ANDRÉ DE SOUZA GOMES

UMA METODOLOGIA PARA IDENTIFICAÇÃO DE MÓDULOS FORMADORES DE SEQUÊNCIAS DE PROTEÍNAS MOSAICAS DO *Trypanosoma cruzi* A PARTIR DO PROTEOMA DO PARASITO UTILIZANDO A FERRAMENTA BLAST

Monografia de graduação apresentada ao Departamento de Ciência da Computação da Universidade Federal de Lavras como parte das exigências do Curso de Ciência da Computação para obtenção do título de Bacharel em Ciência da Computação.

LAVRAS
MINAS GERAIS – BRASIL
2008

ANDRÉ DE SOUZA GOMES

UMA METODOLOGIA PARA IDENTIFICAÇÃO DE MÓDULOS FORMADORES DE SEQUÊNCIAS DE PROTEÍNAS MOSAICAS DO *Trypanosoma cruzi* A PARTIR DO PROTEOMA DO PARASITO UTILIZANDO A FERRAMENTA BLAST

Monografia de graduação apresentada ao Departamento de Ciência da Computação da Universidade Federal de Lavras como parte das exigências do Curso de Ciência da Computação para obtenção do título de Bacharel em Ciência da Computação.

Área de concentração:

Bioinformática

Orientador:

Prof. Dr. Thiago de Souza Rodrigues

LAVRAS
MINAS GERAIS – BRASIL
2008

Ficha Catalográfica preparada pela Divisão de Processo Técnico da Biblioteca Central da UFLA

Gomes, André de Souza

Uma Metodologia para Identificação de Módulos Formadores de Sequências de Proteínas Mosaicas do *Trypanosoma cruzi* a partir do Proteoma do Parasito Utilizando a Ferramenta BLAST / André de Souza Gomes – Minas Gerais, 2008. 47p.

Monografia de Graduação – Universidade Federal de Lavras. Departamento de Ciência da Computação.

1. Bioinformática. 2. Proteínas Mosaicas. 3. *Trypanosoma cruzi*. 4. Proteoma. 5. BLAST. I. GOMES, A. G. II. Universidade Federal de Lavras. III. Título.

ANDRÉ DE SOUZA GOMES

UMA METODOLOGIA PARA IDENTIFICAÇÃO DE MÓDULOS FORMADORES DE SEQUÊNCIAS DE PROTEÍNAS MOSAICAS DO *Trypanosoma cruzi* A PARTIR DO PROTEOMA DO PARASITO UTILIZANDO A FERRAMENTA BLAST

Monografia de graduação apresentada ao Departamento de Ciência da Computação da Universidade Federal de Lavras como parte das exigências do Curso de Ciência da Computação para obtenção do título de Bacharel em Ciência da Computação.

Aprovada em 18 de Novembro de 2008

____
Prof. Dr. Cláudio Fabiano Motta Toledo

____
Prof. Dra. Marluce Rodrigues Pereira

____
Prof. Dr. Thiago de Souza Rodrigues
(Orientador)

LAVRAS
MINAS GERAIS – BRASIL

Aos meus pais Aguinaldo Gomes de Almeida e Rita de Souza Almeida.

A meu irmão Renato de Souza Gomes.

A minha irmã Valéria Gomes de Almeida.

Dedico.

UMA METODOLOGIA PARA IDENTIFICAÇÃO DE MÓDULOS FORMADORES DE SEQUÊNCIAS DE PROTEÍNAS MOSAICAS DO *Trypanosoma cruzi* A PARTIR DO PROTEOMA DO PARASITO UTILIZANDO A FERRAMENTA BLAST

RESUMO

Este trabalho propôs uma metodologia de identificação de módulos formadores de sequências de proteínas mosaicas do *Trypanosoma cruzi* utilizando a ferramenta BLAST. Para o desenvolvimento da metodologia, foi utilizada a família MASP de proteínas e aplicado inicialmente o conjunto de valores padrão dos parâmetros da ferramenta. Posteriormente foram estudadas diferentes combinações de valores de parâmetros a fim de comparação de resultados, incluindo valores indicados pela literatura. A metodologia desenvolvida provou ser eficaz para o objetivo proposto, obtendo melhores resultados quando aplicados valores diferentes dos valores padrão para *E-value*, filtro de regiões de baixa complexidade, tamanho inicial de palavra e matriz de substituição.

Palavras-chave: Bioinformática, Proteínas Mosaicas, *Trypanosoma cruzi*, Proteoma, BLAST.

A METHODOLOGY FOR IDENTIFICATION OF COMPONENT MODULES OF *Trypanosoma cruzi* MOSAIC PROTEIN SEQUENCES FROM THE PARASITE'S PROTEOME USING BLAST

ABSTRACT

This paper proposed a methodology for the identifying component modules of *Trypanosoma cruzi* mosaic proteins sequences using BLAST. For the development of the methodology, MASP protein family was used and a set of default BLAST parameter values was initially applied. Afterwards, different combinations of parameter values were studied for result comparison, including those indicated in literature. The developed methodology proved to be efficient for the proposed objective, obtaining better results when non-default parameter values for *E-value*, low complexity region filter, initial word size and substitution were applied.

Keywords: Bioinformatics, Mosaic Proteins, *Trypanosoma cruzi*, Proteome, BLAST.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

1. INTRODUÇÃO

1.1. Contextualização e Motivação

Genoma é toda a informação genética carregada por uma célula ou organismo. O avanço tecnológico vivido nas últimas décadas possibilitou o Projeto Genoma, cujos principais objetivos são criar mapas físicos de alta resolução, sequenciar todo o Ácido Desoxirribonucléico (DNA) do genoma, criar e depositar as informações obtidas em um banco de dados e aperfeiçoar as técnicas moleculares de modo a melhorar a qualidade dos estudos. Por se tratar de bancos de dados muito extensos, a utilização de plataformas computacionais eficientes para análise dos dados e interpretação dos resultados é indispensável.

Dados biológicos advindos do conhecimento genômico são relativamente complexos em comparação aos provenientes de outras áreas científicas, dada a sua diversidade e seu inter-relacionamento. A partir do conhecimento fundamental do genoma, a comunidade científica objetiva compreender o conjunto de peças que atuam no funcionamento complexo de todo o organismo. Porém, no momento, isso somente é possível por partes. Busca-se entender as estruturas moleculares das proteínas e as interações entre elas e com as demais moléculas biológicas (DNA, carboidratos, lipídios). Também se deseja obter conhecimento sobre as diversas vias metabólicas celulares e o papel da variabilidade genética representada pelas várias formas de cada proteína. Toda essa informação disponibilizada pela ciência genômica só é possível de ser organizada, analisada e interpretada com o apoio da informática.

A bioinformática é imprescindível para a manipulação dos dados biológicos. Ela pode ser definida como uma modalidade que abrange todos os aspectos de aquisição, processamento, armazenamento, distribuição, análise e interpretação da informação biológica. Através da combinação de procedimentos e técnicas da matemática, estatística e ciência da computação são elaboradas várias ferramentas que auxiliam a compreender o significado biológico representado nos dados genômicos. Além disso, através da criação de bancos de dados com as informações já processadas, acelera a investigação em outras áreas como a medicina, a biotecnologia e a agronomia.

O *Trypanosoma cruzi* (*T. cruzi*) é um protozoário parasita causador da doença de Chagas, uma doença incurável e debilitante que afeta milhões de pessoas na América

Latina. O sequênciamento do genoma do *T. cruzi* permitiu o início de análises das sequências de aminoácidos e nucleotídeos derivadas a fim de identificar diversos dados estruturais para estudos funcionais posteriores. Entre esses dados estruturais estão os dados sobre módulos encontrados em determinadas proteínas, formadas pelo rearranjo genético de tais módulos e conhecidas como proteínas mosaicas. Um módulo pode ser definido como um conjunto de aminoácidos invariáveis ou altamente conservados usado repetidamente como "blocos de construção" em diversas proteínas. Cada módulo pode apresentar uma função enzimática, sinalizadora, regulatória ou estrutural diferente, o que faz com que a arquitetura modular de proteínas permita a evolução dessas com funções complexas e altamente especializadas.

O *T. cruzi* infecta o hospedeiro por meio do desenvolvimento de estratégias adaptativas que envolvem diferentes famílias de proteínas de superfície, entre elas a família de proteínas de superfície associadas a mucinas (MASP – *Mucin Associated Surface Protein*) em estudo. Um número de diferentes módulos é encontrado em proteínas desse tipo, no entanto não existem na literatura estudos que verificam a estrutura mosaica das proteínas da família MASP do *T. cruzi*.

Em virtude da grande variabilidade clínica e epidemiológica da doença de Chagas e das características genéticas da população do *T. cruzi*, o tratamento para a doença tem se limitado a medicamentos usados desde o final da década de 1960, com alta taxa de efeitos colaterais e eficácia variável durante a fase crônica da doença. Por esta razão, a identificação dos módulos constituintes das proteínas de famílias protéicas necessárias à sobrevida e à patogenicidade do parasito por meio da análise de seu proteoma abre caminho para a busca de novas estratégias terapêuticas e para a identificação de novos biomarcadores importantes para o desenvolvimento de novas drogas e prognóstico clínico da doença de Chagas.

1.2. Objetivos

O objetivo geral deste trabalho foi o desenvolvimento de uma metodologia para identificação dos módulos formadores de sequências de proteínas mosaicas e, dada uma família de proteínas do *T. cruzi*, a verificação de se essas apresentam estrutura mosaica, ou seja, se são formadas por módulos que se repetem em diferentes proteínas da família. Para o desenvolvimento da metodologia foi utilizada a família MASP de proteínas do *T. cruzi*.

Esse trabalho apresenta os seguintes objetivos específicos:

- Desenvolvimento de um algoritmo para identificação de módulos comuns a várias proteínas de uma família protéica;
- Aplicação do algoritmo desenvolvido para identificação dos módulos presentes nas proteínas da família MASP em estudo;
- Para cada módulo encontrado, identificação das sequências da família de proteínas em questão que o apresentam e mapeamento de sua posição em tais sequências;
- Para cada sequência da família de proteínas MASP, identificação e mapeamento das posições dos módulos que ela apresenta;
- Análise e discussão dos resultados encontrados e da metodologia desenvolvida.

1.3. Estrutura do Trabalho

Os capítulos subsequentes desta monografia estão assim organizados:

- o segundo capítulo explana os conceitos da biologia tomados como necessários para o melhor entendimento deste trabalho e dos ganhos obtidos;
- o terceiro capítulo apresenta conceitos e técnicas da bioinformática utilizadas durante o desenvolvimento deste trabalho;
- o quarto capítulo expõe a classificação da pesquisa e a metodologia utilizada no desenvolvimento do trabalho.
- o quinto capítulo apresenta os resultados obtidos e a discussão destes;
- o sexto capítulo contém a conclusão e propostas de continuidade do trabalho.

2. REFERENCIAL BIOLÓGICO

2.1. O *Trypanosoma cruzi*

O *Trypanosoma cruzi* (*T. cruzi*) pertence à ordem Kinetoplatida, que abrange as famílias Bodonidae Hollande e Trypanosomatidae Kent. Nestas famílias encontram-se flagelados de um ou dois flagelos que se originam de uma abertura conhecida como bolsa flagelar, e normalmente contêm uma estrutura paraflagelar e uma estrutura proeminente, conhecida como cinetoplasto, que corresponde a uma condensação de DNA localizado no interior de uma mitocôndria única e ramificada por todo seu corpo. A família Trypanosomatidae também engloba os seguintes gêneros importantes: *Blastocrithidia, Crithidia, Endotrypanum, Herpetomonas, Leishmania, Leptomonas, Phytomonas* e *Trypanosoma* (SOUZA, 2008).

Por incluir uma série de espécies causadoras de doenças humanas como, por exemplo, o *T. cruzi* (Figura 2.1), agente da doença de Chagas, o gênero *Trypanosoma* é um dos mais importantes dentro da família Trypanosomatidae. O gênero foi dividido em dois grupos com base no comportamento do parasito nos seus hospedeiros, principalmente no vetor. O primeiro chamado Stercoraria, inclui tripanossomos que se desenvolvem no tubo digestivo do vetor, progredindo no sentido da porção intestinal com liberação de formas infectantes pelas fezes. Aqui se tem o *Trypanosoma cruzi* e o *Trypanosoma lewisi*. O segundo, chamado de Salivaria, inclui tripanossomos que se desenvolvem inicialmente no tubo digestivo e que posteriormente atravessam o epitélio digestivo e atingem as glândulas salivares onde podemos encontrar as formas infectantes que são inoculadas mecanicamente. Neste grupo encontramos o *T. brucei*, *T. congolense* e *T. rangeli* (SOUZA, 2008).

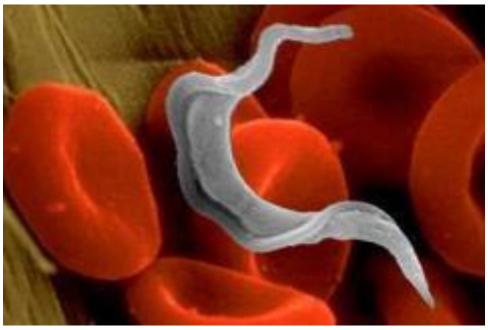


Figura 2.1 – O *Trypanosoma cruzi* rodeado de glóbulos vermelhos
Fonte: Levy (2006)

A doença de Chagas, também conhecida como tripanossomíase americana, foi descoberta em 1909, em Lassance, MG, por Carlos Chagas, um cientista brasileiro que lá se encontrava trabalhando no combate à malária que atingia aquela região por ocasião da construção de uma ferrovia (NEVES et al., 2005).

O *T. cruzi* é um parasita muito antigo, remontando há mais de 150 milhões de anos sua presença no planeta. É um protozoário largamente distribuído na natureza. Sua circulação ocorre entre insetos vetores e mamíferos silvestres. Dotado de grande diversidade genética, de modo geral os clones e populações estudados têm modernamente sido agrupados, mediante estudos de perfil molecular e izoenzimático, em três maiores grupos ou linhagens, denominados GI e GIII (grupos basicamente de origem silvestre, naturalmente vinculado a marsupiais) e Z2 (encontrado na América do Sul, naturalmente ligados a primatas) (DIAS, 2006).

O *T. cruzi* para infectar e se adaptar ao hospedeiro vertebrado explora estratégias evolucionárias de invasão das células alvo e evasão do sistema imunológico (ANDRADE & ANDREWS, 2005). O parasito utiliza diferentes famílias de proteínas de superfície para seu processo de invasão, evasão e (FRASCH, 2000). Uma estratégia chave é a geração e apresentação de antígenos de superfície variáveis (KAHN et al., 1999). O parasito pode tirar vantagem dessa estratégia para aderir a diferentes moléculas na membrana celular e matriz extracelular da célula hospedeira (FRASCH, 2000).

2.2. Expressão Genômica

Todo organismo possui um genoma que contém a informação biológica necessária para construir e manter um exemplar vivo. O genoma é um deposito de informação biológica, mas sozinho é incapaz de liberar tal informação para a célula. A utilização da informação biológica requer uma atividade coordenada de enzimas e outras proteínas, que participam em uma série complexa de reações bioquímicas, chamada expressão genômica (BROWN, 2002).

2.2.1. O Dogma Central da Biologia Molecular

No início da década de 1950, quando a estrutura do DNA foi determinada, tornou-se claro que a informação genética nas células estava codificada na sequência de nucleotídeos do DNA. Mesmo antes da decodificação do DNA se sabia que a informação contida nos genes de algum modo era responsável pelo direcionamento da síntese de proteínas, principais constituintes das células e determinantes não apenas de sua estrutura, mas também de seu funcionamento (ALBERTS et al., 2006).

DNA e proteínas são macromoléculas que desempenham um papel fundamental na vida de uma célula. A informação genética, armazenada no DNA como uma sequência de quatro tipos de nucleotídeos (adenina, guanina, citosina e timina), é transmitida pela replicação do mesmo. No entanto, as proteínas – e não o DNA – são responsáveis pela realização das funções vitais da célula. Deste modo, torna-se necessário que os quatro tipos de nucleotídeos sejam traduzidos para os vinte tipos de aminoácidos componentes das proteínas. Esta etapa é crucial para a expressão da informação genética (KAMOUN et al., 2006).

A informação biológica em cada gene do genoma é dividida em uma série de *exons* separadas por *introns* não-codificantes. Durante a expressão de um gene, o Ácido Ribonucléico (RNA) que é inicialmente sintetizado é uma cópia de todo o gene, incluindo tanto *introns* quanto *exons*. O processo de *splicing* remove os *introns* desse pré-RNA mensageiro (pré-RNAm) e une os *exons* para formar o RNAm que, no fim, dirige a síntese de proteínas (BROWN, 2002).

Há um mecanismo celular que realiza a transcrição do DNA para RNA, gerando um RNA a partir dos códigos do DNA, e posteriormente traduzindo para proteínas. No processo de tradução, cada grupo de três nucleotídeos (um códon) é traduzido para um aminoácido. Estes por sua vez se unem por meio de ligações peptídicas de modo a formarem uma proteína. Os quatro nucleotídeos combinados três a três produzem 64 possíveis combinações, ou seja, considerando que existem apenas vinte tipos de aminoácidos observa-se que existem alguns aminoácidos que podem ser traduzidos por mais de uma sequência de nucleotídeos. O fluxo de informação para gerar um RNA e do RNA gerar uma proteína, juntamente com o fluxo de transmissão da informação de DNA para DNA por meio da replicação (Figura 2.2), formam o dogma central da biologia molecular (KANEHISA, 2000).

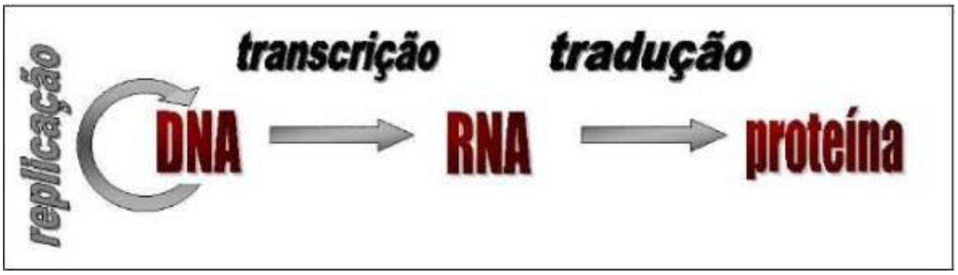


Figura 2.2 – Dogma central da biologia molecular

Durante a evolução, sequências genéticas passam por mudança espontânea ou mutação. As gerações seguintes terão, desse modo, uma sequência diferente. A maioria das mudanças é prejudicial ao organismo e nunca são observadas, e a maior parte das que não são prejudiciais não tem efeito algum. Desse modo, quando proteínas são comparadas observam-se substituições que são compatíveis com a manutenção da estrutura e função. Raramente ocorre uma mudança benéfica ao organismo por prover algum tipo de vantagem biológica. Tais mudanças raras, que também são tidas como variações de sequências, são a base da mudança evolucionária. Outro tipo de variação durante a mudança evolucionária é o movimento de blocos de sequência para criar novos genes e proteínas. (MOUNT, 2004).

Os parasitas da família do *Trypanosoma cruzi* desenvolveram mecanismos próprios de funcionamento que lhes permitem escapar das defesas dos organismos que invadem e se reproduzir com rapidez. No momento de se dividir e originar outra célula idêntica, esses protozoários não seguem a estratégia de outros organismos formados por células com núcleo. Na etapa inicial de produção de proteínas, ao invés de decodificarem um gene por vez, estes lêem todos os genes de uma única vez. Neste momento, a longa molécula espiralada de DNA se espalha pela periferia do núcleo do parasita. Só depois que essa copia simultânea dos genes termina é que a mensagem de cada gene é separada e começa a produção de proteínas que vão formar seus descendentes (ZORZETTO, 2005).

2.2.2. **Proteoma**

O proteoma é o produto final da expressão genômica e engloba todas as proteínas presentes em uma célula em um dado momento (BROWN, 2002).

Uma proteína, como uma molécula de DNA, é um polímero linear não ramificado. Em proteínas, as subunidades monoméricas são chamadas aminoácidos e os polímeros resultantes, ou polipeptídios, têm comprimento raramente maior que duas mil unidades

(BROWN, 2002). O termo proteína vem do grego *proteios* e significa “a mais importante” (ALBERTS et al., 2006).

Os vinte aminoácidos mostrados na Tabela 2.1 são os vistos como especificados pelo código genético. Eles, portanto, são os aminoácidos que são ligados quando polipeptídios são montados durante a fase de síntese protéica da expressão genômica.

Tabela 2.1 – Abreviações aminoácidos

	Abreviações	
Aminoácidos	**Três Letras**	**Uma Letra**
Alanina	Ala	A
Argina	Arg	R
Aspargina	Asn	N
Ácido Aspártico	Asp	D
Cisteína	Cys	C
Ácido Glutâmico	Glu	E
Glutamina	Gln	Q
Glicina	Gly	G
Histidina	His	H
Isoleucina	Ile	I
Leucina	Leu	L
Lisina	Lys	K
Metionina	Met	M
Fenilalanina	Phe	F
Prolina	Pro	P
Serina	Ser	S
Treonina	Thr	T
Triptofano	Trp	W
Tirosina	Tyr	Y
Valina	Val	V

Proteínas são tradicionalmente vistas como tendo quatro níveis estruturais distintos. Tais níveis são hierárquicos, sendo a proteína construída estágio por estágio, com cada nível estrutural dependendo do anterior.

1. A estrutura primaria da proteína é formada pela junção de aminoácidos em um polipeptídio (Figura 2.3a).

2. A estrutura secundaria refere-se às diferentes configurações que podem ser assumidas pelo polipeptídio. A maioria dos polipeptídios é extensa o suficiente para serem dobrados em uma série de estruturas secundárias, uma após a outra ao longo da molécula (Figura 2.3b).

3. A estrutura terciária resulta da dobra dos componentes da estrutura secundaria do polipeptídio em uma configuração tridimensional (Figura 2.3c).

4. A estrutura quaternária envolve a associação de dois ou mais polipeptídios, cada um dobrado em sua estrutura terciária, em uma proteína de múltiplas subunidades. Nem todas as proteínas formam estruturas quaternárias, mas é uma característica de varias proteínas com funções complexas, incluindo várias envolvidas na expressão genômica (Figura 2.3d).

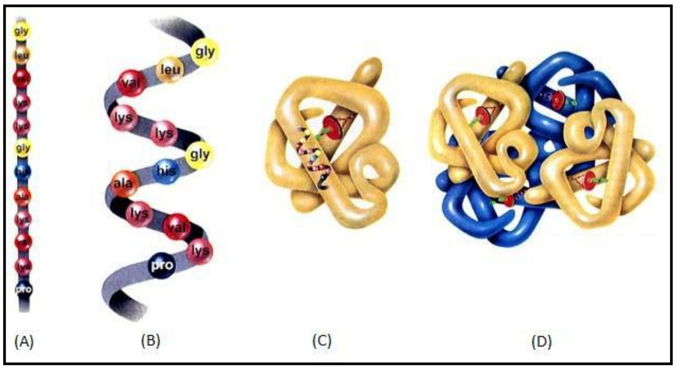


Figura 2.3 – Níveis estruturais das proteínas
Fonte: Biosciences (2008)

A partir das sequências de DNA dos genes, pode-se deduzir a sequência de aminoácidos das proteínas por eles codificadas. Essa informação é de grande importância, já que a sequência de aminoácidos de uma proteína (ou estrutura primária) é a característica primordial que define sua forma e função. Por outro lado, o sequênciamento de genes revela muito pouco sobre como as proteínas de um organismo operam individualmente ou em conjunto para exercer suas funções. Além disso, sabe-se que, após serem sintetizadas, as proteínas podem sofrer importantes modificações chamadas pós-traducionais, como glicosilações e fosforilações. Tais informações não podem ser retiradas exclusivamente da sequência dos genes, havendo necessidade de estudos diretos das proteínas. Do mesmo modo, o estudo do genoma não permite saber que proteínas estão expressas realmente em uma determinada célula em um dado momento. Dentro desse contexto, torna-se importante o estudo em larga escala das proteínas por meio de projetos de análise de proteomas (SOUSA et al., 1999).

A estratégia de escolha mais utilizada atualmente para o estudo de proteomas e que tem demonstrado ser eficiente é a combinação de eletroforese em gel de poliacrilamida bidimensional com espectrometria de massa. A eletroforese permite a separação de proteínas de um dado sistema biológico com alta resolução e reprodutibilidade, enquanto que a espectrometria de massa permite, através de alta demanda e sensibilidade, a identificação de proteínas presente em um *spot* do gel (SODRÉ et al., 2008).

A aplicação da tecnologia proteômica em tripanossomatídeos é particularmente importante para o estudo global da expressão gênica. O *T. cruzi*, assim como outros tripanossomatídeos, regula a expressão de proteínas pós-transcricionalmente através de variações na estabilidade ou na eficiência de tradução dos RNAm's. Entretanto, é notório que o conhecimento do genoma por si só, assim como os mecanismos que controlam a sua expressão não elucida todos os processos biológicos que regulam o ciclo de vida desses parasitos e tão pouco o mecanismo usado para a infecção do hospedeiro. Sendo assim, a investigação do proteoma das várias formas celulares do parasito pode fornecer informações complementar tais como modificações pós-traducionais de aminoácidos, que certamente desempenham um papel crucial na modulação da função protéica nestes parasitos (SODRÉ et al., 2008).

2.3. Sequenciamento do Genoma do *T. cruzi*

A comunidade científica em torno de *T. cruzi*, *Leishmania major* e *T. brucei* começou a discutir a possibilidade de iniciar projetos genoma destes parasitos após o início dos projetos genomas de diversos organismos no início dos anos 1990 no cenário internacional (DEGRAVE, 2008). Com o lançamento dos projetos Genoma do *T. cruzi*, *L. major* e *T. brucei* o conhecimento sobre a genética desses parasitos aumentou consideravelmente. Após alguns anos de execução, o sequênciamento completo do genoma desses parasitos foi concluído em 2005, mas antes mesmo da liberação dos dados e da conclusão do sequênciamento, esses projetos já permitiram aos cientistas identificar diversos novos alvos terapêuticos potenciais, além de fornecer dados estruturais para estudos funcionais posteriores (GUIMARÃES, 2006).

A sequência genômica do *T. cruzi* foi oficialmente publicada junto com as sequências genômicas completas de *L. major* e *T. brucei*, na revista *Science* em 2005. A montagem do genoma foi apenas parcial, devido às muitas dificuldades com sequências repetitivas e a heterozigose do clone. Assim, foram preditas 22.570 proteínas, das quais 12.570 formam pares alélicos (DEGRAVE, 2008).

Assim como outros tripanossomatídeos, esse parasito apresenta algumas características bastante peculiares em termos biológicos que reflem na função e organização de seu genoma. O *T. cruzi* apresenta uma significativa variação na quantidade de DNA nuclear e no número de cromossomos entre diferentes isolados do parasito por apresentar um grande polimorfismo na sua constituição genética. Os genes de *T. cruzi* e dos outros tripanossomatídeos não são em geral interrompidos pelas sequências de inserção, diferentemente da maioria dos organismos eucarióticos (GOLDENBERG, 2008).

Pelo menos 50% de todo o genoma do *T. cruzi* é constituído por sequências repetitivas do DNA e são formadas principalmente pelas famílias de genes que compõem as proteínas de superfície. Estes totalizam 18% dos genes codificadores de proteínas do *T. cruzi*. A família MASP (*Mucin-Associated Surface Protein*) do *T. cruzi*, em estudo nesse trabalho, é uma família de proteínas de superfície associadas à mucina. Ela contém 1377 membros, o que corresponde a aproximadamente 6% do genoma diplóide do *T. cruzi*, e é caracterizada por regiões centrais altamente variáveis (EL-SAYED et al., 2005).

Sugere que proteínas da família MASP podem conter extensivas modificações após o processo de tradução por apresentar um baixo número de peptídeos detectados por abordagens proteômicas. Genes da família MASP podem ser expressos em estágios intermediários não representados nos dados do proteoma ou podem ser expressos de modo mutuamente exclusivo (EL-SAYED et al., 2005).

Ainda existe um grande campo a ser explanado e pesquisado em relação a regulação da expressão gênica em tripanossomatídeos. Com a determinação da sequência genômica do *T. cruzi*, *T. brucei* e *Leishmania major*, que são de relevância para a saúde humana, o uso de ferramentas de análise genômica e pós-genômica e o avanço dos estudos voltados para epigenética, novos mecanismos devem ser evidenciados (GOLDENBERG, 2008).

2.4. Proteínas Mosaicas

Segundo Avery et al. (1993), proteínas mosaicas são um grupo de proteínas que podem ser formadas por um ou mais tipos de uma variedade de diferentes módulos estruturais e que possuem uma extensão diversa de funções.

De acordo com Gaboriaud et al., (1998), a análise comparativa de sequências protéicas tem revelado que muitas proteínas extracelulares são constituídas por um repertório limitado de padrões ou módulos de sequência. Tais proteínas, chamadas proteínas mosaicas, podem então ser descritas como a justaposição linear de módulos contíguos e/ou domínios. Módulos podem ser definidos como subconjuntos de domínios usados repetidamente como “blocos de construção” em diversas proteínas e provavelmente têm aparecido por meio da “mistura de genes” (HEGYI & BORK, 1997). Várias proteínas mosaicas possuem papel essencial na série de reações químicas da biologia extracelular (GABORIAUD, 1998).

Conforme Kolkman & Stemmer (2001), muitas proteínas são compostas por um número de domínios discretos, que frequentemente estão envolvidos em funções específicas que contribuem para a atividade geral da proteína. Uma análise dos genes codificadores de proteínas mosaicas revela uma forte correlação entre organização de domínio e estrutura *intron-exon*. Em outras palavras, cada domínio tende a estar codificado por um ou uma combinação de *exons* que inicia e termina no mesmo quadro de *splice*. Proteínas mosaicas aparentam ser criadas pela junção de múltiplos domínios por meio do embaralhamento de *exons*.

Os domínios encontrados em proteínas mosaicas são evolucionariamente móveis, o que significa que eles se espalharam durante a evolução e agora ocorrem em proteínas que antes não estariam relacionadas (DOOLITTLE, 1995). A maioria das proteínas mosaicas são extracelulares ou constituem as partes extracelulares de proteínas ligadas a membrana e por isso foi proposto que proteínas mosaicas desempenharam um importante papel na evolução da multicelularidade (PATTHY, 1991).

3. TÉCNICAS E FERRAMENTAS

A maioria das ferramentas computacionais utilizadas em bioinformática se baseia em busca por similaridade entre as sequências nucleotídicas ou de aminoácidos. Sequências similares provavelmente possuem uma história evolutiva comum e compartilham funções, de modo que ferramentas baseadas em busca por similaridade podem ser utilizadas para inferir uma função.

3.1. Alinhamento por Pares de Sequências

Alinhamento de sequências é o procedimento de se comparar duas (alinhamento por pares) ou mais (alinhamento múltiplo) sequências de ácidos nucléicos (DNA e RNA) ou proteína por meio da busca de uma série de caracteres individuais ou padrões de caracteres que estão na mesma ordem nas sequências.

O alinhamento de sequências busca possibilitar ao pesquisador determinar se duas sequências apresentam similaridade suficiente tal que uma inferência sobre homologia possa ser justificada. Homologia significa que duas ou mais sequências têm um ancestral comum. Já similaridade é uma medida da qualidade do alinhamento entre duas sequências com base em algum critério. A similaridade não se refere a nenhum processo histórico, sendo apenas uma comparação das sequências com algum método podendo ser definida, por exemplo, contando posições idênticas entre duas sequências.

De acordo com Prosdocimi et al. (2002), existem vários programas de computador que realizam alinhamentos de sequências e a grande maioria deles podem ser utilizadas *on-line*, sem a necessidade de instalação. Os softwares mais utilizados para alinhamentos de sequências são:

- ClustalW – Versão web de um dos programas de alinhamentos múltiplos globais mais utilizados (Clustal).
- Multalin – Programa de alinhamento múltiplo global.
- FASTA – Precursor dos programas de alinhamento. Promove serviço de busca em banco de dados de ácidos nucléicos e proteínas realizando alinhamento local.
- BLAST, BLAST2sequences – Programa de alinhamento mais utilizado no mundo. Realiza busca por sequências homólogas em banco de dados de ácidos nucléicos e proteínas. O programa *BLAST2sequences* consiste no algoritmo BLAST para alinhamento de duas sequências.

Um par de sequências é alinhado escrevendo-as em duas linhas e fazendo com que caracteres idênticos ou similares sejam posicionados na mesma coluna, e caracteres não-idênticos podem ser colocados tanto na mesma coluna como casamento sem êxito quanto em frente a um *gap* (lacuna) da outra sequência. Em um alinhamento ótimo, caracteres não-idênticos e *gaps* são posicionados de modo a trazer o maior número possível de caracteres idênticos ou similares para colunas. Sequências que podem ser alinhadas de imediato dessa maneira são chamadas similares (MOUNT, 2004).

Uma vez que *gaps* são permitidos, o número de possíveis alinhamentos se torna exponencial ao tamanho das sequências, logo não se pode experimentar todos. A introdução de *gaps* também pode levar a alinhamentos sem sentido. Por isso é necessário distinguir entre alinhamentos que ocorreram devido a homologia e alinhamentos que se espera acontecerem ao acaso.

Há duas formas de alinhamento por pares: global e local. No alinhamento global, é feita uma tentativa de alinhar toda a extensão das sequências. Sequências que são bastante semelhantes e que possuem aproximadamente o mesmo tamanho são candidatas ao alinhamento global. No alinhamento local, são alinhadas extensões de sequência com alta densidade de casamentos, gerando desse modo uma ou mais ilhas de casamentos ou sub-alinhamentos nas sequências alinhadas. Alinhamentos locais são mais apropriados para alinhar sequências que são semelhantes apenas em partes de suas extensões, sequências com tamanhos diferentes ou sequências que compartilham um domínio ou região conservada (MOUNT, 2004). A Figura 3.1 exemplifica a diferença entre os dois tipos de alinhamento.

```
SIGRTAQGKEFGVTKARWPAKS
|    |   |||    | |
SAT-VASHIEFGIQVNRYPQVN
```

(a)

```
--------TGKT---------
         |||
--------AGKT---------
```

(b)

Figura 3.1 – Exemplo de um alinhamento global (a) e um alinhamento local em (b)

3.2. Matrizes de Substituição ou *Score*

Matrizes de substituição, também chamadas de matrizes de *score*, são tabelas bidimensionais (i, j) que contêm valores que demonstram a probabilidade de que o aminoácido na posição i sofra mutação para o aminoácido da posição j, quaisquer que sejam as duas sequências envolvidas. Tais matrizes são construídas pelo estudo de diversas amostras de alinhamentos por pares. Se a amostra é grande o suficiente para ser estatisticamente significante, as matrizes devem refletir as verdadeiras possibilidades de mutações que ocorrem ao longo de certo período de mutação (NCBI/Education, 2002 apud CARVALHO, 2002).

Para que seja possível estabelecer um alinhamento, buscando similaridade, entre duas sequências é preciso que um esquema de *score* (pontuação) seja estabelecido. A abordagem utilizada em comparação de sequência de proteínas é a utilização dessas matrizes de substituição, sendo que as duas mais utilizadas são conhecidas como PAM (*Point Accepted Mutation*) e BLOSUM (*Blocks Substitution Matrix*) (HIGA, 2001).

A escolha da matriz de *score* pode ter um efeito considerável sobre os resultados de alinhamentos feitos em bancos de dados de proteínas. Sugere-se que essa escolha seja o elemento técnico mais crítico para o sucesso de uma busca. Idealmente, os valores da matriz devem refletir os fenômenos biológicos que os alinhamentos procuram mostrar, como por exemplo, no caso de motivos conservados ou correlações bem definidas entre estrutura e sequência, os números devem ser derivados de coleções de sequências contendo esses padrões desejados (GUSFIELD, 1997).

3.2.1. **PAM**

A matriz PAM foi desenvolvida em 1978, por um projeto pioneiro da Fundação Nacional de Pesquisa Biomédica dos Estados Unidos (NBRF) e liderado pela pesquisadora Margaret Oakley Dayhoff. Ela e sua equipe fizeram um estudo abrangente das frequências nas quais os aminoácidos se substituem uns pelos outros durante a evolução, baseando-se em três hipóteses: a) os eventos mutacionais são indepentendes do contexto; b) um acontecimento mutacional numa certa posição é independente dos eventos mutacionais anteriores que tiveram lugar nessa posição; c) a probabilidade de substituição de X por Y é a mesma de Y por X (PSC, 2007). Esses estudos envolveram alinhamentos globais de 1572 proteínas de 71 famílias relacionadas com, pelo menos, 85% de similaridade e, em seguida, a construção de árvore filogênica para cada uma dessas famílias. Cada árvore foi examinada pelas substituições encontradas em cada ramo (par de sequência). Essas frequências relativas foram colocadas numa matriz vinte por vinte, representando todas as possíveis combinações de substituição entre os aminoácidos. A matriz foi, então, normalizada para valores que representassem a probabilidade de que 1% dos aminoácidos viesse a experimentar uma mutação, resultando na matriz PAM1 (CARVALHO, 2002).

De maneira mais detalhada, Gusfield (1997), explica que idealmente duas sequências S_1 e S_2 são definidas como sendo divergentes por uma unidade PAM se uma série de mutação pontual aceita (sem inserções ou exclusões) tiver convertido S_1 em S_2 (e vice-versa) com uma média de um ponto por cem aminoácidos envolvidos. Isso não implica que, após cem PAMs, cada aminoácido da sequência será diferente, algumas posições podem mudar várias vezes, revertendo-se até mesmo aos aminoácidos originais, enquanto outras podem nem sofrer qualquer alteração.

As matrizes PAM e outras matrizes de substituição são geralmente apresentadas como matrizes de probabilidades logarítmicas (*log-odds*). Isso porque cada *score* na matriz é o logaritmo de um *odds ratio*. O *odds ratio* usado é a razão de número de vezes que um resíduo (aminoácido) A é observado em substituição ao resíduo B, dividido pelo número de vezes que se esperaria que um resíduo A substituísse o resíduo B de modo aleatório. Assim, um *score* zero significa que a frequência do par de aminoácidos no banco de dados é a mesma esperada “por acaso”; *scores* negativos designam pares de resíduos que se substituem menos frequentemente do que se esperaria “por acaso” e evidenciam o fato de as sequências não serem homólogas; *scores* positivos, por sua vez, designam substituições mais frequentemente do que se esperaria “por acaso”, e isso pode evidenciar homologia entre as sequências (PSC, 2007).

As demais matrizes da família PAM podem ser computadas multiplicando-se a matriz PAM1 por ela mesma *n* vezes e obtendo a frequência de mudanças para proteínas que tenham divergido *n*%, originando uma nova matriz PAM*n*. Dessa forma, se a PAM1 for multiplicada por ela mesma trinta vezes, obtém–se a matriz PAM30 (CARVALHO, 2002).

De acordo com Baxevanis & Ouellette (2001), ao alinhar duas sequências, espera-se que elas compartilhem aproximadamente 20% de identidade, dado que esse valor está no limite para se detectar uma similaridade significativa. Dessa forma, as matrizes PAM200 e PAM250 têm sido largamente utilizadas para alinhamentos do sequências bastantes divergentes. Para alinhamentos de sequências com um grau maior de similaridade, recomenda-se o uso das matrizes PAM de valores mais baixos. Ou seja, uma matriz PAM, em particular, é mais eficiente para alinhar ou encontrar em um banco de dados sequências que tenham divergido pela extensão indicada por sua unidade PAM.

3.2.2. **BLOSUM**

As matrizes BLOSUM possuem uma apresentação similar às matrizes PAM (uma matriz vinte por vinte), mas seus desenvolvedores fizeram uso de uma estratégia diferente e de um conjunto muito maior de dados para estimar as frequências-alvo (CARVALHO, 2002). Os valores das matrizes foram baseados na observação direta de substituição de aminoácidos de um conjunto de aproximadamente dois mil padrões conservados de aminoácidos, chamados blocos, que atuam como assinaturas das famílias de proteínas das quais derivam (BAXEVANIS & OUELLETTE, 2001). Esses blocos foram encontrados no banco de dados BLOCKS, o qual contém alinhamentos múltiplos locais envolvendo sequências distantemente relacionadas, ao contrário do enfoque utilizado pela matriz PAM.

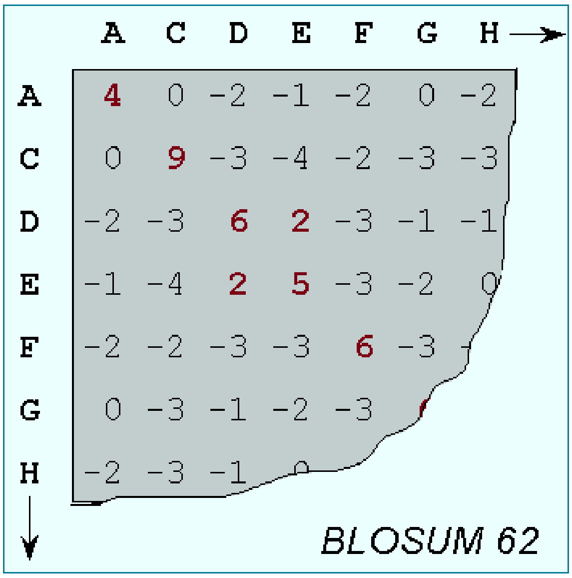
Cada coluna do alinhamento dos blocos provia um conjunto de possíveis substituições de aminoácidos e consideraram-se as hipóteses: a) os eventos mutacionais são independentes do contexto; b) um acontecimento mutacional numa certa posição é independente dos eventos mutacionais anteriores que tiverem lugar nessa posição; c) a probabilidade de substituição de um aminoácido X por Y é a mesma que Y por X; d) substituições mais comuns devem representar uma relação mais próxima entre dois aminoácidos de proteínas relacionadas e, por isso, recebem valores mais favoráveis no alinhamento; e) e contrariamente, substituições raras são menos favorecidas. Este procedimento, no entanto, poderia levar a uma representação excessiva das substituições que ocorrem na maioria dos membros de famílias de proteínas relacionadas (CARAVLHO, 2002).

Para reduzir essa interferência dos membros mais relacionados, as sequências dessas proteínas foram agrupadas em uma única sequência antes de atribuir valores aos alinhamentos das sequências dos blocos. Padrões com 62% de identidade foram novamente reagrupados para formar uma matriz de substituição chamada BLOSUM62 (Figura 3.2), e aquelas com 80% de identidade formaram outra matriz chamada BLOSUM80 e assim por diante. Ou seja, de acordo com Baxevanis & Ouellette (2001), do mesmo modo que acontece no modelo PAM, existe uma série numerada de matrizes BLOSUM, mas o número, neste caso, refere-se ao nível máximo de identidade que as sequências possam ter e ainda contribuir independentemente para o modelo.

Desse modo, para comparar sequências similares, foram construídas matrizes usando altos percentuais, enquanto baixos percentuais são mais apropriados para comparação de sequências altamente divergentes (PEARSON, 2001). De acordo com

Carvalho (2002), as matrizes BLOSUM atuam substancialmente melhor que as matrizes PAM para alinhamentos de sequências de aminoácidos.

As matrizes BLOSUM e PAM diferem não somente no modo pela qual são construídas, mas também em seu uso. Com já foi dito, matrizes PAM de baixos percentuais (PAM1, PAM20, PAM40 etc.) indicam o acontecimento de pouca mudança evolucionária. Já os altos números das matrizes BLOSUM (por exemplo, BLOSUM80), em contraste, é que indicam a mesma situação de pouca mudança evolucionária e alto grau de conservação das sequências (PEARSON, 2001).



	A	C	D	E	F	G	H
A	4	0	-2	-1	-2	0	-2
C	0	9	-3	-4	-2	-3	-3
D	-2	-3	6	2	-3	-1	-1
E	-1	-4	2	5	-3	-2	0
F	-2	-2	-3	-3	6	-3	
G	0	-3	-1	-2	-3		
H	-2	-3	-1				

Figura 3.2 Parte da matriz BLOSUM de tamanho 20 x 20
Fonte: NCBI (2008c)

Na Figura 3.2 cada entrada é a frenquência atual de ocorrência do par de aminoácidos no BLOCKS, agrupado com os demais pares de 62% de identidade, dividido pela probabilidade esperada de ocorrência. O valor esperado é calculado a partir da frequência de ocorrência de cada um dos dois aminoácidos no BLOCKS, e provê uma medida de um alinhamento aleatório dos dois aminoácidos. Um *score* zero significa que a frequência do par de aminoácidos no banco de dados é a mesma esperada "por acaso"; um valor positivo mostra que o par foi encontrado mais frequentemente que "por acaso"; e um *score* negativo significa que o par foi encontrado menos frequentemente que "por acaso".

3.3. Basic Local Alignment Search Tool

O BLAST (*Basic Local Alignment Search Tool*), uma das principais ferramentas utilizadas na bioinformática, utiliza um método heurístico que se baseia na determinação de trechos de similaridade local por meio da comparação de sequências protéicas ou de ácidos nucléicos contra sequências armazenadas em uma base de dados, calculando simultaneamente a significância estatística para os resultados obtidos após estas comparações. Essa ferramenta pode ser utilizada para inferência de relações funcionais e evolutivas de varias sequências, assim como para auxiliar na identificação de membros de uma família gênica.

3.3.1. **O Algoritmo do BLAST**

O algoritmo do BLAST aumenta a velocidade do alinhamento de sequências buscando primeiro por palavras ou *k*-tuplas comuns à sequência buscada (*query*) e a cada sequência de um banco de dados. A busca é delimitada às palavras mais significativas. Em proteínas, a significância é determinada pela avaliação dos casamentos de palavras usando pontuações da matriz de substituição de aminoácidos. No algoritmo BLAST, o tamanho da palavra é, por padrão, três para proteínas e 11 para ácidos nucléicos. Esses tamanhos são o mínimo necessário para alcançar uma pontuação de palavra alta o suficiente para ser significativa, mas não tão elevado que padrões curtos porém significativos sejam perdidos (MOUNT, 2004).

Segundo Mount (2004), os passos do algoritmo do BLAST para alinhar uma sequência *query* (de consulta) com as sequências de um banco de dados de proteínas incluem o seguinte:

1) A sequência *query* é opcionalmente filtrada para remover regiões de baixa complexidade que não são úteis para a produção de alinhamentos de sequência significativos.
2) Uma lista de palavras (*words*) de tamanho padrão três na sequência *query* de proteína é montada começando com as posições 1, 2 e 3; então 2, 3 e 4, etc., até que as três últimas posições disponíveis na sequência sejam alcançadas (Figura 3.3).

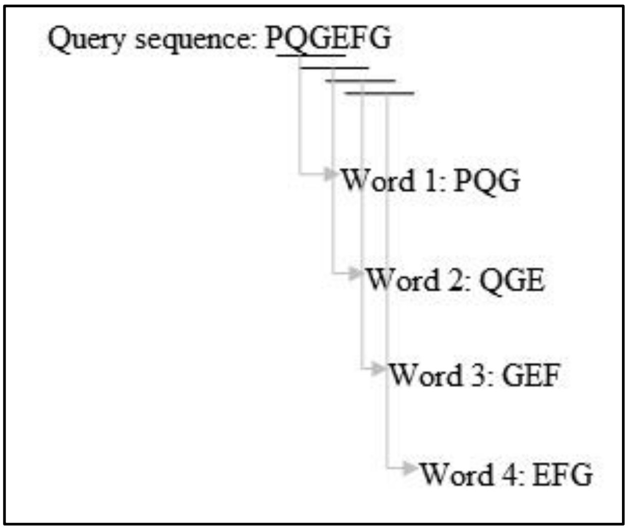


Figura 3.3 – Lista de palavras montada a partir da sequência *query*

3) Usando valores de uma matriz de pontuação como a BLOSUM62, as palavras da sequência *query* são avaliadas buscando um casamento exato com uma palavra em qualquer sequência do banco de dados. As palavras *query* também são analisadas em busca de casamento com palavras contendo qualquer outra combinação de três aminoácidos, com o objetivo de se criar uma lista de possíveis casamentos para cada palavra *query*. Há um total de 20^3 possíveis pontuações de casamento para uma dada posição na sequência, considerando os vinte diferentes aminoácidos. Por exemplo, supondo que a palavra de três letras PQG ocorre em uma sequência *query*. A probabilidade de um casamento com ela mesma é encontrada na matriz BLOSUM62 como a pontuação de um casamento P-P, somada à pontuação de um casamento Q-Q e à pontuação de um casamento G-G = 7 + 5 + 6 = 18. Essas pontuações são somadas porque a matriz BLOSUM62 é feita dos logaritmos das probabilidades de encontrar um casamento em sequências. De modo semelhante, casamentos de PQG com PEG teria pontuação 15, com PRG, 14, com PSG, 13 e com PQA, 12.
4) Uma pontuação de corte chamada limiar de pontuação de palavra (*T*) é selecionado para reduzir o número de possíveis casamentos com PQG para apenas os mais significativos. Por exemplo, se a pontuação de corte *T* é 13, apenas as palavras com pontuação maior ou igual que 13 são mantidas. No exemplo acima, a lista de possíveis casamentos com PQG incluirá PEG (15), mas não PQA (12). A lista de possíveis casamentos da palavra *query* é desse

modo reduzida dos 20³ correspondentes a todas as possibilidades a apenas as de maior pontuação (Figura 3.4).

5) O procedimento acima é repetido para cada palavra de três letras na sequência *query*.
6) As palavras de alta pontuação restantes que constituem possíveis casamentos para cada palavra de três letras na sequência *query* são organizadas em uma árvore de busca eficiente para que sejam comparadas rapidamente às sequências do banco de dados.

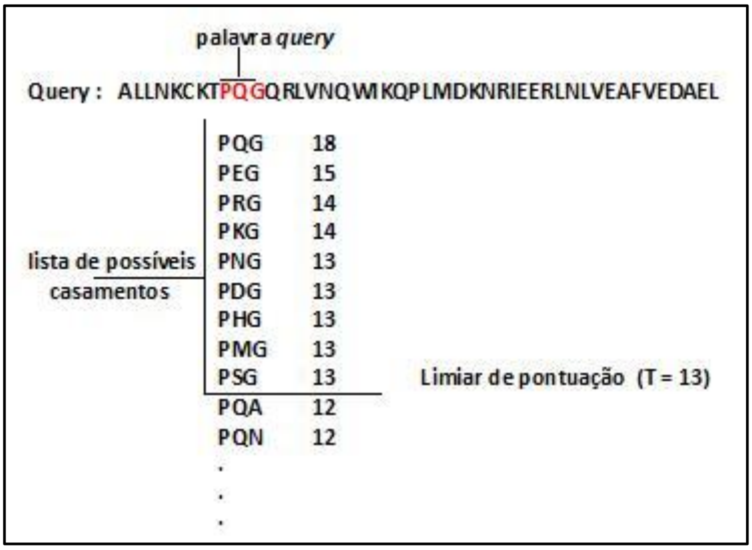


Figura 3.4 – Lista de possíveis casamentos para a palavra *query* utilizando *scores* da matriz BLOSUM62

7) Cada sequência do banco de dados é consultada em busca de um casamento exato com uma das palavras da lista de possíveis casamentos relativa a cada palavra *query* definida no passo quatro. Se um casamento é encontrado, ele é usado como semente de um possível alinhamento sem *gaps* entre a sequência *query* e as sequências do banco de dados.
8) (a) No método BLAST original, é feita uma tentativa para estender um alinhamento a partir de palavras casadas em cada direção ao longo das sequências, continuando enquanto a pontuação permanecer aumentando. O processo de extensão em cada direção é interrompido quando a pontuação acumulada para de aumentar e acaba de começar a cair um pouco abaixo da melhor pontuação encontrada para extensões mais curtas. Nesse ponto, um

trecho maior de sequência (chamada de HSP – *high-scoring segment pair*), que possui uma pontuação maior que a palavra original, pode ter sido encontrado (Figura 3.5).

Figura 3.5 – Extensão do alinhamento de uma palavra *query* definida como semente de um alinhamento

(b) Na versão produzida pelo *National Center for Biotechnology Information* (NCBI) chamada BLAST2 ou BLAST com *gaps*, é usado um método diferente e muito mais eficiente (ALTSCHUL et al., 1997). O método começa montando uma lista de palavras casadas, como nos passos um e quatro acima, com a exceção de que é usado um valor mais baixo de *T* – a pontuação de corte de palavra – como 11 no exemplo acima da palavra PQG, para manter o mesmo nível de sensibilidade de detecção de similaridade de sequências. Essa mudança resulta em uma lista mais longa de palavras e em casamentos com palavras de pontuação mais baixa nas sequências do banco de dados.

9) O próximo passo é determinar se cada pontuação de HSP encontrada por um dos métodos acima tem valor maior que uma pontuação de corte *S*. Um valor apropriado para *S* é determinado empiricamente por meio do exame da faixa de pontuações encontradas pela comparação de sequências aleatórias e pela escolha do valor significativamente maior. Os HSPs casados em todo o banco de dados são identificados e listados.

10) A seguir o BLAST determina a significância estatística de cada pontuação de HSP. É calculada a probabilidade de duas sequências aleatórias atingirem a pontuação (*score*) de HSP.

11) Alinhamentos locais são mostrados para a sequência *query* com cada sequência casada no banco de dados. Versões iniciais do BLAST produziam apenas alinhamentos sem *gaps* que incluíam o HSP inicialmente encontrado. Se dois HSPs eram encontrados, dois alinhamentos separados eram produzidos porque

as duas regiões não podiam ser alinhadas sem *gaps*. O BLAST2 produz um único alinhamento com *gaps* que inclui todas as regiões HSP inicialmente encontradas. A pontuação do alinhamento é obtida e o *expect value* (*E-value*) ou valor esperado *E* (Seção 3.3.3) para aquela pontuação é calculado usando parâmetros estatísticos para alinhamentos com *gaps* que utilizam a mesma combinação de matriz de substituição usada na busca de similaridade.

12) Quando o *E-value* para a pontuação do alinhamento local da sequência *query* com a sequência do banco de dados satisfaz o valor de limiar (que pode ser alterado pelo usuário), o casamento com a sequência do banco de dados é reportado. Os resultados da busca são mostrados como uma lista de casamentos ordenados pela pontuação do alinhamento e valor de *E* seguida pelos alinhamentos de sequências.

3.3.2. **Os Programas BLAST**

O BLAST é basicamente um conjunto de programas que buscam em bancos de dados de sequências por similaridades estatisticamente significativas. Esta busca precisa de vários passos e parâmetros de controle. Os cinco programas tradicionais do BLAST são: BLASTN, BLASTP, BLASTX, TBLASTN e TBLASTX. Os quatro últimos realizam comparação de sequências protéicas, enquanto o BLASTN trabalha com comparação de sequências nucleotídicas (KORF et al., 2003). Neste trabalho foi utilizado o BLASTP.

- BLASTN – Tem como entrada uma sequência de nucleotídeos e a compara com um banco de dados de nucleotídeos.
- BLASTP – Tem como entrada uma sequência de aminoácidos e a compara com um banco de dados de proteínas. Esse programa é muito utilizado quando se tem uma proteína e deseja-se saber se existem, em outros organismos, proteínas similares. É tipicamente utilizado para identificação de regiões comuns entre proteínas e coleta de proteínas relacionadas para análise filogenética.
- BLASTX - Compara uma sequência de nucleotídeos traduzidos em proteína contra um banco de dados de proteínas.
- TBLASTN – Compara uma sequência de aminoácidos contra um banco de dados de nucleotídeos traduzidos em proteínas.
- TBLASTX – Compara uma sequência de nucleotídeos traduzidos em proteína contra um banco de dados de sequências de nucleotídeos traduzidos em proteína.

3.3.3. **Parâmetros do BLAST**

O algoritmo do BLAST contém uma série de parâmetros que controlam o alinhamento, muitos dos quais possuem valores padrão e não precisam ser explicitamente determinados. Os parâmetros utilizados neste trabalho e tidos como mais relevantes, além das matrizes de substituição já descritas na Seção 3.2, são detalhados abaixo. Uma lista de todos os parâmetros pode ser encontrada em NCBI (2008a).

- ***E-value* (-e)**

É útil, particularmente quando buscando grandes bancos de dados, saber a probabilidade de um alinhamento ocorrer por acaso. O BLAST fornece uma medida disso com o *E-value* que ele provê para cada alinhamento. O *E-value* indica a validade do alinhamento: quanto menor, mais provável de ser um bom alinhamento e representar uma similaridade real ao invés de um alinhamento aleatório (MAYER, 2008). Um alinhamento com valor de *E-value* de 1e-63, por exemplo, indica que, pelo menos aproximadamente, a probabilidade de um alinhamento tão bom ou melhor que o primeiro ocorrer ao acaso é mínima (1e-63) (CLARK, 2006). Por padrão, o BLAST mostra alinhamentos com valores de *E-value* de no máximo 10.

- **Tamanho Inicial de Palavra (-W)**

Esse parâmetro define o tamanho inicial de palavra a ser considerado no segundo passo do algoritmo do BLAST (Seção 3.3.1). O tamanho inicial de palavras é um dos parâmetros mais importantes que dirigem a sensibilidade de buscas do BLAST. O valor padrão é três para sequências de proteínas.

- **Estatísticas Baseadas em Composição (-C)**

BLAST permite *E-values* calculados para considerar a composição de aminoácidos do banco de dados de sequências envolvido em alinhamentos reportados. Isso melhora a acurácia, reduzindo assim o número de resultados falsos positivos. As estatísticas melhoradas são alcançadas com um procedimento de escalamento que emprega um sistema de pontuação um pouco diferente para cada sequência do banco de dados. Como resultado, pontuações brutas de alinhamento em geral não corresponderão precisamente àquelas empregadas por qualquer matriz de substituição padrão. Além disso, alinhamentos idênticos podem receber pontuações diferentes, baseado nas composições das sequências que eles envolvem (INCOGEN, 2008).

- **Filtro de Regiões de Baixa Complexidade (-F)**

A filtragem pode eliminar informações estatisticamente significantes, porém biologicamente desinteressantes do relatório do BLAST, deixando apenas as regiões biologicamente mais interessantes da sequência *query* disponíveis para casamento específico contra as sequências do banco de dados. A filtragem só é aplicada à sequência *query*, não às sequências do banco de dados. O BLAST usa a filtragem SEG para BLASTP (MAYER, 2008).

3.3.4. **Relatório do BLAST**

O BLAST produz como saída um relatório contendo as informações de detalhes de similaridade dos alinhamentos (SOUSA & LIFSCHITZ, 2007). O relatório do BLAST consiste de três seções principais: (1) cabeçalho, que contém informação sobre a sequência *query* e o banco de dados buscado; (2) as descrições de linha única de cada sequência do banco de dados alinhada com a *query*; (3) os alinhamentos para cada sequência do banco de dados alinhada, podendo haver mais de um alinhamento para a mesma sequência (KORF et al., 2003).

```
 Score = 87.1 bits (193), Expect = 3e-019,   Method: Composition-based stats.
 Identities = 44/109 (40%), Positives = 64/109 (58%), Gaps = 4/109 (3%)

Query: 1   MAMMMAGRVLLVCALCVLWCGA----SVVLSAAGGFTSDRSDTAENMVLLWYPETNKTCK 56
           MAMMM GRVLLVCALCVLWCGA    +    A+    S ++  A+ ++L W+    + C
Sbjct: 1   MAMMMTGRVLLVCALCVLWCGAGGGFAKEAEASANGVSSKTTPADRIILNWHVLMKEECA 60

Query: 57  ENSTKGGILDESAFKSCMHKSMKEICEVYYNMASTDSDDPEAEEICKKY 105
             +TK G ++  A K C+  +MK IC+ +YN   ++  DPE + +C  Y
Sbjct: 61  TENTKNGTVNVPAMKRCIQVAMKGICDTFYNKTPSEIHDPEVKGMCTYY 109


 Score = 64.2 bits (140), Expect = 3e-012,   Method: Composition-based stats.
 Identities = 30/59 (50%), Positives = 40/59 (67%)

Query: 184 EVMPKNAPESNATGKEEEKRNERNHNNTKTTPLDADAMQRITSSADSDSSTAVSHATSP 242
           E  PKNAPES+  G+ E K +E+ H NTK   +++ AM+ IT +ADSD  TAV+H T P
Sbjct: 138 EGTPKNAPESDDAGRGEGKEDEKEHGNTKQKAVESAAMKHITKTADSDGRTAVTHITFP 196
```

Figura 3.6 – Alinhamento local de par de sequências em relatório do BLAST

Os alinhamentos constituem a maior parte do relatório. Na Figura 3.6 são apresentados dois alinhamentos locais entre uma *query* e uma sequência do banco de dados (*Sbjct*) utilizando o programa BLASTP (Seção 3.3.2). Um conjunto de valores que caracterizam similaridade é apresentado para cada alinhamento: a pontuação (*Score*), o *E-value* (*Expect*), o número e percentual de identidade (*Identities)* e positividade (*positives*) entre os aminoácidos. Além disso, são apresentadas as posições da região alinhada de cada sequência.

3.3.5. **Busca por Casamentos Curtos**

A busca por casamentos curtos é semelhante à busca padrão de sequências protéicas, com os parâmetros definidos de modo a otimizar a busca por sequências curtas. Uma *query* curta é mais provável de ocorrer por acaso no banco de dados. Sendo assim, aumentar o limiar definido pelo *E-value* e diminuir o tamanho da palavra é frequentemente necessário para a obtenção de resultados satisfatórios. O filtro de baixa complexidade também é removido visto que elimina porcentagens maiores de uma sequência curta, podendo até mesmo eliminar a *query*. Além disso, para buscas de sequências protéicas curtas a matriz é mudada para a PAM30, que é mais adequada para encontrar regiões curtas de alta similaridade (INCOGEN, 2008).

Quanto menor o *E-value*, ou mais próximo de zero, mais significativo é o alinhamento. No entanto, buscas com sequências curtas podem ser praticamente idênticas e apresentar um *E-value* relativamente alto. Isso se deve ao fato de que o calculo do *E-value* leva em consideração o tamanho da sequência *query* e ao fato de que sequências curtas tem uma alta probabilidade de ocorrer no banco de dados puramente ao acaso. Essa é a razão pela qual os *E-values* são definidos em valores muito altos quando executando buscas no BLAST usando sequências curtas tanto de nucleotídeos quanto de aminoácidos (MAYER, 2008).

A Tabela 3.1 apresenta um conjunto de parâmetros sugeridos pelo (NCBI, 2008b) para buscas com sequências curtas de proteínas.

Tabela 3.1 – Parâmetros do BLAST para sequências protéicas curtas

Parâmetro	**Valor Padrão**	**Valor Indicado**
E-value	10	30000
Matriz de Substituição	BLOSUM62	PAM30
Tamanho Inicial de Palavra	3	2
Estatísticas Baseadas em Composição	Desativado (F)	Desativado (F)
Filtro de Baixa Complexidade	Ativado (T)	Desativado (F)

4. METODOLOGIA

Este capítulo apresenta inicialmente o tipo de pesquisa em que se enquadra esta monografia. Em seguida são apresentados os dados e os procedimentos metodológicos utilizados ao longo do trabalho.

4.1. Tipo de Pesquisa

De acordo com Jung (2004), a pesquisa desenvolvida é aplicada, visto que se utiliza de conhecimentos e experiências adquiridos por estudiosos e profissionais da área de bioinformática e aplica técnicas já existentes na literatura a fim de gerar novos conhecimentos.

Quanto ao objetivo essa pesquisa é exploratória, uma vez que estuda um assunto atual ainda pouco examinado entre as comunidades e visa à descoberta de teorias e práticas que modificarão as existentes (JUNG, 2004 apud ZAMBALDE, 2008).

Considerando-se os procedimentos adotados, segundo Jung (2004) a pesquisa é experimental, visto que viabiliza descobertas de novos métodos e técnicas e é utilizada para obtenção de novos conhecimentos, além de requerer manipulação imparcial de dados.

4.2. Obtenção dos Dados

Os dados utilizados neste trabalho são sequências de aminoácidos formadores de proteínas da família MASP do *T. cruzi* constituintes do proteoma do parasito. As 810 sequências estudadas, organizadas em um arquivo no formato FASTA (entrada padrão para o BLAST), foram obtidas junto ao Departamento de Parasitologia do Instituto de Ciências Biológicas da Universidade Federal de Minas Gerais, que tem o *T. cruzi* como uma de suas linhas de pesquisa na subárea de Protozologia.

4.3. Procedimentos Metodológicos

Foram utilizados neste trabalho procedimentos para criação de uma estratégia que possibilitasse o desenvolvimento da metodologia proposta de identificação de módulos formadores de sequências de proteínas mosaicas do *T. cruzi*. Ainda não foi proposta na literatura uma metodologia com esse propósito. Este trabalho foi desenvolvido em paralelo com o trabalho de Lima & Souza (2008), que propõe uma metodologia de identificação de módulos de proteínas mosaicas do *T. cruzi* a partir do transcriptoma do parasito.

Este trabalho utilizou a ferramenta BLAST versão 2.2.17, em especial os programas *formatdb* e BLASTP para criação de bancos de dados a partir de arquivos contendo sequências em formato FASTA e para realização de alinhamentos entre pares de sequências protéicas, respectivamente.

Antes que pudesse ser desenvolvida a metodologia de identificação de módulos buscou-se um modo de encontrar tais módulos na família MASP do *T. cruzi* por meio do alinhamento inicial de todas as sequências contra todas e da análise dos resultados obtidos. O BLAST inicialmente foi executado para o conjunto S_1 das 810 sequências da família MASP com os valores padrão dos parâmetros gerando o alinhamento A_1. Os resultados foram filtrados segundo a estratégia detalhada na Seção 4.3.1.2, mantendo-se assim apenas os alinhamentos considerados mais significativos.

```
>Tc00.1047053506965.100
          Length = 325

Score = 55.6 bits (120), Expect = 2e-009
 Identities = 23/30 (76%), Positives = 24/30 (80%), Gaps = 1/30 (3%)

Query: 1  MAMMMTGRVLLVCALCVLW-CGGGVFAEDP 29
          MAMMMTGRVLLVCALCVLW CG  V +  P
Sbjct: 1  MAMMMTGRVLLVCALCVLWCCGAAVVSSIP 30


>Tc00.1047053506599.420
          Length = 318

Score = 56.4 bits (122), Expect = 9e-010
 Identities = 23/30 (76%), Positives = 24/30 (80%)

Query: 1  MAMMMTGRVLLVCALCVLWCGGGVFAEDPA 30
          MAMMMTGRVLLVCALCV+WCG    A D A
Sbjct: 1  MAMMMTGRVLLVCALCVVWCGLSGIAADGA 30


>Tc00.1047053506767.240
          Length = 215

Score = 61.6 bits (134), Expect = 2e-011
 Identities = 28/35 (80%), Positives = 28/35 (80%), Gaps = 5/35 (14%)

Query: 1  MAMMMTGRVLLVCALCVLWCG-GGVFA----EDPA 30
          MAMMMTGRVLLVCALCVLWCG GG FA    E PA
Sbjct: 1  MAMMMTGRVLLVCALCVLWCGVGGGFANEVVEAPA 35
```

Figura 4.1 – Alinhamento de mesma região da *query* com três diferentes sequências do banco de dados

O resultado do BLAST possibilitou verificar visualmente a existência de regiões comuns a várias sequências da família. A Figura 4.1 traz um exemplo, mostrando alinhamentos da mesma região da sequência *query* com três sequências distintas do banco de dados. Considerando a estrutura das proteínas como modular tais módulos estariam presentes nos alinhamentos filtrados. Foi então realizada a extração das regiões envolvidas em tais alinhamentos para que pudesse ser feita uma análise específica das mesmas, o que exigiu a criação de uma estratégia de corte, detalhada na Seção 4.3.1.3. As regiões extraídas compõem o conjunto S_2.

Observou-se em A_1 a existência de transitividade entre os alinhamentos filtrados, ou seja, considerando, por exemplo, regiões A, B e C de sequências, se A alinhou com B, que por sua vez alinhou com C, então A também alinhou com C. Devido a esse fato, a ocorrência de redundância é bastante evidente sendo necessária sua eliminação. Para isso, alinhou-se todas as sequências de S_2 contra elas mesmas realizando, posteriormente, a filtragem dos resultados. Com isso, criou-se uma estratégia de separação dos resultados em grupos (Seção 4.3.1.4), de modo a se ter em cada grupo apenas sequências similares entre si. A separação dos grupos de A_2 possibilitou observar que uma mesma sequência pode estar em diferentes grupos, visto que pode conter mais de uma região comum às demais sequências. Aplicou-se a estratégia de corte aos grupos para se obter uma subsequência representativa para cada um, gerando o conjunto S_3. Desse modo conseguiu-se reduzir o número de sequências presentes em mais de um grupo.

Para chegar aos possíveis módulos, notou-se a necessidade de adoção de uma abordagem iterativa dos passos descritos acima até que cada grupo contenha apenas uma sequência: um possível módulo “escolhido” como representante ao longo do processo. A cada iteração a estratégia de corte define um representante de cada grupo para ser colocado no novo arquivo gerado, sendo que os tamanhos dos grupos vão sendo reduzidos até se tornarem unitários, ou seja, até que a sequência de um grupo só alinhe com ela mesma, sendo assim uma candidata a módulo que representa todas as outras sequências dos grupos dos quais ela participou ao longo do processo. Essa é a idéia da metodologia desenvolvida neste trabalho.

Criou-se uma última estratégia para definição de quais módulos candidatos seriam considerados módulos da família de proteínas mosaicas.

O processo descrito acima para chegar à metodologia iterativa proposta utilizou valores padrão dos parâmetros do BLAST. Criada a metodologia, foi implementado em

C++ um algoritmo com os passos definidos. Executou-se o algoritmo com os valores padrão dos parâmetros e então se passou a estudar outros valores, incluindo aqueles sugeridos na literatura, comparando-se os resultados obtidos para cada combinação de parâmetros utilizados.

4.3.1. **Estratégias Utilizadas**

Nesta seção são detalhadas as estratégias e rotinas criadas para a metodologia desenvolvida.

4.3.1.1. Formatação do Relatório do BLAST

O relatório do BLAST apresenta diversas informações sobre todos os alinhamentos significativos encontrados, como descrito na Seção 3.3.4. Para a metodologia desenvolvida, as informações relevantes ao processamento são apenas a identificação das sequências envolvidas no alinhamento, as posições alinhadas em cada uma e o valor de positividade. A rotina de formatação do relatório extrai essas informações organizando-as de modo a facilitar a manipulação das mesmas, como mostra a Figura 4.2, em que cada linha representa um alinhamento e as colunas representam, respectivamente, a identificação da sequência *query* e da sequência do banco de dados, posições inicial e final do alinhamento na sequência *query* e na sequência do banco de dados e porcentagem de positividade.

Tc00.1047053408547.10	Tc00.1047053506879.10	339	351	290	302	92
Tc00.1047053408547.10	Tc00.1047053508081.30	339	351	287	299	100
Tc00.1047053508869.59	Tc00.1047053510625.54	4	21	3	20	94
Tc00.1047053508869.59	Tc00.1047053508481.20	1	19	1	19	94
Tc00.1047053507071.349	Tc00.1047053510039.30	1	28	3	30	50

Figura 4.2 – Exemplo de formatação do relatório do BLAST

4.3.1.2. Filtragem dos Resultados

Esta estratégia elimina dos resultados aqueles alinhamentos que não obedecem um determinado patamar de positividade, conservando apenas os alinhamentos significativos. Foi definido que para sequências de aminoácidos apenas os alinhamentos com 100% de positividade seriam considerados. O valor de identidade não foi utilizado visto que no processo de tradução podem ocorrer eventos que levam a geração de aminoácidos diferentes dos codificados pelas sequências de nucleotídeos, como por exemplo, deslocamento da janela de leitura de códons.

4.3.1.3. Estratégia de Corte

Esta estratégia é utilizada em duas situações: é aplicada ao alinhamento das sequências originais e aos grupos dos alinhamentos criados nas iterações. Considerando os alinhamentos de uma dada região em uma dada sequência *query*, a idéia central dessa estratégia é encontrar a subsequência da *query* que está presente na maioria dos alinhamentos, senão em todos. Para isso se estuda as posições alinhadas da *query* em cada um desses alinhamentos e se usa como posição de corte aquela que mais se repete entre as posições iniciais e finais dos alinhamentos.

Definidas as posições de corte inicial e final, essas são avaliadas para verificar se limitam uma subsequência com o tamanho mínimo estipulado, que foi definido como quatro aminoácidos. Esse tamanho foi definido por ser o tamanho mínimo de sequência que o BLAST através do BLASTP alinha. Caso a subsequência definida pelas posições de corte obedeça a essa restrição de tamanho, ela é inserida no novo conjunto S_i sendo construído. Caso contrário, as posições de corte são descartadas e busca-se um novo par que limite uma subsequência que obedeça a restrição de tamanho. Na ausência tal par de posições, a região sendo trabalhada é descartada.

4.3.1.4. Separação de Grupos

O objetivo desta estratégia é agrupar sequências similares. Considerando as sequências A, B, C, D, E e F, os alinhamentos A-B, A-C, B-C, C-D, C-F e E-F, a separação em grupos é feita da seguinte forma: como inicialmente ainda não foram criados grupos, cria-se um novo grupo G_1 em que A é cabeça; todas as sequências com as quais A se alinha são inseridas no mesmo grupo. Para o exemplo, neste ponto $G_1 = \{A, B, C\}$.

Passa-se então aos alinhamentos com a *query* B. Inicialmente se busca todos os grupos a que B pertence (G_1 no caso do exemplo). Cada alinhamento de B é analisado a fim de se verificar se a sequência com que B se alinha pertence a algum grupo ao qual B pertence. Caso a sequência não esteja em nenhum grupo de B ela é inserida no grupo em que B for cabeça; se tal grupo não existir, cria-se um novo grupo em que B é cabeça e insere-se a sequência nesse novo grupo. Para o exemplo, verifica-se que C já pertence a G_1.

Continuando o processo passa-se a analisar os alinhamentos da *query* C. O mesmo processo para B é realizado. Para o exemplo, ao analisar os alinhamentos de C verifica-se que D não pertence a G_1. Como C ainda não é cabeça de grupo, é criado um novo grupo G_2 = {C, D}. O próximo alinhamento é C-F. Como F não está em nenhum a que C pertence, mas C é cabeça do grupo G_2, F é inserido em G_2, que passa a ser G_2 = {C, D, F}.

A análise continua com os alinhamentos de E, visto que no exemplo não há alinhamentos cuja *query* é D. Neste ponto atinge-se um novo caso: E não está em nenhum grupo, portanto é criado um novo grupo G_3 = {E, F}.

A Tabela 4.1 apresenta a configuração final dos grupos para o exemplo dado.

Tabela 4.1 – Exemplo de separação em grupos

G_1	G_2	G_3
A	C	E
B	D	F
C	F	

A estratégia de corte será aplicada aos grupos e definirá uma subsequência da cabeça de cada um como representante de todo o grupo. Para o exemplo dado, as posições de corte que definem a subsequência de A representante de G_1 serão definidas com base nas posições dos alinhamentos A-B e A-C; a representante de G_2, com base nas posições de C-D e C-F; e a representante de G_3, com base nas posições do alinhamento E-F.

4.3.1.5. Definição de Módulos

Definidos os possíveis módulos ao fim do algoritmo iterativo, realiza-se o alinhamento desses com as sequências originais. São considerados módulos aqueles módulos candidatos que alinham toda a sua extensão com 100% de positividade com pelo menos 1% das sequências. Essa porcentagem mínima de alinhamentos foi definida considerando que candidatos presentes em menos de 1% das sequências da família têm grande chance de ocorrer ao acaso.

4.3.2. Valores de Parâmetros do BLAST Utilizados

A Seção 3.3.3 apresentou uma descrição dos parâmetros do BLAST utilizados neste trabalho. Buscou-se na literatura os valores de parâmetros mais indicados para se trabalhar com sequências curtas, o que foi descrito na Seção 3.3.5. Os valores utilizados para comparação de resultados da metodologia desenvolvida aplicada à família MASP do *T. cruzi* são apresentados na Tabela 4.2. A coluna “Valor NCBI” corresponde aos valores para sequências curtas de aminoácidos indicados por NCBI (2008b).

Tabela 4.2 – Valores de parâmetros utilizados para comparação de resultados

Parâmetro	Valor Padrão	Valor NCBI (2008b)	Outros Valores Testados
E-value	10	30000	1500
Matriz de Substituição (*Score*)	BLOSUM62	PAM30	BLOSUM80
Tamanho Inicial de Palavra	3	2	---
Estatísticas Baseadas em Composição	Desativado	Desativado	Ativado
Filtro de Regiões de Baixa Complexidade	Ativado: DUST	Desativado	---

O *E-value* 1500 foi selecionado para teste por ser um valor não tão alto quanto o indicado pela literatura. A matriz BLOSUM80 foi escolhida pelo fato de as matrizes BLOSUM terem sido formadas a partir da análise de um conjunto de padrões conservados de aminoácidos e serem consideradas melhores que as matrizes PAM quando utilizado o BLASP. Além disso, a BLOSUM80 foi escolhida por também ser uma matriz utilizada para sequências bastante similares e ser equivalente à PAM30 indicada na tabela por se tratar de matrizes para alinhamentos de sequências com um auto grau de similaridade, o que é o caso de sequências de mesma família.

4.3.3. **Metodologia Desenvolvida**

A metodologia desenvolvida a partir dos procedimentos adotados na Seção 4.3 é apresentada em forma de algoritmo na Figura 4.3. A entrada do algoritmo é um conjunto (S_1) de sequências formadoras de proteínas de uma família do *T. cruzi*.

O processo descrito para chegar à metodologia iterativa proposta utilizou valores padrão dos parâmetros do BLAST. Foi implementado em C++ um algoritmo com os passos definidos. Executou-se o algoritmo com os valores padrão dos parâmetros e então se passou a estudar outros valores, incluindo aqueles sugeridos na literatura, comparando-se os resultados obtidos para cada combinação de parâmetros utilizados.

```
Início do Algoritmo
     Fazer i = 1;
     Fazer u = falso;
     Enquanto u for igual a falso, fazer:
           Executar o BLASTP para obter os alinhamentos de Si com Si;
           Filtrar os alinhamentos utilizando a estratégia de filtragem;
           Se i > 1
                Separar os alinhamentos filtrados em grupos;
                Se todos os grupos forem unitários
                     Fazer u = verdadeiro;
                     Interromper o loop;
                Senão
                     Aplicar a estratégia de corte nos grupos gerando
                     o arquivo Si+1;
           Senão
                Aplicar a estratégia de corte aos alinhamentos
                filtrados gerando o arquivo Si+1;
           Fazer i = i + 1;
     Fim do Enquanto;
     Executar o BLASTP para obter os alinhamentos de Si contra S1;
     Definir os módulos utilizando a estratégia de definição de módulos;
Fim do Algoritmo.
```

Figura 4.3 – Algoritmo da metodologia proposta

5. RESULTADOS

O algoritmo proposto na Seção 4.3.3 foi implementado em C++ e executado para as sequências da família MASP do *T. cruzi* com diferentes conjuntos de valores para os parâmetros do BLAST. Para uma melhor avaliação dos resultados obtidos utilizou-se o alinhamento dos módulos com as sequências originais para mapear, para cada módulo, as posições em que se apresenta nas sequências e calcular a sua frequência de ocorrência.

Posteriormente alinhou-se o conjunto de sequências originais com o conjunto de módulos, mapeando para cada sequência as posições dos módulos que ela apresenta e calculando a frequência com que ocorrem. Tanto para o alinhamento dos módulos com as sequências originais quanto para o alinhamento usado para mapear as sequências foi utilizado um valor alto de *E-value*, para se obter alinhamentos curtos, e a matriz padrão BLOSUM62.

A Tabela 5.1 apresenta os resultados encontrados para cada combinação de parâmetros testada, onde os códigos dos parâmetros são: *-e* para *E-value*, *-M* para matriz de substituição, *-W* para tamanho de palavra, *-C* para estatística de composição e *-F* para filtro de regiões de baixa complexidade.

Os conjuntos de valores de parâmetros utilizados foram:

- **C_1:** -e 10, -M BLOSUM62, -W 3, -C T, -F T;
- **C_2:** -e 30000, -M PAM30, -W 2, -C F, -F F;
- **C_3:** -e 1500, -M PAM30, -W 2, -C F, -F F;
- **C_4:** -e 30000, -M BLOSUM80, -W 2, -C F, -F F;
- **C_5:** -e 30000, -M PAM30, -W 3, -C F, -F F;
- **C_6:** -e 30000, -M PAM30, -W 2, -C T, -F F.

Tabela 5.1 – Comparativa de resultados de combinação de valores de parâmetros do BLAST

Conjunto de Valores de Parâmetros	Total de Módulos Encontrados	Média de Ocorrências de Módulos	Média de Módulos por Sequência	Máximo de Módulos por Sequência
C_1	17	119	1	5
C_2	1182	62	44	116
C_3	896	68	39	88
C_4	181	116	17	53
C_5	414	42	6	24
C_6	2239	60	78	187

Para comparação de resultados foram utilizados o número de módulos encontrados e os valores médios de ocorrência de módulos e de ocorrência de módulos por sequência, considerando-se melhores os maiores valores. Inicialmente o código foi executado com os valores padrão do BLAST, conjunto C_1 de parâmetros, e com os valores indicados por NCBI (2008b), parâmetros C_2. Comparando os resultados verificou-se que os parâmetros indicados (C_2) apresentaram melhores resultados que os valores padrão, com um maior número de módulos encontrados e uma maior média de módulos por sequência.

A partir dessa comparação inicial, outros conjuntos de valores de parâmetros, C_3, C_4 e C_5, foram testados com objetivo de verificar se os valores indicados por NCBI (2008b) realmente levavam a produção dos melhores resultados. Para isso, cada novo conjunto variou um dos parâmetros indicados. A desativação do filtro de baixa complexidade foi adotada para todos os novos conjuntos de alinhamentos, visto que sua ativação elimina porcentagens maiores de uma sequência curta, podendo até mesmo eliminar a *query*.

Comparando os resultados de C_2 e C_3 foi observado que diminuindo o valor de *E-value* de 30000 para 1500, ocorreu uma pequena diminuição nos valores analisados. Mudando a matriz de substituição de PAM30 em $C_{2,}$ para BLOSUM80 em C_4, ocorreu uma queda no número de módulos encontrados e na média de módulos por sequência. Para um aumento do valor da palavra inicial representado na comparação dos parâmetros de C_2 e $C_{5,}$ mostrou-se que com o tamanho três encontra-se, também, um menor número de módulos e uma menor média de módulos presentes em sequências do que com o valor dois.

O melhor conjunto de valores obtido foi o C_6 após cinco iterações do algoritmo, onde a utilização do parâmetro de estatísticas baseadas em composição apresentou melhores resultados que em $C_{2,}$ onde esse parâmetro foi desativado pela indicação de NCBI (2008b) para alinhamento de sequências curtas. Comparando resultados de C_2 e $C_{6,}$ obteve-se um maior número de módulos encontrados e também uma maior média de módulos presentes em sequências.

A análise dos resultados permitiu verificar que a alteração dos parâmetros padrões para a utilização da PAM30, de um alto valor de *E-value* e da desativação do filtro de regiões de baixa complexidade a partir de C_6, apresentou melhores resultados na execução da metodologia proposta para o caso da família de proteínas MASP do *T. cruzi*.

Tabela 5.2 – Módulos de maior incidência nas sequências da família MASP

Módulo	Número de Incidências
AAAA	1695
LLLL	1441
LLLV	1250
AAVV	771
RVLL	751

O processamento da família MASP com o conjunto de valores C_6 levou seis iterações do algoritmo. Aplicada a estratégia de definição de módulos, foram definidos 2239 módulos, sendo que cada um ocorreu em média 60 vezes ao longo das sequências da família. A Tabela 5.2 apresenta os cinco módulos de maior incidência na família MASP.

O Apêndice apresenta o mapeamento para a sequência de maior incidência de módulos (187). A Figura 5.1 mostra uma visualização gráfica desse mapeamento, com os módulos em destaque, onde se verificou a presença de módulos em 78,74% da sequência.

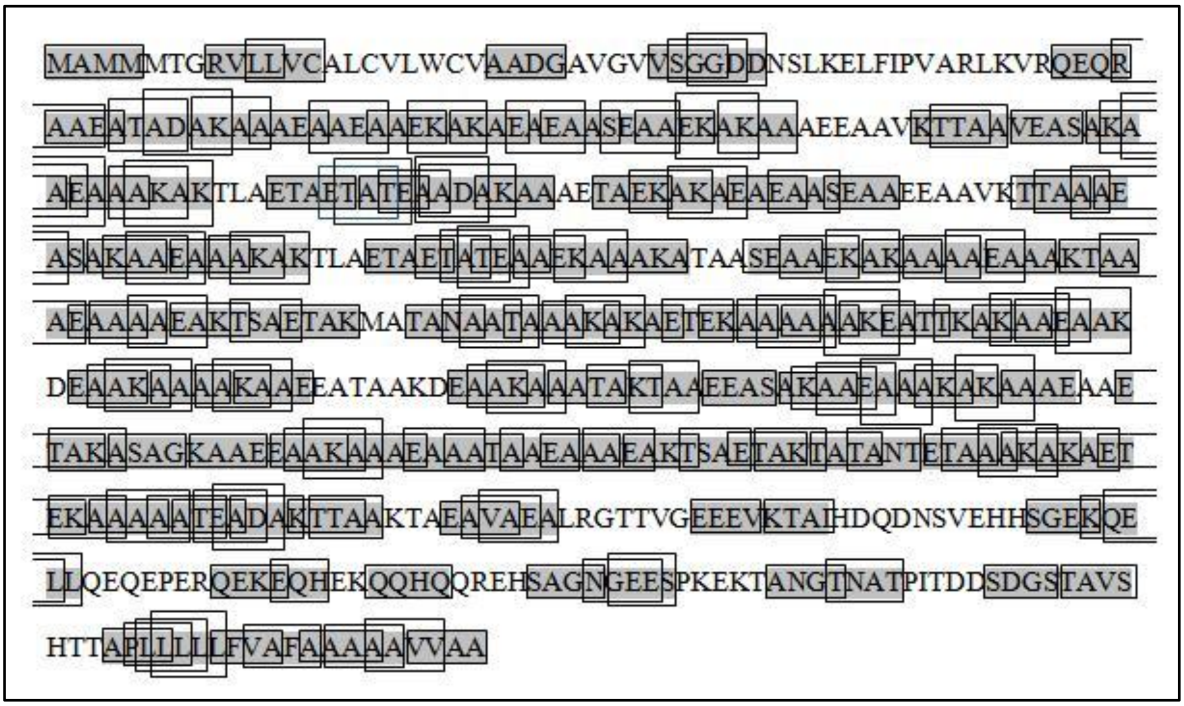


Figura 5.1 – Mapeamento de módulos na sequência Tc00. 1047053507957.200

Na Figura 5.1 pode-se observar a ocorrência de sobreposição de módulos, o que ocorre devido ao próprio BLAST relatar alinhamentos sobrepostos por alinhar uma mesma região da *query* mais de uma vez com a mesma sequência do banco de dados. A sobreposição sugere que os módulos obtidos ao fim do algoritmo não constituem necessariamente módulos individuais, havendo a possibilidade de serem combinados para formar outros módulos em um processo de refinamento dos resultados.

6. CONCLUSÃO

Este trabalho se propôs a criar uma metodologia para identificação de módulos formadores de sequências de proteínas mosaicas do *Trypanosoma cruzi* a partir do proteoma utilizando a ferramenta BLAST.

O algoritmo para a metodologia foi implementado e executado com diferentes combinações de parâmetros do BLAST a fim de comparação dos resultados obtidos. Como medidas de comparação, foram dados maiores pesos aos valores médios de ocorrência de módulos e de módulos por sequência e número total de módulos encontrados. Pela observação dos resultados conclui-se que a metodologia provou ser eficaz para identificação de módulos formadores de proteínas mosaicas e que a combinação do uso da PAM30 aliado a valor alto de *E-value*, desativação do Filtro de Regiões de Baixa Complexidade, diminuição do tamanho da Palavra Inicial e ativação da Estatística Baseada em Composição em relação aos valores padrão apresentou melhores resultados.

A partir dos resultados obtidos conclui-se também que foi confirmada a estrutura mosaica das proteínas da família MASP do *T. cruzi* visto que o mapeamento dos módulos encontrados possibilitou a visualização desses em todas as sequências da família com uma média de 78 módulos por sequência.

Como foi observada a sobreposição e ocorrência em série de alguns módulos, propõe-se como trabalho futuro o estudo da possibilidade de ocorrência de um módulo estar condicionada à ocorrência de outro, o que possibilitaria o refinamento dos resultados obtidos neste trabalho por meio da redefinição como módulo único de módulos que se sobrepõem ou que ocorrem sempre em série e também o estudo da ocorrência condicional de módulos e da presença de um mesmo conjunto de módulos em diferentes sequências pode trazer informações importantes para estudos sobre o *T. cruzi*. Além disso, fica com trabalho futuro a comparação dos resultados obtidos com os resultados de Lima & Souza (2008), cuja identificação de módulos se baseia no transcriptoma do parasito.

APÊNDICE – Mapeamento de Sequência

Este apêndice apresenta o mapeamento dos módulos encontrados na família MASP do *T. cruzi* e das posições em que ocorrem na sequência que apresenta maior número de módulos (187).

Sequência: Tc00.1047053507957.200

MAMM	1	4
RVLL	8	11
LLVC	10	13
AADG	22	25
VSGG	30	33
SGGD	31	34
GGDD	32	35
QEQR	52	55
RAAE	55	58
AAEA	56	59
ATAD	59	62
ADAK	61	64
AKAA	63	66
AAEA	66	69
AAEA	69	72
AAEK	72	75
EKAK	74	77
AKAE	76	79
EAEA	79	82
EAAS	81	84
SEAA	84	87
AAEK	86	89
EKAK	88	91
AKAA	90	93
KTTA	100	103
TTAA	101	104
VEAS	105	108
AKAA	109	112
KAAE	110	113
AAEA	111	114
EAAA	113	116
AAKA	115	118
AKAK	116	119
ETAET	123	127
ETAT	126	129
TATE	127	130
ATEA	128	131
TEAA	129	132
AADA	131	134
ADAK	132	135
AKAA	134	137
TAEK	140	143
EKAK	142	145
AKAE	144	147
EAEA	147	150
EAAS	149	152
SEAA	152	155
KTTA	161	164
TTAA	162	165
TAAA	163	166
AAEA	165	168
AEAS	166	169
AKAA	170	173
KAAE	171	174
AAEA	172	175
EAAA	174	177
AAKA	176	179
AKAK	177	180
ETAET	184	188
ETAT	187	190
TATE	188	191
ATEA	189	192
TEAA	190	193
AAEK	192	195
EKAA	194	197
AAKA	197	200
SEAA	204	207
AAEK	206	209
EKAK	208	211
AKAA	210	213
AAAA	212	215
AAEA	214	217
EAAA	216	219
KTAA	220	223
TAAA	221	224
AAEA	223	226
EAAA	225	228
AAAA	226	229
AAEA	228	231
EAKT	230	233
TSAE	233	236
ETAK	236	239

TANA	242	245	AAEA	350	353
NAAT	244	247	EAAA	352	355
AATA	245	248	AATA	354	357
TAAA	247	250	AAEA	357	360
AAKA	249	252	EAAA	359	362
AKAK	250	253	AAEA	361	364
AKAE	252	255	EAKT	363	366
ETEK	255	258	TSAE	366	369
EKAA	257	260	ETAK	369	372
AAAA	259	262	TAKT	370	373
AAAA	260	263	TATA	373	376
AAAA	261	264	TANT	375	378
AAKE	263	266	ETAA	379	382
AKEA	264	267	TAAA	380	383
EATT	266	269	AAKA	382	385
TKAK	269	272	AKAK	383	386
AKAA	271	274	AKAE	385	388
KAAE	272	275	ETEK	388	391
AAEA	273	276	EKAA	390	393
EAAK	275	278	AAAA	392	395
EAAK	280	283	AAAA	393	396
AAKA	281	284	AATE	395	398
AKAA	282	285	ATEA	396	399
AAAA	284	287	TEAD	397	400
AAKA	286	289	EADA	398	401
AKAA	287	290	ADAK	399	402
KAAE	288	291	KTTA	402	405
EAAK	299	302	TTAA	403	406
AAKA	300	303	EAVA	410	413
AKAA	301	304	AVAE	411	414
AATA	304	307	VAEA	412	415
TAKT	306	309	EEEV	423	426
KTAA	308	311	KTAI	427	430
EEAS	312	315	SGEK	441	444
AKAA	316	319	KQEL	444	447
KAAE	317	320	QELL	445	448
AAEA	318	321	QEKE	456	459
EAAA	320	323	EQHE	459	462
AAKA	322	325	QQHQ	464	467
AKAK	323	326	SAGN	472	475
AKAA	325	328	NGEE	475	478
AAEA	328	331	GEES	476	479
ETAK	333	336	ANGT	485	488
TAKA	334	337	TNAT	488	491
ASAG	337	340	SDGS	497	500
KAAE	341	344	TAVS	501	504
EAAK	345	348	APLL	508	511
AAKA	346	349	PLLL	509	512
AKAA	347	350	LLLL	510	513

LLLL	511	514
LFVA	514	517
VAFA	516	519
AAAA	519	522
AAAA	520	523
AAVV	522	525
VVAA	524	527

REFERENCIAL BIBLIOGRÁFICO

ALBERTS, B.; BRAY, D.; HOPKIN, K.; JOHNSON A.; LEWIS, J.; RAFF M.; ROBERTS, K.; WALTER, P. **Fundamentos da Biologia Celular.** 2.ed. Porto Alegre: Artmed, 2006. 740 p.

ALTSCHUL, S. F.; MADDEN, T. L.; SCHAFFER, A. A.; ZHANG, J.; ZHANG, Z.; MILLER, W.; LIPMAN, D. L. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. **Nucleic Acids Research**, v. 25, n. 17, p. 3389-3402, sep. 1997.

ANDRADE, L. O.; ANDREWS, N. W. The *Trypanosoma cruzi* host cell interplay: location, invasion, retention. **Nature Reviews Microbiology**, v. 3, n. 10, p. 819-823, oct. 2005.

AVERY, V. M.; ADRIAN, D. L.; GORDON, D. L. Detection of mosaic protein mRNA in human astrocytes, **Immunology and Cell Biology**, v. 71, n. 3, p. 215-219, june 1993.

BAXEVANIS, A.D.; OUELLETTE, B.F.F. **Bioinformatics: A pratical Guide to the Analysis of Genes and Proteins.** Wiley-Interscience. USA, 2001.

BIOSCIENCES. **Protein Structure Analysis.** Disponível em: <http://www.gbiosciences.com/EducationalProducts/Protein-Structure-Analysis.aspx>. Acesso em: 01 nov. 2008.

BROWN, T. A. **Genomes**. 2. ed. Oxford: BIOS Scientific Publishers, 2002. 572 p.

CARVALHO, L.F.S. BLOOM – **BLAST Object Oriented Management: uma solução integrada para gerenciamento de resultados do BLAST por meio de um paradigma orientado a objetos.** 203 f. Dissertação (Mestrado em Gestão do Conhecimento e da Tecnologia da Informação) – Universidade Católica de Brasília, Brasília-DF 2002.

CLARK, F. **An Introduction to BLAST.** 2006. Disponível em: <http://clarkfrancis.com/blast/Blast_what_and_how.html>. Acesso em: 05 out. 2008.

DEGRAVE, W. ***Trypanosoma cruzi*:** o genoma. Rio de Janeiro. Disponível em: <http://www.fiocruz.br/chagas/cgi/cgilua.exe/sys/start.htm?sid=14>. Acesso em: 05 maio 2008.

DIAS, J. C. P. Notas sobre o *Trypanosoma cruzi* e suas características bio-ecológicas, como agente de enfermidades transmitidas por alimentos. **Revista da Sociedade Brasileira de Medicina Tropical**, v. 39, n. 4, p. 370-375, jul/ago 2006.

DOOLITTLE, R.F. The multiplicity of domains in proteins. **Annual Review of Biochemistry**, v. 64, p. 287-314, july 1995.

EL-SAYED N.M.; MYLER, P.J.; BARTHOLOMEU, D.C.; NILSSON, D.; AGGARWAL, G.; TRAN, A.N.; GHEDIN, E.; WORTHEY, E.A.; DELCHER, A.L.; BLANDIN, G.; WESTENBERGER, S.J.; CALER, E.; CERQUEIRA, G.C.; BRANCHE, C.; HAAS, B.; ANUPAMA, A.; ARNER, E.; ASLUND, L.; ATTIPOE, P.; BONTEMPI, E.; BRINGAUD, F.; BURTON, P.; CADAG, E.; CAMPBELL, D.A.; CARRINGTON, M.; CRABTREE, J.; DARBAN, H.; DA SILVEIRA, J.F.; DE JONG, P.; EDWARDS, K.; ENGLUND, P.T.; FAZELINA, G.; FELDBLYUM, T.; FERELLA, M.; FRASCH, A.C.; GULL, K.; HORN, D.; HOU, L.; HUANG, Y.; KINDLUND, E.; KLINGBEIL, M.; KLUGE, S.; KOO, H.; LACERDA, D.; LEVIN, M.J.; LORENZI, H.; LOUIE, T.; MACHADO, C.R.; MCCULLOCH, R.; MCKENNA, A.; MIZUNO, Y.; MOTTRAM, J.C.; NELSON, S.; OCHAYA, S.; OSOEGAWA, K.; PAI, G.; PARSONS, M.; PENTONY, M.; PETTERSSON, U.; POP, M.; RAMIREZ, J.L.; RINTA, J.; ROBERTSON, L.; SALZBERG, S.L.; SANCHEZ, D.O.; SEYLER, A.; SHARMA, R.; SHETTY, J.; SIMPSON, A.J.; SISK, E.; TAMMI, M.T.; TARLETON, R.; TEIXEIRA, S.; VAN AKEN, S.; VOGT, C.; WARD, P.N.; WICKSTEAD, B.; WORTMAN, J.; WHITE, O.; FRASER, C.M.; STUART, K.D.; ANDERSSON, B. The genome sequence of *Trypanosoma cruzi*, etiologic agent of Chagas disease. **Science**, v. 309, n. 5733, p. 409-415, july 2005.

FRASCH, A. A. C. Functional diversity in the trans-sialidase and mucin families in *Trypanosoma cruzi*. **Parasitology Today**, v. 16, n. 7, p. 282-286, july 2000.

GABORIAUD, C.; ROSSI, V.; FONTECILLA-CAMPS, J. C.; ARLAUD, G. J. Evolutionary Conserved Rigid Module-domain Interactions can be Detected at the Sequence Level: The Examples of Complement and Blood Coagulation Proteases. **Journal of Molecular Biology**, v. 282, n. 2, p. 459-470, sep 1998.

GOLDENBERG, S. ***Trypanosoma cruzi****:* Regulação da expressão gênica. Rio de Janeiro. Disponível em: <http://www.fiocruz.br/chagas/cgi/cgilua.exe/sys/start.htm?sid=14>. Acesso em: 05 maio 08.

GUIMARÃES, A. C. R **Identificação, Classificação e Anotação de Enzimas Análogas em Tripanosomatídeos**. 2006. 122 p. Dissertação (Mestrado em Ciências) – Instituto Oswaldo Cruz/Fundação Oswaldo Cruz, Rio de Janeiro.

GUSFIELD, D. **Algoritms on Strings, Trees and Sequences: Computer** Science and Computational Biology. New York: Cambridge University Press, 1997. 554 p.

HEGYI, H.; BORK, P. On the classification and evolution of protein modules. **Journal of Protein Chemistry**, v. 16, n. 5, p. 545-551, july 1997.

HIGA, R. H. **Entendendo e Interpretando os Parâmetros Utilizados por BLAST**. Campinas, dez, 2001. Instruções Técnicas.

INCOGEN. **NCBI Blastp**. Disponível em <http://www.incogen.com/public_documents/vibe/details/NcbiBlastp.html>. Acesso em: 13 ago. 2008.

JUNG, C. F. **Metodologia Para Pesquisa & Desenvolvimento**. Rio de Janeiro: Axcel Books do Brasil, 2004. 312 p.

KAHN, S. J.; NGUYEN D.; NORSEN, J.; WLEKLINSKI, M.; GRANSTON, T.; KAHN, M. *Trypanosoma cruzi*: monoclonal antibodies to the surface glycoprotein superfamily differentiate subsets of the 85-kDa surface glycoproteins and confirm simultaneous expression of variant 85-kDa surface glycoproteins. **Experimental Parasitology**, v. 92, n. 1, p. 48-56, may 1999.

KAMOUN, P.; LAVOINNE, A.; VERNEUIL, H de. **Bioquímica e Biologia Molecular**. Rio de Janeiro: Guanabara Koogan, 2006. 444 p.

KANEHISA, M. **Post-genome Informatics**. Oxford: Oxford University Press, 2000. 148 p.

KOLKMAN, J. A.; STEMMER, W. P. C. Directed evolution of proteins by *exon* shuffling. **Nature Biotechnology**, v. 19, n. 5, p. 423-428, may 2001.

KORF, I.; YANDELL, M.; BEDELL, J. **BLAST:** An essential guide to the Basic Local Alignment Search Tool. Sebastopol: O'Reilly, 2003. 339 p.

LEVY, B. **Estudo Aponta Possibilidade de Quimioterapia Natural para Doença de Chagas.** Rio de Janeiro, 2006. Disponível em: <http://www.ioc.fiocruz.br/pages/informerede/corpo/noticia/2006/fevereiro/23_02_06_02.htm>. Acesso em: 05 maio 2008.

LIMA, E. B.; SOUZA, T. R. **Uma Metodologia para Identificação de Módulos Formadores de Sequências de Proteínas Mosaicas do *Trypanosoma cruzi* a partir do Transcriptoma do Parasito Utilizando a Ferramenta BLAST**. 2008. 53p. Monografia (Graduação em Ciência da Computação) – Universidade Federal de Lavras, Lavras, MG.

MAYER, H. **A collection of evaluated bioinformatics programs and databases**: Sequence Similarity. Disponível em <http://homepage.univie.ac.at/herbert.mayer/>. Acesso em 11 ago. 2008.

MOUNT, D. W. **Bioinformatics:** sequence and genome analysis. 2. ed. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press, 2004. 692 p.

NCBI – NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION. **Program Parameters for Blastall**. Disponível em: <http://www.ncbi.nlm.nih.gov/staff/tao/URLAPI/blastall/blastall_node21.html>. Acesso em: 16 ago. 2008a.

NCBI – NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION. **Basic Local Alignment Search Tool.** Disponível em: <http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?CMD=Web&PAGE_TYPE=BlastFAQs>. Acesso em : 13 ago. 2008b.

NCBI – NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION. **Scoring Systems.** Disponível em: < http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/Scoring2.html>. Acesso em : 13 ago. 2008c.

NEVES, D. P.; MELO, A. L. de; LINARDI, P. M.; VITOR, R. W. A. **Parasitologia Humana**. 11. ed. São Paulo: Atheneu, 2005. 494 p.

PATTHY, L. Modular exchange principles in proteins. **Current Opinions in Structural Biology**, v. 1, p. 351-361, 1991.

PEARSON, W.R. **Protein Sequence comparison and Protein Evolution**. Charlottesville: University of Virginia, 2001 53p. Tutorial.

PROSDOCIMI, F.; CERQUEIRA, G. C.; BINNECK, E.; SILVA, A. F.; REIS, A. N.; JUNQUEIRA, A. C. M.; SANTOS, A. C. F.; NBANI JÚNIOR, A.; WUST, C. I.; CAMARGO FILHO, F.; KESSEDJIAN, J. L.; PETRETSKI, J. H.; CAMARGO, L. P.; FERREIRA, R. G. M.; LIMA, R. P.; PEREIRA, R. M.; JARDIM, S.; SAMPAIO, V. S. FOLGUERAS-FLATSCHART, A. V. Bioinformática: Manual do Usuário. **Biotecnologia Ciência & Desenvolvimento**, v. 29, p. 12-25, 2002.

PSC – PITTSBURGH SUPERCOMPUTING CENTER. **Sequence Analysis: Which scoring method should I use?** Pittsburgh 2007. Disponível em: <http://www.psc.edu/research/biomed/homologous/scoring_primer.html>. Acesso em: 16 set. 2008.

SODRÉ, C.L; KALUME, D.E.; SILVA, M.E.R.; FERNANDES O. ***Trypanosoma cruzi*:** Proteoma. Disponível em: <http://www.fiocruz.br/chagas/cgi/cgilua.exe/sys/start.htm?sid=81>. Acesso em: 05 maio 08.

SOUSA, D. X.; LIFSCHITZ, S. **A avaliação do *E-value* para execução do BLAST sobre bases de dados fragmentadas.** 2007. 15 p. Monografia (Graduação em Ciência da Computação) – Pontífica Universidade Católica, Rio de Janeiro.

SOUSA, M. V.; RICART, C. A. O. ; FONTES, W. Análise de Proteomas: O Despertar da Era Pós-Genômica. **Biotecnologia Ciência e Desenvolvimento**. Brasília, v. 7, p. 12-24, 1999.

SOUZA, W. **Morfologia:** Métodos morfológico. Rio de Janeiro. Disponível em: <http://www.fiocruz.br/chagas/cgi/cgilua.exe/sys/start.htm?sid=12>. Acesso em: 05 maio 2008.

ZAMBALDE, A. L.; PÁDUA, C. I. P. S.; ALVES, R. M. **O documento científico em Ciência da Computação e Sistemas de Informação**. Lavras, MG: DCC/UFLA, 2008.

ZORZETTO, R. **Reprodução desvendada:** Identificação de região do núcleo do *Trypanosoma cruzi* pode facilitar o combate ao mal de Chagas. 2005. Disponível em: <http://revistapesquisa.fapesp.br/index.php?art=2763&bd=1&pg=1&lg=>. Acesso em: 05 maio 08.