João Paulo Müller da Silva

Construção e análise de modelos topológicos de redes biológicas usando a ontologia MONET

João Paulo Müller da Silva

Construção e análise de modelos topológicos de redes biológicas usando a ontologia MONET

Dissertação submetida à avaliação como requisito parcial a obtenção do grau de mestre em computação aplicada

Orientador:
Ney Lemke

Universidade do Vale do Rio dos Sinos
Ciências Exatas e Tecnológicas
Programa Interdisciplinar de Pós-Graduação em Computação Aplicada

São Leopoldo

2006

Dedico... aos meus pais, Mário e Maria Helena.

AGRADECIMENTOS

Aos meus pais, por estarem sempre ao meu lado.

Ao meu orientador Prof. Dr. Ney Lemke, pelos ensinamentos, e pela orientação neste duro trabalho.

A minha co-orientadora Profa. Dra. Renata Vieira, pela co-orientação e auxílio no assunto de ontologias e pelas correções do texto.

Aos colegas de laboratório que fizeram das horas intermináveis, sempre um ambiente divertido e descontraído, onde se tem uma grande família e sem demérito aos outros, é o melhor laboratório para se fazer pesquisa.

Ao Eduardo Battistella, por iniciar este trabalho, pelo aprendizado e permitir que hoje se tornasse uma dissertação de mestrado.

Ao José Guilherme, pelo auxílio na parte de normalização dos dados e pela geração do arquivo OWL da ontologia.

A Norma e a Meg, pelos ensinamentos e discussões na parte biológica deste trabalho, assim como nas correções do texto.

A Rejane Weissheimer, secretária do mestrado pela disponibilidade de realizar as questões burocráticas.

Aos meus amigos que sempre me ajudaram quando precisei e me deram força para continuar a luta, e sem esquecer dos companheiros inseparáveis nas festas e nos agitos na noite pelotense.

E a todos aqueles que participaram diretamente ou indiretamente na realização deste trabalho.

A HP, pela apoio financeiro para a realização deste estudo.

A todos vocês meu muito obrigado!!!

Meu objetivo é vencer, e, para isso, preciso dar sempre o melhor de mim.
Ayrton Senna da Silva

RESUMO

Um dos mais importantes desafios para a biologia pós-genômica é entender a estrutura e o comportamento das interações moleculares complexas que controlam o comportamento celular. Para tanto é essencial à integração dos dados biológicos referentes a estas interações armazenados em diversos bancos de dados. Este é um problema difícil, pois estes dados estão disponíveis em bancos de dados públicos espalhados geograficamente na rede mundial de computadores e cada um destes possui um sistema diferente de gerenciamento, formato e visão de como representar os dados. Os principais problemas para a realização desta tarefa são: a necessidade de se desenvolver e aplicar *parsers* para cada banco de dados e a ausência de um vocabulário unificado. Como uma alternativa para facilitar estes problemas, este trabalho propõe a ontologia MONET (*Molecular Network Ontology*) que tem como objetivo ser um modelo integrado para a *rede de redes* que existe dentro da célula. Tal visão integrada ajuda a entender as interações de larga escala responsáveis pelo comportamento da célula, e permite a predição do comportamento celular que pode ser experimentalmente testado. A ontologia engloba o metabolismo e a interação proteína-proteína para os organismos procariotos e eucariotos, e regulação gênica para seres procariotos. Como resultado, este trabalho proporcionou uma padronização dos termos usados nas três áreas abarcadas pela ontologia e a população da ontologia com dados referentes à bactéria *E. coli*. A partir desta integração construímos a rede integrada da bactéria, e com o conhecimento representado realizamos experimentos de aprendizado de máquina para a predição da essencialidade de um gene com base na análise topológica da rede de interações, utilizando o algoritmo J48, obteve-se uma cobertura de $85,7\%$ para o melhor resultado. Além disto, caracterizamos a rede integrada da *E. coli*, como uma rede livre de escala e hierárquica.

Palavras-chave: integração de dados, ontologias, metabolismo, regulação gênica e in-

teração proteína-proteína.

ABSTRACT

One of the most important challenges for biology in the post-genomic is to understand the structure and behavior of the molecular interactions that controls cell behavior. Therefore is essential to integrate biological data concerning these interactions, which are stored in different databases. The integration task is difficult because these data are distributed in public databases on the world wide web and each database has different management systems, formats and views of how to represent biological data. The two main problems involved here are the difficulty in parsing the data when dealing with heterogeneous flat file formats and the inconsistencies due to the absence of an unified vocabulary. As an alternative to facilitate these problems this work proposes MONET (*the Molecular Network*) ontology, an integration model for the unifying of different molecular networks that exist inside the cell. Such integrated view facilitates the understanding of the large-scale interactions responsible for the behavior of the cells, and the prediction of cellular behavior that can be tested experimentally. The ontology integrates metabolic data and protein-protein interaction for prokaryote and eukaryote organisms and also transcriptional-regulatory data only for prokaryote organisms. As result, this work provides a standardization of the terms used in these areas of the ontology and the population of the ontology with data referring to *E. coli*. Using these data we build a network model for *E. coli* molecular interactions. We characterized the resulting graph as an hierarchical free-scale network and by applying machine learning techniques we could predict gene essentiality with 85.7 recall.

Keywords: Data integration, ontology, metabolic pathways, regulation and protein-protein interaction.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

A - *adenina*

ARPA - *Advanced Research Projetcs Agency*

AUG - *metionina*

BIND - *Biomolecular Interaction Network Database*

BioPAX - *Biological Pathways Exchange Format*

C - *citosina*

COOH - *grupo carboxila*

DAML+OIL - *Darpa Markup Language*

DNA - *ácido desoxiribonucléico*

DW - *Data Warehouse*

G - *guanina*

GO - *Gene Ontology*

H - *hidrogênio*

H_2O - *molécula de água*

HUPO - *Human Proteome Organization*

IA - *Inteligência Artifical*

KEGG - *Kyoto Encyclopedia of Genes and Genomes*

KIF - *Knowledge Interchange Format*

MGED - *Microarray Gene Expression Data*

MONET - *MOlecular NETwork*

mRNA - *RNA mensageiro*

ODE - *Ontology Design Environment*

ORF - *Open Reading Frame*

OWL - *Ontology Web Language*

P - *fosfato*

PSI-MI - *Proteomics Standards Initiate*

R - *cadeia lateral*

RDF - *Resource Description Format*

RDF-S - *RDF-Schema*

RNA - *ácido ribonucléico*

RNAp - *RNA polimerase*

SBML - *Systems Biology Markup Language*

SGBD - *Sistema Gerenciador de Banco de Dados*

SO - *Sequence Ontology Project*

T - *timina*

U - *uracila*

UAA - *códon de terminação*

UAG - *códon de terminação*

UGA - *códon de terminação*

W3C - *World Wide Web Consortium*

WEKA - *Waikato Environment for Knowledge Analysis*

XML - *eXtensible Markup Language*

SUMÁRIO

1 INTRODUÇÃO

A bioinformática é a área da ciência onde a Biologia, a Ciência da Computação e Tecnologia da Informação se unem para compor uma única disciplina com o objetivo de tornar possível a extração de conhecimento relevante a partir de informações biológicas. No começo da *revolução genômica*, o interesse da bioinformática foi a criação e manutenção de bancos de dados para armazenar informações biológicas, como, por exemplo, seqüências de nucleotídeos e aminoácidos.

A área evoluiu e hoje abarca a análise e interpretação de vários tipos de dados, incluindo seqüências de nucleotídeos e aminoácidos, domínios e estrutura de proteínas, regulação gênica, redes metabólicas, interação proteína-proteína, entre outras.

Neste contexto, um exemplo da junção da Biologia com a Informática é a integração de dados biológicos, onde de um lado se tem os dados ligados à Biologia, que são fruto de diversos experimentos como o Projeto Genoma, Microarranjos e Dois-híbridos e do outro, as técnicas de integração pertencentes à Informática. Este procedimento é complexo, pois os dados estão distribuídos geograficamente na Internet e armazenados em diversos bancos de dados. Um dos problemas enfrentados é que os bancos de dados possuem diversos sistemas de gerenciamento, assim como formato e visões de como representar seus dados. Além disso, existe também o problema de acesso aos dados, porque alguns bancos de dados são acessíveis por um único mecanismo de consulta através de uma interface *web* ou disponibilizam seus dados em arquivos texto. Relacionado a estes problemas, encontra-se o problema de redundância de informação por parte dos bancos de dados, visto que alguns repositórios apresentam a mesma informação só que de maneira diferente,

como, por exemplo: o banco de dados KEGG *(http://www.genome.jp/kegg)* apresenta a informação de ORF como *dbget* e o nome de gene como *gene*, enquanto que o banco de dados NBCI *(http://www.ncbi.ncm.nih.gov/)* apresenta o código da ORF como *synonym* e o nome de gene como *name*.

No campo da Bioinformática, as ontologias se apresentam como cruciais para a manutenção da coerência de uma larga coleção de conceitos complexos e seus relacionamentos (BAKER et al., 1999). Este trabalho propõe a ontologia MONET (*Molecular Network*).

A ontologia MONET (BATTISTELLA et al., 2004) (BATTISTELLA et al., 2005) é um modelo integrado para a *rede de redes* (metabolismo, regulação gênica e interação proteína-proteína) que existe dentro da célula (BARABáSI; OLTVAI, 2004). Tal visão integrada ajuda a entender as interações de larga escala responsáveis pelo comportamento da célula, para predição do comportamento celular que pode ser experimentalmente testado (IDEKER et al., 2001) e gerar hipóteses testáveis.

A ontologia abarca o metabolismo, regulação gênica e interação proteína-proteína, sendo que para o metabolismo e interação proteína-proteína engloba organismos procariotos e eucariotos, enquanto que a regulação gênica compreende apenas os organismos eucariotos, através de uma visão que permite estabilizar um modelo capaz de minimizar a redundância e inconsistência de dados.

O objetivo geral deste trabalho, é a revisão da especificação da modelagem da ontologia MONET, iniciada em (BATTISTELLA et al., 2004) (BATTISTELLA et al., 2005), e em um segundo momento popular e utilizar a ontologia para a construção da rede integrada da *E. coli* e a predição da essencialidade de uma enzima. Além disso, têm-se os seguintes objetivos.

- avaliação da ontologia.
- integração dos dados biológicos.

- criação das instâncias para a ontologia através de consultas a bases heterogêneas.
- construção e análise dos modelos topológicos.
- construção da rede integrada da *E. coli*.
- predição da essencialidade de uma enzima através de técnicas de aprendizado de máquina.

O texto desta dissertação encontra-se organizado da seguinte maneira: o Capítulo 2 apresenta uma revisão bibliográfica sobre os conceitos de biologia molecular necessários para que este trabalho possa ser melhor compreendido. O Capítulo 3 aborda o tema das redes biológicas, assim como o modelo de grafos aleatórios e apresenta os principais parâmetros para sua descrição. No Capítulo 4 são apresentadas as abordagens de integração de dados existentes, assim como as suas respectivas arquiteturas e ainda salienta as suas vantagens e desvantagens. Neste capítulo também é abordado o tema ontologias, desde os seus conceitos, seus tipos e formalismos, além de uma revisão sobre as ontologias na área de bioinformática. O Capítulo 5 apresenta a ontologia MONET. No Capítulo 6 são apresentados os resultados obtidos por este trabalho, através das aplicações desenvolvidas. E por fim o Capítulo 7 apresenta as conclusões finais.

2 BIOLOGIA MOLECULAR

Neste capítulo serão abordados os conceitos de biologia molecular utilizados neste trabalho, para que o mesmo possa ser mais bem compreendido. A redação deste capítulo está baseada em (LODISH, 1999) e (LEWIN, 2001).

2.1 A célula de organismos procariotos e eucariotos

A célula é a responsável pelos processos metabólicos que ocorrem em todos os seres vivos, pois carrega consigo o material genético (DNA). A propriedade fundamental da célula está na sua capacidade de replicar-se, gerando assim células descendentes contendo cópias do seu material genético. Isto é resultado de uma série de processos metabólicos desencadeados dentro dela.

De acordo com o domínio ao qual a célula pertence (procariotos e eucariotos), a mesma é constituída de forma diferente. No caso de seres procariotos, que possuem uma única célula, como, as bactérias, a célula apresenta um único compartimento composto pela membrana plasmática e pelo citoplasma, conforme ilustra a Figura 1. Já no caso de seres eucariotos, que apresentam uma ou mais células, como, os seres humanos, a célula é constituída pela membrana plasmática, citoplasma e núcleo, de acordo com a Figura 2. As células de eucariotos, diferentemente das de procariotos, possuem regiões bem definidas, separadas do citoplasma por membranas internas formando assim compartimentos, denominados de organelas (exemplos: mitocôndria, retículo endoplasmático, etc.), as quais realizam funções especializadas. O material genético dos seres procariotos está localizado

no citoplasma, enquanto que o dos seres eucariotos encontra-se no núcleo.

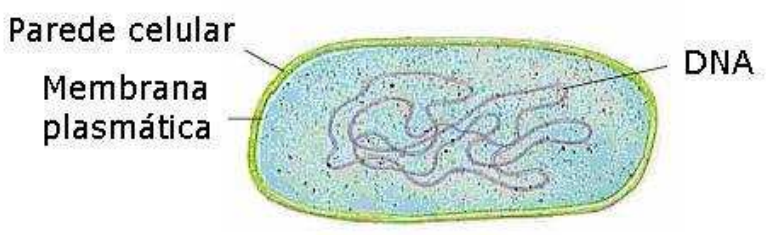

Figura 1: Representação de uma célula dos organismos procariotos. Como exemplo deste tipo de organismo, tem-se a *Escherichia coli*.

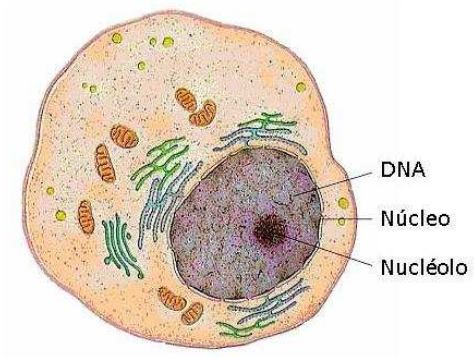

Figura 2: Representação de uma célula dos organismos eucariotos. Como exemplos deste tipo de organismo, tem-se a *Saccharomyces cerevisiae* e os seres humanos.

2.2 DNA e RNA

Na natureza há dois tipos de ácidos nucléicos: DNA (ácido desoxiribonucléico) e RNA (ácido ribonucléico). O DNA é uma molécula composta por duas cadeias ou fitas que se entrelaçam em torno do mesmo eixo formando uma dupla hélice (LEWIN, 2001). Esta molécula armazena as informações relativas ao desenvolvimento e divisão da célula. Por sua vez, a molécula de RNA possui uma única fita e esta é uma intermediária na produção (síntese) de proteínas. Esta molécula é produzida a partir de um gene que foi "expresso" e contém a informação que será usada para construir a cadeia de aminoácidos produzindo, na seqüência, as proteínas. Analogamente a um sistema de comunicação,

essas informações são mantidas dentro da célula em forma de código, que no caso é denominado código genético. Em sua estrutura, os ácidos nucléicos DNA e RNA podem ser vistos como uma cadeia linear composta de unidades químicas chamadas nucleotídeos.

Um nucleotídeo é um composto químico formado por uma base nitrogenada, um grupo fosfato (P) e uma pentose (molécula de açúcar com cinco carbonos). Os nucleotídeos são ricos em energia e direcionam os processos metabólicos no interior das células.

A base nitrogenada é quem caracteriza cada um dos nucleotídeos, sendo eles: adenina (A), citosina (C), guanina (G), timina (T) e uracila (U). As duas primeiras bases, A e C, são denominadas purinas e as outras três, G, T e U, são chamadas pirimidinas. No DNA encontram-se as bases A, C, G e T, enquanto que no RNA encontram-se as bases A, C, G e U.

Cada nucleotídeo de uma fita de DNA se liga ao complementar na outra fita, conforme a regra de pareamento, que é construída da seguinte forma: a base A liga-se com a T e base C liga-se com a G, e da seguinte forma para o RNA: a base A liga-se com a U e base C liga-se com a G. A Figura 3 apresenta a forma de ligação dos nucleotídeos entre as duas fitas do DNA.

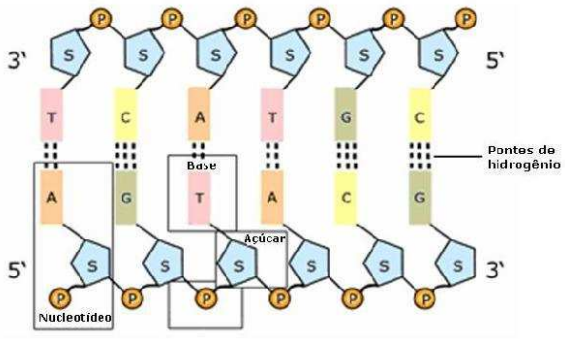

Figura 3: Ligação entre as bases da dupla fita do DNA conforme a regra de pareamento: a base A liga-se com a T por meio de duas pontes de hidrogênio, enquanto que a base C liga-se com a G através de três pontes de hidrogênio.

Com relação a ligação nas bases do DNA, a ligação AT é conceituada como fraca por ocorrer através de duas pontes de hidrogênio, enquanto que a ligação CG é forte em função das suas três pontes de hidrogênio.

Esta estrutura de fitas á antiparalela e suas ligações ocorrem no sentido $5' \rightarrow 3'$. A ligação entre os nucleotídeos (ligações fosfodiéster) de uma cadeia linear (ou seja, entre os nucleotídeos de uma mesma fita de DNA) é feita entre o grupo químico hidroxil ligado ao terceiro carbono da pentose de um nucleotídeo e o fosfato do nucleotídeo seguinte ligado ao carbono cinco da pentose, conforme ilustra a Figura 4. Por convenção, as seqüências são representadas na orientação $5' \rightarrow 3'$.

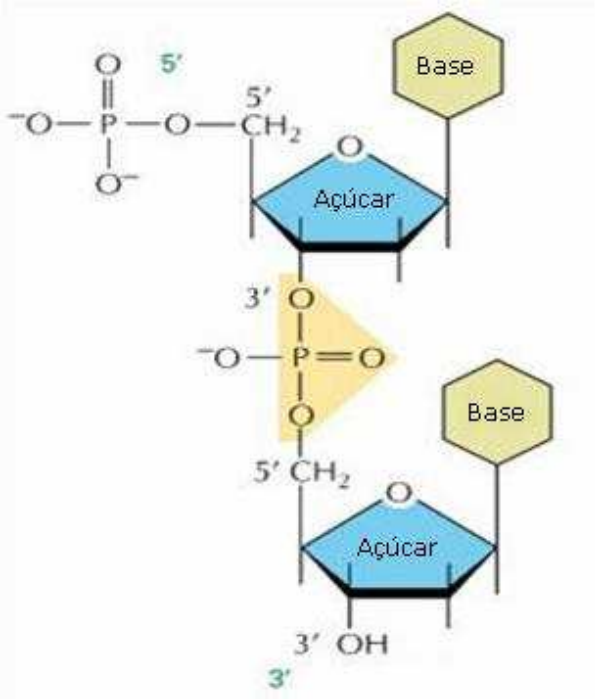

Figura 4: Ligação entre os nucleotídeos de uma mesma fita de DNA.

2.3 Genoma, gene, orf e operon

A informação genética contida na molécula de DNA na célula de um ser vivo, procariotos ou eucariotos, é chamada genoma. Já os genes são os responsáveis pela determinação dos traços hereditários de todos os seres vivos, como, por exemplo, a cor dos

olhos em um ser humano. O gene compreende uma região nucleotídica presente no DNA, que comporta a informação necessária para a produção de uma unidade biomolecular específica (onde a maioria são proteínas), a qual realiza alguma função dentro da célula. A estrutura de um gene apresenta uma parte denominada região promotora, que apresenta uma função na regulação dos genes. O promotor é um segmento de DNA cuja seqüência de nucleotídeos é conservada, o que explica o fato de como a enzima RNA polimerase (RNAp) reconhece o lugar onde deve se ligar. Uma vez ligada à região promotora inicia o processo de síntese da molécula de mRNA (RNA mensageiro). Além dos promotores, os genes possuem na sua estrutura uma região codificadora e uma região terminadora, onde a primeira contém a informação necessária para a fabricação de uma proteína e a segunda é o segmento que sinaliza o término do processo de síntese da molécula de mRNA. A unidade de transcrição (molécula de mRNA) por sua vez, é uma seqüência do DNA transcrito pela RNAp que estende-se desde o primeiro nucleotídeo transcrito até a região terminadora.

A molécula de mRNA transcrita será posteriormente usada para a montagem das proteínas. O processo de montagem de uma proteína implica na decodificação de trincas ou códons de nucleotídeos do mRNA, ou seja, cada conjunto de três nucleotídeos específica um determinado aminoácido. Existem trincas que especificam cada um dos vinte aminoácidos presentes na natureza e trincas que simplesmente significam códigos de término na montagem das proteínas. Via de regra, a construção de uma proteína inicia a partir de um códon de iniciação AUG, que específica o aminoácido *metionina* e conseqüentemente este será o primeiro na seqüência de aminoácidos que farão parte da proteína, e termina em um dos três possíveis códons de terminação que são: UAG, UGA e UAA.

Um outro aspecto relevante a produção das proteínas, é o conceito de ORF (*Open Reading Frame*). Uma fase de leitura que inicie com um códon de início e que não seja encerrada prematuramente por um códon de terminação é denominada ORF. Uma seqüência de DNA tem diferentes possíveis fases de leitura (*frames*) e a determinação da fase de leitura correta segue determinados critérios. Depois de encerrada, a seqüência de DNA pode então ser traduzida para seus aminoácidos correspondentes. Embora seja

comum na prática o uso dos termos gene e ORF indistintamente, é importante frisar sua diferença. Toda região codificante de um gene é uma ORF, entretanto nem toda ORF é um gene.

Em seres procariotos, a grande maioria dos genes estão organizados numa estrutura chamada *operon*, que se refere a uma seqüência de genes adjacentes sob o controle de um mesmo promotor, conforme apresenta a Figura 5. O mesmo não acontece em organismos eucariotos. Neste tipo de organismo, cada gene tem sua própria região promotora.

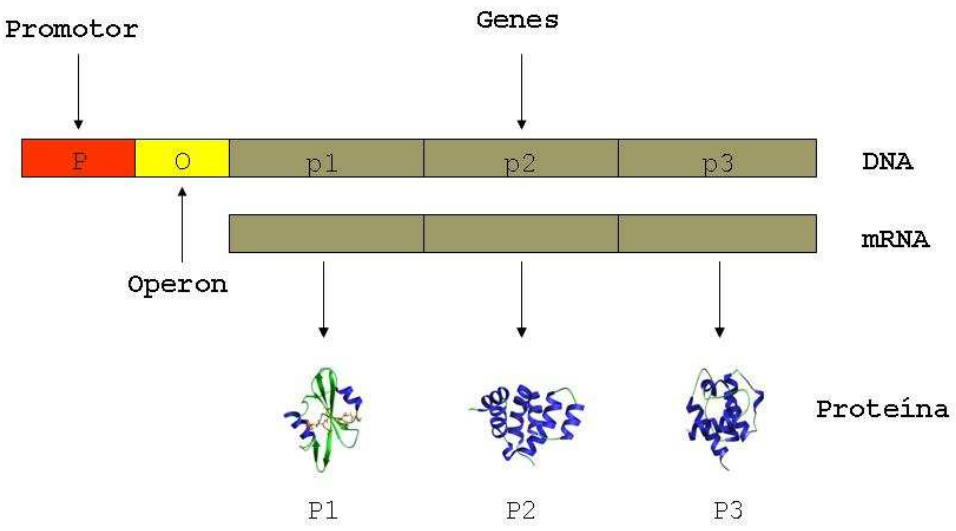

Figura 5: Representação ilustrativa da estrutura de um *operon*, cuja existência ocorre apenas em organismos procariotos.

2.4 Expressão gênica

Conforme apresentado na Seção 2.3, um gene é um segmento de DNA que contém a informação genética. Essa informação torna-se disponível para a célula através de um processo denominado expressão gênica. Quando isso acontece, uma cópia do gene é decodificada em uma molécula de mRNA, que por sua vez é utilizada na síntese (produção) de uma proteína, através de um processo chamado tradução. Este fluxo de informação é considerado o dogma central da biologia molecular.

2.4.1 Regulação da expressão gênica

Se todos os genes funcionassem de forma contínua isso representaria um gasto de energia absurdo para a célula. Para evitar isso, a célula precisa ativar ou desativar determinados genes ao longo do seu ciclo de vida. Genes que permanecem ativos continuamente são denominados genes de expressão constitutiva. O produto desses genes é necessário durante toda a existência da célula, como, por exemplo, os genes que comandam a síntese dos componentes dos ribossomos. Outros genes, só são ativados em circunstâncias muito especiais, onde seus produtos são necessários, sendo que a expressão destes genes, é, portanto, regulada. A capacidade da célula em regular quais dos genes são expressos é referida como regulação da expressão gênica e os diferentes tipos celulares em um organismo multicelular surgem porque diferentes genes estão sendo expressos em suas células. A regulação da expressão gênica pode ocorrer em qualquer etapa do fluxo de informação genética do DNA para a proteína e varia de acordo com o tipo de organismo.

em organismos procariotos : os genes são ativados ou desativados de acordo com as influências do meio ambiente. O genoma pequeno, sem *introns* (pequenos pedaços de um gene que são transcritos, porém não participam do processo de tradução), com pouco DNA extragênico e com seus genes organizados em *operons*, resultam em respostas mais rápidas, promovendo uma adaptação destes organismos mais imediata ao meio em constantes mudanças. Em bactérias, como, por exemplo, a *E. coli*, a atividade gênica é controlada, predominantemente, em nível de transcrição (momento em que a enzima RNAp se liga ao promotor para decodificar um gene em uma molécula de mRNA), onde as proteínas regulatórias se ligam a um sítio específico no DNA próximo ao promotor do(s) gene(s) que irá controlar;

em organismos eucariotos : há uma maior complexidade nos mecanismos de controle devido ao grande conjunto gênico. Várias proteínas e trechos extragênicos do DNA estão envolvidos na regulação dos genes. Existem interações entre genes e o meio e também entre setores localizados a certas distâncias dentro do genoma que podem

atuar como ativadores ou inibidores da transcrição de determinados genes. Nos organismos eucariotos, o produto da transcrição RNAm (também conhecido como transcrito primário) sofre modificações antes de ser traduzido. Este transcrito sofre mudanças e, somente após estas mudanças, este RNAm agora "maduro" pode ser transportado do núcleo da célula para o citoplasma onde ocorrerá o processo de tradução. Uma importante mudança sofrida pelo RNAm antes de sair do núcleo é a retirada dos *introns*.

2.5 Proteínas

As proteínas localizam-se no interior das células e são moléculas que possuem função específica (ou atividade biológica) no organismo dos seres vivos. Essas funções incluem catálise enzimática, função estrutural, regulação, dentre outras.

Uma proteína é formada por unidades denominadas aminoácidos, os quais se ligam de forma linear formando assim uma cadeia polipeptídica. Um aminoácido é constituído de um grupo central (carbono α), um hidrogênio (H), um grupo carboxila (COOH), um grupo amino (H_2O) e uma cadeia lateral (R), conforme a representação ilustrada na Figura 6.

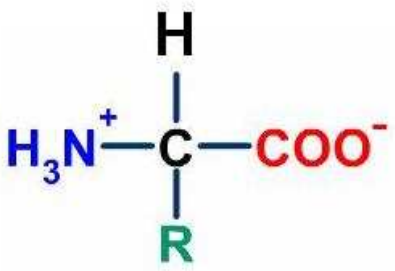

Figura 6: Representação da estrutura de um aminóacido. A cadeia lateral (R) distingue cada um dos vinte aminoácidos existentes na natureza.

A cadeia lateral (R) distingue cada um dos vinte aminoácidos presentes na natureza. A cadeia de aminoácidos é formada por meio de ligações peptídicas, pela união

do grupo carboxila do primeiro aminoácido com o grupo amino do segundo aminoácido, havendo assim a liberação de uma molécula de água (H_2O). A existência de vinte diferentes aminoácidos propicia uma grande e complexa variedade de formas (estrutura tridimensional) às proteínas. A forma das proteínas determina a sua ligação com outras moléculas, ou seja, aquelas que possuem uma forma complementar à da proteína.

2.6 Interação proteína-proteína

Proteínas podem interagir fisicamente umas com as outras, através de sítios de ligação, chamados domínios. Uma mesma proteína pode participar de diversas interações, porém esta interação ocorre em momentos distintos. Via de regra, proteínas que interagem estão relacionadas de alguma forma aos mesmos processos biológicos, o que permite a proposição da função hipotética de uma proteína desconhecida (DENG et al., 2002).

Há uma técnica experimental conhecida como dois-híbridos que se baseia na detecção da interação entre uma proteína desconhecida, chamada de *isca* e uma proteína conhecida, denominada de *presa*. Nesta técnica, a interação é facilmente detectada pela expressão de um gene que será ativado pela ligação das proteínas interagentes na sua região regulatória (seqüência específica do DNA que regula este gene). A partir desta detecção, é possível predizer a função da proteína que inicialmente tinha sua função desconhecida e também construir a rede de interação de proteínas de uma célula ou de um organismo (SAFFI; REVERS; HENRIQUES, 2001).

2.7 Metabolismo

O metabolismo é uma rede complexa de processos físico-químicos, que permite a sobrevivência e reprodução das células. A maioria desses processos são catalisados por enzimas que atuam como eficientes catalisadores e reagem seletivamente nos compostos definidos como substratos.

Considere a reação exemplo a seguir:

$$n_a A + n_b B \xrightarrow{k} n_c C + n_d D \tag{2.1}$$

Tem-se neste exemplo, n_a moléculas da espécie A que reagem seletivamente com n_b moléculas da espécie B para transformar em n_c moléculas da espécie C e n_d moléculas da espécie D. Neste exemplo, tem-se quatro espécies envolvidas, porém o número de espécies em uma reação é variável. A seta ($\rightarrow$) indica o sentido em que a reação ocorre. Neste exemplo, a reação ocorre no sentido $esquerda \rightarrow direita$. Entretanto, existem reações que ocorrem no sentido $esquerda \leftarrow direita$. Nestes dois casos as reações são consideradas irreversíveis. Existem também situações em que as reações ocorrem em ambos os sentidos $esquerda \rightleftharpoons direita$, neste caso são consideradas reversíveis. Considerando o exemplo, ao lado esquerdo estão os compostos demoninados substratos e ao lado direito encontram-se os produtos, que são substâncias químicas formadas durante uma reação. O n indica o número de moléculas ou coeficiente estequiométrico de cada espécie, enquanto que o k indica a taxa em que a reação ocorre. Dentro de uma reação química podem ser encontrados dois tipos de substâncias, chamadas de ativadores e inibidores. Enquanto o ativador é uma substância, com exceção do catalisador e de um dos substratos, que aumenta a taxa de uma reação catalisada, sem que esta seja consumida. Por outro lado, tem-se o inibidor que faz exatamente o oposto do ativador, é uma substância molecular que interfere, diminuindo ou parando uma reação química (LEHNINGER; COX; NELSON, 2000).

Além destas substâncias em uma reação química, existem também as enzimas que são importantes na regulação de processos biológicos e atuam como ativadores ou inibidores de uma reação. Para entender o seu papel, é necessário estudar a cinética química subjacente que prevê o comportamento temporal dos seus reagentes e as suas condições de influência (MURRAY, 1993).

O estudo de rotas bioquímicas através de modelos cinéticos e simulação computacional depende do conhecimento das taxas das reações e objetiva entender a dinâmica

de uma célula viva, em termos de interação entre seus componentes celulares (KIERZEK, 2002).

Nos organismos, as reações estão organizadas em módulos, chamados de rotas metabólicas com funções anabólicas e catabólicas específicas. As reações anabólicas sintetizam moléculas complexas e consomem grande quantidade de energia, já as reações catabólicas quebram as moléculas complexas em moléculas mais simples e precursoras e liberam energia.

Neste trabalho estamos interessados em reações bioquímicas envolvidas no metabolismo de pequenas moléculas (como, por exemplo, Pyrophospate, D-Fructose, ATP, glicose, água, dentre inúmeras outras moléculas), as quais representam um subconjunto do metabolismo completo, que exclui as reações envolvidas no processo de replicação do DNA e síntese de proteínas (ARITA, 2004).

3 REDES BIOLÓGICAS

Este capítulo aborda o tema redes biológicas, assim como o modelo de grafos aleatórios e apresenta os principais parâmetros para sua descrição.

As redes encontram-se presentes em nossas vidas hoje em dia. Somos influenciados por elas a cada momento. Uma destas é a Internet (rede mundial de computadores). Ela se torna presente em boa parte de nossas vidas, como, por exemplo, auxiliando-nos em certos aspectos como, compra de produtos, uso em pesquisa acadêmica ou ainda em linhas gerais e também comunicando pessoas através de programas, tais como, Skype, MSN e ICQ, além de muitas outras atividades. Outros exemplos de redes que podem ser comentadas são: a telefonia móvel, as redes sociais, ecológicas, interação intracelular (HALLINAN, 2004) e a grande explosão do momento o ORKUT (*http://www.orkut.com*), uma rede de relacionamentos.

Conforme apresentado por (HALLINAN, 2004), as redes envolvem uma variedade de contextos e a partir destes, estudos revelam que estes contextos apresentam características dinâmicas e topológicas comuns, o que sugere que estes pontos em comum envolvam processos similares quanto à operação e desenvolvimento da rede.

Muitos estudos têm sido desenvolvidos explorando o contexto de redes e destes uma questão surge: *Por que a estrutura de uma rede deve ser caracterizada?* É pelo simples fato de que a estrutura de uma rede está diretamente associada a sua função (STROGATZ, 2001). Um exemplo, a topologia de redes sociais afeta a propagação da informação. Dentro deste contexto, o interesse nas redes é um movimento grande em pesquisa sobre sistemas complexos, visto que as redes constituem o esqueleto destes sistemas.

Entretanto, as redes são difíceis de entender pelas razões descritas abaixo (STROGATZ, 2001):

complexidade estrutural : devido a sua forma estrutural;

evolução da rede : refere-se a sua evolução, por exemplo, ao longo das espécies, novas enzimas vão surgindo, gerando novos nodos nas redes metabólicas;

diversidade na conexão : as ligações (conexões) entre os nodos podem ser de tamanhos distintos;

complexidade dinâmica : nodos podem ser sistemas dinâmicos não-lineares com variação ao longo do tempo;

diversidade nos nodos : nodos podem ser de diferentes tipos;

meta-complicações : complicações na rede afetam o seu comportamento, como, por exemplo, seu crescimento afeta sua evolução, que por sua vez, afeta seu comportamento.

Com a variedade de contextos apresentados pelas redes, um dos que pode ser melhor explorado é o das redes pertencentes ao domínio da biologia molecular.

Neste contexto de rede é intrínseco que as funções biológicas não sejam atribuídas a uma molécula individual mas sim, que sejam caracterizadas por complexas interações entre os componentes celulares associados, tais como, interação entre proteínas, DNA e RNA e pequenos metabólitos como água, glicose, ATP, ADP, dentre outros. A partir destes processos de interação é possível entender a estrutura e as interações complexas que contribuem para a composição da estrutura e para a função de uma célula viva (BARABáSI; OLTVAI, 2004).

O processo de caracterizar e identificar as características pertinentes em nível de organização biológica é a chave de entrada para estudos na biologia pós-genômica (RAVASZ et al., 2002).

A utilização de experimentos de microarranjos permite a interrogação simultânea do estado dos componentes celulares em um dado momento. No entanto, outros tipos de ensaios genéticos também podem ser utilizados. Os microarranjos de DNA revolucionaram a maneira de analisar a expressão gênica, permitindo que os produtos de RNA de milhares de genes sejam monitorados de uma só vez, provendo informações isoladas e detalhadas dos padrões dinâmicos da expressão gênica que fundamentam os processos celulares complexos. Os experimentos de microarranjos de DNA nada mais são do que lâminas de microscópio crivadas com uma grande seqüência de fragmentos de DNA, cada uma contendo uma seqüência de nucleotídeos que serve como sonda para um gene específico, sendo a seqüência exata e a posição de cada sonda no *chip* conhecidas. Dessa maneira, qualquer fragmento de nucleotídeo que se torne híbrido com uma sonda no arranjo pode ser identificado como produto de um gene específico, simplesmente, detectando-se a posição à qual ela se liga.

A partir de experimentos como este, várias redes se apresentam, como, por exemplo, interação proteína-proteína, metabólicas, de sinalização e transcrição regulatória. Porém, nenhuma destas redes é independente, formando a *rede de redes*, que são as responsáveis pelo comportamento da célula (BARABáSI; OLTVAI, 2004).

O comportamento de muitos sistemas complexos surge da atividade de comunicação entre seus componentes através de interações. Em um nível mais elevado de abstração, os componentes podem ser reduzidos a uma série de nodos interligados através de ligações que representam as interações entre quaisquer dois destes componentes. Os nodos e as ligações (conexões) juntos formam uma rede, ou em uma linguagem mais formal, um grafo.

Um grafo é definido como sendo um par $g = (v, e)$, onde v é um conjunto arbitrário finito ($v \neq 0$) e e um subconjunto com no mínimo dois elementos de v. Os elementos do conjunto v são denominados vértices, enquanto que os do conjunto e são chamados de arestas.

A interligação dos nodos em um modelo de grafos ocorre quando uma aresta interliga dois vértices, sendo que estes são chamados vértices adjacentes.

O caminho em um grafo é visto como um conjunto de nodos interconectados a partir de um nodo inicial a até um nodo final b. Sendo assim, o comprimento de um caminho é definido como o número de nodos visitados, e a distância $d(a, b)$ em um grafo, é o comprimento do menor caminho entre um nodo a e um nodo b.

A partir das definições de caminho e distância entre os nodos de um grafo, é possível a definição do diâmetro, que é caracterizado pela maior distância entre dois vértices de um grafo (JUNGNICKEL, 2002).

Dependendo da natureza das interações, as redes podem ser direcionadas ou não direcionadas. Em redes do tipo direcionadas, a interação ocorre com direção definida, como, por exemplo, a direção de um substrato para um produto em uma rede metabólica. Já no caso de redes não direcionadas isto não ocorre, visto que não existe direção definida e sim, mais de uma direção possível, como, por exemplo, em redes de interação proteína-proteína, onde se a proteína A se liga com a proteína B então a proteína B se liga com a proteína A (BARABáSI; OLTVAI, 2004).

Em muitas redes, se um nodo a é conectado ao nodo b, e este é conectado ao nodo c então, é alta a probabilidade de que o nodo a tenha conexão com o nodo c. Este fenômeno pode ser quantificado usando o coeficiente de clusterização

$$C_i = 2n_i/k(k-1), \tag{3.1}$$

onde n_i é o número de conexões interligando o vizinho k_1 do nodo I a cada outro (BARABáSI; OLTVAI, 2004).

A clusterização caracteriza as tendências dos nodos de formar grupos ou agregados conectados no interior de um grafo. Uma importante medida da estrutura das redes é a função $C(k)$, que é definida como a média do coeficiente de clusterização de todos os nodos com k conexões (BARABáSI; OLTVAI, 2004).

A característica mais elementar de um nodo é seu grau de conectividade k, que enumera o número de ligações de um nodo para outro. Com isso, o grau da distribuição $P(k)$ apresenta a probabilidade de que um nodo selecionado tenha exatamente k ligações. O valor de $P(k)$ é obtido pela soma do número de nodos $N(k)$, com $k = 1, 2, .., n$ conexões. Esta soma é dividida pelo número total de nodos, resultando na freqüência de nodos com k conexões.

3.1 Redes de Erdös e Rényi

O modelo de redes aleatórias proposto por (ERDöS; RéNYI, 1960) assume que um número fixo de nodos n é conectado de forma aleatória a cada outro nodo, com uma probabilidade p. A Figura 7 apresenta o modelo de redes aleatórias.

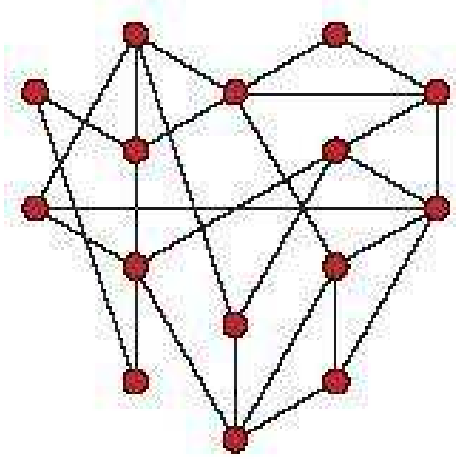

Figura 7: Representação da estrutura de uma rede aleatória.

Neste tipo de rede, o grau dos nodos segue uma distribuição de Poisson. Este tipo de distribuição estatística indica que a maioria dos nodos tem aproximadamente o mesmo número de conexões, e estas se aproximam do grau médio da rede. O grau $P(k)$ diminui exponencialmente quando os nodos desviam de forma significativa da média, o que é bastante raro em redes aleatórias.

3.2 Redes livres de escala

Uma rede livre de escala não possui um número característico de conexões por nodo, como ocorre em uma rede do tipo aleatória (ERDöS; RéNYI, 1960). A probabilidade de que um nodo seja altamente conectado é estatisticamente mais significativa do que em redes do tipo aleatórias. As propriedades topológicas deste tipo de rede determinam que existe um pequeno número de nodos altamente conectados, que são conhecidos por *hubs*, conforme (BARABáSI; OLTVAI, 2004). A Figura 8 ilustra um exemplo deste tipo de rede.

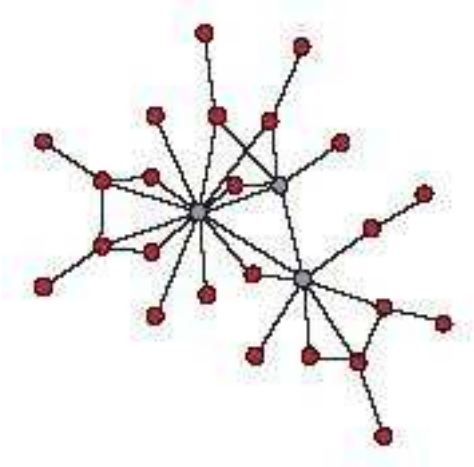

Figura 8: Representação da estrutura de uma rede livre de escala. Os nodos na cor cinza representam os nodos mais conectados, os denominados *hubs*.

A probabilidade estatística $P(k)$ de se encontrar um nodo com k conexões segue uma lei de potência:

$$P(k) \sim k^{-y}, \tag{3.2}$$

onde y é o expoente de grau (RAVASZ et al., 2002) (BARABáSI; OLTVAI, 2004). O valor de y determina muitas propriedades do sistema. Quanto menor for o valor de y, mais importante é o papel dos *hubs* na rede. Visto que para $y = 3$, os *hubs* não são relevantes, para $2 > y > 3$ existe uma hierarquia de *hubs*, com os mais conectados em contato com uma fração menor de todos os nodos e para $y = 2$, os *hubs* e o raio da rede emergem, com

os maiores *hubs* em contato com uma fração grande de todos os nodos.

Neste modelo de redes, a cada instante que um nodo com M ligações é adicionado a rede, este se conecta a um nodo i já existente. A probabilidade de conexão com o nodo i é proporcional a conectividade k do nodo. A rede construída por este processo tem uma distribuição de grau $P(k)$ que segue uma lei de potência, conforme apresentado anteriormente, e não possui modularidade inerente, assim o coeficiente de clusterização (caracteriza todas as tendências de formar grupos ou clusters) $C(k)$ é independente do número de ligações k.

O conceito de modularidade assume que as funções celulares podem ser agregadas em um conjunto de módulos, onde cada módulo é uma entidade discreta composta por diversos componentes e que executa uma tarefa específica separada de outros módulos.

É sabido que milhares de componentes celulares são interconectados dinamicamente, de modo que as propriedades fundamentais da célula são codificadas em uma complexa rede intracelular de interações moleculares. Isto se refere mais ao metabolismo celular, conectado inteiramente em um rede bioquímica na qual centenas de substratos metabólicos são densamente integrados através de reações bioquímicas (RAVASZ et al., 2002).

Dentro desta rede, entretanto, a organização modular não é imediatamente aparente. Estudos demonstram que a probabilidade que um substrato reaja com k outros substratos decai de acordo com a lei de potência deste modelo e com $\cong 2.2$ em todos os organismos, sugerindo assim que as redes metabólicas apresentam um topologia livre de escala (BARABáSI; OLTVAI, 2004).

3.3 Redes hierárquicas

Muitos processos intracelulares são realizados em estruturas modulares. Em nível molecular, a modularidade é uma característica associada a grupos de moléculas que tra-

balham em conjunto para realizar uma determinada função. Nas células existem módulos em complexos protéicos e complexos envolvendo RNA e proteínas que são essenciais nos processos de síntese de proteínas, replicação do DNA e outros.

A construção de um modelo hierárquico combina propriedades das redes livres de escala com um alto grau de clusterização. O ponto inicial para a construção deste modelo de rede é um pequeno cluster. Para melhor explicar esta construção, esta redação será baseada em um cluster inicial de quatro nodos, conforme consta em (BARABáSI; OLTVAI, 2004). Na etapa seguinte são construídas três réplicas do módulo inicial e os três nodos externos de cada módulo criado são conectados ao nodo central do conjunto anterior (nodo inicial), o que resultará em um novo módulo com dezesseis nodos. Na próxima etapa, mais três réplicas são elaboradas, só que estas réplicas são agora construídas a partir do novo módulo (módulo de dezesseis nodos) e não mais sobre o conjunto inicial de quatro nodos, e esta nova replicação produzirá um módulo com sessenta e quatro nodos e os seus nodos externos conectados ao nodo central do módulo anterior (módulo com dezesseis nodos) e assim sucessivamente. A Figura 9 demonstra uma representação deste tipo de rede.

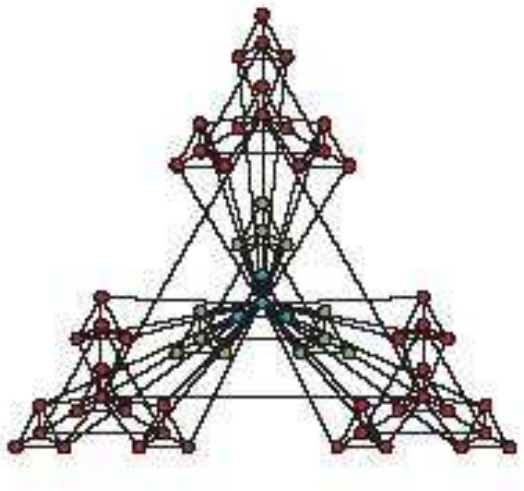

Figura 9: Representação da estrutura de uma rede hierárquica. Neste tipo de rede cada módulo é identificado por um conjunto de triângulos.

Este é um modelo de redes que é significativamente diferente dos modelos tradicionais de redes (BARABáSI et al., 2002). O modelo integra uma topologia livre de escala com uma estrutura modular, isto é, existem pequenos módulos formados por conjuntos de triângulos. A rede produzida possui uma distribuição estatística que segue uma lei de potência, assim como, as redes livres de escala, no entanto, o expoente de grau é diferente. O expoente para este modelo é:

$$y = 1 + (ln4)/(ln3) = 2.26, \tag{3.3}$$

e um coeficiente de clusterização $C \cong 0.6$ (RAVASZ et al., 2002).

A característica mais importante deste modelo de redes não compartilhada pelos modelos de redes aleatórias e pelas redes livres de escala é a arquitetura hierárquica. A rede é construída de pequenos, mas numerosos e integrados módulos, que são ligados aos outros nodos da rede. Estes módulos correspondem a regiões com alta conectividade entre nodos, que são identificados pela alta densidade de motivos de ordem três (3) (triângulos, conforme apresenta a Figura 9). Essas regiões são evidenciadas pelo coeficiente de clusterização C que é em função de k.

A hierarquia deste modelo de rede pode ser caracterizada de maneira quantitativa, isso determina que $C(k)$ obedeça a lei:

$$C(k) \sim k^{-1} \tag{3.4}$$

A arquitetura hierárquica implica que áreas altamente conectadas com nodos altamente conectados se comuniquem com outras áreas que possuem diferenças na conectividade e esta comunicação é mantida por nodos altamente conectados (*hubs*). Esta arquitetura implica que nodos com poucas ligações tenham um elevado valor de C e pertencem a pequenos módulos altamente interconectados. Em contraste, *hubs* altamente conectados tem um valor baixo de C, porque possuem um papel diferente, que é interligar diferentes módulos.

O campo da biologia é repleto de exemplos de modularidade. A interação proteína-proteína e os complexos relativamente invariantes de proteína-RNA (módulos físicos) estão no núcleo de muitas funções biológicas básicas, como, por exemplo, síntese de ácidos nucléicos à degradação de proteínas (BARABáSI; OLTVAI, 2004).

3.4 Redes biológicas e ontologias

As redes biológicas neste trabalho estão diretamente relacionadas ao tema ontologias, que será apresentado na Seção 4.4. A ontologia será usada para a definição semântica e padronização de conceitos biológicos referentes à regulação, interação proteína-proteína e metabolismo. Estes dados serão usados para a construção das redes biológicas com o intuito de facilitar a descoberta de conhecimento relevante para o funcionamento cellular.

4 INTEGRAÇÃO DE DADOS

Neste capítulo será abordado o tema integração de dados, com ênfase em dados biológicos. Será apresentada uma breve descrição sobre as abordagens existentes, assim como as suas respectivas arquiteturas, e ainda realçadas suas vantagens e desvantagens.

Além da integração de dados, será abordado o tema ontologias, desde os seus conceitos, seus tipos e formalismos para representação, e ainda um estudo sobre as ontologias existentes na área de Bioinformática que é o escopo deste trabalho, assim como, os benefícios que uma ontologia pode oferecer.

4.1 Abordagens existentes

Os dados biológicos se encontram hoje em dia, disseminados geograficamente em diversos bancos de dados na Internet (rede mundial de computadores). Este fato implica que cada um destes bancos de dados está em um formato com características particulares, e estes podem ser gerenciados por SGBDs (Sistema Gerenciador de Banco de Dados) diferentes. Outro problema envolvido é quanto à semântica (definição dos dados), visto que não existe um vocabulário comum para padronizar estes dados, isto é, pode haver termos diferentes fazendo referência a um mesmo conceito ou então conceitos diferentes referenciando um mesmo termo. Sem contar que muitos bancos de dados são acessíveis por arquivos texto ou por interfaces *web* que permitem apenas um único mecanismo de consulta.

Além destes problemas, existe uma questão relacionada à confiabilidade dos dados,

isto é, o quanto estes dados disponíveis são confiáveis, visto que os mesmos se encontram disponíveis em bancos de dados públicos. Entretanto, este é um problema que foge ao domínio do processo de integração de dados. Outra questão relacionada à inconsistência dos dados, é que nem sempre os dados disponibilizados por meio de arquivos texto condizem com os dados para consulta *on-line* nos bancos de dados que disponibilizam este serviço.

Levando em conta que os dados não se encontram em um formato padrão e que são distribuídos geograficamente pela rede, é necessário um mecanismo que seja capaz de organizar estes dados, de forma a reuní-los e armazená-los em um único local.

Dentro deste contexto, é necessária a utilização de uma abordagem de integração de dados. Com isso, duas abordagens principais se destacam, e são elas: abordagem materializada e abordagem virtual.

Na abordagem materializada, os dados são adquiridos e integrados em um local físico de armazenamento, como por exemplo, um SGBD, com as consultas (buscas) sendo manipuladas diretamente no local onde os dados estão armazenados, sem necessidade do uso das fontes originais dos dados. Diferentemente do que acontece com a abordagem virtual, onde a recuperação de informações é elaborada a partir de consultas submetidas às fontes originais, sendo assim, neste tipo de abordagem é desnecessário armazenar os dados em um local físico.

Estas duas abordagens apresentam vantagens e desvantagens. No caso da abordagem materializada, os dados são organizados em um local físico de armazenamento único, isto implica que a busca por informações é executada neste repositório, sendo portanto desnecessário o uso das fontes originais para as consultas. Esta abordagem, contudo apresenta uma desvantagem com relação aos dados, isto é, não garante que os dados armazenados são dados constantemente atualizados, ou seja, nem sempre refletem o estado atual das fontes originais. Em contrapartida, a abordagem virtual contempla dados constantemente atualizados e com as consultas manipuladas sobre estes dados, con-

seqüentemente, a resposta implica em dados mais novos (dados mais atualizados) do que a abordagem materializada. A desvantagem da abordagem virtual é não garantir que os dados estarão sempre disponíveis, isto é, que os bancos de dados permanecerão sempre ativos (*on-line*).

4.2 Arquiteturas de integração

Quando é necessário executar o processo de integração de dados, uma questão que surge e que precisa ser respondida é: *Como fazer esta integração?* A resposta é simples, bastar fazer o uso de uma abordagem de integração e implementar a sua arquitetura, e neste quesito, três se destacam: abordagem federada, de mediadores (WIEDERHOLD, 1992) e *data warehouse* (INMON, 1997).

A abordagem federada é composta por um conjunto de bancos de dados que trabalham em cooperação e de forma autônoma e possibilitam o compartilhamento controlado dos dados. Esta troca de informações é exclusivamente para os bancos de dados pertencentes à federação e a sua principal característica é a troca de dados por parte de sistemas completamente diferentes.

Levando em conta que os dados biológicos podem ser adquiridos e integrados, e que os mesmos se encontram armazenados em bancos de dados públicos distribuídos na *web*, estes se enquadram no contexto de uma arquitetura cliente/servidor. De um lado têm-se os usuários, como o lado cliente da arquitetura que a todo o momento buscam por informações através de mecanismos de consultas, e do outro lado, tem-se os servidores, que como o próprio nome diz, representam o lado servidor da arquitetura, onde as informações estão armazenadas e que normalmente estão em bancos de dados. Na comunicação entre o lado cliente e o lado servidor, existe uma camada denominada de *middleware*, cuja única e exclusiva função é garantir que a consulta elaborada pelo usuário chegue de forma correta e precisa ao lado servidor e o mesmo aconteça no retorno desta consulta por parte do servidor ao lado cliente.

Dentro deste contexto, duas arquiteturas se destacam e são elas: arquitetura de mediadores, que implementa a abordagem virtual e a arquitetura de *data warehouse* que faz uso da abordagem materializada.

O conceito de mediadores foi definido em (WIEDERHOLD, 1992), como sendo módulos de *software* que exploram o conhecimento representado sobre um conjunto ou um subconjunto de dados para criar informações para uma camada de alto nível.

Os mediadores constituem uma arquitetura radicalmente diferente. Esta arquitetura é usada nas situações de integração de dados, onde a atualidade dos dados é crítica ou quando é impossível carregar os dados por inteiro das fontes originais. Neste modelo de arquitetura, os dados não são armazenados, quando o lado cliente solicita uma consulta, o mediador simplesmente localiza a fonte apropriada e submete à consulta a fonte. Esta arquitetura é particularmente atrativa para a integração de dados, quando não é possível realizar a aquisição dos dados por meio de *download* e também não é possível aguardar a notificação de quando novas atualizações ocorrem. A tecnologia aqui difere radicalmente dos servidores de dados tradicionais, já que envolve mais manipulações algébricas nas consultas que processamento de dados. Primeiro, o mediador tem que decidir que fontes contribuem para a consulta; isto pode não ser trivial quando o mediador integra dados de dezenas ou centenas de fontes. Segundo, uma vez que as fontes relevantes tenham sido identificadas, o mediador executa uma transformação de consulta fonte-a-fonte, um processo por vezes chamado de *reescrita de consulta*. Quando dados de duas ou mais fontes devem ser extraídos, o mediador precisa produzir um plano global de execução, determinando em que ordem consultar as fontes (ABITEBOUL; BUNEMAN; SUCIU, 2000). A Figura 10 apresenta a representação da arquitetura de mediadores.

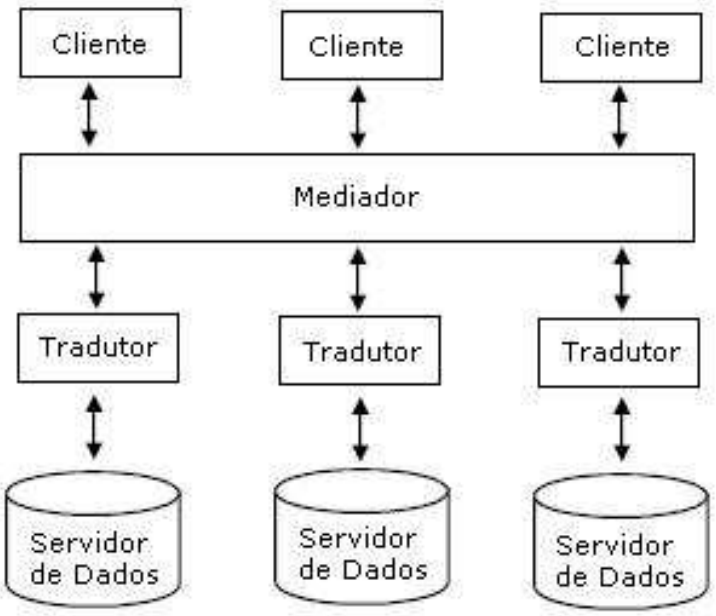

Figura 10: Representação da arquitetura de mediadores, que implementa a abordagem virtual para a integração de dados.

Conforme já apresentado nesta seção, neste tipo de arquitetura os dados não são armazenados em um local físico de armazenamento, isto é, os mesmos se encontram armazenados nas fontes originais por meio de um ou mais servidores centrais, onde as consultas são aplicadas, sendo este processo completamente transparente ao usuário. O processo de consulta é construído da seguinte forma: a consulta é elaborada pelo usuário como se os dados estivessem armazenados localmente, isto é, em uma base de dados local, porém isto não ocorre, estes dados estão armazenados nos diversos bancos de dados distribuídos pela *web*, com isso, o mediador necessita localizar o banco de dados apropriado para submeter à consulta e o retorno desta é apresentado ao usuário. Porém, caso haja necessidade do uso de mais de uma fonte para construir a resposta, estas são agrupadas em uma única resposta, ou seja, mesmo que o mediador necessite usar diferentes fontes para elaborar a resposta, esta é sempre agrupada de acordo com a ordem das fontes e em uma única resposta ao usuário. Para o usuário todo este processo é transparente.

O conceito de *Data Warehousing* é um processo, não um produto, para montar e gerenciar repositórios de dados a partir de várias fontes de dados, com o propósito de ter uma visão detalhada e singular de parte ou todo de um negócio (GARDNER, 1998).

Data Warehouse é uma coleção de dados orientada por assuntos, integrada, variante no tempo, e não volátil, que tem por objetivo dar suporte aos processos de tomada de decisão (INMON, 1997).

Contudo, a abordagem de DW tem como objetivo centralizar os dados oriundos de diversas fontes de dados em um único e centralizado repositório. É um mecanismo que objetiva trabalhar em um ambiente onde exista uma grande quantidade de dados e precisam ser integrados para que seja possível uma análise mais detalhada sobre estes dados. A Figura 11 ilustra uma representação deste tipo de arquitetura.

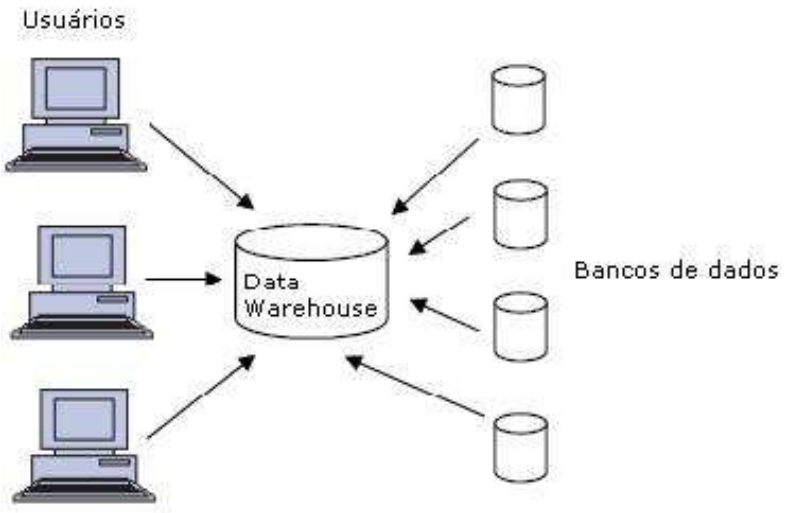

Figura 11: Representação da arquitetura de *data warehouse*, que implementa a abordagem materializada para a integração de dados.

Diferentemente da arquitetura de mediadores, no DW os dados são logicamente e fisicamente transformados, atualizados e armazenados pelo tempo que for conveniente à aplicação (INMON, 1997).

Ainda tem como intuito ser um ambiente que forneça dados integrados com qua-

lidade, possibilitando que sejam manipuladas inferências sobre estes dados.

4.3 Comparação entre as arquiteturas de mediadores e data warehouse

Analisando de uma forma comparativa as vantagens e desvantagens de cada uma das arquiteturas, conclui-se que cada uma delas é apropriada para uma determinada aplicação, ou seja, é necessário estudar as necessidades da aplicação para predizer qual é a arquitetura que melhor se enquadra.

A arquitetura de mediadores é mais aconselhável quando não é possível a aquisição dos dados e nem o armazenamento em um único local físico. No entanto, a arquitetura de *data warehouse* é mais apropriada quando é possível adquirir os dados e centralizá-los em um único repositório para uma análise posterior mais detalhada e minuciosa.

A arquitetura de mediadores possui a vantagem de apresentar os dados constantemente atualizados, porém, apresenta a desvantagem de não poder garantir que os bancos de dados estarão sempre acessíveis (*on-line*) no exato momento em que a consulta é submetida. A arquitetura de *data warehouse*, por sua vez, apresenta a vantagem dos dados estarem armazenados em um único e centralizado repositório, com as consultas realizadas sobre este repositório, sendo desnecessário o uso dos bancos de dados originais no acesso aos dados. Em contrapartida, este tipo de arquitetura não garante que os dados armazenados reflitam o estado atual das fontes originais, isto é, os dados armazenados podem estar desatualizados. Porém, este problema é solucionável, desde que certos princípios sejam executados para garantir a manutenção de um *data warehouse*, e são eles:

reconstrução : periodicamente é necessário atualizar o *data warehouse* com o propósito que o mesmo esteja sempre atualizado;

atualização incremental : somente as modificações transcorridas nas fontes originais são atualizadas;

atualização imediata : atualiza o *data warehouse* sempre que a fonte original é atualizada.

4.4 Ontologias

A palavra ontologia teve origem na Filosofia, onde é a área da Metafísica que investiga a natureza dos seres. Do grego $Ontos + Logia$ ($Ser + Estudo$). No campo que a IA (Inteligência Artificial) abrange, especialistas definiram como sendo uma maneira formal de representar o conhecimento (GóMEZ-PéREZ, 1999). Porém, na literatura, a mais famosa e referenciada definição é a de Gruber: "*uma ontologia é uma especificação explícita de uma conceitualização*" (GRUBER, 1993).

No entanto, em 1997, a definição de Gruber sofre algumas modificações por parte de Borst e passa a ser: "*ontologias são definidas como uma especificação formal de uma conceitualização*" (BORST, 1997).

Estas duas definições foram explicadas em (STUDER; BENJAMINS; FENSEL, 1998), como sendo: Conceitualização se refere a um modelo abstrato de algum fenômeno do mundo identificado por conceitos relevantes deste fenômeno. Explícito significa que o tipo de conceito usado e as restrições são explicitamente definidos. Formal se refere ao fato que uma ontologia deve ser compreendida por uma máquina. Compartilhada reflete a noção que uma ontologia captura o conhecimento consensual, isto é, não é privado para um indivíduo, mas aceito por um grupo.

4.4.1 Componentes

As ontologias provêem um vocabulário comum em uma determinada área e definem, em diferentes níveis de formalidade, os termos e as suas respectivas relações. Os conhecimentos em ontologias são formalizados sob a ótica de cinco tipos de componentes, e são eles: conceitos, relações, funções, axiomas e instâncias (GRUBER, 1993) (GóMEZ-PéREZ, 1999).

conceitos : usados pelo senso comum, podem ser abstratos ou comuns, elementares ou compostos, reais ou fictícios;

relações : representam o tipo de interação entre os conceitos do domínio, como por exemplo, classe e subclasse;

funções : é um tipo de relação, onde o n-ésimo elemento da relação é único para os n-1 elementos anteriores;

axiomas : usados pela modelagem para as sentenças que são sempre verdadeiras;

instâncias : utilizadas para representar os elementos da ontologia, ou seja, os dados.

4.4.2 Especificação de uma ontologia

Ao se fazer uso de uma ontologia, é essencial que a mesma seja formalmente especificada (BORST, 1997). Há inúmeros formalismos que podem ser usados para este processo, tanto os baseados em *frames* quanto os baseados em lógica de predicados ou ainda em ambos os paradigmas. Dentre os formalismos existentes, há os tradicionais (GóMEZ-PéREZ, 1999) e os padrões *web* (SU; LARS, 2002). Os que mais se destacam nos métodos tradicionais são a Ontolíngua, CycL, LOOM e FLogic (GóMEZ-PéREZ, 1999).

A Ontolíngua é uma linguagem baseada no KIF (*Knowledge Interchange Format*) e na *Frame Ontology*, e é uma linguagem para a construção de ontologias pelo servidor da Ontolíngua. O KIF é uma interlíngua, isto é, uma linguagem para tradução entre formalismos de representação, que incorpora declarativas semânticas (definição dos termos), tem força expressiva suficiente para representar o conhecimento declarativo contido em aplicações típicas de sistemas de base de conhecimento. Todavia, apresenta um problema, que é a ausência de um motor de inferência. A *Frame Ontology* é uma ontologia para representação de conhecimento para a modelagem em uma abordagem baseada em *frames* e foi construída a partir do KIF e uma série de extensões desta linguagem. Através da Ontolíngua é possível elaborar ontologias de três maneiras:

- fazendo o uso de expressões do tipo KIF;
- uso exclusivo do vocabulário da *Frame Ontology*, porém desta maneira não é possível representar os axiomas;
- usando as duas linguagens ao mesmo tempo, dependendo da preferência do desenvolvedor.

De forma independente da abordagem, a definição na Ontolíngua segue um padrão e sempre é composta por um cabeçalho, uma definição formal em linguagem natural e uma descrição formal escrita em KIF ou no vocabulário controlado da *Frame Ontology*.

CycL é uma linguagem para representação de conhecimento. É declarativa e expressiva, similar ao cálculo de predicados de primeira ordem com o acréscimo de algumas extensões. O seu motor de inferência executa desde lógicas genéricas até procura *best-fit*, fazendo uso de um conjunto de heurísticas proprietárias, uso de microteorias para otimizar inferências de domínios restritos e inclui diversos módulos para classes de inferências específicas.

LOOM é uma linguagem de programação de alto nível baseada na lógica de primeira ordem. Provê um modelo declarativo de linguagem de especificação expressivo e explícito, suporte dedutivo, checagem de consistência automática, diversos paradigmas de programação que atuam como uma interface com o modelo de especificação declarativo e serviços de base de conhecimento.

FLogic é uma integração de linguagens baseadas em *frames* e cálculo de predicados de primeira ordem, inclui objetos (simples ou complexos), herança, tipos polimórficos, consulta de métodos e encapsulamento. É um sistema dedutivo que trabalha com a teoria do cálculo de predicados e herança estrutural e comportamental.

Com relação aos formalismos baseados no padrão *web* tem-se o XML, RDF e a OWL, conforme apresenta (SU; LARS, 2002).

O XML (*eXtensible Markup Language*) é um formato universal para a estruturação

de documentos e dados na *web* proposto pelo W3C (*http://www.w3c.org*). A sua principal contribuição é a capacidade de prover uma sintaxe comum e fácil para documentos *web*. Entretanto, o XML sozinho não é uma linguagem de ontologias, porém, o XML-Schema pode ser estendido e usado para especificar uma ontologia. No entanto, o XML-Schema foi criado principalmente para verificação de documentos XML e modelagem de primitivas que são as suas maiores aplicações.

O RDF (*Resource Description Format*) (*http://www.w3c.org*) é uma infra-estrutura para conversão, troca e reuso de metadados estruturados, e assim como o XML também foi proposto pelo W3C. O RDF fornece um formulário padrão para representar os metadados em XML. O modelo de dados em RDF consiste em três tipos de objetos:

resources : são descritos por expressões RDF;

properties : definem aspectos específicos, características, atributos ou relações para descrição de um recurso;

statements : atribui um valor para uma propriedade em um recurso específico (pode ser outra indicação de RDF).

O RDF não possui mecanismos para definir o relacionamento entre processos, atributos e recursos, este é o papel do RDF-S (*http://www.w3c.org*). O RDF-S pode ser usado diretamente para descrever ontologias, embora sua função principal não seja a especificação de uma ontologia. O RDF-S fornece um conjunto fixo de primitivas para a definição de uma ontologia (classes, subclasses, propriedades, *is-a*, elementos de relacionamentos, dentre outros) e uma maneira padrão para converter para dentro do XML. No entanto, o RDF-S tem poder de expressão limitado, visto que os axiomas não podem ser definidos diretamente.

Neste caso, observa-se que a relação entre RDF-S e ontologias é muito mais próxima do que entre XML e ontologias.

A OWL (*Ontology Web Language*) (*http://www.w3c.org*) é baseada em XML, RDF e RDF-S, e pode ser usada para representar explicitamente o significado dos termos nos vocabulários e nos relacionamentos entre os termos. Esta representação de termos e seus inter-relacionamentos são chamados ontologia. A OWL possui mais facilidades para expressar o significado e a semântica do que XML, RDF, RDF-S. A OWL é uma revisão da DAML+OIL (*http://www.w3c.org*).

A OWL possui três sub-linguagens, e são elas:

OWL lite : suporte para usuários que necessitam de uma hierarquia de classificação;

OWL DL : suporte aos usuários que desejam o máximo de expressividade;

OWL full : para usuários que desejam o máximo de expressividade e a liberdade semântica do RDF sem nenhuma garantia computacional.

4.4.3 Tipos de ontologias

Na literatura existem diferentes caracterizações sobre os tipos de ontologias. No entanto, neste trabalho será adotada a caracterização segundo a visão de (STUDER; BENJAMINS; FENSEL, 1998).

ontologias de domínio : compreendem o conhecimento que é válido para um domínio em particular, provendo assim um vocabulário específico dentro deste domínio, como por exemplo, eletrônica, medicina e mecânica;

ontologias de aplicação : contém todo o conhecimento necessário para a modelagem de um domínio particular;

ontologias de representação de conhecimento : compreendem a representação para formalizar o conhecimento em paradigmas de representação, por exemplo, a *Frame Ontology*, que faz uso da representação primitiva, usando uma linguagem baseada em *frames*;

ontologias genéricas : faz uso de um domínio genérico, isto é, a ontologia pode ser reaproveitada em diversos domínios.

4.4.4 Princípios de construção de uma ontologia

Quando existe a necessidade de se construir uma ontologia, algumas questões surgem e, portanto, precisam ser respondidas, como por exemplo: *Existe algum conjunto de princípios para se construir uma ontologia? Se sim, quais são? E por onde começar?* A resposta para todas estas perguntas é sim. Analogamente à produção de um *software*, onde é necessário o correto cumprimento de certas atividades para que no final se obtenha um produto com qualidade, como por exemplo, especificar o domínio e o escopo da aplicação, elaborar o levantamento de requisitos junto a um especialista, o desenvolvimento da modelagem do sistema, dentre outras etapas. Na construção de uma ontologia ocorre da mesma maneira e um conjunto de princípios deve ser seguido, conforme apresentam (GRUBER, 1993) (GóMEZ-PéREZ, 1999):

clareza e objetividade : uma ontologia deve apresentar de forma clara e objetiva o significado dos termos, por meio de sua definição, assim como uma documentação em linguagem natural;

completitude : a definição expressa por uma condição necessária e suficiente é preferida em relação a uma definição parcial, isto é, a definição completa sobre uma incompleta;

coerência : uma ontologia deve ser coerente, isto significa permitir inferências que sejam consistentes com as definições. A definição dos axiomas deve ser lógica, e a coerência deve atingir também os conceitos definidos informalmente, tais que os mesmos devem ser expressos em linguagem natural e exemplos;

extensibilidade : permitir a inclusão de novos termos, ou ainda, a adição de termos especializados, de forma que não seja necessária a revisão de definições já existentes;

mínima codificação : a conceitualização deve ser especificada em um nível de conhecimento sem que exista a dependência de um padrão;

mínimo compromisso ontológico : com objetivo de aumentar o reuso, apenas o conhecimento essencial deve ser incluído, injetando assim a menor teoria possível sobre um determinado conceito, tornando possível a inclusão de novos conceitos para especializar o assunto.

4.4.5 Metodologias

O processo de desenvolvimento de uma ontologia se refere a quais etapas devem ser executadas, e são de três tipos:

atividades de gerenciamento : tem como objetivo assegurar o bom funcionamento da ontologia e isto inclui tarefas de planejamento, controle e garantia de qualidade;

atividades orientadas ao desenvolvimento : tem como intuito construir a ontologia, executando tarefas de especificação, conceitualização, formalização, implementação e tarefas de manutenção;

atividades integrais : o foco é dar sustentações sólidas às atividades de desenvolvimento e compreende, aquisição de conhecimento, integração, avaliação, documentação e configuração.

Se a construção da ontologia se der em pequena escala, algumas destas etapas podem ser abstraídas, porém, se a construção se der em larga escala é necessário que todas estas etapas sejam executadas (GóMEZ-PéREZ, 1999).

4.4.6 Ferramentas de desenvolvimento

Com relação à construção de uma ontologia, existe uma enorme gama de ferramentas disponíveis. Dentre estas, aparecem o servidor da Ontolíngua, Ontosaurus, ODE,

Tadzebao e WebOnto, Protégé, dentre inúmeras outras ferramentas, conforme apresentam (GóMEZ-PéREZ, 1999) (ALMEIDA; BAX, 2003).

O servidor da Ontolíngua compreende um conjunto de ferramentas e serviços com suporte a criação de ontologias compartilhadas entre grupos distribuídos geograficamente. Elaborada no contexto da ARPA (*Knowledge Sharing*) pela *Stanford University*, a sua arquitetura provê uma biblioteca de ontologias, tradutores para as linguagens Prolog, CORBA's IDL, Clips, LOOM e KIF e ainda um editor para se criar e se navegar pela ontologia.

Ontosaurus foi desenvolvido pelo Instituto de Ciências da Informação na *University of South California* e é dividido em duas partes: um servidor de ontologias que faz uso do LOOM como sistema de representação de conhecimento e um servidor de navegação de ontologias que dinamicamente cria páginas html, incluindo imagem e documentação textual que mostra a hierarquia da ontologia e usa o formato html para permitir que o usuário possa editar a ontologia. Ontosaurus converte de LOOM para Ontolíngua, KIF, KRSS e C++.

ODE (*Ontology Design Environment*) foi desenvolvido pela Escola de Ciência da Computação da Universidade Politécnica de Madrid e a sua principal vantagem são os módulos de conceitualização para a construção de modelos conceituais *ad hoc*. O módulo de conceitualização permite desenvolver a ontologia ao nível do conhecimento fazendo uso de um conjunto intermediário de representação que são independentes da linguagem em que a ontologia está sendo implementada. Uma vez que a conceitualização está completa, o código é gerado automaticamente usando o gerador de códigos da ODE, sendo que o mesmo inclui a Ontolíngua, FLogic e bancos de dados relacionais.

Tadzebao e WebOnto são ferramentas complementares. Tadzebao permite discussões síncronas e assíncronas sobre ontologias, enquanto que WebOnto suporta navegação colaborativa, construção e edição de ontologias na *web*.

Protégé (*http://protege.stanford.edu*) é um ambiente interativo para o projeto de

ontologias, possui código aberto (*open source*), além de ser multiplataforma. Desenvolvido em Java pela *Stanford University*, a mesma que desenvolveu o servidor da Ontolíngua. Oferece uma interface gráfica para a edição de ontologias e provê uma arquitetura para a criação de ferramentas baseadas em conhecimento. Sua arquitetura modular permite a inserção de novos recursos (*plugins*), como por exemplo, um motor de inferência. As aplicações desenvolvidas no Protégé são usadas em resolução de problemas e tomadas de decisão em um domínio particular e também está sempre em constante desenvolvimento. A sua última versão até o momento é a 3.1, e é voltada para a construção de ontologias OWL.

4.4.7 Benefícios das ontologias

Com relação aos seus benefícios, uma ontologia proporciona a capacidade de reaproveitamento de outras ontologias, visto que sempre que uma nova ontologia é proposta, é analisada dentro ou até mesmo fora do domínio, a existência de ontologias semelhantes, para que se tenha uma base por onde começar o seu desenvolvimento. Outro ponto relevante, é que uma ontologia torna o conhecimento padrão para um determinado domínio, eliminando assim o problema da redundância de informação.

As ontologias proporcionam ainda melhorias na recuperação de informações, ao organizar o conteúdo de diversas fontes de dados que compõem um domínio (ALMEIDA; BAX, 2003). Além disso, as ontologias permitem formas de representação baseadas em *frames*, lógica de predicados ou ainda em ambos os paradigmas. A forma de representação baseada em lógica possibilita o uso de mecanismos de inferência para construir novos conhecimentos, a partir de conhecimentos já existentes, representando assim uma evolução em relação às técnicas tradicionais.

4.4.8 Aplicações em Bioinformática

As ontologias podem ser utilizadas para a comunicação entre sistemas, pessoas e organizações, suportar o projeto e o desenvolvimento de sistemas genéricos baseados em conhecimento. Entretanto, o número de aplicações que faz uso de ontologias para modelagem de aplicações ainda é relativamente modesto. Muitas vezes, as ontologias são construídas para modelar uma aplicação específica, sem uma consideração especial pelo compartilhamento e reuso. Diversos são os problemas que dificultam o reuso das ontologias nas aplicações, como por exemplo, os formalismos de representação diferem dependendo de onde a ontologia se encontra. Ontologias no mesmo servidor normalmente são descritas com diferentes níveis de detalhe e também não existe um formato comum para a representação relevante da informação sobre as quais, o usuário pode decidir qual delas é a que melhor se enquadra no seu propósito. Estes são os principais pontos apontados por (GóMEZ-PéREZ, 1999), como os problemas que causam este baixo número de aplicações conhecidas até o momento nas áreas de gerenciamento do conhecimento, geração de linguagem natural, sistemas baseados em conhecimento, dentre outras.

Existem diversas áreas e domínios usando ontologias hoje em dia, como por exemplo, gestão do conhecimento, comércio eletrônico, processamento de linguagem natural, recuperação de informações na *web*, projetos relacionados à educação e Bioinformática. Dentro do contexto da Bioinformática, que é o domínio e o escopo deste trabalho, algumas ontologias se destacam e são detalhadas a seguir:

GO (*Gene Ontology*) (*http://www.geneontology.org*) é um dos projetos mais ambiciosos aplicados à Biologia. É um esforço colaborativo dirigido à necessidade de descrições consistentes de produtos de genes em diferentes bancos de dados. O projeto começou com uma colaboração entre três bancos de dados de organismos em 1998. Desde então, o GO *Consortium* cresceu para incluir novos bancos de dados.

Os colaboradores da GO estão desenvolvendo três estruturas, vocabulários controlados (ontologias) que descrevem os produtos dos genes nos termos de seus processos

biológicos, funções moleculares e componentes celulares.

processos biológicos : formado por um ou mais conjuntos de funções moleculares;

funções moleculares : descreve as atividades no nível molecular;

componentes celulares : enumera a localização na célula, considerando subestruturas celulares.

Cada um destes domínios possui a sua própria organização hierárquica (YEH et al., 2003). A utilização destas sub-ontologias é na anotação de genes, produtos de genes e seqüências.

Neste esforço há três aspectos separados: primeiro, escrevem e mantém a ontologia sozinhos, segundo, fazem associações entre ontologias e genes, e produtos de genes em colaboração com os bancos de dados e por último, desenvolvem ferramentas que facilitam a criação, manutenção e uso das ontologias.

SO (*Sequence Ontology Project*) (*http://song.sourceforge.net*) é um conjunto de termos usados para descrever características sobre uma seqüência de nucleotídeos ou proteínas. Abrange características "cruas", tais como, batida de similaridade de nucleotídeos e interpretações, tais como modelo de genes.

Provê recursos para a comunidade de Bioinformática, que são:

- um vocabulário controlado estruturado para a descrição de anotações preliminares de seqüência de ácidos nucléicos;
- uma representação estruturada destas anotações dentro dos bancos de dados genômicos;
- um vocabulário controlado estruturado para a descrição das mutações na seqüência e no nível mais bruto no contexto dos bancos de dados genômicos.

PSI-MI (*Proteomics Standards Initiate*) (*http://psidev.sourceforge.net*) é uma ontologia de interações moleculares com enfoque na interação proteína-proteína. A PSI-MI é um esforço do HUPO (*Human Proteome Organization*) implementada em XML-Schema através de uma especificação ontológica. O estado corrente da ontologia implementa representações declarativas de interações moleculares, divididas em cinco conceitos:

detecção de característica : método utilizado para determinar as características envolvidas na interação, por exemplo, estrutura tridimensional de uma proteína;

tipo de característica : são as propriedades de subseqüências que interferem na ligação das proteínas, por exemplo, determinar o início e o fim dos sítios de ligação, onde o sítio é o local onde uma proteína se liga à outra;

detecção de interação : é o método para identificar o modo de interação das proteínas (*in silico* (simulação por computador) ou através de experimentos de bancada);

tipo de interação : é o método de interação física entre as proteínas, por exemplo, a maneira como duas proteínas interagem fisicamente em uma interação;

detecção de participantes : método para detectar as proteínas envolvidas em uma interação.

A PSI-MI permite a definição de comunidades-alvo padrões para representação de dados em *proteomics* para facilitar a comparação, troca e verificação nos dados, e também define um conjunto mínimo de dados padrão que permite a cientistas fornecer um conjunto central de dados, porém, para a informação completa é necessário consultar a fonte original dos dados. Além disso, a ontologia trabalha com dados não sincronizados entre as diversas bases de dados que a compõem.

MGED (*Microarray Gene Expression Data*) (*http://www.mged.org*) tem o propósito de ser uma ontologia para prover termos padrões para a anotação de experimentos de microarranjos. Sua modelagem de representação necessita de uma estrutura de dados

complexos, entretanto, a inexistência de um formato universal complica este processo, tais como, documentação e troca de dados (SPELLMAN et al., 2002). Artigos na área têm demonstrado que a reprodução de experimentos de microarranjos é uma tarefa problemática (BRAZMA et al., 2001).

BioPAX (*Biological Pathways Exchange Format*) (*http://www.biopax.org*) tem como objetivo facilitar a integração e a troca de dados armazenados em diversos bancos de dados biológicos referentes às vias metabólicas. Normalmente, a integração de dados de diversos bancos de dados biológicos é visto como um desafio em Bioinformática. Uma solução para este problema é definir um formato padrão de representar estes dados para uma determinada comunidade. Atualmente, não existe um formato padrão aplicável aos dados biológicos de vias, apesar destes dados estarem disponíveis em cerca de 100 bancos de dados distribuídos na *web*.

O projeto BioPAX tem como intuito fornecer um formato para a troca de dados de vias, para representar os elementos chave do modelo de dados para os bancos de dados mais populares, e para alcançar este objetivo, a ontologia BioPAX foi desenvolvida para suportar os modelos de vias existentes, tais como, BioCyc (*http://www.biocyc.org*), BIND (*http://www.bind.ca*), KEGG (*http://www.genome.jp/kegg*), além de muitos outros bancos de dados.

Quando projetada para o nível 1, a equipe de desenvolvimento esforçou-se para encontrar um denominador comum, devido às muitas necessidades diferentes de representação, aderindo aos princípios do projeto que promovem a interoperabilidade. Estes princípios incluem flexibilidade, extensibilidade e compatibilidade com outros padrões. Devido ao fato dos dados de vias serem complexos e poderem ser representados em muitos níveis de detalhe, o BioPAX está fazendo o uso de uma abordagem de desenvolvimento nivelada, similar à SBML (*http://sbml.org/index.psp*).

O nível 1 da ontologia BioPAX representa a informação sobre os caminhos metabólicos. O nível 2 expande o escopo para cobrir interações moleculares. Este can-

didato (nível 2) é liberado para a revisão final e para ser testado pela comunidade, se nenhum erro for detectado, o candidato é liberado tornando-se o nível 2 da ontologia.

As ontologias apresentadas exemplificam como os esforços têm coberto a vasta área da Biologia Molecular. Entretanto, a construção de modelos topológicos integrados de redes moleculares ainda necessita de uma ontologia que seja capaz de integrar dados de experimentos envolvendo as diferentes moléculas presentes dentro da célula. Cada um dos esforços apresentados poderia ser útil nesta tarefa, porém, nenhuma das ontologias possuía todos os conceitos necessários para a construção de uma rede integrada. Baseado neste ponto haveria duas possibilidades, a primeira seria usar uma das ontologias existentes e adequar à mesma para a construção dos modelos topológicos integrados, e a segunda seria a construção de uma nova ontologia para realizar esta tarefa. Neste trabalho optou-se pela segunda possibilidade, com isso apresenta-se a ontologia MONET (*Molecular Network*) (BATTISTELLA et al., 2004).

5 ONTOLOGIA MONET

Uma das mais importantes mudanças para a Biologia na era pós-genômica é entender a estrutura e o comportamento de interações moleculares complexas que controlam o comportamento celular (BARABáSI; OLTVAI, 2004). A enorme e complexa quantidade de dados coletados durante os últimos anos, contém um valor muito grande de informações que necessitam de uma abordagem de integração (UETZ; IDEKER; SCHWIKOWSKI, 2002). Cientistas da computação e biólogos pesquisam por metodologias inovadoras para lidar com estes dados, de forma a aumentar o entendimento a cerca dos processos biológicos fundamentais que operam dentro da célula (BARABáSI; OLTVAI, 2004) (YEGER-LOTEM et al., 2004) (UETZ; IDEKER; SCHWIKOWSKI, 2002) (IDEKER et al., 2001).

Entretanto, a integração é uma tarefa difícil devido ao fato dos dados biológicos estarem distribuídos em diferentes bancos de dados. Estes por sua vez, têm diferentes sistemas de gerenciamento, formato e visão como representar estes dados. Destes, muitos são acessíveis por arquivos texto ou por interfaces *web* que permitem um único mecanismo de consulta ou nem permitem consultas. Neste contexto, dois problemas principais estão envolvidos: necessidade de um *parser* para cada banco de dados e ausência de um vocabulário unificado.

Em Bioinformática, as ontologias são cruciais para a manutenção da coerência de uma larga coleção de conceitos complexos e seus relacionamentos (BAKER et al., 1999).

Uma ontologia é uma especificação explícita de uma conceitualização (GRUBER, 1993), enquanto que os vocabulários controlados apenas restringem os termos utilizados

para descrever um domínio. As ontologias estendem este simples vocabulário controlado e permitem uma especificação formal dos termos e relacionamentos, isto é feito, para possibilitar o compartilhamento e o reuso do conhecimento. Ainda suportam a interoperabilidade entre sistemas, e permitem inferências sobre o conhecimento representado.

Dentro deste contexto, este trabalho apresenta a ontologia MONET, inicialmente proposta (BATTISTELLA et al., 2004) (BATTISTELLA et al., 2005), que é um modelo integrado para a *rede de redes* que existe dentro da célula (BARABáSI; OLTVAI, 2004). Tal visão integrada ajuda a entender as interações de larga escala responsáveis pelo comportamento celular que pode ser experimentalmente testado (IDEKER et al., 2001) e facilitar a formulação de novas hipóteses.

O domínio da ontologia integra informações de caminhos metabólicos, regulação gênica e interação proteína-proteína, sendo que para caminhos metabólicos e interação proteína-proteína engloba organismos procariotos e eucariotos, enquanto que a regulação gênica contempla apenas os organismos procariotos, através de uma visão que permite estabilizar um modelo capaz de minimizar a redundância e inconsistência de dados. Na Figura 12 é apresentado o domínio da ontologia MONET, que é composto por caminhos metabólicos, regulação gênica e interação proteína-proteína. A intersecção mostra os conceitos comuns entre as redes.

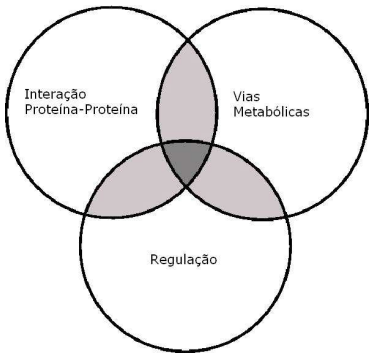

Figura 12: Domínio da ontologia MONET.

5.1 Modelagem e especificação

A modelagem da ontologia MONET se desenvolveu por meio do ambiente Protégé (*http://protege.stanford.edu*) em um ambiente Linux, e foi usado por dois motivos principais: primeiro por não ser somente um editor de ontologias, mas sim um KBMS, visto que um dos objetivos é popular a base de instâncias com diversos organismos de diferentes bancos de dados biológicos, e o segundo motivo é por contemplar uma arquitetura modular, o que permite a extensão de suas funcionalidades, através da adição de novos recursos (*plugins*), como por exemplo, o RACER (*http://www.racer-systems.com/*) que pode ser utilizado para a checagem de inconsistências na ontologia, e ainda possui a vantagem de ser um ambiente multiplataforma. Uma outra característica da ferramenta é permitir que a ontologia seja exportada em diferentes formatos de representação, o que possibilita a especificação da ontologia sob diferentes visões, tais como, OWL, RDF, XML e XML-Schema.

A ontologia é baseada no Protégé. É uma descrição formal explícita de conceitos de um domínio de discurso (classes), propriedades de cada conceito (atributos), que descrevem suas características e restrições (facetas). A Figura 13 mostra o ambiente Protégé no seu formato de gerenciamento para a criação de uma ontologia. Neste exemplo, o ambiente já se encontra no formato OWL, que é o formato adotado para o desenvolvimento de ontologias *web* e apresenta também os conceitos referentes à ontologia MONET. A Figura 14 descreve a modelagem dos conceitos da ontologia, assim com os seus respectivos relacionamentos, e apresenta em diferentes cores os conceitos relacionados ao metabolismo, interação proteína-proteína, regulação gênica e organismos (atributos não são apresentados). É importante destacar que alguns conceitos pertencem a mais de um domínio, como por exemplo, o conceito PROTEIN que participa no metabolismo, na interação proteína-proteína e regulação gênica.

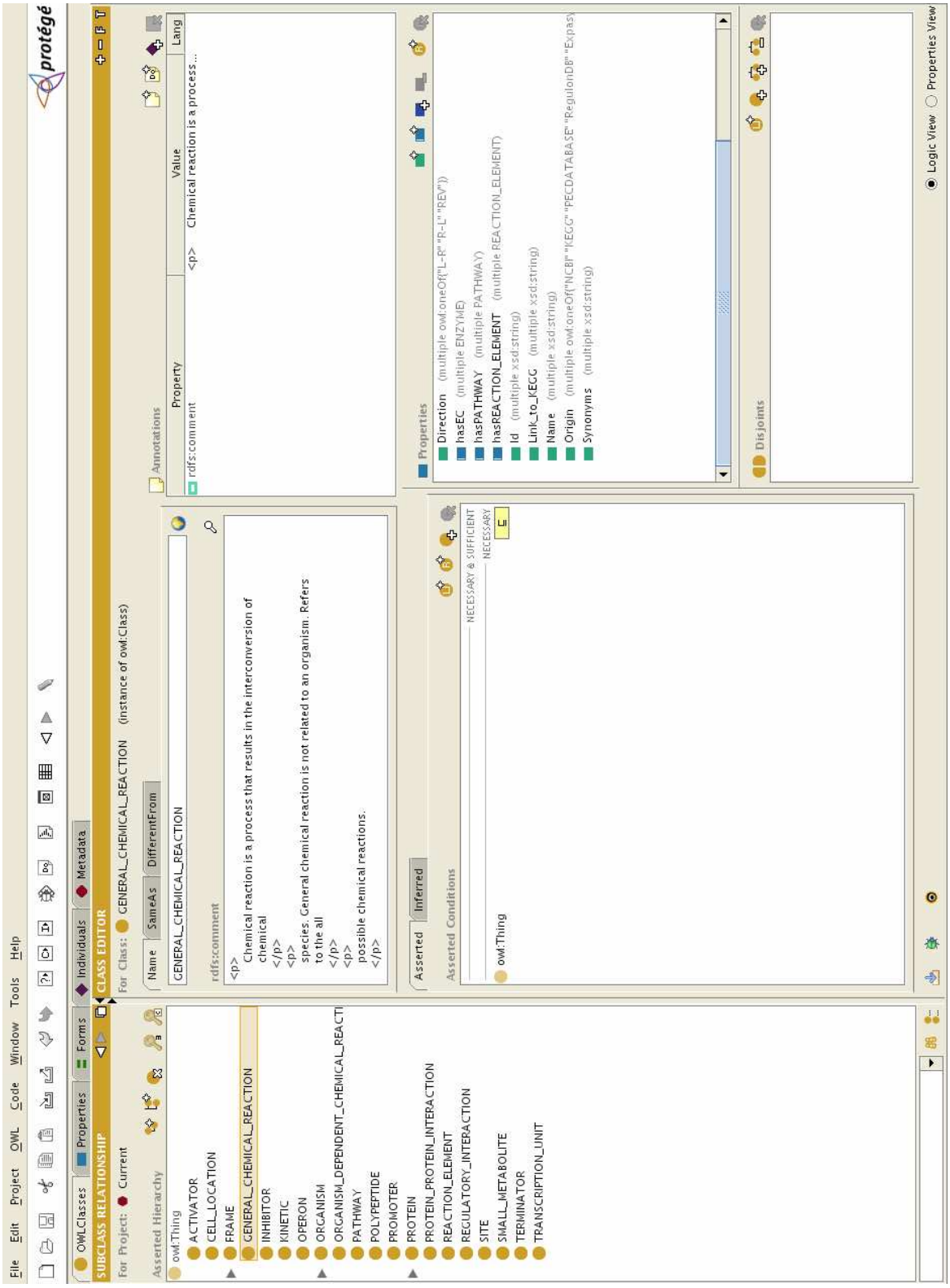

Figura 13: Interface de gerenciamento do ambiente Protégé, no formato OWL, versão 3.1 Beta.

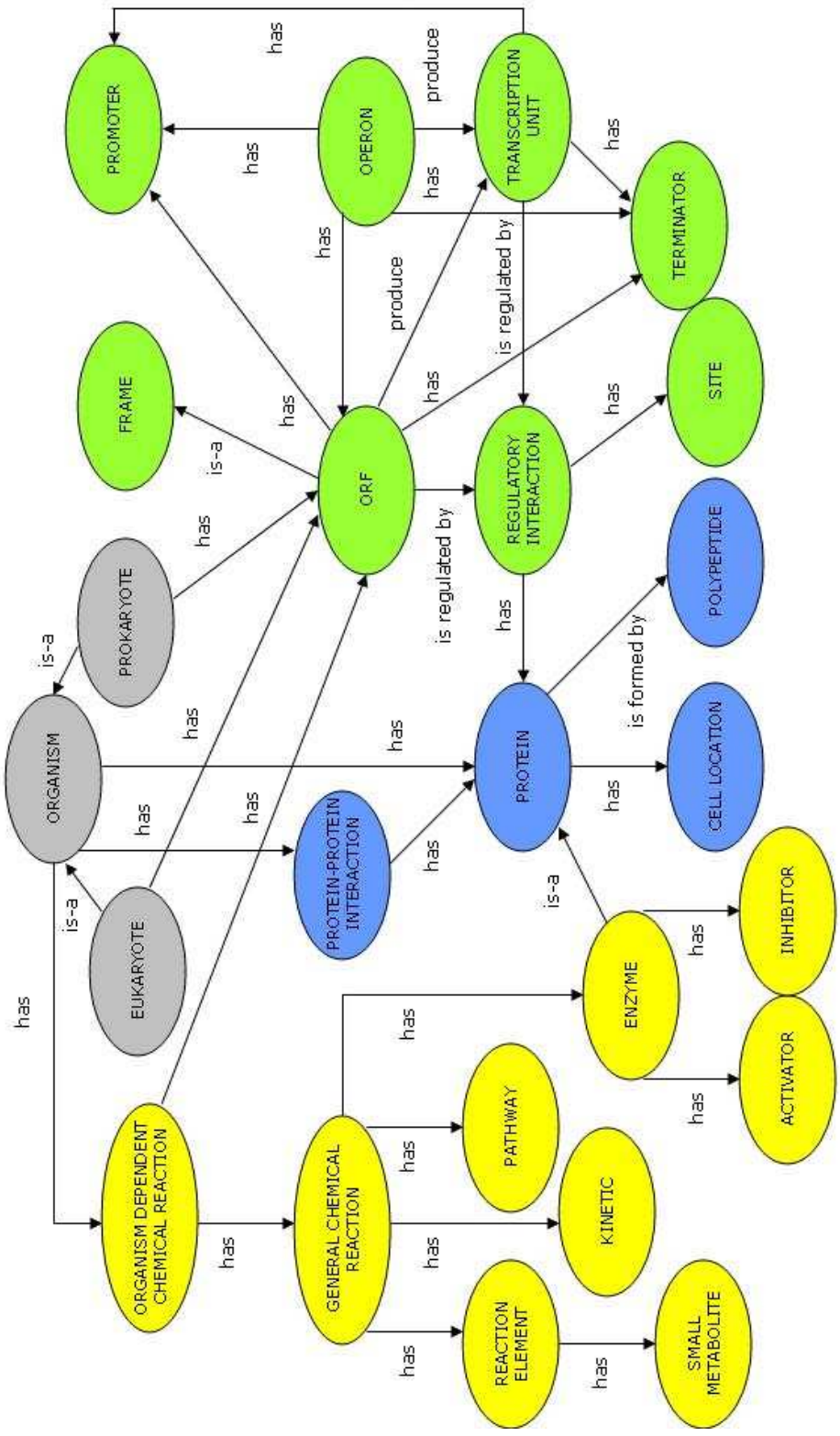

Figura 14: Modelagem da ontologia MONET.

O processo de desenvolvimento da modelagem da ontologia iniciou com estudos

sobre ontologias em um contexto geral e depois com ontologias na área de Bioinformática. A partir destes estudos e da modelagem inicial da ontologia proposta em (BATTISTELLA et al., 2004), iniciou-se a revisão da modelagem. Foram analisados os conceitos existentes, seus relacionamentos e seus nomes, de forma que fosse possível compreender o que cada conceito e cada relacionamento queriam expressar. A partir destes estudos, modificações foram realizadas tais como: a inserção das subclasses PROKARYOTES e EUKARYOTES na classe ORGANISM, estas duas subclasses foram adicionadas para especificar o tipo de organismo. Esta modificação foi necessária porque existem estruturas que os seres procariotos possuem e os eucariotos não, como, a estrutura de *operon*. O conceito REPRESSOR existente na versão inicial da modelagem foi excluído. Houve também casos de classes que tornaram-se atributos de outra classe, como é o caso de SUBSTRATE e PRODUCT que na modelagem inicial estavam modeladas como duas classes se relacionando com a classe SMALL METABOLITE, de forma que os compostos que fossem substratos ficariam armazenados na classe SUBSTRATE e os produtos armazenados na classe PRODUCT, no entanto todos os compostos já estavam armazenados na SMALL METABOLITE, o que representaria redundância de informação, visto que ainda os compostos que fossem substratos seriam armazenados na classe SUBSTRATE e os que fossem produtos na classe PRODUCT, e caso os compostos fossem substratos e produtos ao mesmo tempo, o que ocorre em reações químicas reversíveis, seriam armazenados em ambas as classes, resultando em informações repetidas duas vezes. Para solucionar este problema na classe REACTION ELEMENT, foi adicionada uma propriedade denominada Place para distinguir substratos e produtos. O atributo Place possui dois valores possíveis, L (*left*) que indica os substratos e R (*right*) que indica os produtos. A classe REACTION ELEMENT se relaciona com a classe SMALL METABOLITE, onde a primeira classe indica se o composto é substrato ou produto e a segunda informa qual é o composto evitando duplicação. Alguns conceitos e alguns relacionamentos tiveram seus nomes alterados para garantir sua adequação aos padrões da Biologia. Além disso, para todos os conceitos presentes na ontologia foram elaboradas definições formais, o que na versão inicial da modelagem não existia.

A criação da modelagem da ontologia ocorreu em conjunto com um especialista da Biologia Molecular. Todas as mudanças feitas na ontologia foram apresentadas e discutidas com este especialista. O especialista foi o responsável pela correção dos detalhes biológicos envolvidos na modelagem, tais como, nome de conceitos e de relacionamentos. Estas discussões sobre a ontologia uniram duas áreas diferentes e permitiram que ambas trabalhassem em conjunto para resolver um problema existente dentro da área abarcada pela Bioinformática e conseqüentemente resultaram em uma linguagem comum compartilhada por ambas as áreas.

Além do especialista na área da Biologia Molecular, também foi usado o RACER em conjunto com o ambiente Protégé, como ferramenta para checar a consistência na modelagem da ontologia. E durante a execução desta checagem, uma inconsistência apareceu. Nas subclasses PROKARYOTES e EUKARYOTES havia sido definido um atributo denominado hasORF que faz referência à classe ORF. No entanto, tanto os organismos procariotos como eucariotos possuem ORF, então o RACER indicou que este atributo deveria ser definido na classe ORGANISM, visto que PROKARYOTES e EUKARYOTES são subclasses de ORGANISM, e portanto este atributo é herdado da sua superclasse.

5.2 Inclusão dos dados biológicos

Com a revisão da modelagem, as alterações feitas e explicadas na seção anterior, o processo de modelagem e especificação da ontologia encontra-se finalizado e a etapa seguinte é localizar na *web* bancos de dados biológicos públicos que comportam os dados referentes ao domínio da ontologia.

Conforme já apresentado no início deste Capítulo a ontologia MONET compreende caminhos metabólicos, regulação gênica e interação proteína-proteína.

Para popular a sua base de instâncias com dados biológicos, primeiramente é necessário adquirir estes dados, posteriormente aplicar um *parser* em cada um dos bancos de dados, em seguida centralizá-los em um local físico de armazenamento e por fim

modelá-los de acordo com a modelagem conceitual da ontologia. Cada uma destas etapas é detalhada a seguir:

5.2.1 Aquisição dos dados

A primeira etapa consiste no processo de extração (aquisição) dos dados a partir dos seus respectivos bancos de dados. Os dados são adquiridos por meio de *download*, visto que são disponibilizados em arquivos texto. A Tabela 1 apresenta a lista dos bancos de dados biológicos utilizados para a aquisição dos dados.

Banco de Dados	Endereço de URL
Brite	*http://www.genome.ad.jp/brite/*
Expasy	*http://bo.expasy.org*
KEGG	*http://www.genome.jp/kegg/*
Nature Feb. 2005	(BUTLAND et al., 2005)
NCBI	*http://www.ncbi.ncm.nih.gov/*
PEC	*http://gottani.lab.nig.ac.jp/ecoli/pec/index.jsp*
RegulonDB	*http://www.cifn.unam.mx/Computational_Genomics/regulondb*

Tabela 1: Bancos de dados biológicos usados na aquisição dos dados para a geração da base de instâncias da ontologia MONET.

5.2.2 Normalização e integração dos dados

A segunda etapa corresponde ao processo de normalização e integração dos dados. Nesta etapa, primeiro é necessário à criação de um *parser*, que é um programa de computador e tem como finalidade manipular os dados oriundos dos diversos bancos de dados biológicos por meio de arquivos texto. Este *parser* é baseado em um conjunto de regras de acordo com a necessidade do arquivo disponibilizado e a saída gerada por este *parser* pode ser um ou mais arquivos normalizados, dependendo da estrutura do(s) arquivo(s) de entrada (arquivo(s) de *download*) e dos dados presentes nestes arquivos. A complexidade da programação do *parser* é diferente para cada banco de dados e depende também da forma como os dados estão organizados nos arquivos disponibilizados pelo banco de dados.

O objetivo da normalização é colocar os dados em um formato padrão único de

forma que os mesmos possam ser integrados. A Figura 15 apresenta de uma forma visual como ocorre o processo de normalização dos dados. Na mesma figura é apresentado um exemplo extraído do banco de dados do KEGG, onde em (a) tem-se o arquivo original, em (b) a aplicação do *parser*, em (c), (d), (e) e (f) são apresentados os novos arquivos gerados pelo processo de normalização.

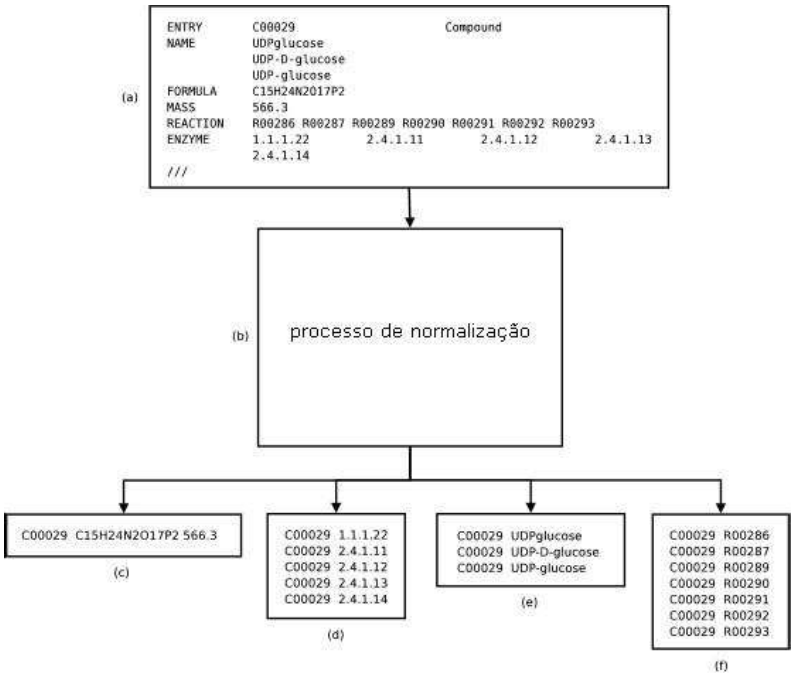

Figura 15: Representação do processo de normalização.

Com o processo de normalização concluído é então executado o próximo passo que é a integração dos dados. Para a execução desta etapa é necessária à utilização de uma abordagem de integração, e a abordagem utilizada neste trabalho é a abordagem materializada, onde os dados são extraídos dos bancos de dados e posteriormente centralizados em um único repositório, neste caso, o local é o SGBD PostgreSQL (*http://www.postgresql.org*). Esta abordagem implementa a arquitetura de *Data Warehouse* (INMON, 1997), e foi escolhida por ser a mais apropriada quando há necessidade de aquisição e o armazenamento dos dados em um repositório único e centralizado, de forma a permitir o estudo e análise destes dados em um processo minucioso.

Com a abordagem de integração escolhida é o momento de dar início ao processo de integração, e este é executado da seguinte maneira: Para cada arquivo gerado pelo processo de normalização é criada uma tabela no ambiente PostgreSQL, através da linguagem SQL, que é a linguagem usada pelo ambiente para a criação de tabelas e manipulação dos dados. Juntamente com a criação de cada uma das tabelas, os dados referentes a cada tabela são carregados. Este processo é executado para todos os arquivos gerados pelo processo de normalização de todos os bancos de dados.

Com a criação das tabelas e a sua respectiva carga (colocação dos dados dos arquivos normalizados nas tabelas criadas), ou seja, cada tabela representa um arquivo do processo de normalização com todos os seus dados, assim o processo de integração dos dados está completo.

No entanto, ainda é necessária a criação das tabelas referentes à ontologia MONET. A criação destas tabelas é realizada de acordo com a modelagem conceitual desenvolvida no ambiente Protégé. Para cada classe presente na modelagem, é criada uma tabela no banco de dados, e cada um dos atributos da classe será um campo da tabela do banco de dados. Algumas tabelas auxiliares são necessárias para representar os dados que somente com as tabelas referentes aos conceitos não são possíveis de representação. Os relacionamentos que existem na modelagem conceitual são preservados e também construídos no banco de dados, isto é, a modelagem conceitual será codificada em forma de tabelas.

Da mesma forma que no processo de integração, a cada tabela criada para representar os conceitos da ontologia MONET os dados são imediatamente inseridos, só que agora não são mais inseridos a partir dos arquivos texto, mas sim das tabelas que referenciam estes arquivos. No entanto, esta carga de dados é diferente, isto porque, os dados para compor um determinado conceito da ontologia, podem vir de várias tabelas, mas isso nem sempre ocorre. Para exemplificar, quando são adicionados os dados para o conceito PROTEIN-PROTEIN INTERACTION, estes são oriundos do banco de dados BRITE

(*http://www.genome.ad.jp/brite/*) e dados publicados na Revista Nature de Fevereiro de 2005 (BUTLAND et al., 2005), e estes dados encontram-se em forma de tabelas, que foram criadas no momento em que foi executada a integração de dados. Então estes dados são manipulados através da linguagem SQL para que os mesmos sejam inseridos na tabela que faz referência ao conceito PROTEIN-PROTEIN INTERACTION. Já alguns conceitos como por exemplo, o SMALL METABOLITE possui dados apenas do banco de dados do KEGG (*http://www.genome.jp/kegg/*), estes são inseridos através da tabela criada com dados do KEGG que tratam os dados de compostos. A Figura 16 ilustra de forma visual todo o processo, desde a etapa de aquisição dos dados, passando pelo processo de normalização e chegando ao processo de integração para o SGBD. Esta figura mostra também a saída dos dados para o ambiente Protégé.

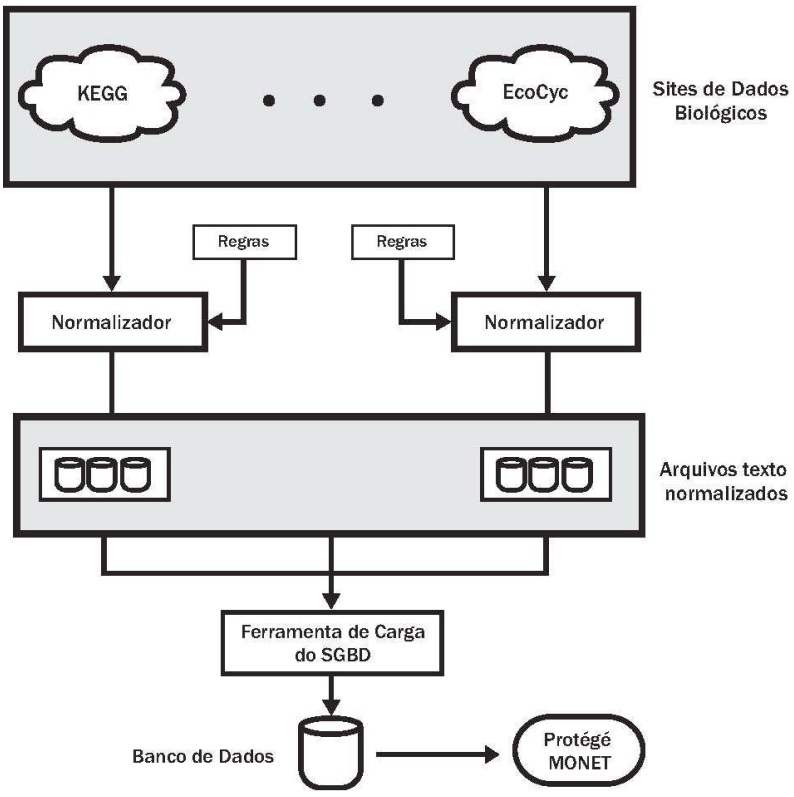

Figura 16: Representação completa do processo de aquisição, normalização e integração dos diversos bancos de dados biológicos utilizados para dentro do ambiente PostgreSQL, assim como a saída dos dados para a ferramenta Protégé, gerando assim a ontologia MONET.

5.2.3 Limpeza dos dados

A terceira e última etapa é a limpeza dos dados. Esta tarefa tem por objetivo eliminar ou corrigir dados incorretos e imprecisos. Esta é uma etapa executada junto ao especialista da área biológica. O especialista era o responsável por analisar os dados incorretos. No entanto, esta etapa não é essencial como às duas anteriores, porém, sem a sua execução os dados podem ficar incorretos na base de instâncias da ontologia, e isso pode influenciar em resultados futuros extraídos a partir destes dados.

5.3 Criação da base de instâncias da ontologia MONET

Com os dados biológicos dentro do ambiente PostgreSQL, a próxima etapa é agrupar estes dados para formar as instâncias da ontologia MONET. Este processo é executado em parte dentro do ambiente do banco de dados e parte fora dele. Na etapa executada dentro, os dados armazenados em forma de tabelas que compõem as classes, subclasses, atributos e instâncias são exportados para fora do ambiente PostgreSQL em arquivos texto. É um arquivo para cada tabela do banco de dados.

Na etapa realizada fora do ambiente PostgreSQL, todos estes arquivos exportados são interpretados por um *parser*, cujo é baseado em conjunto de regras que possibilita ler cada um dos arquivos exportados individualmente, juntar estes arquivos baseado nos relacionamentos definidos na modelagem conceitual da ontologia em um único arquivo e colocá-los em um arquivo no padrão OWL, cujo arquivo tem o tamanho de 64MB. O processo de carga da ontologia no formato OWL no ambiente Protégé demora cerca de $19,36$ minutos e é necessário um computador com memória superior à 512MB de memória RAM. No final este arquivo OWL conterá todas as classes, subclasses, atributos, relacionamentos e instâncias da ontologia MONET, assim como todas as definições formais de cada conceito (classe). A Tabela 2 apresenta todos os conceitos presentes na ontologia, assim como o seu respectivo número de instâncias. No entanto, alguns conceitos não

possuem instâncias, visto que não existe dados disponíveis para tais, como é o caso dos conceitos ACTIVATOR, INHIBITOR e KINETIC.

Conceito	Número de instâncias
ENZYME	4.560
GENERAL CHEMICAL REACTION	6.469
OPERON	784
ORF	10.615
ORGANISM	3
ORGANISM DEPENDENT CHEMICAL REACTION	5.838
PATHWAY	238
PROMOTER	973
PROTEIN	9.977
PROTEIN-PROTEIN INTERACTION	12.248
REACTION ELEMENT	18.194
REGULATORY INTERACTION	1.376
SITE	1.216
SMALL METABOLITE	23.954
TERMINATOR	137
TRANSCRIPTION UNIT	833

Tabela 2: Lista dos conceitos presentes na ontologia MONET bem como a sua respectiva quantidade de instâncias.

6 REDE INTEGRADA DA E. COLI

Com intuito de avaliar a ontologia MONET propomos e construímos uma rede gênica integrada para a bactéria *E. coli*, compreendendo dados de metabolismo, regulação gênica e interação proteína-proteína. A escolha por este organismo justifica-se por dois motivos principais: (a) o fato de ser um organismo procarioto com regulação bem compreendida e (b) apresentar o conjunto de dados mais extenso de regulação gênica e metabolismo. A Tabela 3 apresenta os bancos de dados utilizados para a construção da rede.

Rede de interação	Fonte dos dados
Metabolismo	*http://www.genome.jp/keg/*
Regulação	*http://www.cin.unam.mx/Computational_Genomics/regulondb*
Interação Proteína-Proteína	(BUTLAND et al., 2005)

Tabela 3: Fontes originais dos dados.

Na rede proposta os nodos são genes e os genes g1 e g2 que codificam as proteínas p1 e p2, estão conectados, se:

interação proteína-proteína : p1 e p2 interagem fisicamente;

regulação : g1 regula a transcrição do gene g2 ou

metabolismo : um produto gerado pela reação catalisada pela proteína p1 é consumido na reação catalisada pela proteína p2.

A Figura 17 apresenta um esquema para a rede integrada proposta. Na literatura não foi encontrada nenhuma referência a uma rede desta natureza.

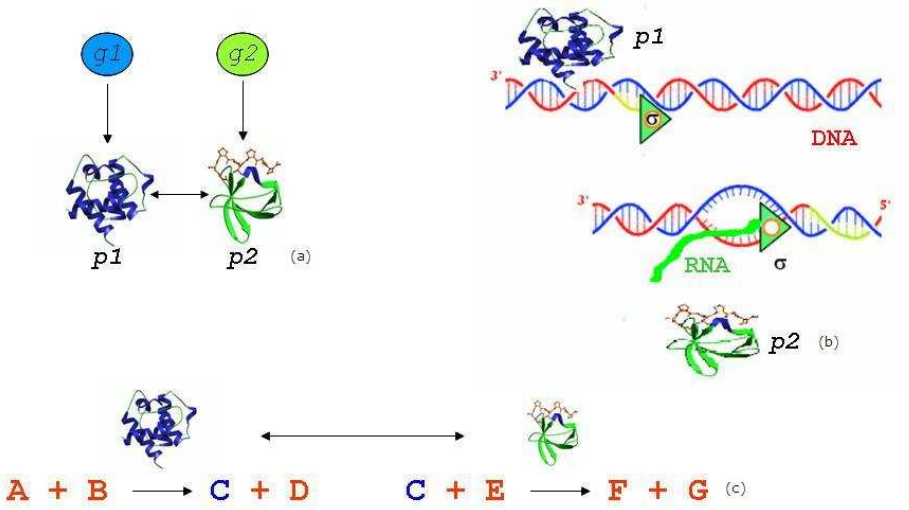

Figura 17: Estrutura da rede integrada da *E. coli*. Os três possíveis mecanismos de conexão da rede integrada: (a) interação proteína-proteína, (b) regulação gênica e (c) metabolismo.

Com base no procedimento descrito, obteve-se uma rede composta por um conjunto de 51.642 interações. A Figura 18 apresenta a distribuição dos genes de acordo com o seu respectivo número de interações. Esta rede apresenta a maioria dos seus genes, um conjunto de 1.938 genes com poucas conexões (1 a 50) e um grupo reduzido de 5 genes, altamente conectados, com mais de 400 interações (ver Figura 18). Desde o trabalho clássico de Barabási (BARABáSI; ALBERT, 1999) (JEONG et al., 2001) os nodos mais importantes de uma rede são considerados aqueles com o maior número de interações, a Tabela 4 apresenta a lista dos 10 genes mais conectados.

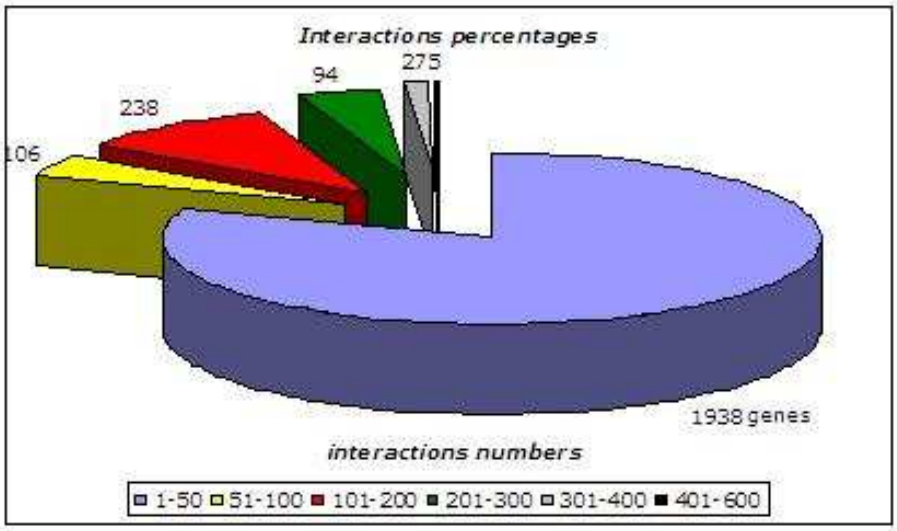

Figura 18: Distribuição do número de interações para os genes na rede da *E. coli*.

Genes	Número de interações
metK	516
pdxB	467
rspB	467
wecC	467
sixA	454
ygdP	392
ntpA	354
astD	351
nudG	351
ybbF	351

Tabela 4: Lista dos 10 genes mais conectados na rede integrada da *E. coli*, considerando todos os compostos.

Uma parte importante das interações desta rede é devido a interações metabólicas, que ocorrem devido à produção-consumo de compostos como ATP, ADP, NADH que apesar de serem importantes, dada a sua presença maciça não são tão importantes para determinar a essencialidade de um gene.

Por este motivo consideramos uma nova rede da qual foram excluídos os 10 compostos que participam de reações metabólicas. Esta nova rede é composta por um conjunto de 21.338 interações. A Figura 19 apresenta a distribuição dos genes de acordo com o

número de interações. Da mesma forma da rede anterior, esta rede apresenta a maioria dos seus genes com poucas conexões, 2.221 possuem de (1 a 50) conexões e um grupo reduzido de genes (4) é altamente conectado, com mais de 200 conexões. Conforme o esperado observa-se que ocorreu uma diminuição significativa do número de interações dos nodos mais conectados. A Tabela 5 apresenta a nova lista dos 10 genes mais conectados para a rede da *E. coli*, considerando a exclusão dos 10 compostos mais conectados no metabolismo.

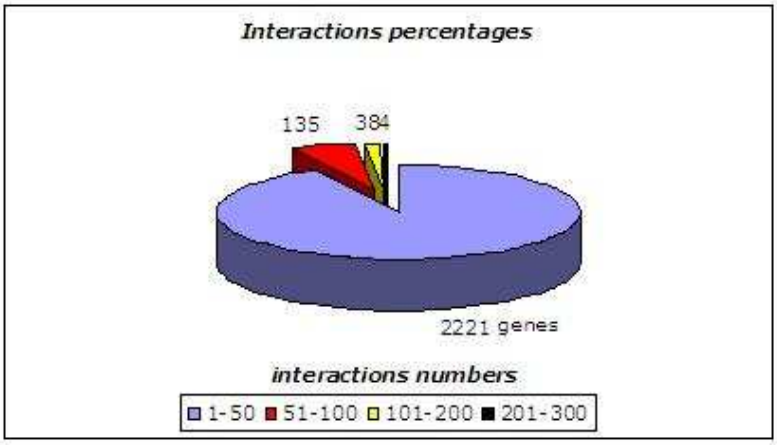

Figura 19: Distribuição do número de interações para os genes na rede da *E. coli*.

Genes	Número de interações
crp	272
pdxB	221
rspB	221
wecC	221
aceE	200
ihf	169
nadC	167
fis	156
trpD	155
lpdA	154

Tabela 5: Lista dos 10 genes mais conectados na rede integrada da *E. coli*. Para a construção desta rede, foram excluídos os 10 compostos que mais aparecem no metabolismo.

Já foi demonstrado (BARABáSI; OLTVAI, 2004) que as redes metabólicas e a rede de interação proteína-proteína são livres de escala, ou seja, $P(k) \sim k^{-b}$ (veja Capítulo 3),

já a rede regulatória não é livre de escala. Na Figura 20 apresentamos o $P(k)$ para três redes possíveis: com todos os metabólicos, sem os cinco mais conectados e sem os dez mais conectados. As linhas apresentam os melhores ajustes para os dados. Observa-se que o comportamento dos dados é qualitativamente o mesmo nos três casos.

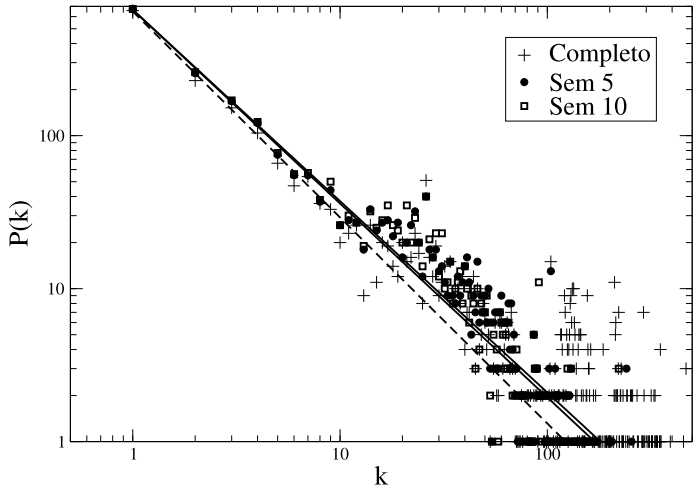

Figura 20: Distribuição do $P(k)$ das redes integradas da *E. coli*, sem os 5 compostos e sem os 10 compostos mais conectados no metabolismo. Em todos os casos a rede é livre de escala.

Uma questão relevante é determinar se a rede construída é uma rede hierárquica, isto é, analisar se a rede integrada possui pequenos mais integrados módulos que são ligados aos outros nodos da rede, formando assim uma estrutura de hierarquia. Com este intuito medimos $C(k)$, os resultados são apresentados na Figura 21 que mostra o coeficiente de clusterização da rede. Os resultados neste caso mostram que a rede completa não apresenta hierarquia, enquanto que nos demais casos $C(k) \sim k^{-a}$. Na Figura 22 apresentamos a dependência dos parâmetros a e b com o número de metabólicos excluídos da rede metabólica, as barras de erro representam o intervalo de confiança do ajuste. Note que para a rede completa, estatisticamente não se pode descartar a possibilidade de $C = cte$ e a consideramos, portanto como uma rede não hierárquica. Os ajustes foram

realizados utilizando o *software* Mathematica 5.1.

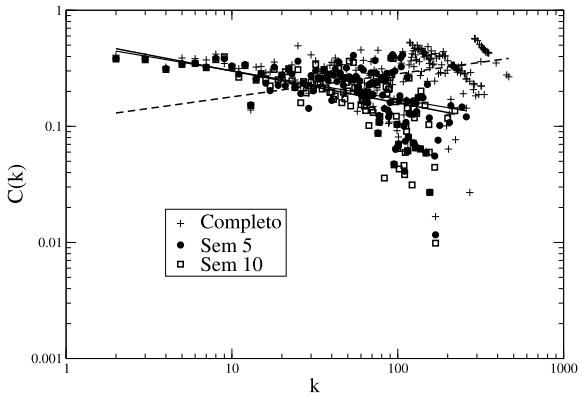

Figura 21: Coeficiente de clusterização $C(k)$ das redes da *E. coli*: rede completa, rede sem os 5 e sem os 10 compostos mais utilizados no metabolismo. As linhas representam o melhor ajuste nos dados. Os dados indicam que a rede completa é não hierárquica, enquanto que as outras redes possuem esta propriedade.

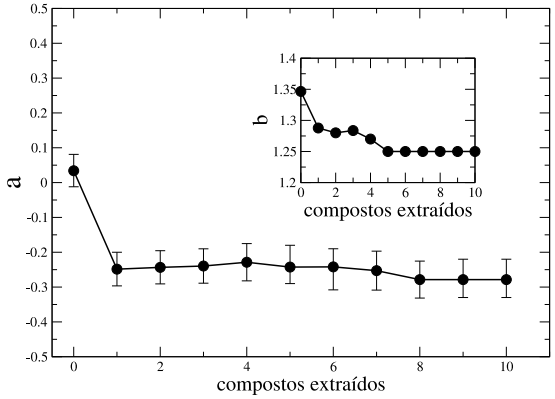

Figura 22: Parâmetro de ajuste para $P(k)$ (detalhe) e $C(k)$ em relação ao número de compostos excluídos da rede integrada da *E. coli*. Observe que a rede completa aparentemente é não hierárquica e que os parâmetros de ajuste se estabilizam para as redes com mais de 5 compostos excluídos.

6.1 Essencialidade dos genes

Uma questão relevante do ponto de vista biológico é a determinação de quais genes são essenciais, ou seja, aqueles que se forem excluídos do genoma implicam na incapacidade do organismo de sobreviver. Diferentes metodologias foram propostas (SEGRE; VITKUP; CHURCH, 2002) (IMIELINSKI et al., 2005) (PALUMBO et al., 2005) para determinar com base em informações topológicas quais genes possuem esta característica. A rede proposta na seção anterior pode ser utilizada também com esta finalidade. O conjunto de genes essenciais é em geral uma pequena parte dos genes de um organismo, isto pode parecer surpreendente, mas é facilmente explicável pelo fato de que a evolução tende a garantir que todos os sistemas biológicos sejam redundantes, pois isto é um mecanismo importante para garantir a sobrevivência dos organismos. As metodologias propostas em geral visam maximizar a cobertura da classe de genes essenciais ao mesmo tempo em que minimizam o número de falsos positivos.

Tradicionalmente os genes mais conectados são considerados aqueles com maior probabilidade de serem essenciais (BARABáSI; ALBERT, 1999) (JEONG et al., 2001). Para a rede completa, o gene *metK* é o mais conectado possuindo um total de 516 conexões com outros genes dentro da rede. Este gene codifica a enzima *metionina-adenosil transferase* (*EC: 2.5.1.6*), que por sua vez atua em várias vias metabólicas, como, a degradação de *metionina*, *treonina*, *isoleucina* e *valina* no metabolismo do selenoaminoácido. Esta enzima catalisa a formação de *S-adenosilmetionina*. A *adenosilmetionina* ocupa uma posição metabólica central e atua como maior doador do grupo *metil* em sistemas biológicos.

Para a rede cujo os 10 compostos mais conectados no metabolismo foram desconsiderados, o gene que mais se destaca é o gene *crp* que codifica a proteína CRP (proteína receptora de adenosina monofosfato-cíclico (AMPc)). Este gene codifica um fator de transcrição muito importante envolvido principalmente no catabolismo, que é a quebra de nutrientes para gerar energia e produzir moléculas mais simples, neste caso, quebra de outras fontes de carbono diferentes da glicose. A *E. coli* faz uso preferencialmente da

glicose como fonte de carbono e energia e somente utiliza outros açúcares quando a glicose começa a faltar. A presença de glicose previne o catabolismo de outros açúcares através de alguns mecanismos, e um deles a glicose baixa o nível de AMPc no interior da célula, e como o AMPc é o indutor da proteína CRP, ela não consegue ativar a transcrição dos genes envolvidos no catabolismo de fontes alternativas de carbono. A proteína CRP atua também no controle da transcrição de genes necessários para a produção de energia, metabolismo de aminoácidos, de nucleotídeos e sistema de transporte de íons. Além disso, CRP pode regular a transcrição de outros fatores de transcrição, como, MelR, RpoH, BlgC, Fis e PdhR.

A rede integrada contudo, nos permite realizar análises mais sofisticadas baseadas em Inteligência Artificial visando à proposição de métodos mais confiáveis que os tradicionais para determinar quais genes são essenciais. Com este objetivo utilizamos a base de dados PEC (*http://www.shigen.nig.ac.jp/ecoli/pec/*) que contém informação experimental sobre a essencialidade do gene, quando esta é conhecida. Os genes com essencialidade desconhecida foram desconsiderados.

Para a realização do experimento para a predição da essencialidade dos genes, além dos dados da rede integrada da *E. coli* foi agregada à informação de dano dos genes presentes na rede. Estes dados foram retirados de (LEMKE et al., 2004) e não se encontram presentes na ontologia. O dano é definido como um critério quantitativo para enumerar o efeito da deleção de uma enzima, e foi demonstrado como um parâmetro útil para detectar a essencialidade de enzimas (LEMKE et al., 2004).

Esta aplicação se desenvolveu no ambiente WEKA (*Waikato Environment for Knowledge Analysis*) (WITTEN; FRANK, 2000). O WEKA é uma coleção de algoritmos de aprendizado de máquina para tarefas de mineração de dados. Contém ferramentas para pré-processamento, classificação, regressão, clusterização, regras de associação e visualização. Neste trabalho, foi utilizada uma ferramenta de classificação através do algoritmo J48 que implementa o método de árvores de decisão.

Segundo a definição apresentada em (REZENDE, 2003), uma árvore de decisão é uma estrutura de dados definida recursivamente como:

- um *nó folha* que corresponde a uma classe ou
- um *nó de decisão* que contém um teste sobre algum atributo. Para cada resultado do teste existe uma aresta para uma subárvore. Cada subárvore tem a mesma estrutura que a árvore.

Uma árvore de decisão pode ser representada como um conjunto de regras, isto é, a regra inicia pelo topo da árvore e segue até uma de suas folhas. Como as regras que representam uma árvore de decisão são disjuntas, apenas uma única regra pode ser executada quando um novo exemplo é classificado.

As regras nada mais são do que a implementação em qualquer ambiente de programação de um conjunto de *if's*, isto é, uma árvore de decisão é facilmente transposta, basta a implementação das suas regras, partindo do topo da árvore até as regras mais específicas (nós de decisão).

O processo de análise foi constituído com a realização de cinco experimentos utilizando os seguintes atributos:

- nome do gene;
- número de interações entre as proteínas;
- número de enzimas que produzem compostos consumidos pela enzima codificada pelo gene (`metabolismo_in`);
- número de enzimas que consomem compostos produzidos pela enzima codificada pelo gene (`metabolismo_out`);
- número de genes que regulam o gene de interesse (`regulacao_in`);

- número de genes que um gene regula (`regulacao_out`);
- dano: número de enzimas deletadas até que um organismo não consiga sobreviver;
- essencialidade: informa se o gene é indispensável (essencial) para que o organismo continue vivo.

Destes cinco experimentos realizados, dois deles contemplam o uso do atributo dano, dois utilizaram a rede sem os 10 compostos mais conectados e um deles foi realizado sem a replicação dos dados (explicação na análise dos dados).

Para a execução destes experimentos no ambiente WEKA foi usado o algoritmo J48, com *10 fold cross-validation* e com os parâmetros apresentados na Tabela 6, apresentados aqui com o objetivo de garantir a reprodutibilidade dos experimentos. Para garantir árvores mais simples de serem interpretadas e com maior probabilidade de poderem ser generalizadas para outros organismos, utilizamos minNumObj=100, o que garante um número mínimo de instâncias por regras.

Parâmetro	Valor
binarySplit	False
confidenceFactor	0.25
debug	False
minNumObj	100
numFolds	10
reduceErrorPruning	False
saveInstanceData	False
seed	1
subtreeRaising	True
unpruned	False
useLaplace	False

Tabela 6: Lista dos parâmetros e seus respectivos valores para a geração dos resultados apresentados para a predição da essencialidade de um gene.

6.2 Análise dos resultados

Nestas análises houve a necessidade da replicação dos dados para a classe E (*essential*), isto foi necessário pelo desbalanceamento nos dados entre as classes E e N (*non-essential*). O número de genes da classe N é muito superior aos da classe E. De qualquer forma, é apresentado o melhor resultado obtido sem a replicação dos dados.

Os resultados conseguidos pelas análises dos cinco experimentos realizados são apresentados na Tabela 7. Nesta tabela, na coluna replicação há uma indicação de sim ou não para informar se houve replicação dos dados nas análises. Na coluna Atrib. Dano, há dois valores possíveis, sim e não, onde o sim corresponde que o atributo dano foi considerado e não que o atributo dano não foi considerado na realização do experimento, a coluna completo com a indicação de sim ou não informa se a rede completa foi utilizada no experimento. A sigla ICC corresponde a Instâncias Classificadas Corretamente e ICI significa Instâncias Classificadas Incorretamente. Para a análise dos dados foram usadas as seguintes medidas: Precisão, Cobertura e Medida-F (WITTEN; FRANK, 2000).

A precisão é a proporção dos exemplos verdadeiros da classe x entre todos aqueles que foram considerados como pertencentes à classe x. Na matriz de confusão é o elemento da diagonal dividido pela soma da coluna relevante.

A cobertura é a proporção de exemplos que são classificados como sendo da classe x entre todos os exemplos que verdadeiramente são da classe x, isto é, quantos exemplos foram capturados. Na matriz de confusão é a diagonal dividida pela soma dos exemplos classificados corretamente mais ou classificados incorretamente.

A medida-f é uma medida combinada pela precisão e pela cobertura, e é representada pela seguinte fórmula:

$$medida - f = 2 * precisao * cobertura/(precisao + cobertura) \tag{6.1}$$

Resultados	Análise 1	Análise 2	Análise 3	Análise 4	Análise 5
Num. Instâncias	3.879	3.868	3.879	3.868	1.998
Replicação	sim	sim	sim	sim	não
Atrib. Dano	sim	sim	não	não	não
Completo	sim	não	sim	não	sim
ICC	3.161	3.138	3.172	3.138	1.794
ICI	718	730	707	730	204
Medida-F (N)	0.794	0.787	0.797	0.787	0.943
Medida-F (E)	0.832	0.831	0.834	0.831	0.49
Cobertura (N)	0.774	0.758	0.778	0.757	0.948
Cobertura (E)	0.85	0.856	0.852	0.857	0.469
Precisão (N)	0.816	0.818	0.818	0.819	0.939
Precisão (E)	0.814	0.806	0.818	0.806	0.513

Tabela 7: Resultados gerados pelas análises dentro do ambiente WEKA.

Os resultados apresentados nas próximas tabelas, correspondem as matrizes de confusão para cada uma destas análises apresentadas na Tabela 7, juntamente com uma breve análise dos resultados obtidos por elas. As Tabelas 8, 9, 10, 11 e 12 apresentam as matrizes de confusão para as análises 1, 2, 3, 4 e 5, respectivamente.

Classe N	Classe E	classificada como
1.384	405	N
313	1.777	E

Tabela 8: Matriz de confusão da análise 1.

Classe N	Classe E	classificada como
1.348	430	N
300	1.790	E

Tabela 9: Matriz de confusão da análise 2.

Classe N	Classe E	classificada como
1.392	397	N
310	1.780	E

Tabela 10: Matriz de confusão da análise 3.

Classe N	Classe E	classificada como
1.346	432	N
298	1.792	E

Tabela 11: Matriz de confusão da análise 4.

Classe N	Classe E	classificada como
1.696	93	N
111	98	E

Tabela 12: Matriz de confusão da análise 5.

De acordo com os resultados apresentados pelas diferentes análises, os mesmos são praticamente equivalentes, com uma variação pequena na predição, onde a pior classificação obteve 85% de cobertura, encontrando 1.777 genes para a classe E e na melhor classificação, uma cobertura de $85,7\%$, ou seja, 1.792 genes. A diferença entre a pior e a melhor classificação foi de 15 genes. A melhor classificação apresenta replicação nos dados, faz uso da rede sem os 10 compostos mais conectados no metabolismo e não utiliza o atributo dano. Com relação à quinta análise na qual não é considerada a replicação dos dados para a classe E, observa-se que a cobertura da classe essencial é baixa, $46,9\%$, considerando a rede completa da *E. coli*. Apesar de que a exatidão nestes casos seja maior a cobertura da classe das essenciais é muito baixa, tornando esta metodologia de pouco interesse prático.

A Figura 23 apresenta o conjunto de regras que representa a árvore de decisão de melhor resultado com replicação dos dados, e a Figura 24 apresenta a árvore de decisão referente ao melhor resultado das análises sem a replicação dos dados.

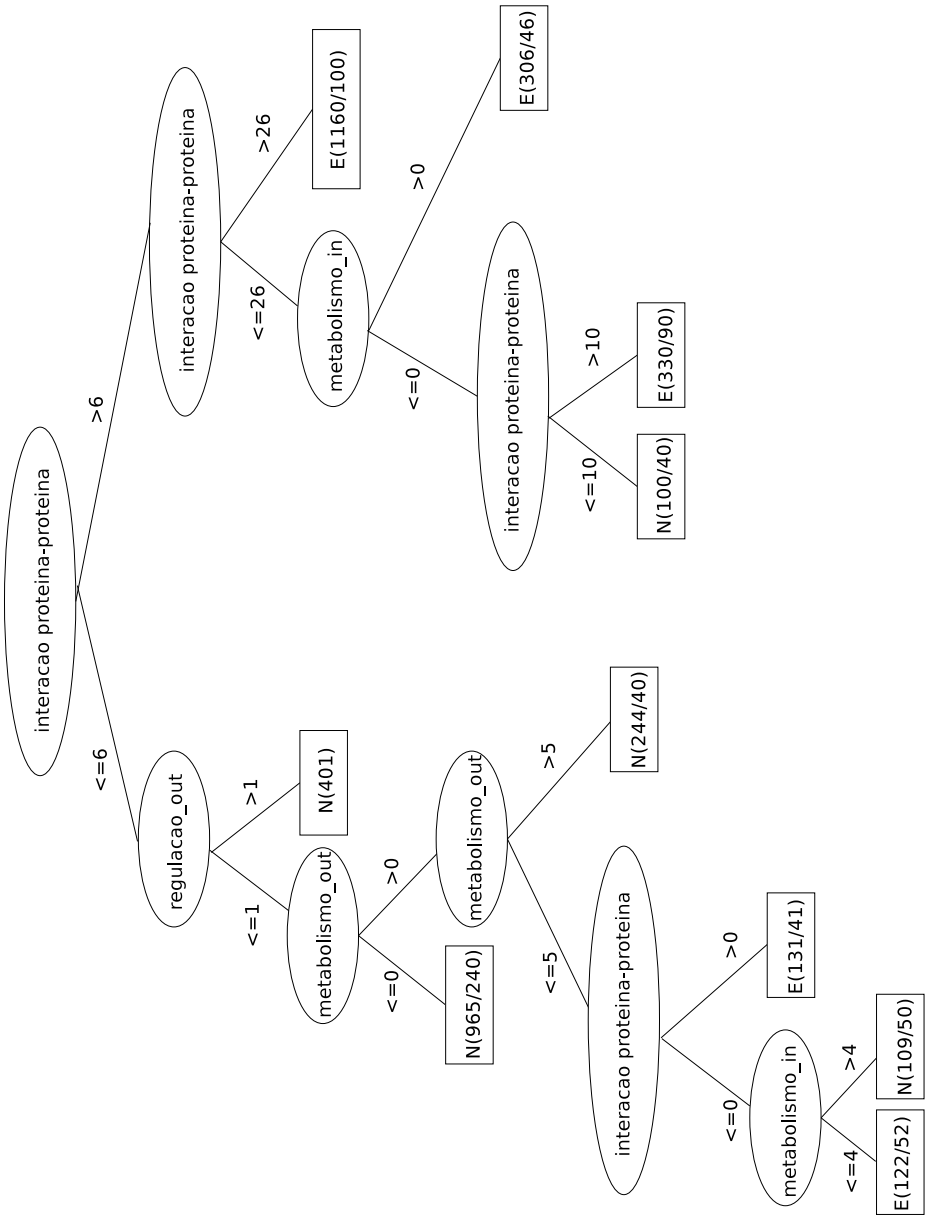

Figura 23: Árvore de decisão gerada pela melhor análise, a qual apresenta uma cobertura de $87,5\%$.

Conforme a árvore gerada pela melhor classificação, observa-se que o topo da árvore é a interação proteína-proteína, considerando essenciais os genes com no mínimo 6 interações. Na verdade, todas as árvores geradas pelos diversos testes realizados apresen-

tam a interação proteína-proteína como o topo da árvore. No entanto, em alguns casos o número de interações entre os genes variou.

Na rede da *E. coli*, genes com grande quantidade de interação proteína-proteína tendem a ser essenciais. Além disso, proteínas com número intermediário de interação proteína-proteína e que sejam enzimas também tendem a ser essenciais. A `regulacao_in` é um atributo não significativo para a predição da essencialidade, tanto é que o mesmo não aparece na árvore. Já a `regulacao_out` indica que se o gene possuir mais de 1 interação, o mesmo já é classificado como não essencial. Se o gene possuir 0 ou mais do que 5 interações no `metabolismo_out`, o gene é classificado como não essencial. Esta última característica é interessante pois mostra que enzimas que produzem compostos não muito utilizados podem ser essenciais.

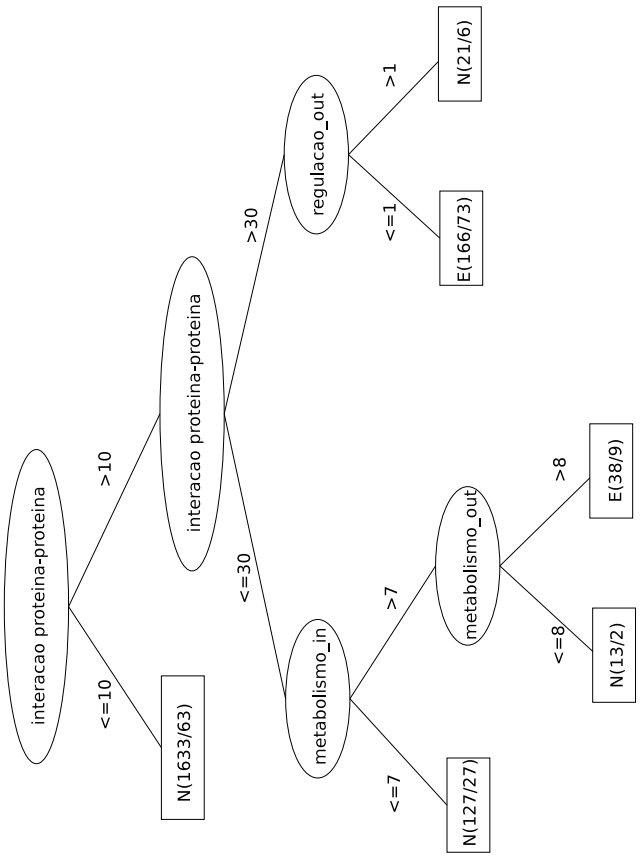

Figura 24: Árvore de decisão gerada pela melhor análise, sem a replicação dos dados para a classe E e que apresenta uma cobertura de $46,9\%$.

De forma equivalente, a análise da rede sem replicação dos dados, apresenta como topo da árvore a interação proteína-proteína. Se o gene possuir 10 ou menos interações na interação proteína-proteína, este gene é classificado como não essencial. Genes com interações no `metabolismo_in` maiores do que 7 e menores do que 8 no `metabolismo_out` são classificados como essenciais. Na `regulacao_out` genes com número de interações 0 ou 1 são apontados como essenciais e a `regulacao_in` é irrelevante.

Em ambas as análises observa-se uma concordância qualitativa dos resultados, apesar de que no segundo caso foi possível recuperar um número muito maior de genes essenciais, o que para todos os fins práticos é mais relevante. O dano não se mostrou determinante para a determinação da essencialidade dos genes, aparentemente por que esta informação é redundante com os dados do metabolismo.

7 CONCLUSÕES E CONSIDERAÇÕES FINAIS

Este trabalho apresentou uma metodologia para o desenvolvimento de uma ontologia biológica. Começando pelo processo de revisão da modelagem e especificação da ontologia biológica MONET (BATTISTELLA et al., 2004) (BATTISTELLA et al., 2005), que incluiu as seguintes etapas: (a) estudos sobre ontologias em um contexto geral e ontologias na área de Bioinformática; (b) revisão da modelagem inicial; (c) modificações na modelagem da ontologia; (d) especificação no formato OWL e (d) definições formais dos conceitos presentes na ontologia. Esta ontologia tem como objetivo ser um modelo integrado para a *rede de redes* que existe dentro da célula (BARABáSI; OLTVAI, 2004).

Agregados ao desenvolvimento da ontologia, foram agrupados dados biológicos para que fosse possível a criação das instâncias, e conseqüentemente o desenvolvimento de aplicações para a extração de conhecimento biológico.

Com o desenvolvimento da ontologia resolveram-se os problemas de padronização, incoerência e organização encontrada nos dados biológicos, visto que estes são problemas existentes na área de Bioinformática devido aos dados estarem distribuídos geograficamente na *web*, e pela enorme e complexa quantidade de dados coletados nos últimos anos e ainda pelo fato de que cada banco de dados trata os dados da forma que melhor lhe convém, sem a preocupação com uma possível utilização dos mesmos por outras aplicações.

A ontologia apresenta ainda uma definição formal sobre cada um dos conceitos, o que torna possível a qualquer aplicação que futuramente faça uso (reuso e compartil-

hamento são características fundamentais das ontologias) da ontologia, o conhecimento sobre estes conceitos, ou seja, o conhecimento que cada um representa. Atualmente isto não ocorre com alguns bancos de dados, que simplesmente disponibilizam os dados e não expõem o que estes dados representam.

A partir da especificação da ontologia, aquisição e integração dos dados biológicos referente a cada um dos domínios da ontologia, foi possível à realização de experimentos, tais como, a construção da rede integrada da *E. coli* e a predição da essencialidade de um gene, os quais visavam validar a ontologia.

Com a construção da rede integrada da *E. coli* foi possível validar e avaliar a ontologia com relação ao seu domínio e a sua modelagem, isto porque para esta aplicação foram utilizados dados dos três domínios (metabolismo, interação proteína-proteína e regulação gênica) existentes na ontologia MONET. Com este experimento foi possível mostrar as interações dos genes nos três domínios e descobrir quais eram os genes mais conectados para este organismo.

A predição da essencialidade de um gene por sua vez se baseou na rede integrada da *E. coli* com objetivo de prever a qualidade dos dados armazenados na ontologia para predizer a essencialidade de um gene baseado em todas as suas interações na rede integrada, e o melhor resultado obtido foi uma cobertura de $85,7\%$ na taxa de acerto.

A ontologia serviu como banco de dados para a construção das duas aplicações desenvolvidas. Desde modo, serviu como fonte de informação, permitindo que inferências fossem realizadas em cima dos dados armazenados. A partir deste ponto de vista, observa-se que não é mais necessário o uso das fontes originais para o desenvolvimento de aplicações do porte das realizadas neste trabalho. Assim as fontes originais ficam sendo necessárias apenas para a atualização dos dados da ontologia.

Através do desenvolvimento das aplicações e baseado nos resultados obtidos por elas, foi possível demonstrar a utilidade do conhecimento biológico modelado à execução destes experimentos e conseqüentemente a extração de resultados. Outra questão perti-

nente da ontologia foi que aproximou duas áreas diferentes que são a Biologia e a Computação, para resolver um problema na área de Bioinformática, unindo assim estas duas áreas e fazendo com que biológicos e cientistas da computação compartilhassem uma única linguagem.

Com relação aos trabalhos futuros pretende-se adicionar ao escopo da ontologia MONET sinalização celular e motivos topológicos. A sinalização celular se refere aos processos que permitem que a célula mude seu comportamento em função de mudanças no meio exterior. Por exemplo, a *E. coli* pode passar a metabolizar lactose em um meio rico em lactose. Para tanto ela deve reconhecer esta característica e passar a produzir as enzimas necessárias para metabolizar este açúcar. Os motivos topológicos são subgrafos (subconjuntos de arestas e nodos de um dado grafo) que ocorrem com freqüência maior que a esperada. É importante localizar os motivos, pois eles podem estar relacionados à função dos genes.

Além disso, estudar a possibilidade da utilização conjunta com bancos de dados e OWL, isto é, poder consultar os dados em OWL a partir de acesso via banco de dados. Assim como, comparar a especificação OWL da ontologia MONET com a do BioPAX (*http://www.biopax.org*) para o banco de dados do KEGG (*http://www.genome.jp/kegg*) e ainda comparar o OWL da MONET com a da GO (*http://www.geneontology.org*).

Objetiva-se ainda incluir *web services* para a aquisição dos dados biológicos para a ontologia, visto que a partir deste serviço é possível extrair os dados de forma direta para a ontologia, facilitando assim o processo de aquisição, normalização e integração de dados.

REFERÊNCIAS

ABITEBOUL, S.; BUNEMAN, P.; SUCIU, D. *Data on the Web: from Relations to Semistructured Data and XML*. San Francisco: Morgan Kaufmann, 2000. 258 p.

ALMEIDA, M. B.; BAX, M. P. Uma visão geral sobre ontologias: pesquisa sobre definições, métodos de avaliação e de construção. *Ci. Inf.*, Brasília, v. 2, n. 3, p. 7–20, 2003.

ARITA, M. The metabolic world of *Escherichia coli* is not small. *Proc Natl Acad Sci USA*, v. 101, n. 6, p. 1543–1547, Feb 2004.

BAKER, P. G. et al. An ontology for bioinformatics applications. *Bioinformatics*, v. 15, n. 6, p. 510–520, 1999.

BARABáSI, A.-L.; ALBERT, R. Emergence of scaling in random networks. *Science*, v. 286, n. 5439, p. 509–512, Oct 1999. This paper introduced the concept of scale-free networks and proposed a mechanism for their emergence.

BARABáSI, A.-L. et al. Scale-free and hierarchical structures in complex networks. Nov 2002. Department of Physics, University of Notre Dame, Notre Dame, IN 46556, USA and Department of Pathology, Northwestern University, Illions 60611, USA.

BARABáSI, A.-L.; OLTVAI, Z. N. NETWORK BIOLOGY: understanding the cell's functional organization. *Nat Rev Genet*, v. 5, n. 2, p. 101–113, Feb 2004.

BATTISTELLA, E. et al. An Integrated Model for Celullar Analysis. *III Brazilian Workshop on Bioinformatics*, Brasília, Brasil, p. 1–8, 2004.

BATTISTELLA, E. et al. Using protégé to build a Molecular Network Ontology. *8th International Protégé Conference*, Madrid, Spain, p. 122, Jul 2005.

BORST, W. N. *Construction of Engineering Ontologies for Knowledge Sharing and Reuse*. 227 p. Tese (Doutorado) — Universiteit Twente, Enschede, Netherlands, Sep 1997. Disponível em: <http://doc.twente.nl/fid/1392>.

BRAZMA, A. et al. Minimum information about a microarray experiment (MIAME) - toward standards of microarray data. *Nature*, v. 29, n. 4, p. 365–371, 2001.

BUTLAND, G. et al. Interaction network containing and essential protein complexes in *Escherichia coli*. *Nature*, v. 3, n. 433, p. 531–537, Feb 2005.

DENG, M. et al. Inferring domain-domain interactions from protein-protein interactions. *Genome Res*, v. 12, n. 10, p. 1540–1548, Oct 2002.

ERDöS, P.; RéNYI, A. On the evolution of random graphs. *Math. Inst. Hung. Acad. Sci.*, v. 5, p. 17–61, 1960.

GARDNER, S. R. Building the data warehouse. *Commun. ACM*, New York, v. 41, n. 9, p. 52–60, 1998.

GóMEZ-PéREZ, A. ONTOLOGIAL ENGINEERING: A STATE OF THE ART. *Expert Update*, v. 3, n. 2, p. 33–43, 1999.

GRUBER, T. R. Towards principles for the design of ontologies used knowledge sharing. *International Journal of Human Computer Studies*, v. 43, p. 907–928, 1993.

HALLINAN, J. Gene duplication and hierarchical modularity in intracelullar interaction networks. *Bio Systems*, v. 74, n. 1-3, p. 51–62, Apr 2004. Evaluation Studies.

IDEKER, T. et al. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, v. 292, p. 929–934, 2001.

IMIELINSKI, M. et al. Investigating metabolite essentiality through genome-scale analysis of escherichia coli production capabilities. *Bioinformatics*, v. 21, n. 9, p. 2008–2016, May 2005. Evaluation Studies.

INMON, W. H. *Como construir o Data Warehouse*. Rio de Janeiro: Campus, 1997. 238 p.

JEONG, H. et al. Lethality and centrality in protein networks. *Nature*, v. 6833, n. 411, p. 41–42, May 2001.

JUNGNICKEL, D. *Graphs, Networks and Algorithms*. Berlin: Springer, 2002. 589 p. (Algorithms and Computation in Mathematics, v. 5).

KIERZEK, A. M. STOCKS: STOChastic Kinetic Simulations of biochemical systems with Gillespie algorithm. *Bioinformatics*, v. 18, n. 3, p. 470–481, Mar 2002.

LEHNINGER, A. L.; COX, M. M.; NELSON, D. L. *Lehninger Principles of Biochemistry*. New York: Worth, 2000. 1152 p.

LEMKE, N. et al. Essentiality and damage in metabolic networks. *Bioinformatics*, v. 20, n. 1, p. 115–119, 2004.

LEWIN, B. *Genes VII*. Porto Alegre: Artes Médicas, 2001. 955 p.

LODISH, H. *Molecular Cell Biology*. New York: Scientific American, 1999. 1344 p.

MURRAY, J. D. *Mathematical Biology*. Second. Berlin: Springer, 1993. 767 p. (Bioinformatics, v. 19).

PALUMBO, M. C. et al. Functional essentiality from topology features in metabolic networks: a case study in yeast. *FEBS Lett*, v. 579, n. 21, p. 4642–4646, Aug 2005.

RAVASZ, E. et al. Hierarchical Organization of Modularity in Metabolic Network. *Science*, v. 297, n. 5586, p. 1551–1555, Aug 2002.

REZENDE, S. O. *Sistemas Inteligentes: Fundamentos e Aplicações*. Baureri, SP: Manole, 2003. 525 p.

SAFFI, J.; REVERS, L. F.; HENRIQUES, J. A. O sistema dois-híbridos de *Saccharomyces cerevisiae*. Biotecnologia Ciência e Desenvolvimento, n. 21, p. 22–26, 2001.

SEGRE, D.; VITKUP, D.; CHURCH, G. M. Analysis of optimality in natural and perturbed metabolic networks. *Proc Natl Acad Sci USA*, v. 99, n. 23, p. 15112–15117, Nov 2002.

SPELLMAN, P. T. et al. Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biology*, v. 3, n. 9, 2002.

STROGATZ, S. H. Exploring complex networks. *Nature*, v. 410, p. 268–276, 2001.

STUDER, R.; BENJAMINS, R.; FENSEL, D. Knowledge Engineering: Principles and Methods. *Data and Knowledge Engineering*, v. 25, p. 161–197, 1998.

SU, X.; LARS, I. A Comparative Study of Ontology Language and Tools. In: *Proceedings of Conference on Advanced Information System Engineering*. Toronto, Canadá: [s.n.], 2002.

UETZ, P.; IDEKER, T.; SCHWIKOWSKI, B. Visualization and integration of protein-protein interactions. *Golemis*, Cold Spring Harbor Laboratory Press, p. 623–646, 2002.

WIEDERHOLD, G. Mediators in the Architecture of Future Information Systems. *IEEE Computer Society Press*, v. 25, n. 3, p. 38–49, 1992.

WITTEN, I. H.; FRANK, E. *Data Mining: Pratical Machine Learning Tools and Techniques with Java Implementations*. San Francisco: Morgan Kaufmann, 2000. 369 p.

YEGER-LOTEM, E. et al. Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction. *PNAS*, v. 101, p. 5934–5939, 2004.

YEH, I. et al. Knowledge acquistion, consistency, checking and concurrency control for Gene Ontology. *Bioinformatics*, v. 19, n. 2, p. 241–248, 2003.