



Universidade de Brasília

Instituto de Ciências Exatas
Departamento de Ciência da Computação

Visualização de Vias Metabólicas com Banco de Dados em Grafo

Gabriella de Oliveira Esteves

Monografia apresentada como requisito parcial
para conclusão do Bacharelado em Ciência da Computação

Orientador

Prof. Dr.^a Maria Emília Machado Telles Walter

Brasília
2016

Universidade de Brasília — UnB
Instituto de Ciências Exatas
Departamento de Ciência da Computação
Bacharelado em Ciência da Computação

Coordenador: Prof. Dr. Rodrigo Bonifácio de Almeida

Banca examinadora composta por:

Prof. Dr.^a Maria Emília Machado Telles Walter (Orientador) — CIC/UnB
Prof. Dr. Professor I — CIC/UnB
Prof. Dr. Professor II — CIC/UnB

CIP — Catalogação Internacional na Publicação

Esteves, Gabriella de Oliveira.

Visualização de Vias Metabólicas com Banco de Dados em Grafo /
Gabriella de Oliveira Esteves. Brasília : UnB, 2016.

83 p. : il. ; 29,5 cm.

Monografia (Graduação) — Universidade de Brasília, Brasília, 2016.

1. Bioinformática, 2. Redes Metabólicas, 3. Banco de Dados Não
Relacional, 4. Grafo, 5. OrientDB

CDU 004.4

Endereço: Universidade de Brasília
Campus Universitário Darcy Ribeiro — Asa Norte
CEP 70910-900
Brasília-DF — Brasil



Universidade de Brasília

Instituto de Ciências Exatas
Departamento de Ciência da Computação

Visualização de Vias Metabólicas com Banco de Dados em Grafo

Gabriella de Oliveira Esteves

Monografia apresentada como requisito parcial
para conclusão do Bacharelado em Ciência da Computação

Prof. Dr.^a Maria Emília Machado Telles Walter (Orientador)
CIC/UnB

Prof. Dr. Professor I Prof. Dr. Professor II
CIC/UnB CIC/UnB

Prof. Dr. Rodrigo Bonifácio de Almeida
Coordenador do Bacharelado em Ciência da Computação

Brasília, 08 de Julho de 2016

Dedicatória

Dedicatória

Agradecimentos

Agradecimento

Resumo

Resumo em português

Palavras-chave: Bioinformática, Redes Metabólicas, Banco de Dados Não Relacional, Grafo, OrientDB

Abstract

Abstract in english

Keywords: Bioinformatics, Metabolic Networks, Non-Relational Database, Graph, OrientDB

Sumário

| | |
|---|-----------|
| Introdução | 1 |
| Justificativa | 1 |
| Problema | 1 |
| Objetivo | 1 |
| Descrição dos Capítulos | 2 |
| 1 Avaliação em Interação Humano-Computador | 3 |
| 1.1 Conceitos Básicos de IHC | 3 |
| 1.2 Avaliação Através de Observação | 4 |
| 1.2.1 Método de Avaliação de Comunicabilidade | 5 |
| 2 Redes Metabólicas | 10 |
| 2.1 Conceitos Básicos de Biologia Molecular | 10 |
| 2.1.1 Ácidos Nucléicos | 10 |
| 2.1.2 Síntese de Proteína | 12 |
| 2.2 Conceitos Básicos de Metabolismo | 13 |
| 2.2.1 Metabolismo Primário | 14 |
| 2.2.2 Metabolismo Secundário | 14 |
| 2.3 Banco de Dados de Redes Metabólicas | 15 |
| 2.3.1 Banco de Dados NoSQL | 16 |
| 2.3.2 Reactome | 18 |
| 2.3.3 KEGG | 19 |
| 2.3.4 BioCyc | 21 |
| 2.3.5 2Path | 22 |
| 3 Projeto da Interface | 26 |
| 3.1 Projeto | 26 |
| 3.2 Detalhes da Implementação | 26 |
| 3.3 Desafios | 27 |
| 4 Resultados | 28 |
| 4.1 Interface | 28 |
| 4.2 Avaliação e Discussão | 28 |
| 5 Método e Resultados | 29 |

| | | |
|----------|--------------------------------------|-----------|
| 6 | Conclusão e Trabalhos Futuros | 30 |
| 6.1 | Conclusão | 30 |
| 6.2 | Trabalhos Futuros | 30 |
| 7 | Cronograma | 31 |
| | Referências | 32 |

Lista de Figuras

| | | |
|-----|--|----|
| 1.1 | Exemplo de tipos de caixa de texto que representam o mesmo campo (Nome) porém com níveis de <i>affordance</i> diferentes, sendo que o último apresenta melhor o objetivo da caixa e, portanto, ela é reconhecida mais facilmente [14]. | 4 |
| 2.1 | imagem de um nucleotídeo e das bases nitrogenadas. Mostrar backbone da pentose 1'...5'. Adaptado de : [17]. | 10 |
| 2.2 | Adaptado de : [17] | 12 |
| 2.3 | Exemplo de via metabólica retirado do site MetaCyC. Esta representa o processo de produção da bioluminescência de bactérias [9]. | 14 |
| 2.4 | Teorema CAP. Na primeira coluna, o banco de dados não suporta partição de rede, na segunda existe replicação parcial de dados entre os servidores e na terceira, replicação total [6]. | 17 |
| 2.5 | Visão geral da rede metabólica do <i>Homo sapiens</i> do Reactome, com destaque na via metabólica de remoção do Orc1 da cromatina do processo de regulação da replicação do DNA. A interface pode ser dividida nas seções [1], [2], [3] e [4]. | 18 |
| 2.6 | Visão geral da rede metabólica de referência do KEGG, com destaque na via metabólica de biossíntese do <i>backbone</i> de terpenóides. A interface pode ser dividida nas seções [1], [2] e [3]. | 20 |
| 2.7 | Via metabólica de remoção do Orc1 da cromatina. | 23 |
| 2.8 | Via metabólica de biossíntese do <i>backbone</i> de terpenóides. | 24 |
| 2.9 | Representação modificada da Figura 2.9 com maior nível de detalhes. . . . | 25 |

Lista de Tabelas

| | | |
|-----|---|----|
| 1.1 | Descrição das etiquetas de acordo com a falha de comunicação que representam [7, p. 355-356]. | 7 |
| 7.1 | Cronograma | 31 |

Introdução

No início dos anos 50, uma química britânica chamada Rosalind Frankling usou a técnica de difração de raios-X para determinação da estrutura da biomolécula do DNA e concluiu que sua forma era helicoidal. Seu trabalho foi empregado nos experimentos de dois pesquisadores, Francis Crick e James Watson, em um laboratório em Cambridge, e assim fizeram a grande descoberta que desencadearam várias linhas de pesquisas atuais: A partir do DNA, o processo de *transcrição* fornece uma fita de RNA, que por sua vez, a partir do processo de *tradução*, fornecem a proteína. Esta sequência de processos ficou conhecida como Dogma Central da biologia molecular.

A partir de então, pesquisadores já sequenciaram cadeias de DNA, RNA e proteína de vários organismos, criando uma quantidade de informação tão extensa que apenas ferramentas de Big Data podem ser usadas para análise, visualização, busca, etc, para um tratamento eficiente. Estes dados, denominados dados ômicos, são armazenados em bancos de dados específicos hoje em dia, mais vários deles colaboram entre si. O foco deste trabalho é apresentar os bancos de dados que representam redes metabólicas em modelo de grafo já existentes bem como suas ferramentas de visualização de vias metabólicas, e apresentar a ferramenta de visualização 2Path formulizada e desenvolvida neste projeto.

Justificativa

Atualmente, a quantidade de dados ômicos estudados pelos pesquisadores é extensa e complexa. Uma maneira de amenizar o esforço feito para analisá-los e compreendê-los é oferecer uma ferramenta que aproxime o usuário (pesquisador) e os dados e a maneira mais natural é representar tais dados em forma de grafo (redes metabólicas). Esta ferramenta deverá permitir que o usuário visualize e interaja com os dados dinamicamente.

Problema

Construir uma visualização interativa de redes metabólicas armazenadas em banco de dados de grafos que permita ao pesquisador explorar os aspectos biológicos do organismo estudado.

Objetivo

Construir um sistema que acesse redes metabólicas armazenadas em bancos de dados em grafo e gere uma visualização interativa

- Implementar uma busca das vias metabólicas de interesse a a partir de parâmetros informados pelo pesquisador no sistema
- Recuperar a informação desejada e exibi-la para o pesquisador de forma ergonômica

Descrição dos Capítulos

No Capítulo 1 serão descritos os conceitos básicos de biologia molecular, metabolismo primário e secundário e os bancos de dados mais utilizados para armazenar as informações referentes às redes metabólicas. No Capítulo 2 serão apresentados detalhadamente quatro ferramentas de visualização de redes metabólicas: *KEGG Pathway*, *MetaCyc*, *Reactome Browser* e *Cytoscape*.

Os Capítulos 3 e 4 já são relacionados ao sistema 2Path desenvolvido neste projeto. Enquanto o Capítulo 3 aborda o tema de interação humano-computador para a concepção de uma interface auto-explicativa e consistente, o Capítulo 4 descreve as linguagens e ambientes utilizados para a construção de tal. O Capítulo 5 especifica como são feitas as buscas por vias metabólicas no sistema pelo usuário e apresenta os resultados obtidos.

O Capítulo 6 finaliza o trabalho com a conclusão e sugestão de trabalhos futuros.

Capítulo 1

Avaliação em Interação Humano-Computador

Atualmente a tecnologia está presente mais do que nunca em grande parte das atividades das pessoas, modernizando as casas, ruas, escolas, diferentes ambientes de trabalho, integrando-se na vida pessoal de cada um muitas vezes desmedidamente, se difundindo cada vez mais rápido. Estas tecnologias são chamadas TCIs (Tecnologias da Informação e Comunicação) quando interferem direta e indiretamente nos processos de comunicação e informação dos indivíduos. Para satisfazer certos critérios de qualidade, estes sistemas devem possuir algumas características que facilitem seu uso de alguma maneira para seus usuários específicos. Ainda, vários métodos de avaliação podem ser aplicados sobre eles, cada um com um foco diferente a ser analisado. Em Ciência da Computação, a área de Interação Humano-Computador (IHC) é responsável por verificar a qualidade destes sistemas mensurando certas propriedades e avaliar o impacto dos mesmos na vida dos usuários. Neste capítulo serão apresentados os conceitos básicos de IHC e um tipo de avaliação de sistemas, a avaliação através de observação, bem como um exemplo deste tipo, o Método de Avaliação de Comunicabilidade, que será utilizado neste projeto.

1.1 Conceitos Básicos de IHC

Esta seção apresenta os componentes básicos envolvidos na interação usuário-sistema utilizados na matéria interdisciplinar Interação Humano-Computador. A **interação** engloba todo o contato realizado pelo usuário com o objetivo de exercer uma tarefa através do sistema. Segundo John Kammersgaard, 1988, existem quatro perspectivas de interação usuário-sistema [7, p. 21]:

- 1 **Sistema:** Usuário conhece linguagem específica voltada para o sistema e seu objetivo é a transmissão correta dos dados da maneira mais rápida possível;
- 2 **Parceiro de discurso:** Usuário interage com sistema através de uma inteligência artificial que personifica um interlocutor especialista naquilo que usuário procura;
- 3 **Mídia:** Usuários interagem entre si por meio do sistema, que pode oferecer diversos recursos como jogos e portal de notícias, porém é focado na qualidade da comunicação entre pessoas;

- 4 **Ferramenta:** Usuário utiliza sistema de maneira automática como instrumento de propósito geral para realizar suas tarefas.

Em um sistema interativo, a **interface** é responsável por manter o contato motor (como a *webcam* e o teclado), perceptivo (como o monitor) e conceitual (interpretação do contato físico) do usuário [7, p. 25]. O conjunto dos elementos da interface que expõem seu funcionamento de maneira implícita é chamado de *affordance* [7, p. 27]. Esta palavra não possui tradução para o português, mas entende-se que significa "reconhecimento", pois uma vez que o usuário é levado a realizar os passos corretos para cumprir seus objetivos em uma interface, mesmo sem nunca ter interagido com ela antes, considera-se que ele já conhecia algumas características que o sistema oferece e obteve respostas já esperadas. Nesse sentido, a *affordance* é geralmente uma característica muito desejável nos sistemas.

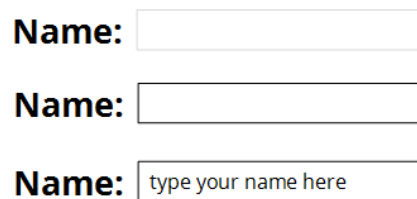


Figura 1.1: Exemplo de tipos de caixa de texto que representam o mesmo campo (Nome) porém com níveis de *affordance* diferentes, sendo que o último apresenta melhor o objetivo da caixa e, portanto, ela é reconhecida mais facilmente [14].

Em Interação Humano-Computador, a qualidade de um certo sistema está fortemente relacionada à sua interface e interação [7, p. 28]. Os critérios de qualidade são:

- 1 **Usabilidade:** Medida de complexidade no aprendizado do uso da interface para atingir objetivos de maneira eficiente (no menor tempo possível, utilizando o menor número de recursos do sistema) e eficaz (o mais automático possível);
- 2 **Experiência do usuário:** Medida de satisfação do usuário em relação ao sistema;
- 3 **Acessibilidade:** Medida da flexibilidade do sistema, ou seja, da capacidade de usuários interagirem com o mesmo;
- 4 **Comunicabilidade:** Medida da capacidade de transmissão das intenções do designer do sistema para o usuário por meio da interface.

1.2 Avaliação Através de Observação

Para manter os critérios básicos de qualidades é importante que exista algum tipo de avaliação do sistema antes de sua entrega ao(s) usuário(s). Na perspectiva de quem desenvolve o sistema, a avaliação deve verificar se o sistema recebe as entradas, processa os dados e gera o resultado na saída corretamente, enquanto que na perspectiva do usuário, a avaliação deve verificar como o comportamento da interface afeta a experiência de uso do sistema [7, p. 286], recaiando nos quatro critérios de qualidade descritos na seção 1.1. A

etapa de avaliação pode se dar durante (avaliação formativa) ou após (avaliação somativa) o processo de design. Na primeira opção, o foco é a identificação de possíveis problemas com a vantagem do baixo custo de correção. Na segunda opção, o foco é a verificação dos níveis de qualidade do protótipo de escopo definido. [7, p. 294]

Os tópicos mais avaliados em IHC são os problemas na interação e na interface do sistema, que são classificados de acordo com a frequência com que ocorrem, com sua gravidade ou com os quatro critérios de qualidade. Um exemplo de questionário associado à este tópico de avaliação se encontra na parte de consolidação dos resultados da subseção 1.2.1.

A fundamentação por trás dos métodos de avaliação é dada pela teoria da Engenharia Semiótica, focada em dois tipos de comunicações: usuário-sistema e designer-usuário através do sistema (metacomunicação). Segundo esta teoria, toda aplicação computacional é um artefato de metacomunicação por onde o designer, como interlocutor, se comunica com o usuário. Quando o objetivo é avaliar a quantidade de emissão de metacomunicação, utiliza-se o Método de Inspeção Semiótica [7, p. 330]. Quando é avaliar a quantidade de recepção de metacomunicação, utiliza-se o Método de Avaliação de Comunicabilidade [7, p. 344]. A metamensagem transmitida possui um modelo superficial que pode ser utilizado para facilitar o entendimento dos problemas do usuário, de o que pode ser feito para solucioná-los e como fazê-lo. Os locais demarcados por bicudos devem ser substituídos de acordo com o sistema desenvolvido:

Este é o meu entendimento, como *<designer do sistema>*, de quem você, *<usuário específico / grupo de usuários>*, é, do que aprendi que você quer ou precisa *<fazer / executar / visualizar / se comunicar>*, de que maneiras *<prefere / é mais natural>* fazer, e *<por quê / por quê não outras>*. O *<nome do sistema>*, portanto, é o sistema que projetei para você, e *<esta navegação específica>* é a forma como você pode ou deve utilizá-lo para alcançar uma gama de objetivos que se encaixam nesta visão.

A Semiótica também estuda a semiose: processo de significação dos signos, tudo o que é percebido pelos seres humanos e produz sensação, na mente dos intérpretes. Este é um processo que transforma os fenômenos perceptíveis através dos sentidos em experiências dos indivíduos [11]. Analogamente, em IHC a semiose é utilizada para compreender as ações tomadas, bem como as reações de resposta, pelos usuários que interagem com sistemas cobertos de signos. Nesse sentido, é possível elaborar um perfil semiótico do sistema para retratar os problemas de comunicabilidade, e, para isso, a metamensagem pode ser utilizada como base.

1.2.1 Método de Avaliação de Comunicabilidade

O MAC (Método de Avaliação de Comunicabilidade) é um método qualitativo de avaliação somativa cujo objeto são os problemas na interação e interface do sistema com foco na percepção do usuário. Ele consiste em uma análise da gravação de representantes dos usuários utilizando o sistema com base em 13 etiquetas (expressões linguísticas que caracterizam ruptura na comunicação) e é composto por cinco etapas: preparação, coleta de dados, interpretação, consolidação dos resultados e relato dos resultados.

1.2.1.1 Preparação

Nesta etapa o avaliador é responsável por definir o perfil dos usuários, selecionar os representantes dos mesmos e elaborar o ambiente e tarefa realizada por eles. O material da gravação é preparado e verificado fazendo-se um teste piloto. Um termo de consentimento da avaliação deve ser redigido para assinarem o avaliador e cada representante. [7, p. 305] Devem ser elaborados dois questionários: um pré-teste, para coletar informações de cada representante tais como conhecimento sobre o domínio do sistema e um pós-teste, para coletar informações referentes à opinião dos participantes sobre suas experiências com o sistema [7, p. 306].

1.2.1.2 Coleta de Dados

Aqui o avaliador deve receber os representantes de usuários, explicar o procedimento da tarefa realizada, entregar uma via do termo de consentimento para cada representante assinar, já com a assinatura do avaliador. Os questionários pré-teste são entregues e cabe ao avaliador decidir se é necessário ler em voz alta as perguntas, caso alguma possa conter mais de uma interpretação [7, p. 309]. Não é especificado tempo mínimo nem máximo para esta tarefa. Ao final do questionário, o avaliador posiciona os equipamentos de gravação perto de cada participante, de maneira que a imagem gravada seja a tela, o *mouse/mousepad*, o teclado e as mãos dos mesmos, e o som captado seja apenas a voz dos mesmos.

Uma técnica muito conhecida de coleta de dados na gravação é chamada *think aloud*, onde os participantes devem relatar em voz alta tudo aquilo que estão pensando em relação ao sistema, como execução, planejamento de execução, reação às respostas da interface, etc. O resultado final facilita bastante a análise do avaliador, porém o ato de falar enquanto desempenha a tarefa pode descentralizar o pensamento dos participantes [7, p. 309]. Uma vez que estes sintam-se seguros em realizar as tarefas ao mesmo tempo em que se expressam verbalmente, os avaliadores terão muito mais dados (etiquetas) para interpretar na fase seguinte.

Terminada a gravação, ocorre uma entrevista para coletar a opinião de todos os participantes em relação ao sistema, bem como tirar dúvidas sobre seus desempenhos.

1.2.1.3 Interpretação

Nesta etapa o avaliador assiste todos os vídeos contabilizando certas expressões linguísticas chamadas de etiquetas apresentadas na tabela 1.1. Deve-se levar em consideração os perfis dos usuários traçados pelo questionário pré-teste, bem como as entrevistas pós-teste, uma vez que as experiências passadas de cada um pode afetar seu comportamento no sistema. As expressões linguísticas possuem na maioria conotação ruim com respeito ao sistema, tais como indagação sobre seu comportamento e aversão às respostas. Observando os vídeos, o avaliador pode perceber as barreiras de comunicabilidade e gargalos do sistema do ponto de vista do usuário, por isso o nome de método de avaliação de *comunicabilidade*. Quando o número de participantes é muito baixo, o avaliador pode fazer a interpretação ao mesmo tempo em que coleta dados, sem precisar de gravação.

Tabela 1.1: Descrição das etiquetas de acordo com a falha de comunicação que representam [7, p. 355-356].

| Falhas de comunicação completas: efeito obtido é inconsistente com a intenção comunicativa do usuário | | |
|--|---|-----------------------|
| aspecto semiótico | característica específica | etiqueta |
| O usuário termina uma semiose malsucedida, mas não inicia outra para obter o resultado esperado | porque, mesmo percebendo que não obteve o resultado esperado, não possui mais recursos, capacidade ou vontade de continuar tentando | Desisto |
| | porque não percebe que não obteve o resultado esperado | Para mim está bom... |
| Falhas de comunicação parciais: o efeito obtido é somente parte do efeito pretendido de acordo com a intenção do usuário | | |
| aspecto semiótico | característica específica | etiqueta |
| O usuário abandona uma semiose antes de obter o resultado esperado, e inicia outra com o mesmo propósito | porque, embora entenda a solução de IHC proposta, prefere seguir por outro caminho no momento | Não, obrigado |
| | porque não entende a solução de IHC proposta | Vai de outro jeito |
| Falhas de comunicação temporárias: o efeito parcial do processo de interpretação (semiose) e de comunicação (interação) do usuário é inconsistente e incoerente com sua intenção de comunicação | | |
| aspecto semiótico | característica específica | etiqueta |
| O usuário interrompe temporariamente sua semiose | porque não encontra uma expressão apropriada para sua intenção de comunicação | Cadê? |
| | porque não percebe ou não entende a expressão do sistema (preposto do designer) | Ué, o que houve? |
| | porque não consegue formular sua próxima intenção de comunicação | E agora? |
| O usuário percebe que seu ato comunicativo não foi bem-sucedido | porque percebeu que havia "falado" algo no contexto errado | Onde estou? |
| | porque percebeu que havia "falado" algo errado | Epa! |
| | porque não obteve o resultado esperado depois de conversar com o sistema (preposto do designer) por algum tempo, alternando vários turnos de fala com ele | Assim não dá. |
| O usuário procura compreender o ato comunicativo do sistema (preposto do designer) | através da metacomunicação implícita | O que é isto? |
| | através da metacomunicação explícita | Socorro! |
| | testando várias hipóteses sobre o significado do que o sistema comunicou | Por que não funciona? |

Após a análise dos dados, o avaliador deve redefinir um protótipo da próxima versão do sistema já com a solução para os problemas mais simples e listar todos os obstáculos na interface que podem ser corrigidos.

1.2.1.4 Consolidação dos Resultados

Este é o momento em que o avaliador busca diferenciar as características do grupo das características individuais observando a recorrência das etiquetas em certos pontos da navegação no sistema. Pode-se atribuir significado às etiquetas de acordo com frequência de uso em um certo contexto, com sequência de uso e com o nível de problemas, por exemplo. Além disso, como o foco desta avaliação é verificar os problemas relacionados à interação e interface, nesta etapa devem ser respondidas as seguintes questões [7, p. 293]:

- O usuário consegue operar o sistema?
- Ele atinge seu objetivo? Com quanta eficiência? Em quanto tempo? Após cometer quantos erros?
- Que parte da interface e da interação o deixa insatisfeito?
- Que parte da interface o desmotiva a explorar novas funcionalidades?
- Ele entende o que significa e para que serve cada elemento da interface?
- Ele vai entender o que deve fazer em seguida?
- Que problemas de IHC dificultam ou impedem o usuário de alcançar seus objetivos?
- Onde esses problemas se manifestam? Com que frequência tendem a ocorrer? Qual é a gravidade desses problemas?
- Quais barreiras o usuário encontra para atingir seus objetivos?
- Ele tem acesso a todas as informações oferecidas pelo sistema?

Respondidas as perguntas, o avaliador pode criar então um perfil semiótico para o sistema com base na metamensagem:

Este é o meu entendimento, como designer, de **quem você, usuário, é** (1), do que aprendi que você **quer ou precisa fazer** (2), de **que maneiras prefere fazer** (3), e **por quê** (4). Este, portanto, é o sistema que projetei para você, e esta é **a forma como você pode ou deve utilizá-lo** (5) para alcançar uma gama de objetivos que se encaixam nesta visão. [7, p. 356]

- 1 Qual é o perfil dos usuários?
- 2 Quais são seus desejos e o que a metacomunicação realiza para satisfazê-los?
- 3 Quais as maneiras de realizar seus desejos e de que maneira eles preferem fazer?
- 4 O que os leva a ter esta preferência?
- 5 Quão bem o conteúdo da metacomunicação é transmitidos aos usuários?

1.2.1.5 Relato dos Resultados

Ao final, o avaliador terá apresentar aos *stakeholders* responsáveis pelo desenvolvimento do sistema os seguintes resultados:

- Os objetos da avaliação;
- Breve descrição do método de avaliação;
- Número e perfil dos avaliadores e representantes de usuários;
- Descrição das tarefas executadas pelos participantes;
- Resultado das etiquetagens, em geral contabilizando as etiquetas por usuário e tarefa;
- Problemas de comunicabilidade encontrados;
- Sugestões de melhoria;
- Perfil semiótico do sistema.

Capítulo 2

Redes Metabólicas

Neste capítulo serão descritos os conceitos básicos da biologia molecular, metabolismo e bancos de dados específicos para redes metabólicas. A primeira seção detalha a origem das principais estruturas que promovem o metabolismo tais como DNA e enzima, enquanto que a segunda seção descreve de fato como ocorre o processo. Por fim, a última seção apresenta os principais bancos de dados em grafo voltados para redes metabólicas: *KEGG*, *MetaCyc* e *Reactome*.

2.1 Conceitos Básicos de Biologia Molecular

2.1.1 Ácidos Nucléicos

Os ácidos nucleicos são biomoléculas responsáveis pelo armazenamento, transmissão e tradução das informações genéticas dos seres vivos. Isto é possível devido ao processo de síntese de proteínas que permite, assim, a base da herança biológica. Os ácidos nucleicos são polímeros, macromoléculas formadas por estruturas menores chamadas monômeros, que nesse caso são nucleotídeos. Nucleotídeos são compostos de três elementos: um radical fosfato (HPO_4), uma pentose, ou seja, um monossacarídeo formado por cinco átomos de carbono, e uma base nitrogenada. Existem cinco tipos de bases nitrogenadas que podem compor um nucleotídeo: Adenina(A), Timina(T), Citosina(C), Guanina(G) e Uracila(U).

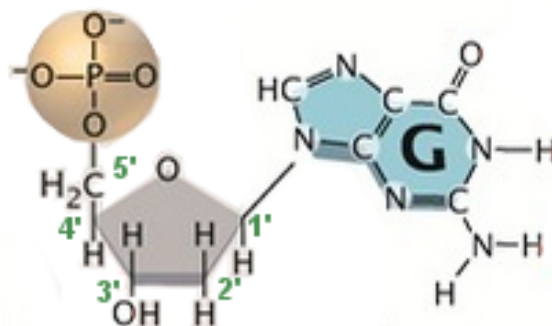


Figura 2.1: imagem de um nucleotídeo e das bases nitrogenadas. Mostrar backbone da pentose 1'...5'. Adaptado de : [17].

Na figura 2.1, observa-se que no *backbone* do nucleotídeo existe uma numeração de 1' à 5', que representam os carbonos presentes na pentose. Para a criação de uma fita de ácido nucleico, no processo de polimerização formar-se uma ligação fosfodiéster entre o carbono da posição 5' do *backbone* de um nucleotídeo e o carbono de posição 3' do *backbone* de outro [18]. Por definição o sentido da leitura de uma fita de ácido nucleico é $5' \rightarrow 3'$, o que é deve ser levado em consideração ao se fazer interpretação de dados do material genético.

Dois tipos de ácidos nucleicos são encontrados nos seres vivos: ácido desoxirribonucleico (DNA ou ADN) e ácido ribonucleico (RNA ou ARN). Eles diferenciam-se tanto na estrutura do *backbone* e nas bases nitrogenadas, quanto em suas funções. Os DNAs são as biomoléculas que armazenam as informações referentes ao funcionamento de todas as células dos seres vivos de maneira específica: sequências de pares de bases nitrogenadas. Nesse sentido, além de haver a ligação fosfodiéster entre os nucleotídeos, cada um também se liga a partir de suas bases nitrogenadas, formando assim um eixo helicoidal tridimensional chamada de dupla hélice [18]. Esta estrutura foi descoberta em 1953, pelo biólogo James Watson e pelo físico Francis Crick [17], porém os ácidos nucleicos já eram estudado desde 1869 na Suíça pelo químico-fisiológico Friedrich Miescher.

Em relação à estrutura dos monômeros do DNA, o *backbone* dos nucleotídeos é uma desoxirribose, indicada na figura 2.2. Para a formação da dupla hélice, os pares são feitos com uma base nitrogenada do grupo de purinas, composto orgânico que possui um anel duplo de carbono, e outra base do grupo de pirimidinas, composto orgânico que possui um anel simples de carbono. No caso do DNA, somente quatro das cinco bases são empregadas: as purinas Adenina(A) e Guanina(G), que se ligam com as pirimidinas Timina(T) e Citosina(C) respectivamente. Desta forma, A e T são bases complementares, assim como G e C. Uma fita de DNA pode conter centenas de milhões de nucleotídeos.

A representação do DNA, seja nos livros ou computacionalmente, é dada por um par em paralelo de strings de letras A, T, G e C. Como explicado no início dessa seção, o sentido padrão da leitura de uma fita é de $5' \rightarrow 3'$, mas no caso do DNA, as hélices são dispostas de maneira antiparalela, ou seja, uma é lida de $5' \rightarrow 3'$ e a outra, de $3' \rightarrow 5'$. Observa-se que a partir de uma hélice, pode-se inferir a sequência de sua hélice complementar. Seja, por exemplo, uma hélice H1 igual a AGTAAGC; então H2 em seu sentido oposto é H2' igual a TCATTCG, e no sentido regular, igual a GCTTACT. A figura 2.2 apresenta a estrutura do DNA como explicada nesta seção.

Os RNAs são biomoléculas semelhantes ao DNA, porém contam com três diferenças básicas. A primeira é a estrutura do *backbone* dos nucleotídeos, que é composta por uma ribose ao invés de um desoxirribose. A segunda diferença é em relação às bases nitrogenadas, onde a pirimidina Uracila(U) substitui a Timina(T). Por fim, o RNA é formado por apenas uma hélice tridimensional.

Existem três tipos de RNAs presentes no citoplasma - espaço entre a membrana plasmática e o núcleo da célula. Cada um possui funções específicas que serão detalhadas na seção 2.1.2. Em suma, O RNA mensageiro (mRNA) é responsável pela transferência de informação do DNA para o RNA ribossômico (rRNA), que por sua vez irá desanexar a proteína do RNA transportador (tRNA) combinando-o com o rRNA, executando assim, a síntese de proteína.

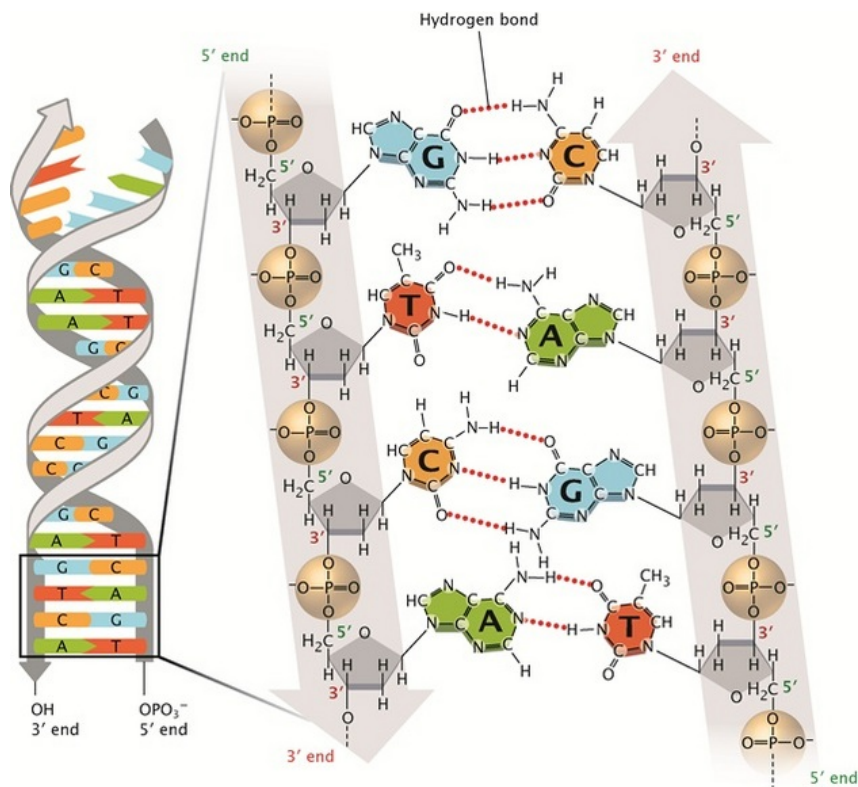


Figura 2.2: Adaptado de : [17]

2.1.2 Síntese de Proteína

As proteínas são biomoléculas com diversas responsabilidades no corpo dos seres vivos. Se fizerem parte do no grupo de proteínas fibrosas, como o colágeno, irão compor a estrutura do corpo e para isso precisam ser resistentes e insolúveis em água. Caso estejam no grupo de proteínas globulares, como a hemoglobina, realizarão processos dinâmico pelo corpo tais como transportações e cataliações [3]. Cada tarefa é realizada por um proteína com uma estrutura específica e otimizada pra tal.

Assim como os ácidos nucleicos, as proteínas são polímeros, macromoléculas cujos monômeros são aminoácidos. Aminoácidos são moléculas que possuem cinco componentes: amina (NH_2), carbono (C), hidrogênio (H), ácido carboxílico (COOH) e uma cadeia lateral que funciona como identificador de cada um dos 20 tipos de aminoácidos presentes nos seres vivos. A maneira como eles são criados será explicada com mais detalhes mais à frente, pois envolve um processo complexo de síntese de proteína executado pelo ribossomo. A ligação, ou polimerização, de dois aminoácidos é feita unindo a amida de um com o ácido carboxílico do outro, liberando uma molécula de água (H_2O) e formando uma cadeia chamada de dipeptídeo. Como houve liberação de água na ligação, o dipeptídeo não é formado por aminoácidos, mas sim resíduos dos mesmos. Nesse sentido, cadeias peptídicas de 100 à 5000 diferentes resíduos aminoácidos, ou cadeia polipeptídicas, constituem a proteína.

Existem quatro estruturas para caracterização de uma proteína [18]. A mais simples é chamada de estrutura primária e é composta por uma sequência linear de resíduos aminoácidos. A estrutura secundária é tridimensional e estabiliza-se por meio de ligações

de hidrogênio na cadeia principal, chamada de *backbone*. Dependendo da disposição dos resíduos de aminoácidos, esta cadeia pode se dar forma de hélice (α -Helix) ou em forma de folha (β -Helix). A estrutura terciária é dada pela união de várias estruturas secundárias e, por fim, a estrutura quaternária é composta de múltiplas estruturas terciárias [5].

A **transcrição** é o processo de produção de mRNA a partir do DNA e ele ocorre da seguinte forma: O início de cada gene possui um identificador em uma das fitas para indicar o local da codificação e, a partir dali, uma cópia inversa (A, T, C, G são traduzidos para U, A, G, C respectivamente) do mesmo é feita sob forma de molécula de mRNA que, por consequência, obterá a mesma sequência que a cadeia codificadora (a qual não possui o identificador), porém trocando os U's por Ts.

O mRNA deixa, então, o núcleo celular e inicia a **tradução** no citoplasma. O processo ocorre no interior de uma organela celular chamada de ribossomo, constituído de proteínas e rRNA e cuja função é construir a molécula de proteína a partir de duas entradas, o mRNA e tRNA. A estrutura do tRNA é tal que de um lado se encaixa exatamente um códon¹ e no oposto, seu aminoácido correspondente. O processo de tradução se dá da seguinte forma: a medida em que o mRNA passa pelo interior do ribossomo, este atrai quaisquer tRNAs das proximidades cujos códons sejam correspondentes ao da subsequência corrente do mRNA. No momento em que o códon do tRNA se conecta com um dos códons do mRNA, a molécula de proteína em desenvolvimento é liberada e, com o auxílio da catálise de uma enzima, agregada no aminoácido que estava fixado naquele tRNA. Esta fase é finalmente completa quando o mRNA apresenta um códon de parada, pois nenhum tRNA possui correspondência para tal [18]. Uma proteína simples é, então, formada.

2.2 Conceitos Básicos de Metabolismo

As reações bioquímicas são alterações químicas que fornecem um ou mais produtos a partir de um ou mais substratos. O conjunto de todas as reações bioquímicas que ocorrem dentro de um organismo vivo é chamado de metabolismo, e ele podem ser dividido em dois subconjuntos: catabolismo, quando ocorre a quebra de moléculas complexas produzindo energia, e anabolismo, quando ocorre a síntese de moléculas complexas, o que requer energia. Geralmente a energia liberada pelas reações catabólicas é usada para impulsionar as reações anabólicas[8]. Uma via metabólica é uma sequência de reações bioquímicas, cujo produto e substrato são denominados metabólitos, que podem ser catalisados por enzimas, as quais muitas vezes necessitam de compostos químicos chamados de co-fatores para realizarem suas atividades na célula. O conjunto de vias metabólicas de um organismo é chamado de rede metabólica.

Enzimas são proteínas responsáveis por auxiliar a realização de biossíntese (construção) e biodegradação de moléculas no metabolismo com o propósito de catalisar (acelerar) reações bioquímicas. As enzimas possuem um local pré-determinado em formato côncavo chamado de sítio ativo, que comporta um ou mais substratos. Se a enzima comporta apenas um substrato, a estrutura que se forma com o preenchimento do sítio ativo é um complexo enzima-substrato, porém se ela comporta mais de um substrato, a estrutura é chamada de complexo ternário intermediário [15]. Quando a atividade catalítica não

¹Sequência de três nucleotídeos.

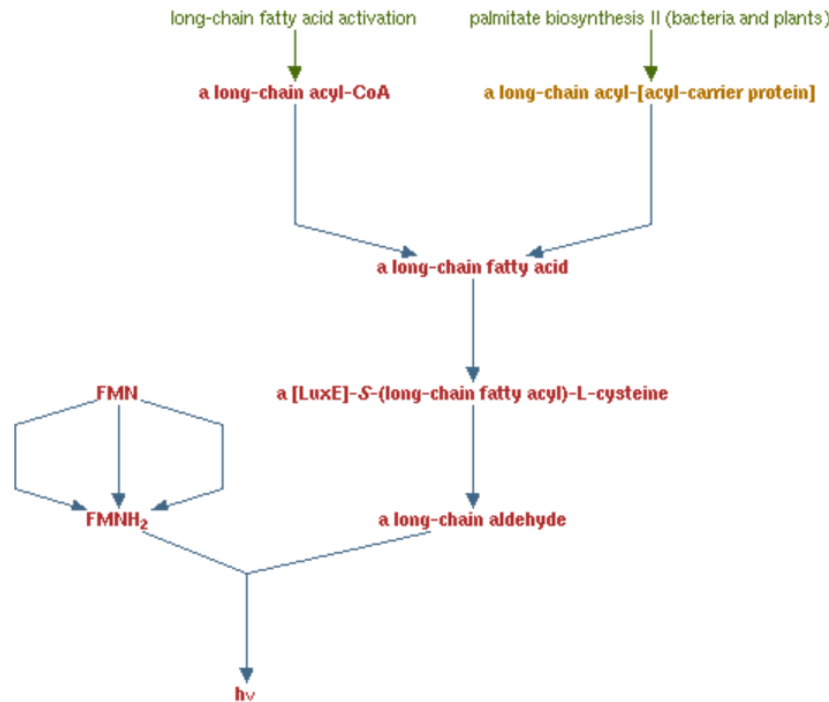


Figura 2.3: Exemplo de via metabólica retirado do site MetaCyC. Esta representa o processo de produção da bioluminescência de bactérias [9].

pode ser realizada apenas pela enzima, co-fatores auxiliam o processo. Eles podem ser coenzimas, associadas momentaneamente às enzimas, ou grupos prostéticos, associados firmemente a elas [15]. Quando duas enzimas possuem a mesma atividade enzimática porém estruturas físicas diferentes são chamadas isoenzimas[15].

2.2.1 Metabolismo Primário

O metabolismo, além de ser subdividido em catabolismo e anabolismo, também pode ser classificado em relação à função que exerce no organismo. Se a função realizada é fundamental no organismo, como crescimento, desenvolvimento e reprodução, ele é denominado metabolismo primário [20]. Mitose e meiose são exemplos de metabolismos primários.

2.2.2 Metabolismo Secundário

No caso do metabolismo não realizar função essencial no organismo, ele é classificado como metabolismo secundário. Estes são caracterizados pela vasta diversidade química e, desta forma, são responsáveis pela sobrevivência do organismo em diferentes meios ambientes de acordo com os fatores bióticos (elementos causados pela interação entre organismos como, por exemplo, cadeia alimentar) e abióticos (elementos naturais independente de organismos como, por exemplo, luz e temperatura) [20]. Enquanto 20% dos metabólitos secundários são encontrados em bactérias, fungos e organismos sésseis² marinhos, os ou-

²Que vivem fixos, sem capacidade de locomoção.

tros 80% encontram-se em plantas vasculares [20] e estes podem ser subdivididos em três classes: terpenóides, alcalóides e fenólicos [13].

- Os **terpenóides** constituem no grupo mais abundante de produtos naturais, apresentando uma grande variedade estrutural e funcional, principalmente no Reino *Plantae*. No metabolismo secundário, possuem funções como produção de óleos, esteroides, cera, resinas e borracha natural, produção de compostos usados para defesa contra herbívoros ou aromas usados para atrair polinizadores [20].
- Os **alcalóides** são majoritariamente tóxicos à outros organismos diferentes daquele que os produz. Nesse sentido, eles possuem nas plantas função de defesa contra herbívoros e podem ser encontrados principalmente nos locais mais propícios à ataques como, por exemplo, nas sementes, flores e tecidos periféricos em crescimento. Para o consumo dos seres humanos, são usados na fabricação de estimulantes, como cafeína e nicotina, e drogas, como morfina[20]. Por apresentarem alta diversidade estrutural, é difícil classificá-los; a tentativa mais recente se baseia na semelhanças entre os esqueletos carbônicos [13].
- Os **fenólicos** são caracterizados por suas propriedades anti-oxidantes, anti-inflamatória e anti-cancerígena e muitos deles são bactericidas, antisépticos e vermífugo. Eles estão presentes em praticamente todas as plantas e são utilizados na química, biologia, agricultura e medicina [13].

2.3 Banco de Dados de Redes Metabólicas

Desde o descobrimento da estrutura do DNA por Crick e Watson, o número de sequências de proteínas descobertas cresceu, aumentando também a necessidade de criar-se um banco de dados para indexá-las. A físico-química norte-americana Margaret Dayhoff, com colaboração de alguns membros do *National Biomedical Research Foundation* em Washington, foi a primeira a construir um banco de dados com este propósito em um tipo de atlas de proteínas na década de 60. Somente em 1984 esta coleção foi intitulada de *Protein Information Resource* [16]. Os dados eram organizados de acordo com o grau de similaridade das sequências, onde o agrupamento das mesmas era dado em forma de árvore filogenética representando famílias e superfamílias de proteínas. Caso a semelhança seja alta, é provável que tenham as mesmas funções bioquímicas e estrutura tridimensional. A partir da árvore gerada, foi possível calcular as mutações que ocorreram nos aminoácidos durante a evolução genética e, então, produzir uma tabela utilizada até hoje, chamada PAM (*Percent Acept Mutation*), que apresenta tais dados³. Outro banco de dados de grande porte e bastante utilizado nos dias de hoje é o GenBank, estabelecido em 1982 por Walter Goad e demais colaboradores com o objetivo de catalogar sequências genéticas e coleções de anotações de todos os DNAs públicos, agora, com o patrocínio do *National Center for Biotechnology Information*. Os dois bancos são públicos e continuam crescendo exponencialmente [16].

Nos dias de hoje, a quantidade de dados é tão grande e que os biólogos enfrentam dificuldades em tarefas como análise, busca, armazenamento, visualização e atualização de dados. Nesse sentido, eles utilizam agora ferramentas de Big Data em suas pesquisas.

³1 PAM é uma medida de tempo para representar 1 mutação para cada 100 aminoácidos.

Existem grandes áreas da biologia voltadas para estudo deste dados (chamados de dados ômicos), tais como, por exemplo, genoma⁴, transcritoma⁵, proteoma⁶, metaboloma⁷ e interactoma⁸. Nesta seção serão apresentados três bancos de dados utilizados no estudo do metaboloma cuja estrutura de dados que representam as redes metabólicas são grafos.

2.3.1 Banco de Dados NoSQL

Para comportar tamanha quantidade de dados e ao mesmo tempo providenciar alta performance, atualmente é possível utilizar um banco de dados NoSQL como alternativa ao tradicional modelo relacional. NoSQL pode significar Não-Relacional ou *Not Only SQL* [10], para ressaltar as vantagens sobre o modelo tradicional, que são rápida leitura e escrita dos dados, suporte à armazenamento em larga escala, fácil escalabilidade, fácil distribuição (replicação ou fragmentação) de dados entre vários servidores, e baixo custo. [12]. Entretanto, uma grande desvantagem é sua falta de suporte à linguagem SQL, o que obriga o usuário a aprender a recuperar seus dados de outra maneira⁹.

Em 2000, o Professor Eric Brewer da Universidade da Califórnia em Berkeley estabeleceu o Teorema CAP (*Consistency, Availability and Partition tolerance*), que afirma que quando os dados estão particionados¹⁰ em rede, os BDs NoSQL que são distribuídos devem tolerar tal falha e oferecer, assim, apenas duas escolhas para o usuário: consistência ou disponibilidade de dados, nunca os dois ao mesmo tempo. [12] [10]. Caso o banco decida fornecer mais consistência do que disponibilidade, o sistema tentará sempre retornar a *query* com dados atualizados para o usuário, mesmo que a operação demore. A Figura 2.4 ilustra o teorema também conhecido como Teorema de Brewer. Caso decida fornecer mais disponibilidade do que consistência, o sistema tentará retornar a *query* o mais rápido possível, mesmo se o dados estiverem desatualizados. Esta *trade-off* inerente na maioria dos BDs NoSQL pode, na verdade, ser decidida pelo próprio usuário ao configurar seu banco de dados.

2.3.1.1 Banco de Dados em Grafo

Grafos são as estruturas mais úteis para se representar interações entre objetos, portanto são de grande interesse no campo biológico, na modelagem de dados ômicos [19]. Bancos de dados em grafo podem ser utilizados, por exemplo, quando a finalidade é buscar via *query* a procedência dos objetos¹¹, pois realizar uma travessia em grafo é bem mais rápido computacionalmente do que realizar múltiplos *joins* num modelo relacional.

Ao se trabalhar com banco de dados em grafos, é importante ter conhecimento das principais terminologias a seguir:

- Adjacência: Nós que dividem uma aresta incidente ou arestas que dividem um nó incidente;

⁴Material genético de um organismo.

⁵Conjunto dos transcritos mRNA, tRNA, rRNA e microRNA.

⁶Conjunto de proteínas e suas variantes em um organismo.

⁷Conjunto de metabólitos de um organismo.

⁸Conjunto de interações moleculares em um organismo.

⁹No banco de dados Neo4j, por exemplo, usa-se a linguagem CYPHER.

¹⁰Falha em algum dispositivo de rede que causa a separação entre dois conjuntos de computadores.

¹¹A procedência é a linhagem do dado, ou seja, os detalhes de sua criação desde a origem conhecida.

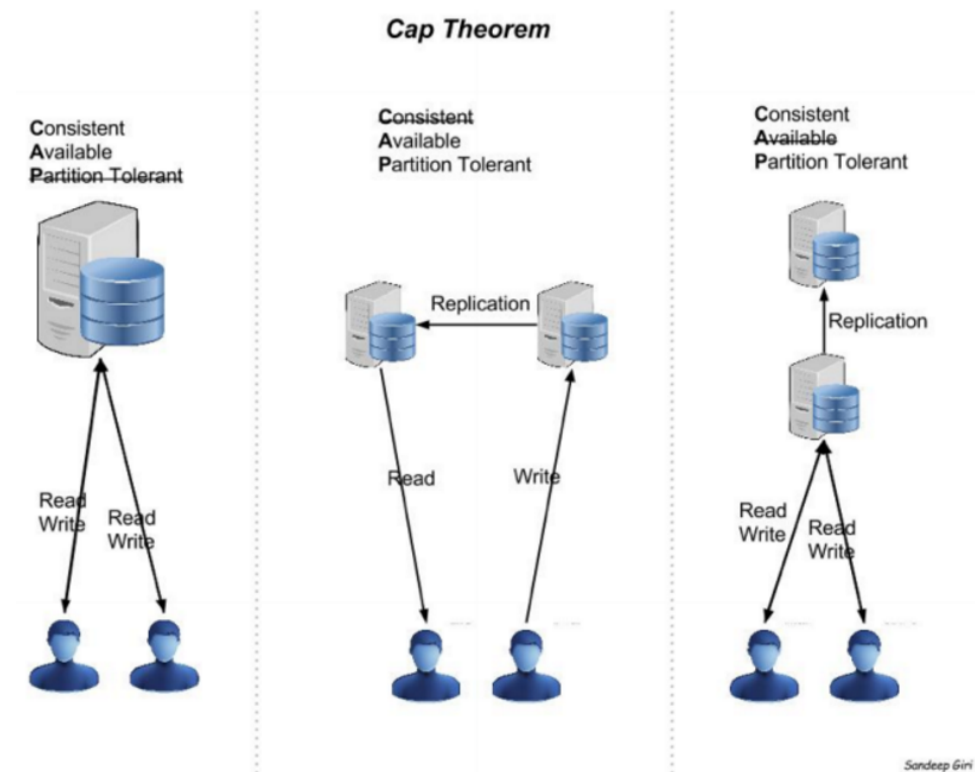


Figura 2.4: Teorema CAP. Na primeira coluna, o banco de dados não suporta partição de rede, na segunda existe replicação parcial de dados entre os servidores e na terceira, replicação total [6].

- Aresta: Uma registro que representa relação entre um par de objetos. Pode possuir direção, quando a relação possui origem e destino, ou não;
- Caminho: Coleção de alternados nós e arestas;
- *Constraint*: Restrição definida pelo usuário que o banco deve impor aos dados;
- DAG: Um grafo direcionado e acíclico;
- Incidência: Aresta adjacente associada à um nó, ou nó associado à uma aresta;
- *Label*: Rótulo de um nó que o agrupa em algum subconjunto;
- Loop: Uma aresta que conecta um nó à ele mesmo;
- Nó: Um registro que representa um objeto e possui um número indefinido de propriedades, labels e arestas incidentes;
- Propriedade: Atributo armazenado em um nó ou aresta;
- Registro: Unidade de armazenamento;
- Vizinho: Nó conectado por uma aresta em comum.

2.3.2 Reactome

Reactome¹² é um banco de dados de reações de mudança de estado, ou seja, além de reações bioquímicas, ele também abrange reações de ativação, de degradação e de ligação, por exemplo [4]. Ele faz uma ligação sistemática entre as proteínas de um certo organismo e as funções moleculares do mesmo, fornecendo uma base de funções que pode ser utilizada para pesquisas sobre expressão de genes ou mutações somáticas. O Reactome disponibiliza o *Pathway Browser*, uma rede geral para cada organismo, que representa os vários seus sistemas, como reprodução e metabolismo, por exemplo. Algumas sub-redes estão conectadas (por exemplo, replicação de DNA e ciclo de célula), outra não (por exemplo, contração muscular e reprodução). Nesta rede, cada nó representa uma via cujo número de entidade se reflete no raio do nó, e cada aresta representa a relação entre estas vias. O site ainda possui uma ferramenta de análise de dados baseada nas correspondências entre as reações nas redes dos organismos comparados.

Para acessar sua ferramenta de visualização, basta navegar a partir da *home page*:

<http://www.reactome.org/> ⇒
Browse Pathways

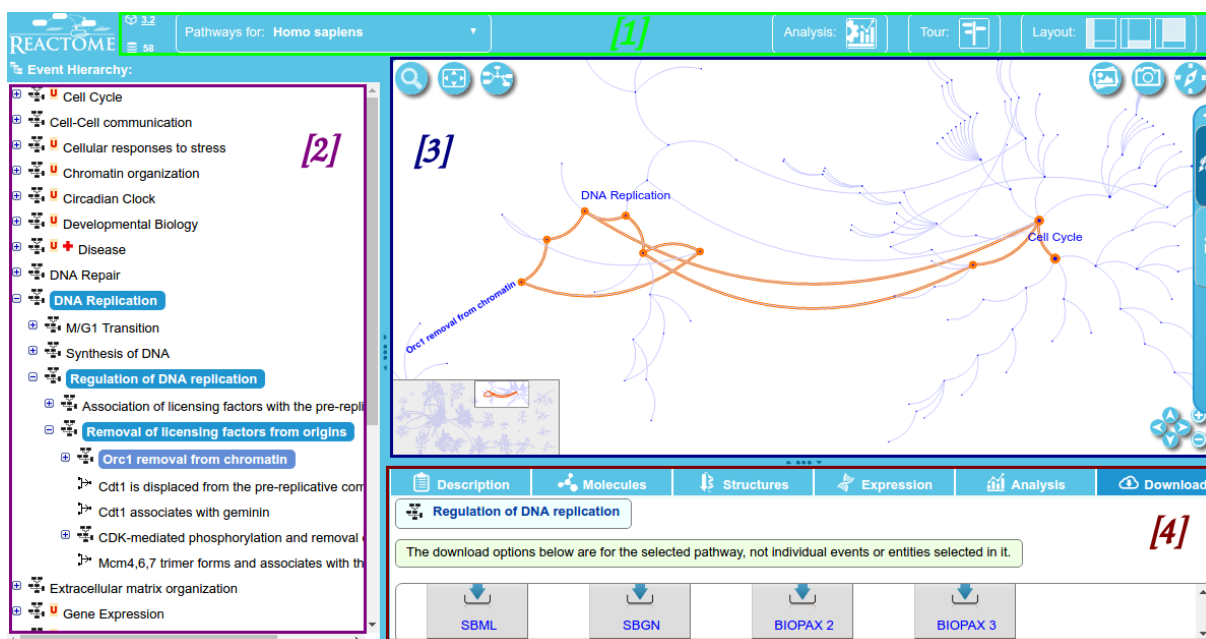


Figura 2.5: Visão geral da rede metabólica do *Homo sapiens* do Reactome, com destaque na via metabólica de remoção do Orc1 da cromatina do processo de regulação da replicação do DNA. A interface pode ser dividida nas seções [1], [2], [3] e [4].

Segue abaixo a descrição de cada elemento da interface da ferramenta de visualização do Reactome, subdivididos na Figura 2.5.

- [1] Menu de navegação direta e indiretamente relacionado às vias metabólicas. À esquerda tem-se a versão do *Pathway Browser* e de seu banco de dados e um menu

¹²Disponível pela web através do site <http://www.reactome.org>.

dropdown para se selecionar o organismo mostrado na interface. À direita tem-se um *link* (*Analysis*) para a uma página modal onde é possível analisar os dados do usuário e fazer comparação entre espécies, um *link* (*Tour*) para um vídeo explicando como *Pathway Browser* funciona e um conjunto de *layouts* que alteram a visualização geral da página;

- [2] Lista de redes metabólicas ordenadas alfabeticamente;
- [3] Mapa gerada em forma de grafo interativo. Quando uma rede metabólica da lista [2] é selecionada, ela se destaca no mapa. No topo, à esquerda é possível buscar por palavras chaves no mapa e visualizá-lo em tela cheia e à direita é possível visualizar imagens detalhadas das redes metabólicas, capturar uma imagem do mapa gerado e ler sobre a notação utilizada na página. Abaixo, à esquerda tem-se um mini-mapa por onde o usuário pode se localizar no mapa geral e à direita tem-se botões com o mesmo propósito e mais dois: (+) *Zoom in* e (-) *Zoom out*. Por fim, à direita do mapa existe um painel que se abre da direita para a esquerda que permite que o usuário edite as cores do mapa e obtenha informação sobre o mesmo.
- [4] Menu de operação e informações sobre as redes metabólicas selecionadas em [2]. O *Description* contém seus detalhes tais como compartimento celular e referências externas. O *Molecules* contém uma lista de compostos químicos e proteínas associadas às vias. O *Structures* oferece uma lista de imagens que representam a via e seus componentes. O *Expression* apresenta detalhes sobre o local no organismo onde ocorrem as reações das vias selecionadas. O *Analysis* apresenta o resultado das análise feita em [1]. Finalmente o botão *Downloads* oferece arquivos das vias para baixar em diversos formatos tais como SBML e PDF.

Caso o usuário queira visualizar a(s) via(s) metabólica detalhadamente, basta clicar nos nós do mapa, cujo tamanho está diretamente relacionado ao número de entidades que compõem a via. A imagem do grafo desaparece e então um conjunto dos elementos que o compõe aparece em seu lugar. A Figura 2.7 apresenta a via da remoção do Orc1 da cromatina do processo de regulação da replicação do DNA.

2.3.3 KEGG

O KEGG¹³ (*Kyoto Encyclopedia of Genes and Genomes*) é uma base de informações sobre sistemas biológicos em nível molecular, sobretudo sobre conjuntos de dados em larga escala gerados por sequenciamento de genoma [2]. As informações sobre os sistemas podem ser dadas em forma de módulos, unidades funcionais com identificação otimizada para análise dos dados, em forma de *brite*, coleção de arquivos estruturados hierarquicamente sobre as funções das entidades biológicos, ou em forma de vias, mapa de interações moleculares e reações químicas. Dado que o metabolismo é um conjunto de reações e transformações químicas, a maneira natural de representá-lo é por meio de uma rede de interações, ou seja, em forma de vias. O KEGG oferece uma ferramenta de busca de vias metabólicas sobre várias redes metabólicas, dos vários organismos que constituem o banco de dados.

¹³Disponível pela *web* através do site <http://www.kegg.jp/>.

Para acessar sua ferramenta de visualização, basta navegar a partir da *home page*:

<http://www.kegg.jp/> ⇒

Data-oriented entry points ⇒ KEGG PATHWAY ⇒

Pathway Maps ⇒ 1. Metabolism ⇒

1.0 Global and overview maps ⇒ Metabolic pathways

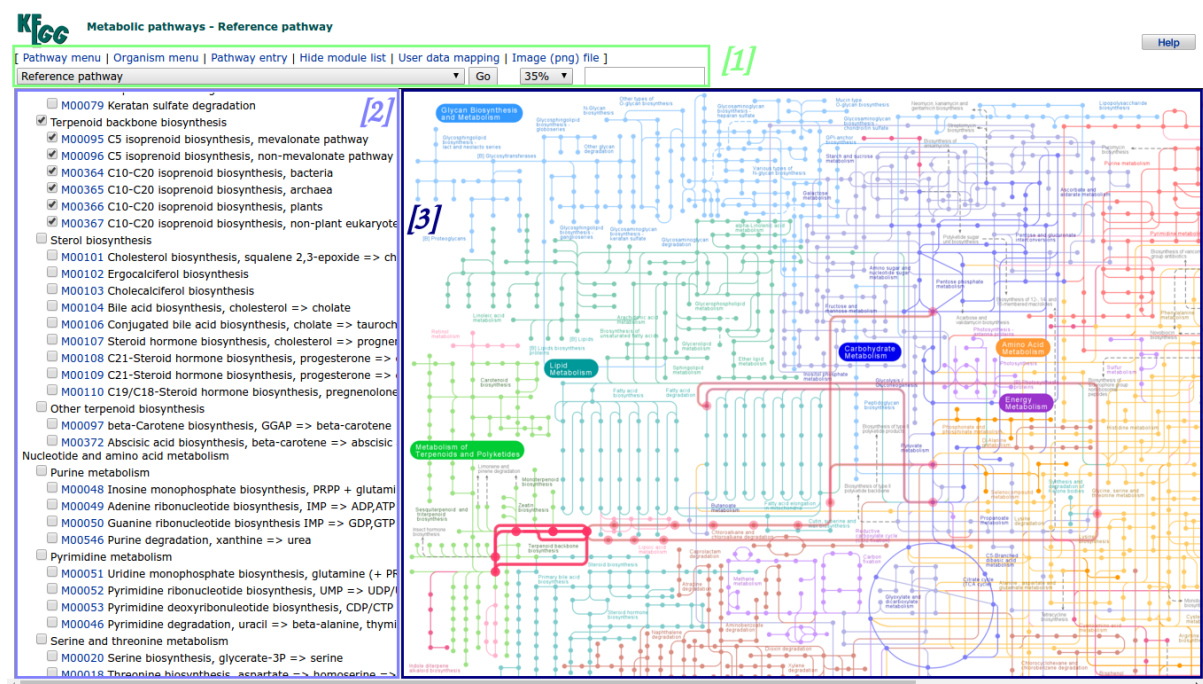


Figura 2.6: Visão geral da rede metabólica de referência do KEGG, com destaque na via metabólica de biossíntese do *backbone* de terpenóides. A interface pode ser dividida nas seções [1], [2] e [3].

Segue abaixo a descrição de cada elemento da interface da ferramenta de visualização do KEGG, subdivididos na Figura 2.6.

[1] Menu de navegação direta e indiretamente relacionado às vias metabólicas. O *Pathway menu* redireciona para uma página que contém uma lista *links* para todos os diagramas mostrados no mapa geral. O *Organism menu* redireciona para à uma página com uma lista de todos os organismos conhecidos pelo KEGG (346 eucariotas e 4159 procariotas¹⁴). O *Pathway entry* redireciona para a descrição da via metabólica selecionada. O *Hide module list*, quando clicado, oculta a área [2] da página. O *User data mapping* abre uma nova janela que solicita um objeto seguido de uma cor escolhida pelo o usuário para dar sua própria coloração às vias. Por fim, o *Image (png) file* oferece a imagem do mapa para *download*. Abaixo é possível escolher o organismo apresentado no mapa, sua resolução e buscar por palavras chaves;

[2] Lista de redes metabólicas organizadas de acordo com suas funções;

¹⁴Visitado em 2016-10-06

- [3] Mapa gerada em forma de grafo interativo. Quando uma rede metabólica da lista [2] é selecionada, ela se destaca no mapa.

Caso o usuário queira visualizar a(s) via(s) metabólica detalhadamente, o KEGG oferece uma lista de diagramas¹⁵ desenhados manualmente em formato de grafo interativo cujos nós, quando clicados, redirecionam para uma página contendo a descrição do objeto, que pode ser enzima, composto químico ou outra rede metabólica, por exemplo. Continuando com o exemplo da biossíntese do *backbone* de terpenóides, para acessar seu diagrama, apresentado na figura 2.8, basta navegar a partir da *home page*:

<http://www.kegg.jp/> \Rightarrow
Data-oriented entry points \mapsto KEGG PATHWAY \Rightarrow
Pathway Maps \mapsto 1. Metabolism \Rightarrow
1.9 Metabolism of terpenoids and polyketides \mapsto Terpenoid backbone biosynthesis

2.3.4 BioCyc

O BioCyc¹⁶ é um sistema de coleção de aproximadamente 7 mil bancos de dados chamados PGDBs (*Pathway/Genome Databases*) pois possuem duas maneiras diferentes de representar as informações: modelo de vias metabólicas, que enfatiza as sequências de reações, substratos e produtos de múltiplos organismos, ou modelo de sequência genômica, que destaca a localização e descrição dos genes de cada organismo específico [1]. Os bancos PGDBs são organizado em três camadas de acordo com a frequência de atualizações/refinações e da maneira com que os dados foram obtidos. O BioCyc possui um banco de dados específico para redes metabólicas determinadas experimentalmente, chamado MetaCyc¹⁷. Este é o único banco de dados multi-organismos do grupo BioCyc e ele é referência na ferramenta gratuita *Pathway Tools* desenvolvida pelo instituto de pesquisa *SRI International*.

Diferente das ferramentas do KEGG e Reactome, o Pathway Tools não possui um mapa global de visualização das vias metabólicas. Para acessá-las, o usuário deve buscar por alguma palavra chave conhecida ou navegar pelo sumário de vias¹⁸ classificado hierarquicamente com base em suas funções biológicas e na classe dos metabólitos produzidos/consumidos. O acesso à via da Figura 2.3, por exemplo, se deu pela navegação a partir da *home page*:

<http://metacyc.org/> \Rightarrow
Metabolism \mapsto Browse Pathway Ontology \Rightarrow
Pathways \mapsto Bioluminescence (1 instances) \mapsto bacterial bioluminescence

Cada nó das vias metabólicas quando clicado redireciona o usuário para uma página com mais detalhes sobre o objeto, que pode ser outra via, um composto químico ou

¹⁵Detalhes sobre a notação dos diagramas: http://www.genome.jp/kegg/document/help_pathway.html

¹⁶Disponível pela *web* através do site <http://biocyc.org>.

¹⁷Disponível pela *web* através do site <http://metacyc.org/>.

¹⁸Na última visita, em 2016-10-06, a página possuía 3461 vias metabólicas.

uma enzima, por exemplo. A Figura 2.9 é uma modificação da Figura 2.3 apresentando diretamente o conjunto destes elementos.

2.3.5 2Path

Conteúdo estará disponível quando site estiver pronto.

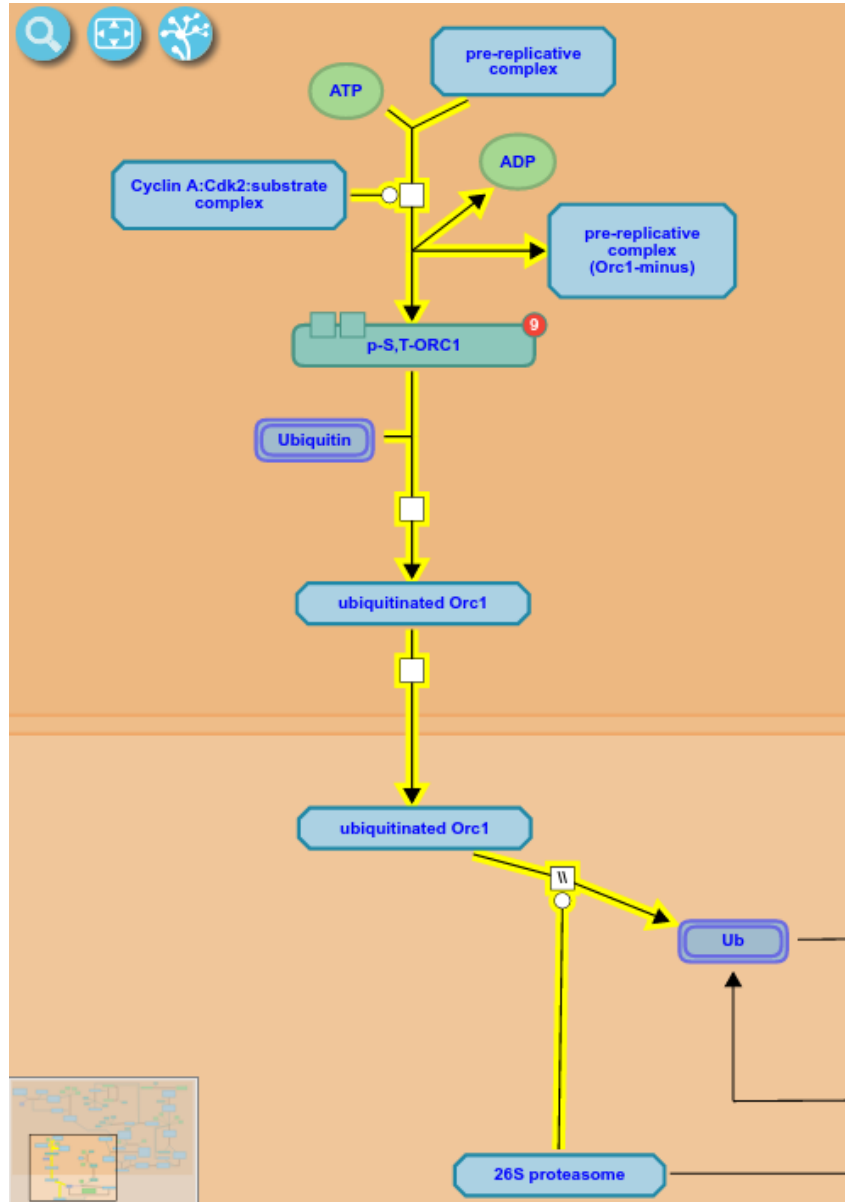


Figura 2.7: Via metabólica de remoção do Orc1 da cromatina.

TERPENOID BACKBONE BIOSYNTHESIS

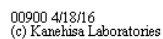


Figura 2.8: Via metabólica de biossíntese do *backbone* de terpenóides.

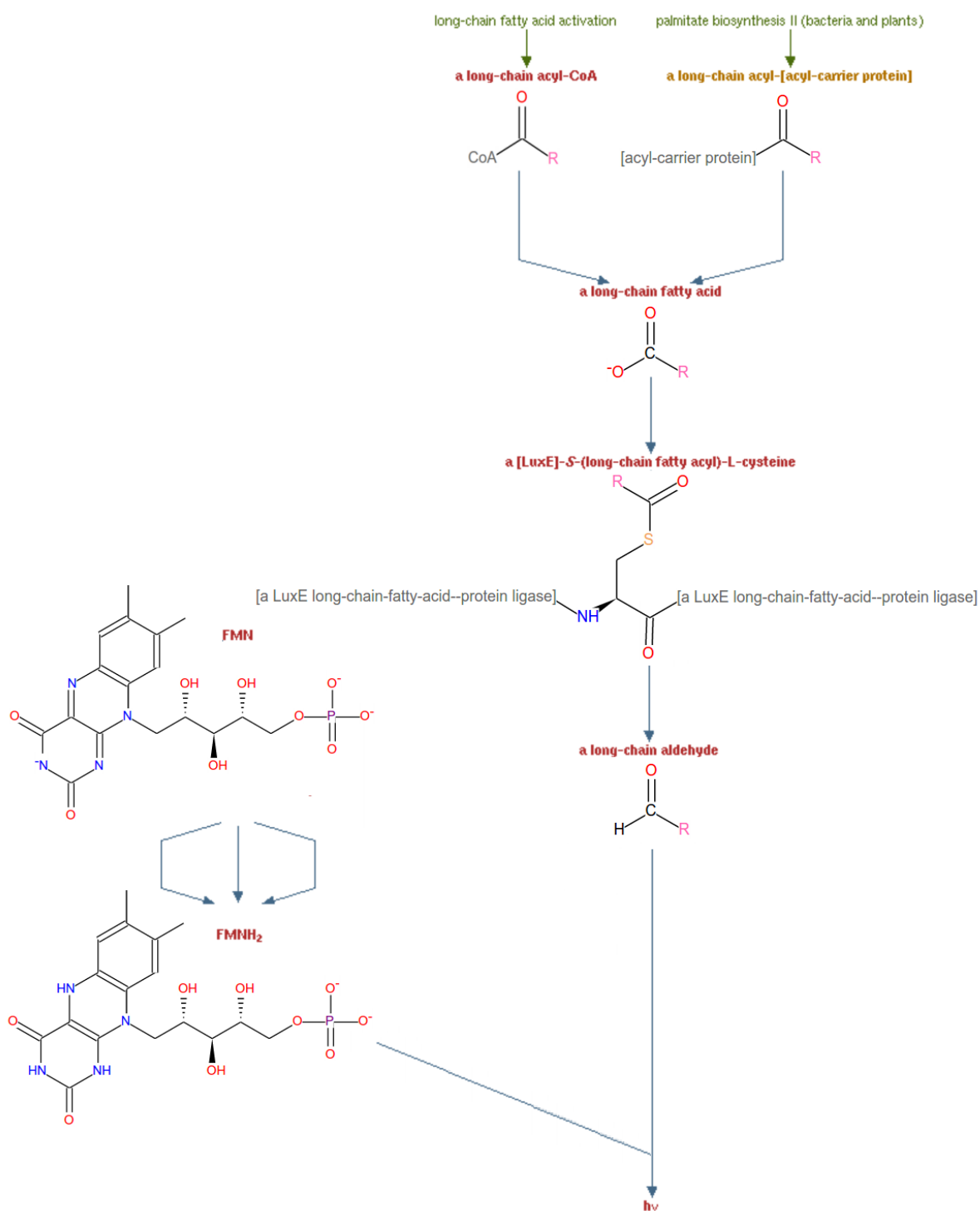


Figura 2.9: Representação modificada da Figura 2.9 com maior nível de detalhes.

Capítulo 3

Projeto da Interface

3.1 Projeto

Aplicação do método da avaliação da comunicabilidade

3.2 Detalhes da Implementação

O sistema desenvolvido para este projeto é uma aplicação web chamada *2Path*. O usuário deve se cadastrar no *website* para ter acesso às redes metabólicas do banco de dados do sistema, bem como pesquisar por palavras chaves no mesmo. Nesta seção serão apresentadas as linguagens e ferramentas utilizadas no desenvolvimento do *website*, as características, funcionalidades e limites do sistema e, por fim, as dificuldades enfrentadas na implementação do projeto.

O sistema foi desenvolvido no ambiente de desenvolvimento integrado *open source* Eclipse Java EE - *Java Platform, Enterprise Edition*, versão Mars 4.5.2. Para simplificar a obtenção das dependências do projeto, ou seja, pacotes de arquivos java (extensão .jar), foi utilizada o Apache Maven, *software* de gerenciamento de projeto e ferramenta de compreensão de programa. Este *software* opera sobre o arquivo *pom.xml*, onde POM significa *Project Object Model* e contém as especificações de cada projeto que se tornará dependência do sistema em desenvolvimento, além de outros aspectos do código. O servidor selecionado para hospedagem local, *localhost* porta 8080, do sistema foi o Apache TomCat versão 7.0.

As páginas da aplicação foram desenvolvidas na linguagem de marcação XHTML, *Extensible Hypertext Markup Language*, e a estilização em CSS, *Cascading Style Sheets*.

JSF
PRIMEFACES
JAVASCRIPT
ANGULARJS
ORIENTBD

3.3 Desafios

O que foi o trabalho. Descrever todo o ambiente usado Neste capítulo serão apresentados os primeiros resultados experimentais obtidos.

Capítulo 4

Resultados

4.1 Interface

4.2 Avaliação e Discussão

Capítulo 5

Método e Resultados

Capítulo 6

Conclusão e Trabalhos Futuros

6.1 Conclusão

Neste capítulo serão apresentadas as considerações finais do trabalho, assim como as limitações e dificuldades encontradas.

6.2 Trabalhos Futuros

A partir deste trabalho, foi possível identificar os seguintes pontos a serem melhorados:

- x

Capítulo 7

Cronograma

O cronograma está apresentado na Tabela a seguir, mostrando o início das atividades em Janeiro de 2016 com a revisão literária e com término previsto para Junho de 2016, juntamente com a defesa do Trabalho de Conclusão de Curso.

Tabela 7.1: Cronograma

| Atividades | 2016 | | | | | |
|---|------|-----|-----|-----|-----|-----|
| | Jul | Ago | Set | Out | Nov | Dez |
| Revisão bibliográfica | X | X | | | | |
| Familiaridade com ambiente de desenvolvimento | | X | X | | | |
| Implementação da aplicação | | X | X | X | | |
| Interpretação dos resultado | | | | X | X | X |
| Defesa | | | | | | X |

Referências

- [1] Introduction to biocyc. <http://biocyc.org/intro.shtml>, visitado em 2016-10-01. 21
- [2] Kegg overview. <http://www.kegg.jp/kegg/kegg1a.html>, visitado em 2016-10-01. 19
- [3] Proteínas. <http://www.professoraangela.net/documents/proteinas.html>, visitado em 2016-01-02. 12
- [4] Usersguide. <http://wiki.reactome.org/index.php/Usersguide>, visitado em 2016-10-01. 18
- [5] Protein structure, 2009. <http://www.particlesciences.com/news/technical-briefs/2009/protein-structure.html>, visitado em 2016-01-02. 13
- [6] Abhishek Bansal. Nosql, cap theorem (part of mongodb course). <https://www.knowbigdata.com/blog/nosql-cap-theorem-part-mongodb-course>, visitado em 2016-10-13. vii, 17
- [7] Simone Diniz Junqueira Barbosa and Bruno Santana da Silva. *Interação Humano-computador*. Elsevier Editora Ltda., 2010. viii, 3, 4, 5, 6, 7, 8
- [8] Perry Carter. Catabolic and anabolic reactions. <http://classes.midlandstech.edu/carterp/courses/bio225/chap05/lecture1.htm>, visitado em 2016-10-12. 13
- [9] Ron Caspi. Metacyc pathway: bacterial bioluminescence. <http://metacyc.org/META/new-image?object=PWY-7723>, visitado em 2016-10-13. vii, 14
- [10] Rick Cattell. Scalable sql and nosql data stores. *SIGMOD Rec.*, 39(4):12–27, May 2011. 16
- [11] Claudio Manoel de Carvalho Correia. Semiose e desenvolvimento cognitivo: Estudo sobre as estratégias de construção dos processos sógnicos em seqüências lógicas, 2001. <http://www.filologia.org.br/vicnlf/anais/semiose.html>, visitado em 2016-10-06. 5
- [12] Jing Han, E. Haihong, Guan Le, and Jian Du. Survey on NoSQL database. In *Pervasive Computing and Applications (ICPCA), 2011 6th International Conference on*, pages 363–366. IEEE, October 2011. 16

- [13] Justin N. Kabera, Edmond Semana, Ally R. Mussa, and Xin He. Plant secondary metabolites: Biosynthesis, classification, function and pharmacological properties. *Journal of Pharmacy and Pharmacology*, 2:377–392, 2014. 15
- [14] Nicole Legault. Affordance: What does it mean for e-learning? <https://community.articulate.com/articles/affordance-what-does-it-mean-for-e-learning>, visitado em 2016-10-06. vii, 4
- [15] Gerhard Michal and Dietmar Schomburg. *The Cell and Its Contents*, pages 14–36. John Wiley & Sons, Inc., 2012. 13, 14
- [16] David W. Mount. *Bioinformatics : sequence and genome analysis*. Cold Spring Harbor, N.Y. Cold Spring Harbor Laboratory Press, 2001. 15
- [17] Leslie A. Pray. Discovery of dna structure and function: Watson and crick, 2008. <http://www.nature.com/scitable/topicpage/discovery-of-dna-structure-and-function-watson-397>, visitado em 2016-01-15. vii, 10, 11, 12
- [18] João Carlos Setubal and João Meidanis. *Introduction to computational molecular biology*. PWS Publishing Company, 1997. 11, 12, 13
- [19] Chad Vicknair, Michael Macias, Zhendong Zhao, Xiaofei Nan, Yixin Chen, and Dawn Wilkins. A comparison of a graph database and a relational database: A data provenance perspective. In *Proceedings of the 48th Annual Southeast Regional Conference*, ACM SE '10, pages 42:1–42:6, New York, NY, USA, 2010. ACM. 16
- [20] Röbbbe Wünschiers, Martina Jahn, Dieter Jahn, Ida Schomburg, Susanne Peifer, Elmar Heinzle, Helmut Burtscher, Julia Garbe, Annika Steen, Max Schobert, Dieter Oesterhelt, Josef Wachtveitl, and Antje Chang. *Metabolism*, pages 37–209. John Wiley & Sons, Inc., 2012. 14, 15