada uma tabela comparando as características dos trabalhos correlatos com as funcionalidades do modelo de integração de dados biológicos proposto neste trabalho de mestrado.

4.1 Considerações Iniciais

Nas próximas seções são descritas as principais características de alguns sistemas que realizam integração de esquemas e integração de instâncias de dados biológicos. Ao final, uma tabela comparativa entre eles é mostrada comparando as características dos trabalhos correlatos com as características do modelo de integração de dados biológicos proposto neste trabalho de mestrado.

4.2 YeastMed

O YeastMed (BRIACHE et al., 2010) é um sistema de integração de dados biológicos baseado em mediador, que recebe uma consulta do usuário e decompõe-na em subconsultas em *XQuery* (linguagem de consulta ao XML) que são disparadas contra cinco banco de dados do organismo *Saccharomyces cerevisiae*.

Para isso, o YeastMed conta com o uso do SB-KOM (NAVAS-DELGADO e ALDANA-MONTES, 2009), um mediador baseado em ontologia que faz a decomposição da consulta em tempo de execução, com base em um conjunto de regras de mapeamento.

Cada fonte de dados foi modelada em um esquema XML, que traduz o esquema fonte na ontologia própria do YeastMed. Assim, a ontologia interna do YestMed representa os esquemas dos cinco banco de dados reconciliados em um único esquema global.

Dessa forma, quando o usuário realiza uma consulta utilizando os termos da ontologia, o YeastMed converte-a para requisições *XQuery* para as fontes de dados necessárias. Essas requisições são executadas por *web services* que retornam dado XML. Um avaliador/integrador estabelece relacionamentos entre esses dados retornados, baseado em propriedades da ontologia para fazer a integração.

Por se tratar de um mediador que utiliza a linguagem *Global As a View* (GAV), esse sistema tem dificuldades em ser estendido para integrar um novo banco de dados. Para incluir outro BDB seria necessário recriar a ontologia interna.

4.3 Algoritmo PIC

O Algoritmo PIC (GREENE, BRYAN e CUNNINGHAM, 2008) é um sistema de integração que utiliza clusterização para integrar os dados de duas ou mais fontes heterogêneas.

Formalmente: seja X o conjunto de todos os objetos de dados do domínio. Nesse domínio um usuário possui acesso a um conjunto de visões v, onde X_n é um subconjunto de objetos presentes em X na visão n.

A partir dessas visões de dados, executa-se um algoritmo de agrupamento que gera uma coleção de clusters (também denominados agrupamentos) base para cada visão, onde C_n representa a coleção de clusters gerados na visão X_n, conforme representado na Figura 17.

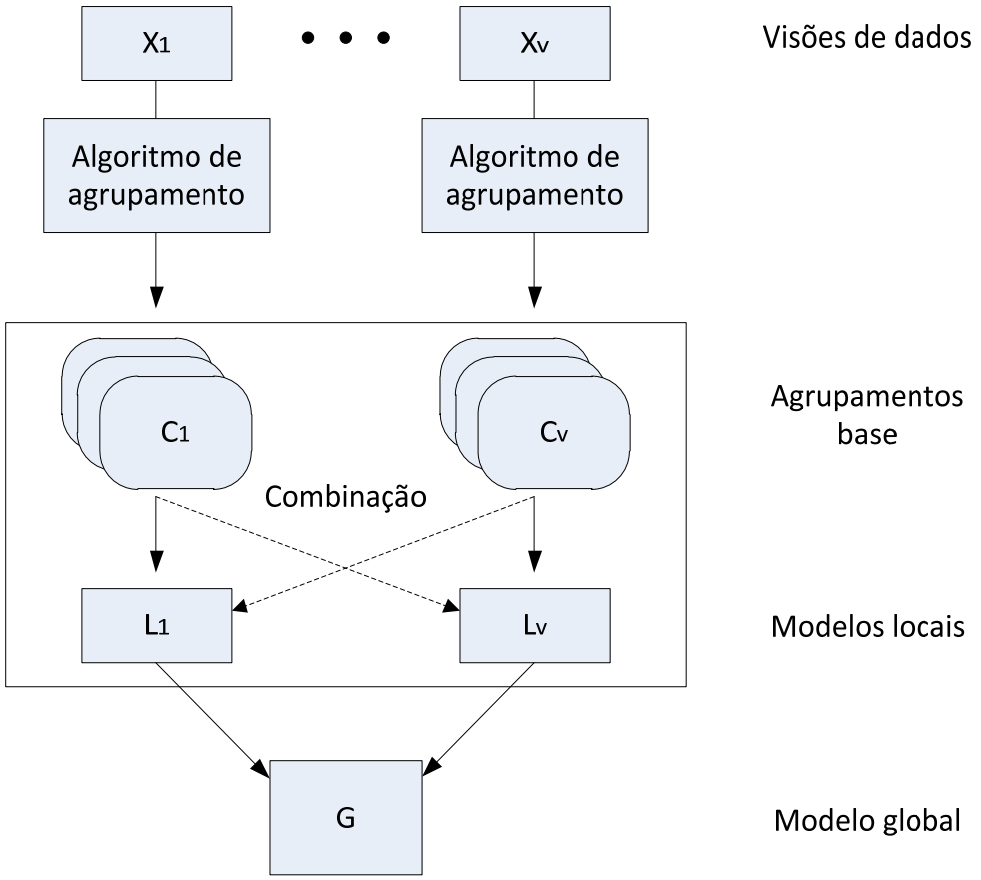

Figura 17 Visão geral do funcionamento do algoritmo PIC, adaptada de Greene,Bryan e Cunningham (2008).

Nesse cenário, o algoritmo PIC executa duas fases:

- Fase 1: Geração de um conjunto de modelos locais *L*, que são o resultado da combinação entre os clusters base da visão X_n e outras visões. Os clusters gerados possuem sobreposição de alguns elementos.
- Fase 2: Criação do modelo global *G* por meio da fusão de elementos comuns entre os modelos locais. Os clusters únicos a cada modelo são preservados.

Na fase 2, considera-se cada par de clusters por meio de todos os modelos locais e realiza-se a fusão caso o **coeficiente de sobreposição binário** seja maior que 0,5, com base em experimentos. O coeficiente é dado pela fórmula abaixo:

$$\text{Coeficiente de sobreposição binário } (A, B) = \frac{|A \cap B|}{\min(|A|, |B|)}$$

Onde A e B são conjuntos. O valor de 1 indica que B está contido integralmente ou é idêntico a A. O valor de 0 indica que não há nenhum elemento similar entre A e B.

4.4 Bio-AXS

O Bio-AXS (SEIBEL, LEMOS e LIFSCHITZ, 2003) é um *framework* orientado a objetos que visa integrar dados biológicos. Os *hot spots* (partes extensíveis do *framework*, que podem ser editadas pelo usuário) são os encapsuladores (*wrappers*) utilizados para fazer interface com as bases de dados.

Esse sistema é dividido em módulos, onde o módulo de administração contém um modelo de classe da biologia compatível com as fontes de dados e os métodos da classe, chamado de Modelo Biológico. O Bio-AXS também está equipado com um módulo conversor que traduz o esquema das fontes para esquema XML e converte os dados para XML.

O Bio-AXS faz a integração por meio de mediadores, usando a linguagem *Local As a View* (LAV), que é extensível. A integração pode ser feita com qualquer fonte de dados biológicos, desde que se crie um encapsulador para a tradução do esquema do banco de dados para esquema XML e a conversão dos dados para XML. O administrador do sistema é responsável por fazer a incorporação desse novo esquema XML no esquema global.

Por exemplo, o administrador pode criar um esquema de dados para cada projeto de pesquisa existente e, no momento da consulta, um dos esquemas pode ser escolhido. De acordo com a escolha, a ontologia utilizada é diferente (i.e. a ontologia muda de acordo com o esquema global selecionado).

Para integrar os dados de diferentes fontes são utilizadas descrições dos objetos, definição de relacionamentos, identificação de sinônimos e regras de conversão.

4.5 Atlas

No sistema Atlas (SHAH et al., 2005), os dados biológicos das fontes são armazenados em um *data warehouse* que é consultado por meio de uma *Application Programming Interface* (API), a qual encapsula comandos SQL. A interface com o usuário utiliza os métodos da API para realizar as interações com o sistema.

As fontes de dados utilizadas pelo Atlas são de quatro categorias: sequência, interações moleculares, recursos relacionados a genes, e ontologias.

O Atlas utiliza ontologias externas e uma ontologia própria que descreve o modelo de características de sequência do Genbank. A ontologia é melhorada com termos mais específicos como *is-synonym-of*.

Esquemas tabulares foram criados para cada um dos quatro tipos de dados que o Atlas provê suporte, o que dificulta a integração de novas fontes de dados, pois os esquemas e mapeamentos terão que ser alterados.

O Atlas tem como desafio lidar com conflitos de integração nos quais decisões precisam ser tomadas para realizar uma resolução de conflito.

A integração do Atlas é realizada em dois níveis. O primeiro nível integra dados de tipos similares de diferentes fontes por meio de um modelo de dados em comum. O segundo nível utiliza as ontologias e referência cruzada para fazer a resolução de conflitos.

4.6 Considerações Finais

Considerando o enfoque deste projeto de pesquisa e as suas especificações, a Tabela 2 foi elaborada para comparar o contexto de aplicação do modelo proposto neste trabalho de mestrado com o contexto dos trabalhos correlatos descritos nas seções anteriores. Dentre os trabalhos correlatos, o YeastMed e o Bio-AXS utilizam a técnica de mediadores, sendo que o primeiro faz uso da linguagem GAV (não

extensível) e o segundo, da linguagem LAV (que é extensível). O algoritmo PIC usa a técnica de clusterização, enquanto o Atlas implementa um *data warehouse*.

Tabela 2 Elementos abordados pelos trabalhos correlatos ao modelo proposto.

Ferramenta	Técnica	Domínio	Usa dicionário?	Usa XML?	Facilmente extensível?
YeastMed Briache et al., 2010	Mediador GAV	Dados do *Saccharomyces cerevisiae*	Não	Sim	Não
Algoritmo PIC GREENE, BRYAN e CUNNINGHAM, 2008	Clusterização	Gene, publicações e proteínas	Não	Não	Sim
Bio-AXS SEIBEL, LEMOS e LIFSCHITZ, 2003	Mediador LAV	Dados genômicos	Sim	Sim	Sim
Atlas SHAH et al., 2005	*Data warehouse*	Sequências, interações moleculares, genes e ontologias	Sim	Não	Não
MIDB	Categorização	Sequências de nucleotídeos	Sim	Sim	Sim

Dentre os trabalhos apresentados, apenas o YeastMed e o Bio-AXS utilizam o XML como forma a facilitar a integração dos dados, devido ao caráter semi-estruturado desse documento.

O uso de um dicionário de sinônimos é importante pois ajuda a distinguir que dois termos referem-se a uma mesma entidade do mundo real. Os sistemas que fazem uso dessa técnica são o Bio-AXS e o Atlas.

Os trabalhos correlatos a este são de diversos domínios das ciências biológicas. O YeastMed trabalha com o domínio de dados do *Saccharomyces cerevisiae*. O algoritmo PIC usa dados de genes, publicações e proteínas. O Bio-AXS, por sua vez, foca em dados genômicos, enquanto o Atlas trabalha com quatro tipos de dados: sequências, interações moleculares, genes e ontologias.

O modelo proposto (MIDB) usa a abordagem de Categorização para fazer a integração dos esquemas, utilizando XML para armazenamento de dados de diferentes BDBs (com diferentes esquemas), facilitando no momento da integração. Ademais, o uso de um dicionário auxilia na detecção de sinônimos.

Como a integração de esquemas no MIDB é realizada automaticamente via *software* (sem intervenção humana), o sistema é mais extensível em relação à adição de outros bancos de dados biológicos.

Além disso, outra novidade do modelo proposto é a interação com o usuário para realizar decisões de integração. Ou seja, um especialista de domínio pode auxiliar na resolução de conflitos no momento da integração de instâncias (ou resolução de entidades), o que garante uma melhor qualidade do processo de integração.

Capítulo 5

MODELO DE INTEGRAÇÃO DE BANCOS DE DADOS BIOLÓGICOS (MIDB)

Neste capítulo é apresentada a proposta do Modelo de Integração de Dados Biológicos denominado MIDB, assim como são apresentados os detalhes sobre a metodologia desta pesquisa.

5.1 Considerações iniciais

Este trabalho tem o objetivo de propor um **Modelo de Integração de Dados Biológicos** (MIDB), contemplando especificamente a integração de sequências de nucleotídeos. Esse modelo deve permitir a integração de diferentes fontes de dados e auxiliar biólogos na identificação de similaridade de dados de sequências. A seção seguinte provê um detalhamento do trabalho de pesquisa, mostrando quais as etapas seguidas pelo mesmo.

5.2 Detalhamento do trabalho de pesquisa

O modelo proposto por este trabalho pode ser observado na Figura 18. Ela é composta por quatro etapas: Carregamento de Dados (Etapa 1); Categorização dos

Atributos (Etapa 2); Resolução de Entidades (Etapa 3); e Gerenciamento de Clusters (Etapa 4).

Os dados extraídos das fontes de dados são armazenados localmente (Etapa 1) e seus atributos são mapeados em categorias, para identificação de atributos similares (Etapa 2). Em seguida, os valores de alguns atributos importantes são comparados para verificar se correspondem à mesma entidade do mundo real (Etapa 3). Caso não haja identificação automática de similaridade após essa comparação, é solicitada a revisão de um especialista de domínio e após sua decisão os dados são armazenados no banco de dados local (Etapa 4).

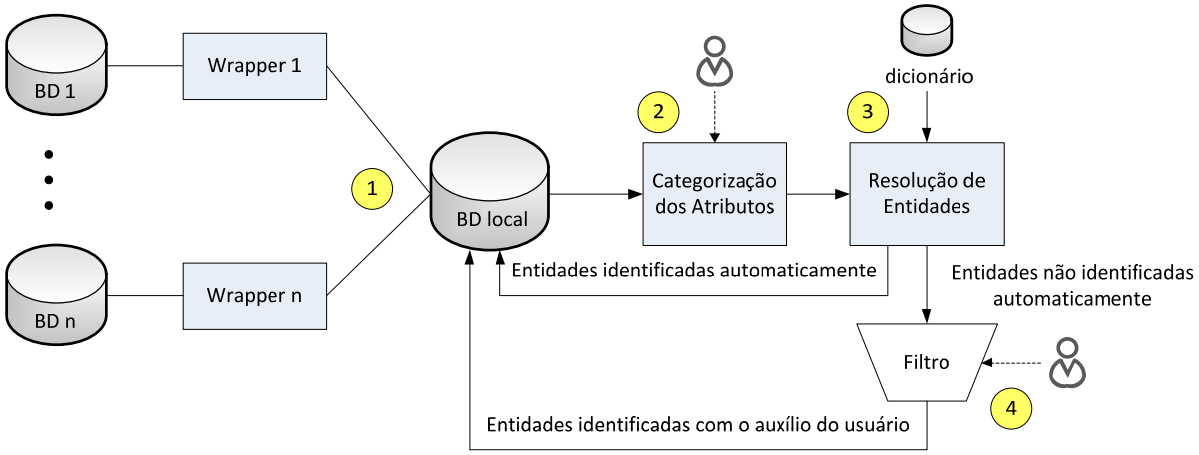

Figura 18 Arquitetura do sistema MIDB.

Na etapa de **Carregamento de Dados**, os dados de sequência de nucleotídeos são extraídos de um conjunto de bancos de dados biológicos desse domínio, por meio do uso de encapsuladores (*wrappers*) que fazem a conexão e consulta nas bases de dados e retornam um documento XML, que é armazenado no banco de dados local, devido à disponibilidade e privacidade que podem ser oferecidas por esse tipo de armazenamento.

Na etapa de **Categorização dos Atributos**, os atributos dos BDBs são divididos em categorias, indicando seu escopo. Exemplos de categorias são Locus, Descrição e SeqNuc. Essa etapa serve para realizar a integração do esquema das fontes de dados envolvidas.

Em seguida, na etapa de **Resolução de Entidades**, utilizam-se os valores provenientes de duas Categorias de atributos (Descrição e SeqNuc) para identificar similaridade entre registros. Um algoritmo de clusterização proposto (que conta com

o uso de dicionário de dados), bem como o BLAST são utilizados e em seguida é realizado um emparelhamento dos resultados obtidos.

Em alguns casos o emparelhamento produz resultados de similaridade, mas em outros é necessária uma fase seguinte, a etapa de **Gerenciamento de Clusters**, na qual o especialista de domínio toma a decisão final em relação a um registro ter similaridade ou não com outros, considerando alguns fatores, tais como sequência do organismo, descrição, e demais atributos.

A seguir, são descritas cada uma das etapas do modelo proposto.

5.3 Etapa 1: Carregamento dos Dados

Em primeiro lugar é preciso definir quais bancos de dados biológicos serão utilizados. Esses BDBs devem ter dados de sequência de nucleotídeos.

Para fazer a extração dos dados é necessário criar um encapsulador para cada fonte de dados. Ele é responsável por realizar a conexão com o banco de dados, buscar os dados de sequência desejados e retornar ao banco de dados local.

Para facilitar a integração dos dados na etapa posterior, esses dados devem estar organizados no formato XML, da seguinte forma: cada atributo da tabela deve ser mapeado como nome do elemento XML; e cada valor de atributo deve ser mapeado como conteúdo do elemento correspondente.

A Listagem 1 apresenta a formatação de um registro no Genbank. Na Figura 19 tem-se esse mesmo registro formatado em XML. Observe que o atributo *locus* foi transformado no elemento *<locus>...</locus>* e o valor desse atributo está entre as *tags* *<locus>* e *</locus>* na Figura 19.

```
<?xml version="1.0" encoding="UTF-8"?>
<locus>AF048778                 315 bp    DNA     linear   BCT 09-FEB-1999</locus>
<definition>Actinomyces naeslundii urease gamma-subunit (ureA) gene, complete cds.</definition>
<accession>AF048778  </accession>
<version>AF048778.1  GI:4249596</version>
<keywords>.</keywords>
<source>Actinomyces naeslundii Actinomyces naeslundii Bacteria; Actinobacteria;
Actinobacteridae; Actinomycetales; Actinomycineae; Actinomycetaceae; Actinomyces.</source>
<reference>1  (bases 1 to 315) REFERENCE   1  (bases 1 to 315)   AUTHORS   Morou-Bermudez,E.
and Burne,R.A.   TITLE     Genetic and physiologic characterization of urease of Actinomyces
naeslundii   JOURNAL   Infect. Immun. 67 (2), 504-512 (1999)    PUBMED   9916052 REFERENCE
2  (bases 1 to 315)   AUTHORS   Morou-Bermudez,E. and Burne,R.A.   TITLE     Direct
Submission   JOURNAL   Submitted (17-FEB-1998) Center for Oral Biology, Univ. of Rochester,
601 Elmwood Ave, Rochester, NY 14642, USA</reference>
<features>Location/Qualifiers source1..315 /organism=Actinomyces naeslundii /mol_type=genomic
DNA /strain=WVU45 /db_xref=taxon:1655 gene  1..315 /gene=ureA CDS   1..315 /gene=ureA
/codon_start=1 /transl_table=11 /product=urease gamma-subunit /protein_id=AAD13723.1
/db_xref=GI:4249597 /translation=MHLTPREQEKLLIVVAADLARRRKDRGIRLNHPEAVAYITAEIL
EGAREGRTVTDLMAYGTTLLTYDDVMEGVPEMIRAVQVEATFPDGTKLVSVHDPIRRR LP</features>
<origin> 1 atgcacctca ccccgcgtga gcaggagaag ctactcatcg tcgtcgccgc agacctggca 61 cgaagacgca
aggacagagg gatccggctc aaccaccccg aggcagtcgc ctacatcacc 121 gctgagatcc tcgagggagc ccgagagggg
cgcacggtca cggatctcat ggcctacggg 181 accaccctgc tgacctacga cgacgtcatg gagggagtcc ccgagatgat
ccgcgcggtc 241 caggtggagg cgaccttccc cgacggaacc aagctcgtct ccgtccacga cccgattcgc 301
aggaggctgc catga</origin>
```

Figura 19 Exemplo de registro do Genbank formatado em XML.

Esse mapeamento de atributos para elementos XML, e de valor de atributo para conteúdo de elemento pode ser realizado pelo próprio encapsulador ou por um outro programa que seja executado em seguida.

Deve ser criado um documento XML para cada BDB, utilizando a estrutura proposta, contendo os registros retornados pelo encapsulador. O documento XML de um determinado BDB deve ser nomeado com o nome do BDB e a extensão *xml*. Por exemplo, o arquivo *Genbank.xml* deve conter os registros do BDB Genbank em formato XML.

Nesse arquivo, os registros devem ser delimitados pelas *tags* <*register*> e </*register*> e deve haver um elemento XML raiz, hierarquicamente superior ao elemento *register*, chamado de *root*. Assim, todos os registros devem estar hierarquicamente abaixo de <*root*>, conforme mostrado na Figura 20.

```
<root>
    <register>
        <!--registro 1-->
    </register>
    .
    .
    .
    <register>
        <!--registro n-->
    </register>
</root>
```

Figura 20 Estrutura dos registros no documento XML.

Utilizando as regras de formatação do documento XML mencionadas acima é possível criar um XML *schema* para os bancos de dados biológicos utilizados. Por exemplo, a Listagem 5 mostra o XML *schema* para o Genbank.

Listagem 5 XML *schema* referente aos dados do banco de dados Genbank.

```
<?xml version="1.0"?>
<xs:schema>
   <xs:element name="root">
      <xs:element name="register">
         <xs:complexType>
            <xs:sequence>
               <xs:element name="locus" type="xs:string"/>
               <xs:element name="definition" type="xs:string"/>
               <xs:element name="accession" type="xs:string"/>
               <xs:element name="version" type="xs:string"/>
               <xs:element name="keywords" type="xs:string"/>
               <xs:element name="source" type="xs:string"/>
               <xs:element name="reference" type="xs:string"/>
               <xs:element name="features" type="xs:string"/>
               <xs:element name="origin" type="xs:string"/>
            </xs:sequence>
         </xs:complexType>
      </xs:element>
   </xs:element>
</xs:schema>
```

A Listagem 6 mostra um exemplo de documento XML gerado contendo registros do BDB Genbank no XML *schema* definido pela Listagem 5.

Listagem 6 Arquivo Genbank.xml contendo dois registros.

```
<?xml version="1.0" encoding="UTF-8"?>
<root>
        <register>
                <locus>AF048778                 315 bp    DNA     linear   BCT 09-
                       FEB-1999</locus>
                <definition>Actinomyces naeslundii urease gamma-subunit (ureA)
                       gene, complete cds.</definition>
                <accession>AF048778  </accession>
                <version>AF048778.1  GI:4249596</version>
                <keywords>.</keywords>
```

```
                <source>Actinomyces naeslundii Actinomyces naeslundii Bacteria;
                        Actinobacteria; Actinobacteridae; Actinomycetales;
                        Actinomycineae; Actinomycetaceae; Actinomyces.</source>
                <reference>1  (bases 1 to 315) REFERENCE   1  (bases 1 to 315)
                        AUTHORS   Morou-Bermudez,E. and Burne,R.A.   TITLE
                        Genetic and physiologic characterization of urease of
                        Actinomyces naeslundii   JOURNAL   Infect. Immun. 67
                        (2), 504-512 (1999)    PUBMED   9916052 REFERENCE   2
                        (bases 1 to 315)   AUTHORS   Morou-Bermudez,E. and
                        Burne,R.A.   TITLE     Direct Submission   JOURNAL
                        Submitted (17-FEB-1998) Center for Oral Biology, Univ.
                        of Rochester, 601 Elmwood Ave, Rochester, NY 14642,
                        USA</reference>
                <features>Location/Qualifiers source1..315 /organism=Actinomyces
                        naeslundii /mol_type=genomic DNA /strain=WVU45
                        /db_xref=taxon:1655 gene  1..315 /gene=ureA CDS   1..315
                        /gene=ureA /codon_start=1 /transl_table=11 /product=urease
                        gamma-subunit /protein_id=AAD13723.1 /db_xref=GI:4249597
                        /translation=MHLTPREQEKLLIVVAADLARRRKDRGIRLNHPEAVAYITAEIL
                        EGAREGRTVTDLMAYGTTLLTYDDVMEGVPEMIRAVQVEATFPDGTKLVSVHDPIRRR
                        LP</features>
                <origin> 1 atgcacctca ccccgcgtga gcaggagaag ctactcatcg tcgtcgccgc
                        agacctggca 61 cgaagacgca aggacagagg gatccggctc aaccaccccg
                        aggcagtcgc ctacatcacc 121 gctgagatcc tcgagggagc ccgagagggg
                        cgcacggtca cggatctcat ggcctacggg 181 accaccctgc tgacctacga
                        cgacgtcatg gagggagtcc ccgagatgat ccgcgcggtc 241 caggtggagg
                        cgaccttccc cgacggaacc aagctcgtct ccgtccacga cccgattcgc 301
                        aggaggctgc catga</origin>
        </register>
        <register>
                <locus>AF048782                 492 bp    DNA     linear   BCT 09-
                        FEB-1999</locus>
                <definition>Actinomyces naeslundii urease accessory protein (ureE)
                        gene, complete cds.</definition>
                <accession>AF048782  </accession>
                <version>AF048782.1  GI:4249604</version>
                <keywords>.</keywords>
                <source>Actinomyces naeslundii Actinomyces naeslundii Bacteria;
                        Actinobacteria; Actinobacteridae; Actinomycetales;
                        Actinomycineae; Actinomycetaceae; Actinomyces.</source>
                <reference>1  (bases 1 to 492) REFERENCE   1  (bases 1 to 492)
                        AUTHORS   Morou-Bermudez,E. and Burne,R.A.   TITLE
                        Genetic and physiologic characterization of urease of
                        Actinomyces naeslundii   JOURNAL   Infect. Immun. 67 (2),
                        504-512 (1999)    PUBMED   9916052 REFERENCE   2  (bases 1
                        to 492)   AUTHORS   Morou-Bermudez,E. and Burne,R.A.
                        TITLE     Direct Submission   JOURNAL   Submitted (17-FEB-
                        1998) Center for Oral Biology, Univ. of Rochester, 601
                        Elmwood Ave, Rochester, NY 14642, USA</reference>
                <features>Location/Qualifiers source1..492 /organism=Actinomyces
                        naeslundii /mol_type=genomic DNA /strain=WVU45
                        /db_xref=taxon:1655 gene  1..492 /gene=ureE CDS   1..492
                        /gene=ureE /codon_start=1 /transl_table=11 /product=urease
                        accessory protein /protein_id=AAD13727.1
                        /db_xref=GI:4249605
                        /translation=MIIESISGNIHDLPGEDLEDVHVESVVLPLADLTKRIQRVRSDH
                        GTELGIRLAPGAPDLREGDILLRNERGIVVVRLEPTDVLVIAPVTVREMGVVAHNLGN
                        RHLPAQFFGPAEPFPGLEGHDGVMVIQYDHTAEHYLEHLGVRHARMERSMPVPFRHAE
                        HTH</features>
                <origin> 1 atgattatcg agtccatctc aggcaatatc cacgacctgc ccggcgagga
                        cctggaggac 61 gtgcacgttg agagcgtggt cctcccccctg gccgacctca
                        cgaaacgcat ccaaagagtg 121 cgctcggacc acggcaccga gctcggcatc
                        cggctcgcgc cgggcgcacc ggacctgcgc 181 gagggagaca ttctgcttcg
                        caacgaacgc ggaatcgtcg tcgtgcgcct ggagcccacc 241 gacgtcctgg
                        tcatcgcgcc cgtcacggtg cgcgagatgg gagtcgtcgc ccacaacctg 301
                        ggcaaccggc acctgcctgc ccagttcttc ggacccgctg agccccttccc
                        ggggcttgag 361 gggcatgacg gtgtcatggt catccagtac gaccacaccg
                        ccgagcacta cctggagcac 421 ctgggggtgc gtcacgcgcg gatggagcgt
```

```
                        tccatgcccg tccccttccg ccatgccgag 481 cacacccact ga</origin>
        </register>
</root>
```

O motivo de escolher o formato XML como armazenamento é em decorrência da diferença de modelagem existente nos diversos BDBs e também com o objetivo de facilitar a integração de dados, que ocorre nas etapas seguintes.

O XML fornece as vantagens de interoperabilidade, independência de aplicação, é extensível e auto-descritivo.

5.4 Etapa 2: Categorização dos Atributos

Para fazer a integração de dados de diferentes fontes, o primeiro passo é conhecer a estrutura das fontes, com o intuito de evitar a comparação de atributos que possuem escopo diferentes. Por exemplo, suponha que em um BDB1 tem-se um atributo *nome1*, que armazena o nome de um organismo; enquanto em um BDB2 tem-se um atributo *nome2* que armazena o nome de um gene. Apesar de um termo similar ser o nome de ambos os atributos, não podemos compará-los, pois seus escopos são diferentes.

Com o objetivo de evitar tais conflitos durante a integração de dados, realiza-se uma fase de integração de esquemas, por meio da classificação dos atributos em **Categorias** que indicam similaridade de escopo. Nessa fase, atributos de um mesmo escopo são classificados na mesma categoria. Por exemplo, na categoria *SeqNuc* encontram-se os atributos de todas as fontes de dados que armazenam sequências de nucleotídeos, independente do formato utilizado para seu armazenamento em cada BDB.

Conforme visto no exemplo acima cujos nomes de atributos são similares, comparar apenas o nome do atributo é insuficiente em alguns casos. Sendo assim, foram considerados os valores dos atributos para identificar o seu escopo e sua semântica, de uma forma automática.

Antes da integração de esquemas, o projetista ou especialista de domínio realiza uma análise detalhada dos bancos de dados que serão integrados. Essa análise compreende avaliar os valores de registros provenientes de tais bancos de

dados e identificar similaridades léxicas na formação dos dados: conjunto de valores (números ou caracteres utilizados) possíveis para um atributo, ordem de formação de identificadores utilizados, palavras-chave presentes em determinados escopos de atributos. Devido a essa análise detalhada, esse processo é complexo e demanda tempo, porém deve ser realizado com minúncia, pois é a base para a integração dos esquemas dos bancos de dados.

Como resultado da análise dos valores e escopo dos atributos, criou-se um conjunto de **Regras de Mapeamento das Categorias**. Trata-se de uma coleção de regras léxicas que são utilizadas por um analisador léxico (*parser*) para identificar em qual Categoria um determinado atributo se enquadra. Essas regras são apresentadas na Listagem 7.

Listagem 7 Regras de Mapeamento das Categorias.

```
Referência: AUTHOR|JOURNAL|PMID
Features: translation
Cogs: COGS
TopPDBHits: PDB
PfamSummary: E_value
Prodom: ProDom|protein domain
Descrição: ase|fimbrial|protein|factor|transporter
Ontologia: query=GO
Link: http
Fonte: eria
Locus: [A-Z]+[_|]*[0-9]+[ ]*[ |.[0-9]][0-9]*[ ][a-zA-z]+
Versão:[A-Z]+[_|]*[0-9]+[.][0-9]
ID: [A-Z]+[_|]*[0-9]+
NomeOrganismo: [A-Z][a-z]+[ ][a-z]+
LocalizaçãoCelular: Membrane|Cytoplasm|Extracellular
SeqProt:
[A|C|D|E|F|G|H|I|K|L|M|N|P|Q|R|S|T|V|W|Y][A|C|D|E|F|G|H|I|K|L|M|N|P|Q|R|S|T|V|W|Y]+
[$]
SeqNuc: Sequence|[A|T|C|G|a|t|c|g][A|T|C|G|a|t|c|g]+
ContagemBase: [A|C|T|G|a|c|t|g]
Número: [0-9]*
Ec: [0-9]+[.][-]*
NomeGene:[a-z][a-z][a-z][A-Z]
```

Suponha que deseja-se categorizar o atributo *reference* do Genbank, do primeiro registro da Listagem 6. Ao executar a primeira regra de mapeamento:

Referência: AUTHOR|JOURNAL|PMID

há uma correspondência, pois os termos AUTHOR e JOURNAL são encontrados no valor do atributo cujo elemento correspondente é o <*reference*> da Listagem 6.

Assim, as Regras de Mapeamento de Categorias são utilizadas para realizar a integração de esquemas dos bancos de dados envolvidos, categorizando cada atributo do banco de dados de acordo com a regra em que ele se enquadra. Devido

à dependência do escopo dos atributos envolvidos, essas regras são dependentes do domínio da aplicação.

As Regras de Mapeamento das Categorias devem ser executadas obrigatoriamente na ordem sequencial presente na Listagem 7, pois as primeiras regras da listagem são mais gerais e as últimas são mais específicas, ou seja, na ordem ascendente as regras vão se generalizando e na ordem descendente elas vão se restringindo.

Atributos cujos valores armazenados sejam de tipo numérico não são cobertos por tais regras. Elas valem apenas para atributos literais. Para atributos numéricos, a regra apenas classifica-os na Categoria Número.

Após serem classificados em categorias, os atributos são armazenados internamente no MIDB em uma estrutura de grafos, pois ela permite transitividade e eliminação de inconsistências. Por exemplo, sejam os seguintes BDBs e atributos:

BDB1, atributo accession1

BDB2, atributo ac

BDB3, atributo accession 2

BDB4, atributo id

Suponha que os atributos dos BDBs 1, 3 e 4 foram classificados como pertencentes à mesma Categoria. A representação interna deles é apresentada na Figura 21, na qual o grafo representa os atributos equivalentes entre as fontes de dados consideradas.

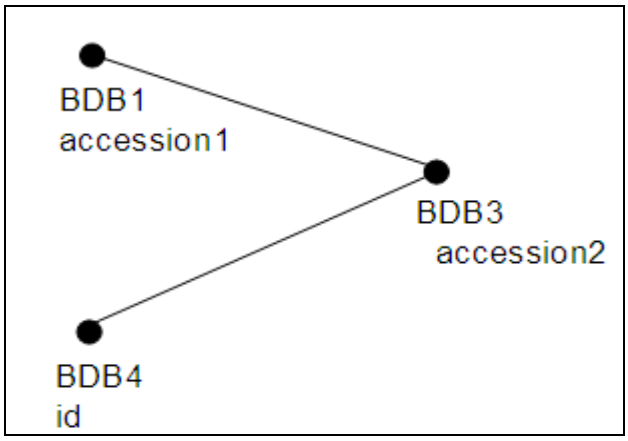

Figura 21 Exemplo do grafo da representação interna dos atributos pertencentes a uma determinada categoria.

Formalizando, seja um grafo G cujos nós <n_1,..., n_x> representam atributos de bancos de dados. Se há um caminho do nó n_a ao nó n_b, então os atributos representados por n_a e n_b são equivalentes. Isso assegura a transitividade entre os nós e a eliminação de inconsistências.

O especialista de domínio pode participar da etapa de Categorização dos Atributos, indicando quais atributos são da mesma Categoria. Para isso, o especialista deve fazer as identificações de atributos da mesma Categoria manualmente, antes da realização da categorização automática.

5.5 Etapa 3: Resolução de Entidades

Na etapa anterior tem-se como resultado os conjuntos de atributos similares, que foram classificados em cada uma das Categorias. Na etapa de Resolução de Entidades, o objetivo é identificar as instâncias de dados similares, ou seja, identificar quais registros de dados referem-se à mesma entidade do mundo real.

Para isso, necessita-se escolher os atributos que são chaves naturais dos BDBs, pois se o registro r_1 do BDB1 possui um atributo chave igual ao registro r_2 do BDB2, então r_1 e r_2 são registros que se referem à mesma entidade do mundo real. Essa é a forma mais simples de identificação de similaridade entre registros.

Assim, foram analisadas todas as Categorias criadas (que representam os principais atributos presentes nos BDBs que armazenam dados de sequências de nucleotídeos), com o intuito de identificar as possíveis chaves candidatas. Essas chaves candidatas podem ser formadas por apenas um atributo (apenas uma Categoria) ou um conjunto de atributos (um conjunto de Categorias), ou seja, uma chave composta.

Após uma análise detalhada dos atributos e de alguns registros, notou-se que, para os dados de sequência de nucleotídeos, existe uma chave candidata:

Categoria Descrição + Categoria SeqNuc

A utilização da chave composta (Categoria Descrição + Categoria SeqNuc) foi identificada como chave candidata. A Categoria Descrição armazena termos que descrevem o nome de um determinado gene. Alguns exemplos de descrição são:

"*Actinomyces naeslundii* hypothetical protein" e "*Actinomyces naeslundii* beta-glucosidase". Assim, é possível perceber que há mais chance de pesquisadores "batizarem" o mesmo gene com uma descrição similar.

Porém, considerar somente a descrição como atributo único na resolução de entidades não é o suficiente, pois pode-se ter dois registros com a mesma descrição, mas sequenciando diferentes partes de um gene. Assim, torna-se importante considerar também o atributo que representa o sequenciamento genético, e tal atributo é representado pela Categoria SeqNuc (categoria de sequência de nucleotídeos).

Ao comparar duas sequências por meio do algoritmo do BLAST (BLAST, 2010) e elas tiverem similaridade significativa (100% de similaridade e mesmo tamanho de sequência), elas tem muita probabilidade de se referirem ao mesmo gene. Assim, é possível que referem-se à mesma entidade do mundo real.

Assim, a chave candidata selecionada para realizar a resolução de entidades é a composição das categorias Descrição e SeqNuc. Essa composição permite que seja identificada similaridade nos registros de dados de sequências de nucleotídeos.

Como os dados das categorias Descrição e SeqNuc tem natureza distinta, é necessário que elas tenham algoritmos diferentes que façam a comparação de similaridade. Para tal comparação, no caso da Categoria SeqNuc, o BLAST é utilizado para gerar agrupamentos de similaridade. Já para a comparação de similaridade de dados da Categoria Descrição foi criado o Algoritmo de Clusterização.

Assim sendo, nesta etapa existem três processos: Clusterização por Descrição (onde é apresentado o Algoritmo de Clusterização criado por este trabalho), Clusterização por Sequência (onde o BLAST é utilizado para gerar clusters) e Emparelhamento de Clusters, como apresentado na Figura 22.

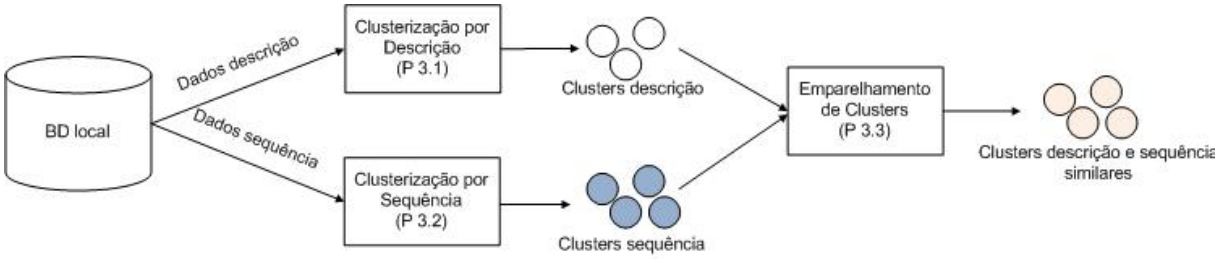

Figura 22 Processos da etapa de Resolução de Entidades.

O Algoritmo de Clusterização e os agrupamentos gerados pelo uso do BLAST podem ter resultado de similaridade diferente. Para evitar inconsistência, ao final da execução de ambos os algoritmos há um processo de emparelhamento, que verifica se os resultados foram iguais.

Os algoritmos citados, bem como o processo de emparelhamento são descritos nas subseções a seguir.

5.5.1 Processo 3.1: Clusterização por Descrição

No processo de Clusterização por Descrição, utiliza-se o Algoritmo de Clusterização. A partir da descrição dos registros (atributo da Categoria Descrição das fontes de dados), esse algoritmo gera um conjunto de agrupamentos (*clusters*), onde cada cluster contém elementos com alta similaridade entre si. Esse algoritmo tem uma fase de pré-processamento, onde é feita a preparação dos dados, e uma fase de processamento, onde os clusters são criados.

Na fase de **pré-processamento** utiliza-se um dicionário de sinônimos e abreviações. Esse dicionário faz substituições de sinônimos e abreviações de termos por suas formas canônicas. Além disso, ele possui uma lista de termos que podem ser ignorados, ou seja, são termos irrelevantes nas comparações, e que em alguns casos poderiam até afetar o resultado desejado. O dicionário substitui esses termos por vazio, removendo-os.

Por exemplo, a grafia do nome do organismo *Actinomyces naeslundii* escrita com gênero e espécie e em itálico e latim é a forma mais aceita, sendo assim considerada a sua forma canônica. Suponha que um determinado BDB escreve *A. naeslundii* para referir-se ao mesmo organismo. Para facilitar comparações de

similaridade, o dicionário de sinônimos e abreviações substituiria o termo *A. naeslundii* por sua forma canônica *Actinomyces naeslundii*.

Após a utilização do dicionário na fase de pré-processamento dos dados da Categoria Descrição, ocorre a fase seguinte, onde a clusterização propriamente dita é realizada.

Na fase de **processamento da clusterização** acontece a geração dos clusters. Nessa fase, armazena-se o tamanho dos valores dos dados da Categoria Descrição de todos os registros e ordena-os em ordem crescente. O registro com menor tamanho será o registro base para as comparações e formação do cluster. Esse registro é comparado com uma lista de registros que ainda não foram agrupados.

Nessas comparações precisa-se usar um critério ou métrica de similaridade, para poder medir o quão similar os registros são, por meio de um resultado numérico. São utilizadas duas métricas de similaridade (ou distâncias de edição): Distância de Edição de Caracteres e a Distância de Edição de Palavras.

O Algoritmo 1 calcula a Distância de Edição de Caracteres (DEC), comparando a igualdade dos caracteres entre a descrição de dois registros e retornando a razão entre o número de caracteres iguais e o tamanho do campo descrição do menor registro. A DEC retorna a porcentagem da similaridade de duas *strings* utilizando a Distância de Hamming como base e comparando com o menor registro.

Algoritmo 1 Procedimento que calcula a Distância de Edição de Caracteres.

```
1  iguais <- 0
2  tam1 <- tamanho(descricao1)
3  tam2 <- tamanho(descricao2)
4  se (tam1 < tam2) então
5     para (i <- 0; i < tam1; i <- i+1) faça
6        se (descricao1[i] = descricao2[i]) então
7           iguais <- iguais + 1
8        fim-se
9     fim-para
10    retorna (iguais / tam1)
11 senão
12    para (i <- 0; i < tam2; i <- i+1) faça
13       se (descricao1[i] = descricao2[i]) então
14          iguais <- iguais + 1
15       fim-se
16    fim-para
17    retorna (iguais / tam2)
18 fim-se
```

Por exemplo, suponha que a *descricao1* seja "*Actinomyces naeslundii* ribosomal protein S7" e a *descricao2* seja "*Actinomyces naeslundii* ribosomal protein L11". Calculando o número de caracteres iguais entre ambas as descrições, utilizando a Distância de Edição de Caracteres tem-se o valor 41/43 ≈ 0,95.

Em uma outra situação, suponha que a *descricao1* seja "*Actinomyces naeslundii* partial DNA gyrase subunit A" e que a *descricao2* seja "*Actinomyces naeslundii* strain E1-20 DNA gyrase subunit A". Calculando a taxa de similaridade apenas por meio da semelhança e diferença de caracteres proposta pela Distância de Edição de Caracteres, obtém-se o valor 25/51 ≈ 0,49. Porém, percebe-se que existem muitas palavras em comum entre tais descrições e que a taxa de similaridade seria mais alta se as palavras fossem consideradas ao invés dos caracteres, neste caso.

Como solução para isso, resolveu-se utilizar também a Distância de Edição de Palavras (DEP). Ela busca quantas palavras são iguais entre dois registros e retorna a razão entre o número de caracteres iguais das palavras e o tamanho do campo descrição do menor registro. Tais instruções podem ser observadas no Algoritmo 2, que é um procedimento que tem como objetivo fazer o cálculo da Distância de Edição de Palavras. As variáveis *Descricao1[]* e *Descricao2[]* são vetores que armazenam uma palavra em cada índice. Nas linhas 6 e 13, há chamadas ao procedimento *busca_palavra*. Esse procedimento busca uma palavra em um vetor e retorna *true* caso a palavra seja encontrada, e *false* caso contrário.

Algoritmo 2 Procedimento que calcula a Distância de Edição de Palavras.

```
1  iguais <- 0
2  tam1 <- tamanho(Descricao1[])
3  tam2 <- tamanho(Descricao2[])
4  se (tam1 < tam2) então
5     para (i <- 0; i < tam1; i <- i+1) faça
6        se (busca_palavra(Descricao1[i], Descricao2[] = true) então
7           iguais <- iguais + tam(Descricao1[i]) + 1 //espaco em branco
8        fim-se
9     fim-para
10    retorna ((iguais - 1) / tam1)
11 senão
12    para (i <- 0; i < tam2; i <- i+1) faça
13       se (busca_palavra(Descricao2[i], Descricao1[] = true) então
14          iguais <- iguais + tam(Descricao2[i]) + 1
15       fim-se
16    fim-para
17    retorna ((iguais - 1) / tam2)
18 fim-se
```

Retomando a situação anterior, em que a *descricao1* é "*Actinomyces naeslundii* partial DNA gyrase subunit A" e que a *descricao2* "*Actinomyces naeslundii* strain E1-20 DNA gyrase subunit A", o novo resultado de similaridade calculado por meio do Algoritmo 2 (DEP) é 43/51 ≈ 0,84, uma diferença significativa em relação ao uso da Distância de Edição de Caracteres.

Para clusterizar os registros são utilizadas ambas as métricas (Distância de Edição de Caracteres e Distância de Edição de Palavras). O valor a ser considerado nas comparações será o primeiro resultado dessas distâncias a atingir o valor do limiar de comparação.

Conforme mencionado anteriormente, o registro base (registro cuja descrição é a menor) é comparado com os demais registros que ainda não foram clusterizados. Caso o resultado de uma das distâncias de edição seja maior ou igual a 0,83 (limiar de comparação), o i-ésimo registro é considerado similar ao registro base. Isso significa que o i-ésimo registro deve pertencer ao mesmo cluster que o registro base. Assim sendo, ele sai da lista de registros não clusterizados e entra no cluster no qual o registro-base está contido.

Ao final de cada iteração na lista de registros, tem-se um cluster sendo gerado, com o registro base do cluster sendo a base para todas as comparações. Se ainda existirem registros que não foram clusterizados, o processo inicia novamente, escolhendo-se um novo registro base (que tenha o menor tamanho do valor do atributo da Categoria Descrição), e executando sucessivamente as instruções citadas.

O Algoritmo 3 sintetiza os passos descritos nesta seção. Note que ele remove as *stopwords* (lista de caracteres que são ignorados no pré-processamento, por exemplo parênteses e hífen entre dois caracteres não numéricos) na linha 2, utiliza o dicionário de sinônimos e abreviações (linha 3) e as distâncias de edição de caracteres e palavras (linhas 15 e 19, respectivamente). O ciclo determinado pela palavra-chave *enquanto* só é finalizado quando todos os registros forem clusterizados. O *registro_base* é a base para todas as comparações e, também, base para a criação do cluster.

Algoritmo 3 Algoritmo de clusterização.

```
1  //Fase de pré-processamento
2  Remoção de stopwords
3  Consulta dicionário de sinônimos e abreviações
4  //Fase de processamento
```

```
5  Calcula tamanho dos registros
6  nroCluster <- 0
7  //limiar obtido experimentalmente com base em estudo de dados biológicos
8  limiar <- 0.83
9  enquanto há registros a serem clusterizados faça
10    registro_base <- registro com menor tamanho
11    remove(registro[], registro_base)
12    nroRegistros <- nroRegistros - 1
13    cluster[nro_cluster] <- registro_base
14    para (i <- 0; i < nroRegistros; i <- i+1) faça
15       se distancia_edicao_caracteres(registro_base, registro[i]) >=
   limiar entao
16          adiciona(cluster[nro_cluster], registro[i])
17          remove(registro[], registro[i])
18       senao
19           se distancia_edicao_palavras(registro_base, registro_i) >=
   limiar entao
20          adiciona(cluster[nro_cluster], registro[i])
21          remove(registro[], registro[i])
22       fim-se
23    fim-para
24    nroCluster <- nroCluster + 1
25 fim-enquanto
```

Note que o limiar de 0,83 (presente nas comparações das linhas 15 e 19) foi obtido experimentalmente com base em um conjunto de registros de dados biológicos. Os trinta registros que fazem parte desse conjunto foram selecionados cautelosamente considerando vários casos de teste que poderiam ocorrer: pares de registros que possuem descrição similar e devem estar no mesmo cluster, pares de registros que possuem descrição diferente e devem estar em clusters distintos, e pares de registros com descrições parcialmente similares. Um especialista organizou manualmente os registros em clusters de acordo com a similaridade das descrições. O limiar de comparação do algoritmo foi ajustado para que os resultados fossem iguais à clusterização realizada pelo especialista.

Por exemplo, suponha os seguintes registros, que fazem parte do conjunto de teste utilizado como base para identificar o valor do limiar de comparação:

1. *Actinomyces naeslundii* strain E1-20 DNA gyrase subunit A (gyrA) gene, partial cds.
2. *Actinomyces naeslundii* strain R24330 DNA-directed RNA polymerase beta-subunit (rpoB) gene, partial cds.
3. *Actinomyces naeslundii* strain F6E1 DNA gyrase subunit A (gyrA) gene, partial cds.
4. *Actinomyces naeslundii* beta glucosidase.

Observa-se que os registros 1 e 2 possuem descrição parcialmente similar, os registros 1 e 3 possuem descrição similares e os registros 1 e 4 possuem descrições diferentes.

Comparando-se o registro 1 em relação aos demais registros utilizando as medidas de Distância de Edição de Caracteres (DEC) e Distância de Edição de Palavras (DEP), tem-se os seguintes resultados:

DEC(1,2) = 0,39

DEP(1,2) = 0,64

DEC(1,3) = 0,38

DEP(1,3) = 0,83

DEC(1,4) = 0,61

DEP(1,4) = 0,56

Considerando as descrições dos registros 1 e 3 e os valores obtidos das distâncias de edição utilizadas, observa-se que apenas esse par de registros são similares, pois é maior ou igual ao valor do limiar de comparação obtido experimentalmente.

5.5.2 Processo 3.2: Clusterização por Sequência

A Clusterização por Sequência utiliza o BLAST, um algoritmo que compara sequências biológicas de nucleotídeos (provenientes do DNA) e de aminoácidos (provenientes de proteínas) por meio da técnica de alinhamento local de sequências.

Neste processo, o objeto de comparação é as sequências de nucleotídeos. Uma das vantagens de utilizar o BLAST é poder buscar uma sequência de entrada em um repositório de dados de sequência, onde são retornadas as sequências que possuem similaridade com a entrada.

A Figura 23 mostra os passos realizados no processo de Clusterização por Sequência. É realizada uma busca para cada sequência do repositório BLAST. O resultado dessas buscas é analisado lexicamente, gerando clusters, onde cada cluster contém registros identificados como similares pelo BLAST.

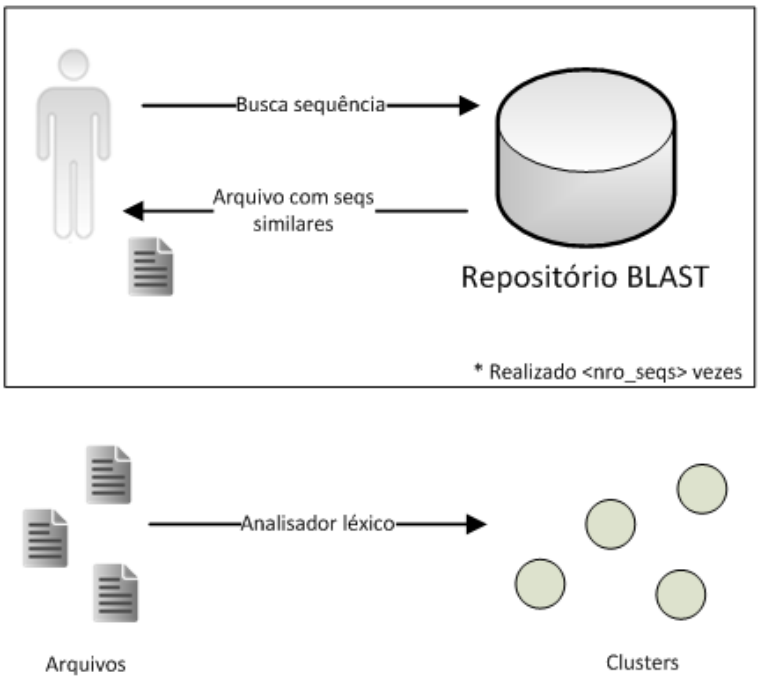

Figura 23 Sequência de passos realizados no processo de Clusterização por Sequência.

Primeiro, é necessário popular o BLAST com os registros das fontes de dados. Depois disso, para buscar os registros similares, é necessário criar arquivos de consulta, que são arquivos do formato FASTA (arquivo de formato textual contendo dados de sequência de nucleotídeos ou aminoácidos) que contém apenas o registro base que está sendo buscado.

Em seguida, executam-se consultas usando como entrada cada um dos arquivos criados anteriormente. O comando utilizado para criar as consultas foi utilizando os parâmetros de *expect* (parâmetro que indica acaso) igual a zero, *identities* (similaridade de alinhamento local) igual a 100% e sequências de tamanho igual. Para cada consulta, o BLAST cria um arquivo de saída que contém os registros que possuem similaridade considerável, além do registro base da comparação.

Com o auxílio de um analisador léxico, é feita uma varredura nos arquivos de saída identificando os registros cuja similaridade é 100% e cujos tamanhos são iguais, criando clusters que abranjam esses registros.

5.5.3 Processo 3.3: Emparelhamento de Clusters

Durante o Processo 3.1, os clusters de descrição foram gerados a partir do Algoritmo de Clusterização, e no Processo 3.2 obtiveram-se os clusters de sequência, gerado a partir da execução do algoritmo BLAST e análise léxica do seu resultado. Sendo assim, nesse momento existem dois conjuntos de clusters: um conjunto de clusters descrição e outro conjunto de clusters sequência.

Neste processo de emparelhamento, os conjuntos de clusters são comparados (bem como seus elementos), para identificar se há similiaridade 100%, ou seja, todos os elementos do cluster descrição são iguais a todos os elementos do cluster sequência.

Para isso, o algoritmo de emparelhamento utiliza como entrada os dois conjuntos de clusters. Para cada cluster de descrição, há uma comparação com todos os clusters de sequência. Isso ocorre na linha 5 do Algoritmo 4. Caso a comparação seja de 100% (valor 1), o cluster sequência é removido do conjunto de clusters sequência e há o início de um novo ciclo com um novo cluster descrição. Quando a comparação é de 100%, o cluster é chamado de *identidade*.

Algoritmo 4 Verifica emparelhamento de clusters.

```
1  para cada cluster descricao //indice = i
2     cd = cluster_desc[i]
3     para cada cluster sequencia // indice = j
4        cs = cluster_seq[j]
5        se (compara_clusters(cluster_desc[i], cluster_seq[j]) = 1) então
6           escreva("clusters similares: "cluster_desc[i], cluster_seq[j])
7           remove(cluster_seq, cluster_seq[j])
8           j <- nro clusters seq // para finalizar o loop e ir p/ prox
9        fim-se
10    fim-para
11 fim-para
```

O procedimento que compara um cluster de descrição com um cluster de sequência verifica o menor cluster (em relação à quantidade de elementos) e inicia com o primeiro elemento de tal cluster, buscando-o no outro cluster. Existe um contador que vai monitorando quantos elementos são diferentes entre ambos os clusters.

A fórmula criada para calcular o valor da porcentagem de similaridade entre os dois clusters é:

$$\text{Porcentagem de similaridade } (c1, c2) = \left| \frac{\text{elementosDif} - \text{tam(maior(c1, c2))}}{\text{tam(maior(c1, c2))}} \right|,$$

Onde *c1* e *c2* são os clusters de entrada, *elementosDif* é o número de elementos diferentes entre ambos os clusters e *tam(maior(c1, c2))* calcula o tamanho do maior cluster entre c1 e c2. Note que o numerador da fórmula calcula o módulo resultante da subtração dos elementos diferentes pelo tamanho do maior cluster.

O resultado retornado por essa fórmula é um valor numérico entre 0 e 1, onde 0 indica que não há similaridade alguma e 1 indica 100% de similaridade.

Se o valor retornado for 1, significa que o cluster gerado pelo Algoritmo de Clusterização da descrição e o cluster gerado pela Clusterização por Sequência estão iguais. Nesse caso, foi possível realizar uma correspondência automática onde nota-se acerto das duas partes. Os clusters desse cenário são armazenados integrados no banco de dados local.

Em alguns casos o valor retornado não é 1. Caso o valor da fórmula seja 0, significa que nenhum elemento é igual, então não existe nenhuma ação que possa ser tomada. Porém, se o valor retornado pela função estiver entre 0 e 1 (não incluindo esses valores), significa que pelo menos um elemento é igual entre os dois clusters.

5.6 Etapa 4: Gerenciamento de clusters

Suponha que no processo de emparelhamento de clusters descrição e clusters sequência, o valor retornado é um valor numérico decimal entre 0 e 1 (mas não incluindo esses valores). Nessa situação, qual seria a melhor alternativa? Deve-se considerar o cluster gerado pela Clusterização por Descrição, utilizando o Algoritmo de Clusterização, ou o cluster criado pela Clusterização por Sequência, usando o BLAST?

Esses tipos de conflitos são grandes desafios da integração de dados. Existem algumas técnicas de como resolvê-los:

a) Armazenar todos os resultados obtidos e adicionar anotações nos dados. O benefício é que nenhum dado será perdido, porém a tomada de decisão para resolução do conflito é protelada.

b) Fundir parcialmente os dados em comum, eliminando os fatos que julgam-se incorretos.

c) Contar com o auxílio do usuário (especialista de domínio) para tomar decisões de resolução de conflitos.

No presente trabalho, as opções *b* e *c* são utilizadas para resolver esse problema.

Assim, no caso anterior em que o cluster descrição e o cluster sequência não possuem similaridade 100%, mas 0 < *similaridade* < 1, cria-se um novo cluster, contendo apenas os elementos iguais entre o cluster sequência e o cluster descrição. Formalmente, seja C_s o cluster sequência e C_d o cluster descrição. Os elementos a serem fundidos são $C_s \cap C_d$.

Os demais elementos que restam no cluster são:

- Elementos restantes no cluster descrição: $C_d – (C_s \cap C_d)$
- Elementos restantes no cluster sequência: $C_s – (C_s \cap C_d)$

Assim, considerando os dois clusters, é necessário, ainda, tomar uma decisão em relação aos elementos de $C_d – (C_s \cap C_d)$ e $C_s – (C_s \cap C_d)$.

Nesse caso, o especialista de domínio é consultado a fim de tomar uma decisão sobre os elementos remanescentes. Caso ele queira incluir algum elemento dos conjuntos $C_d – (C_s \cap C_d)$ e/ou $C_s – (C_s \cap C_d)$ no cluster criado, é feita essa adição. Caso contrário, se o especialista optar por não adicionar nenhum elemento, o resultado será composto pelos clusters $C_s \cap C_d$, $C_d – (C_s \cap C_d)$ e $C_s – (C_s \cap C_d)$.

Dessa forma, o especialista de domínio realiza o Gerenciamento dos Clusters, conforme representado pelas instruções do Algoritmo 5. Note que esse algoritmo é o Algoritmo de Emparelhamento de Clusters (Algoritmo 4) com a opção de fazer gerenciamento de clusters que não são identidade.

Os clusters gerados por essa fase são armazenados no banco de dados local.

Algoritmo 5 Algoritmo de emparelhamento com resolução de conflitos.

```
1   para cada cluster descricao //indice = i
2      cd = cluster_desc[i]
```

```
3      para cada cluster sequencia // indice = j
4         cs = cluster_seq[j]
5         se (compara_clusters(cluster_desc[i], cluster_seq[j]) = 1) então
6            escreva("clusters similares: "cluster_desc[i], cluster_seq[j])
7            remove(cluster_seq, cluster_seq[j])
8            j <- nro clusters seq // para finalizar o loop e ir p/ prox
9         senão
10           se (compara_clusters(cluster_desc[i], cluster_seq[j]) > 0)então
11              Gerenciamento_Clusters()
12           fim-se
13        fim-se
14     fim-para
15  fim-para
```

5.7 Considerações Finais

Neste capítulo foi descrito o modelo utilizado para integração de fontes de dados biológicos para dados de sequência, denominado MIDB. Esse modelo é composto por quatro etapas: Carregamento de Dados (Etapa 1); Categorização dos Atributos (Etapa 2); Resolução de Entidades (Etapa 3); e Gerenciamento de Clusters (Etapa 4).

Na Etapa 1 ocorre o carregamento de dados, os quais são provenientes de fontes de dados biológicos do domínio de sequência de nucleotídeos. Um encapsulador obtém os dados de uma determinada fonte e transforma-os em um documento XML, que pode ser manipulado com mais facilidade nas etapas posteriores. Em seguida, na Etapa 2, os documentos XML são analisados lexicamente e seus atributos são classificados em categorias, ocorrendo a integração dos esquemas dos BDBs fonte. A Etapa 3 é essencial, pois nela ocorre a integração das instâncias dos dados. Os atributos das categorias Descrição e SeqNuc que forem similares são agrupados em clusters, que são comparados. Se um determinado cluster descrição e um dado cluster sequência tiverem correspondência para todos os seus elementos, eles são nomeados de cluster identidade e armazenados. Nesse caso, a resolução de conflitos foi realizada automaticamente pelos algoritmos de clusterização. Caso contrário, segue-se para a Etapa 4, onde o especialista de domínio toma decisão em relação a clusters não-identidade.

No próximo capítulo é apresentado um estudo de caso realizado a partir do uso do modelo MIDB proposto.

Capítulo 6

ESTUDO DE CASO UTILIZANDO O MODELO MIDB

Neste capítulo são realizados estudos de casos baseados no uso do Modelo de Integração de Dados Biológicos (MIDB), visando avaliação da metodologia proposta. Esses estudos de caso são apresentados e discutidos neste capítulo.

6.1 Considerações iniciais

Neste capítulo são apresentados os resultados de dois estudos de caso relativos a duas etapas do modelo MIDB, a saber: etapa de Categorização dos Atributos (Etapa 2) e etapa de Resolução de Entidades (Etapa 3). Os experimentos realizados neste capítulo utilizaram dados biológicos do domínio de sequências de nucleotídeos e tiveram a finalidade de avaliar a Etapa 2 e a Etapa 3 separadamente. Essas etapas contribuem para a integração de dados biológicos de sequências de nucleotídeos, que é o principal objetivo do modelo proposto nesta dissertação.

Para os estudos de caso, o modelo MIDB foi aplicado a um cenário com quatro fontes de dados: DDBJ, EMBL, Genbank e Oralgen. Todos esses BDBs possuem dados de sequência de nucleotídeos, e, mais especificamente, consultaram-se sequências referentes a organismos do gênero *Actinomyces*, como *Actinomyces naeslundii*, *Actinomyces oris* e *Actinomyces johnsonni*, que são bactérias presentes na cavidade oral dos seres humanos.

Encapsuladores foram criados para fazer a extração dos dados de sequência das fontes de dados e transformá-los em arquivos do formato XML, os quais foram

armazenados num banco de dados local (Etapa 1). Em seguida ocorre a etapa de Categorização dos Atributos (Etapa 2).

6.2 Estudo de caso 1: Etapa de Categorização dos Atributos

Na etapa de Categorização dos Atributos, os atributos das fontes de dados são classificados nas Categorias criadas. Essa etapa pode ser realizada com o auxílio de um especialista de domínio, mas neste estudo de caso a intenção é que ela seja realizada automaticamente (sem intervenção humana), para análise dos resultados.

A Categorização dos Atributos foi implementada com a tecnologia Java JSE 6.24 e o ambiente de desenvolvimento Netbeans IDE 6.9.1.

As fontes de dados DDBJ, EMBL e Genbank possuem 9 atributos cada, enquanto o Oralgen possui 26 atributos, totalizando 53 atributos. Desses atributos, 5 não possuem valor (estão vazios ou NULL), e assim não foi possível serem categorizados.

Desconsiderando os atributos sem valor, os demais foram categorizados corretamente, exceto um único atributo que foi classificado na Categoria incorreta. Refere-se ao atributo *AC* do EMBL, que foi categorizado como *Versão* ao invés de *ID*. Dessa forma, a taxa de acerto, desconsiderando os atributos sem valor, foi de aproximadamente 97,91% (i.e. 47/48).

A Listagem 8 mostra o trecho de um registro no arquivo embl.xml. Nela pode-se observar o formato do valor do elemento *ac*: DQ658412.1. Ele é formado por duas letras, uma sequência de números, um ponto, seguido de um número.

Listagem 8 Trecho de um registro do arquivo embl.xml.

```
<root>
   <register>
      <id>DQ658412.1 SV 1 linear genomic DNA STD PRO 1435 BP.</id>
      <ac>DQ658412.1</ac>
      <de>Actinomyces naeslundii partial methyltransferase</de>
      <os>Actinomyces naeslundii</os>
      <oc>Bacteria Actinobacteria Actinobacteridae Actinomycetales Actinomycineae
          Actinomycetaceae Actinomyces.</oc>
      …
   </register>
</root>
```

Executando as Regras de Mapeamento das Categorias na ordem em que estão representadas na Listagem 7, a primeira regra que faz correspondência ao valor do elemento *ac* é a regra Versão. Tal regra expressa que um atributo da Categoria Versão deve ter um ou mais caracteres maiúsculos, seguido ou não pelo caracter *underline* (“_”) , mais uma sequência de números, seguido por um ponto e um número.

De fato, o atributo *ac* poderia ser considerado um atributo da Categoria Versão, pois além de seguir a regra de formação, possui semântica de versão. Por outro lado, se ele não tivesse o ponto seguido de número, poderia perfeitamente ser considerado um atributo da Categoria ID. Olhando o registro de uma forma mais abrangente, observa-se que o atributo que seria sua chave primária (atributo ou conjunto de atributos único que distingue um registro dos demais) é o atributo *ac*. Dessa forma, ele foi considerado como um atributo que deveria ser da Categoria ID.

Considerando as categorias que serão utilizadas na Etapa 3, que são a Categoria Descrição e Categoria SeqNuc, o acerto é de 100% dos atributos, o que demonstra que os dados que serão utilizados pela próxima etapa são confiáveis e não irão causar erros em etapas e processos posteriores.

Além disso, o sistema criado para realizar a Categorização de Atributos conseguiu superar a restrição que existia para atributos numéricos. Por meio do uso do nome do atributo em questão e de uma lista de termos sinônimos a esse nome foi possível identificar e categorizar corretamente todos os atributos numéricos.

6.3 Estudo de caso 2: Etapa de Resolução de Entidades

Na etapa de Resolução de Entidades, os dados da Categoria Descrição e da Categoria SeqNuc são clusterizados, sendo que a estratégia de clusterização depende da Categoria do dado. Em seguida esses dados são emparelhados.

No Processo de Clusterização por Descrição (Processo 3.1), foi criado um analisador léxico para extrair os dados da Categoria Descrição que foi identificada na etapa anterior. Como os dados da Categoria Descrição do BDB Oralgen não

possuem o nome do organismo a ser comparado, isso foi adicionado a todos os registros dessa fonte de dados. Além disso, a pontuação entre caracteres não numéricos foi removida para todos os BDBs de entrada. A pontuação entre caracteres numéricos não foi removida porque ela faz a diferença na descrição de um dado de sequência, por exemplo no termo sublinhado em "*Actinomyces naeslundii* strain MMRC12-1 DNA-directed RNA polymerase beta subunit (rpoB) gene, partial cds".

O Algoritmo 1, Algoritmo 2 e Algoritmo 3 do Processo de Clusterização por Descrição foram implementados utilizando a linguagem de programação Perl.

Os 605 registros provenientes das fontes de dados foram agrupados em 83 clusters, de acordo com o Algoritmo de Clusterização que utiliza os dados advindos da Categoria Descrição. As métricas utilizadas para decidir em que cluster um determinado registro será classificado foram a Distância de Edição de Caracteres e a Distância de Edição de Palavras.

A Listagem 9 mostra o exemplo de um cluster gerado pelo Algoritmo de Clusterização. O formato nela apresentado é de um contador interno (utilizado para facilitar as manipulações dos dados, pelo fato de ser único) seguido da descrição da sequência de nucleotídeos (Categoria Descrição). Observe que o registro base é o registro 475 e os demais registros possuem similaridade maior ou igual a 0,83 quando aplicados às distâncias de edição de palavras ou caracteres.

Listagem 9 Exemplo de um cluster gerado pelo Algoritmo de Clusterização.

```
475 Actinomyces naeslundii 16S ribosomal RNA gene partial sequence
515 Actinomyces naeslundii strain TG6 16S ribosomal RNA gene partial sequence
518 Actinomyces naeslundii strain TeJ7 16S ribosomal RNA gene partial sequence
476 Actinomyces naeslundii strain GiTB 16S ribosomal RNA gene partial sequence
516 Actinomyces naeslundii strain GRG14 16S ribosomal RNA gene partial sequence
517 Actinomyces naeslundii strain GumJ6B 16S ribosomal RNA gene partial sequence
```

No Processo 3.2, Clusterização por Sequência, foi necessário criar um arquivo FASTA contendo todos os registros provenientes das fontes de dados. Esse arquivo tem o formato mostrado pela Listagem 10.

Listagem 10 Exemplo de arquivo no formato FASTA.

```
>gi|1| EU620999  | DDBJ| Actinomyces oris strain P5N citrate synthase I (gltA)
gene, partial cds.
ggcctgccgctgctctaccccgacccgcagcgctcctacgtcgaggacttcatccgcctgaccttcgggatgccctaccagtc
ctacgacatcgacccggccgtggtgcgcgccctggacatgctcctcatcctgcacgccgaccacgagcagaactgctcgacct
ccacggtgcgcctcgtgggctcggccgacgccaacatgtacgcctccgtggccgcgggtgtgggcgccctgtccgggccgctg
cacggcggcgccaacgaggccgtcctgcggatgctggacacgatccagagctcgggaatgagcacggccgagttcgtccgcaa
ggtcaaggacaaggaggacggcgtgcggctcatgggcttcggccaccgggtctacaagaactacgacccgcgcgccgccatcg
tcaaggagaccgcccacgacgtcctgacccgcctgggctccgatgacggcgaccgcaagctcgagatcgccatggagctcgag
gagacggcgctgcgcgacgagtacttcgtctcgcgcagcctctacccgaacgtc
>gi|2| EU621000  | DDBJ| Actinomyces oris strain CCUG 33920 citrate synthase I
(gltA) gene, partial cds.
Ggtctgccgctgctctaccccgacccgcagcgctcctacgtcgaggacttcatccgcctgaccttcgggatgccctaccagtc
ctacgacatcgacccggccgtggtgcgggccctggacatgctgctcatcctgcatgccgaccacgagcagaactgctcgacct
ccacggtgcgcctcgtgggctcggccgacgccaacatgtacgcctccgtggccgcgggcgtgggcgccctgtccgggccgctg
cacggcggcgccaacgaggcggtcctgcggatgctggacacgatccagagctcggggatgagcacggccgagttcgtccgcaa
ggtcaaggacaaggaggacggcgtccggctcatgggcttcggccaccgggtctacaagaactacgacccgcgcgccgcgatcg
tcaaggagaccgcccacgacgtcctgacccgcctggggtccgacgacggcgaccgcaagctcgagatcgccatggagctggag
gagacggcgctgcgcgacgagtacttcgtctcccgcagcctctacccgaatgtc
```

Após criar o arquivo FASTA, utiliza-se esse mesmo arquivo para fazer a criação de uma base de dados no BLAST. Essa base será consultada por um arquivo de entrada contendo qual o registro que se deseja consultar, bem como sua sequência.

Como o objetivo é fazer uma consulta de todos os registros com todos os registros, foi necessário criar um arquivo de consulta para cada registro da base de dados. Em seguida, um comando BLAST realiza a consulta na base de dados a partir do arquivo de consulta e gera um arquivo de saída contendo todos os registros cuja sequência de nucleotídeos possui similaridade (tamanho iguais das sequências, *expect* igual a zero e *identities* igual a 100%) com o registro consultado.

Essa atividade de emitir o comando de consulta para cada registro foi realizada por um arquivo de processamento em lote criado para esse fim.

Em seguida, os arquivos de saída foram visitados um a um para identificar em cada arquivo quais eram os registros que possuem sequência similar, que estão na seção de “alinhamento significativo” do arquivo de saída da consulta BLAST. Um exemplo de arquivo de saída pode ser encontrado no APÊNDICE A – SAÍDA DE ARQUIVO BLAST.

Foi criado um agrupamento para cada registro base e seus similares, totalizando 112 clusters. Os clusters gerados por esse processo possuem o formato semelhante aos gerados pelo processo anterior, sendo que a diferença é o processo de geração dos clusters em si.

No Processo 3.3, ocorre o Emparelhamento de clusters descrição e clusters sequência. O conjunto de clusters gerados por ambos os processos foram comparados e analisados. No resultado, foram analisados os clusters identidade e os clusters de descrição e sequência que tiveram pelo menos um elemento em comum.

Neste estudo de caso, 41 clusters são clusters identidade. Ao comparar esse número com o número de clusters descrição (83 clusters), tem-se que aproximadamente 49,39% dos clusters são idênticos. Ou seja, quase metade dos clusters gerados pelo Algoritmo de Clusterização tiveram elementos iguais aos clusters que usaram o BLAST como processo de geração. Um exemplo de cluster idêntico é o cluster da Listagem 9.

Nos demais casos de clusters não enquadrados em cluster identidade, estão os clusters descrição e sequência que possuem pelo menos um elemento em comum.

Por exemplo tem-se o cluster sequência número 167, ou CS167; e os clusters descrição CD530 e CD589, apresentados, respectivamente, na Listagem 11, Listagem 12 e Listagem 13. Uma listagem mais completa dos elementos desses clusters pode ser encontrada no APÊNDICE B – ELEMENTOS DOS CLUSTERS CS167, CD530 E CD589.

Listagem 11 Elementos do cluster sequência CS167 e suas descrições.

```
CS167 = [305, 304, 271, 267, 168, 270]

305 Actinomyces naeslundii hypothetical protein
304 Actinomyces naeslundii hypothetical protein
271 Actinomyces naeslundii hypothetical protein
267 Actinomyces naeslundii hypothetical protein
168 Actinomyces naeslundii beta-glucosidase gene, complete cds
270 Actinomyces naeslundii beta-glucosidase
```

Listagem 12 Elementos do cluster descrição CD530 e suas descrições.

```
CD530 = [271, 272, 269, 268, 304, 267, 305]

271 Actinomyces naeslundii hypothetical protein
272 Actinomyces naeslundii hypothetical protein
269 Actinomyces naeslundii hypothetical protein
268 Actinomyces naeslundii hypothetical protein
304 Actinomyces naeslundii hypothetical protein
267 Actinomyces naeslundii hypothetical protein
305 Actinomyces naeslundii hypothetical protein
```

Listagem 13 Elementos do cluster descrição CD589 e suas descrições.

```
CD589 = [270, 168]

270 Actinomyces naeslundii beta-glucosidase
168 Actinomyces naeslundii beta-glucosidase gene, complete cds
```

O cluster sequência CS167 teve 57,14% de similaridade com o cluster CD530 e 33,33% de similaridade com o cluster CD589. Lembrando que o cálculo de similaridade de dois clusters faz a comparação dos elementos diferentes em relação ao maior dos dois clusters, por isso esses números não somam o valor 100%.

Os registros 305, 304, 271, 267, 168 e 270 compõem o cluster sequência CS167, mas estão contidos em diferentes clusters descrição: 305, 304, 271 e 267 estão no cluster descrição CD530 (pois todos do cluster possuem a descrição *Actinomyces naeslundii* hypothetical protein), enquanto o 168 e o 270 estão no cluster descrição CD589 por ter descrição similar a *Actinomyces naeslundii* beta-glucosidase. Esse é um caso curioso, pois apesar dos registros 305, 304, 271, 267, 168 e 270 possuírem sequência similar, a descrição dos últimos registros (168 e 270) é diferente da descrição dos demais, indicando que os registros 168 e 270 poderiam possuir sequência diferente dos demais.

Esse é um caso onde é necessário que o especialista de domínio gerencie os clusters. Ele toma decisão em relação a cada elemento do cluster apresentado, para os clusters que não foram identificados como identidade. Em sua decisão final, resultado da decisão de cada elemento diferente entre uma dupla de cluster descrição e cluster sequência, ele pode decidir manter o cluster descrição, ou decidir manter o cluster sequência, ou ainda criar novas divisões de clusters e modificar elementos entre os clusters.

Durante a fase de gerenciamento de clusters, caso o especialista de domínio decida manter um elemento no cluster descrição ao invés do cluster sequência, o elemento será removido do cluster sequência, e vice-versa, caso ele opte por manter o elemento no cluster sequência.

O mesmo conjunto de dados de entrada foi utilizado para implementação do Algoritmo PIC, descrito no capítulo de Trabalhos Correlatos. Os clusters de entrada desse algoritmo foram obtidos por meio do algoritmo *Hierarchical Cluster* do Weka

(WEKA, 2012). Em seguida, o PIC foi implementado utilizando a tecnologia Java JSE 6.24 e o ambiente de desenvolvimento Netbeans IDE 6.9.1.

Foram obtidos 65 clusters, dos quais 55 possuem apenas registros com a mesma descrição, e os demais são clusters sobrepostos, que incluem vários registros que não possuem similaridade de descrição nem de sequência, como pode ser visto na Listagem 14, que mostra um trecho dos elementos do cluster 3 gerado pelo Algoritmo PIC.

Os quatro registros mostrados não possuem similaridade de descrição nem de sequência. Isso ocorre porque o Algoritmo PIC realiza um *merge* dando confiabilidade para a primeira clusterização, pois o cluster acaba absorvendo elementos que não tem *match*, devido ao limiar de 0,50 utilizado por essa técnica, que perde a identidade dos clusters.

Listagem 14 Elementos do cluster 3 do algoritmo PIC.

```
...
   • Actinomyces naeslundii strain CCUG 34725 phenylalanyl-tRNA synthetase alpha
     subunit (pheS) gene, partial cds.
   • Actinomyces naeslundii fimbrial structural subunit (fimA) and putative
     fimbria-associated protein genes, complete cds.
   • Actinomyces naeslundii Permease for cytosine/purines, uracil, thiamine,
     allantoin
   • Actinomyces oris strain A18A-3 methionyl-tRNA synthetase (metG) gene, partial
     cds.
...
```

Comparando o Algoritmo PIC com o MIDB, observa-se que o MIDB constrói clusters com escopo mais restrito, realizando resolução de entidades que são mais similares no mundo real. Além disso, o MIDB permite que o especialista de domínio participe da integração, oferecendo dados para escolha entre similaridade de sequência ou descrição dos dados.

6.4 Considerações Finais

Neste capítulo foram descritos dois estudos de caso: um estudo de caso com a Etapa 2 – Categorização dos Atributos, e outro estudo de caso da Etapa 3 – Resolução de Entidades.

As avaliações foram realizadas em relação a dados de sequências de organismos do gênero *Actinomyces*, com o objetivo de integrar registros de genes que correspondem à mesma entidade do mundo real.

Também foi realizada uma comparação com o Algoritmo PIC, um dos trabalhos correlatos deste trabalho. O MIDB apresentou-se superior em realação à integração de instâncias mais similares ao mundo real, além de permitir a participação do usuário no processo de integração.

Capítulo 7

CONCLUSÃO

Este capítulo apresenta a conclusão do presente trabalho de mestrado e nele são destacadas as principais contribuições deste trabalho, bem como sugestões de trabalhos futuros e novos desafios.

7.1 Considerações iniciais

Neste trabalho foi proposto um modelo de integração de bancos de dados biológicos chamado MIDB. O modelo é composto por quatro etapas: **Carregamento de Dados** (Etapa 1); **Categorização dos Atributos** (Etapa 2); **Resolução de Entidades** (Etapa 3); e **Gerenciamento de Clusters** (Etapa 4). A partir dos dados extraídos das fontes de dados selecionadas (Etapa 1), os atributos dos esquemas de tais fontes são categorizados (Etapa 2). O objetivo da Etapa 2 é realizar a integração de esquemas para facilitar a próxima etapa, a Etapa 3, onde ocorre a integração de instâncias. Para realizar a Etapa 3, foi proposto o Algoritmo de Clusterização, baseado nos dados da Categoria Descrição e nas distâncias de edição de caracteres e de palavras; e a utilização do BLAST para agrupar os dados da Categoria SeqNuc. Nessa etapa ainda é realizado um Emparelhamento entre os clusters descrição e clusters sequência. Os clusters que não foram identificados como identidade vão para a etapa posterior, a Etapa 4, onde o especialista de domínio realiza o Gerenciamento de Clusters, decidindo em quais clusters alguns elementos devem estar contidos.

A Etapa 2 e a Etapa 3 foram avaliadas por meio de estudos de caso sobre dados de sequência de nucleotídeos de genes do gênero *Actinomyces*, provenientes

de quatro fontes de dados: DDBJ, EMBL, Genbank e Oralgen. O estudo de caso da Etapa 2 tem como objetivo classificar os atributos do esquema dessas bases de dados entre as Categorias, usando como base de comparação as Regras de Mapeamento de Categorias que foram propostas neste trabalho de mestrado. Num total de 53 atributos, sendo 5 com valores vazios que por esse motivo não puderam ser categorizados, o resultado foi de 47 atributos categorizados corretamente, num montante de 48 atributos, ou seja, aproximadamente 97,91% de acerto. Além disso, foi implementada uma forma de superar a restrição de categorização de dados numéricos, por meio de um dicionário de sinônimos que é consultado para buscar sinônimos ao nome do atributo.

A partir do experimento realizado no primeiro estudo de caso (referente à Etapa 2), foi realizado o estudo de caso da Etapa 3. Nele, os registros provenientes das fontes de dados utilizam os atributos da Categoria Descrição e da Categoria SeqNuc para realizar a integração de instâncias. O Algoritmo de Clusterização foi aplicado nos dados da Categoria Descrição, gerando 83 agrupamentos. Em paralelo, foram realizadas consultas BLAST para todos os registros e os dados resultantes foram analisados lexicamente, dando origem a 112 agrupamentos na Clusterização por Sequência. Os agrupamentos provenientes de ambos os processos foram comparados em um processo seguinte, denominado Emparelhamento de Clusters, dando um resultado de aproximadamente 49,39% dos clusters advindos dos dois processos anteriores sendo idênticos.

Os demais clusters foram apresentados para o especialista de domínio fazer uma análise e tomar decisão em relação a cada elemento não comum entre clusters sequência e descrição. Assim, o modelo MIDB conseguiu identificar que aproximadamente 50% dos clusters necessitavam do tratamento provido por um especialista de domínio e que agrupar somente por sequência ou agrupar somente por descrição pode levar a erro.

Em seguida são destacadas as principais contribuições deste trabalho (Seção 7.2), o quão adaptável é o modelo proposto (Seção 7.3) e, por fim, os trabalhos futuros (Seção 7.4).

7.2 Contribuições

Neste trabalho, as principais contribuições foram a Categorização de Atributos, o Algoritmo de Clusterização, a Distância de Edição de Caracteres e a Distância de Edição de Palavras e a proposta de modelo de integração de dados biológicos (MIDB).

A Categorização de Atributos é um método inovador de realizar a integração de esquemas, diferenciado dos demais citados no capítulo de integração de dados. Ela é realizada automaticamente e utiliza como base de comparação os valores dos próprios atributos para identificar sua Categoria. Isso reduz o tempo que especialistas gastariam para integrar esquemas manualmente, tendo que estudar esquemas de uma fonte de dados e uma estratégia de integrá-los às demais fontes. A Categorização de Atributos reduz drasticamente o tempo utilizado nessa tarefa dispendiosa para minutos ou até segundos (manualmente essa atividade pode levar dias, dependendo do tamanho do banco de dados). Cabe salientar que a Categorização de Atributos deste trabalho foi proposta para dados do domínio de sequência de nucleotídeos, mas com Regras de Mapeamento de Categorias corretamente criadas, ela pode ser extensível a outros domínios de dados.

O Algoritmo de Clusterização é responsável por realizar a integração de instâncias cujos dados sejam descritivos (literais). Com base na métrica Distância de Edição de Caracteres e na métrica Distância de Edição de Palavras (que também são contribuições deste trabalho), é possível comparar registros e estabelecer uma medida de distância entre eles, verificando a similaridade de caracteres e palavras entre dois literais. Um limiar de distância define se um registro deve estar contido ou não em um determinado cluster. Esses agrupamentos (clusters) representam uma instância do mundo real sendo a estrutura onde ocorre de fato a integração das instâncias. Nesse cenário foi necessário utilizar dados de sequência, então foram realizados processos posteriores de comparações de dados de sequência e emparelhamento; porém se houver apenas dados do tipo literal, o Algoritmo de Clusterização pode ser aplicado sem fases posteriores, como uma forma de integração de instância dos dados.

A proposta do modelo MIDB também é uma contribuição deste trabalho. Ela realiza a integração de esquemas baseada na Categorização dos Atributos, que é

inovadora; e a integração de instâncias é feita com base no Algoritmo de Clusterização, utilização do BLAST e Emparelhamento, na qual a geração de clusters é utilizada por poucos trabalhos da literatura. Além disso, o especialista de domínio tem espaço para tomar decisão em relação às instâncias que não foram iguais entre o processo do Algoritmo de Clusterização e o BLAST, o que é muito importante, pois permite que os dados sejam “curados”.

O MIDB apresentou-se superior tanto ao BLAST quanto ao Algoritmo PIC. Em relação ao primeiro, o MIDB apresenta também a clusterização por meio de descrição, que permite observar que em algumas situações o BLAST pode levar a erro. Em relação ao Algoritmo PIC, o MIDB agrupa as instâncias com uma similaridade maior em relação ao mundo real, restringindo para que apenas elementos de um escopo mais próximo estejam presentes no mesmo cluster. Além disso, uma vantagem a mais do MIDB é a participação do especialista do domínio na análise dos resultados, permitindo que ele possa fazer adequações nos resultados de acordo com dados obtidos em experimentos reais.

7.3 Adaptabilidade do Modelo Proposto

O modelo proposto possui quatro etapas conforme apresentado neste trabalho. A Etapa 1 e a Etapa 4 desse modelo são independentes do domínio e podem ser aplicadas a outros domínios sem nenhuma modificação. A Etapa 2 e a Etapa 3 também são aplicáveis a outros domínios, porém necessitam de algumas alterações.

A Etapa 1 é independente de domínio, pois é a etapa onde os encapsuladores são construídos para fazer extração de dados de uma fonte de dados e formatá-los em um arquivo XML.

A Etapa 2 tem a restrição que os atributos sejam do tipo literal para que as Regras de Mapeamento de Categorias possam ser aplicadas. Essas regras devem ser validadas e novas devem ser criadas dependendo do domínio de dados que sofrerá a integração de esquemas. A criação das regras exige um esforço maior compatível com qualquer sistema baseado no uso de regras.

A Etapa 3 é dependente de um dicionário, do Algoritmo de Clusterização (Processo de Clusterização por Descrição), do uso do BLAST (Processo de Clusterização por Sequência) e do Processo de Emparelhamento de Clusters. O dicionário, o Algoritmo de Clusterização e o Processo de Emparelhamento de Clusters podem ser aplicados a outros domínios, com algumas alterações nos dois primeiros. O dicionário deve ser alterado com termos do domínio desejado, o Algoritmo de Clusterização eventualmente pode necessitar de alteração do limiar de distância para definir se um elemento está contido ou não em um cluster, e o Processo de Emparelhamento de Clusters não necessita nenhuma alteração. A criação do Dicionário depende do especialista de domínio e também requer esforço adicional.

A Etapa 4 é independente do domínio, ou seja, os termos podem ser gerenciados por especialistas sem existir restrição às informações armazenadas no banco de dados.

7.4 Trabalhos Futuros

A seguir são enumeradas sugestões de trabalhos futuros:

- Extensão do trabalho para integrar outros dados biológicos. Seria interessante fazer a integração de dados como proteínas, vias metabólicas, publicações, etc. Para isso novas fontes de dados precisam ser adicionadas e novas Regras de Mapeamento de Categorias necessitarão ser criadas. Além disso, a etapa de Resolução de Entidades precisará ser estendida para poder lidar com esses novos dados biológicos.
- Além do dicionário, utilizar uma ontologia de domínio (por exemplo, *Gene Ontology* e *Sequence Ontology*, que são ontologias difundidas entre a comunidade) para poder identificar sinônimos e integrar mais facilmente os dados.
- Utilização de procedência de dados para mapear a origem e as transformações que podem ser aplicadas a um dado. Quando ocorrem as decisões do especialista de domínio, ele pode sugerir modificações

nos dados, e os valores anteriores podem ser perdidos. Ao usar a procedência, pode-se ter uma linha de tempo das transformações que o dado sofreu.

REFERÊNCIAS

AMES, R. M.; MONEY, D.; GHATGE, V. P.; WHELAN, S.; LOVELL, S. C. Determining the evolutionary history of gene families. **Bioinformatics**, v. 28, n. 1, p. 48-55, 2012.

BENJELLOUN, O.; GARCIA-MOLINA, H.; MENESTRINA, D.; SU, Q.; EUIJONG, S. W.; WIDOM, J. Swoosh: a generic approach to entity resolution. **The VLDB Journal**, v. 18, n.1, p. 255-276, 2009.

BENSON, D. A.; KARSCH-MIZRACHI, I.; LIPMAN, D. J.; OSTELL, J.; WHEELER, D. L. Genbank. **Nucleic Acids Research**, v. 34 (Database Issue), p. D16-D20, 2006.

BHATTACHARYA, I.; GETOOR, L. Collective Entity Resolution in Relational Data. **ACM Transactions on Knowledge Discovery from Data**, v. 1, article 5, 36p, 2007.

BLAST – Basic Local Alignment Search Tool. Disponível em < http://blast.ncbi.nlm.nih.gov/>. Acesso em 20 jan 2010.

BRIACHE, A.; MARRAKCHI, K.; KERZAZI, A.; NAVAS-DELGADO, I.; ALDANA-MONTES, J. F.; HASSANI, B. D. R.; LAIRINI, K. YeastMed: an XML-Based System for Biological Data Integration of Yeast. In: International Workshop on Semantic Web Applications and Tools for the Life Sciences, 3., 2010, Berlin. **Proceedings**. 2010.

CHEN, Z.; KALASHNIKOV, D.V.; MEHROTRA, S. Exploiting relationships for object consolidation. In **Proceedings of the International Workshop on Information Quality in Information Systems** (IQIS), p. 47-58, 2005.

DNA Data Bank of Japan (DDBJ). Disponível em <http://www.ddbj.nig.ac.jp/Welcome-e.html> . Acesso em 20 jan 2010.

European Nucleotide Sequence Database (EMBL). Disponível em <http://www.ebi.ac.uk/embl/index.html>. Acesso em 25 mai 2010.

GALPERIN, M. Y. The Molecular Biology Database Collection: 2005 update. **Nucleic Acids Research**, v. 33 (Database Issue), p. D5-D24, 2005.

GALPERIN, M. Y.; COCHRANE, G. R. The 2011 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection. **Nucleic Acids Research**, v. 39 (Database Issue), p. D1-D6, 2011.

Genbank. Disponível em <http://www.ncbi.nlm.nih.gov/Genbank/>. Acesso em 20 jan 2010.

GenProtEC, *E.coli* Genome and proteome database. Disponível em <http://genprotec.mbl.edu/>. Acesso em 20 jan 2010.

GREENE, D.; BRYAN, K.; CUNNINGHAM, P. Parallel integration of heterogeneous genome-wide data sources. In: IEEE International Conference on Bioinformatics and Bioengineering, 8., 2008, Atenas. **Proceedings**. 2008.

GUSNANTO, A.; WOOD, H. M.; PAWITAN, Y.; RABBITTS, P.; BERRI, S. Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data. **Bioinformatics**, v. 28, n. 1, p. 40-47, 2012.

HAMMING, R. W. Error detecting and error correcting codes. **Bell System Technical Journal**, v. 26, n. 2, p. 147-160, 1950.

LANDER, E. S. et al. Initial sequencing and analysis of the human genome. **Nature**, v. 409, p. 860-921, 2001.

LEVENSHTEIN, V. I. Binary codes capable of correcting deletions, insertions, and reversals. **Soviet Physics Doklady**, v. 10, 1966.

LIU, L.; ÖZSU, M. T. **Encyclopedia of Database Systems**. 1. ed.: Springer, 2009. 3752 p.

MEDCRAFT, P. S. **Integração de Bancos de Dados Federados na Web Usando Agentes Móveis**. 117 f. Dissertação de Mestrado em Informática: Universidade Federal da Paraíba, 2003.

MENESTRINA, D.; BENJELLOUN, O.; GARCIA-MOLINA, H. Generic entity resolution with data confidences. In: **VLDB Workshop on Clean Databases**, 2006, p. 25-32.

NAKANO, M. **Um Ambiente de Teste para Pesquisa de Similaridade em Bancos de Dados Biológicos**. 90 f. Dissertação de Mestrado em Ciência da Computação: Universidade Estadual do Maringá, 2005.

NATIONAL Library of Medicine. **Free Web-Based Access to NLM Databases**. NLM Technical Bulletin. Bethesda, EUA. 1997.

NAVAS-DELGADO, I.; ALDANA-MONTES, J. F. Extending SD-Core for Ontology-based Data Integration. **Journal of Universal Computer Science** v. 15, n. 17, p. 3201-3230, 2009.

NCBI-Genbank Flat File Release 187.0 Distribution Release Notes. Disponível em <ftp://ftp.ncbi.nih.gov/Genbank/gbrel.txt>. Acesso em 23 dez 2011.

ON, B-W.; ELMACIOGLU, E.; LEE, D.; KANG, J.; PEI, J. Improving Grouped-Entity Resolution Using Quasi-Cliques. In: IEEE International Conference on Data Mining, 6., 2006, Hong Kong. **Proceedings**. 2006. p. 18-22.

PREDICTIVE Testing for Huntington's Disease. Disponível em <http://www.healthsystem.virginia.edu/internet/huntdisease/geninfo.cfm>. Acesso em 25 jan 2010.

SEIBEL, L. F. B.; LEMOS, M.; LIFSCHITZ, S. Implementation Issues of Bio-AXS: an Object-oriented Framework for Integrating Biological Data and Applications. In: International Database Engineering and Applications Symposium, 7., 2003, Hong Kong. **Proceedings**. 2003.

SHAH, S. P.; HUANG, Y.; XU, T.; YUEN, M. M. S.; LING, J.; OUELETTE, B. F. Atlas – a data warehouse for integrative bioinformatics. **BMC Bioinformatics**, v. 6, 2005.

SILVA JUNIOR, W. A. **Laboratório de Genética Molecular e Bioinformática**. Disponível em: <http://lgmb.fmrp.usp.br>. Acesso em 2006.

TEIXEIRA, M. V. C. **Gerenciamento de anotações de biossequências utilizando associações entre ontologias e esquemas XML**. 105 f. Dissertação de Mestrado em Ciência da Computação. Universidade Federal de São Carlos, 2008.

UNIPROT. *Keratin, type II cytoskeletal 2 epidermal – Homo sapiens*. Disponível em < http://www.uniprot.org/uniprot/P35908>. Acesso em 25 jan 2010.

VEGA - Vertebrate Genome Annotation database. Disponível em <http://vega.sanger.ac.uk/index.html. >. Acesso em 18 jan 2010.

Weka **Waikato Environment for Knowledge Analysis**, versão 3. Universidade de Waikato, 2012.

WESCHE, P. L.; GAFFNEY, D. J.; KEIGHTLEY, P. D. DNA sequence error rates in Genbank records estimated using the mouse genome as a reference. **DNA Sequence - The Journal of Sequencing and Mapping** v. 15, n. 5-6, p. 362-364, 2004.

wwPDB – World Wide Protein Data Bank. Disponível em <http://www.wwpdb.org/>. Acesso em 18 jan 2010.

YANAGA, F. **Um Estudo de Desempenho da Estrutura de Indexação MRS em Memória Secundária**. 116 f. Dissertação de Mestrado em Ciência da Computação. Universidade Estadual do Maringá, 2006.

ZIEGLER, P.; DITTRICH, K. R. Three Decades of Data Integration - All Problems Solved?, **18th IFIP World Computer Congress** (WCC), v. 12, p. 3-12, 2004.

APÊNDICE A – SAÍDA DE ARQUIVO BLAST

Na Listagem 15 é mostrado um exemplo de saída de uma consulta no BLAST. O registro que está sendo consultado é o registro 18, conforme mostra no trecho *query* da listagem (Query= gi|18| AF048780 | DDBJ| *Actinomyces naeslundii* urease alpha-subunit (ureC) gene, complete cds.).

Note que os alinhamentos significativos foram os dos registros 473 (proveniente do Genbank), 310 (proveniente do EMBL) e o próprio registro 18 (proveniente do DDBJ) também é mostrado como alinhamento significativo.

Listagem 15 Exemplo de saída de uma consulta no BLAST.

```
BLASTN 2.2.25+


Reference: Zheng Zhang, Scott Schwartz, Lukas Wagner, and Webb
Miller (2000), "A greedy algorithm for aligning DNA sequences", J
Comput Biol 2000; 7(1-2):203-14.



Database: db\bdbs.fasta
           606 sequences; 433,469 total letters



Query= gi|18| AF048780  | DDBJ| Actinomyces naeslundii urease alpha-subunit
(ureC) gene, complete cds.

Length=1713
                                                                      Score     E
Sequences producing significant alignments:                          (Bits)  Value

gi|473  AF048780  | Genbank| Actinomyces naeslundii urease alpha-...  3164    0.0
gi|310  AF048780.1| EMBL| Actinomyces naeslundii urease alpha-sub...  3164    0.0
gi|18  AF048780  | DDBJ| Actinomyces naeslundii urease alpha-subu...  3164    0.0


>gi|473 AF048780  | Genbank| Actinomyces naeslundii urease alpha-subunit
(ureC) gene, complete cds.
Length=1713

 Score = 3164 bits (1713),  Expect = 0.0
```

```
 Identities = 1713/1713 (100%), Gaps = 0/1713 (0%)
 Strand=Plus/Plus

Query  1     ATGAGCAAGGAACTCAGCCGTCGGGAGCACGCCGCTCTGTACGGCCCCACCACTGGGGAC  60
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1     ATGAGCAAGGAACTCAGCCGTCGGGAGCACGCCGCTCTGTACGGCCCCACCACTGGGGAC  60

Query  61    GCCGTCCGCCTGGCCGATACCGGTCTCTTCGCCCAGATCGAGCGGGACCTGACCCACCGT  120
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  61    GCCGTCCGCCTGGCCGATACCGGTCTCTTCGCCCAGATCGAGCGGGACCTGACCCACCGT  120

Query  121   GGCGACGAGGCCGTCTTCGGAGGCGGCAAGGTCATCCGCGACGGAATGGGCCACAACGGC  180
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  121   GGCGACGAGGCCGTCTTCGGAGGCGGCAAGGTCATCCGCGACGGAATGGGCCACAACGGC  180

Query  181   CAGCGCACGCGAGACGAGGACATCCCCGACACCGTCATCACCAACGCGATCATCATCGAC  240
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  181   CAGCGCACGCGAGACGAGGACATCCCCGACACCGTCATCACCAACGCGATCATCATCGAC  240

Query  241   CACACCGGGGTCTACAAGGCCGACGTCGCCATCCGCGACGGCGTCATCAGCGCTATCGGC  300
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  241   CACACCGGGGTCTACAAGGCCGACGTCGCCATCCGCGACGGCGTCATCAGCGCTATCGGC  300

Query  301   GCAGCGGGCAACCCGGACATCATGGACGACGTCGACATCGTCATCGGCACCTCCACGGAG  360
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  301   GCAGCGGGCAACCCGGACATCATGGACGACGTCGACATCGTCATCGGCACCTCCACGGAG  360

Query  361   GTGATCGCCGGCGAGCACCGCATCCTGACGGCCGGAGGCATCGACTCCCACATTCACTTC  420
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  361   GTGATCGCCGGCGAGCACCGCATCCTGACGGCCGGAGGCATCGACTCCCACATTCACTTC  420

Query  421   ATCTCCCCCACCCAGGTCGCCACCGCCCTGGCCTCGGGCGTGACCACCATGATCGGAGGA  480
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  421   ATCTCCCCCACCCAGGTCGCCACCGCCCTGGCCTCGGGCGTGACCACCATGATCGGAGGA  480

Query  481   GGGACAGGACCCAGCGACGGCACCAACGCCACCACGATCACTCCCGGGGCGTGGAACCTG  540
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  481   GGGACAGGACCCAGCGACGGCACCAACGCCACCACGATCACTCCCGGGGCGTGGAACCTG  540

Query  541   GCCCGTATGCTCCAGGCGGTCGAGGACTTCCCCATGAATATCGGCCTGCTGGGCAAGGGC  600
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  541   GCCCGTATGCTCCAGGCGGTCGAGGACTTCCCCATGAATATCGGCCTGCTGGGCAAGGGC  600

Query  601   CACGCCTCGGCCCGGGAGCCTCTGGCCGAGCAGCTGCGCGCCGGTGCGGTCGGCTTCAAG  660
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  601   CACGCCTCGGCCCGGGAGCCTCTGGCCGAGCAGCTGCGCGCCGGTGCGGTCGGCTTCAAG  660

Query  661   ATCCACGAGGACTGGGGTGCCACGCACGCCGTCATCGACGAGGCGCTCAAGGTCGCCGAC  720
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  661   ATCCACGAGGACTGGGGTGCCACGCACGCCGTCATCGACGAGGCGCTCAAGGTCGCCGAC  720

Query  721   GAGTTCGACGTGCAGGTCGCCATCCACACCGACACGCTCAACGAGTGCGGTTTCGTCGAG  780
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  721   GAGTTCGACGTGCAGGTCGCCATCCACACCGACACGCTCAACGAGTGCGGTTTCGTCGAG  780

Query  781   GACACCCGCCGCGCGATCGGCGGACGGGTCATCCACACCTTCCACACCGAGGGCGCCGGC  840
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  781   GACACCCGCCGCGCGATCGGCGGACGGGTCATCCACACCTTCCACACCGAGGGCGCCGGC  840

Query  841   GGCGGCCACGCCCCCGACATCATCACCCTGGCGCAGGACTCCAACATCCTGCCGTCCTCG  900
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  841   GGCGGCCACGCCCCCGACATCATCACCCTGGCGCAGGACTCCAACATCCTGCCGTCCTCG  900

Query  901   ACCAACCCGACGCTGCCCTTCACGCGCAACACCGCTGAGGAGCATCTCGACATGCTCATG  960
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  901   ACCAACCCGACGCTGCCCTTCACGCGCAACACCGCTGAGGAGCATCTCGACATGCTCATG  960

Query  961   GTGTGCCACCACCTCAACCCCGCCATCCCGGAGGACGTCGCCTTCGCCGACTCGCGCATC  1020
```

```
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  961   GTGTGCCACCACCTCAACCCCGCCATCCCGGAGGACGTCGCCTTCGCCGACTCGCGCATC  1020

Query  1021  CGCCCGGAGACGATCGCGGCCGAGGACGTTCTGCACGACCTGGGCGTGTTCTCGATGACG  1080
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1021  CGCCCGGAGACGATCGCGGCCGAGGACGTTCTGCACGACCTGGGCGTGTTCTCGATGACG  1080

Query  1081  TCGTCGGACTCCCAGGCCATGGGGCGCGTCGGTGAGGTCATCATCCGCACCTGGCAGGTG  1140
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1081  TCGTCGGACTCCCAGGCCATGGGGCGCGTCGGTGAGGTCATCATCCGCACCTGGCAGGTG  1140

Query  1141  GCCGACCAGATGAAGAAGGCGCGCGGCAGGCTCGCCGGGGACCCGGAGGACGGCGACAAC  1200
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1141  GCCGACCAGATGAAGAAGGCGCGCGGCAGGCTCGCCGGGGACCCGGAGGACGGCGACAAC  1200

Query  1201  CTGCGCATCAAGCGCTACGTGTCGAAGTACACGATCAACCCGGCCCGAGCCAACGGGATC  1260
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1201  CTGCGCATCAAGCGCTACGTGTCGAAGTACACGATCAACCCGGCCCGAGCCAACGGGATC  1260

Query  1261  GCCGAGGTCGTGGGCAGCGTGGAGGTGGGCAAGTGGGCCGACCTGGTGCTGTGGGACCCC  1320
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1261  GCCGAGGTCGTGGGCAGCGTGGAGGTGGGCAAGTGGGCCGACCTGGTGCTGTGGGACCCC  1320

Query  1321  GCCTTCTTCGGGGTCAAGCCGTCTTTGATCCTCAAGGGGGGCCAGATCGCCTCCGCGGTC  1380
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1321  GCCTTCTTCGGGGTCAAGCCGTCTTTGATCCTCAAGGGGGGCCAGATCGCCTCCGCGGTC  1380

Query  1381  ATGGGCGACGCCAACGCCTCGATCCCCACCCCGGAGCCCACGCTCATGCGCACCATGTTC  1440
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1381  ATGGGCGACGCCAACGCCTCGATCCCCACCCCGGAGCCCACGCTCATGCGCACCATGTTC  1440

Query  1441  GGCGGCCACGGGGCGGCACCGGCGTCGAACTCGATCACCTTCATGTCCCAGGCGGCCATC  1500
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1441  GGCGGCCACGGGGCGGCACCGGCGTCGAACTCGATCACCTTCATGTCCCAGGCGGCCATC  1500

Query  1501  GATGCGGGAGTGCCCCAGAGCCTGGGGCTGCGCAAGAAGGTCTGTCCGGCCCACGGGGTC  1560
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1501  GATGCGGGAGTGCCCCAGAGCCTGGGGCTGCGCAAGAAGGTCTGTCCGGCCCACGGGGTC  1560

Query  1561  AGGCGGCTGACGAAAGCGGACATGGCCTTCAACGACGCCACTCCGGCGCTCACCGTCGAT  1620
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1561  AGGCGGCTGACGAAAGCGGACATGGCCTTCAACGACGCCACTCCGGCGCTCACCGTCGAT  1620

Query  1621  CCCGAGACCTACGAGGTCACCGTCGACGGCGAGAAGGTCACCTGCGAGCCTGCCGAGGTC  1680
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1621  CCCGAGACCTACGAGGTCACCGTCGACGGCGAGAAGGTCACCTGCGAGCCTGCCGAGGTC  1680

Query  1681  CTCGCCATGGCCCAGCGCTACTTCCTCTTCTGA  1713
            ||||||||||||||||||||||||||||||||
Sbjct  1681  CTCGCCATGGCCCAGCGCTACTTCCTCTTCTGA  1713


>gi|310 AF048780.1| EMBL| Actinomyces naeslundii urease alpha-subunit
Length=1713

 Score = 3164 bits (1713),  Expect = 0.0
 Identities = 1713/1713 (100%), Gaps = 0/1713 (0%)
 Strand=Plus/Plus

Query  1     ATGAGCAAGGAACTCAGCCGTCGGGAGCACGCCGCTCTGTACGGCCCCACCACTGGGGAC  60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1     ATGAGCAAGGAACTCAGCCGTCGGGAGCACGCCGCTCTGTACGGCCCCACCACTGGGGAC  60

Query  61    GCCGTCCGCCTGGCCGATACCGGTCTCTTCGCCCAGATCGAGCGGGACCTGACCCACCGT  120
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  61    GCCGTCCGCCTGGCCGATACCGGTCTCTTCGCCCAGATCGAGCGGGACCTGACCCACCGT  120

Query  121   GGCGACGAGGCCGTCTTCGGAGGCGGCAAGGTCATCCGCGACGGAATGGGCCACAACGGC  180
```

```
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  121  GGCGACGAGGCCGTCTTCGGAGGCGGCAAGGTCATCCGCGACGGAATGGGCCACAACGGC  180

Query  181  CAGCGCACGCGAGACGAGGACATCCCCGACACCGTCATCACCAACGCGATCATCATCGAC  240
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  181  CAGCGCACGCGAGACGAGGACATCCCCGACACCGTCATCACCAACGCGATCATCATCGAC  240

Query  241  CACACCGGGGTCTACAAGGCCGACGTCGCCATCCGCGACGGCGTCATCAGCGCTATCGGC  300
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  241  CACACCGGGGTCTACAAGGCCGACGTCGCCATCCGCGACGGCGTCATCAGCGCTATCGGC  300

Query  301  GCAGCGGGCAACCCGGACATCATGGACGACGTCGACATCGTCATCGGCACCTCCACGGAG  360
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  301  GCAGCGGGCAACCCGGACATCATGGACGACGTCGACATCGTCATCGGCACCTCCACGGAG  360

Query  361  GTGATCGCCGGCGAGCACCGCATCCTGACGGCCGGAGGCATCGACTCCCACATTCACTTC  420
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  361  GTGATCGCCGGCGAGCACCGCATCCTGACGGCCGGAGGCATCGACTCCCACATTCACTTC  420

Query  421  ATCTCCCCCACCCAGGTCGCCACCGCCCTGGCCTCGGGCGTGACCACCATGATCGGAGGA  480
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  421  ATCTCCCCCACCCAGGTCGCCACCGCCCTGGCCTCGGGCGTGACCACCATGATCGGAGGA  480

Query  481  GGGACAGGACCCAGCGACGGCACCAACGCCACCACGATCACTCCCGGGGCGTGGAACCTG  540
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  481  GGGACAGGACCCAGCGACGGCACCAACGCCACCACGATCACTCCCGGGGCGTGGAACCTG  540

Query  541  GCCCGTATGCTCCAGGCGGTCGAGGACTTCCCCATGAATATCGGCCTGCTGGGCAAGGGC  600
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  541  GCCCGTATGCTCCAGGCGGTCGAGGACTTCCCCATGAATATCGGCCTGCTGGGCAAGGGC  600

Query  601  CACGCCTCGGCCCGGGAGCCTCTGGCCGAGCAGCTGCGCGCCGGTGCGGTCGGCTTCAAG  660
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  601  CACGCCTCGGCCCGGGAGCCTCTGGCCGAGCAGCTGCGCGCCGGTGCGGTCGGCTTCAAG  660

Query  661  ATCCACGAGGACTGGGGTGCCACGCACGCCGTCATCGACGAGGCGCTCAAGGTCGCCGAC  720
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  661  ATCCACGAGGACTGGGGTGCCACGCACGCCGTCATCGACGAGGCGCTCAAGGTCGCCGAC  720

Query  721  GAGTTCGACGTGCAGGTCGCCATCCACACCGACACGCTCAACGAGTGCGGTTTCGTCGAG  780
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  721  GAGTTCGACGTGCAGGTCGCCATCCACACCGACACGCTCAACGAGTGCGGTTTCGTCGAG  780

Query  781  GACACCCGCCGCGCGATCGGCGGACGGGTCATCCACACCTTCCACACCGAGGGCGCCGGC  840
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  781  GACACCCGCCGCGCGATCGGCGGACGGGTCATCCACACCTTCCACACCGAGGGCGCCGGC  840

Query  841  GGCGGCCACGCCCCCGACATCATCACCCTGGCGCAGGACTCCAACATCCTGCCGTCCTCG  900
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  841  GGCGGCCACGCCCCCGACATCATCACCCTGGCGCAGGACTCCAACATCCTGCCGTCCTCG  900

Query  901  ACCAACCCGACGCTGCCCTTCACGCGCAACACCGCTGAGGAGCATCTCGACATGCTCATG  960
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  901  ACCAACCCGACGCTGCCCTTCACGCGCAACACCGCTGAGGAGCATCTCGACATGCTCATG  960

Query  961  GTGTGCCACCACCTCAACCCCGCCATCCCGGAGGACGTCGCCTTCGCCGACTCGCGCATC  1020
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  961  GTGTGCCACCACCTCAACCCCGCCATCCCGGAGGACGTCGCCTTCGCCGACTCGCGCATC  1020

Query  1021 CGCCCGGAGACGATCGCGGCCGAGGACGTTCTGCACGACCTGGGCGTGTTCTCGATGACG  1080
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1021 CGCCCGGAGACGATCGCGGCCGAGGACGTTCTGCACGACCTGGGCGTGTTCTCGATGACG  1080

Query  1081 TCGTCGGACTCCCAGGCCATGGGGCGCGTCGGTGAGGTCATCATCCGCACCTGGCAGGTG  1140
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1081 TCGTCGGACTCCCAGGCCATGGGGCGCGTCGGTGAGGTCATCATCCGCACCTGGCAGGTG  1140

Query  1141 GCCGACCAGATGAAGAAGGCGCGCGGCAGGCTCGCCGGGGACCCGGAGGACGGCGACAAC  1200
```

```
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1141  GCCGACCAGATGAAGAAGGCGCGCGGCAGGCTCGCCGGGGACCCGGAGGACGGCGACAAC  1200

Query  1201  CTGCGCATCAAGCGCTACGTGTCGAAGTACACGATCAACCCGGCCCGAGCCAACGGGATC  1260
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1201  CTGCGCATCAAGCGCTACGTGTCGAAGTACACGATCAACCCGGCCCGAGCCAACGGGATC  1260

Query  1261  GCCGAGGTCGTGGGCAGCGTGGAGGTGGGCAAGTGGGCCGACCTGGTGCTGTGGGACCCC  1320
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1261  GCCGAGGTCGTGGGCAGCGTGGAGGTGGGCAAGTGGGCCGACCTGGTGCTGTGGGACCCC  1320

Query  1321  GCCTTCTTCGGGGTCAAGCCGTCTTTGATCCTCAAGGGGGGCCAGATCGCCTCCGCGGTC  1380
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1321  GCCTTCTTCGGGGTCAAGCCGTCTTTGATCCTCAAGGGGGGCCAGATCGCCTCCGCGGTC  1380

Query  1381  ATGGGCGACGCCAACGCCTCGATCCCCACCCCGGAGCCCACGCTCATGCGCACCATGTTC  1440
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1381  ATGGGCGACGCCAACGCCTCGATCCCCACCCCGGAGCCCACGCTCATGCGCACCATGTTC  1440

Query  1441  GGCGGCCACGGGGCGGCACCGGCGTCGAACTCGATCACCTTCATGTCCCAGGCGGCCATC  1500
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1441  GGCGGCCACGGGGCGGCACCGGCGTCGAACTCGATCACCTTCATGTCCCAGGCGGCCATC  1500

Query  1501  GATGCGGGAGTGCCCCAGAGCCTGGGGCTGCGCAAGAAGGTCTGTCCGGCCCACGGGGTC  1560
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1501  GATGCGGGAGTGCCCCAGAGCCTGGGGCTGCGCAAGAAGGTCTGTCCGGCCCACGGGGTC  1560

Query  1561  AGGCGGCTGACGAAAGCGGACATGGCCTTCAACGACGCCACTCCGGCGCTCACCGTCGAT  1620
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1561  AGGCGGCTGACGAAAGCGGACATGGCCTTCAACGACGCCACTCCGGCGCTCACCGTCGAT  1620

Query  1621  CCCGAGACCTACGAGGTCACCGTCGACGGCGAGAAGGTCACCTGCGAGCCTGCCGAGGTC  1680
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1621  CCCGAGACCTACGAGGTCACCGTCGACGGCGAGAAGGTCACCTGCGAGCCTGCCGAGGTC  1680

Query  1681  CTCGCCATGGCCCAGCGCTACTTCCTCTTCTGA  1713
            ||||||||||||||||||||||||||||||||
Sbjct  1681  CTCGCCATGGCCCAGCGCTACTTCCTCTTCTGA  1713


>gi|18 AF048780  | DDBJ| Actinomyces naeslundii urease alpha-subunit
(ureC) gene, complete cds.
Length=1713

 Score = 3164 bits (1713),  Expect = 0.0
 Identities = 1713/1713 (100%), Gaps = 0/1713 (0%)
 Strand=Plus/Plus

Query  1     ATGAGCAAGGAACTCAGCCGTCGGGAGCACGCCGCTCTGTACGGCCCCACCACTGGGGAC  60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1     ATGAGCAAGGAACTCAGCCGTCGGGAGCACGCCGCTCTGTACGGCCCCACCACTGGGGAC  60

Query  61    GCCGTCCGCCTGGCCGATACCGGTCTCTTCGCCCAGATCGAGCGGGACCTGACCCACCGT  120
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  61    GCCGTCCGCCTGGCCGATACCGGTCTCTTCGCCCAGATCGAGCGGGACCTGACCCACCGT  120

Query  121   GGCGACGAGGCCGTCTTCGGAGGCGGCAAGGTCATCCGCGACGGAATGGGCCACAACGGC  180
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  121   GGCGACGAGGCCGTCTTCGGAGGCGGCAAGGTCATCCGCGACGGAATGGGCCACAACGGC  180

Query  181   CAGCGCACGCGAGACGAGGACATCCCCGACACCGTCATCACCAACGCGATCATCATCGAC  240
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  181   CAGCGCACGCGAGACGAGGACATCCCCGACACCGTCATCACCAACGCGATCATCATCGAC  240

Query  241   CACACCGGGGTCTACAAGGCCGACGTCGCCATCCGCGACGGCGTCATCAGCGCTATCGGC  300
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  241   CACACCGGGGTCTACAAGGCCGACGTCGCCATCCGCGACGGCGTCATCAGCGCTATCGGC  300
```

```
Query  301   GCAGCGGGCAACCCGGACATCATGGACGACGTCGACATCGTCATCGGCACCTCCACGGAG  360
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  301   GCAGCGGGCAACCCGGACATCATGGACGACGTCGACATCGTCATCGGCACCTCCACGGAG  360

Query  361   GTGATCGCCGGCGAGCACCGCATCCTGACGGCCGGAGGCATCGACTCCCACATTCACTTC  420
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  361   GTGATCGCCGGCGAGCACCGCATCCTGACGGCCGGAGGCATCGACTCCCACATTCACTTC  420

Query  421   ATCTCCCCCACCCAGGTCGCCACCGCCCTGGCCTCGGGCGTGACCACCATGATCGGAGGA  480
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  421   ATCTCCCCCACCCAGGTCGCCACCGCCCTGGCCTCGGGCGTGACCACCATGATCGGAGGA  480

Query  481   GGGACAGGACCCAGCGACGGCACCAACGCCACCACGATCACTCCCGGGGCGTGGAACCTG  540
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  481   GGGACAGGACCCAGCGACGGCACCAACGCCACCACGATCACTCCCGGGGCGTGGAACCTG  540

Query  541   GCCCGTATGCTCCAGGCGGTCGAGGACTTCCCCATGAATATCGGCCTGCTGGGCAAGGGC  600
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  541   GCCCGTATGCTCCAGGCGGTCGAGGACTTCCCCATGAATATCGGCCTGCTGGGCAAGGGC  600

Query  601   CACGCCTCGGCCCGGGAGCCTCTGGCCGAGCAGCTGCGCGCCGGTGCGGTCGGCTTCAAG  660
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  601   CACGCCTCGGCCCGGGAGCCTCTGGCCGAGCAGCTGCGCGCCGGTGCGGTCGGCTTCAAG  660

Query  661   ATCCACGAGGACTGGGGTGCCACGCACGCCGTCATCGACGAGGCGCTCAAGGTCGCCGAC  720
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  661   ATCCACGAGGACTGGGGTGCCACGCACGCCGTCATCGACGAGGCGCTCAAGGTCGCCGAC  720

Query  721   GAGTTCGACGTGCAGGTCGCCATCCACACCGACACGCTCAACGAGTGCGGTTTCGTCGAG  780
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  721   GAGTTCGACGTGCAGGTCGCCATCCACACCGACACGCTCAACGAGTGCGGTTTCGTCGAG  780

Query  781   GACACCCGCCGCGCGATCGGCGGACGGGTCATCCACACCTTCCACACCGAGGGCGCCGGC  840
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  781   GACACCCGCCGCGCGATCGGCGGACGGGTCATCCACACCTTCCACACCGAGGGCGCCGGC  840

Query  841   GGCGGCCACGCCCCCGACATCATCACCCTGGCGCAGGACTCCAACATCCTGCCGTCCTCG  900
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  841   GGCGGCCACGCCCCCGACATCATCACCCTGGCGCAGGACTCCAACATCCTGCCGTCCTCG  900

Query  901   ACCAACCCGACGCTGCCCTTCACGCGCAACACCGCTGAGGAGCATCTCGACATGCTCATG  960
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  901   ACCAACCCGACGCTGCCCTTCACGCGCAACACCGCTGAGGAGCATCTCGACATGCTCATG  960

Query  961   GTGTGCCACCACCTCAACCCCGCCATCCCGGAGGACGTCGCCTTCGCCGACTCGCGCATC  1020
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  961   GTGTGCCACCACCTCAACCCCGCCATCCCGGAGGACGTCGCCTTCGCCGACTCGCGCATC  1020

Query  1021  CGCCCGGAGACGATCGCGGCCGAGGACGTTCTGCACGACCTGGGCGTGTTCTCGATGACG  1080
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1021  CGCCCGGAGACGATCGCGGCCGAGGACGTTCTGCACGACCTGGGCGTGTTCTCGATGACG  1080

Query  1081  TCGTCGGACTCCCAGGCCATGGGGCGCGTCGGTGAGGTCATCATCCGCACCTGGCAGGTG  1140
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1081  TCGTCGGACTCCCAGGCCATGGGGCGCGTCGGTGAGGTCATCATCCGCACCTGGCAGGTG  1140

Query  1141  GCCGACCAGATGAAGAAGGCGCGCGGCAGGCTCGCCGGGGACCCGGAGGACGGCGACAAC  1200
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1141  GCCGACCAGATGAAGAAGGCGCGCGGCAGGCTCGCCGGGGACCCGGAGGACGGCGACAAC  1200

Query  1201  CTGCGCATCAAGCGCTACGTGTCGAAGTACACGATCAACCCGGCCCGAGCCAACGGGATC  1260
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1201  CTGCGCATCAAGCGCTACGTGTCGAAGTACACGATCAACCCGGCCCGAGCCAACGGGATC  1260

Query  1261  GCCGAGGTCGTGGGCAGCGTGGAGGTGGGCAAGTGGGCCGACCTGGTGCTGTGGGACCCC  1320
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1261  GCCGAGGTCGTGGGCAGCGTGGAGGTGGGCAAGTGGGCCGACCTGGTGCTGTGGGACCCC  1320
```

```
Query  1321  GCCTTCTTCGGGGTCAAGCCGTCTTTGATCCTCAAGGGGGGCCAGATCGCCTCCGCGGTC  1380
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1321  GCCTTCTTCGGGGTCAAGCCGTCTTTGATCCTCAAGGGGGGCCAGATCGCCTCCGCGGTC  1380

Query  1381  ATGGGCGACGCCAACGCCTCGATCCCCACCCCGGAGCCCACGCTCATGCGCACCATGTTC  1440
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1381  ATGGGCGACGCCAACGCCTCGATCCCCACCCCGGAGCCCACGCTCATGCGCACCATGTTC  1440

Query  1441  GGCGGCCACGGGGCGGCACCGGCGTCGAACTCGATCACCTTCATGTCCCAGGCGGCCATC  1500
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1441  GGCGGCCACGGGGCGGCACCGGCGTCGAACTCGATCACCTTCATGTCCCAGGCGGCCATC  1500

Query  1501  GATGCGGGAGTGCCCCAGAGCCTGGGGCTGCGCAAGAAGGTCTGTCCGGCCCACGGGGTC  1560
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1501  GATGCGGGAGTGCCCCAGAGCCTGGGGCTGCGCAAGAAGGTCTGTCCGGCCCACGGGGTC  1560

Query  1561  AGGCGGCTGACGAAAGCGGACATGGCCTTCAACGACGCCACTCCGGCGCTCACCGTCGAT  1620
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1561  AGGCGGCTGACGAAAGCGGACATGGCCTTCAACGACGCCACTCCGGCGCTCACCGTCGAT  1620

Query  1621  CCCGAGACCTACGAGGTCACCGTCGACGGCGAGAAGGTCACCTGCGAGCCTGCCGAGGTC  1680
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1621  CCCGAGACCTACGAGGTCACCGTCGACGGCGAGAAGGTCACCTGCGAGCCTGCCGAGGTC  1680

Query  1681  CTCGCCATGGCCCAGCGCTACTTCCTCTTCTGA  1713
             ||||||||||||||||||||||||||||||||
Sbjct  1681  CTCGCCATGGCCCAGCGCTACTTCCTCTTCTGA  1713



Lambda     K      H
    1.33    0.621     1.12

Gapped
Lambda     K      H
    1.28    0.460    0.850

Effective search space used: 713377717


  Database: db\bdbs.fasta
    Posted date:  Dec 11, 2011  1:38 PM
  Number of letters in database: 433,469
  Number of sequences in database:  605



Matrix: blastn matrix 1 -2
Gap Penalties: Existence: 0, Extension: 2.5
```

APÊNDICE B – ELEMENTOS DOS CLUSTERS CS167, CD530 E CD589

Na Listagem 16 são mostrados os elementos dos clusters CS167, CD530 e CD589. O formato da listagem é o formato FASTA. Nele podem-se encontrar o identificador único, o acesso, o BDB proveniente, o nome do gene e sua sequência.

Listagem 16 Elementos dos clusters CS167, CD530, CD589.

```
>gi|168| AY029505  | DDBJ| Actinomyces naeslundii beta-glucosidase gene, complete
cds.
atgaccgccacgtccactacttctaagagcaatccgaacttccccgacggcttcctgtggggcggggccaccgccgccaacca
gatcgagggcgcttacaacgaggacggcaagggcctgtccgtccaggacgtcatgcctcggggcatcatggcccaccccaccc
aggctcccacaccggataaccttcaagctcgaggcgatcgaccttctaccaccgcttacgccgaggacatctccctgttcgcg
gagatgggtttcaaggtcttccgcttctccatcgcctggagccgcatcttcccgctcggcgacgagaccgagcccaatgagga
aggactngccttctacgaccgggtcctcgacgagctcgagaagcacgggatcgagccactggtcaccatcagccactacgaga
ccccgctgcacctggcgcgcacctacgncggctggaccgaccgccgcctcatcggcttcttcgagcgctacgcccgcaccctg
ttcgagcgctatggcaagcgggtcaagtactggctcaccttcaacgagatcaactccgtgctccatgagcccttcctatctgg
gggcgtcgccacgcccaaggacaggcccccgagcaggacctctaccaggccatccaaaacgagctcgtcgcctccgcggccg
cgaccaggatcgcccatgagaccaaccccgacatccaggtcggctgcatgatcctggccgatcccacctacccgctcacccct
gatccccgggacgtgtgggcggccaagcaggcagagcgcgccaactacgccttcggagacctccacgtacgtggtgagtaccc
cggatacctgcggcggaccctgcgggacaagggcatcgagctggagatcaccgaggaggaccgcgtgctgctgcgggagcaca
ccgtcgacttcgtctccttctcctactacatgtncgtgtgcgagaccgtcacccagtcggccgaggccggccggggcaacctc
atgggcggcgtccccaatcccaccctcgaggcctccgagtggggatggcagatcgacccggcgggcctgcgcaccatcctgaa
cgactactgggaccgctggggcaagcctctgttcatcgtcgagaacggcctgggagccaaggacgtcctcgttgacggaccca
acggtcccacggtcgaggacgactaccgcatcgcctacatgaacgaccacctggtccaggtcgccgaggccattgccgacggc
gtcgaggtcctgggctacacctcctggggctgcatcgacctggtctcggcctccaccgcccagatgtccaagcgctacgggtt
catctacgtggaccgtgacgacggcggcaacggcaccctggcccgctaccgcaagaagtccttcggctggtaccgcgacgtca
tcgcctccaacggtgcctccctcgtgcctccggtgcaggaaccgccgcgggggtag
>gi|267| HB973106.1| EMBL| Actinomyces naeslundii hypothetical protein
atgaccgccacgtccactacttctaagagcaatccgaacttccccgacggcttcctgtggggcggggccaccgccgccaacca
gatcgagggcgcttacaacgaggacggcaagggcctgtccgtccaggacgtcatgcctcggggcatcatggcccaccccaccc
aggctcccacaccggataaccttcaagctcgaggcgatcgaccttctaccaccgcttacgccgaggacatctccctgttcgcg
gagatgggtttcaaggtcttccgcttctccatcgcctggagccgcatcttcccgctcggcgacgagaccgagcccaatgagga
aggactngccttctacgaccgggtcctcgacgagctcgagaagcacgggatcgagccactggtcaccatcagccactacgaga
ccccgctgcacctggcgcgcacctacgncggctggaccgaccgccgcctcatcggcttcttcgagcgctacgcccgcaccctg
ttcgagcgctatggcaagcgggtcaagtactggctcaccttcaacgagatcaactccgtgctccatgagcccttcctatctgg
gggcgtcgccacgcccaaggacaggcccccgagcaggacctctaccaggccatccaaaacgagctcgtcgcctccgcggccg
cgaccaggatcgcccatgagaccaaccccgacatccaggtcggctgcatgatcctggccgatcccacctacccgctcacccct
gatccccgggacgtgtgggcggccaagcaggcagagcgcgccaactacgccttcggagacctccacgtacgtggtgagtaccc
cggatacctgcggcggaccctgcgggacaagggcatcgagctggagatcaccgaggaggaccgcgtgctgctgcgggagcaca
ccgtcgacttcgtctccttctcctactacatgtncgtgtgcgagaccgtcacccagtcggccgaggccggccggggcaacctc
atgggcggcgtccccaatcccaccctcgaggcctccgagtggggatggcagatcgacccggcgggcctgcgcaccatcctgaa
cgactactgggaccgctggggcaagcctctgttcatcgtcgagaacggcctgggagccaaggacgtcctcgttgacggaccca
```

acggtcccacggtcgaggacgactaccgcatcgcctacatgaacgaccacctggtccaggtcgccgaggccattgccgacggc
gtcgaggtcctgggctacacctcctggggctgcatcgacctggtctcggcctccaccgcccagatgtccaagcgctacgggtt
catctacgtggaccgtgacgacggcggcaacggcaccctggcccgctaccgcaagaagtccttcggctggtaccgcgacgtca
tcgcctccaacggtgcctccctcgtgcctccggtgcaggaaccgccgcgggggtag
>gi|268| M21976.1| EMBL| Actinomyces naeslundii hypothetical protein
atgaagtacaacaccagcacgctggggcgtcgggctgcagccgcagccggtgtcctcaccctggccgtgcttggtctggcccc
catggctcaggccgagaacgccaaccacggagacatcaacaccgaggcgttaggctccctcaccatccacaagcacctcaacg
gagacggcaaccccatcggtgctcctgacggcaccgcttccaacgacgatggcaagggtgcaccggtctccggagtgcagttc
acggcctacgagatcaatgggatcgacctgaagacctcagagggatgggccaaggtcaacgccctgaccaataccggagcgat
tcccgacaacgcctgcgccaaccctgggcagccgacgctccccaactacaccttccgctcgagcagggtctcgggcgacactg
atcgcgacggagaggccaagatcgagagcctgccggtcaaggcctatctcgtgtgcgagaccaagactcctggcaacatcgtc
cagaaggccaagcccttcgtggtcacgattcctcacccgaacactgccgcgaaggcggatggaacgtggctctacgacgtcca
cgtctaccccaagaacgagaagatcgaagtcgccaagaccatcgaggaccagaggaacaacggctacatcgtcggatccaaag
tccgcttcccggtctcctcgacgctgccgaagctggatgacaattcctactacaagtactaccagttcaaggacaccctggac
aatcgtctgaagcaggtgaccgctacagacgtcactctcgggggacgcggctggatgagggcaccgactacaccctgggcac
cgacgggcagaccgtgaccgtgaccttcaaccagaatggtctgagcaagctcaagggcaatccgggtcagaagcttcaggcgg
tcttcgagggagtcgtctccgaagtcggcgacggcagcatcaacaacaccgcccagctcatctccgacacgacctacgccgag
cagccccccggcacctgagacgcctcccgccaaccccgacaaccctccgacgaccgaacaggtgacctcgaagtggggcgacct
gacgatcaagaaggtcgacggcaacgaccggtccggtgacaaggacggcctcaagggtgctgagttccagatctacaaggcca
aggacgcctacgcggacacatgctccccggaggccgatgggcagcccctcaccatcaacggtgagagcaccttcaccaccggt
gagggcggcacgatcaacttcaaggccctgttcgtctccgactccgtccaggacaccggtcgtgacaaccgggtggacgcgcc
ccaccgttgctacgtgctggtggagaccaaggctcctgccggctacgtgcttcccgcagacgccagtcgggccatcaccgtgg
agcctggggccggtgtgacgcagcaggtcgtgatcgacaacgtcaagcagtcggttccgggcctgcccctgaccggtgccaac
ggcatgctcatcctgaccgcctccggcgcggctctgctgatgatcgccgtcggctccgtcctcgtggcccgctaccgcgagcg
caagcggaaccgggacctcgccgcctga
>gi|269| AJ401093.1| EMBL| Actinomyces naeslundii hypothetical protein
atggagtcgacgcctaagagagggcggtcccagcgctgttggttcctcatgaattggagtgtgttccctttgctgtctcgtcg
tgccatcaaacctcgcctgcgtaggctcgtcgccggcgctgcggcactcgcctgcgtcgctgcggcgctcgtgggcactgagc
agccttctgccggggcggccccgagcgggctccccgccgacgcgcttccggcggtcttcgctcagggcggcaccagctggtac
cgggacaccatccagtggctccagtgggcggactacgaccagaacttcgccggccagaccaagccgaatgtccccgtgctcga
ctacgggcagcgcaggaccttcaccaactaccgcgacctgggagaggccggatacctggtcaccacctgcaatctgtccaatc
tcaagcaccttggtcaccagaagggcttccccgacgatctttcccgcggcccgctggtcgccactattccgggcacctgggcc
ggtgacatcctcgacaacctctacaacaccggtggcgccgggggctggagcgatggcagctcggagtggcacaacggcctgaa
gtacccggacaactacaccaaccacaaccggatggtcatcggactggccaacggctacgcctacaacggggacaagacctggg
acggcagggacaagaacgaccgccgggccaaccgcacgcccaccggcggatactcacgcatcagctttgacgtctcctgctcg
gcagctgtccaggcccccgacggcagctccacaccggtggccctcaacggcctggtcttcgccgacgcggaggcctccaaccc
cggcagcaccaatgagcccttcgacggcgagtggatccaggccgaggtccccagcaaccagcgagtcacctggcgcctcctgg
acggtgggcgctccagtaactgcacggacaccagaggacgtaccgggcggccggtcacgacccgcgcgtccttctccaacggg
aaccacaccctgaggctggacaatacggcacaggagtgcgtctaccagaacggcggaggttactccaagcccaacggcatcgg
tggtcccgcggtctccatgttcatggaaggggccaccagcgccaccatcaccatgcagggatcgggctactccgctgtcgccc
tcgggctggtcctggtcaccgacttcggtgacgcgccggtctcctacggcagcgcctccgccctcctgcagcccaggtgggac
ggtggtgaggtcacgaaccgcaacggctatgacctgttcggtggccgcctcatcaacaccacccgcatgcagcccaccgggcc
ctacctgggctcgggtactgacgccgagtcccagcagcgcttcagcgacggggccgacggcgacgacaacgacggctggtacg
gcaacgacgaggacggtgtctccatccccgcctcaggcattgagaccgccccgggacagcagtacaccaccaacgtccgctgc
ggcggccggggagcggtggccggttggatcgactggaaccacaacggcgtctttgacgccaccgagaagagcggtcagaccac
ctgtgacggctggggcaatgccacactgcgttggaccgtccccgaggacgtcgtgcgcagcatcgatggagaggccggatccc
agccccacacctacctacgggtgaggacagctgaggcaggagtgaacctcaagcccaccggcagcacgatgggcggcgaggtg
gaggactaccggatcgcggtgcgcgtccccacgatccggctcgtgaagaacgtccaggcccccttacaccgggcaggtcagggc
cctcggggccgaccagtggaccctcaaggcccagcagggcaacggtggggcggcgtcccagcaggtgaccggcaacgttgaca
ccgggatcaaggcggtgcgtccaggacagtacgtgctcagcgagtcctccacgaaccccaggcaccgggctacgaggcgagc
gcatggcagtgcgcccagaccccgggacgcgcggcagctggagcagctcgagcctgacaggatccaacatccaggtgcgggg
caccgatcgcctcacctgctgggtcaccaacacgaccaagcccggagcgctgtcctggaccaaggtcgaccaggacggtcaga
caccctgggcggtaccagctggaccctgaccggtcccggcgtccccgccggcaccgtcgtcgaggactgccaggcggcgggc
tgccgcggcggcgcctaccgggacaccaacccggctccgggtgtcttcgaggtcagtggactggcgtgggggagctactccgt
caccgagagaaccgcccccagtggataccagcgcctggacaagaccttggccttcaacgacgtctccggcgccaacctcaagg
ctgggctcaaggacaccaccggtgtcaccaggggagccgtgaccaaccagcgtctgaccggagccgtgtcctggaagaagcag
gacacgagcggacatgccctgggcggctcggagtggacgctcagcggccccggagtcccggccaggaccaccgtcaccgactg
cgtcatcgccggtggcaggggacggtgcccgggcggcccctacgccgatacggaccccgcggccggttccttcaccgtgaccg
gcctgccctgggacacgcgctcctactccctggtggagaagcgggctcccgcgggataccggcttgacacgacgagccgcagc
ttcgccatcaagcccaacgccctccagtactccttctcgaaggccttcaccaacgagaaggcggccacccccgcttgcccct
gacaggcgggcgcggtgcgcacatcttcctgatcgcaggagcggtcgtctccgctctggcgatctccaccggactcctcagga
ggcgccgtcactgcaacctcgactga
>gi|270| AY029505.1| EMBL| Actinomyces naeslundii beta-glucosidase
atgaccgccacgtccactacttctaagagcaatccgaacttccccgacggcttcctgtggggcggggccaccgccgccaacca
gatcgagggcgcttacaacgaggacggcaagggcctgtccgtccaggacgtcatgcctcggggcatcatggcccaccccaccc
aggctcccacaccggataaccttcaagctcgaggcgatcgaccttctaccaccgcttacgccgaggacatctccctgttcgcg
gagatgggtttcaaggtcttccgcttctccatcgcctggagccgcatcttcccgctcggcgacgagaccgagcccaatgagga
aggactngccttctacgaccgggtcctcgacgagctcgagaagcacgggatcgagccactggtcaccatcagccactacgaga

ccccgctgcacctggcgcgcacctacgncggctggaccgaccgccgcctcatcggcttcttcgagcgctacgcccgcaccctg
ttcgagcgctatggcaagcgggtcaagtactggctcaccttcaacgagatcaactccgtgctccatgagcccttcctatctgg
gggcgtcgccacgcccaaggacaggcccccgagcaggacctctaccaggccatccaaaacgagctcgtcgcctccgcggccg
cgaccaggatcgcccatgagaccaaccccgacatccaggtcggctgcatgatcctggccgatcccacctacccgctcacccct
gatccccgggacgtgtgggcggccaagcaggcagagcgcgccaactacgccttcggagacctccacgtacgtggtgagtaccc
cggatacctgcggcggaccctgcgggacaagggcatcgagctggagatcaccgaggaggaccgcgtgctgctgcgggagcaca
ccgtcgacttcgtctccttctcctactacatgtncgtgtgcgagaccgtcacccagtcggccgaggccggccggggcaacctc
atgggcggcgtccccaatcccaccctcgaggcctccgagtggggatggcagatcgacccggcgggcctgcgcaccatcctgaa
cgactactgggaccgctggggcaagcctctgttcatcgtcgagaacggcctgggagccaaggacgtcctcgttgacggaccca
acggtcccacggtcgaggacgactaccgcatcgcctacatgaacgaccacctggtccaggtcgccgaggccattgccgacggc
gtcgaggtcctgggctacacctcctggggctgcatcgacctggtctcggcctccaccgcccagatgtccaagcgctacgggtt
catctacgtggaccgtgacgacggcggcaacggcaccctggcccgctaccgcaagaagtccttcggctggtaccgcgacgtca
tcgcctccaacggtgcctccctcgtgcctccggtgcaggaaccgccgcgggggtag
>gi|271| HD030516.1| EMBL| Actinomyces naeslundii hypothetical protein
atgaccgccacgtccactacttctaagagcaatccgaacttccccgacggcttcctgtggggcggggccaccgccgccaacca
gatcgagggcgcttacaacgaggacggcaagggcctgtccgtccaggacgtcatgcctcggggcatcatggcccaccccaccc
aggctcccacaccggataaccttcaagctcgaggcgatcgaccttctaccaccgcttacgccgaggacatctccctgttcgcg
gagatgggtttcaaggtcttccgcttctccatcgcctggagccgcatcttcccgctcggcgacgagaccgagcccaatgagga
aggactngccttctacgaccgggtcctcgacgagctcgagaagcacgggatcgagccactggtcaccatcagccactacgaga
ccccgctgcacctggcgcgcacctacgncggctggaccgaccgccgcctcatcggcttcttcgagcgctacgcccgcaccctg
ttcgagcgctatggcaagcgggtcaagtactggctcaccttcaacgagatcaactccgtgctccatgagcccttcctatctgg
gggcgtcgccacgcccaaggacaggcccccgagcaggacctctaccaggccatccaaaacgagctcgtcgcctccgcggccg
cgaccaggatcgcccatgagaccaaccccgacatccaggtcggctgcatgatcctggccgatcccacctacccgctcacccct
gatccccgggacgtgtgggcggccaagcaggcagagcgcgccaactacgccttcggagacctccacgtacgtggtgagtaccc
cggatacctgcggcggaccctgcgggacaagggcatcgagctggagatcaccgaggaggaccgcgtgctgctgcgggagcaca
ccgtcgacttcgtctccttctcctactacatgtncgtgtgcgagaccgtcacccagtcggccgaggccggccggggcaacctc
atgggcggcgtccccaatcccaccctcgaggcctccgagtggggatggcagatcgacccggcgggcctgcgcaccatcctgaa
cgactactgggaccgctggggcaagcctctgttcatcgtcgagaacggcctgggagccaaggacgtcctcgttgacggaccca
acggtcccacggtcgaggacgactaccgcatcgcctacatgaacgaccacctggtccaggtcgccgaggccattgccgacggc
gtcgaggtcctgggctacacctcctggggctgcatcgacctggtctcggcctccaccgcccagatgtccaagcgctacgggtt
catctacgtggaccgtgacgacggcggcaacggcaccctggcccgctaccgcaagaagtccttcggctggtaccgcgacgtca
tcgcctccaacggtgcctccctcgtgcctccggtgcaggaaccgccgcgggggtag
>gi|272| U85709.1| EMBL| Actinomyces naeslundii hypothetical protein
atgccggccgcggtcgtcctggtggtgacgatcctcgtcatcggtctcgtgcgcggcagccaccggcgcgcggtgtgggccgt
cgtcgccgtggtcggggccaacctgaccacccagatcctcaagtaccagttcctgtggcggcccgacttcaacatcaccgagc
gctgggacaacgccaacaccctgccctcggggcacaccaccatggcggcctcggccgcggtcgccctcatcctgctgtccggg
cggcgctggcggccgctgtcggcctgggccgggggggctgttcaccgtcgccatgggctactcccaccctcgtgtgccagtggc
accgcccctccgacgtcctggccgggatcttcgtgcccgtgtcctggggcgccctcgccgtagccggcggggcctggcgcgcc
ccctggtcgctgcgcgaggggaccggggaggagagctccggtga
>gi|304| AX467596.1| EMBL| Actinomyces naeslundii hypothetical protein
atgaccgccacgtccactacttctaagagcaatccgaacttccccgacggcttcctgtggggcggggccaccgccgccaacca
gatcgagggcgcttacaacgaggacggcaagggcctgtccgtccaggacgtcatgcctcggggcatcatggcccaccccaccc
aggctcccacaccggataaccttcaagctcgaggcgatcgaccttctaccaccgcttacgccgaggacatctccctgttcgcg
gagatgggtttcaaggtcttccgcttctccatcgcctggagccgcatcttcccgctcggcgacgagaccgagcccaatgagga
aggactngccttctacgaccgggtcctcgacgagctcgagaagcacgggatcgagccactggtcaccatcagccactacgaga
ccccgctgcacctggcgcgcacctacgncggctggaccgaccgccgcctcatcggcttcttcgagcgctacgcccgcaccctg
ttcgagcgctatggcaagcgggtcaagtactggctcaccttcaacgagatcaactccgtgctccatgagcccttcctatctgg
gggcgtcgccacgcccaaggacaggcccccgagcaggacctctaccaggccatccaaaacgagctcgtcgcctccgcggccg
cgaccaggatcgcccatgagaccaaccccgacatccaggtcggctgcatgatcctggccgatcccacctacccgctcacccct
gatccccgggacgtgtgggcggccaagcaggcagagcgcgccaactacgccttcggagacctccacgtacgtggtgagtaccc
cggatacctgcggcggaccctgcgggacaagggcatcgagctggagatcaccgaggaggaccgcgtgctgctgcgggagcaca
ccgtcgacttcgtctccttctcctactacatgtncgtgtgcgagaccgtcacccagtcggccgaggccggccggggcaacctc
atgggcggcgtccccaatcccaccctcgaggcctccgagtggggatggcagatcgacccggcgggcctgcgcaccatcctgaa
cgactactgggaccgctggggcaagcctctgttcatcgtcgagaacggcctgggagccaaggacgtcctcgttgacggaccca
acggtcccacggtcgaggacgactaccgcatcgcctacatgaacgaccacctggtccaggtcgccgaggccattgccgacggc
gtcgaggtcctgggctacacctcctggggctgcatcgacctggtctcggcctccaccgcccagatgtccaagcgctacgggtt
catctacgtggaccgtgacgacggcggcaacggcaccctggcccgctaccgcaagaagtccttcggctggtaccgcgacgtca
tcgcctccaacggtgcctccctcgtgcctccggtgcaggaaccgccgcgggggtag
>gi|305| FB584124.1| EMBL| Actinomyces naeslundii hypothetical protein
atgaccgccacgtccactacttctaagagcaatccgaacttccccgacggcttcctgtggggcggggccaccgccgccaacca
gatcgagggcgcttacaacgaggacggcaagggcctgtccgtccaggacgtcatgcctcggggcatcatggcccaccccaccc
aggctcccacaccggataaccttcaagctcgaggcgatcgaccttctaccaccgcttacgccgaggacatctccctgttcgcg
gagatgggtttcaaggtcttccgcttctccatcgcctggagccgcatcttcccgctcggcgacgagaccgagcccaatgagga
aggactngccttctacgaccgggtcctcgacgagctcgagaagcacgggatcgagccactggtcaccatcagccactacgaga
ccccgctgcacctggcgcgcacctacgncggctggaccgaccgccgcctcatcggcttcttcgagcgctacgcccgcaccctg
ttcgagcgctatggcaagcgggtcaagtactggctcaccttcaacgagatcaactccgtgctccatgagcccttcctatctgg
gggcgtcgccacgcccaaggacaggcccccgagcaggacctctaccaggccatccaaaacgagctcgtcgcctccgcggccg
cgaccaggatcgcccatgagaccaaccccgacatccaggtcggctgcatgatcctggccgatcccacctacccgctcacccct

gatccccgggacgtgtgggcggccaagcaggcagagcgcgccaactacgccttcggagacctccacgtacgtggtgagtaccc
cggatacctgcggcggaccctgcgggacaagggcatcgagctggagatcaccgaggaggaccgcgtgctgctgcgggagcaca
ccgtcgacttcgtctccttctcctactacatgtncgtgtgcgagaccgtcacccagtcggccgaggccggccggggcaacctc
atgggcggcgtccccaatcccaccctcgaggcctccgagtggggatggcagatcgacccggcgggcctgcgcaccatcctgaa
cgactactgggaccgctggggcaagcctctgttcatcgtcgagaacggcctgggagccaaggacgtcctcgttgacggaccca
acggtcccacggtcgaggacgactaccgcatcgcctacatgaacgaccacctggtccaggtcgccgaggccattgccgacggc
gtcgaggtcctgggctacacctcctggggctgcatcgacctggtctcggcctccaccgcccagatgtccaagcgctacgggtt
catctacgtggaccgtgacgacggcggcaacggcaccctggcccgctaccgcaagaagtccttcggctggtaccgcgacgtca
tcgcctccaacggtgcctccctcgtgcctccggtgcaggaaccgccgcgggggtag