UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

"UMA ABORDAGEM PARA OBTENÇÃO DE REGIÕES ORTÓLOGAS EM MÚLTIPLOS PROTEOMAS"

ANDERSON PEGORARO SILVA

São Carlos
Agosto/2006

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

S586ao

Silva, Anderson Pegoraro.
Uma abordagem para obtenção de regiões ortólogas em múltiplos proteomas / Anderson Pegoraro Silva. -- São Carlos : UFSCar, 2008.
67 f.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2006.

1. Biologia computacional. 2. Genomas. 3. Banco de dados. 4. Bioinformática. 5. Teoria dos grafos. I. Título.

CDD: 574.0285 (20ª)

Universidade Federal de São Carlos

Centro de Ciências Exatas e de Tecnologia

Programa de Pós-Graduação em Ciência da Computação

"Uma abordagem para Obtenção de Regiões Ortólogas em Múltiplos Proteomas"

ANDERSON PEGORARO SILVA

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

Membros da Banca:

Prof. Dr. Mauro Biajiz
(Orientador – DC/UFSCar)

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

Prof. Dr. Marcelo Augusto Santos Turine
(DCT/UFMS)

São Carlos
Agosto/2006

Dedico este trabalho:

Aos meus pais,
pelo exemplo de amor,
trabalho e dedicação.

AGRADECIMENTOS

Não poderia deixar de agradecer a algumas pessoas que, de alguma forma, com mais ou menos intensidade, me auxiliaram para a conclusão deste trabalho.

Agradeço especialmente ao meu orientador, Prof. Dr. Mauro Biajiz, pela amizade, paciência e, principalmente, pela liberdade e incentivo concedidos para a escolha do trabalho.

Agradeço ao Prof. Dr. Nalvo Franco de Almeida Junior pelo suporte dado no desenvolvimento do projeto.

Agradeço aos meus pais e irmãos por sempre apoiarem as minhas decisões, especialmente a decisão de continuar os estudos.

Agradeço a dois grandes amigos, José Leite (Rato) e Daniel Gatera (Morto), que me acompanham desde os tempos de graduação. O primeiro pelas idéias valiosas e dicas de implementação, o segundo pela revisão da tese.

Agradeço também ao meu primo Daniel pela companhia quase que diária e pela ajuda na parte de biologia.

Agradeço ao meu amigo Vinicius Martelli (Vica), com o qual eu compartilhei grande parte dos momentos vividos no mestrado.

Agradeço ao meu sobrinho Bruno pelos momentos de descontração.

Por fim, agradeço aos meus grandes amigos Jonas, Thiago Jabur, Felipe e Camila Camargo, que não poderiam ficar fora dessa lista.

PENSAMENTO

"A imaginação é mais importante que o conhecimento”.

Albert Einstein.

RESUMO

Com o avanço das técnicas de seqüenciamento e o inevitável crescimento do número de genomas seqüenciados, torna-se necessário o uso de técnicas computacionais para analisar e gerenciar essa grande massa de dados. Uma das análises possíveis diz respeito à extração de características funcionais e evolutivas dos organismos estudados. Assim, nesta linha de pesquisa, este estudo preocupa-se em identificar regiões comuns, em termos dos genes que elas contêm, em múltiplos proteomas, que conservam a ordem e o conteúdo gênico. O problema é modelado com o auxílio de um grafo colorido, sendo que regiões comuns entre proteomas são como cliques no grafo. Aproveitando-se de peculiaridades do grafo construído, um algoritmo para redução do espaço de busca foi desenvolvido, possibilitando obter resultados completos em um pequeno espaço de tempo. Portanto, a contribuição deste trabalho constitui numa abordagem para encontrar regiões comuns em múltiplos proteomas, acrescida de uma implementação, que originou a ferramenta *Multiple Proteome Comparison* (MPC).

ABSTRACT

With the progress of the sequencing techniques and the inevitable increasing number of sequenced genomes, it becomes interesting studying computational techniques to analyze these data. One of the possible analyses refers to the extraction of functional and evolutionary characteristics of the studied organisms. Thus, in this line of research, this study is motivated in identifying common regions, in terms of the genes that they contain, in multiple proteomes that keep the order and the gene content. The problem is modeled with a colored graph, and the common regions between proteomes are like clicks in the graph. Using peculiarities of the constructed graph, an algorithm for search space reduction was developed, making it possible to get complete results in a short time. Therefore, the contribution of this work constitutes an approach to find common regions in multiple proteomes, added of an implementation from which the Multiple Proteome Comparison (MPC) tool originated.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

BBH	*Bidirectional Best Hit*
BLAST	*Basic Local Alignment Search Tool*
BLOSSUM	*Blocks Substitution Matrix*
DDBJ	*DNA DataBank of Japan*
DNA	*Deoxyribonucleic acid* (Ácido desoxirribonucleico)
EGG	*Extended Genome-Genome Comparison*
EMBL	*European Molecular Biology Laboratory*
KEGG	*Kyoto Encyclopedia of Genes and Genomes*
MBGD	*Microbial Genome Database*
MPC	*Multiple Proteome Comparison*
NCBI	*National Center For Biotechnology Information*
NIH	*National Institute of Health*
PAM	*Point Accepted Mutation*
PCDB	*Pairwise Comparison Database*
RGC	Região de Genes Consecutivos
RNA	*Ribonucleic acid* (Ácido ribonucleico)
RO	Região Ortóloga
ROM	Região Ortóloga Múltipla
SGBD	Sistema Gerenciador de Banco de dados
UNIPROT	*Universal Protein Resource*

CAPÍTULO 1

1. INTRODUÇÃO

1.1. Motivação e Objetivos

O genoma humano foi seqüenciado em 2000, estabelecendo o que foi denominado "fim da era pré-genômica". Antes desse marco, muitos imaginavam que o trabalho envolvendo o genoma humano seria concluído assim que o seu seqüenciamento fosse finalizado e, conseqüentemente, os benefícios desse estudo estariam à disposição de todos.

Na realidade, a etapa de seqüenciamento de um genoma é apenas um primeiro passo, pois na era pós-genômica faz-se necessária a junção de experimentos biológicos e análise computacional para interpretar os dados previamente gerados, objetivando-se extrair informações biológicas. A tarefa computacional é chamada de biologia computacional ou Bioinformática (LENGAUER, 2000).

Além do genoma humano, outros organismos estão sendo seqüenciados a cada dia. Como conseqüência dessa "nova" demanda, novas tecnologias de mapeamento, seqüenciamento e análise de seqüências têm sido desenvolvidas, sendo esta última a de maior interesse para os bioinformatas.

Sabendo-se que os "Projetos Genoma" têm como objetivo a determinação das seqüências de bases do DNA de espécies biológicas e o mapeamento de seus genes, seqüenciar outros organismos facilita o descobrimento de funções de genes e seqüências do genoma humano, pois se duas seqüências são similares, pode-se supor que suas funções também as sejam.

Com o desenvolvimento das tecnologias de seqüenciamento, bancos de dados robustos foram construídos para armazenar e gerenciar a explosão no número de seqüências obtidas pelos pesquisadores. Como exemplo desses bancos de dados, destaca-se o GenBank, que é um banco de dados público de seqüências de DNA e proteínas mantido pelo NCBI *(National Center For Biotechnology Information)* e serve como referência para verificar se um organismo foi seqüenciado (NCBI, 2006).

Visto o constante crescimento no número de seqüenciamentos, como ilustrado na Figura 1.1 o aumento exponencial do número de seqüências contidas do GenBank e, conseqüentemente, a grande quantidade de dados que necessita ser analisada para se extrair

significados biológicos dos mesmos, ferramentas de comparação de genomas são construídas e aprimoradas constantemente.

Com o surgimento dos bancos de dados públicos do genoma, houve uma disseminação do conhecimento adquirido com o seqüenciamento de vários organismos, devido ao acesso gratuito e fácil por meio da *web*. Isto motivou o desenvolvimento de ferramentas de bioinformática que realizem buscas por pequenas variações nos genomas de diferentes indivíduos e espécies, a partir de dados armazenados em grandes bancos de dados públicos.

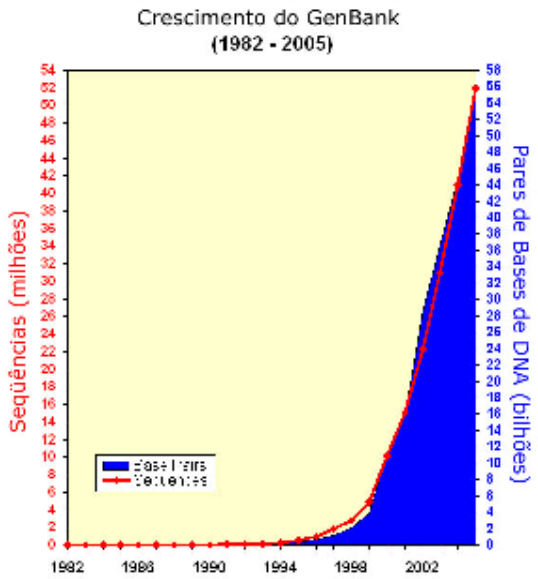


Figura 1.1 – Crescimento exponencial do número de seqüências contidas no GENBANK (GENBANK, 2006)

Com uma disponibilidade maior no número de seqüências genômicas, a comparação de genomas completos com dados organizados e com a utilização de diversas técnicas computacionais, tem se tornado útil não apenas como um mecanismo para encontrar características comuns entre diversos genomas, mas também como uma abordagem para entender processos evolutivos entre múltiplos genomas (CHOI *et al.*, 2005).

Em Almeida (2002) uma ferramenta de comparação de genomas, denominada EGG, foi desenvolvida. EGG proporciona encontrar, dentre outras coisas, genes ortólogos[1] e suas regiões, através de comparações dois a dois. Com base nessas comparações e com a necessidade de se realizar comparações entre múltiplos genomas, foi desenvolvida a ferramenta BAGRE (MONTERA, 2004).

[1] O conceito de genes ortólogos será apresentado na seção 2.4.

O objetivo de BAGRE é proporcionar, através de uma interface *web,* a realização de comparações simultâneas de vários genomas, com o propósito de encontrar grupos de genes em vários genomas que preservam a ordem e o conteúdo gênico. Para isso, BAGRE utiliza-se de heurísticas para a determinação dos resultados, uma vez que o problema é np-completo, o que pode provocar resultados imprecisos ou incompletos.

A justificativa para a escolha do tema deste trabalho dá-se pela necessidade de evolução das ferramentas existentes de comparação de genomas, com atenção especial à ferramenta BAGRE. Neste sentido, este trabalho destina-se a encontrar regiões ortólogas múltiplas[2] (ROM), em termos dos genes que elas contêm, através da comparação de proteomas[3] de seres procariotos, preocupando-se em buscar resultados completos e precisos. O principal objetivo é encontrar "pistas" de funcionalidades comuns a partir de regiões ortólogas conservadas entre os vários organismos envolvidos.

A escolha de seres procariotos deve-se à proximidade filogenética, que por conseqüência tendem a preservar a ordem e a funcionalidade dos genes. A conservação da ordem de genes entre diferentes organismos é utilizada na predição de funções e também no estudo de relacionamentos evolucionários (TAMAMES, 2001).

Para a modelagem da solução proposta, utiliza-se um grafo colorido[4]. Cada Região de Genes Consecutivos[5] (RGCs) é representada por um vértice no grafo, e a ortologia de duas regiões por uma aresta ligando esses vértices, sendo que os vértices referentes ao mesmo proteoma pertencem a mesma classe de coloração. As Regiões Ortólogas Múltiplas (ROMs) são expressas no grafo através de cliques.

Aproveitando-se de peculiaridades do grafo construído, um algoritmo de eliminação de vértices foi desenvolvido. Como uma RGC é um conjunto de genes, utiliza-se de uma função para medir a similaridade entre conjuntos e decidir se haverá ou não um agrupamento de RGCs. Para possibilitar a realização de testes, outra contribuição do trabalho destina-se ao desenvolvimento de uma ferramenta para a comparação de múltiplos proteomas, denominada MPC (*Multiple Proteome Comparison*).

---

[2] O conceito de Região Ortóloga Múltipla será apresentado no capítulo 3.
[3] Proteoma é o conjunto de genes de um genoma que codificam proteína.
[4] O conceito de grafo colorido será apresentado no capítulo 4.
[5] Uma região de genes consecutivos é um conjunto de genes consecutivos de um proteoma.

1.2. Organização da Dissertação

O trabalho está organizado da seguinte forma. O Capítulo 2 apresenta conceitos básicos de biologia molecular, abordando os principais temas necessários para a contextualização e entendimento do trabalho. No Capítulo 3 é feito um levantamento dos principais bancos de dados do genoma e de ferramentas de comparação de genomas. O Capítulo 4 descreve a principal contribuição deste trabalho, que se trata de uma abordagem para encontrar ROMs. Por fim, o Capítulo 5 é destinado à conclusão do trabalho, apresentando as principais contribuições e propostas de trabalhos futuros.

CAPÍTULO 2

2. CONCEITOS BÁSICOS DE BIOLOGIA MOLECULAR

2.1. Considerações Iniciais

Biologia molecular é o ramo da biologia que trata da formação, estrutura e função de macromoléculas essenciais à vida, como ácidos nucléicos (DNA e RNA) e proteínas (ANSWERS, 2006). A biologia molecular apresenta uma multidisciplinaridade por se relacionar diretamente com outras áreas da biologia, como genética e bioquímica.

2.2. Células

Todos os seres vivos são formados de células- compartimentos envolvidos por membrana, preenchidos com uma solução aquosa concentrada de substâncias químicas (ALBERTS *et al.,* 2002a*;* ALBERTS *et al.,* 2002b).

Organismos superiores, como os humanos, são como cidades celulares nas quais grupos de células performam tarefas especializadas e são ligadas por um complexo sistema de comunicações. Nelas são realizadas a maior parte dos processos metabólicos dos seres vivos e nelas está contido o material genético. A membrana celular contém poros que permitem a interação com o ambiente. Dessa forma, a célula recebe nutrientes necessários para desenvolver-se e expele substâncias.

Um aspecto importante na evolução dos organismos deve ter ocorrido há 1,5 milhão de anos, quando houve a transição de células pequenas com estruturas simples – os procariotos – para células eucarióticas com estruturas mais complexas (ALBERTS *et al.,* 2002a*;* ALBERTS *et al.,* 2002b).

Os seres procariotos incluem os diversos tipos de bactérias. Estas são células esféricas ou em forma de bastonetes curtos de tamanhos que variam de 0,1 a 600 μm (ABEDON, 2006). Na maioria dessas espécies, a proteção da célula é feita por uma camada altamente resistente: a parede celular. Logo abaixo encontra-se a membrana citoplasmática, que delimita um único compartimento contendo DNA, RNA, proteína e moléculas. Um exemplo de célula procariótica pode ser visto na Figura 2.1(a).

Em contraste com as células procarióticas, as células eucarióticas possuem um núcleo, que contém a maioria do DNA celular, envolvido por uma membrana de dupla camada, sendo que os outros componentes celulares encontram-se no citoplasma. No citoplasma, organelas distintas podem ser reconhecidas, dentre elas as mitocôndrias, o complexo de Golgi e os cloroplastos, este no caso de células capazes de fotossíntese. Um exemplo de célula eucariótica pode ser visto na Figura 2.1(b).

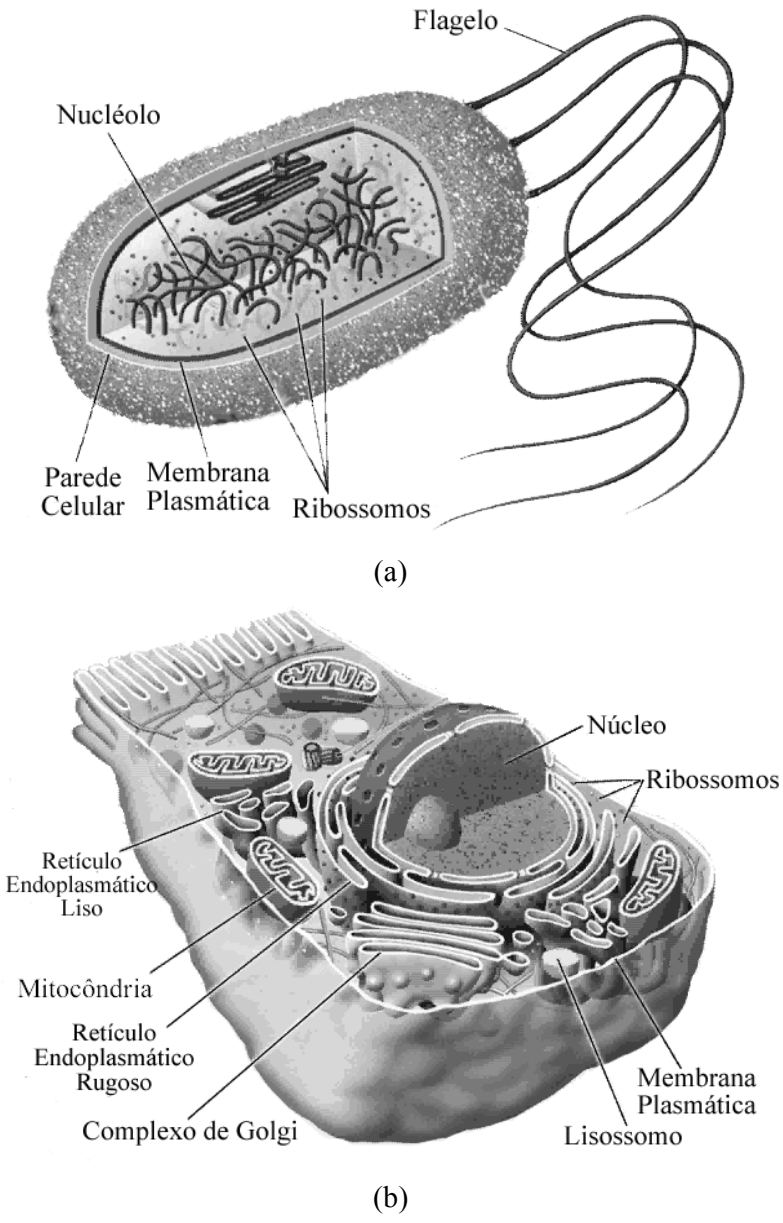


Figura 2.1 – Organização Celular. (a) Célula Procariótica; (b) Célula Eucariótica (COOPER, 1996).

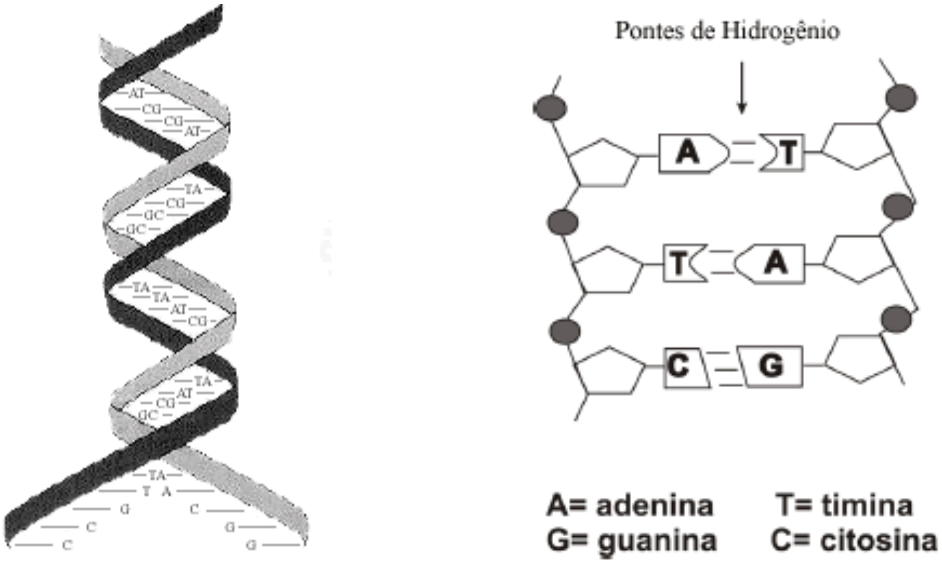
A propriedade fundamental de uma célula está na sua capacidade de crescer e replicar-se, gerando células descendentes contendo cópias do seu material genético (através das biomoléculas de ácidos nucléicos, ou DNA). Isso é resultado de uma série de processos metabólicos intrínsecos e complexos desencadeados dentro da célula.

2.3. Ácidos Nucléicos

Na natureza há dois tipos de ácidos nucléicos: DNA, ou ácido desoxirribonucléico, e RNA, ou ácido ribonucléico. A composição dos ácidos nucléicos dá-se por uma longa cadeia de nucleotídeos, sendo que cada nucleotídeo é composto de um açúcar, um fosfato e uma base nitrogenada.

Existem quatro tipos de bases nitrogenadas no nucleotídeo, as purinas: adenina (A) e guanina (G), e as pirimidinas: citosina (C), timina (T) e uracila (U). No DNA são encontradas as bases A, G, C e T e no RNA encontram-se as bases A, G, C e U. Cada conjunto de três bases, denominado códon, na seqüência ao longo da fita de DNA, sinaliza às células um determinado aminoácido a ser usado na síntese de proteína (descrita na seção 2.5).

O DNA está representado na Figura 2.2(a). É a molécula base do material genético encontrado em todas as células, sendo responsável por carregar as informações genéticas de uma geração para a próxima. Trata-se de uma molécula muito longa e fina, formada por uma dupla fita de orientações opostas. A interação entre as duas fitas é feita através de pontes de hidrogênio, sendo que as guaninas pareiam-se com as citosinas através de três pontes de hidrogênio, enquanto as adeninas pareiam-se com as timinas através de duas, como ilustra a Figura 2.2(b).



(a) (b)

Figura 2.2 – Molécula de DNA. (a) Exemplo de Representação da dupla fita de DNA; (b) Interação entre duas fitas de DNA através de pontes de hidrogênio (SILVA, 2001).

Nas fitas do DNA, segue-se uma regra de pareamento, sendo as orientações classificadas de extremidade 5' e extremidade 3', na qual cada fita é o complemento reverso da outra, conforme ilustra a Figura 2.3.

```
5'   ... A T T T C C G A T G C ... 3'
         | | | | | | | | | | |
3'   ... T A A A G G C T A C G ... 5'
```

Figura 2.3 - Pareamento de uma molécula de DNA

As moléculas de RNA são compostas por uma fita simples, e possuem tamanho reduzido se comparadas às moléculas de DNA. As células possuem três tipos diferentes de RNA: RNA mensageiro (mRNA), RNA transportador (tRNA) e RNA ribossomal (rRNA).

- mRNA: Contêm a informação para a síntese de proteínas;
- tRNA: Transportam aminoácidos para que ocorra a síntese de proteínas;
- rRNA: Componentes da maquinaria de síntese de proteínas presente nos ribossomos.

Cada tipo tem um papel específico a desempenhar, juntamente com o DNA, nas reações de síntese protéica (seção 2.5).

2.4. Relações Gênicas

Sob o ponto de vista da aparência exterior, a evolução transformou o universo dos seres vivos em tal grau que eles não são mais reconhecidos como parentes. O ser humano, uma mosca, uma margarida, uma levedura, uma bactéria parecem tão diferentes que parece inconcebível compará-los. Ainda assim, todos descendem de um ancestral comum, e quanto maior a profundidade de investigação, mais e mais evidências de uma mesma origem é encontrada (ALBERTS *et al.,* 2002a*;* ALBERTS *et al.,* 2002b).

Genes evolucionariamente relacionados e sua existência revelam uma estratégia básica pela qual, organismos mais complexos surgiram, ou seja, genes ou porções de genes duplicaram-se e as novas cópias divergiram das originais por mutações e recombinações, para se ajustar a tarefas adicionais.

Começando com uma coleção de genes nas células primitivas, para se adaptar a tarefas adicionais, as formas de vida mais complexas foram capazes de desenvolverem mais de 50.000 genes, hoje presentes em uma célula de um animal ou vegetal (ALBERTS *et al.,* 2002a*;* ALBERTS *et al.,* 2002b).

Um gene pode ser classificado como uma pequena parte codificante do genoma, responsável pela determinação dos traços hereditários dos organismos vivos, e um genoma pode ser entendido como o material hereditário total de uma célula de uma determinada espécie (SILVA, 2001).

Uma Região de Genes Consecutivos (RGC) é um conjunto de genes consecutivos num genoma. Sendo que o próprio genoma consiste em uma RGC. O conjunto de genes de um genoma que codificam proteína denomina-se proteoma (ALMEIDA, 2002).

O genoma de uma célula pode ser facilmente visualizado em longas seqüências de DNA que são compactadas durante o final da mitose na forma de cromossomos individualizados (SILVA, 2001).

Uma diferença notável entre seres eucariotos e procariotos acontece também entre seus genes. O material genético de um eucarioto é dividido em regiões codificadoras (*exons*) alternando-se com regiões não-codificadoras (*introns*). No caso dos procariotos, o material genético é conhecido como genes codificadores de proteínas sem *introns*. Assim,

genes de bactérias são consideravelmente mais simples do que genes de organismos mais complexos, como é o caso dos seres humanos.

Os genes de bactérias que codificam proteínas podem ser descritos conforme a Figura 2.4. Existem três seqüências em um gene: (i) promotora; (ii) codificante e; (iii) terminadora (*Stop codon*) (MIR, 2004).

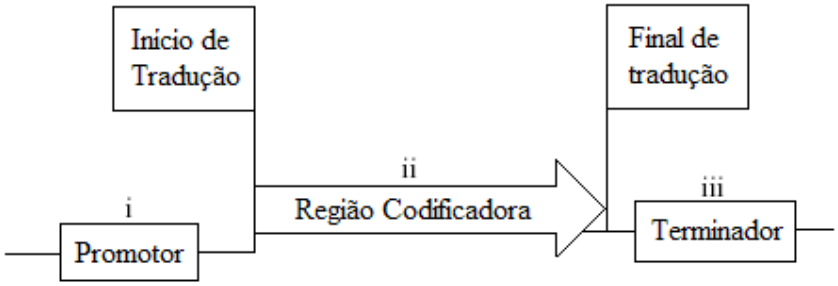


Figura 2.4 – Estrutura esquemática de um gene de um procarioto (MIR, 2004).

Toda relação entre genes pode ser classificada como ortologia ou paralogia. Essas relações dizem respeito a eventos de evolução gênica. Genes que descendem de um mesmo ancestral são homólogos. Já os genes ortólogos, são genes que além de homólogos, pertencem a genomas diferentes. Por fim, os genes parálogos são genes que além de homólogos, pertencem a um mesmo genoma.

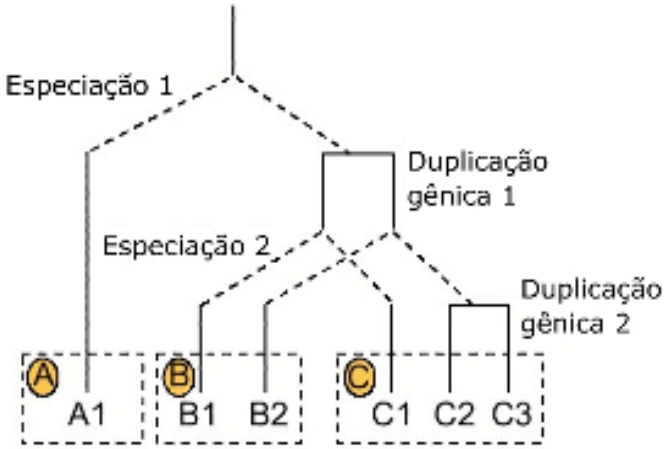


Figura 2.5– Ortologia e Paralogia (JENSEN, 2001)

Deve-se perceber também, que é possível que um dado gene em uma espécie tenha mais do que um ortólogo em outra. Além disso, detectam-se genes parálogos em espécies

diferentes. Isto pode ser constatado na Figura 2.5, pois A1 tem três ortólogos na espécie C, mas somente C1 é ortólogo de B1. Já B2 tem dois ortólogos na espécie C (C2 e C3), onde C1 e B2 são parálogos (JENSEN, 2001).

2.5. Síntese de Proteínas

Proteínas são biomoléculas longas, formadas por combinações dos 20 diferentes tipos de aminoácidos (Tabela 1) ligados por cadeias polipeptídicas. São moléculas extremamente versáteis que podem possuir dezenas ou centenas de aminoácidos.

A síntese das proteínas inicia-se pela transcrição do DNA para mRNA no núcleo da célula. Os nucleotídeos adequados pareiam-se (DNA:mRNA= A:U, T:A, C:G e G:C). Dessa forma, a molécula de RNA formada copia a mensagem contida no DNA (intervém nessa cópia uma enzima, a RNA-polimerase), processo conhecido como transcrição.

O mRNA migra do núcleo para o citoplasma, onde interage com um ou mais ribossomos, formando um molde para a síntese de proteínas. O próximo passo é a "leitura" (tradução) do mRNA para a proteína, intermediado por diversos tipos de tRNA, que se encontram ligados a aminoácidos simples que estão no citoplasma. Cada aminoácido interage com um tipo específico de tRNA, que o encaixa no códon adequado do mRNA.

Genericamente, o mRNA indica qual a seqüência de aminoácidos terá uma determinada proteína. Após a síntese, as novas moléculas de proteína recém-sintetizadas separam-se do ribossomo e são direcionadas para a sua localização específica na célula. A Figura 2.6 ilustra, sucintamente, como ocorre a expressão gênica.

Desse modo, as mensagens do DNA servem para que a célula fabrique proteínas específicas, processo este que repete continuamente. Em resumo: DNA "produz" RNA e RNA "produz" proteína.

Tabela 1 - Tabela de códons (DALTON, 2006)

Nome	Abreviação (3 letras)	Abreviação (1 letra)	Códons de DNA para cada Aminoácido
Alanine	Ala	A	GCA, GCC, GCG, GCU
Cysteine	Cys	C	UGC, UGU
Aspartic Acid	Asp	D	GCA, GAU
Glutamic Acid	Glu	E	GAA, GAG
Phenylalanine	Phe	F	UUC, UUU
Glycine	Gly	G	GGA, GGC, GGG, GGU

Histidine	His	H	CAC, CAU
Isoleucine	Ile	I	AUA, AUC, AUU
Lysine	Lys	K	AAA, AAG
Leucine	Leu	L	UUA, UUG, CUA, CUC, CUG, CUU
Methionine	Met	M	AUG
Asparagine	Asn	N	AAC, AAU
Proline	Pro	P	CCA, CCC, CCG, CCU
Glutamine	Gln	Q	CAA, CAG
Arginine	Arg	R	CGA, CGC, CGG, CGU, AGA, AGG
Serine	Ser	S	UCA, UCC, UCG, UCU, AGC, AGU
Threonine	Thr	T	ACA, ACC, ACG, ACU
Valine	Val	V	GUA, GUC, GUG, GUU
Tryptophan	Trp	W	UGG
Tyrosine	Tyr	Y	UAC, UAU
Stop Codons			UAA, UAG, UGA

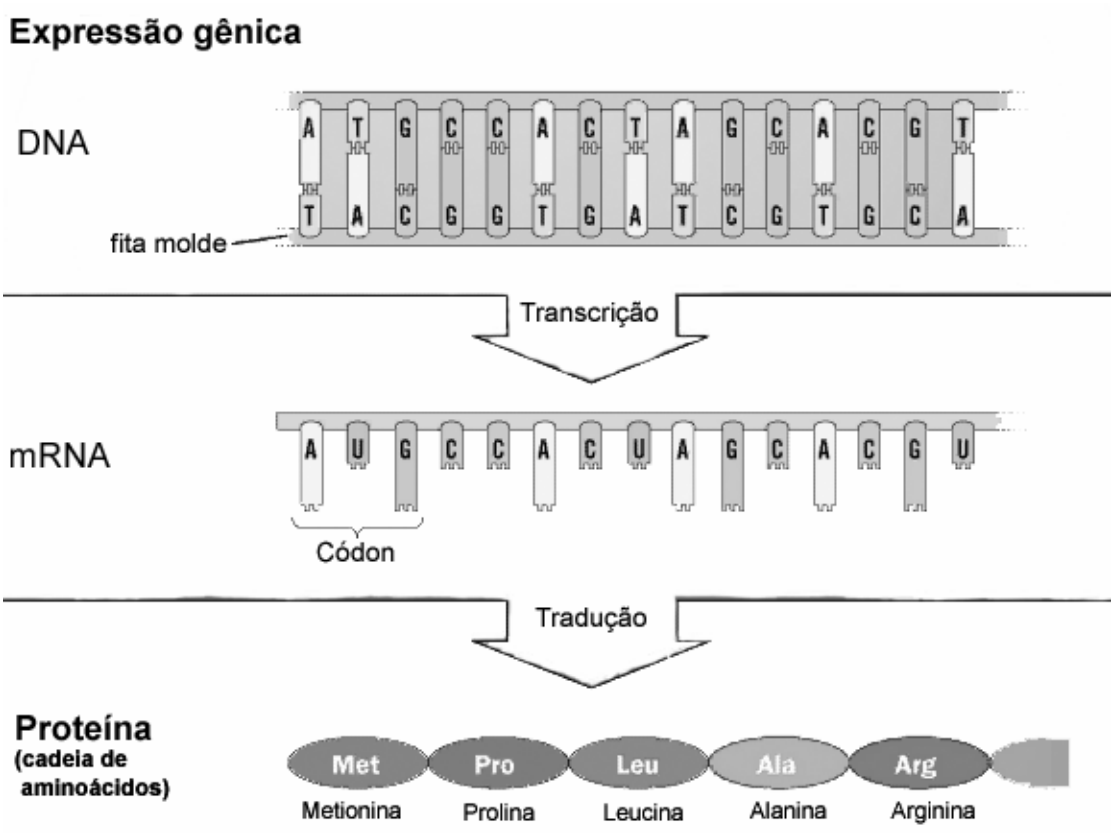


Figura 2.6 - Esquema simplificado da síntese de proteínas. A biomolécula de DNA é transcrita para mRNA (RNA mensageiro), o qual por sua vez é traduzido em uma seqüência de aminoácidos (proteína) (COOPER *et al.*, 1996).

2.6. Considerações Finais

Sucintamente, descrevemos neste capítulo os principais conceitos de biologia necessários para a compreensão desse estudo.

No Capítulo 3, inicialmente são discutidos aspectos relativos ao seqüenciamento de genomas. Logo após, faz-se um levantamento dos principais bancos de dados de genomas e de ferramentas de comparação de seqüências de DNA e proteínas, destacando as principais características.

CAPÍTULO 3

3. BIOINFORMÁTICA E GENÔMICA COMPARATIVA

3.1. Considerações Iniciais

O mapeamento de genomas gera diariamente um volume elevado de informações que são sistematicamente armazenadas em bancos de dados computacionais[6], que servem de fontes de estudo para biologia e medicina. Para analisar estes dados surgiu uma área de pesquisa chamada Bioinformática.

Sucintamente, pode-se dizer que a Bioinformática é a utilização de computadores para tarefas de biologia. Numa definição mais ampla, bioinformática consiste em um conjunto de técnicas advindas da matemática, da estatística e da computação, aplicadas a problemas de biologia molecular, em particular aos problemas de genômica (MIR, 2004).

A Bioinformática é uma área da informática que auxilia os pesquisadores em biologia a criar, melhorar, desenvolver e manipular os bancos de dados de biologia molecular e outras ferramentas computacionais para coletar, organizar e interpretar os dados experimentais obtidos em laboratórios.

O propósito deste capítulo resume-se em apresentar de forma objetiva conceitos relativos à Bioinformática, com um enfoque maior na comparação de genomas.

3.2. Genoma: Seqüenciamento, Montagem e Anotação

Os dados existentes em bancos de dados do genoma são, em grande parte, provenientes de seqüenciamentos. Uma das formas de seqüenciar um genoma é fazê-lo por completo, ou seja, determinar a sua seqüência de bases, podendo então ter seus genes identificados.

O seqüenciamento do genoma se dá por meio de clonagem de fragmentos de DNA extraídos do núcleo das células. Existem diferentes métodos para o seqüenciamento, um dos mais utilizados, é conhecido como shotgun (ADAMS, 2000).

O método de seqüenciamento shotgun caracteriza-se pela construção de uma biblioteca genômica. Para isso, utiliza-se o método de fragmentação aleatória do DNA, que

[6] Os principais bancos de dados serão apresentados na seção 3.3.

gera fragmentos de 500 a 800 pares de base. A seguir, os fragmentos de DNA são clonados em plasmídeos (onde são inseridos material genético, no processo de clonagem) e seqüenciados massivamente, garantindo um grau mínimo de confiabilidade nos resultados, sendo este processo conhecido como cobertura. Logo após o seqüenciamento, é preciso determinar a ordem dos fragmentos, sendo esta etapa denominada montagem.

A etapa de montagem é puramente computacional. Nela utilizam-se as seqüências individuais de cada fragmento, chamadas de *reads*, na reconstrução da seqüência completa da molécula original ou dos fragmentos longos. No entanto, o processo de montagem de fragmentos não é simples, é comum ocorrerem erros no seqüenciamento e problemas durante a quebra e clonagem de *read*s. Quando os *reads* são agrupados formam-se os *contigs* e a união de todos os *contigs* constitui o genoma.

Tendo em mãos a seqüência completa de DNA de um cromossomo ou de um genoma, pode-se dar início ao processo de análise e determinação dos genes dessa seqüência, processo conhecido como anotação.

O primeiro passo da anotação, denominado "identificação", pode ser realizado com o auxílio de ferramentas computacionais, que têm o propósito de automatizar o processo de reconhecimento dos genes em uma seqüência. O segundo passo, denominado "descrição", pode ser parcialmente automatizado, pois necessita de muita intervenção humana para julgar e atribuir a cada seqüência sua função bioquímica.

3.3. Banco de dados do Genoma

Inicialmente, as seqüências genômicas ou bioseqüências eram armazenadas em formato texto, pois a principal fonte de dados eram publicações em artigos científicos. Com o aumento do volume de dados, as bioseqüências passaram a ser gerenciadas por meio de Sistemas Gerenciadores de Banco de dados (SGBDs).

Com uma quantidade significativa de dados oriundos de seqüenciamentos armazenados em bancos de dados, o próximo passo foi disponibilizá-los na *web*, o que proporcionou um acesso fácil por parte da comunidade científica e interessados. Assim, os bancos de dados de genoma representam hoje uma ferramenta de suporte indispensável para biólogos e geneticistas. Nesta seção serão enumerados alguns dos principais bancos de dados do genoma.

3.3.1. GenBank

O GenBank (BENSON, 2002) é um banco de dados mantido pelo *National Institute of Health* (NIH), o qual armazena dados de seqüências genéticas e uma coleção de anotações relacionadas às seqüências de DNA disponíveis. É um banco de dados público, que tem sua atualização realizada por pesquisadores ao redor do mundo.

O Genbank mantém-se sincronizado com o *European Molecular Biology Laboratory* (EMBL, 2004) e o DNA *DataBank of Japan* (DDBJ, 2006). Desta forma, os pesquisadores podem consultar qualquer uma destas fontes para obter o mesmo conjunto de dados como resposta.

A submissão de dados no GenBank pode ser feita de duas formas:

- Banklt

Banklt é uma ferramenta para submissão de dados, tendo o seu uso recomendado para as submissões simples. Pode-se indicar regiões que codificam as proteínas no mRNA e o nome do gene. Genericamente, a ferramenta transforma os dados submetidos pelo usuário para o formato do GenBank. Além disso, o usuário tem a opção de adicionar anotações sobre a fonte da seqüência e as características biológicas da mesma utilizando suas próprias palavras, através de uma interface.

- Sequin

Sequin é uma ferramenta apropriada para submissões muito longas e complexas, ou para quando for necessário um controle sobre as anotações feitas. É um programa gráfico, que guia o usuário através de formulários com um controle de vocabulário a ser utilizado durante o processo de submissão de seqüências e o fornecimento de anotações biológicas e bibliográficas. Possui também uma funcionalidade que tem como objetivo detectar possíveis erros.

3.3.2. EMBL

Similar ao GenBank no que se refere ao armazenamento de seqüências, o EMBL é parte integrante do consórcio internacional que reúne o GenBank e o DDBJ (EMBL, 2004).

Os principais contribuintes do EMBL são autores e grupos individuais de projetos de pesquisa genômica, sendo que a troca de dados acontece diariamente entre os bancos de dados colaboradores.

Devido à sincronização com o GenBank e o DDBJ, o formato dos dados do EMBL possui compatibilidade com esses bancos e muitas ferramentas de análise são comuns. Para a análise das seqüências têm-se ferramentas de alinhamentos múltiplos, pesquisas por palavras-chave, identificação de padrões específicos etc.

3.3.3. Swiss-Prot

O SWISS-PROT é um banco de dados de proteínas que também possui integração com o EMBL, disponibilizando domínios, descrições de funções e estruturas de proteínas. Seu armazenamento iniciou-se em 1986 e vem sendo mantido de forma colaborativa desde 1987 pelo *Departament of Medical Biochemistry of the University of Geneva* e a EMBL *Data Library* (SWISS-PROT, 2006). Esforços são concentrados para manter um nível mínimo de redundância no Swiss-Prot.

Nesta base de dados, assim como na maioria das outras, os dados podem ser divididos em duas classes, a principal e a de anotações. A primeira diz respeito à seqüência em si, nas informações de citações, ou seja, referências bibliográficas e, por fim, nos dados taxonômicos que se referem à descrição da fonte biológica da proteína. A última consiste numa série de informações enumeradas a seguir:

- Funções da proteína;
- Domínios e sites: são agrupamentos de proteínas que compartilham funções ou são derivadas de um antecessor comum;
- Estrutura secundária: é a estrutura que diz respeito aos padrões regulares e repetitivos (BIOPROJECT, 2006);
- Estrutura quaternária: muitas proteínas são constituídas por mais de uma cadeia polipeptídica. A estrutura quaternária descreve a forma com que as diferentes subunidades se agrupam e se ajustam para formar a estrutura total da proteína (BIOPROJECT, 2006);
- Similaridades com outras proteínas;
- Doenças associadas com a deficiência dessa proteína;
- Conflitos de seqüência, variações etc.

3.3.4. PDB

O *Protein DataBank* (PDB) (PDB, 2006) é um banco de dados que armazena estruturas tridimensionais de macromoléculas biológicas. Contudo, com o passar dos anos e, conseqüentemente, com o avanço da tecnologia e mudanças na questão de compartilhamento de conhecimento na comunidade científica, fez com que o banco crescesse.

Atualmente, o PDB é o maior repositório do mundo para o processamento e distribuição de dados da estrutura tridimensional de grandes moléculas de proteínas e ácidos nucléicos. As estruturas tridimensionais de macromoléculas biológicas do PDB foram determinadas experimentalmente, sendo que a disposição espacial dos aminoácidos que formam as proteínas é relacionada à função das mesmas. A submissão dos dados ao PBD é realizada pela *web*, e os resultados de pesquisas realizadas neste banco de dados são facilmente acessíveis.

3.3.5. Enzyme

Enzyme é um repositório de dados que contém informações relacionadas à nomenclatura das enzimas. Além do número de identificação da enzima são armazenados outros dados como, por exemplo, nome recomendado, nomes alternativos, atividade catalítica, cofatores, ponteiros para as entradas no Swiss-Prot, ponteiros para o ProSite (SIGRIST, 2002), que descreve a família da proteína que a enzima pertence.

Enzyme é fortemente acoplado ao Swiss-Prot, estabelecendo grande benefício para ambos, pois se permite que correções e atualizações sejam propagadas eficientemente (ENZYME, 2006).

3.3.6. MBGD (*Microbial Genoma Database*)

O banco de dados MBGD é um sistema para a análise comparativa entre genomas microbianos que tenham sido completamente seqüenciados. A função central do MBGD é criar uma tabela usando a classificação de genes ortólogos, comparando todos contra todos, relacionando as similaridades entre genes em múltiplos genomas (UCHIYAMA, 2003).

A criação da tabela é feita utilizando-se de um algoritmo de classificação automatizado. Desta forma, cada usuário pode criar uma tabela de classificação própria, especificando os organismos e parâmetros. A tabela de classificação gerada é armazenada

num banco de dados, e pode ser explorada com a combinação de dados de genomas individuais para descobrir relações de similaridade entre genomas.

O usuário pode analisar os dados de vários pontos de vista, como a análise de padrões filogenéticos, a comparação da ordem dos genes e a comparação detalhada da estrutura dos genes.

3.3.7. Outros Bancos de Dados

KEGG (*Kyoto Encyclopedia of Genes and Genomes*) é um exemplo de banco de dados funcional, ou seja, que permite a análise das funções dos genes. É utilizado na comparação de genomas, e pode-se obter mapas metabólicos e identificar quais enzimas estão presumidamente codificadas por um determinado genoma. Possibilita encontrar também grupos de genes ortólogos e mapas genômicos (KANEHISA e GOTO, 2000).

COG *(Clusters of Orthologous Groups)* é um banco de dados de *clusters* de grupos de proteínas ortólogas, que provém do seqüenciamento de genomas de seres eucariotos e procariotos. Essa coleção de conjuntos de proteínas ortólogas consiste em uma plataforma útil na anotação de genomas seqüenciados recentemente e, em estudos evolucionários (TATUSOV *et al.*, 2003).

Por fim, um banco de dados que merece destaque é o UniProt *(Universal Protein Resource)*(APWEILER, 2004). Trata-se de um repositório central de seqüências de proteínas e suas funções, criado através da junção de informações de três bancos de dados protéicos: Swiss-Prot, TrEMBL, and PIR (PIR, 2006).

3.4. Alinhamento de Seqüências

Em algumas ferramentas de comparação de seqüências biológicas como BLAST (*Basic Local Alignment Search Tool*)(seção 3.5.2) e FASTA (seção 3.5.1), descritos adiante neste capítulo, faz-se uso de alinhamentos de seqüências para a análise de similaridades.

Um alinhamento é uma forma de se comparar seqüências biológicas (DNA ou proteína). Ao se estabelecer um alinhamento entre essas seqüências, é possível descobrir se elas estão evolutivamente relacionadas ou não.

Uma seqüência biológica é representada através de uma cadeia de caracteres escrita com um determinado alfabeto[7]. Assim, comparar duas seqüências biológicas é equivalente a comparar duas cadeias de caracteres. Exemplo:

ACCTATGCAC
ACCATGCAC

Neste exemplo as seqüências possuem tamanhos diferentes. Por isso, é preciso igualar os tamanhos usando um caractere nulo "-" conhecido como *gap* ou espaço. Com a inserção do caractere nulo obtém-se:

A C C T A T G C A C
A C C - A T G C A C

A seguir, alinhamento é formalmente definido, de acordo com Almeida (2002).

3.1 Definição de Alinhamento

Dadas as seqüências $s = s_1 \ldots s_m$ e $t = t_1 \ldots t_n$, com símbolos pertencentes ao alfabeto Σ, e com $m, n \geq 0$, um alinhamento de s e t é um mapeamento de s e t nas seqüências s' e t', respectivamente, cujos símbolos pertencem ao alfabeto $\Sigma' = \Sigma \cup \{ - \}$, onde o símbolo '-' é chamado de espaço, tal que:

1. $|s'| = |t'| = l$;
2. a remoção dos espaços de s' e t' leva a s e t, respectivamente; e
3. não é permitida a condição $s_i' = - = t_i'$, $1 \leq i \leq l$.

3.4.1. Esquemas de Pontuação

No alinhamento entre duas seqüências, o pareamento (de bases no caso de DNA e de aminoácidos no caso de proteínas) é estabelecido com base em um esquema de pontuação, onde se procura alinhar a seqüência com o objetivo de obter a pontuação mais alta para medidas de similaridade.

[7] Uma seqüência de DNA é escrita com um alfabeto de quatro letras, referentes às bases nitrogenadas. Uma seqüência de proteína é escrita com um alfabeto de vinte letras, sendo que cada letra corresponde a um dos vinte diferentes aminoácidos existentes na natureza.

Considerando as seqüências $s = s_1 \ldots s_m$ e $t = t_1 \ldots t_n$, com símbolos pertencentes ao alfabeto Σ, um esquema de pontuação é dado por uma tupla (p, g) onde a função $p : \Sigma x \Sigma \rightarrow \mathbb{R}$ determina a pontuação de cada par de caracteres alinhados e g é usado para penalizar *gaps* ($g < 0$ na maioria das vezes). O esquema de pontuação fornece um valor numérico a cada alinhamento possível. Sendo (s', t') o alinhamento das seqüências s e t, se adiciona $p(a, b)$ cada vez que um caractere a de s' está pareado com um caractere b de t'. Quando um caractere de s' ou t' está alinhado a um *gap*, g é agregado à pontuação. A pontuação total denotada como *score* (s', t') é a soma de todas as pontuações em todas as posições do alinhamento (s', t').

Sendo ψ o conjunto dos possíveis alinhamentos, a semelhança entre s e t é dada por:

$$sem(s,t) = \max_{(s',t') \in \psi} score(s',t')$$

Para uma seqüência de DNA, a pontuação de um alinhamento é ditada por uma matriz de substituição de base. A matriz apresentada na Figura 3.1(a) se preocupa apenas com os *matches* (equivalência entre duas bases de seqüências diferentes), enquanto que a matriz apresentada Figura 3.1(b) leva em consideração *matches* e *mismatches* (não equivalência entre duas bases de seqüências diferentes), além de uma penalidade para *gaps*.

	A	**G**	**C**	**T**
A	1	0	0	0
G	0	1	0	0
C	0	0	1	0
T	0	0	0	1

(a)

	A	**G**	**C**	**T**	**-**
A	1	-1	-1	-1	-2
G	-1	1	-1	-1	-2
C	-1	-1	1	-1	-2
T	-1	-1	-1	1	-2

(b)

Figura 3.1 – Matrizes de substituição de base (a) Matriz de Substituição de Bases com *matches*; (b) Matriz de Substituição de Bases com *matches, mismatches* e penalidade para *gaps*.

Um exemplo de utilização da matriz da Figura 3.1(b) é apresentado a seguir na Figura 3.2.

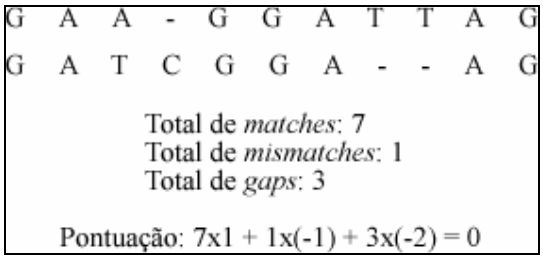


Figura 3.2 – Exemplo do cálculo da pontuação de um alinhamento através da utilização de uma matriz de substituição de base.

Na comparação de seqüências de DNA, as matrizes de substituição são relativamente simples, sendo que o mesmo não ocorre quando se deseja comparar seqüências de proteínas. A comparação destas seqüências leva em conta critérios evolutivos, por isso, é preciso um esquema de pontuação mais elaborado. As matrizes de substituição PAM (*Point Accepted Mutation*) e BLOSUM (*Blocks Substitution Matrix*) são muito utilizadas para a comparação de proteínas.

As matrizes PAM (DAYHOFF, SCHWARTZ e ORCUTT, 1978) estabelecem esquemas de pontuação para grupos de seqüências relacionadas. Foram construídas observando-se substituições de aminoácidos através de alinhamentos. Para isso utilizou-se de um amplo conjunto de proteínas relacionadas, as quais tinham sofrido uma certa divergência evolucionária.

Cada matriz PAM é determinada por um número que indica o grau de divergência entre as seqüências usadas. Dessa forma, uma matriz PAM1 é uma unidade de divergência evolutiva na qual ocorreu 1% de substituição dos aminoácidos.

As matrizes BLOSUM (HENIKOFF e HENIKOFF, 1991) foram construídas através da extração de segmentos sem *gaps*, chamados blocos, de múltiplos alinhamentos de famílias de proteínas, envolvendo seqüências com pouca relação (diferentemente das matrizes PAM, nas quais foram usadas seqüências relacionadas), e então agrupados com base na porcentagem de identidade. Da mesma forma como nas matrizes PAM, as matrizes BLOSUM são seguidas de números que se referem ao nível máximo de identidade entre as seqüências.

3.4.2. Tipos de Alinhamento

Existem dois tipos de alinhamento, o global e o local. O primeiro se estende por toda seqüência, já o segundo localiza fragmentos de seqüências que são mais similares. A definição de ambos os alinhamentos é apresentada a seguir:

3.2 Definição de Alinhamento Global

Dado um alfabeto A com uma matriz de substituição M, o alinhamento global para duas seqüências $s=\{s_1s_2s_3...s_m \mid s_i \in A\}$ e $t=\{t_1t_2t_3...t_n \mid t_j \in A\}$ sendo $1 \leq i \leq m$ e $1 \leq j \leq n$, consiste em encontrar cadeias de caracteres α e β, as quais são obtidas de s e t inserindo espaços no início ou no final de s e t, e cuja pontuação calculada usando M é máxima (GUSFIELD, 1997) *apud* (XU *et al.*, 2003).

3.3 Definição de Alinhamento Local

Dado um alfabeto A com uma matriz de substituição M, o alinhamento local para duas seqüências $s=\{s_1s_2s_3...s_m \mid s_i \in A\}$ e $t=\{t_1t_2t_3...t_n \mid t_j \in A\}$ sendo $1 \leq i \leq m$ e $1 \leq j \leq n$, consiste em encontrar subcadeias de caracteres de s e t, cujo valor da similaridade (alinhamento global ótimo) é máximo (GUSFIELD, 1997) *apud* (XU *et al.*, 2003).

3.5. Ferramentas de Bioinformática

O grande volume de dados gerados pelos diversos projetos de seqüenciamento de genomas torna necessária a utilização de ferramentas para análise computacional destas informações. As ferramentas de bioinformática são os programas de software projetados para extrair informações significativas da grande massa de dados biológicos.

A existência de uma grande quantidade de ferramentas disponíveis para análise genômica mostra a necessidade de se atender a diferentes objetivos e a dificuldade que se tem em englobar várias funcionalidades em uma única ferramenta. É nesse sentido que esforços são gastos continuamente, pensando em atender cada vez mais e melhor a necessidade de biólogos e afins, desenvolvendo novas ferramentas e integrando as já existentes.

3.5.1. FASTA

FASTA (PEARSON, 1985; PEARSON e LIPMAN, 1988) foi o primeiro programa utilizado em larga escala para busca de similaridades. É um método heurístico, utilizado para realizar alinhamentos locais de seqüências.

O algoritmo utilizado por FASTA pesquisa primeiro por alinhamentos exatos e sem inserção de *gaps* entre subseqüências de tamanho *k*, k-tuplas, comuns à seqüência de consultada e à seqüência do banco de dados de seqüências.

Após identificar as k-tuplas comuns, o programa constrói regiões. Uma região é formada por uma ou mais *k*-tuplas. Posteriormente, as regiões recebem uma pontuação conforme uma matriz de substituição utilizada, como BLOSUM ou PAM. Este processo é repetido para cada seqüência do banco de dados.

Finamente, as regiões de maior pontuação são utilizadas para produzir um alinhamento usando um algoritmo de programação dinâmica restrito a uma banda ao redor das regiões.

Existem diferentes programas que utilizam FASTA. Abaixo segue uma descrição de cada um deles:

- **Fasta3** - Faz uma varredura numa biblioteca de proteínas ou de seqüências de DNA por seqüências similares;
- **Fastx/y3** - Compara uma seqüência de DNA com um banco de dados de seqüências de proteínas, comparando a seqüência traduzida do DNA em *frames* para diante e trás;
- **Tfastx/y3** - Compara uma proteína com um banco de dados de DNA traduzido;
- **Fasts3** - Compara fragmentos de peptídeos com um banco de dados de proteínas ou DNA;
- **Fastf3** - Compara uma mistura de peptídeos ordenados com um banco de dados de proteínas ou DNA.

3.5.2. BLAST (*Basic Local Alignment Search Tool*)

BLAST é uma ferramenta de comparação e alinhamento local de seqüências de nucleotídeos ou de aminoácidos depositadas em bancos de dados. É de muita importância para quem trabalha com bioinformática, pois possibilita realizar análises de similaridades

de seqüências como DNA, RNA e proteínas. Como utiliza métodos heurísticos, consegue resultados significativos em um tempo satisfatório.

Essencialmente, a partir de uma seqüência de consulta (*query)* introduzida pelo usuário, BLAST tenta achar todas as seqüências em bancos (*subject*) que possuem alinhamentos locais estatisticamente significativos. O usuário pode especificar também um limiar para o alinhamento, denominado *score*.

As seqüências similares retornadas de uma pesquisa são conhecidas como *hits* e são acompanhados de alinhamentos, juntamente com o *score*, e de uma estimativa de significância, denominada de *e-value*. O *e-value* é proporcional à probabilidade de um *hit* com o seu alinhamento ser encontrado ao acaso. Assim, quanto menor o *e-value*, mais significante é o *hit*.

A estratégia usada no BLAST para a determinação dos *hits* é a busca de "sementes", que consistem em pares de seqüências muito curtas entre as seqüências em estudo, as quais são estendidas em ambos os lados. A extensão prossegue até o alcance dos escores máximos. Nem todas as extensões são investigadas porque o programa compara os escores destas extensões com um limiar cuidadosamente escolhido. Assim, extensões significativas podem não ser localizadas. Contudo, é uma margem de erro aceitável.

De acordo com o tipo de seqüência de entrada (nucleotídeo ou aminoácido) e com o tipo de resultado esperado, existe um programa BLAST específico:

- **BLASTP:** Compara uma seqüência *query* de aminoácidos contra um banco de dados de seqüências de proteínas;
- **BLASTN:** Compara uma seqüência *query* de nucleotídeos contra um banco de dados de seqüências de nucleotídeos;
- **BLASTX:** Compara uma seqüência *query* de nucleotídeos contra um banco de dados de seqüências de proteínas;
- **TBLASTN:** Traduz uma seqüência de aminoácidos para nucleotídeo e compara com o banco de dados de genes;
- **TBLASTX:** Traduz uma seqüência de nucleotídeo para aminoácidos e compara com o banco de proteínas.

3.5.3. EGG (*Extended Genome-Genome Comparison*)

Basicamente, EGG (ALMEIDA, 2002) compara o conjunto de genes de dois proteomas utilizando resultados do BLAST (seção 3.5.2), com o propósito de identificar

relacionamentos entre esses genes. Para realizar a comparação entre os proteomas, EGG utiliza-se do programa BLASTP, que pertence à família BLAST, um dos programas mais populares de busca em bases de dados biológicos.

A ortologia entre dois genes, também chamada de *match*, é definida por EGG seguindo alguns parâmetros. Por exemplo, um *match* entre os genes g_i e h_i ocorre se g_i encontrou h_i como *hit*, sendo o *e-value* encontrado menor que 10^{-5} e um alinhamento de no mínimo 60%, e vice-versa. Esses valores são sugeridos pelo EGG, mas podem ser alterados pelo usuário.

Com o propósito de encontrar pares de genes ortólogos, EGG compara cada gene g_i de um genoma G com todos os genes de um genoma H, sendo a próxima etapa a comparação de todos os genes h_j de H com todos os genes de G.

As ortologias resultantes da comparação de dois genomas G e H que possuem o número de genes igual a m e n respectivamente, são armazenadas em uma matriz binária A_{mxn}, tal que $A_{i,j} = 1$ se, e somente se, g_i e h_i formam um *match*.

O passo que segue após a definição da matriz de *match*es é investigá-la para encontrar os *runs*, as regiões ortólogas (ROs) e a espinha dorsal de dois proteomas. Os *runs* (definição 3.4) são a base para se encontrar as outras estruturas e consiste basicamente de uma seqüência contígua de *matches*. A existência de muitos *runs* caracteriza um alto grau de relacionamentos e, portanto, alta similaridade.

3.4 Definição de *Runs*

Sejam dois genomas G e H. Para uma RGC α de G formada pelos genes $g_i, \ldots, g_k$ e uma RGC β de H formada pelos genes $h_j, \ldots, h_l$, tais que $k - i + 1 = l - j + 1$, $k > i$, e $l > j$, dizemos que α e β formam um *run* se uma das seqüências de pares de genes ortólogos acontece:

- $(g_i, h_j), (g_{i+1}, h_{j+1}), \ldots, (g_k, h_l)$; ou
- $(g_i, h_l), (g_{i+1}, h_{l-1}), \ldots, (g_k, h_j)$.

A primeira opção caracteriza um *run* paralelo e a segunda um *run* anti-paralelo.

Para cada *run* encontrado, EGG gera automaticamente um código. Os primeiros quatro símbolos do código identificam o par de genomas comparados, logo em seguida tem-se o ano, o mês e o dia (dois dígitos para cada um) e um número seqüencial do *run*,

naquela comparação genômica. Finalmente aparece a descrição se o *run* é paralelo ou anti-paralelo. Um exemplo de um *run* é ilustrado na Figura 3.3.

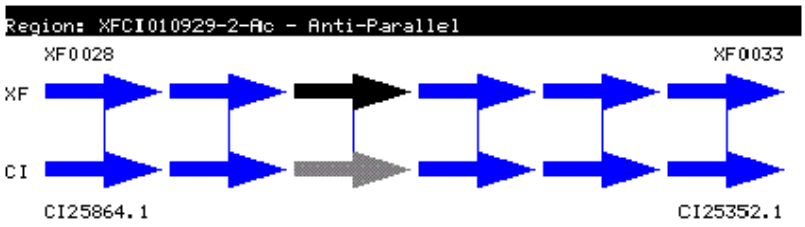


Figura 3.3 – Exemplo de um *Run* anti-paralelo (ALMEIDA, 2002)

Com a definição de *runs*, é possível definir o conceito de região ortóloga.

3.5 Definição de Região Ortóloga

Segundo Almeida (2002), uma região ortóloga *R* consiste em um *run* isolado com pelo menos *M* pares de ortólogos, ou uma união de *run*s com um total de pelo menos *M* pares de ortólogos cuja distância entre os genes extremos do *run*s não ultrapasse um valor fixo *k* em números de genes.

Em uma região ortóloga, a união de *runs* faz-se necessária devido à possível existência de inserções, remoções ou substituições de genes ao longo do tempo, e que pode acarretar em "buracos". O número mínimo *M* de pares de ortólogos reduz a possibilidade de *run* ter sido encontrado ao acaso.

Como ilustra a Figura 3.4, uma região ortóloga pode ser representada com o auxílio de um grafo bipartido[8], sendo os genes representados pelos pontos e as arestas do grafo representando a ortologia entre os genes.

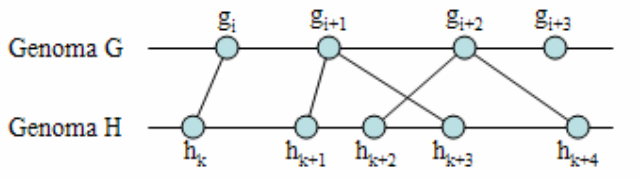


Figura 3.4 – Região Ortóloga

[8] Um grafo $G = (V,E)$, onde V é o conjunto de vértices e E o conjunto de arestas, é bipartido se V pode ser particionado em dois conjuntos X e Y tais que qualquer aresta (u, v) é tal que: ou $u \in X$ e $v \in Y$, ou $u \in Y$ e $v \in X$.

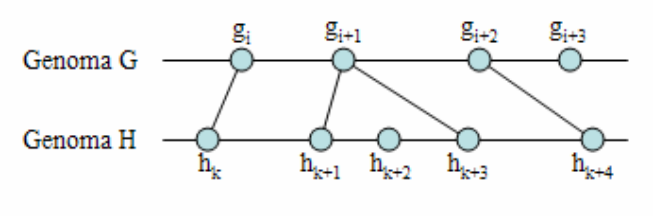


Figura 3.5 – Espinha Dorsal

A espinha dorsal de dois proteomas considera apenas genes fortemente ortólogos. Para isso, EGG usa o conceito de BBH (*Bidirectional Best Hit*). Dois genes formam um BBH se um encontrou o outro como melhor *hit*, menor *e-value*, e vice-versa. Com isso, a interferência de genes parálogos é reduzida, e o objetivo consiste em encontrar o maior alinhamento possível sem que haja cruzamentos, como ilustra a Figura 3.5.

É possível visualizar as ROs encontradas por EGG através de arquivos texto, como ilustra a Figura 3.6, referente a uma RO encontrada na comparação de *Leifsonia xyli subsp. xyli* e *Corynebacterium glutamicum*, contendo 5 matches. Para cada RO encontrada, EGG disponibiliza um código, a quantidade de *matches*, o número aproximado de bases em cada genoma na região e quais genes participam dela. Sobre os genes são informados o identificador no genoma, o identificador (gi), o tamanho e o produto. Por fim, *matches* da RO. A respeito do código, os primeiros quatro símbolos identificam o par de genomas comparados, logo em seguida tem-se o ano, o mês e o dia (dois dígitos para cada um) e um número seqüencial da RO, naquela comparação genômica. Finalmente aparece o símbolo R, de região, seguido de c (consistente) ou i (inconsistente).

```
>LECG030825-1-Rc
5 matches
7kb in le - 7kb in cg
================================================
Gene (le)      gi        size  product
================================================
+LE10.1      1011       472aa  chromosomal replication initiator protein
+LE20.1      2011       380aa  DNA polymerase III, beta subunit
+LE30.1      3011       294aa  6-phosphogluconate dehydrogenase
+LE40.1      4011       384aa  DNA replication and repair protein RecF
+LE50.1      5011       157aa  conserved hypothetical protein
+LE60.1      6011       662aa  DNA gyrase subunit B
================================================
Gene (cg)      gi        size  product
================================================
+Cg10001     19551251   524aa  COG0593:ATPase involved in DNA replication initiation
+Cg10002     19551252   394aa  COG0592:DNA polymerase sliding clamp subunit (PCNA homolog)
+Cg10003     19551253   394aa  COG1195:Recombinational DNA repair ATPase (RecF pathway)
+Cg10004     19551254   178aa  conserved hypothetical protein, predicted by GeneMark hmm
+Cg10005     19551255   684aa  COG0187:DNA gyrase (topoisomerase II) B subunit

======
matches
======
================================================================================
Gene           start size evalue[     best hit     ] product
================================================================================
+LE60.1         5914 662      0 [best              ] DNA gyrase subunit B
+Cg10005        5435 684      0 [best              ] COG0187:DNA gyrase (topoisomerase II) B subunit
----------------------------------------
+LE50.1         5314 157  2e-20 [best              ] conserved hypothetical protein
+Cg10004        4766 178  2e-20 [best              ] conserved hypothetical protein, predicted by GeneMark
hmm
----------------------------------------
+LE40.1         4160 384  5e-82 [best              ] DNA replication and repair protein RecF
+Cg10003        3585 394  4e-82 [best              ] COG1195:Recombinational DNA repair ATPase (RecF pathway)
----------------------------------------
+LE20.1         2083 380  5e-61 [best              ] DNA polymerase III, beta subunit
+Cg10002        2292 394  4e-61 [best              ] COG0592:DNA polymerase sliding clamp subunit (PCNA
homolog)
----------------------------------------
+LE10.1          225 472 1e-125 [best              ] chromosomal replication initiator protein
+Cg10001            1 524 1e-125 [best              ] COG0593:ATPase involved in DNA replication initiation
```

Figura 3.6 - RO encontrada na comparação *Leifsonia xyli subsp. xyli* e *Corynebacterium glutamicum*

A ferramenta EGG é muito útil no estudo de relacionamentos entre genes. No entanto, limita-se a encontrar semelhanças entre apenas dois proteomas. Assim, utilizando-se das comparações dois a dois realizadas por EGG surgiu a ferramenta BAGRE (seção 3.5.4), que realiza a comparação entre múltiplos proteomas, com o objetivo de se encontrar regiões de genes que se conservam entre esses organismos.

3.5.4. Bagre

A ferramenta BAGRE compara vários genomas com o propósito de encontrar regiões que se mantêm conservadas, ou seja, Regiões Ortólogas Múltiplas (ROMs). A definição de ROMs segundo Montera (2004) será apresentada a seguir.

3.6 Definição de Região Ortóloga Múltipla

Uma região ortóloga múltipla *R* é um conjunto de RGCs de dois ou mais proteomas distintos, tais que qualquer par de RGCs deste conjunto forma uma região ortóloga.

O processo utilizado por BAGRE para obter ROMs inicia-se com a realização de comparações dois-a-dois entre todos os pares possíveis de proteomas, utilizando a ferramenta EGG (seção 3.5.3) para realizar tais comparações. A etapa seguinte consiste na construção de um grafo, com base nas comparações realizadas anteriormente. O problema é modelado considerando cada vértice do grafo uma RGC que tenha participado de uma RO. Assim, uma aresta no grafo corresponde à ligação de duas RGCs, ou seja, uma RO.

No momento da criação dos vértices do grafo, é preciso atentar para o fato de que pode haver sobreposição entre duas RGCs distintas de um mesmo genoma. Neste caso é preciso decidir se estas duas RGCs serão representadas por um ou dois vértices no grafo.

A sobreposição é tratada na ferramenta BAGRE considerando-se o tamanho da menor RGC, ou seja, caso a sobreposição seja maior ou igual a P% do tamanho da menor RGC, tem-se que as duas RGCs serão representadas por um único vértice no grafo, caso contrário, um vértice será criado para cada uma das RGC. Sendo que P é um parâmetro definido pelo usuário.

Com o grafo construído, a estratégia para obtenção de ROs existentes entre os diversos genomas consiste na obtenção de cliques maximais (clique com maior número de vértices em um grafo) presentes neste grafo. Assim, uma clique de tamanho *k* representa uma ROM entre *k* genomas envolvidos na comparação.

Encontrar clique máxima em um grafo é um problema NP-Difícil. Para solucionar o problema optou-se pela implementação de uma heurística, mesmo não se garantindo que todas as cliques maximais sejam encontradas (MONTERA, 2004). O resultado é apresentado através da construção de uma alinhamento múltiplo entre RGCs pertencentes a ROM, sendo todo o processo apresentado num sistema via *web*.

3.5.5. Outras Ferramentas

A coleção de ferramentas destinadas a comparações de genomas é extensa. Constantemente novas ferramentas são construídas ou aprimoradas, sempre se preocupando em atingir um objetivo em especial. As ferramentas criadas sempre visam realizar uma

análise cada vez mais minuciosa dos dados provenientes de seqüenciamentos de seqüências de DNA e proteínas.

GenomeComp é uma ferramenta de comparação genômica que utiliza a saída textual do BLAST. Com a ferramenta pode-se visualizar variações entre genomas, como regiões repetidas, inserções, remoções e reagrupamentos de segmentos de genomas entre espécies ou entre culturas de um determinado microorganismo (YANG *et al.*, 2003).

GenomeBlast é uma ferramenta que foi desenvolvida para realizar a análise comparativa múltipla de pequenos genomas. É utilizado o programa BLASTP para comparação de seqüências similares e o método *maximum parsymony* baseado no conteúdo dos genes para inferir na filogenia dos genomas. Dentre outros objetivos desta ferramenta, é possível encontrar: genes homólogos candidatos entre genomas comparados; uma tabela informativa de presença ou ausência de genes; e uma filogenia de genomas.

Outra ferramenta para análise múltipla de genomas denomina-se M-GCAT (TREANGEN e MESSEGUER, 2005). Constitui-se de um pacote de comparação de genomas que pode analisar, agrupar e alinhar múltiplos genomas, e em muitos casos produz alinhamentos consistentes com BLAST.

Para realizar comparações entre genomas de seres eucariotos, pode-se utilizar uma ferramenta denominada Procom (*Proteome Comparison*) (LI, 2005). É uma ferramenta destinada a encontrar genes que desempenham funções específicas em um organismo. Através de comparações de genomas, genes que ocorrem em organismos sem as características de interesse, são removidos da coleção de genes em comum entre organismos com as características. Esta comparação é útil quando genes codificam proteínas associadas a uma característica especifica somente a uma coleção de organismos.

3.6. Considerações Finais

Neste capítulo, inicialmente explicou-se todo o processo que envolve o seqüenciamento de genomas. A seguir, foram enumerados os principais bancos de dados que armazenam informações provenientes de seqüenciamentos. As principais características desses bancos foram destacadas e quais tipos de dados (proteínas, seqüências, enzimas, etc) são armazenados em cada um.

Foram enumeradas também ferramentas que envolvem comparações de genomas. O objetivo de cada ferramenta foi colocado de forma simples, abordando as principais características e peculiaridades de cada uma.

O próximo capítulo tem o objetivo de apresentar a principal contribuição deste trabalho. Trata-se de uma abordagem para encontrar regiões ortólogas em múltiplos proteomas. A modelagem do problema é semelhante a da ferramenta BAGRE, no entanto, utiliza-se de um grafo colorido, o que possibilitou a construção de algoritmos mais precisos, garantindo um conjunto de respostas confiável e completo.

CAPÍTULO 4

4. MPC (*Multiple Proteome Comparison*)

4.1. Considerações Iniciais

O objetivo deste trabalho, como já foi descrito anteriormente, volta-se para a extração de características funcionais e evolutivas, através da comparação de proteomas de seres procariotos, com o objetivo de encontrar regiões em múltiplos genomas, em termos de seus genes, que conservam a ordem e o conteúdo gênico.

Como pôde ser visto no Capítulo 3, existem várias ferramentas para analisar dados provenientes de seqüenciamentos de genomas. A preocupação em se construir a ferramenta MPC deve-se à busca de melhorias no resultado com relação às regiões encontradas e também à diminuição de tempo de processamento.

A seguir são apresentadas algumas definições sobre conceitos utilizados na implementação do projeto.

4.1 Definição de Grafo

Um grafo $G(V, E)$ é definido pelo par de conjuntos V e E, onde V é um conjunto não vazio de vértices ou nodos do grafo; e E é um conjunto de pares ordenados $a = (v,w)$, v e $w \in V$, tal que v e w são arestas do grafo.

4.2 Definição de Grafo Colorido

Uma coloração para um grafo $G = (V,E)$ é uma função sobrejetora $\varphi : V \rightarrow \{1, ..., k\}$, na qual o conjunto imagem constitui-se das cores utilizadas no grafo. Se um grafo G está associado a uma coloração φ, diz-se que este é um grafo colorido, e denota-se por G_φ. Dois vértices u e $v \in G_\varphi$ têm a mesma cor se $\varphi(u) = \varphi(v)$. Uma classe de coloração $c(j)$ de G_φ é o conjunto constituído por todos aqueles vértices cuja cor é j. Formalmente, $c(j) = \{v \in V \mid \varphi(v) = j\}$.

4.3 Definição de Clique

Uma clique de um grafo $G = (V, E)$ é constituído do subconjunto $C \subseteq V$, tal que para todo par (v, w) de vértices distintos em C, existe a aresta $(v,w) \in E$. Uma clique C é

maximal quando o acréscimo de qualquer vértice à *C* faz com que *C* deixe de ser clique. Uma clique *C* é máxima se não existir clique no grafo contendo mais vértices que em *C*.

Na seção 3.5.3, o conceito de região ortóloga foi definido em termos de *runs*. A seguir o mesmo é redefinido para uma melhor interpretação nas próximas seções.

4.4 Definição de Região Ortóloga (ALMEIDA, 2002)

Em uma comparação entre dois proteomas *G* e *H*, uma região ortóloga (RO) é um par (α,β) tal que:

- α é uma RGC em *G*;
- β é uma RGC em *H*;
- α e β são descendentes de uma mesma região ancestral; e
- α e β contêm aproximadamente o mesmo número de genes.

Nas seções seguintes será apresentado todo o processo seguido para a obtenção de ROMs, bem como a ferramenta MPC que o engloba.

4.2. Obtenção de Regiões Ortólogas Múltiplas

O conceito de ROMs utilizado pela ferramenta MPC é o mesmo da definição **3.6**. O objetivo desta seção é apresentar a abordagem utilizada pela ferramenta MPC na obtenção de ROMs. A Figura 4.1 ilustra uma visão geral do processo.

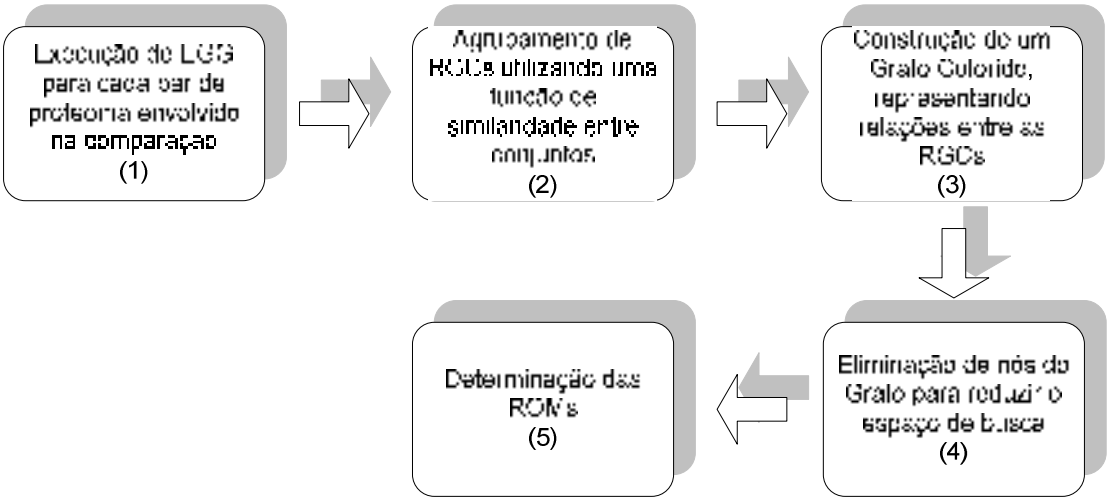


Figura 4.1 – Visão Geral do processo para Obtenção de Regiões Ortólogas Múltiplas

Passo 1

A ferramenta MPC é alimentada com os dados oriundos do EGG (seção 3.5.3). Como EGG realiza comparações de dois proteomas apenas, é preciso executá-lo diversas vezes. O número de comparações é obtido utilizando-se de combinação simples, onde *n* é o número de proteomas e *k* o número de elementos dos grupos.

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Como a comparação é sempre feita dois a dois, a fórmula para encontrar o número de comparações é:

$$\binom{n}{2} = \frac{n!}{2!(n-2)!}$$

Ao final desta etapa, tem-se um conjunto de arquivos referentes à similaridade dois a dois dos proteomas envolvidos. Dos arquivos gerados pelo EGG, é utilizado aquele que mostra as regiões ortólogas encontradas; um exemplo desse arquivo é ilustrado na Figura 3.6.

Passo 2

Nesse passo, o objetivo é explicar como RGCs similares de um mesmo proteoma são agrupadas. Além disso, discute-se a importância do agrupamento de RGCs para a determinação de ROMs.

A Figura 4.2 ilustra parte do resultado de comparações de proteomas realizadas utilizando o EGG. Na figura, cada elipse representa um gene de um proteoma, sendo que genes de um mesmo proteoma estão sobre uma mesma linha vertical. O número dentro de cada elipse é a identificação do gene, fornecido pelo BLAST, e a linha ligando duas elipses de proteomas diferentes representa uma ortologia entre dois genes.

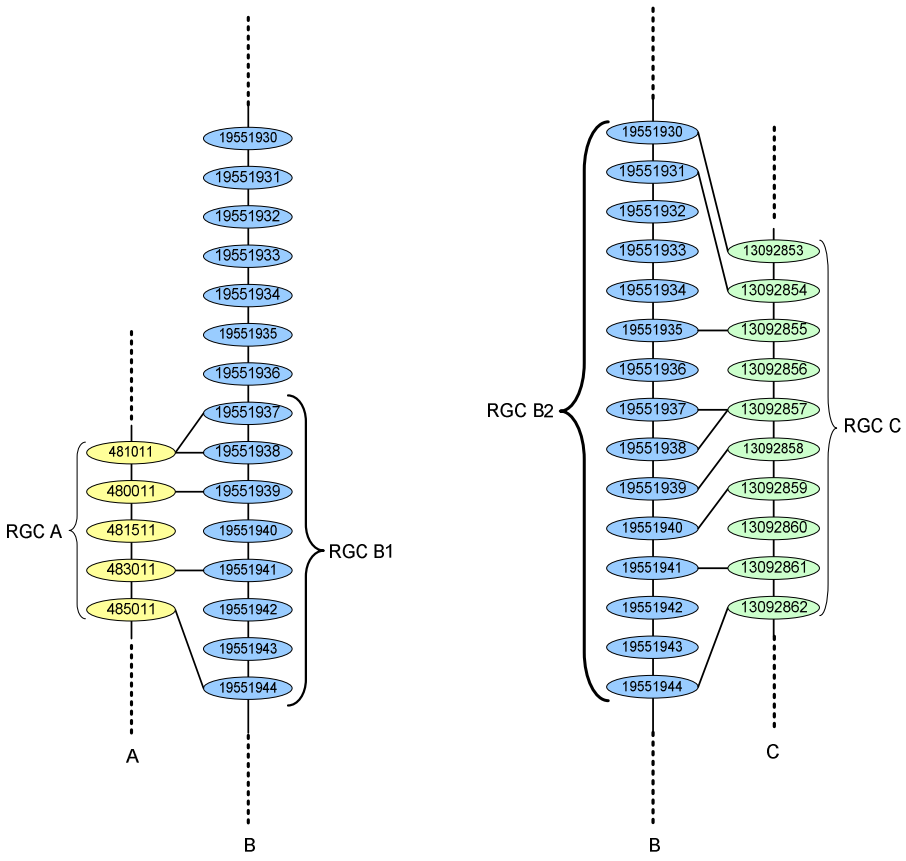


Figura 4.2 – Obtenção de ROs: (a) RO encontrada através da comparação das bactérias *Leifsonia xyli subsp. Xyli* e *Corynebacterium glutamicum;* (b) RO encontrada através da comparação das bactérias *Corynebacterium glutamicum* e *Mycobacterium leprea strain TN.*

Pode-se perceber na Figura 4.2 a presença de quatro RGCs e de duas ROs. A RO formada pelas RGCs *A = {481011, 480011, 481511, 483011, 485011}* e *B1 = {19551937, 19551938, 19551939, 19551940, 19551941, 19551942, 19551943, 19551944}*, ilustradas na parte *(a)* da figura, são resultantes da comparação entre as bactérias *Leifsonia xyli subsp. Xyli* e *Corynebacterium glutamicum*. Na parte *(b)* da figura, ilustra-se uma RO formada pelas RGCs *B2={19551930, 19551931, 19551932, 19551933, 19551934, 19551935, 19551936, 19551937, 19551938, 19551939, 19551940, 19551941, 19551942, 19551943, 19551944}* e *C={13092853, 13092854, 13092855, 13092856, 13092857, 13092858, 13092859, 13092860, 13092861, 13092862}*, resultantes da comparação entre as bactérias *Corynebacterium glutamicum* e *Mycobacterium leprea strain TN*.

Alguns dos genes das RGCs *B1* e *B2* são iguais, o que caracteriza uma sobreposição de RGCs. A mecânica utilizada para decidir se estas RGCs devem ser agrupadas ou não, é baseada na similaridade entre elas, e será discutida logo a seguir.

Para medir a similaridade entre duas RGCs, utilizou-se uma função que mede a similaridade entre conjuntos. Sendo RGC_a e RGC_b as RGCs que se deseja estimar a similaridade, a similaridade *S* é medida da seguinte forma:

$$S(RGCa, RGCb) = \frac{| RGCa \cap RGCb |}{| RGCa \cup RGCb |}$$

A decisão de se agrupar RGCs toma como base a similaridade *S* e uma similaridade mínima *m*, definida pelo usuário, sendo $0 < m \leq 1$. Logo, duas RGCs *A* e *B* são agrupadas caso $S(A,B) \geq m$. Como mostra o exemplo logo abaixo.

Exemplo 4.1: Deseja-se agrupar RGCs similares, tendo como base as comparações duas a duas das bactérias *Leifsonia xyli subsp. Xyli*, *Corynebacterium glutamicum* e *Mycobacterium leprea strain TN,* considerando apenas as RGCs da Figura 4.2.

Na Figura 4.2, as únicas RGCs que possuem genes em comum são as RGCs *B1* e *B2*. A similaridade entre estas RGCs é dada da seguinte forma:

$$S(RGCB1, RGCB2) = \frac{| RGCB1 \cap RGCB2 |}{| RGCB1 \cup RGCB2 |} = \frac{8}{15} = 0.53$$

Assim, de acordo com a decisão de agrupar RGCs similares, caso $0.53 \geq m$, as RGCs serão agrupadas.

Para que ocorram agrupamentos de RGCs, caso seja preciso, todas as RGCs de um mesmo proteoma precisam ser comparadas entre si, a fim de se descobrir similaridades. Esta tarefa é a que consome mais tempo na determinação de ROMs, como pode ser visualizado no Apêndice A. Isto acontece devido à elevada quantidade de RGCs de um proteoma, que faz com que o número de comparações seja elevado.

Na Figura 4.3 a seguir, tem-se a descrição do algoritmo, criado neste trabalho, para agrupar RGCs similares de um proteoma *P*, considerando uma similaridade mínima para

agrupamentos de RGCs *m*. A similaridade entre duas RGCs é conseguida pela função getSimilarity.

Início Algoritmo

1. **para** cada Rgc[i] do proteoma *P* **faça**

   2. **para** cada Rgc[j] do proteoma *P*, sendo j>i **faça**

      Se getSimilarity(Rgc[i],Rgc[j]) ≥ *m* **faça**
      Criar uma nova Rgc através da função:
      joinRgcs(Rgc[i],Rgc[j])

3. Eliminar da coleção de Rgcs do Proteoma *P*, as Rgcs que foram agrupadas, através da função:
   removeRgc(Rgc)

4. Adicionar na coleção de Rgcs do Proteoma *P*, as Rgcs que foram criadas através de agrupamentos, usando a função:
   addRgc(Rgc)

Fim Algoritmo

Figura 4.3 – Algoritmo para agrupar RGCs similares de um proteoma *P*

Quanto maior a similaridade mínima *m* para agrupar RGCs definida pelo usuário, menor é o número de agrupamentos e, conseqüentemente, maior é a quantidade de RGCs de cada proteoma, como pode-se perceber na Figura 4.4.

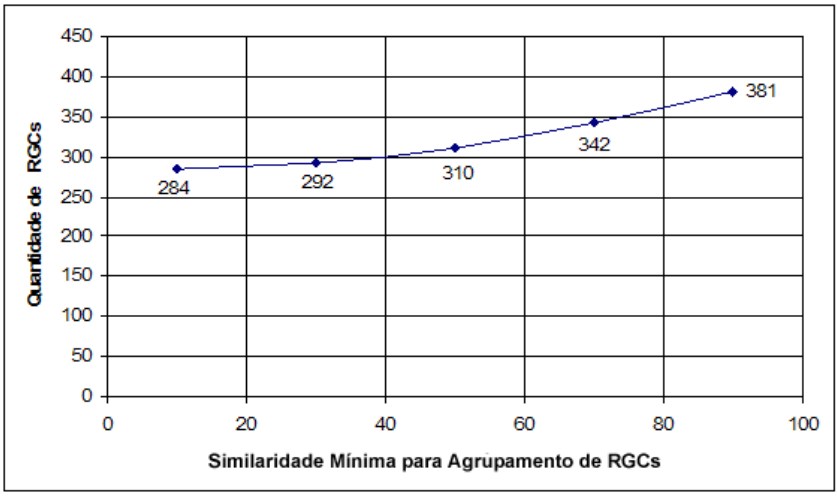


Figura 4.4 - Gráfico da quantidades de RGCs pela similaridade mínima para agrupamento de Rgcs. Dados gerados através da comparação dos proteomas das bactérias: Leifsonia xyli subsp. xyli, Corynebacterium glutamicum e Mycobacterium leprea strain TN.

O parâmentro *m* não se limita em influenciar a quantidade de RGCs de um proteoma, mais do que isso, quanto maior o *m,* menor é a quantidade de ROMs encontrada (Figura 4.5). Isto mostra a grande influência do parâmetro *m* na determinação de ROMs.

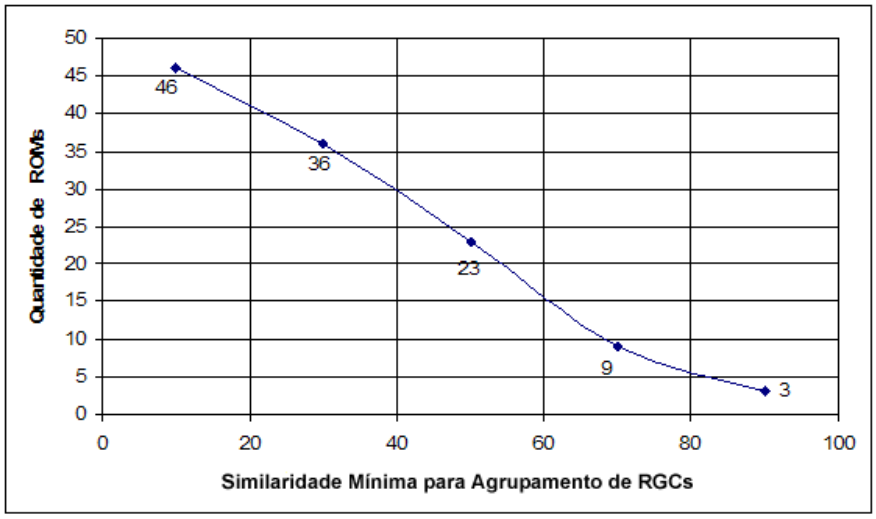


Figura 4.5 - Gráfico da quantidades de ROMs por similaridade mínima para agrupamento de RGCs. Dados gerados através da comparação dos proteomas das bactérias: Leifsonia xyli subsp. xyli, Corynebacterium glutamicum e Mycobacterium leprea strain TN.

Após o agrupamento de RGCs similares, a etapa que se segue diz respeito à construção de um grafo colorido, tarefa que será apresentada no Passo 3.

Passo 3

Neste trabalho um grafo colorido é utilizado para representar relacionamentos entre RGCs de diferentes proteomas. A escolha por um grafo colorido deveu-se à necessidade de particionar os vértices do grafo em conjuntos independentes, com o intuito de dividir um grupo em subgrupos que contenham apenas elementos compatíveis.

O conceito de coloração utilizado aqui difere da definição de *k*-coloração, no qual exige-se que vértices adjacentes tenham cores distintas, sendo *k* o número de cores, e o menor *k* possível denominado número cromático.

Como exemplo, tem-se na Figura 4.6(a) a ilustração de três proteomas com suas respectivas RGCs, e na Figura 4.6(b), o grafo que ilustra como esses proteomas se relacionam. Cada vértice do grafo corresponde a uma RGC que participa de uma RO, sendo que existe uma aresta ligando dois vértices se existe ortologia entre essas regiões.

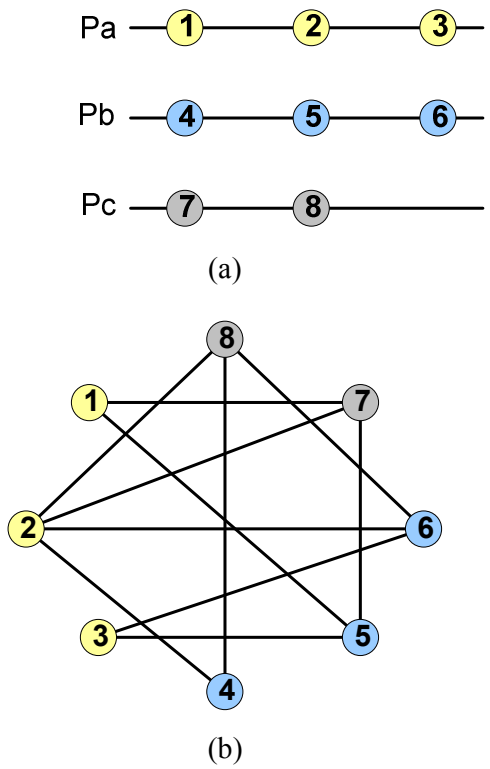


Figura 4.6 – Representação de relacionamentos de RGCs de diferentes proteomas, utilizando um grafo colorido (a) Representação de três proteomas com suas RGCs; (b) Grafo *G* com os relacionamentos entre as RGCs.

Formalmente, na Figura 4.6(b) tem-se um grafo $G_{\varphi} = (V,E)$, onde $V = \{1, 2, 3, 4, 5, 6, 7, 8\}$ e $\varphi : V \rightarrow \{Amarelo, Azul, Cinza\}$, constituindo 3 classes de coloração, uma para cada proteoma. Assim, é possível separar conjuntos de RGCs de cada proteoma. É importante verificar também, que vértices de uma mesma classe de coloração não possuem arestas entre si.

A modelagem do problema utilizando um grafo colorido foi essencial para a criação de um algoritmo de redução do espaço de busca, que será apresentado logo a seguir no Passo 4.

Passo 4

Esta etapa do processo de obtenção de ROMs constitui basicamente de um algoritmo de refinamento desenvolvido neste trabalho. No algoritmo, iterativamente, são eliminados os vértices do grafo construído no passo anterior, que não participarão de cliques com uma quantidade de vértices *n,* sendo *n* o número de proteomas selecionados

para comparação. Isto é possível, pois se busca ROMs entre todos os proteomas envolvidos na comparação.

Na Figura 4.7 a seguir, tem-se a descrição do algoritmo utilizado para eliminação de vértices do grafo e, conseqüentemente, redução do espaço de busca. No algoritmo, a quantidade de vértices adjacentes a um vértive *v* e pertencentes a classes de coloração distintas, é conseguida através da função getQtAdjColor, e a variável qtComparePtrm refere-se à quantidade de proteomas selecionados para comparação.

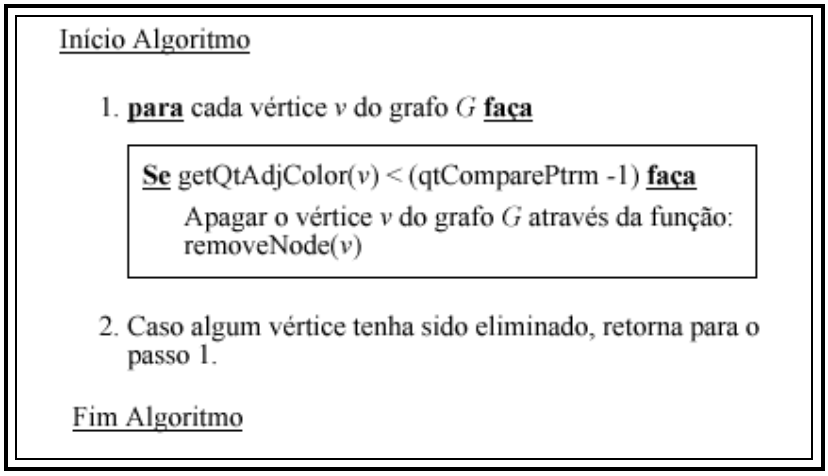


Figura 4.7 - Algoritmo para redução do Espaço de Busca

Na Tabela 2 a seguir são exibidos os proteomas cadastrados na ferramenta MPC com suas respectivas siglas, os quais são a fonte para a execução mostrada na Figura 4.8.

Tabela 2 – Proteomas de Bactérias

Sigla	Proteomas de Bactérias
le	*Leifsonia xyli subsp. Xyli*
cg	*Corynebacterium glutamicum*
ml	*Mycobacterium leprea strain TN*
mt	*Mycobacterium tuberculosis H37Rv*
sc	*Streptomyces coelicolor A3*
bi	*Bifidobacterium longum NCC2705*

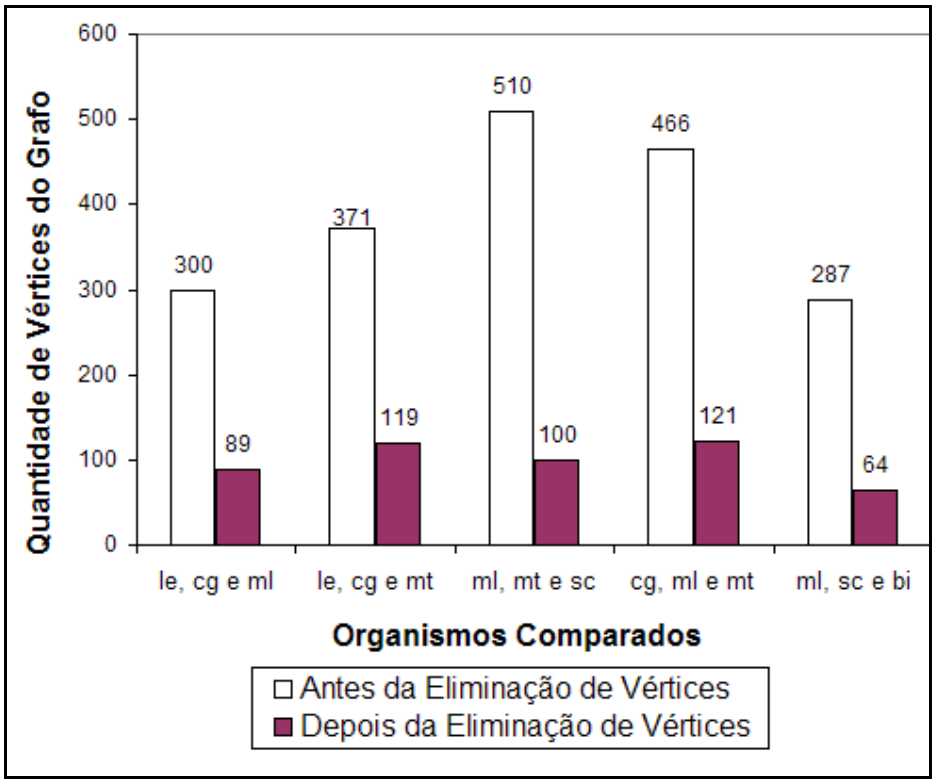


Figura 4.8 – Redução do Espaço de Busca: Quantidade de vértices do grafo antes e depois da execução do algoritmo de eliminação de vértices do grafo

A Figura 4.8 ilustra exemplos reais de redução do espaço de busca. Os dados foram conseguidos através da comparação dos organismos da Tabela 2, utilizando a ferramenta MPC.

Passo 5

Com o espaço de busca reduzido consideravelmente, o último passo da abordagem para encontrar ROMs consiste na execução de um algoritmo para este fim, o qual é apresentado na Figura 4.9.

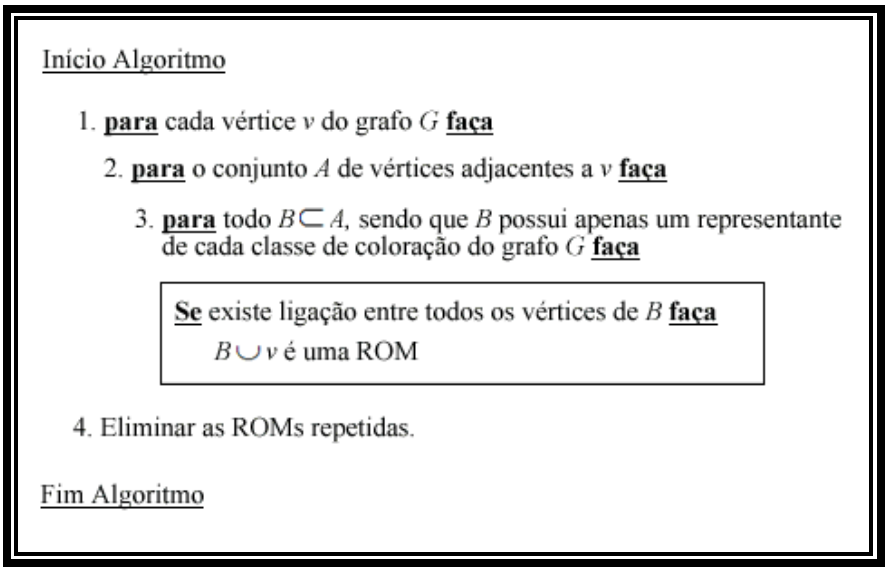


Figura 4.9 –Algoritmo para encontrar Regiões Ortólogas Múltiplas

Nesta seção descreveu-se todo o processo utilizado pela ferramenta MPC na busca por ROMs, além dos três principais algoritmos utilizados para este fim. O algoritmo que realiza o agrupamento de RGCs similares é o que consome mais tempo, pois todas RGCs de um mesmo proteoma precisam ser testadas entre si. O tempo gasto pelo algoritmo para encontrar regiões ortólogas múltiplas é reduzido, pois o número de vértices no grafo é diminuído consideravelmente após a execução do algoritmo de redução do espaço de busca. Isto pode ser constatado no apêndice A, o qual mostra vários testes realizados pela ferramenta MPC.

A seção seguinte apresenta detalhes da implementação da ferramenta MPC.

4.3. Implementação da ferramenta MPC

4.3.1. Arquitetura

A arquitetura definida para a ferramenta é ilustrada na Figura 4.10, e cada módulo participante será definido a seguir.

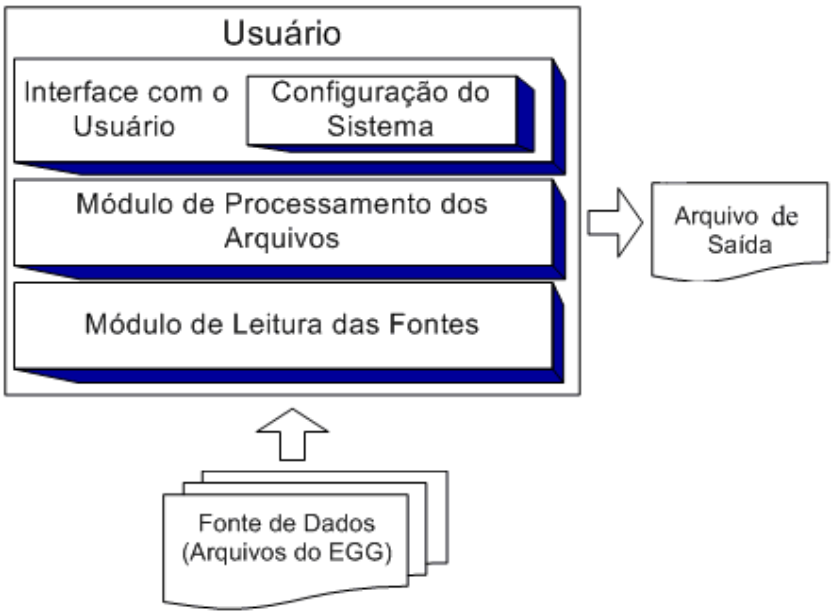


Figura 4.10 – Arquitetura da Ferramenta MPC

- **Módulo de Leitura das Fontes**: Na camada inferior encontram-se os arquivos oriundos da ferramenta EGG, que informam o relacionamento entre dois proteomas e, servem como fonte de dados. O módulo de leitura das fontes é responsável pelo pré-processamento desses dados de entrada;

- **Módulo de Processamento dos Arquivos:** Este módulo é responsável por colocar em prática a execução dos principais algoritmos, dentre eles, algoritmo de diminuição do espaço de busca e o algoritmo para encontrar ROMs;

- **Módulo de Interface:** Basicamente o módulo de interface fornecerá duas funcionalidades:

    - Configuração do Sistema: disponibiliza meios que permitem ajustar parâmetros para encontrar ROMs; definir os conjuntos de dados que servirão de fonte para o sistema; definir a similaridade mínima para o agrupamento de RGCs; e especificar o local de saída dos dados.
    - Processamento e visualização de dados: permite a realização do processamento dos dados de entrada e disponibiliza os resultados em formato texto.

A linguagem escolhida para modelar as demais metodologias de captura e racionalização de processo da ferramenta é a *Unified Modeling Language* (UML) a qual é adequada para especificar, visualizar, construir, e documentar artefatos de sistemas de *softwares*, bem como para modelagem de negócios e outros sistemas que não são *softwares*. Ela representa uma coleção das melhores práticas de engenharia que têm sido aprovadas com sucesso na modelagem de sistemas complexos e de grande escala (BOOCH, RUMBAUGH e JACOBSON, 2005; OMG, 2006).

A seguir tem-se a Figura 4.11 apresentando o diagrama de Caso de Uso geral da ferramenta.

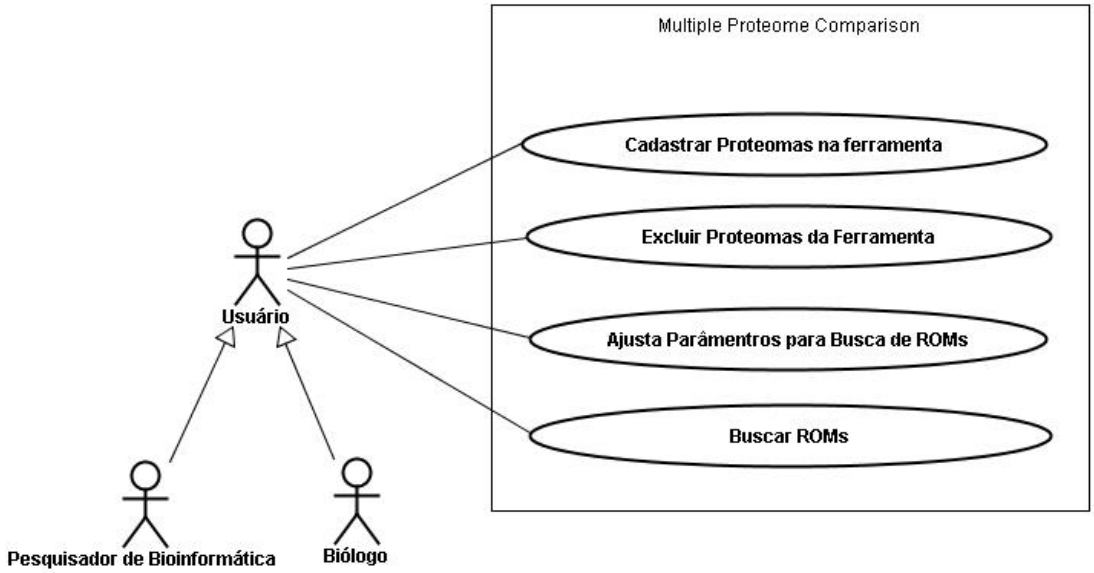


Figura 4.11 – Diagrama de Caso de Uso da ferramenta *Multiple Proteome Comparison*

Esse diagrama mostra um conjunto de atores e seus relacionamentos, ilustrando uma visão estática global da ferramenta. Outra visualização de modelagem importante é um diagrama de classes da ferramenta MPC, ilustrado na Figura 4.12. Esse diagrama é mostrado de maneira simplificada e contém alguns detalhes de implementação, representando bem a visão estática e os relacionamentos das classes da ferramenta, com as principais funções.

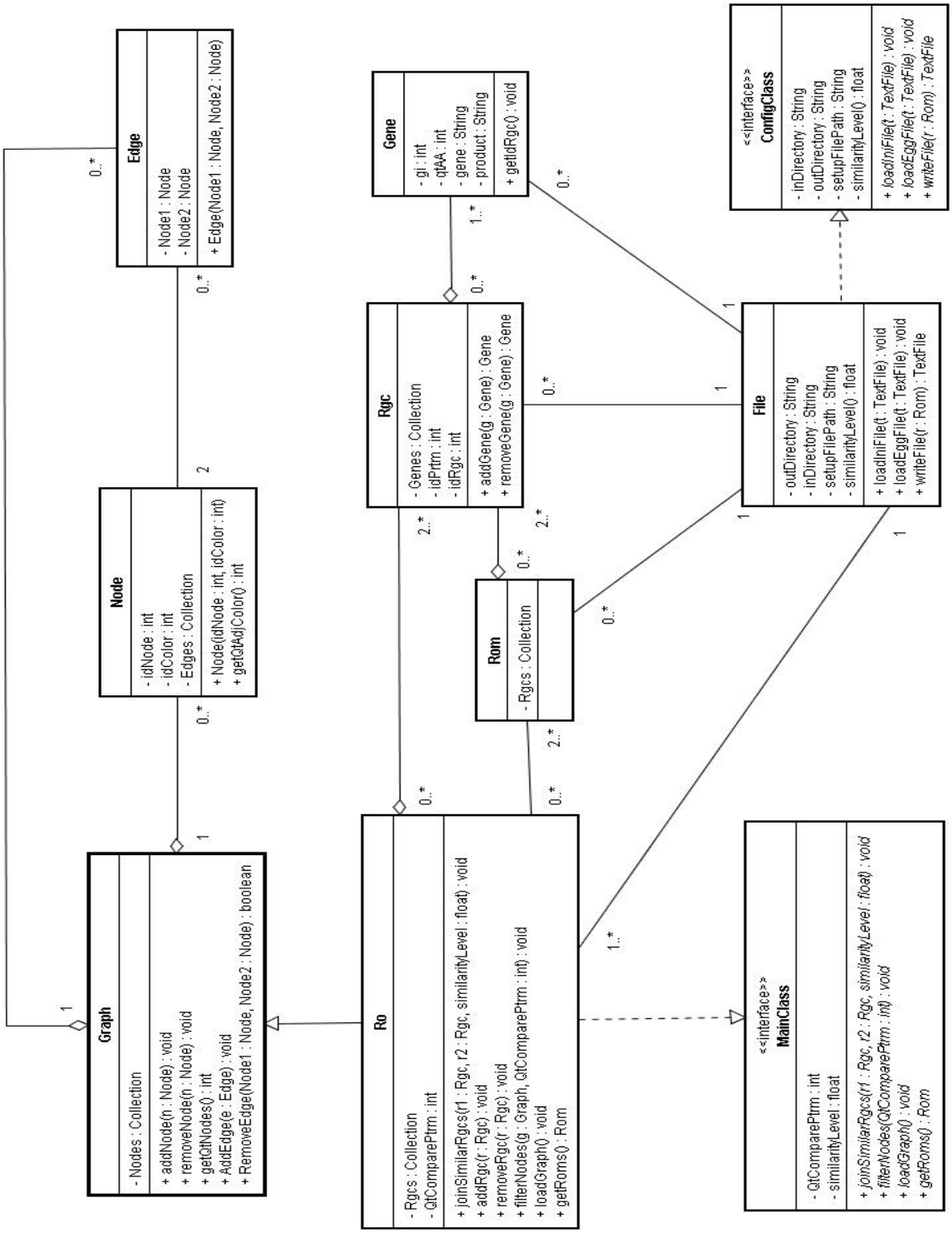


Figura 4.12 – Diagrama de Classes da Ferramenta MPC

4.3.2. Interface

Nesta seção será apresentada a interface da ferramenta MCP. Resumidamente, trata-se de uma ferramenta destinada a encontrar ROMs entre diversos proteomas previamente cadastrados, com base em comparações dois a dois feitas pelo EGG.

A Figura 4.13 ilustra a tela inicial da ferramenta. A parte de cima é destinada à exibição do estado atual dos parâmetros de configuração; para alterá-los é só entrar no menu "Configurações" e depois em "Ajustar Parâmetros". Na parte de baixo estão os proteomas cadastrados e passíveis de se realizar comparações.

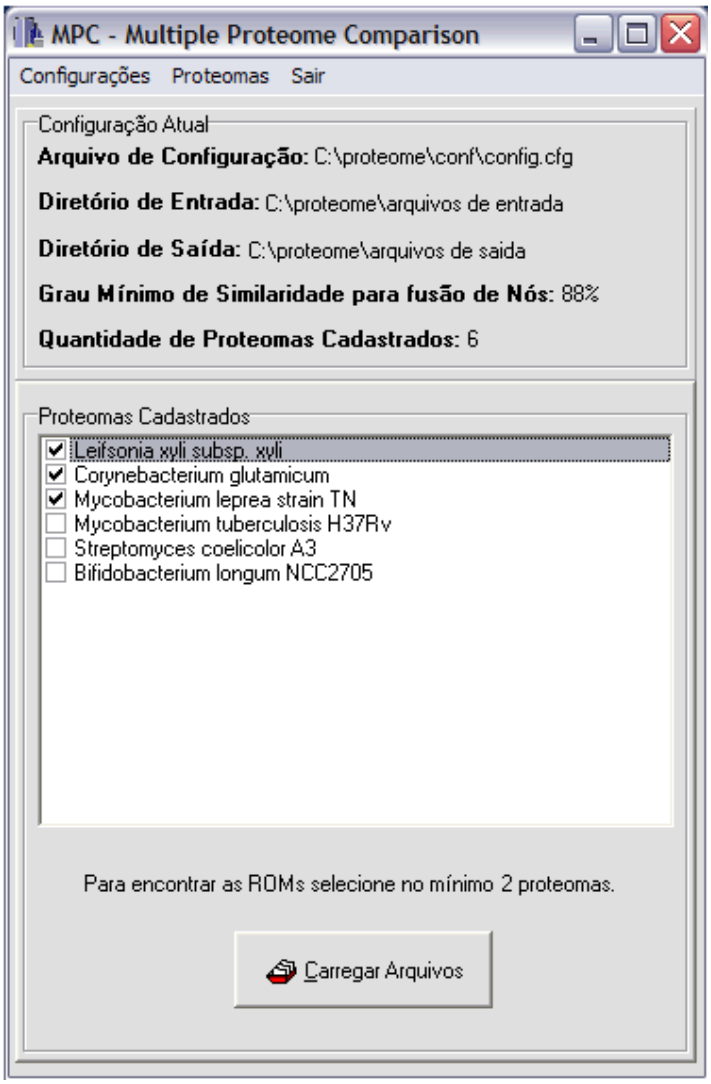


Figura 4.13 – Tela inicial da Ferramenta MPC.

Uma etapa muito importante antes da execução da ferramenta diz respeito ao ajuste dos parâmetros de configuração que é realizado com a ajuda da interface exibida na Figura

4.14. Logo no topo da tela existe um espaço reservado para o ajuste do "Arquivo de Configuração". Este arquivo nada mais é que uma lista contendo todos os proteomas cadastrados, juntamente com um identificador e uma sigla. Assim que o programa é iniciado, este arquivo é lido, e os proteomas são mostrados na tela inicial. Não é preciso se preocupar com o conteúdo deste arquivo, a ferramenta MPC disponibiliza de uma interface para gerenciar os proteomas cadastrados, que será vista mais adiante nesta seção.

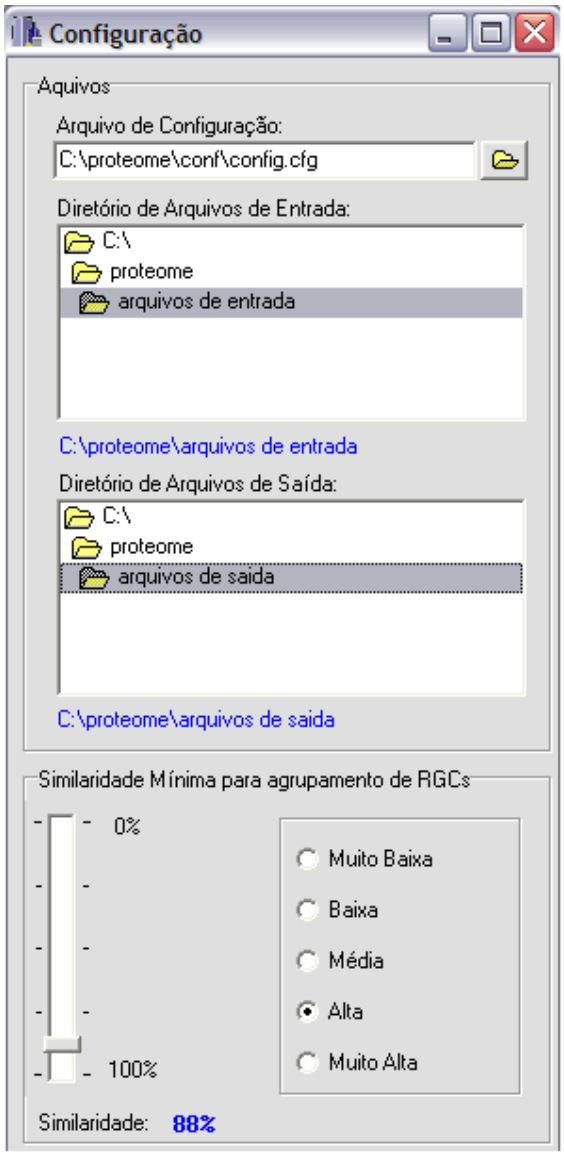


Figura 4.14 – Tela de Configuração da Ferramenta MPC.

Após ajustar o arquivo de configuração, é preciso especificar o diretório onde os arquivos do EGG, referentes a comparações dois a dois dos proteomas, estão depositados.

O espaço reservado para este fim está descrito como “Diretório de Arquivos de Entrada”. Logo abaixo, encontra-se o local disponibilizado para especificar o diretório de saída de dados da ferramenta MPC, denominado “Diretório de Arquivos de Saída”.

O usuário também precisa calibrar a similaridade mínima necessária para agrupar RGCs, e a escolha pode ser feita de duas formas. A primeira através da barra vertical, o que possibilita escolher um valor de similaridade mais preciso. Na segunda forma, o usuário pode escolher diretamente a intensidade mínima de similaridade. Neste caso, o valor médio do intervalo de similaridade é escolhido como a similaridade mínima necessária para realizar um agrupamento de RGCs, como pode-se observar na Tabela 3.

Tabela 3– Intervalos de Similaridades para agrupamento de RGCs

Intensidade de Mínima de Similaridade	Intervalo de Similaridade	Valor Médio
Muito Baixa	]0,20[	10
Baixa	[20,40[	30
Média	[40,60[	50
Alta	[60,80[	70
Muito Alta	[80,100]	90

Como dito anteriormente, a ferramenta MPC disponibiliza meios para gerenciar os proteomas cadastrados. Para cadastrar o proteoma de um organismo na ferramenta, basta entrar no menu “Proteomas” e depois em “Cadastrar”. A tela exibida é ilustrada na Figura 4.15 abaixo.

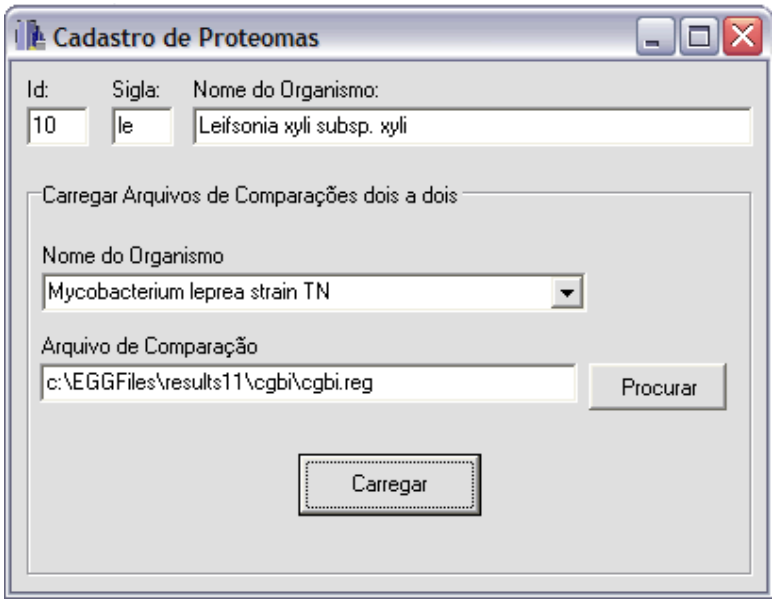


Figura 4.15 – Interface para Cadastro de Proteomas

O campo “Id” é identificador único para cada proteoma cadastrado. Após especificar o nome do organismo que será inserido na ferramenta juntamente com sua sigla e seu identificador, é preciso carregar os arquivos de comparações, um para cada organismo que já estava cadastrado na ferramenta.

A Figura 4.16 ilustra a interface destina a excluir proteomas cadastrados na ferramenta. Para a exibição da interface, basta entrar no menu “Proteomas” e depois em “Cadastrar”. Para excluir proteomas da ferramenta, é preciso selecioná-lo e pressionar o botão “Excluir Proteoma(s)”.

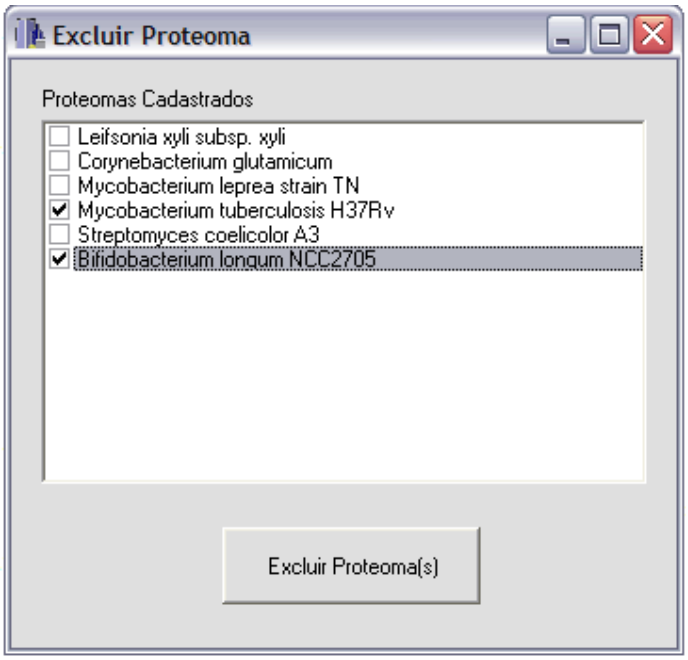


Figura 4.16 - Interface para Exclusão de Proteomas

Com os parâmetros de configuração ajustados e os proteomas devidamente cadastrados, a ferramenta está pronta para iniciar o cálculo das ROMs. Para isso, o usuário precisa escolher quais proteomas deseja investigar com o objetivo de encontrar ROMs. Isso é possível logo na tela inicial (Figura 4.13). Com os proteomas selecionados, ao pressionar o botão “Carregar Arquivos”, os arquivos presentes no diretório de entrada são lidos e carregados na memória. A próxima tela é a da Figura 4.17.

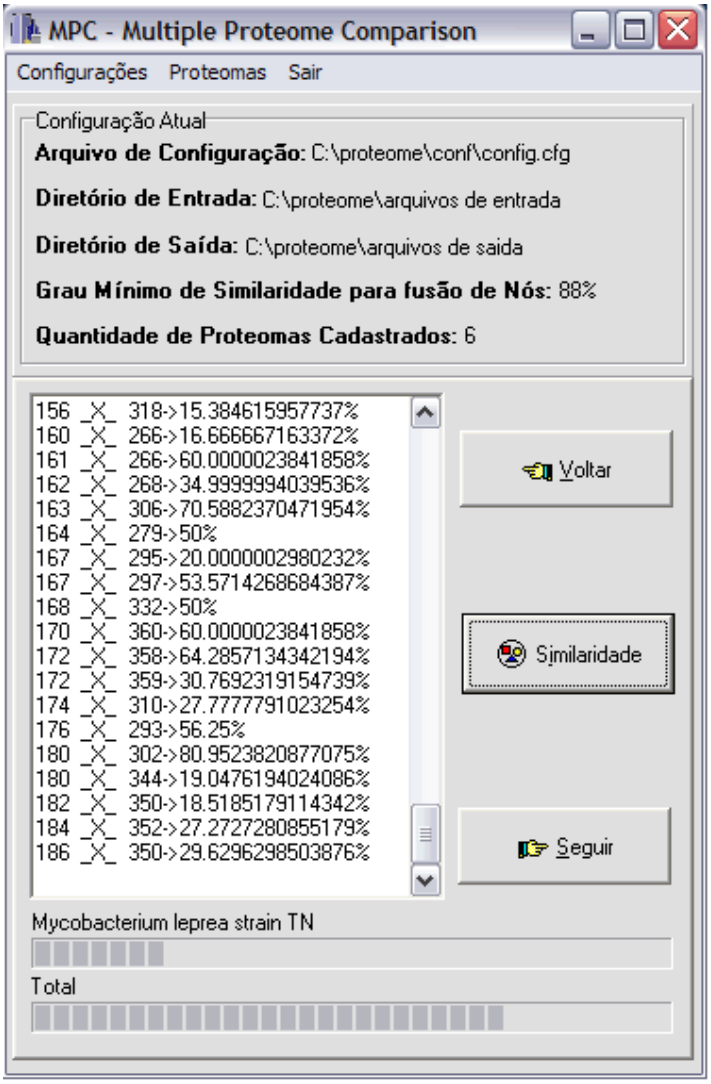


Figura 4.17 – Tela de Cálculo de Similaridades entre RGCs

A tela em questão possibilita disparar os cálculos de similaridades entre as RGCs através do botão “Similaridade”, e agrupa RGCs similares de acordo com a similaridade mínima definida previamente pelo usuário.

Finalmente, após agrupar RGCs similares, é possível encontrar ROMs entre os organismos selecionados anteriormente. Para isso, basta pressionar o botão “Encontrar ROMs” da tela ilustrada na Figura 4.18.

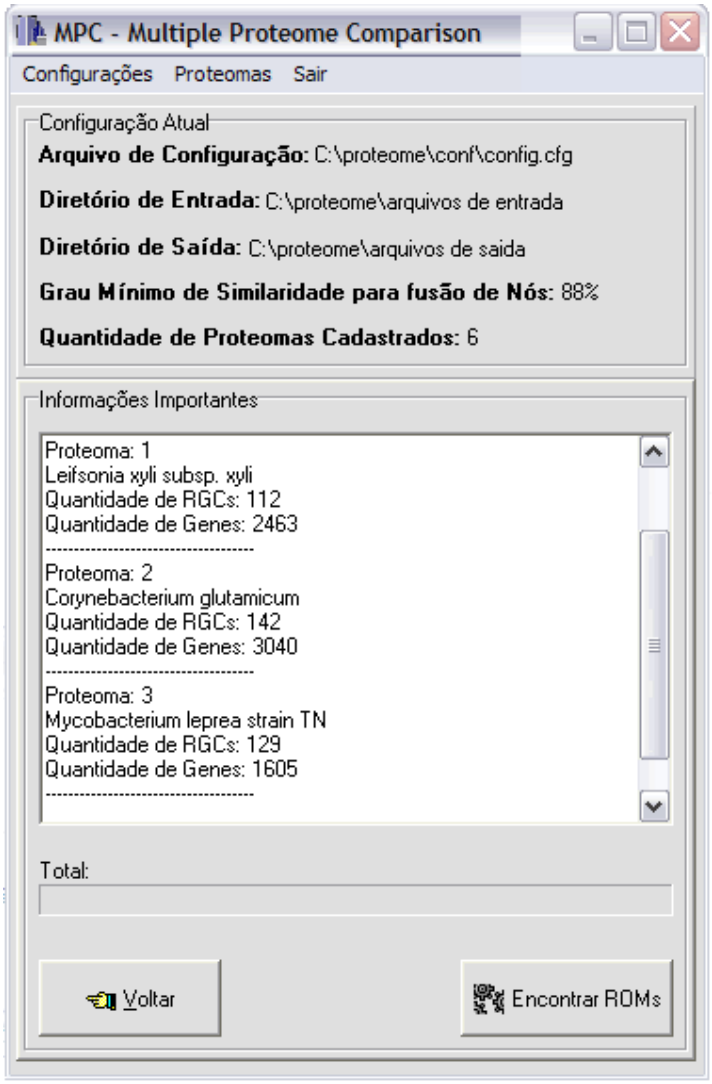


Figura 4.18 – Tela para Iniciar o Cálculo de ROMs

A Figura 4.19 ilustra parte do arquivo de saída resultante da comparação dos proteomas das bactérias Leifsonia xyli subsp. xyli, Corynebacterium glutamicum e Mycobacterium leprea strain TN, efetuados pela ferramenta MPC. O arquivo inicia-se com a sigla de todos os proteomas envolvidos na comparação, seguido de um contador de ROMs, da data de execução e da similaridade mínima escolhida pelo usuário para agrupamento de RGCs (no exemplo 77%). Nas próximas linhas, enumera-se o conjunto de genes que compõe a ROM encontrada, separado por proteoma. Por fim, exibem-se os *matchs* dois a dois, ou seja, a relação dos genes pertencentes a ROM encontrada.

```
>LE-CG-ML-3-060702-S77
=============================================================================
Gene (LE)     gi          size      product
=============================================================================
-LE2130.1      213011   617aa     ABC transporter, NBP/MSD fusion protein (pimaricin)
-LE2140.1      214011   575aa     ABC transporter, ATP-binding protein
=============================================================================
Gene (CG)     gi          size      product
=============================================================================
+Cgl0928       19552178466aa     COG1132:ABC-type multidrug/protein/lipid transport system, ATPase com
+Cgl0929       19552179530aa     COG1132:ABC-type multidrug/protein/lipid transport system, ATPase com
=============================================================================
Gene (ML)     gi          size      product
=============================================================================
-ML1113        13093097 629aa     probable ABC transporter, ATP-binding component
-ML1114        13093098 584aa     probable ABC transporter protein, ATP-binding component
================
matches (LE-CG)
================

=============================================================================
Gene          start       size    e-value  [    best hit   ]          product
=============================================================================
-LE2140.1   211987     575    2e-34    [-Cgl0940   /1e-174]   ABC transporter, ATP-binding protein
+Cgl0929    1010201    530    2e-34    [-LE2130.1  /2e-55 ]   COG1132:ABC-type multidrug/protein/lipid transport system, ATPase com
-----------------------------------------------------------------------------
-LE2140.1   211987     575    6e-14    [-Cgl0940   /1e-174]   ABC transporter, ATP-binding protein
+Cgl0928    1008664    466    4e-14    [-LE2130.1  /2e-22 ]   COG1132:ABC-type multidrug/protein/lipid transport system, ATPase com

-----------------------------------------------------------------------------
-LE2130.1   210006     617    3e-55    [-Cgl0939   /0    ]    ABC transporter, NBP/MSD fusion protein (pimaricin)
+Cgl0929    1010201    530    2e-55    [best             ]    COG1132:ABC-type multidrug/protein/lipid transport system, ATPase com
-----------------------------------------------------------------------------
-LE2130.1   210006     617    4e-22    [-Cgl0939   /0    ]    ABC transporter, NBP/MSD fusion protein (pimaricin)
+Cgl0928    1008664    466    2e-22    [best             ]    COG1132:ABC-type multidrug/protein/lipid transport system, ATPase com
-----------------------------------------------------------------------------

================
matches (LE-ML)
================

=============================================================================
Gene          start       size    e-value  [    best hit   ]          product
=============================================================================
-LE2140.1   211987     575    1e-151   [best             ]    ABC transporter, ATP-binding protein
-ML1114     1289623    584    1e-151   [best             ]    probable ABC transporter protein, ATP-binding component
-----------------------------------------------------------------------------
-LE2140.1   211987     575    9e-60    [-ML1114    /1e-151]   ABC transporter, ATP-binding protein
-ML1113     1287737    629    1e-59    [-LE2130.1  /1e-171]   probable ABC transporter, ATP-binding component
-----------------------------------------------------------------------------
-LE2130.1   210006     617    8e-61    [-ML1113    /1e-171]   ABC transporter, NBP/MSD fusion protein (pimaricin)
-ML1114     1289623    584    1e-60    [-LE2140.1  /1e-151]   probable ABC transporter protein, ATP-binding component
-----------------------------------------------------------------------------
-LE2130.1   210006     617    1e-171   [best             ]    ABC transporter, NBP/MSD fusion protein (pimaricin)
-ML1113     1287737    629    1e-171   [best             ]    probable ABC transporter, ATP-binding component
-----------------------------------------------------------------------------

================
matches (CG-ML)
================

=============================================================================
Gene          start       size    e-value  [    best hit   ]          product
=============================================================================
+Cgl0929    1010201    530    3e-25    [-ML1113    /3e-45 ]   COG1132:ABC-type multidrug/protein/lipid transport system, ATPase com
-ML1114     1289623    584    6e-25    [-Cgl0940   /1e-164]    probable ABC transporter protein, ATP-binding component
-----------------------------------------------------------------------------
+Cgl0929    1010201    530    3e-45    [best             ]    COG1132:ABC-type multidrug/protein/lipid transport system, ATPase com
-ML1113     1287737    629    6e-45    [-Cgl0939   /0    ]    probable ABC transporter, ATP-binding component
-----------------------------------------------------------------------------
+Cgl0928    1008664    466    2e-09    [-ML1113    /3e-14 ]   COG1132:ABC-type multidrug/protein/lipid transport system, ATPase com
-ML1114     1289623    584    4e-09    [-Cgl0940   /1e-164]    probable ABC transporter protein, ATP-binding component
-----------------------------------------------------------------------------
+Cgl0928    1008664    466    3e-14    [best             ]    COG1132:ABC-type multidrug/protein/lipid transport system, ATPase com
-ML1113     1287737    629    7e-14    [-Cgl0939   /0    ]    probable ABC transporter, ATP-binding component
-----------------------------------------------------------------------------
```

Figura 4.19 – Arquivo de Saída da Ferramenta MCP

4.3.3. Ambiente de Implementação

O ambiente de *software* utilizado para implementação foi o *Microsoft Windows XP Professional Edition* como sistema operacional. A implementação em si foi realizada utilizando-se a versão 5 do compilador Borland C++ Builder.

Como se utilizou a tecnologia orientada a objetos para a implementação do sistema, houve grande preocupação com a utilização das características desta tecnologia, isto é, reutilização de código, herança, polimorfismo, abstração e encapsulamento.

4.4. Considerações Finais

Este capítulo apresentou a definição e a implementação de uma abordagem para encontrar ROMs, que deu origem à ferramenta MPC. Sobre a ferramenta, descreveu-se a interface, arquitetura e implementação.

O próximo capítulo é a conclusão do trabalho, sendo apresentadas as principais contribuições e trabalhos futuros.

5. CONCLUSÃO

5.1. Considerações Finais

Devido ao aumento no número de seqüenciamentos e, conseqüentemente, da grande quantidade de dados que necessita ser analisada para se extrair significados biológicos dos mesmos, ferramentas de comparação de genomas são construídas e aprimoradas constantemente.

Este trabalho teve como foco a implementação de uma abordagem para encontrar regiões ortólogas múltiplas, que deu origem a uma ferramenta para a comparação de múltiplos proteomas, denominada MPC. A principal preocupação foi a busca de resultados completos e precisos, uma vez que a obtenção de um conjunto de respostas completas pode ser onerosa devido à quantidade de dados envolvida.

Para a modelagem da solução proposta, utilizou-se de um grafo colorido, que permitiu, dentre outras coisas, a criação de algoritmo para redução do espaço de busca, tornando a tarefa de encontrar ROMs menos onerosa.

5.2. Contribuições

Dentre as principais contribuições deste trabalho, pode-se citar:

1. **Tratamento da sobreposição de RGCs através de uma medida de similaridade entre conjuntos**

   Como foi visto, pode ocorrer de RGCs compartilharem um mesmo subconjunto de genes, o que se faz necessário, em alguns casos, agrupar essas RGCs. Como essa tarefa é crucial na resolução do problema, optou-se por uma medida de similaridade entre as RGCs para inferir nos agrupamentos. Essa adoção torna mais intuitiva, por parte do usuário, a decisão de um limiar para os agrupamentos.

2. **Desenvolvimento de um algoritmo de redução do espaço de busca**

   Como a obtenção de ROMs é uma tarefa que possui um custo elevado, devido a quantidade de dados envolvida, optou-se por solucionar o problema de uma forma que o espaço de busca fosse reduzido, para possibilitar trazer resultados completos em um

tempo hábil. Assim, utilizou-se de um grafo colorido, sendo as ROMs cliques no grafo. Isso permitiu a criação de um algoritmo para a eliminação de vértices e, conseqüentemente, redução do espaço de busca.

3. Desenvolvimento de um algoritmo para encontrar ROMs

O algoritmo desenvolvido para encontrar ROMs é muito rápido, pois sua execução acontece com um número pequeno de vértices no grafo, devido ao algoritmo de eliminação de nós.

4. Implementação de uma ferramenta para comparação múltipla de Proteomas, denominada MPC

A implementação da ferramenta MPC teve como objetivo colocar em prática a abordagem proposta para esse fim.

5.3. Trabalhos Futuros

A seguir são descritos alguns trabalhos que deverão ser posteriormente desenvolvidos.

1. Utilização de uma base de dados na Ferramenta MPC

A principal função de uma base de dados seria o armazenamento de resultados, visando-se evitar reprocessamentos. Por exemplo, uma tarefa que consome um tempo considerável na obtenção do ROMs é o agrupamento de RGCs semelhantes, já que todas as RGCs de um proteoma precisam ser testadas entre si. A inserção de um banco de dados possibilitaria armazenar previamente a similaridade entre todas as RGCs. Assim, o tempo total gasto na obtenção de ROMs teria uma diminuição considerável.

2. Criação de uma versão *web* da ferramenta MPC

Para facilitar o acesso e disseminar a utilização da ferramenta MCP, uma tarefa interessante consiste em desenvolver uma versão *web* da ferramenta MPC.

3. Adição de novas funcionalidades à ferramenta MPC

3.1. Criar diferentes níveis de acesso aos usuários da ferramenta. Por exemplo, apenas o usuário administrador teria a permissão para cadastrar, remover ou alterar usuários e proteomas na ferramenta;

3.2. O trabalho aborda um mecanismo para encontrar regiões ortólogas em múltiplos proteomas, mas uma tarefa importante e não realizada neste trabalho consiste no estudo e desenvolvimento de mecanismos para realizar alinhamentos múltiplos entre essas regiões;

3.3. Apresentar os resultados da ferramenta graficamente, para possibilitar uma melhor análise dos mesmos;

3.4. Uma limitação da ferramenta consiste na busca de ROMs que contenham todos os organismos selecionados para comparação. Uma tarefa adicional seria encontrar todas as ROMs dentre os organismos selecionados.

4. Análise Comparativa da ferramenta MPC

Não foram realizadas análises comparativas com ferramentas que realizam a mesma tarefa. Pretende-se realizar testes comparativos com a ferramenta Bagre, com o propósito de se analisar desempenho e o conjunto resposta.

BIBLIOGRAFIA

ABEDON S. T. **Bacteria Cell Shapes and Arrangements**. Disponível em: < http://www.mansfield.ohio-state.edu/~sabedon/biol2010.htm >. Acesso em: jul.2006.

ADAMS, M. D. *et al.* **The genome sequence of Drosophila melanogaster.** Science, 287(5461), mar. 2000.

ALBERTS *et al.* **Fundamentos de biologia celular - Uma introdução à biologia molecular da célula**, 1ª Edição, Editora Artmed, Porto Alegre: RS, 2002.

ALBERTS *et al.* **Molecular Biology of the Cell**., 4th. Edição, Garland Science, Nova Iorque: NY, 2002.

ALMEIDA, N.F. **Ferramentas para Comparação Genômica**. Dissertação (Doutorado). IC-UNICAMP, 2002.

ANSWERS. **Answers – Molecular Biology**. Disponível em: < http://www.answers.com/ topic/molecular-biology >. Acesso em: jun.2006.

APWEILER, R. *et al.* **UniProt: the Universal Protein knowledgebase**. Nucleic Acids Res., 2004.

BENSON, D. A. *et al.* **GenBank**, Nucleic Acids, 2002.

BIOPROJECT. **The Biology Project**. Large Molecules Problems Set. Disponível em <http://www.biology.arizona.edu/the_biology_project/the_biology_project.html> . Acesso em: fev.2006.

BOOCH, G; RUMBAUGH, J e JACOBSON, I. **UML: Guia do Usuário.** 1. São Paulo: Editora Campus. V.1.2005. 474 p.

CHOI, K. *et al.* **PLATCOM: a Platform for Computational Comparative Genomics**.

Bioinformatics. 2005.

COOPER, G. M. **The Cell- A molecular approach**. 1st edition. ASM Press. Washington, DC, 1996.

DALTON, M., **Codon Table**. Disponível em: <http://ccgb.umn.edu/~mwd/cell_www/chapter2/codon_table.html>. Acesso em: jun.2006.

DAYHOFF M.O., SCHWARTZ R. e ORCUTT B.C. **Atlas of Protein Sequence and Struct**ure. Vol. 5. Suppl. 3, 345-358, 1978.

DDBJ. **DNA DataBank of Japan**. Disponível em: <http://www.ddbj.nig.ac.jp/Welcome-j.html>. Acessado em: jan.2006.

EMBL. **The EMBL Nucleotide Sequence Database.** Nucleic Acids Res., jan. 2004.

ENZYME. Enzyme nomenclature database. Disponível em: <http://us.expasy.org/enzyme/>. Acesso em: jan. 2006.

GENBANK. **Crescimento do GenBank**. Disponível em: <http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html>. Acesso em: maio. 2006.

GUSFIELD, D., **Algorithms on *String*s, Trees and Sequences Computer Science and Computational Biology**. Press Syndicate of the University of Cambridge. USA, 1997.

HENIKOFF, S. e HENIKOFF, J. **Amino acid substitution matrices from protein blocks**. in 'Proceedings of the National Academy of Sciences of the USA', Vol. 19, pp. 6565–6572, 1991.

JENSEN, R.A. **Orthologs and paralogs – we need to get it right.** Genome Biol.2, 2001.

KANEHISA, M. e GOTO, S. **Kegg: Kyoto encyclopedia of genes and genomes**. Nucleic Acids Res., 2000.

LENGAUER, T. **Computational Biology at the Beginning of the Post-genomic Era**. LNCS. Vol. 2000. Informatics: 10 Years Back - 10 Years Ahead. R. Wilhelm(Ed.), Springler, Berlin, p. 341-355. 2000.

LI, J. B. **Procom: a web-based tool to compare multiple eukaryotic proteomes**. Bioinformatics, 2005.

LU, G. **GenomeBlast: a web tool for small genome comparison**. Symposium of Computations in Bioinformatics and Bioscience, 2006.

MIR, L. **Genômica**. São Paulo: Atheneu, 2004.

MONTERA, L. **Regiões Ortólogas em Múltiplos Genomas**, Dissertação (Mestrado). UFMS, Campo Grande, 2004.

NCBI. **National Center For Biotechnology Information (NCBI)**. Disponível em: <http://www.ncbi.nlm.nih.gov/>. Acesso em: maio. 2006.

O'BRIEN, K.P.; REMM, M. e SONNHAMMER, E.L.L. **Inparanoid: a comprehensive database of eukaryotic orthologs.** Nucleic Acids, 2005.

PDB. **Protein DataBank**. Disponível em: <http://www.rcsb.org/pdb/>. Acessado em: jan. 2005.

PEARSON, W. R. **Rapid and sensitive sequence comparisions with FASTP and FAST**. Methods in Enzymology, 1985.

PEARSON, W.R. e LIPMAN, D.J. **Improved tools for biological sequence comparison.** Proceedings of the National Academy of Sciences, 1988.

PIR. **Protein Information Resource**. Disponível em: <http://pir.georgetown.edu/> . Acesso em jun. 2006.

SEIBEL, L. **Bio-AXS: Uma Arquitetura para Integração de Fontes de Dados e Aplicações de Biologia Molecular**. Dissertação (Doutorado). PUC-Rio, Dep. de Informática, 2000.

SIGRIST, C. J. *et al.* **PROSITE: a documented database using patterns and profiles as motif descriptors**. Brief Bioinform, 265-274, 2002.

SILVA, F.H. **Fundamentos em Biotecnologia**. Relatório Técnico. DGE UFSCar. São Carlos, 2001.

SWISS-PROT. **Swiss-Prot Protein knowledgebase**. Disponível em: <http://us.expasy.org/sprot/>. Acessado em: Janeiro/2006.

TAMAMES, J. **Evolution of gene order conservation in prokaryotes**. Genome Biology, 2001.

TATUSOV, R.L. *et al.* **The COG database: an updated version includes eukaryotes.** Bioinformatics, 2003.

TREANGEN, T e MESSEGUER, X. **M-GCAT: Multiple genome comparison and alignment tool**. In Academic Demo at the Annual Meeting for the International Society For Computational Biology, ISMB, 2005.

UCHIYAMA, I. **MBGD: microbial genome database for comparative analysis**, Nucleic Acids Res, v. 31, n. 1, p. 58-62, 2003.

XIE, T. e HOOD, L. **ACGT – a comparative genomics tool.** Bioinformatics, 2003.

XU, W. *et al.* **Indexing Protein Sequences in Metric Space**. Technical Report TR-04-06. The University of Texas at Austin. Department of Computer Sciences, 2003.

YANG, J. *et al.* **GenomeComp: a visualization tool for microbial genome comparison**.

Microbiol Methods, 2003.

OMG. **Unified Modeling Language Specification Version 2.0**. 2006. Disponível em: <http://www.omg.org/technology/documents/formal/uml.htm>. Acesso em: mar. 2006.

GLOSSÁRIO

Ácido Nucléico	Molécula biológica composta por uma longa cadeia de nucleotídeos. O DNA é formado por quatro tipos de nucleotídeos repetidos aleatoriamente milhares de vezes.
Adenina	Base nitrogenada encontrada no DNA e RNA.
Alelo	Uma das diversas formas alternativas de um gene específico que ocupa uma certa região do cromossomo.
Aminoácidos	Unidades fundamentais de uma molécula de proteína. Nosso corpo pode sintetizar a maioria dos aminoácidos a partir de seus componentes (carbono, nitrogênio, oxigênio, hidrogênio e, algumas vezes, enxofre). Entretanto, oito aminoácidos (chamados aminoácidos essenciais) devem ser obtidos através da dieta alimentar.
Árvore filogenética	Uma árvore filogenética mostra as relações de origem e parentesco entre as espécies atuais e seus ancestrais. A primeira árvore filogenética que se tem notícia foi elaborada pelo naturalista alemão Ernest Haeckel (1834-1919).
Bioinformática	É o uso de técnicas computacionais para análises de caracterização molecular, integrando modelos matemáticos e estatísticos utilizados na interpretação e análise dos problemas biológicos.
Biotecnologia	Utilização de organismos vivos para resolução de problemas e geração de produtos de interesse. Atualmente é definido como o conjunto de tecnologias que utilizam células vivas e/ou moléculas biológicas para resolução de problemas e geração de produtos de interesse.
Célula	Unidade estrutural e funcional de todos os seres vivos. De acordo com a organização estrutural, as células são divididas em células procariontes e células eucariontes.
Célula Eucariótica	Caracterizada pela presença de uma membrana que define e protege o núcleo celular.
Célula Procariótica	Caracterizada pela ausência de uma membrana individualizando o núcleo celular.
Citoplasma	Material celular contido dentro de uma

	membrana celular e que circunda o núcleo.
Citosina	Base nitrogenada encontrada no DNA ou RNA.
Contig	Conjunto de reads.
Complexo de Golgi	Conjunto de bolsas ou cisternas membranosas, achatadas e empilhadas, tendo ao redor pequenos vacúolos ou vesículas que se desprendem por brotamento (os lisossomos).
Cromossomo	Componentes celulares que contém as informações genéticas. Cada cromossomo contém inúmeros genes. Os cromossomos ocorrem aos pares: um proveniente da mãe e outro proveniente do pai. Cromossomos de casais diferentes são visivelmente diferentes.
DNA (ácido desoxirribonucléico)	Molécula que é a base do material genético encontrado em todas as células. O DNA carrega as informações genéticas de uma geração para a próxima. Como o DNA é uma molécula muito longa e fina, ele é arranjado em unidades, chamadas cromossomos. O DNA pertence a uma classe de moléculas biológicas, chamada de ácidos nucléicos.
Dupla hélice	Termo usado para descrever a configuração da molécula de DNA. A hélice consiste de duas fitas de nucleotídeos, espiralizadas, ligadas uma a outra por pontes de hidrogênio entre as bases.
Enzima	Proteínas que aceleram a velocidade das reações químicas. As enzimas são catalisadores que promovem repetidamente as reações sem serem modificadas por elas.
Gene	Unidade de informação hereditária que define características especificas de cada individuo. Um gene é uma seção da molécula do DNA que especifica a produção de uma proteína em particular.
Gene específico	Um gene g de um genoma G é considerado específico em relação a outro genoma H caso não haja nenhum gene h' pertencente a H ortólogo a g.
Genes homólogos	São genes que descendem de um mesmo ancestral.
Genes ortólogos	São genes que além de homólogos pertencem a genomas diferentes.
Genes parálogos	São genes homólogos pertencentes a um mesmo genoma.

Genética	Ramo da Biologia que estuda os mecanismos de transmissão das características estruturais e funcionais de um organismo para as gerações subseqüentes.
Genoma	É o conjunto de informação genética de um ser vivo, contido no DNA. Sinônimo de genótipo, patrimônio genético ou patrimônio hereditário. Formado por um conjunto de contigs.
Guanina	Base nitrogenada encontrada no DNA ou RNA.
Mitocôndria	Estrutura membranosa de forma variável (esférica ou de bastonete), constituída de duas membranas, cristas mitocondriais e matriz mitocondrial, com capacidade de auto-reprodução.
Moléculas biológicas	Moléculas grandes e complexas, como proteínas, lipídeos e carboidratos, que são produzidas somente por organismos vivos. As moléculas biológicas são normalmente chamadas de macromoléculas ou biopolímeros.
Núcleo	A estrutura dentro das células eucarióticas, circundada por uma membrana que contém os cromossomos de um organismo.
Nucleotídeo	Constituinte elementar dos ácidos nucléicos (DNA e RNA). É composto por quatro bases nitrogenadas (adenina, guanina, citosina e tiamina), de um fosfato e de um açúcar. A seqüência de nucleotídeos constitui o código genético de um organismo.
Polímero	Grande molécula formada por uma série de ligações covalentes, que unem várias unidades idênticas ou semelhantes (monômeros).
Polipeptídeo	Polímero linear composto por múltiplos aminoácidos. Proteínas são grandes polipeptídeos, e os dois termos podem ser usados como sinônimos.
Proteína	Uma molécula composta de aminoácidos. Existem muitos tipos de proteínas, todas desenvolvem um número diferente de funções essenciais para o crescimento da célula.
Read	Fragmento de DNA seqüenciado que contém de 500 a 800 bases.
Região Específica	É considerada uma região rica em genes específicos comparando-se dois genomas G

	e H, por exemplo.
Região Promotora	Uma determinada seqüência de nucleotídeos onde a RNA polimerase se liga e inicia a trabscrição.
RGC (Região de Genes Consecutivos)	É o conjunto de genes consecutivos num genoma, de acordo com suas coordenadas de início, independente da fita.
Ribossomo	Grânulo citoplasmático constituído por RNA e proteínas. É o responsável pela síntese de proteínas.
RNA (ácido ribonucléico)	Assim com o DNA, o RNA é um tipo de ácido nucléico. O RNA se diferencia do DNA em três aspectos: os nucleotídeos do RNA contem o açúcar ribose ao invés do desoxirribose; o RNA contém a base uracila ao invés da timina; e o RNA é uma molécula de fita simples ao invés de uma hélice dupla fita.
RNAm (RNA mensageiro)	Ácido nucléico, fita simples, que carrega a instrução para o ribossomo sintetizar uma proteína em particular.
RNAt (RNA transportador)	Molécula de RNA que carrega aminoácidos para sítios nos ribossomos para síntese das proteínas.
RO (Região Ortóloga):	É composta por um par de RGCs de genomas diferentes que são ortólogas entre si.Possuem aproximadamente o mesmo número de genes, além de serem descendentes de uma mesma região ancestral.
Seqüência de DNA	A ordem de bases na molécula de DNA.
Seqüenciamento de DNA	Determinação da seqüência de bases do DNA.
Timina	Base nitrogenada encontrada no DNA.
Tradução	Processo que utiliza um RNA mensageiro molde para sintetizar uma proteína.
Transcrição	Processo de utilização de um DNA molde para fazer uma molécula de RNA complementar.
Uracila	Base nitrogenada encontrada no RNA.

APÊNDICE A

Testes de Desempenho da Ferramenta MPC

Para a realização dos testes da ferramenta MPC, utilizou-se do mesmo ambiente descrito na seção 4.3. O nome de cada organismo comparado é encontrado na Tabela 2, e a similaridade mínima utilizada para o agrupamento de RGCs foi de 40%.

Tabela 4 – Testes de Desempenho da Ferramenta MPC

Organismos Comparados	Quantidades			Cálculo de Tempo (milisegundos)		
	Vértices do Grafo antes da execução do algoritmo de eliminação	Vértices do Grafo após execução do algoritmo de eliminação	ROMs	Agrupar RGCs	Eliminar Vértices	Cálculo de Roms
le☒cg☒ml☒mt☐sc☐bi☐	300	89	30	64189	46	30
le☒cg☒ml☐mt☒sc☐bi☐	371	119	36	135829	78	62
le☒cg☒ml☐mt☐sc☒bi☐	409	137	26	131015	141	30
le☒cg☒ml☐mt☐sc☐bi☒	238	70	22	16688	14	15
le☒cg☐ml☒mt☒sc☐bi☐	368	96	32	272906	78	30
le☒cg☐ml☒mt☐sc☒bi☐	321	101	28	133156	78	30
le☒cg☐ml☒mt☐sc☐bi☒	211	49	18	13530	14	13
le☒cg☐ml☐mt☒sc☒bi☐	448	126	31	242827	156	30
le☒cg☐ml☐mt☒sc☐bi☒	223	49	17	12624	15	13
le☒cg☐ml☐mt☐sc☒bi☒	329	124	17	87249	92	46
le☐cg☒ml☒mt☒sc☐bi☐	466	121	39	613250	124	46
le☐cg☒ml☒mt☐sc☒bi☐	393	88	27	244827	109	14
le☐cg☒ml☒mt☐sc☐bi☒	255	57	21	24187	14	15
le☐cg☒ml☐mt☒sc☒bi☐	533	167	44	305468	92	31
le☐cg☒ml☐mt☒sc☐bi☒	317	68	20	43249	78	15
le☐cg☒ml☐mt☐sc☒bi☒	356	110	18	56296	45	15
le☐cg☐ml☒mt☒sc☒bi☐	510	100	27	324031	78	18
le☐cg☐ml☒mt☐sc☒bi☒	287	64	15	33797	15	15
le☐cg☐ml☐mt☒sc☒bi☒	396	69	12	66828	46	15
le☒cg☒ml☒mt☒sc☐bi☐	724	38	6	1659342	343	13
le☒cg☒ml☒mt☐sc☒bi☐	677	58	7	1031327	359	14
le☒cg☒ml☒mt☐sc☐bi☒	464	35	9	194280	141	14
le☒cg☐ml☒mt☒sc☒bi☐	788	56	4	969530	218	13
le☒cg☐ml☒mt☒sc☐bi☒	504	28	5	295125	93	12
le☒cg☐ml☒mt☐sc☒bi☒	537	52	4	259983	109	14
le☒cg☐ml☐mt☒sc☒bi☒	634	64	3	393796	171	14
le☐cg☒ml☒mt☒sc☒bi☐	903	51	5	1573312	266	13
le☐cg☒ml☒mt☒sc☐bi☒	599	34	5	642734	124	13
le☐cg☒ml☒mt☐sc☒bi☒	590	51	4	619187	296	15
le☐cg☒ml☒mt☒sc☐bi☒	599	34	5	612438	108	15
le☐cg☐ml☒mt☒sc☒bi☒	688	50	2	574922	343	15
le☒cg☒ml☒mt☒sc☒bi☐	1344	53	1	4060264	3312	45
le☒cg☒ml☒mt☒sc☐bi☒	930	32	3	1380905	281	13
le☒cg☒ml☒mt☐sc☒bi☒	955	66	1	1119875	374	30
le☒cg☒ml☐mt☒sc☒bi☒	1099	81	3	1565327	516	30
le☒cg☐ml☒mt☒sc☒bi☒	1032	44	0	1576999	890	14
le☐cg☒ml☒mt☒sc☒bi☒	1137	49	1	2366922	516	13
le☒cg☒ml☒mt☒sc☒bi☒	1667	54	0	7673046	2437	62