INSTITUTO OSWALDO CRUZ
Mestrado em Biologia Celular e Molecular

Geração e análise comparativa de seqüências genômicas de *Trypanosoma rangeli*

Glauber Wagner

RIO DE JANEIRO
2006

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ
Pós-Graduação em Biologia Celular e Molecular

Glauber Wagner

Geração e análise comparativa de seqüências genômicas de *Trypanosoma rangeli*

Dissertação apresentada ao Instituto Oswaldo Cruz como parte dos requisitos para obtenção do título de Mestre em Biologia Celular e Molecular

Orientadores: Prof. Dr. Alberto Martin Rivera Dávila
Prof. Dr. Edmundo Carlos Grisard

RIO DE JANEIRO
2006

W133 Wagner, Glauber

Geração e análise comparativa de seqüências genômicas de Trypanosoma rangeli / Glauber Wagner. – Rio de Janeiro, 2006.
xiv, 105 f. : il.

Dissertação (mestrado) – Instituto Oswaldo Cruz, Biologia Celular e Molecular, 2006
Bibliografia: f. 66-79.

1. Trypanosoma. 2. Genoma. 3. Biologia computacional. 4. GSS. I. Título.

CDD: 616.9363

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

Pós-Graduação em Biologia Celular e Molecular

Glauber Wagner

Geração e análise comparativa de seqüências genômicas de *Trypanosoma rangeli*

ORIENTADORES: Prof. Dr. Alberto Martin Rivera Dávila
Prof. Dr. Edmundo Carlos Grisard

Aprovada em: 08/ 05 /2006

EXAMINADORES:

Prof. Dr. Alexandre Afrânio Peixoto (DBBM / IOC / FIOCRUZ) **- Presidente**
Prof. Dr. Mário Steindel (MIP / CCB / UFSC)
Prof. Dra. Maria Cláudia Reis Cavalcanti (IME)

SUPLENTES:

Prof. Dra. Maria Luiza Machado Campos (DCC / NCE / UFRJ)
Prof. Dra. Yara Maria Traub-Cseko (DBBM / IOC / FIOCRUZ)

Rio de Janeiro, 08 de maio de 2006

Este trabalho foi desenvolvido no Laboratório de Biologia Molecular de Tripanosomatideos do Departamento de Bioquímica de Biologia Molecular (IOC / FIOCRUZ) e no Laboratório de Protozoologia do Departamento de Microbiologia e Parasitologia (CBB / UFSC).

Dedico este trabalho ao meu irmão Willian Wagner (*in memorium*),
que me apoiou durante todo o tempo que esteve entre nós e ao
nos deixar no último ano, deixou um vazio e uma lição: “Procure
fazer sempre aquilo que você mais ama, até o final de seus dias”.
Mas tenho certeza que esteve e estará me apoiando
nos momentos em que mais necessitar.
Saudades meu irmão ....

AGRADECIMENTOS

Aos meus orientadores, Dr. Alberto e Dr. Edmundo, pelos conselhos, apoio, conversas, amizade, e pela aposta no meu trabalho, mesmo sabendo que jamais havia trabalhado com biologia molecular e bioinformática. Vocês são pessoas em que me espelho e afirmo sem nenhum constrangimento, que se chegar a ter o conhecimento que vocês possuem, estarei realizado. Obrigado por tudo!!!!!!.

Agradeço aos meus pais, Almir e Claudete, pelo apoio, carinho, ajuda nos momentos difíceis, pelo apoio financeiro é claro, pelas conversar e por me ouvir falar desse trabalho, mesmo sem entender nada do que eu estava falando, vocês são demais!!!!!

Ao meu irmão Willian, que esteve comigo até a metade dessa jornada de corpo presente, e que após sua partida esteve ao meu lado sempre, nos meus pensamentos e no meu coração e assim sempre ficará. Você é um anjo que deu um "rolé" pela terra e agora está cumprido seu papel ai em cima, dando força a todos…. Obrigado por ter sido meu irmão… sempre!!!!!

A minha noiva e futura esposa, Adriana, pelo amor e carinho, amizade, conversas e principalmente pela compreensão, onde mesmo a distância esteve sempre do meu lado me apoiando. Você sempre foi e será muito importante para mim, sem você eu não chegaria aqui …. Te amo demais, obrigado por tudo!!!

Ao Dr. Mário, que além de participar da banca, aceitou em ser o revisor deste trabalho e também pela ajuda e conselhos durante estes anos de convívio. Muito Obrigado...

Agradeço a Dra. Maria Cláudia (Yoko), Dra. Maria Luiza, Dra. Yara e Dr. Alexandre por terem aceito participar dessa banca e pelo grande auxílio de todos durante a execução deste trabalho.

Aos meus familiares, tios, primos, sogro, sogra e cunhadas, pelas conversas e churrascos que sempre me fizeram bem, principalmente nos momentos difíceis dessa caminhada.

Aos meus colegas de Laboratório de biologia molecular do Rio, Priscila, Luana e Silvana pela ajuda interminável na construção da biblioteca, pela amizade e conversas no momento do café....

Meus amigos e colegas do laboratório de Protozoologia, Patrícia, Juliana, Rodrigo, Gianina, Cris e Taís, pela ajuda durante o seqüenciamento, pela amizade, pelas conversas, pelas brincadeiras. Devo muito a vocês!!!!.

A minha colega de bioinformática Kary, pela ajuda, amizade, paciência e acima de tudo pelas longas horas de conversa nos momentos em que eu mais precisa bater um papo!!!! Você sempre será bem vinda na minha casa..... com certeza posso afirmar que você foi uma irmã para mim no Rio!!! Gracias!!!!

Aos colegas de bioinformática, Pablo, Rafael, Pedro, Ricardo, Ildefonso, Henrique, Diogo, Zé, Juliana, Tiago e Linair pela ajuda, amizade, risadas e incentivo nos momentos mais difíceis.

Aos amigos do laboratório do Rio, Igor, Letícia, Cris, Anissa, Luanda, Erich, João, Marcel R., Marcel M., André, Adriana e todos os que já passaram por lá durante o período que estive no laboratório. Obrigado pela força de todos!!!!

A todo o pessoal do Laboratório de Protozoologia (não vou citar nomes senão cometo uma injustiça), muito obrigado pela amizade e ajuda, em especial a Letícia, valeu pelas células….

Ao pessoal do NCE, IME e COPPE que me ajudaram com os problemas de banco de dados, programação e pela paciência para me ensinar um pouco de computação, além da amizade que formamos durante este tempo.

Ao grande amigo Igor, pelo convívio e pelas longas conversas que muito me ajudaram a entender a biologia molecular... Com certeza teremos muitos trabalhos a realizar juntos!!!! E muitas conversas inacabadas para terminar!!!!!

Aos meus eternos amigos e irmãos, Kurt, Marcos, Luiz, Edivan e Fernandinho, que mesmo longe me apoiaram e ajudaram de alguma forma que eu continuasse esta jornada. Quando eu perdi um irmão, eu ganhei cinco!!!!!!!!

É claro que não posso esquecer do casal de grandes amigos, André e Luísa, pela companhia e grande amizade que formamos no Rio, que com certeza levarei para todo o lugar que for o carinho de vocês!!!!!

Agradeço a todos os amigos que conheci no Rio, Marli, Pablo, Dioguinho, Serginho, Carol, Marcos, Antônio, Paulo, Thomas, João, Critiano, Lívia, Cláudia pela amizade e momentos de descontração que passamos juntos..... Valeu!!!!

A todo o pessoal que utilizaram o GARSA (os "*bugs*") e que me ajudaram, e continuarão ajudando, a resolver os problemas do sistema, Juliana, Renata, Denise e Patrícia. Obrigado!!!!!

A todos aqueles que tive o prazer de conhecer na Casa Amarela, em especial, Mary, Francisco, Gio, Érica, Paulo, Dona Vera, que sempre estavam dispostos a conversar sobre os mais diversos assuntos…. Valeu pelo apoio de todos…..

Às secretárias do curso, Daniele, Cleide e Eliete, pela grande ajuda que me prestaram desde o início....

Ao Dream Theater, que além de embalar a redação desta tese e de muitas linhas de código do GARSA, vieram para o Brasil pra que pudesse ver a melhor banda tocar !!!!

A todos aqueles que não foram citados, por falta de memória..... valeu!!!!

Ao CNPq pelo apoio financeiro e bolsa de estudos…..

MUITO OBRIGADO !!!!

“Você não pode ensinar nada a um homem”;
você pode apenas ajudá-lo a encontrar a resposta dentro dele mesmo."
Galileu Galilei

RESUMO

O protozoário hemoflagelado *Trypanosoma* (*Herpetosoma) rangeli* Tejera, 1920, (Kinetoplastida: Trypanosomatidae) compartilha diversas espécies de hospedeiros invertebrados e vertebrados com *T. cruzi*, agente etiológico da doença de Chagas. Recentemente, foram publicados os genomas de 3 espécies de tripanosomatídeos de alta relevância em saúde humana (*Tri-Tryps*). Porém, espécies não-patogênicas não possuem o mesmo status, e como *T. rangeli* não determina nenhuma patogenia ao homem, poucos trabalhos no âmbito genômico tem sido desenvolvidos. Duas abordagens metodológicas têm sido utilizadas na busca de genes em diversas espécies, a GSS (*Genome Sequence Survey*) que visa a geração de seqüências de clones de DNA genômico gerados aleatoriamente e a EST (*Expressed Sequence Tags*) que visa a geração de seqüências a partir de bibliotecas de cDNA. Neste trabalho seqüenciamos 1.720 seqüências genômicas de *T. rangeli* cepa SC58 através de GSS. Foi também desenvolvido no âmbito do presente estudo um sistema de anotação de seqüências, chamado GARSA (*Genomic Analysis Resources for Sequence Annotation*). Neste sistema, é possível executar 21 programas de bioinformática, que vão desde a avaliação de qualidade e limpeza das seqüências até análise filogenética e domínios protéicos, numa forma simples e intuitiva. Após a limpeza dos 1.720 cromatogramas, um total de 915 seqüências foi agrupado em 375 seqüências não redundantes (GSS-nr). O conteúdo G+C das regiões codificantes foi de 55%. Análises de similaridade utilizando os programas BLAST e Interpro, identificaram similaridade em 68% das seqüências, sendo 53% proteínas hipotéticas de organismos pertencentes à mesma família, notadamente o *T. cruzi*. Também foram encontradas seqüências associadas ao processo de edição de mRNA (DEAD box helicase), bem como seqüências relacionadas a superfície do parasito, como trans-sialidase, metaloproteases e mucinas. Foram realizadas anotações funcionais baseadas no vocabulário proposto pelo Consórcio *Gene Ontology*, sendo que a maior parte das anotações dentro da categoria de função molecular está relacionada com RNA helicase, serino peptidases e proteínas ligantes. Para 31% das seqüências não foi possível inferir as funções com base na similaridade com genes já determinados, podendo estas serem seqüências ainda não determinadas, seqüências específicas de *T. rangeli* ou regiões intergênicas. Até o presente momento nenhum trabalho com a finalidade de seqüenciar o genoma de *T. rangeli* foi desenvolvido, portanto este trabalho pode ser considerado como o primeiro com o objetivo de explorar em maior escala o genoma desta espécie.

Palavras chaves: *Trypanosoma rangeli*; GSS; genoma; anotação; bioinformática.

ABSTRACT

The hemoflagellate protozoan parasite *Trypanosoma* (*Herpetosoma) rangeli* Tejera, 1920 (Kinetoplastida: Tryponosomatidae) share several species of invertebrate and vertebrate hosts with *T. cruzi*, etiological agent of Chagas' disease. Recently, the genome of 3 trypanosomatid species of major importance on human health (*Tri-Tryps*) were described but non-pathogenic species has not been well studied, among which we include *T. rangeli*. Two distinct approaches have been used on genomics of several species, the GSS (*Genome Sequence Survey*) which aims the generation of sequences from randomly generated genomic DNA clones and EST (*Expressed Sequence Tags*), directed to the generation of sequences from cDNA libraries. In the present study 1,720 genomic sequences from *T. rangeli* SC58 were generated by GSS. Furthermore, an integrated system for sequence analysis and annotation named GARSA (*Genomic Analysis Resources for Sequence Annotation*) was also developed. Through this system it is possible to run 21 bioinformatics softwares from simple sequence analysis and trimming to phylogenetic and protein domain analyses in a user-friendly and intuitive manner. After analysis of the 1,720 sequences, a total of 915 were grouped in 375 non-redundant sequences (GSS-nr). The G+C content of the coding regions was of 55%. Similarity searches based on BLAST and Interpro revealed positive for 68% of the sequences, being 53% hypothetical proteins of organisms belonging to the same family, especially *T. cruzi*. Also, sequences related to the mRNA editing process (DEAD box helicase), as well as from the parasite coat as trans-sialidase, metaloproteases and mucinas were found. Functional annotation based on the *Gene Ontology* consortia vocabulary were carried out, mostly related to molecular function and related to RNA helicase, serino-peptidases and ligands. For 31% of the generated sequences was not possible to infer functions based on similarity searches. Thus, these sequences may represent unknown sequences, *T. rangeli* specific sequences or even intergenic regions. Up to now there are no reports concerning the *T. rangeli* genome, indicating that the present work is the first one addressing a large scale exploration of the parasite genome.

Key words: *Trypanosoma rangeli*; GSS; genome; annotation; bioinformatic.

Índice

	Dedicatória	v
	Agradecimentos	vi
	Resumo	x
	Abstract	xi
	Índice	xii
1	Introdução	1
1.1	A Ordem Kinetoplastida e o Gênero *Trypanosoma*	1
1.2	*Trypanosoma rangeli*	2
1.3	Genômica de parasitos	6
1.3.1	Genomas de Tripanosomatídeos	7
1.3.2	Metodologias de seqüenciamento	10
1.4	Bioinformática e anotação de seqüências	12
2	Objetivos	20
2.1	Objetivo Geral	20
2.2	Objetivos Específicos	20
3	Materiais e Métodos	21
3.1	Crescimento e extração do DNA total da cepa *T. rangeli* SC58	21
3.2	Construção da biblioteca genômica	22
3.3	Extração do DNA plasmidial e seqüenciamento do DNA	22
3.4	Desenvolvimento da Sistema GARSA	23
3.4.1	Desenvolvimento e equipamentos utilizados	23
3.4.2	Banco de Dados GARSA	24
3.4.3	Desenvolvimento e programas utilizados no *pipeline*	24
3.5	Análise dos resultados	26
4	Resultados	34
4.1	Sistema GARSA	34
4.1.1	Desenvolvimento Geral	34
4.1.2	Banco de dados	34
4.1.3	*Pipeline* GARSA	34
4.2	Seqüenciamento e análises da Biblioteca de GSS da cepa SC58 de *T. rangeli*	43

4.2.1 Construção e seqüenciamento da biblioteca 101 (*GSS*) 43
4.2.2 Análises das seqüências 43
5 Discussão 55
5.1 Sistema de anotação 55
5.2 Análise das seqüências de *T. rangeli* 57
6 Conclusões 65
7 Referências Bibliográficas 66
8 Anexos 80

1 - INTRODUÇÃO

1.1 – A Ordem Kinetoplastida e o Gênero *Trypanosoma*

A ordem Kinetoplastida compreende organismos flagelados (1 a 2 flagelos) que possuem uma única mitocôndria e uma estrutura denominada cinetoplasto que contêm o DNA mitocondrial ou kDNA. O kDNA é formado por uma "rede" concatenada composta por milhares de minicírculos[1] e centenas de maxicírculos[2] (Barret *et al*., 2003). Além disso, apresentam o processo de edição do mRNA[3] em suas mitocôndrias (Simpson *et al.*, 2000b), bem como a formação de transcritos policistrônicos, ou seja, a transcrição seqüencial de vários genes formando um mRNA em que os genes estão organizados *in tandem*, que posteriormente são separados através do mecanismo de *trans-splicing*, em que os mRNA são formados pela junção de cada região codificante do mRNA policistrônico com uma região 5' não codificante conhecida como *spliced-leader* ou mini-éxon (Nilsen 1995; Campbell *et al.,* 2003) e uma extensão de Adeninas ("cauda poli A") na porção 3' dessa região codificante, as quais determinam a maturação do mRNA.

Os organismos pertencentes a esta Ordem possuem uma organela chamada de peroxissomo (glicossomo em Trypanosomatidae) espalhada por todo o citosol do organismo. Foram encontradas enzimas envolvidas na via glicolítica e no metabolismo de carboidratos (Michels *et al.,* 2000), sugerindo que esta organela possui papel essencial nestes metabolismos. Estes organismos também apresentam um ou dois flagelos que emergem de uma bolsa flagelar, situada no corpúsculo basal flagelar, próximo ao cinetoplasto.

Segundo Vickerman (1976) esta ordem é dividida em duas Famílias, Bodonidae e Trypanosomatidae. A primeira compreende os organismos de vida livre dos gêneros *Bodo, Rynchobodo*, *Ichtyobodo*, *Parabodo*, *Herotomita, Pleuromonas, Phyllomitus, Cruzella, Amastigomonas, Dimastigella*, *Cephalothamnium, Phanerobia, Lamellasoma)* ou parasitos obrigatórios do gênero (*Cryptobia, Trypanoplasma*), que apresentam dois flagelos e um

[1] Minicírculo: pequenas moléculas de DNA, entre 465bp – 10,00bp que compõe a rede de kDNA em kinetoplastideos, especula-se que estes contém seqüências conhecidas como gRNA (RNA guias), estas atuam na edição de RNAm nas mitocôndrias (Simpson *et al.,* 2000b)

[2] Maxicírculo: moléculas de DNA encontradas no cinetoplasto de kinetoplastídeos, estas moléculas possuem tamanho médio de 10kbp, e contem informações homólogas ao DNAmt dos demais eucariotos (Simpson *et al*., 2000b)

[3] mRNA *editing* ou edição de RNAm é o processo de controle pós-transcricional típico de kinetoplastideos onde alguns transcritos sofrem inserções ou deleções de resíduos de Uracila (U), ou a troca de uma Citosina (C_{34}) por uma Uracila (U_{34}) modificando um códon de parada para um codificador de triptofano (Simpson *et al.,* 2000b)

cinetoplasto extenso. Já a segunda Família, compreende os organismos que apresentam um flagelo e com cinetoplasto pequeno, todos são parasitos obrigatórios (*Trypanosoma, Leishmania, Phytomonas*, *Crithidia*, *Leptomonas*, *Blastocrithidia, Herpetomonas, Rhynchoidomonas, Endotrypanum*).

Já a Família Trypanosomatidae apresenta algumas características peculiares que a difere das demais Famílias da Ordem Kinetoplastida, como a presença de uma camada subpelicular de microtúbulos, que confere maior rigidez à célula e a presença de um flagelo que emerge de uma bolsa flagelar localizada próximo ao cinetoplasto (De Souza, 2002).

Os organismos pertencentes ao gênero *Trypanosoma* apresentam um ciclo biológico heteroxeno e possuem formas tripomastigotas no sangue do hospedeiro vertebrado. Os hospedeiros invertebrados insetos são hematófagos, e no caso de hospedeiros vertebrados aquáticos, são transmitidos por hirudíneos (sanguessugas). Este gênero é dividido em duas secções: Stercoraria e Salivaria. A secção Stercoraria inclui espécies, que se desenvolvem no intestino posterior de seus vetores como o *T. cruzi, T. lewisi* e *T. theileri*, enquanto as espécies pertencentes à secção Salivaria se desenvolvem no intestino anterior e são transmitidas, ou seja, via saliva dos vetores como o *T. brucei, T. vivax* e *T. congolense* (Hoare, 1972; Vickerman, 1976).

A posição taxonômica do *T. rangeli* tem sido objeto de muita discussão. A maioria dos autores classifica *T. rangeli* como pertencente à Secção Stercoraria, (Stevens & Gibson, 1999; Briones, 1999; Hughes & Piontkivska, 2003), mostrando maior similaridade deste com *T. cruzi*, principalmente quando são utilizados para a comparação marcadores clássicos, como gene ribossomal 18S RNA. Entretanto trabalhos utilizando outros marcadores moleculares evidenciam considerável similaridade entre *T. rangeli* e *T. brucei*, alocando assim este parasito na secção Salivaria (Amorin *et al*., 1993, Henriksson *et al.,* 1996).

Porém até o presente momento, nenhum trabalho realizado utilizou estes ou outros marcadores filogenéticos de forma concatenada. Segundo Baldauf (2003), utilizar um grupo de genes ordenados de forma concatenada nas análises filogenéticas pode ser uma boa alternativa para resolver filogenias controversas, como é o caso do *T. rangeli*.

1.2 - O *Trypanosoma rangeli*

O *Trypanosoma (Herpetosoma) rangeli* Tejera, 1920, é um protozoário hemoflagelado pertencente à Ordem Kinetoplastida, Família Trypanosomatidae (D'Alessandro 1976) secção

Stercoraria (Hoare, 1972). Possui como hospedeiros invertebrados várias espécies da Sub-Família Triatominae, popularmente conhecidos como "barbeiros", além de diversas espécies de mamíferos, incluindo o homem (D'Alessandro & Saraiva, 1999).

Ao longo do ciclo nos hospedeiros vertebrado e invertebrado, o *T. rangeli* apresenta formas epimastigotas e tripomastigotas com elevado pleomorfismo. As formas tripomastigotas sangüíneas apresentam tamanhos variando entre 26 a 34μm e uma membrana ondulante bem desenvolvida, além de apresentar um cinetoplasto pontual e sub-terminal (Guhl & Vallejo, 2003) (Figura 1.1).

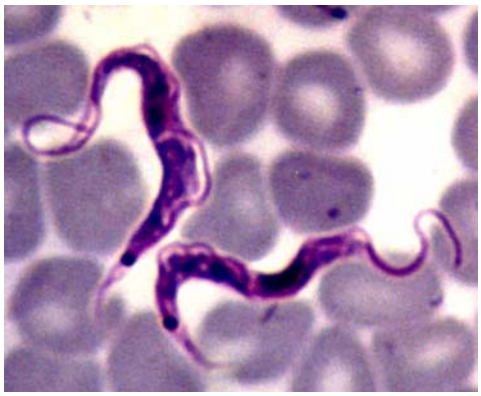


Figura 1.1: Forma tripomastigota sangüínea do *Trypanosoma rangeli* em sangue de camundongo experimentalmente infectado (aumento de 1.000X).

O *T. rangeli* tem uma ampla distribuição geográfica nas Américas, ocorrendo em simpatria com o *T. cruzi* Chagas, 1909 agente etiológico da Doença de Chagas (D'Alessandro & Saraiva, 1999; Grisard *et al*., 1999) (Figura 1.2). No Brasil, há casos da ocorrência de *T. rangeli* descritos em diferentes regiões, como na Amazônia (Coura *et al*., 1996; D'Alessandro & Saraiva, 1999) e na região Sul do Brasil (Steindel *et al.,* 1991), sendo os primeiros casos de infecção em humanos descritos por Coura *et al.* (1996).

Vallejo *et al.* (2002), utilizando kDNA, demonstraram a existência de duas linhagens genéticas do parasito na América Latina, pois as cepas apresentam variações no tamanho e no número de regiões conservadas do minicírculo do kDNA, podendo apresentar uma (KP1), duas (KP2) ou 4 (KP3) regiões (Vallejo *et al*., 1994).

Além da sobreposição geográfica, estes parasitos compartilham 60% de seus determinantes antigênicos solúveis (Afchain *et al*., 1979), o que pode estar influenciando as estimativas da incidência do *T. cruzi*, através da ocorrência de resultados falso-positivos em exames sorológicos (Guhl *et al.*, 2002). Infecção mista de *T. rangeli* e *T. cruzi* em triatomíneos têm sido demonstrada

por diferentes autores (Chiurillo *et al.*, 2003; Steindel *et al.,* 1994). Além disso, é elevado o número de descarte de bolsas de sangue devido a resultados inconclusivos para a doença de Chagas, sendo que este descarte atingiu em 1992 cerca de 1,6% de bolsas descartadas no maior banco de sangue do Brasil (Silva, 2002).

Mesmo com a vasta diversidade de espécies de mamíferos infectadas pelo *T. rangeli*, pouco se conhece sobre o ciclo evolutivo do parasito nestes hospedeiros (D'Alessandro & Saraiva, 1999; Grisard *et al.,* 1999), o que ocorre de maneira inversa nos hospedeiros invertebrados, cujo ciclo é bem conhecido. A capacidade das formas epimastigotas em escapar do trato digestivo médio para a hemocele é a principal característica biológica deste ciclo. São encontradas formas epimastigotas e tripomastigotas longas de forma livre na hemolinfa, onde passam a se multiplicar intensamente por divisão binária, além de formas do parasito dentro de hemócitos que confundem com formas amastigotas (Meirelles *et al.*, 2005). Após a colonização da hemolinfa, os parasitos migram para as glândulas salivares do vetor onde as formas epimastigotas acabam aderindo à superfície, atravessando a parede e atingindo a luz das glândulas salivares. No lúmen glandular ocorre a diferenciação dos parasitos em metatripanosomas, ou tripomastigotas metacíclicos, que são transmitidos aos hospedeiros vertebrados no próximo repasto sangüíneo (D'Alessandro 1976) (Figura 1.3).

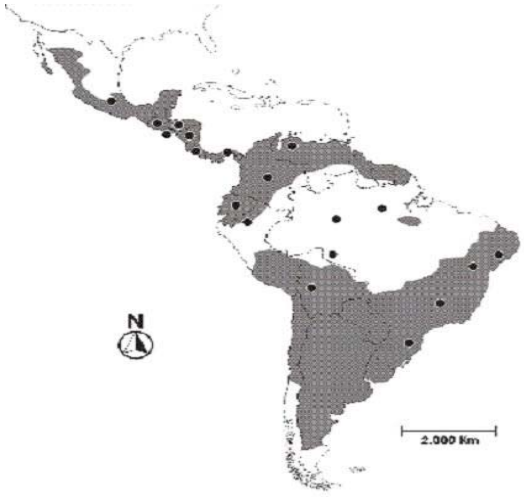


Figura 1.2: Distribuição geográfica de *Trypanosoma cruzi* (área sombreada) nas Américas e locais onde já foram descritos casos de *Trypanosoma rangeli* (pontos escuros) (Fonte: Grisard & Steindel, 2004).

Nos hospedeiros vertebrados as formas tripomastigotas são predominantes e experimentos com monócitos humanos mostraram que este pode invadir e multiplicar-se lentamente no interior destas células (Osório *et al.*, 1995), entretanto trabalhos mais recentes mostram que este parasito não possui a capacidade de se dividir dentro das células dos hospedeiros vertebrados (Eger-Mangrich *et al.*, 2001). Na primeira semana de infecção é observado um pequeno aumento da parasitemia, porém após 15 dias torna-se extremamente baixa, impossibilitando a detecção deste através de microscopia ótica.

Apesar da forma inoculativa ser a principal forma de transmissão dos metatripanosomas para os hospedeiros vertebrados, a transmissão contaminativa (através das fezes dos triatomíneos) é considerada possível, mas sem grande relevância epidemiológica (Steindel *et al.,* 1991b; D'Alessandro & Saraiva, 1999).

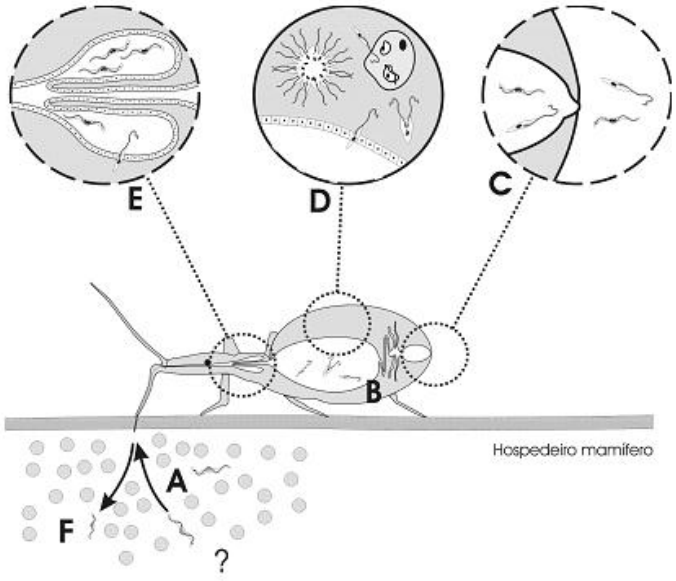


Figura 1.3: Ciclo biológico do *Trypanosoma rangeli* no hospedeiro invertebrado. A – Ingestão das formas tripomastigotas de *T. rangeli* durante o repasto sangüíneo do triatomíneo, B – Formas tripomastigotas e epimastigotas no trato intestinal do triatomíneo, C – Parasitos alcançam a ampola retal e formas epimastigotas e tripomastigotas podem ser excretadas juntamente com as fezes e urina, D – Adesão de formas epimastigotas do parasito ao epitélio intestinal e transposição para a hemocele onde se multiplicam, E – Invasão pelo *T. rangeli* das glândulas salivares do triatomíneo e diferenciação para formas infectivas (metatripanosomas), as quais são inoculadas com a saliva (F) durante o repasto sangüíneo (Fonte: Grisard & Steindel, 2004).

Devido à necessidade de diferenciar o *T. cruzi* do *T. rangeli*, diferentes metodologias estão sendo empregadas atualmente. O esfregaço sangüíneo é a forma mais tradicional e rápida para esta finalidade, entretanto torna-se pouco conclusiva devido às características biológicas do *T. rangeli*, como a baixa parasitemia nos hospedeiros vertebrados, tornado muito curto o período em que as formas tripomastigotas podem ser observadas no sangue do hospedeiro vertebrado (Grisard & Steindel, 2004).

O xenodiagnóstico apresenta-se como uma alternativa, porém o fator complicador é a susceptibilidade diferencial de certas espécies de triatomíneos à infecção por diferentes cepas de *T. rangeli* (Machado *et al*., 2001), bem como a necessidade altos níveis de parasitemia nos hospedeiros mamíferos para que ocorra a infecção (Guhl *et al*., 2002) e a semelhança morfológica das formas de *T. cruzi* e *T. rangeli* presentes no intestino do barbeiro (Grisard & Steindel, 2004).

Assim, novas metodologias vêem sendo estudadas e aplicadas em diferentes centros de diagnóstico, entre as quais podemos citar a imunofluorescência indireta, anticorpos monoclonais, aglutinação por lectinas (Santos & Pereira, 1984; Acosta *et al*., 1991; Steindel *et al*., 1991), perfil de restrição de DNA cinetoplástico (Vallejo *et al*., 1994), análises de perfis de RAPD (Steindel *et al.*, 1994), análise do gene do mini-éxon por reação em cadeia da polimerase (PCR) (Murthy *et al.,* 1992; Grisard *et al*., 1999; Vallejo *et al*., 2002), perfis de RFLP do gene de cisteína proteinase (Tanaka, 1997).

1.3 – Genômica de parasitos

Após o advento do seqüenciamento automatizado de DNA, houve um aumento no número de seqüências depositadas em base de dados públicas, principalmente seqüências de vírus (Degrave *et al.,* 2001). Porém somente após o início do Projeto Genoma Humano e a publicação do primeiro genoma completo de um organismo celular, o *Haemophilus influenzae* (Fleischmann *et al*., 1995), houve um aumento considerável no número de genomas completamente seqüenciados, principalmente aqueles que determinam patogenia em humanos ou em animais e plantas de importância econômica.

O Brasil entrou definitivamente para o grupo de países com a capacidade de seqüenciar genomas completos a partir do projeto genoma de uma bactéria patogênica de laranja, *Xylella fastidiosa* (Simpson *et al*., 2000a), sendo o primeiro fitopatógeno seqüenciado no mundo. Após isso, outros genomas foram e estão sendo seqüenciados no país por grupos multi-institucionais:

Chromobacterium violaceum (Brazilian National Genome Project Consortium 2003) e *Mycoplasma synoviae* (Vasconcelos *et al.,* 2005) seqüenciados pelo Projeto Genoma Nacional Brasileiro (http://www.brgene.lncc.br/); *M. hyopneumoniae* (Vasconcelos *et al.,* 2005) pelo Programa de Investigação de Genomas Sul –PIGS (http://www.genesul.lncc.br) *Herbaspirullum seropediace* pelo Programa de Seqüenciamento Genômico do Estado do Paraná, GENOPAR (http://www.genopar.com.br/); *Schistosoma mansoni* pela FAPESP (http://www.watson.fapesp.br/onsa/Genoma3.htm) e pela Rede de Seqüenciamento de Minas Gerais (http://rgmg.cpqrr.fiocruz.br/); *Gluconacetobacter diazotrophicus* pelo Consórcio RioGene (http://www.riogene.lncc.br/), entre outros.

Atualmente diversas espécies de parasitos eucariotos estão tendo ou já tiveram o genoma explorado ou completamente seqüenciado, entre eles *Giardia lamblia* (Smith *et al.,* 1998), *Plasmodium vivax*[*], *P. falciparum*[*], *Entamoaeba histolytica*[*], *Ascaris suum*[*], *Crypotosporidium parvum* (Puiu *et al*., 2004), *Brugia malayi* (Ghedin *et al.,* 2004a) entre outros.

1.3.1 – Genomas de Tripanosomatídeos

A necessidade de acelerar o processo de descoberta de genes em *T. cruzi,* para a identificação de novos marcadores para diagnóstico, quimioterapia e vacinas, iniciou o mapeamento genético do *T. cruzi* em meados da década de 90 (Hanke *et al*., 1996, Ferrari *et al*., 1997). Entretanto somente em junho de 2005 foi publicado o genoma completo desta espécie (El-Sayed *et al*., 2005a), juntamente com outras 2 espécies de parasitos pertencentes à Família Trypanosomatidae, *T. brucei* (Berriman *et al*., 2005) e *L. major* (Ivens *et al.,* 2005), conhecidos como *TriTryps*. Juntas, estas espécies causam a morte de milhares de pessoas no mundo, sendo os países tropicais os mais afetados (Barret *et al.,* 2003).

Além destes, várias outras espécies pertencentes à Família Trypanosomatidae estão tendo seus genomas completamente ou parcialmente seqüenciados, como *L. braziliensis* (Laurentino *et al*., 2004), *T. vivax* (http://www.sanger.ac.uk/Projects/T_vivax; Dávila *et al.*, 2003b; Guerreiro *et al.,* 2005) e *T. congolense* (http://www.sanger.ac.uk/Projects/T_congolense).

Segundo El-Sayed *et al*. (2005a), o tamanho estimando do genoma (diplóide) de *T. cruzi* é de aproximadamente 106,4 MB, outros trabalhos estimam em 87Mb (Cano *et al.*, 1995), porém há evidências dos próprios autores e de outros (Sturm *et al.,* 2003) que a cepa utilizada, *T. cruzi*

[*] http://www.sanger.ac.uk/Projects

CL-Brener, trata-se de uma cepa híbrida, ocasionando assim problemas durante o seqüenciamento e montagem deste genoma, e devido a isso foi seqüenciado apenas 67 MB do genoma desta espécie. Foram preditos cerca de 12 mil genes por genoma haplóide, 3.500 pseudogenes[1], com tamanho médio dos genes de 1,1 kb e conteúdo G+C de 53,4%. Entretanto, aproximadamente metade desse genoma é composto por seqüências repetitivas, principalmente retrotransposons, regiões teloméricas e proteínas de superfície.

Por outro lado, o *T. brucei* (Berriman *et al*., 2005) possui um genoma pequeno de 26 MB dividido em 11 cromossomos, contendo aproximadamente 9 mil genes e 900 pseudogenes, com tamanho médio de 1,2 kb e 50,9% de conteúdo G+C. Já o genoma de *L. major* (Ivens *et al.,* 2005), apresenta um tamanho aproximado de 32 MB em 36 pequenos cromossomos, 8 mil genes e 39 pseudogenes, porém com conteúdo G+C de 60%.

Importantes evidências sobre a organização e composição dos genomas dos tripanosomatídeos podem ser esclarecidas com a comparação dos genomas destas 3 espécies (El-Sayed *et al.,* 2005b). Como o grande número de genes ortólogos[2] entre estas espécies, sendo as porcentagens de COG[3] (Tatusov *et al*., 2001) muito similares, refletindo as verdadeiras relações filogenéticas entre estas espécies, maior entre *T. cruzi* e *T. brucei* do que quando comparado com *L. major*. Outro fato interessante evidenciado por El-Sayed *et al.* (2005b), é a manutenção da sintenia[4] entre os genomas, suportando a baixa taxa de recombinação e reorganização dos genomas como evidenciado por Ghedin *et al.* (2004b), as regiões não sintênicas poderiam ter sido ocasionadas pelos rearranjos dos elementos móveis presentes nas espécies desta família.

Não foram identificados domínios de proteínas referentes a organismos fotossintéticos (protistas, como *Euglena* sp.), não suportando a teoria de transferência horizontal de genes associados a plantas (Saas *et al.,* 2000; Wilkinson *et al.,* 2002; Hannaert *et al.,* 2003). Segundo estes autores, a origem de alguns genes relacionados ao metabolismo energético nos organismos da ordem Kinetoplastida, deve-se a eventos de transferência horizontal (HGT[5]) desses genes de um endosimbionte, possivelmente da mesma alga que originou o cloroplasto em Euglenida. Estes dados suportam a teoria de que a aquisição deste endosimbionte ocorreu em algum ancestral

[1]Pseudogene são genes não funcionais originados de genes funcionais que apresentam em sua seqüência um códon de parada e/ou mudança na janela de leitura (Balakirev & Ayala 2003)

[2] Genes Ortólogos são genes originados de um único gene presente em uma espécie ancestral comum das espécies comparadas, através do processo de especiação (Koonin 2005).

[3] Abreviação para: *Clusters of Orthologous Groups* (COG)

[4] Sintenia é o termo utilizado quando se quer dizer sobre a ordem seqüencial em que os genes estão organizados no genoma de uma espécie.

[5] HGT (*Horizontal Gene Transfer*) ou LGT (*Lateral Gene Transfer*) é um evento em que genes são adquiridos a partir de um organismo não ancestral, como por exemplo, de um parasito para seu hopedeiro.

comum entre Kinetoplastida e Euglenida, permanecendo até hoje na ordem Euglenida e perdido em Kinetoplastida. Entretanto Kinetoplastida apresenta algumas características destes antigos endosimbiontes, tanto em nível genômico quanto organelas, como os glicossomos (Hannaert *et al.,* 2003). Guerreiro *et al.* (2005) encontraram em *T. vivax,* um homólogo do gene fosfopantotenoil-cisteina sintetase de *Arabdopsis thaliana*, gene este relacionado com o crescimento celular de plantas.

Estudos relacionados ao genoma de *T. rangeli* são escassos*,* sendo a maioria destes focados em marcadores clássicos para a diferenciação entre esta espécie e *T. cruzi* (Vallejo *et al.,* 1994; Grisard *et al*., 1999; Vargas *et al*., 2000; Vallejo *et al.,* 2002), refletindo o baixo número de seqüências deste organismo no GenBank (Tabela 1.1). Assim, pouco se sabe sobre os genes envolvidos nos processos de invasão para a hemocele dos triatomíneos bem como durante seu ciclo dentro dos hospedeiros vertebrados (Snoijer *et al.,* 2004).

Tabela 1.1: Tabela comparativa da quantidade de seqüências nucleotídicas, projetos genomas concluídos ou em andamento, seqüências protéticas, estruturas de proteínas resolvidas e citações no banco de dados PubMed, entre as espécies de *Trypanosoma* disponíveis no GenBank (http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/, acesso: 12/02/2006).

Organismo	***Nucleotídeo***				***Projetos genomas***	***Proteínas***	***Estrutura de Proteínas***	***Citações PubMed***
	Core	***EST***	***GSS***	***Total***				
T. rangeli	111	0	0	111	0	52	9	29
T. vivax	23	0	0	23	1*	25	5	39
T. cruzi	88111	13972	26684	128.767	1	41052	57	1115
T. brucei	10583	5133	91827	107.543	1	23436	59	1291
T. congolense	88	0	0	88	1*	79	0	91
T. evansi	119	0	1	120	0	93	0	18
L. major	1821	2191	18542	22.554	1	10262	25	1038
L. braziliensis	132	0	10419	10.551	1*	68	2	105
L. donovani	854	9839	183	10.876	1*	533	5	705
L. infantum	322	21	0	343	1*	208	0	199

* genomas sendo seqüenciados pelo The Welcome Trust Sanger Institute (http://www.sanger.ac.uk/)

Quando comparamos o número de seqüências entre *T. rangeli* e as principais espécies da Família Trypanosomatidae, tornam-se evidente que pouco se tem feito para o melhor

entendimento desta espécie. Enquanto apenas 111 seqüências nucleotídicas de *T. rangeli* foram submetidas ao GenBank, 128.704 de *T. cruzi*, 107.543 de *T. brucei* e 22.554 de *L. major* foram submetidas ao mesmo repositório até o dia 12/02/2006 (Tabela 1.1), ficando a frente apenas de *T. vivax* e *T. congolense*, ambas com seus genomas sendo seqüenciados pelo *The Wellcome Trust Sanger Institute*.

Uma das possíveis explicações para este baixo número de dados genômicos e de estudos realizados, é o fato deste parasito não causar patogenia em seres humanos. Entretanto seqüenciamento de genomas e identificação de novos genes presentes em espécies não patogênicas ao homem e/ou organismos modelos, são também importantes para estudos comparativos, filogenéticos e de novos conhecimentos biológicos dessas espécies. Este tem sido o caso de espécies como *Paramecium* sp. (Dessen *et al*., 2001), *Plasmodium chabaudi* (Janssen *et al*., 2001), *Caenorhabditis elegans* (Aziz & Blaxter, 2003), *Drosophila melanogaster* (Schaeffer *et al*., 2003), *Saccharomyces cerevisae* (Cliften *et al*., 2001) , *Mus musculus* (Gregory *et al.*, 2002).

Com este intuito, Snoeijer *et al*. (2004) iniciaram o primeiro seqüenciamento em maior escala de *T. rangeli*. Nesse trabalho os autores conseguiram seqüenciar 656 EST (*Expressed Sequence Tags*) de formas epimastigotas e tripomastigotas deste parasito, sendo a maioria destas seqüências similares a *T. cruzi,* entretanto para uma considerável parcela de seqüências não foi encontrada similaridade com nenhum outro organismo. Paralelamente a este estudo, o mesmo grupo realizou o seqüenciamento de 1.295 ORESTES (*Open Reads Expressed Sequence Tags*) das formas tripomastigotas, formando 758 seqüências não redundantes, onde observaram grande expressão de moléculas de superfície (Rodrigues, 2005). Stoco *et al.* (2005), obtiveram mais 1.644 seqüências das formas epimastigotas e tripomastigotas de *T. rangeli*. As seqüências de ambos os projetos ainda estão em análise e não foram submetidas ao GenBank.

Como os trabalhos citados anteriormente tratam do seqüenciamento de EST (RNAm) deste parasito e como não há registros na literatura de nenhum trabalho com a finalidade de seqüenciar o genoma de *T. rangeli*, o presente estudo pode ser considerado o primeiro com o objetivo de explorar o genoma desta espécie.

1.3.2 – Metodologias de seqüenciamento

Existem duas abordagens de seqüenciamento aplicadas para obter informações dos genomas de organismos, o seqüenciamento de porções do DNA (*shotgun* ou BAC – *Bacterial*

Artificial Chromosome) e o seqüenciamento do DNA complementar (cDNA) obtidos a partir de regiões transcritas do genoma, EST ou ORESTES (Passagia & Zaha, 2003).

Segundo El-Sayed *et al*. (2000), projetos de EST tornam-se mais atraentes do que projetos de seqüenciamento de genomas completos, devido ao menor investimento financeiro necessário. Projetos EST são efetivos na obtenção de dados preliminares sobre as regiões codificantes, os genes, proporcionando ampla variedade de marcadores para mapeamento STS (*sequence-tagged-sites*), bem como na obtenção de um perfil de expressão célula-célula ou estágio específico de desenvolvimento. Esta metodologia foi utilizada para a pesquisa de genes em *Phytomonas serpens* (Pappas *et al*., 2005), *T. cruzi* (Porcel *et al.,* 2000), *T. rangeli* Snoeijer *et al*. (2004) e *P. vivax* (Merino *et al.,* 2003).

Esta abordagem baseia-se no seqüenciamento do cDNA - sintetizado a partir do mRNA transcrito pela célula. Após a extração do mRNA total, são utilizados iniciadores para a amplificação deste material através da técnica da Reação da Polimerase em Cadeia (PCR), isto possibilita uma maior eficiência na clonagem dos fragmentos gerados (Adams *et al*., 1991). Entretanto, está técnica de EST favorece o seqüenciamento das regiões 5' e 3' do mRNA, deixando as regiões mais internas desta molécula com pouca representatividade. Para contornar este problema, Dias-Neto *et al.* (1999) realizaram algumas modificações nesta metodologia e conseguiram aumentar a representatividade das seqüências internas dos mRNA. Estes autores descrevem que a utilização de iniciadores não degenerados e condições de PCR com baixa estringência possibilitam o seqüenciamento destas regiões do mRNA, está técnica é conhecida como ORESTES. Esta abordagem tem sido muito utilizada em perfis de expressão de células tumorais em humanos (Dias-Neto *et al.,* 1999; Souza *et al.,* 2002) e está sendo aplicada para *T. rangeli* (Stoco *et al.,* 2005).

Da mesma maneira que a abordagem de EST, ao explorar um genoma utilizando a abordagem de GSS (*Genome Sequence Survey*) (Peterson *et al*., 1991), é possível identificar regiões que possam ser potencialmente usadas como alvos para quimioterapia e marcadores moleculares para diagnóstico específico, como demonstrado em *T. brucei* (El-Sayed & Donelson, 1997), *G. lamblia* (Smith *et al*., 1998), *L. major* (Akapyants *et al*., 2001) e *Plasmodium* sp. (Carlton *et al*., 2001). Entretanto, ao se utilizar esta abordagem, é possível identificar também regiões intergênicas, íntrons, pseudogenes, regiões repetitivas, importantes para a filogenia e genômica comparativa.

O GSS se baseia no seqüenciamento randômico de clones obtidos através da fragmentação total do genoma de um organismo. Esta abordagem de seqüenciamento pode ser utilizada para

dois objetivos distintos: (1) explorar o genoma de uma espécie sem almejar o seqüenciamento completo, como por exemplo, *P. chabaudi* (Janssen *et al*., 2001), *L. braziliensis* (Laurentino *et al*., 2004), *T. vivax* (Guerreiro *et al*., 2005), ou (2) seqüenciar o genoma completo de um organismo, neste caso ela é mais conhecida como *Whole Genome Shotgun* (WGS), esta abordagem foi utilizada para seqüenciar os genomas de *T. cruzi* (El-Sayed *et al*., 2005a), *L. major* (Ivens *et al*., 2005) e *T. brucei* (Berriman *et al*., 2005), entre outros.

Venter *et al.* (2004) utilizaram esta abordagem para realizar estudos de metagenômica (genômica ambiental) do mar de Sargasso. Neste caso, amostras da água do mar foram coletadas e o DNA total dos microorganismos presentes foram extraídos, fragmentados, clonados e seqüenciados de forma randômica. A partir destas seqüências, os genomas de diversas espécies foram parcialmente ou completamente seqüenciados, descobrindo novas espécies.

O uso das abordagens GSS e EST em conjunto gera resultados eficazes na identificação de famílias de genes, regiões intergênicas e regiões repetitivas, como demonstrado em *T. cruzi* (Agüero *et al*., 1996) e na comparação entre *P. berghei*, *P. vivax* e *P. falciparum* (Carlton *et al*., 2001).

Dávila *et al*. (2003a) e Parkinson *et al*. (2002) tem sugerido que projetos genoma de pequena escala, apenas visando à geração de EST e GSS, apresentam um melhor custo-benefício que estudos de genes individuais ou famílias gênicas.

1.4 – Bioinformática e anotação de seqüências

Com o surgimento da "Era Genômica", consolidado com o lançamento do seqüenciamento do genoma humano em 1990 e o seqüenciamento do primeiro genoma completo em 1995 (Fleischmann *et al.,* 1995), tornou-se necessário o uso de ferramentas de informática para as análises em larga escala dos resultados obtidos, surgindo uma nova área multidisciplinar, a Bioinformática. Simultaneamente houve o surgimento de profissionais capacitados, o bioinformata, para analisar e desenvolver aplicativos para a análise desse grande volume de dados.

Existem diversas definições para bioinformática, uma das mais abrangentes diz que "*bioinformática é uma área de pesquisa multidisciplinar que promove a interface entre as áreas biológicas e informática, onde promove a aplicação de técnicas computacionais para a administração e análise de dados biológicos, sendo a principal meta a descoberta de informações biológicas ocultas em uma grande quantidade de dados, auxiliando no*

entendimento da biologia dos organismos" (http://www.ebi.ac.uk/2can/bioinformatics/bioinf_what_1.html).

Os primeiros trabalhos em bioinformática apareceram no final da década de 80 e inicio da década de 90, entre eles, dois programas utilizados na comparação das similaridades entre seqüências, o BLAST (Altschul *et al.,* 1990) e o FASTA (Pearson & Lipman, 1988). Em 1993 alguns trabalhos começaram a evidenciar a necessidade da construção de banco de dados de DNA e proteínas (Kunisawa, 1990; Barker, 1993), desde então o número de seqüências de DNA, proteínas e RNAm depositadas no GenBank, por exemplo, vem crescendo exponencialmente. Em agosto de 2004 chegava a 37,3 milhões de seqüências nucleotídicas (Figura 1.4) e um ano após este número já era de 47 milhões.

Com esse crescente número de seqüências, foram criados diversos bancos de dados tanto de seqüências de nucleotídeos (NT[1], dbEST[1], dbGSS[1], RefSeq[1], DBBJ[2]), seqüências protéicas (NR[1], RefSeq[1], Uniprot[3]), estrutura de proteínas (PDB[4]), vias metabólicas (KEGG[5]), entre outros bancos específicos para cada genoma seqüenciado (TcruziBD, GeneDB, PlasmoDB, entre outros).

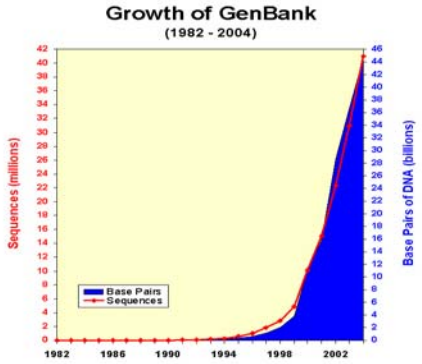


Figura 1.4: Gráfico demonstrando o crescimento no número de seqüências (Vermelho) e pares de bases (Azul) depositadas no Genbank. Onde a partir de 1994, após o seqüenciamento de genomas bacterianos, houve um aumento significativo de entradas neste banco. (Fonte: (http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.htm, acessado em 11/02/06).

O crescente número de projetos genomas, transcriptomas e proteomas vêm fazendo com que novas metodologias, programas, sistemas integradas e bancos de dados sejam desenvolvidos

[1] http://www.ncbi.nih.nlm.gov/entrez
[2] http://www.ddbj.nig.ac.jp/
[3] http://www.pir.uniprot.org/
[4] http://www.pdb.org/
[5] http://www.genome.ad.jp/kegg/

e/ou aprimorados a fim de facilitar e agilizar o trabalho que antes levaria meses ou anos para ser realizado (Dávila, 2002; Hammer & Schneider, 2003). Um bom exemplo disso é a velocidade de anotação dos genomas seqüenciados, o que antes levaria dias ou meses para determinar a seqüência completa do genoma e anotar apenas um gene, atualmente é realizado em horas através da semi-automatização deste processo e por programas cada mais sensíveis e rápidos (Baxevanis & Oullette, 2001).

Anotação de seqüências é um processo múltiplo, pelo qual uma ou mais seqüências brutas de DNA ou de aminoácido são analisadas com a finalidade de atribuir características biológicas para o entendimento do contexto biológico em que estas se inserem, ou seja, sua função (Stein, 2001).

Para isto são usados diversos programas que utilizam algoritmos distintos, como, por exemplo, busca de similaridade local e global, inteligência artificial - HMM (*Hidden Markov Models*), SVM (*Suported Vector Machine*), Redes Neurais, Redes Bayesianas, busca de padrões, e programas ou sistemas de análise que combinam diferentes algoritmos.

Estes programas são utilizados para diversas finalidades, como a análise de qualidade das seqüências, montagem ou agrupamento das seqüências, predição de genes *in silico*, análises de similaridade e homologia, alinhamentos múltiplos de seqüências, análises filogenéticas, predição de estruturas de proteínas *in silico*, entre outras. Na tabela 1.2, listamos alguns dos programas mais utilizados durante o processo de anotação de seqüências.

Um destes programas utilizado na busca de similaridade entre seqüências é o BLAST (Altschul *et al*., 1990; Altschul *et al*., 1997). Este programa utiliza um algoritmo de busca de similaridade local, que consiste em determinar regiões entre as seqüências que possuem uma similaridade mínima e a partir destas regiões é avaliada a similaridade total entre as seqüências.

Um outro programa muito utilizado para esta finalidade e também permite a identificação de domínios e famílias de proteínas é o InterProScan (Mulder *et al.,* 2005). Este programa reúne vários bancos de dados de modelos HMM (ver abaixo) e expressões de domínios de diversas famílias de proteínas em um único banco, conhecido como Interpro, e promove uma busca nestes bancos utilizando diferentes algoritmos.

Existem programas específicos que integram alguns dos programas acima relacionados. Estas sistemas ou programas têm como objetivo facilitar a análise e anotação das seqüências através de *pipelines* de execução dos programas, exemplo destes programas são: ESTAnnotator (Hotz-Wagnblastt *et al.,* 2003) SABIA (Almeida *et al*., 2004), ESTIMA (Kumar *et al.*, 2004) e GATO (Fujita *et al.*, 2005), entre outros. Há também ferramentas que apenas são utilizados para a

visualização e anotações de seqüências, como o Artemis (Rutherford *et al.*, 2000) e Apollo (Lewis *et al.*, 2002).

Tabela 1.2: Lista com as principais etapas durante o processo de anotação e as principais ferramentas utilizadas durante o processo de anotação.

Etapas da Anotação	***Ferramentas***	***Referência***
Avaliação de qualidade	*Phred*	Ewing *et al*. 1998b
Limpeza de vetores	*Crossmatch*	Ewing *et al*. 1998a
Montagem ou agrupamento de seqüências	*Phrap*	Ewing *et al*. 1998a
	CAP3	Huang & Madan 1999
Predição *in silico* de genes	*GLIMMER*	Delcher *et al*. 1999
	GeneMark	Borodovsky & McIninch 1993
Análise de similaridade	*BLAST*	Altschul *et al*. 1997
	FASTA	Pearson and Lipman 1988
	Interpro	Mulder *et al*. 2005
Busca de homologias distantes	*HMMER*	Eddy 2003
	SAM	Karplus *et al.* 1998
Alinhamento de seqüências	*Clustalw*	Thompson *et al*. 1994
	T-Coffee	Notredame *et al*. 2000
Análises filogenéticas	*Phylip*	Felsenstein 2005
	MEGA	Kumar *et al*. 2004
Busca de genes ortólogos/parálogos	*OrthoMCL*	Li *et al*. 2003
Vias metabólicas	*KEGG*	http://www.genome.ad.jp/kegg/
Códons usuais (*Codon Usage*)	*Cusp*	Rice *et al*. 2000
Conteúdo G+C	*Geecee*	Rice *et al*. 2000

O processo de anotação se desenvolve em paralelo com a genômica comparativa, sendo que esta se baseia em um princípio bastante simples: características comuns nas seqüências de DNA são compartilhadas entre as espécies (Hardison, 2006). Quando comparamos genomas ou seqüências de organismos filogeneticamente relacionados, procura-se responder a questões simples, como por exemplo, qual a função de um gene e quais as seqüências que diferenciam estas espécies. Entretanto, ao comparar genomas de organismos filogeneticamente mais distantes,

as questões a serem respondidas tornam-se mais amplas, como a organização desses genomas (sintenia), o número de genes pertencentes a um organismo distinto (genes órfãos), vias metabólicas, conservadas e específicas, sendo o último caso capaz de auxiliar na obtenção de fármacos específicos para parasitos com efeito colateral no hospedeiro bastante reduzido (Volker & Brown, 2002), além de também tentar inferir a função de um gene.

Ao realizar um trabalho em genômica comparativa ou a simples anotação de genes, são utilizados dois termos para expressar a semelhança entre duas ou mais seqüências, homologia e similaridade. Entretanto, muitas vezes estes termos são utilizados erroneamente, apesar de existir diferenças entre estes termos. Quando há necessidade de atribuir apenas a semelhança entre duas ou mais seqüências, em termos quantitativos, usa-se o termo similaridade, enquanto para relacionar filogeneticamente as seqüências (Koonin, 2005), ou seja, uma atribuição qualitativa, usa-se o termo homologia.

Existem dois tipos de homologias: a ortologia e a paralogia. Ortólogos são aqueles genes provindos de um gene ancestral presente em uma espécie ancestral das espécies comparadas através de um evento de especiação, além disto, estes genes não possuem as mesmas funções obrigatoriamente. Parálogos são genes encontrados em uma mesma espécie originados a partir de um evento de duplicação, estes genes podem possuir a mesma função após este evento, porém podem não possuir a mesma função posteriormente, geralmente genes parálogos formam famílias gênicas (Koonin, 2005) (Figura 1.5).

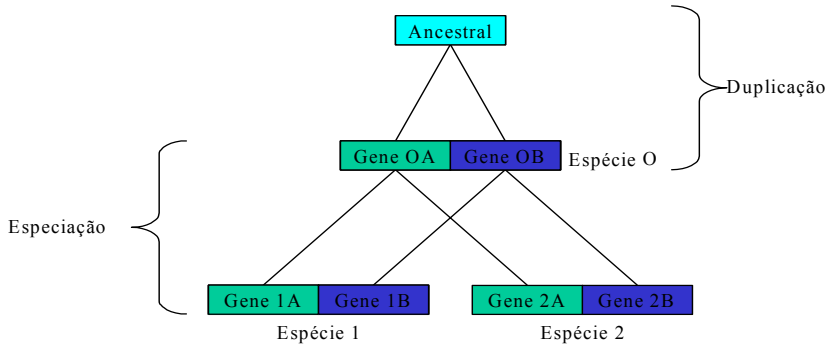


Figura 1.5: Figura demonstrando os eventos relacionados à formação de genes parálogos (Genes OA – OB na espécie O) através de um evento de duplicação, e ortólogos (Genes 1A - 2A e Genes 1B- 2B) por um evento de especiação.

Existem diferentes abordagens para identificar genes homólogos, entretanto há necessidade

dos genomas comparados estarem completamente seqüenciados. Uma das abordagens mais utilizadas é a SymBeT (*Symetrical Best Bi-directional Hit*), onde ao executar o programa "*blastp"* (Altschul *et al*., 1997) genoma contra genoma, dois genes serão considerados ortólogos quando o gene de um genoma for a melhor entrada (similar) e possuir como melhor entrada o mesmo gene no outro genoma (Koonin, 2005) (Figura 1.6).

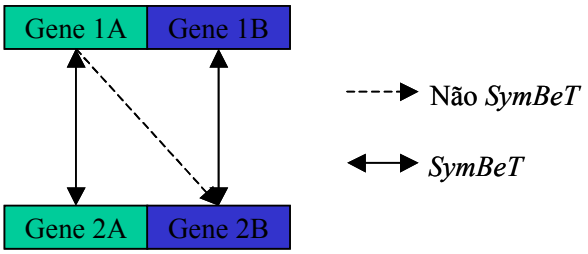


Figura 1.6: Esquema de como se determina um gene ortólogo pela abordagem do *SymBeT*. As setas indicam a melhor entrada, sendo que as bidirecionais indicam onde ocorreu um *SymBeT*, assim os genes que apresentam esta característica são considerados genes ortólogos.

Um exemplo da utilização desta metodologia é o banco de dados COG (*Clusters of Orthologous Groups*) do NCBI[1]. Este banco de dados agrupa seqüências de proteínas de diversos genomas procariotos de acordo com as relações filogenéticas entre seqüências, formando grupos de genes ortólogos através da comparação entre todos os genes (Tatusov *et al.*, 2000). Outro caso é o do KOG (*Eukariotic Orthologous Groups*), que utiliza uma abordagem semelhante ao COG, porém utilizando os genomas de eucariotos disponíveis no GenBank NCBI (Tatusov *et al.,* 2003).

Apesar desta abordagem ser bastante utilizada, existem problemas devido ao tipo de busca de similaridade utilizado, o algoritmo BLAST consegue encontrar seqüências similares quando a similaridade entre elas é alta (por convenção, dois genes podem ser considerados homólogos quando há mais de 30% de similaridade entre suas seqüências protéicas). Assim ao comparar genes homólogos verdadeiros, mas com diferenças significativas entre as seqüências primárias de suas proteínas, esta abordagem não será capaz de identificá-los, estes homólogos são conhecidos como homólogos distantes (Karplus *et al.,* 1998).

Uma das metodologias mais utilizadas para auxiliar o problema de homólogos distantes é a utilização de técnicas de inteligência artificial como os modelos probabilísticos associados à

[1] http://www.ncbi.nlm.nih.gov/cog

seqüência protéica, modelos HMM (Eddy, 2004). Estes modelos são gerados a partir de um alinhamento múltiplo de um determinado gene, onde cada resíduo possui uma probabilidade de ser encontrada na posição específica ao longo da seqüência, então este modelo é confrontado contra um conjunto de dados onde se deseja encontrar os homólogos desse gene (Eddy, 1998). Os programas mais utilizados para esta finalidade são: HMMER (Eddy, 2003) e SAM (Karplus *et al.*, 1998). Estes modelos probabilísticos também são utilizados na identificação *in silico* de genes (Borodovsky & McIninch, 1993; Stanke *et al.*, 2006).

Outra abordagem que vem se discutindo para identificar homólogos distantes é a utilização de seqüências de ancestrais de alguns programas de reconstrução filogenética. Baseados em diferentes métodos evolutivos, como máxima parcimônia e máxima verossimilhança, são preditas as seqüências de cada "nó" de uma árvore filogenética, isto pode ajudar no entendimento de processos evolutivos e função de famílias protéicas (Cai *et al.,* 2004).

Um problema cada vez mais comum nas anotações de diversos genomas é a grande quantidade de "genes hipotéticos", estes são encontrados em organismos de diferentes linhagens filogenéticas, sendo que não se conhece ou não foi determinada sua função biológica, e quando encontrados apenas em organismos relacionados filogeneticamente são chamados de "genes hipotéticos conservados" (Galperin & Koonin, 2004). Já existem iniciativas que visam identificar as funções destes "genes hipotéticos" (Roberts, 2004), foram selecionados àqueles genes encontrados em múltiplos genomas de diferentes linhagens filogenéticas e são possíveis alvos para estudos bioquímicos. Galperin & Koonin (2004) e Roberts (2004) afirmam que o uso conjunto de diferentes algoritmos de busca de similaridade, é essencial na identificação das famílias de "genes hipotéticos", melhorando assim seu entendimento.

Ao tentar descrever a função de um gene, o anotador utiliza termos biológicos conhecidos para fazê-lo. Porém este processo era realizado de acordo com os conhecimentos de cada anotador, assim havia diferenças nas anotações de um mesmo gene, pois eram anotados por diferentes anotadores. Na tentativa de organizar esta multiplicidade de termos e regras, foi criado um consórcio para definir um conjunto de vocabulário controlado, utilizado na anotação dos genes e produtos gênicos, o *Gene Ontology* (GO) (Gene Ontology Consortium, 2006). Este vocabulário é dividido em 3 ontologias: (1) componente celular (2) processo biológico e (3) função molecular. Cada ontologia contém termos e atribuições das seqüências, organizados de forma hierárquica, onde termos de uma ontologia podem estar relacionados a outros em outra ontologia, por exemplo, um gene que possui um termo pertencente a componente celular pode conter mais de um termo dentro de processo biológico que pode apresentar um ou mais termos

relacionado à função molecular[1].

Baseados nas premissas deste consórcio, foram gerados diversos programas que procuram fazer a identificação automática dos termos de uma seqüência (Khan *et al.*, 2003; Chalmel *et al.*, 2005; Jones *et al.*, 2005), entretanto todos possuem limitações, apresentando ainda um número muito elevado de determinações falso-positivas.

Na tentativa de tornar mais rápida as análises comparativas entre diversos genomas e fornecer dados integrados de projetos genoma, transcriptoma e proteoma, vem sendo desenvolvidas plataformas integradas de consulta e busca de genes, como por exemplo, o ESMBL[2]. No Brasil, mais especificamente na Fundação Oswaldo Cruz (FIOCRUZ), há um esforço para a construção de uma plataforma de genômica e transcriptoma comparativa denominada BiowebDB[3] com o apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), no âmbito do qual o presente projeto está inserido. Nesta plataforma dados de diversos genomas que estão sendo seqüenciados por diversos grupos colaboradores, estarão disponíveis para a consulta e análise comparativa entre estes e outros genomas.

[1] http://www.geneontology.org/GO.doc.shtml/#ontologies
[2] http://www.esmbl.org/
[3] http://www.biowebdb.org/

2 – OBJETIVOS

2.1 – Objetivo geral:

Gerar seqüências do genoma da cepa SC58 de *Trypanosoma rangeli*, através da técnica de GSS (*Genome Sequence Survey*) e desenvolver um sistema de anotação genômica objetivando a análise comparativa com os dados dos genomas do *T. cruzi* e do *T. brucei.*

2.2 – Objetivos específicos:

1. Construir biblioteca genômica da cepa SC58 de *T. rangeli* em pUC18 contendo insertos de aproximadamente 1-3Kb;
2. Gerar e anotar 400 seqüências não redundantes do genoma do *T. rangeli*;
3. Desenvolver um sistema de análise e anotação de seqüências genômicas;
4. Comparar e avaliar o polimorfismo nas seqüências obtidas com as seqüências disponíveis dos genomas do *T. cruzi, T. brucei* e de outras espécies de tripanosomatídeos;
5. Quantificar e avaliar o número de genes em *T. rangeli* identificados no presente projeto.
6. Identificar possíveis genes ortólogos entre *T. rangeli* e os genomas completos de *T. cruzi*, *T. brucei* e *L. major.*

3 – MATERIAIS E MÉTODOS

3.1- Crescimento e extração do DNA total da cepa SC58 de *T. rangeli*:

A cepa de *T. rangeli* escolhida para a execução deste trabalho foi a SC58, isolada do roedor *Echymis dasythrix* em Florianópolis, SC (Steindel *et al.,* 1991). Esta cepa é mantida pelo Laboratório de Protozoologia do Departamento de Microbiologia e Parasitologia (MIP) do Centro de Ciências Biológicas (CCB) da Universidade Federal de Santa Catarina (UFSC), gentilmente cedida pelos professores Dr. Mário Steindel e Dr. Edmundo C. Grisard.

A cepa SC58 tem sido mantida através de passagens semanais em meio LIT com 15% de soro bovino fetal (SBF) a 27°C e periodicamente submetida a passagens cíclicas em camundongo/triatomíneo/camundongo.

A extração do DNA total deste parasito foi realizada segundo o método de fenol/clorofórmio (Sambrook & Russel, 2001). Após o crescimento dos parasitos em meio LIT, a cultura contendo parasitos em fase exponencial de crescimento foi centrifugada durante 10 minutos a 2.000 x *g* e lavada 2 vezes com PBS pH 7,2 gelado. Foi adicionado ao precipitado tampão de lise contendo 100 µg de proteinase K (20 mg/ml) e incubado a 42°C durante 3 horas até a degradação total das proteínas. Em seguida foi adicionado um volume igual de fenol equilibrado, sendo a solução centrifugada a 14.000 x *g* em temperatura ambiente. Após a centrifugação a fase aquosa foi transferida para um novo tubo e adicionado igual volume de fenol-clorofórmico (1:1), sendo novamente centrifugada durante 10 minutos a 14.000 x *g* à temperatura de 4°C. Novamente a fase aquosa foi transferida para um novo tubo e adicionada de volume igual de clorofórmio, sendo novamente submetida à centrifugação 14.000 x *g* por 10 minutos à temperatura ambiente. Para precipitar o DNA, a fase aquosa foi transferida para um novo tubo e adicionada de 1/10 do volume da solução de NaCl 3M pH 5,2 e 2,5 vezes o volume de etanol 100%, sendo incubada durante 12 horas a -20°C. Após este período, a solução foi centrifugada durante 30 minutos a 14.000 x *g* à temperatura de 4°C, sendo o sobrenadante descartado e o precipitado lavado com etanol 70%. Foi adicionado 30 µl de água ultrapura estéril ao precipitado contendo DNA e 1µl RNAse durante 1 hora a 37 ^{o}C, para eliminar moléculas de RNA remanescentes.

A dosagem e pureza do DNA foram avaliadas através de espectrofotometria observando-se as absorbâncias a 260nm e 280nm, assim como a relação entre ambas. Após a dosagem, as amostras foram mantidas a –20ºC.

3.2 – Construção da biblioteca genômica:

A biblioteca 101 foi construída a partir da sonicação total do DNA de *T. rangeli*, seguindo o protocolo descrito por Leech *et al.* (2004). Aproximadamente 30μg de DNA total de *T. rangeli* foi fragmentado utilizando sonicador Ultrasonicator Cleaner CL-6 (Branson) durante 10s (potencia total, 10Hz) e os fragmentos gerados foram tratados com 6μl (45U/μl) de *Mung Bean Nuclease* (MBN) durante 10 minutos a 25°C para que todos os fragmentos obtivessem as extremidades abruptas, então foram selecionados os fragmentos entre 1,0-3,0 Kb.

Para a clonagem, foi utilizado 150ng do vetor de clonagem pUC18 (Invitrogen) digerido com 1μl *SmaI* (1U/μl) (New England Biolabs) durante 1 hora a 37°C e desfosforilado com 2,7μl de *Shrimp Alkaline Phosphatase* (SAP) (1U/μl) (New England Biolabs) 30 minutos a 37°C. O vetor e os insertos foram ligados utilizando-se 1 μl de T4 DNA Ligase (1U/μl) (New England Biolabs) por 12 horas a 16°C (indicação do fabricante).

As ligações foram transformadas em células competentes DH5α (*Escherichia coli*), e estas cultivadas em placas contendo 10mL de meio LB-Ágar com ampicilina (200 μg/μl), suplementada com 50μl X-Gal (20mg/μl) e 100 μl de IPTG (200mg/μl), para a seleção dos clones positivos.

As etapas descritas acima foram todas desenvolvidas no Laboratório de Biologia Molecular de Tripanosomatideos (DBBM / IOC / FIOCRUZ) e a biblioteca foi criopreservada no Laboratório de Protozoologia (MIP/CCB/UFSC) na presença de glicerol 33% em placas de 96 poços a -86°C.

3.3 – Extração do DNA plasmidial e seqüenciamento do DNA:

O crescimento dos clones foi realizado em placas do tipo *deep-well* de 96 poços em 1 ml de meio LB com ampicilina (100mg/ml), coberta com adesivo contendo orifícios para a oxigenação das culturas, a 37ºC durante 24h sob agitação de 100rpm. Após isto foi realizada a

extração de DNA plasmidial através da técnica de lise alcalina (*miniprep*) segundo Sambrook & Russel (2001).

Após a verificação da extração através de eletroforese em géis de agarose a 1%, foram feitas as reações de seqüenciamento utilizando o *Kit DyEnamic® ET Dye Terminator* (Amershan Biosciences) de acordo com as especificações do fabricante, com 5 pmol de iniciadores M13 Universal senso (*forward)* (GTA AAA CGA CGG CCA GT) ou anti-senso (*reverse)* (CAG GAA ACA GCT ATG AC) e 500 a 1.000 *n*g de DNA plasmidial, nas seguintes condições: 95°C por 25 segundos, seguidos por 25 ciclos com temperatura de desnaturação a 95°C durante 20 segundos, ligação do iniciador 50°C por 25 segundos e extensão a 60°C por 90 segundos.

Os produtos gerados da reação de seqüenciamento foram precipitados com isopropanol 70% para a retirada dos nucleotídeos e iniciadores não utilizados na reação. Os produtos foram eletroinjetados a 2KV durante 120 segundos e eletroeluídos por 140 minutos a 7KV, em um equipamento *MegaBace 1000® DNA Analysis System* (Amershan Biosciences).

3.4 - Desenvolvimento do Sistema GARSA:

O sistema de análise e anotação de seqüências GARSA (*Genomic Analysis Resources for Sequence Annotation*) vem sendo desenvolvida durante os últimos 3 anos pelo Laboratório de Biologia Molecular de Tripanosomatideos (DBBM / IOC / FIOCRUZ) em parceria com o Laboratório de Bioinformática (MIP / CCB / USFC), Núcleo de Computação Eletrônica (NCE / UFRJ) e Instituto Militar de Engenharia (IME / RJ), sob coordenação do Dr. Alberto M. R. Dávila do IOC/FIOCRUZ.

O principal objetivo deste sistema é facilitar o processo de anotação de seqüências, reduzindo a intervenção de usuários no sistema, cabendo a estes apenas avaliar os resultados fornecidos pelo GARSA e efetuar a anotação das seqüências. Isto é possível pela automatização das análises realizadas pelo sistema através de um organograma (*pipeline*) de programas de bioinformática.

3.4.1 – Desenvolvimento e equipamentos utilizados

Este sistema foi desenvolvido utilizando linguagem de programação PERL (*Pratical Expert Report Language)*, contendo 105 *scripts*[1] fazendo uso de módulos Bioperl 1.4

[1] Arquivos contendo os códigos de programação utilizados para desempenhar uma determinada função do programa

(http://www.bioperl.org/). A interface gráfica do programa foi desenvolvida através de *scripts* em CGI para visualização via navegadores Web, rodando assim sob Servidor HTTP Apache (http://www.apache.org/).

Devido ao grande número de dados, bem como a necessidade de processamento robusto exigido pelos diversos programas utilizados neste sistema, o GARSA foi desenvolvido e seu *pipeline* executado em um servidor Xeon Biprocessado Intel, 4 GB de memória RAM, HD 120 GB e sistema operacional LINUX (Fedora Core 2).

3.4.2 – Banco de Dados GARSA

O GARSA foi desenvolvido para ser um sistema flexível, suportar diversos projetos simultaneamente e armazenar todas as análises geradas pelos programas de seu *pipeline*, bem como as informações inseridas pelos diversos usuários. Para isso, fez-se necessário à utilização de um banco de dados flexível e eficiente. O sistema de gerenciamento de banco de dados (SGBD) escolhido foi o MySQL (http://www.mysql.org) por ser robusto o suficiente e gratuito.

3.4.3 - Desenvolvimento e programas utilizados no *pipeline*

O *pipeline* deste sistema é composto por 21 programas de bioinformática e estruturado de forma que haja comunicação entre eles através de *scritps* em PERL, sem necessidade de interferência do usuário. Além disso, cada programa possui interfaces gráficas para que os usuários possam fazer executar cada programa, bem como a visualização dos resultados, através de scritps CGI. Na tabela 3.2 estão relacionados todos os programas utilizados no *pipeline* do GARSA, bem com sua respectiva função.

Foram integrados a este *pipeline* a utilização de dados provenientes diretamente dos banco Gene_Ontology e EC_Database, para facilitar o processo de anotação. A utilização de dados provenientes do banco Taxonomy também foi incorporada no *pipeline* com o intuito de automatizar a busca das espécies pertencentes às seqüências similares e às seqüências de cada projeto.

A anotação por *Gene Onotology* (GO) foi semi-automatizada neste *pipeline* utilizando as abordagens descritas por Jones *et al.* (2005). Todos os termos da *Gene Ontology* (GO) foram selecionados a partir dos resultados das análises de similaridade contra as base de dados GO (seqdblite-12-2005), uniprot_swissprot, uniprot_trembl e InterPro.

O banco de dados Gene_Ontology possui uma associação entre o número de acesso das seqüências dos bancos citados e os termos da GO. Assim termos associados a cada seqüência

encontrada nas análises de similaridade foram importados pelo sistema e utilizados para a anotação automática da GO.

Tabela 3.2: Programas implementados no *pipeline* GARSA

Programas	***Função***	***Versão***	***Referência***
Phred	Avaliação de qualidade	0.020425.c	Ewing *et al.*, 1998b
Crossmatch	Remoção de seqüências de vetores	0.020425.c	Ewing *et al*., 1998a
CAP3	Agrupamento das seqüências	3.0	Huang & Madan, 1999
*Glimmer3**	Predição de genes	3.0	Delcher *et al*., 1998
*YACOP**	Predição de genes	2.0	Tech & Merkl, 2003
*Critica** [+]	Predição de genes	1.05	Badger *et al.*, 1999
*RBS Finder**[+]	Identificação do sitio de ligação ribossomal (RBS)	1.0	Susek *et al*., 2001
*Zcurve**[+]	Predição de genes	1.02	Guo *et al*., 2003
BLAST	Avaliação de similaridade	2.1.12	Altschul *et al*. 1997
RpsBlast	Busca de domínios conservados	2.1.12	Altschul *et al*. 1997
*Wu-Blast**[+]	Análise de similaridade	2.0	Lopez *et al.,* 2003
*Interpro**	Busca de domínio e famílias protéicas	3.3	Mulder *et al.*, 2005
*HMMER**	Busca de homologia	2.3.2	Eddy, 1998
*geecee**	Conteúdo G+C	2.9.0.6	Rice, *et al*., 2000
*cusp**	Cálculo dos Códons Usuais	2.9.0.6	Rice, *et al*., 2000
*tRNA-Scan**	Identificação de RNA transportados	1.23	Lowe et al., 1997
*Clustalw**	Alinhamento múltiplo	1.83	Thompson, *et al*. 1994
*Muscle**	Alinhamento múltiplo	3.52	Edgar *et al*., 2004
*Probcons**	Alinhamento múltiplo	1.10	Do *et al.,* 2005
*WebLogo**	Visualização dos alinhamentos	2.8	Croocks *et al*. 2004
*Phylip**	Análises filogenéticas	3.61	Felsenstein, 2005

* indicam os programas adicionados durante o andamento desta dissertação; [+] indicam os programas que são dependências do programa YACOP (Tech & Merkl, 2003)

Os termos estão divididos em três categorias ou ontologias: Função Molecular, Componente Celular e Processo Biológico. Uma seqüência pode apresentar mais de um termo dentro de cada categoria, onde geralmente há uma associação entre os termos de uma categoria e

outra. Por exemplo, uma seqüência possui um termo na categoria função molecular e esta função molecular está associada a um ou mais processos biológicos e este está localizado em um componente celular (http://www.geneontology.org/GO.doc.shtml#ontologies, acessado em 19/03/06).

Além das implementações dos programas ao *pipeline*, foram desenvolvidas interfaces gráficas para realizar a anotação manual das seqüências, visualização dos resultados, gráficos contendo informações gerais de cada projeto, como por exemplo, a distribuição dos tamanhos dos cromatogramas e das seqüências não redundantes, além de interfaces que possibilitam o compartilhamento de dados entre os projetos (autorizados por cada administrador).

3.5 – Análise dos resultados

Todas os cromatogramas (*reads*) provindos do seqüenciamento foram submetidos ao *pipeline* do GARSA, onde a qualidade destas seqüências foi avaliada pelo programa *Phred* e as seqüências do vetor foram removidas pelo programa *Crossmatch*. Somente as seqüências que obtiveram qualidade *phred* ≥ 20 e tamanho final ≥ 100 foram aceitas pelo sistema.

Seguindo o *pipeline* do GARSA, as seqüências válidas foram submetidas ao programa *CAP3* para o agrupamento das seqüências similares, formando um conjunto de seqüências não redundantes, denominadas neste trabalho como GSS-nr, e utilizado para as análises posteriores.

Em seguida as seqüências GSS-nr tiveram o conteúdo G+C estimado pelo programa *geecee* do pacote EMBOSS (*European Molecular Biology Open Software Suite*), bem como a busca por seqüências de RNAt pelo programa tRNA-Scan.

A predição de genes foi realizada usando o programa Glimmer3, disponível no *pipeline* do GARSA. Para isso, um conjunto contento 1808 genes de proteínas hipotéticas de *T. cruzi*[1] foi montado para ser usado na construção do modelo IMM (*Interpolated Markov Models*) para a identificação dos possíveis genes, sendo considerados genes àqueles que apresentaram tamanho superior a 150 nucleotídeos.

Análise de similaridade utilizando o programa BLAST foi realizada comparando todas as GSS-nr com as bases de dados descritas na tabela 3.3. Os programas utilizados do pacote BLAST foram escolhidos de acordo com o tipo de base de dados comparada. Desta forma, "*blastn"*, contra bases de nucleotídeos; "*blastx*" contra bases de proteínas, onde as seqüências GSS-nr são

[1] http://www.genedb.org/tcruzi, acessado em 17/03/06

automaticamente traduzidas nas 6 possíveis fases de leituras (*frames*); "*tblastx*" contra base de nucleotídeos, onde tanto as seqüências quanto os bancos são automaticamente traduzidos nas fases de leitura (*frames*).

A busca de domínios e famílias de proteínas foi realizada utilizando o programa *InterProScan* (versão 3.3) e a versão da base de dados Interpro 12.0 (http://www.ebi.ac.uk/interpro, acessado 6/12/05) disponibilizado pelo EBI (*European Bioinformatics Institute*). Foi utilizado o programa *RPSBlast* para a busca nas bases de dados: CDD (*Conserved Domain Database*), Pfam, Smart (*Simple Modular Architecture Research Tool*), COG e KOG (Tabela3. 3).

Em todas as análises que utilizam o pacote BLAST, o ponto de corte de *e-value* foi estipulado em *1e-5,* ou seja, seqüências com *e-value* superior a este valor não foram consideradas nas análises.

As seqüências GSS-nr que não apresentaram similaridade com nenhum dos bancos ou que apresentaram *e-values* não significativo foram re-analisados utilizado o programa disponível no pacote BLAST, chamado "*psiblast*". Este programa promove a busca de similaridade entre seqüências protéicas, a partir de diversas buscas ou interações, utilizando o resultado de cada interação para montar uma matriz de substituição que será usada na interação posterior. Isto possibilita encontrar, após algumas interações, seqüências homólogas, porém bastante divergentes quanto à seqüência primária (Altschul *et al*., 1997). Nesta etapa, analisamos todas as 6 fases de leitura contra o base de dados refseq_protein e Uniref90 utilizando no mínimo duas interações e os resultados foram carregados no banco de dados e utilizados nas análises finais.

Quando encontrado mais de 5 seqüências similares ao final de cada análise por "*psiblast"*, estas também foram utilizadas para a construção de alinhamentos múltiplos, através dos programas *Muscle* (mais de sete seqüências) ou *ProbCons* (menos de 7 seqüências), e estes alinhamentos foram utilizados para a busca de homologia em outros clusters através de modelos de HMM realizados pelo programa *HMMER*.

Também foi realizada a classificação das seqüências utilizando o vocabulário controlado do Consórcio *Gene Ontology* (GO).

Foram realizadas também análises filogenéticas utilizando os programas *Phylip* (através do *pipeline* do GARSA) e *MEGA* (Kumar *et al*., 2004), utilizando respectivamente modelos distancia-*p* e *Kimura-2-parameters,* em ambos foram utilizadas o método de construção deárvore *Neighbor Joinning*, com *bootstrap* 1.000. As seqüências utilizadas nesta etapa foram

selecionadas com base nos resultados de similaridade e os alinhamentos foram construídos com o programa *ProbCon*s.

Após a identificação das regiões codificantes em cada sequência GSS-nr (não necessariamente genes completos), foram confrontadas todas contra todas (utilizando o programa "*blastp"*) para identificar possíveis seqüências repetitivas ou possíveis genes parálogos. Também foi realizada a identificação dos possíveis genes ortólogos compartilhados entre *T. rangeli*, *T. cruzi*, *T. brucei* e *L. major* através do programa *OrthoMCL* (Li *et al.*, 2003). Para isto, foram utilizados todas as seqüências codificantes de *T. rangeli* anotadas neste projeto e todas as 18.939, 7.795 e 8.265 proteínas preditas dos genomas de *T. cruzi*, *T. brucei* e *L. major*, respectivamente.

Tabela 3.3: Tabela contendo as base de dados utilizadas para as análises de similaridade, a descrição de cada base, o tipo de dados, o tamanho em número de seqüências, a versão ou data da última atualização e o endereço *Web* onde as seqüências foram baixadas.

Banco de dados	***Descrição***	***Tipo***	***Tam.***	***Versão***	***Endereço Web***
COG	Constitui uma forma de análise funcional e evolutiva do genoma, apenas de organismos procariotos.	PsMM proteína	4.873	29/09/05	http://www.ncbi.nlm.nih.gov/COG/
KOG	É uma versão do COG para sete genomas eucarióticos completos, *S. cerevisiae*, *S. pombe*, *E. cuniculi*, a planta *A. thaliana*, o nematódeo *C. elegans*, a mosca *D. melanogaster* e *H. sapiens*.	PsMM proteína	4.825	29/09/05	http://www.ncbi.nlm.nih.gov/COG/grace/shokog.cgi
Pfam	Pfam é uma base de dados curado semi-automatizado de famílias e domínios de proteínas, contendo alinhamentos múltiplos de proteínas e modelos HMM	PsMM proteína	7.255	29/09/05	http://www.sanger.ac.uk/Software/Pfam/
SMART	Permite a anotação e identificação de domínios geneticamente móveis e a análise da arquitetura dos domínios	PsMM proteína	663	29/09/05	http://smart.embl-heidelberg.de/
CDD	É uma coleção de alinhamentos	PsMM	11.399	29/09/05	http://www.ncbi.nlm.nih.gov/Structure/cdd/

	múltiplos de seqüências para domínios conservados e proteínas inteiras	proteína			cdd.shtml
Tbrucei_NCBI.Fasta	Contém somente seqüências de proteínas do *T. brucei* referentes ao genomas depositado no GenBank/NCBI	Proteína	7.795	22/11/05	http://www.ncbi.nlm.nih.gov/
Tcruzi_NCBI.fasta	Contém somente seqüências de proteínas do *T. cruzi* referentes ao genoma depositado no GenBank/NCBI	Proteína	18.939	22/11/05	http://www.ncbi.nlm.nih.gov/
Lmajor_NCBI.fasta	Contém somente seqüências de proteínas do *L. major* referentes ao genoma depositado no GenBank/NCBI	proteína	8.265	22/11/05	http://www.ncbi.nlm.nih.gov/
tcruzi-nt.fasta	Contém somente seqüências nucleotídicas de *T. cruzi* depositado no GenBank/NCBI	nucleotídeo	128.761	02/01/06	http://www.ncbi.nlm.nih.gov/
tbrucei-nt.fasta	Contém somente seqüências nucleotídicas de *T. brucei* depositado no GenBank/NCBI	nucleotídeo	107.606	02/01/06	http://www.ncbi.nlm.nih.gov/
lmajor-nt.fasta	Contém somente seqüências	nucleotídeo	1.500	02/01/06	http://www.ncbi.nlm.nih.gov/

	nucleotídicas de *L. major* depositado no GenBank/NCBI				
lbraziliensis-nt.fasta	Contém somente seqüências nucleotídicas de *L. braziliensis* depositado no GenBank/NCBI	nucleotídeo	10.534	02/01/06	http://www.ncbi.nlm.nih.gov/
uniprot_sprot.fasta	É uma base de dados de seqüências de proteínas curado, que fornece um alto nível de anotação, um mínimo nível de redundância e um alto nível de integração com outros bancos.	proteína	184.304	04/10/05	http://www.ebi.ac.uk/swissprot/
uniprot_trembl.fasta	Possui a tradução de todos os CDS presentes no EMBL/GenBank/DDBJ Nucleotide. Complementa o uniprot_sprot	proteína	1.779.481	04/10/05	http://www.ebi.ac.uk/trembl/
uniref90.fasta	Conjunto de grupos formado por seqüências depositadas no banco UnirProt Konwlegdebase, onde as seqüências pertencentes aos clusters compartilham 90% de similaridade.	proteína	187.963	06/01/06	http://www.pir.uniprot.org/database/nref.shtml
euglenozoa-not-kineto-nt.fasta	Seqüências nucleotídicas de organismos pertencentes ao filo	nucleotídeo	731	02/01/06	http://www.ncbi.nlm.nih.gov/

	Euglenozoa, mas que não pertencem ao subgrupo Kinetoplastida que foram depositadas no GenBank/NCBI				
euglenozoa-not-kineto-aa.fasta	Seqüências protéicas de organismos pertencentes ao filo Euglenozoa, mas que não pertencem ao subgrupo Kinetoplastida que foram depositadas no GenBank/NCBI	proteína	759	02/01/06	http://www.ncbi.nlm.nih.gov/
kinetoplastida-aa.fasta	Seqüências protéicas dos organismos pertencentes à ordem kinetoplastida depositadas no GenBank	proteína	76.979	04/10/05	http://www.ncbi.nlm.nih.gov/
ecoli.aa	Conjunto curado de genes de *E coli* K12	proteína	4.289	06/08/05	http://www.ncbi.nlm.nih.gov/
virulence-pathogenicity.fasta	Base de dados de genes responsáveis pela patogenicidade de bactérias e vírus	proteína	369	13/12/05	http://www.ars.usda.gov/research/projects/projects.htm?accn_no=406482
env_nr	Sub conjunto de dados do NCBI onde são depositadas as seqüências genômicas ambientais (metagenômica).	nucleotídeo	994.617	20/11/05	http://www.ncbi.nlm.nih.gov/

minicircles	Contém seqüências dos minicírculo (kDNA) depositadas no GenBank.	nucleotídeo	808	06/08/05	http://www.ncbi.nlm.nih.gov/
repbase1008.fasta	Base de dados de regiões repetitivas.	nucleotídeo	7.588	04/10/05	http://www.ncbi.nlm.nih.gov/
Refseq_protein	Base de dados de proteínas não redundantes do NCBI.	proteína	2.051.681	15/02/06	http://www.ncbi.nlm.nih.gov/RefSeq
Refseq_genomic	Base de dados de seqüências nucleotídicas não redundantes do NCBI.	nucleotídeo	597.921	20/11/05	http://www.ncbi.nlm.nih.gov/RefSeq
kineto-non-tritryps.fasta	Contêm seqüências de kinetoplastidas com exceção das seqüências de *T cruzi, T brucei* e *L major* depositadas no GenBank/NCBI	proteína	1.856	08/12/05	http://www.ncbi.nlm.nih.gov/

4 – RESULTADOS

4.1 – Sistema GARSA

4.1.1 - Desenvolvimento Geral

O sistema GARSA contém 131 arquivos, sendo 31 *scritps* PERL, 74 CGI e 26 arquivos diversos (entre eles arquivos de configuração e figuras), sendo os *scripts* em PERL utilizados para a execução dos programas e os CGI para as interfaces gráficas do GARSA.

Este sistema vem sendo desenvolvido durante 3 anos, onde 6 diferentes programadores já implementaram melhorias neste sistema, como por exemplo, implementação de multi-usuários, multi-projetos, interfaces gráficas, entre outras. Atualmente o *pipeline* GARSA possui 21 programas de bioinformática, sendo 16 dos 21 programas foram implementados durante a execução deste trabalho (Tabela 3.2), como por exemplo os programas de predição de genes, análises filogenéticas, entre outros.

4.1.2 – Banco de dados

Durante a implementação dos diversos programas, fez-se necessário adicionar tabelas ao banco de dados de cada projeto, bem como excluir algumas tabelas que se tornaram redundantes. Atualmente o banco de cada projeto está estruturado com 22 tabelas relacionais como demonstrado na figura 4.1. A descrição da função e a quantidade de campos de cada tabela são apresentadas na tabela A1 na seção de anexos.

Além de haver um banco para cada projeto, foram arquitetados outros 6 bancos de dados para armazenar dados comuns a todos os projetos, apresentados e descritos na Tabela 4.1 e as figuras A2 – A7 na secção de anexos apresentam os esquemas de cada banco.

4.1.3 – *Pipeline* GARSA

O *pipeline* do GARSA (Figura 4.2) pode ser dividido em 5 partes: (1) configuração do projeto e usuários; (2) configuração das bibliotecas; (3) execução dos programas; (4) análise dos resultados; (5) visualização dos dados analisados. Este *pipeline* está representado visualmente na figura 4.2.

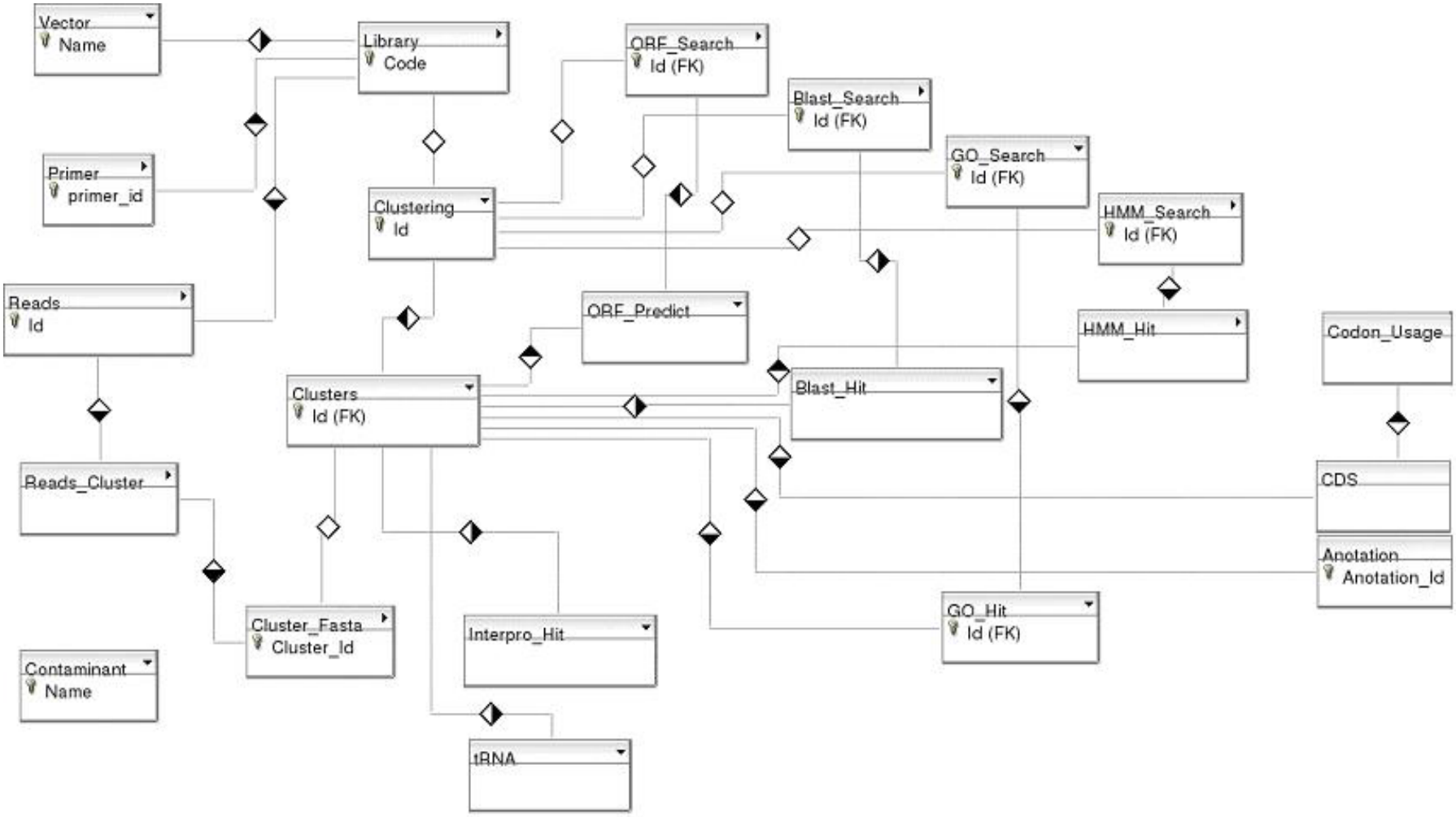


Figura 4.1: Esquema e relacionamento do banco de dados de cada projeto do sistema GARSA. As caixas indicam cada tabela do banco, onde os nomes estão nas áreas escuras de cada tabela. As linhas indicam o relacionamento entre as tabelas: relacionamentos 1:1 estão indicados pelas linhas que contem diamantes claros (◇); os diamantes com metade clara e a outra escura (◆) indicam relacionamento 1: n.

Tabela 4.1: Nome e descrição de cada banco de dados que compõe o sistema GARSA.

Banco	***No. de Tabelas***	***Descrição***
seqonsql	7	Contém as informações de todos os projetos, usuários, e informações compartilhadas entre os projetos
UniVec	2	Armazena as seqüências dos vetores, tanto as seqüências disponibilizadas pelo sistema, quanto às inseridas por cada administrador de cada projeto
Contaminant	1	Armazena as seqüências consideradas contaminantes, como RNAr e seqüências mitocôndriais de alguns organismos para que possam ser excluídas das análises quando selecionadas em um projeto
EC_Database	2	Contém os números e as descrições de cada número EC, utilizado para classificação de enzimas (http://www.chem.qmul.ac.uk/iubmb/enzyme/, acessado em 06/02/06).
Taxonomy	4	Armazena os nomes científicos de algumas espécies do banco de taxonomia do *National Center for Biotechnology* (NCBI), relacionando-os com o código de acesso do GenBank (gi) (http://www.ncbi.nlm.nih.org/taxonomy, acessado em 10/02/06)
Gene_Ontology	27	Banco contendo todos os dados do Gene Ontology Consortium, disponível gratuitamente na página deste consórcio (http://www.geneontology.org/, acessado em 06/02/06)
"projetos" (cada projeto contem um nome de banco distinto)	22	Armazena todas as informações, seqüências, resultados dos programas, das análises e anotações de cada projeto, onde cada projeto possui seu próprio banco de dados.

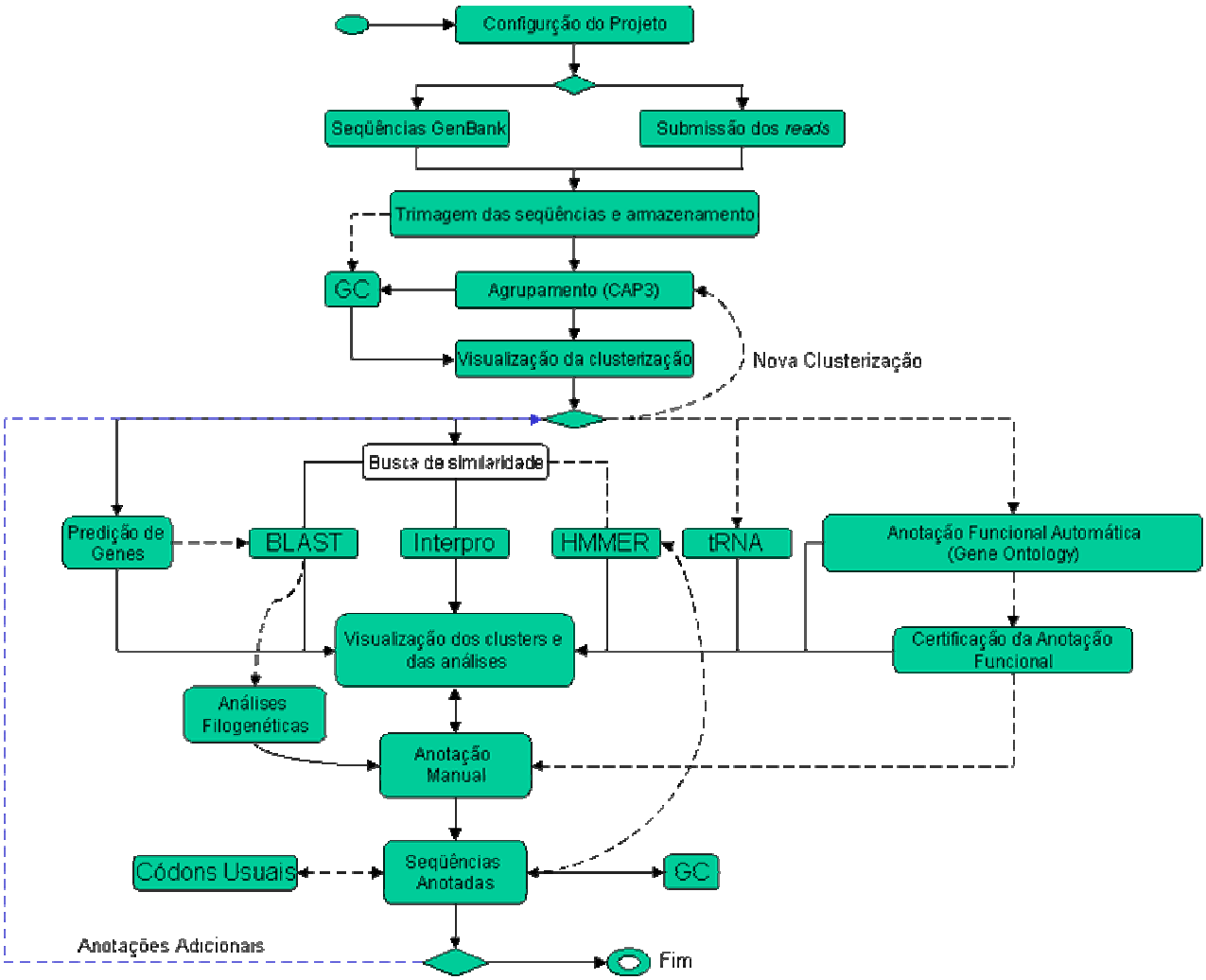


Figura 4.2: *Pipeline* GARSA. As caixas com fundo verde indicam os programas e etapas do *pipeline*; os diamantes indicam as decisões que os usuários tomam durante o *pipeline*; as setas cheias indicam a seqüência mais simples de execução do *pipeline*, ou seja, onde a saída de um programa é visualizada ou utilizada como entrada de outro programa; as setas descontínuas indicam as opções de análises mais acuradas (inseridas exclusivamente durante a execução deste trabalho).

Devemos ressaltar a existência de privilégios dados aos usuários do GARSA, partindo desde a administração do sistema até a simples visualização das estatísticas de um projeto, isto se reflete nas etapas deste *pipeline*. O principal privilégio é aquele exercido pelo administrador do sistema (admin_garsa – *sysadmin*), que possibilita ao usuário criar, apagar e modificar os projetos do sistema, em seguida, o privilégio (*admin*) dos administradores de cada projeto, com permissões de modificar dados referentes aos seus projetos. Devemos ressaltar que estas

modificações foram implementadas durante os 3 anos de desenvolvimento desse sistema com o auxílio de colegas do Núcleo de Computação Eletrônica (NCE / UFRJ) e Instituto Militar de Engenharia (IME / EB) .

4.1.3.1 – Configuração do projeto e usuários

A primeira parte, realizada apenas pelo administrador do sistema, consiste na criação de uma conta projeto no sistema e a configuração do administrador, tornando cada projeto independente. São inseridos no banco de dados administrativo os dados referentes ao novo projeto como o nome, o nome do administrador, o laboratório responsável, a qualidade *Phred* mínima e o tamanho mínimo de cada seqüência aceita.

A partir desta parte do *pipeline* todas as execuções e configurações realizadas por um usuário são restritas a cada projeto, sem interferência nos demais projetos hospedados no sistema. Então, o administrador do projeto pode criar usuários e determinar seus privilégios (Tabela 4.2).

Tabela 4.2: Tabela resumindo os privilégios disponíveis no GARSA e as permissões que estes conferem ao usuário.

Privilégio	***Permissão***
admin	Este usuário é o administrador do projeto, tendo acesso a todos os dados no determinado projeto.
write	Execução de alguns programas, filogenia e "*psiblast"*, além da anotação dos resultados
read	Permite que o usuário veja todos os resultados das análises dos programas, fazer busca no banco de dados, utilizarem as ferramentas comparativas, porém sem a permissão de anotar ou remover qualquer informação
guest	Apenas permite a visualização dos resultados finais das análises
stat	O usuário apenas pode visualizar os dados estatísticos do projeto, não é permitido a visualização de nenhum resultado

4.1.3.2 – Configuração das bibliotecas

Somente o administrador do projeto possui a permissão de criar às bibliotecas no sistema, necessitando informar ao sistema o vetor de clonagem utilizado e/ou as seqüências dos

iniciadores utilizados nas reações de seqüenciamento, bem como o código da biblioteca, nome e descrição da mesma (Anexos figura A8).

A partir disso o usuário poderá carregar as seus cromatogramas (*reads*). Entretanto o GARSA utiliza uma nomenclatura específica para cada cromatograma, como por exemplo, o nome TGEG101007A07.g, as duas primeiras letras indicam o código do projeto (TG), as duas letras seguintes o código do laboratório (EG), os 3 dígitos seguintes, letras ou números, indicam o código da biblioteca (001), 3 dígitos indicando o número da placa (007), e a combinação de uma letra e dois números (A07) refere-se ao posição da placa de seqüenciamento e a letra após o ponto indica a fita seqüenciada (g – *reverse* ou senso e b – *forward* ou anti-senso). Portanto o usuário deverá nomear cada seqüência e o arquivo compactado, de acordo com os padrões do GARSA, para submeter ao sistema. Opcionalmente foi desenvolvida uma interface gráfica para nomear e submeter automaticamente as seqüências ("*Rename and Submit Plate"*).

A cada submissão de uma placa são executados os programas *Phred* para a avaliação de qualidade da seqüência e o programa *Crossmacth* para a limpeza do vetor de clonagem. Após isso, todos os dados e as seqüências limpas ou trimadas são inseridos na tabela *Reads* do banco de dados e o usuário pode visualizar graficamente todas as seqüências através de uma interface gráfica "*Plate Reports"*.

É possível também carregar seqüências diretamente do GenBank, utilizando a interface "*Download from GenBank"* (Anexos figura A8). O administrador do projeto seleciona o tipo de seqüência (GSS, EST, STS e Gene) e o organismo, então o sistema busca as seqüências no GenBank, cria as bibliotecas, avalia o tamanho e submete cada seqüência como se fosse uma seqüência oriunda de um seqüenciamento.

4.1.3.3 – Execução do programas

A terceira parte do pipeline tem início com a execução do programa *CAP3*, para reduzir a redundância do banco de seqüências de cada biblioteca e total (n+1 agrupamentos (*clusterização*), onde n é o número de bibliotecas no sistema), isto agrupará as seqüências similares em um consenso (*cluster*) e identificará as seqüências únicas (*singlets*), originando um conjunto de seqüências não redundantes. Todos os dados são armazenados em 4 tabelas no banco de dados, Clustering, Clusters, Clusters_Fasta e Reads_Cluster.

As seqüências não redundantes (*clusters* e *singlets*) formadas pelo agrupamento total do banco serão utilizados durante todo o processo de análise e anotação, tornando a execução deste programa obrigatória para a continuidade do *pipeline*. Apesar deste programa já ter sido

incorporado no *pipeline* antes do início deste trabalho e sua execução ser simples, não havia a opção para alterar alguns parâmetros que interferem na eficiência do processo de agrupamento, sendo adicionada à interface campos para a alteração dos valores padrões destes parâmetros de acordo com a necessidade do projeto.

Após o término da execução do *CAP3*, o sistema automaticamente executa o programa *geecee* (pacote EMBOSS) para o cálculo do conteúdo G+C de todas as seqüências não redundantes.

A cada nova execução do programa *CAP3* os resultados dos agrupamentos anteriores são excluídos do sistema, havendo a necessidade de executar novamente todos os demais programas.

Com as seqüências não redundantes formadas, todos os demais programas são liberados para a execução. O GARSA não obriga o usuário a executar um programa ou outro, exceto o *CAP3* como salientado anteriormente, entretanto, vamos descrever o pipeline como se todos os programas fossem executados.

Para a predição de genes foram incorporados os programas *Glimmer3* e *YACOP*, ambos adicionados exclusivamente durante a execução deste trabalho. Cada programa possui uma interface gráfica de execução contendo campos para a alteração dos seus principais parâmetros. Deve ser fornecido ao sistema um conjunto de genes conhecidos para a construção dos modelos IMM, que são utilizados para a busca dos genes pelo programa *Glimmer3*. O programa *YACOP* executa 3 outros programas de predição de genes, *Glimmer2.12* CRITICA (requer o programa *Wu-BLAST*) e *Zcurve*, fornecendo como resultado final à combinação destes programas. Os resultados são apresentados em uma tabela na interface "*Show Predicted Genes*" e cada gene predito pode ser visualizado através da interface gráfica "*Gene Details*".

Outro programa disponibilizado no pipeline GARSA é o *tRNA-Scan* que busca seqüências de RNAt (RNA transportadores) nas seqüências não redundantes, mas para isso o usuário deve apenas selecionar, via interface gráfica "*tRNA-Scan*", o modelo de RNAt (Eucarioto, Bactéria, Archea, Mitocondrial/Cloroplasto e Geral) a ser usado pelo programa como padrão de busca. Os resultados são armazenados na tabela RNAt do banco de dados do projeto específico.

Entre os programas de busca de similaridade, foram incorporados ao pipeline os pacotes BLAST, RPSBlast, InterProScan e HMMER. O pacote BLAST foi um dos primeiros a ser implementados no sistema, contudo foram feitas melhorias durante a execução deste trabalho. Através de uma interface gráfica, "*Run Blast*" (Anexos Figura A9), onde o usuário seleciona o programa (*blastn*, *blastx*, *tblastx*), a base de dados para executar o programa e o valor de corte do *e-value* (padrão $1e^{-5}$), o sistema GARSA montará a linha de comando e executará o programa. O

programa *blastp* somente será disponibilizada quando executado algum programa de predição de genes.

As bases de dados são formatadas de acordo com as especificações do pacote BLAST e armazenadas em uma determinada pasta no servidor em que o GARSA está instalado, tornando possível o acesso a estas bases por todos os projetos. Entretanto, cada administrador de projeto pode inserir uma nova base de dados através da interface "*New Blast DB*", assim o sistema formatará e alocará esta base no local apropriado.

Em outra interface, "*Run RPSBlast*", é possível executar o programa de identificação de domínios conservados *RPSBlast*, onde o usuário apenas seleciona uma base de dados das que estão disponíveis no sistema e o valor de corte de *e-value* (padrão < $1e^{-5}$).

Os resultados dos programas do pacote *BLAST* e *RPSBlast* são armazenados na tabela Blast_Hit do banco de dados e na tabela Blast_Search são armazenados os relacionamentos entre bases de dados e a rotina utilizada em cada execução do programa.

O programa para a identificação de famílias e domínios de proteínas *InterProScan* foi também inserido no *pipeline* GARSA durante a execução deste trabalho. Este programa é executado através da interface "*Run Interpro*" apenas executando um comando gráfico, o sistema se encarregará de montar o comando, executar o programa e alimentar os dados no banco de dados na tabela Interpro_Hit.

Também foi disponibilizado o programa para busca de homologia *HMMER*. Na interface para a execução deste programa, "*Run HMMER*", o usuário deve inserir um conjunto de seqüências de genes conhecidos e selecionar o programa de alinhamento (*Muscle* ou *ProbCons*), então o GARSA irá construir um alinhamento múltiplo e este será submetido ao *HMMER*, que por sua vez construirá o modelo HMM e realizará a busca de homólogas ao conjunto de genes iniciais.

Além do programa *HMMER*, o programa *blastpgp* (*psiblast*) está disponível no GARSA. Entretanto, devido à execução desta rotina ser demorada e haver a necessidade de determinar a quantidade máxima de interações, esta etapa é realizada seqüência por seqüência e cada resultado pode ser enviado para a execução do programa *HMMER* ou inserido no banco de dados na tabela Blast_Hit.

A anotação funcional pelo Consórcio *Gene Ontology* ("*Run GO Annotation*") foi implementada neste *pipeline* através da utilização dos resultados de similaridade com as bases de dados uniprot_trembl, uniprot_swissprot, seqdblite (GO) e os resultados do programa InterProScan. As descrições da GO (Termos) associados às seqüências das bases citadas no

banco Gene_Ontology são utilizados para a anotação automática de cada seqüência não redundante. Através de métodos de concordância e distância cada termo recebe uma pontuação e cabe ao usuário definir a melhor anotação funcional para uma seqüência durante a análise dos resultados.

4.1.3.4 – Análise dos resultados

A análise dos resultados de cada seqüência é realizada através de uma interface gráfica contendo os resultados de todos os programas em forma de tabelas e gráficos, possibilitando ao usuário a visualização geral dos resultados dos programas (Anexos figura A10).

Os resultados dos programas *BLAST* e *RPSBlast* podem ser utilizados para a construção de filogenias através do programa *Phylip*. O usuário seleciona as entradas (seqüência similar na base de dados) de uma determinada seqüência não redundante e o tipo de seqüência (nucleotídica ou protéica), o sistema irá montar automaticamente o alinhamento múltiplo utilizando o programa *ClustalW* e solicitará o modelo filogenético, de acordo com o tipo de seqüência selecionado, para executar o método de distância e construir a árvore consenso pelo programa *Phylip*. Automaticamente irá criar também um gráfico do alinhamento múltiplo utilizando o programa *WebLogo*. O usuário poderá analisar todos esses resultados separadamente e utilizá-los na anotação de suas seqüências (Anexos figura A11).

Com relação à anotação das seqüências, uma interface gráfica foi desenvolvida para facilitar este processo (Anexos figura A12). Esta interface está dividida em três etapas: (1) o usuário determina o início e o fim da região da seqüência a ser anotada; (2) o GARSA mostra ao usuário a região selecionada e a seqüência protéica resultante (caso de uma seqüência codificante); (3) informa a descrição (único campo obrigatório), uma anotação funcional para cada categoria funcional da GO (anotações automáticas realizadas pela GO são apresentadas como opções ao usuário), número EC (no caso de enzima), organismo mais similar, base de dados da seqüência mais similar, melhor *e-value* e *score*, além de alguma informação adicional para esta seqüência. Todas estas informações são armazenadas na tabela CDS do banco de dados.

Após a anotação das regiões codificantes de cada seqüência não redundante, o usuário poderá executar o programa *cusp* (pacote EMBOOS) para o cálculo dos códons usuais (*codon usage)* do conjunto distinto das regiões codificantes ou para todos as regiões simultaneamente.

4.1.3.5 – Visualização dos resultados e dos dados analisados

Todos os resultados podem ser consultados e visualizados através de interfaces gráficas intuitivas, por exemplo, na interface "*Blast Hit Queries*" (Anexos figura A13) o usuário possui uma visão geral dos resultados obtidos pelas análises de similaridade, selecionando os resultados de uma ou mais análise realizadas. Outra interface, "*Comparative Queries*", o usuário poderá comparar os resultados obtidos entre duas ou mais base de dados.

Foi desenvolvida uma secção onde estão todos os dados estatísticos de cada projeto (Anexos figura A14), como por exemplo, tamanho médio das seqüências não redundantes, número de seqüências sem entradas, número de seqüências anotadas, conteúdo G+C, entre outros. Nesta mesma secção foram criados gráficos e tabelas exportáveis para o Excel facilitando assim a visualização dos dados de cada projeto.

4.2 – Seqüenciamento e análises da Biblioteca de GSS da cepa SC58 de *T. rangeli*

4.2.1 – Construção e seqüenciamento da biblioteca 101 (*GSS*)

Foram obtidos aproximadamente 800 clones da biblioteca 101, entre 1,0 – 3,0 Kb (Figura 4.3), armazenados em placas de 96 poços contendo solução de glicerol 33% e seqüenciados em duas etapas na Universidade Federal de Santa Catarina. Assim alguns clones foram seqüenciados mais de uma vez e ao final das duas etapas foram geradas 1.720 seqüências (sendo ou *forward* e anti-senso ou *reverse*), somando um total de 716.267 pb (pares de bases) seqüenciados.

4.2.2 – Análises das seqüências

Os 1.720 cromatogramas gerados foram submetidos ao sistema GARSA (Projeto *T. rangeli* GSS) com tamanhos variados (Figura 4.4), foram avaliados quanto à qualidade pelo programa *Phred* e retirada das seqüências o vetor pUC18 (Invitrogen) pelo programa *Crossmatch*. Foram considerados apenas os cromatogramas com qualidade *Phred* ≥ 20 e tamanho superior a 100 pb após a retirada das seqüências de vetor e regiões de baixa qualidade. Das 1.720 seqüências submetidas, 915 (54%) foram aceitos pelo sistema, total de 359.864 pb, com tamanho médio de 209 pb (Figura 4.5).

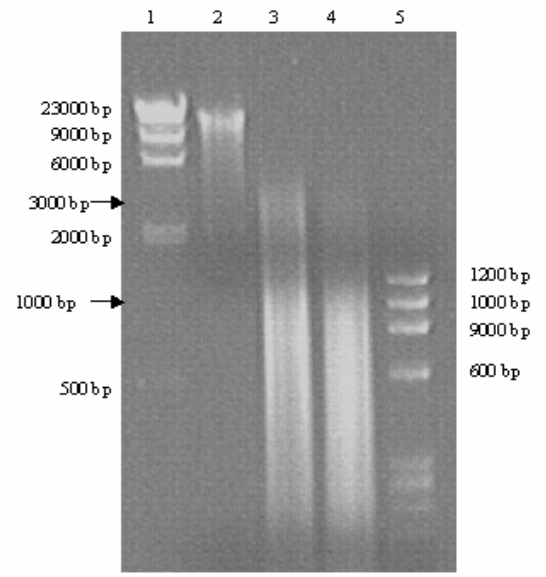


Figura 4.3: Gel de agarose 1% mostrando o DNA de *T. rangeli* SC58 após a sonicação e tratamento com *Mung Bean Nuclease* (1) Marcador λ/HindIII (New England Biolabs, EUA); (2) DNA *T. rangeli*; (3 – 4) DNA *T. rangeli* sonicado e tratado com MBN; (4) Marcador ΦX174 RF/*Hae*III; entre as setas esta o DNA que foi recuperado e usado para a construção da biblioteca.

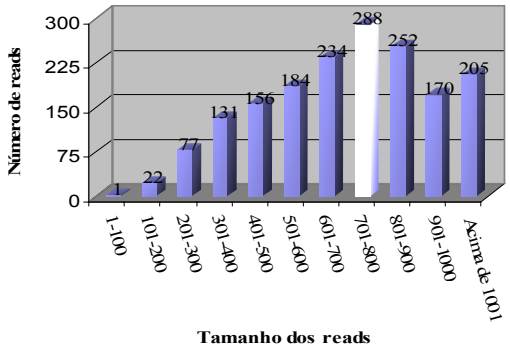


Figura 4.4: Distribuição do tamanho das seqüências submetidos ao GARSA.

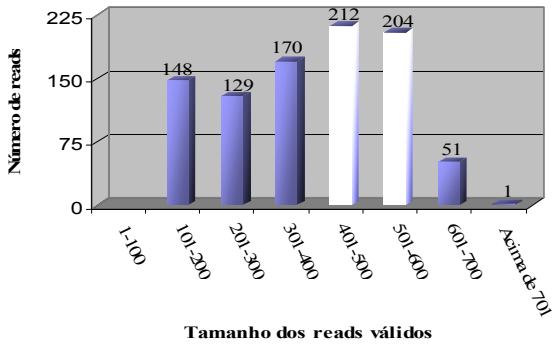


Figura 4.5: Distribuição do tamanho das seqüências após a limpeza de vetor e regiões de baixa qualidade.

Das 915 seqüências limpas que foram submetidas ao programa *CAP3* (via GARSA) para normalizar o banco de seqüências, foram obtidas 375 seqüências não redundantes (GSS-nr) com tamanho médio de 377 pb (Figura 4.6), sendo 258 seqüências únicas e 117 grupos formados por 657 seqüências (Tabela 4.3).

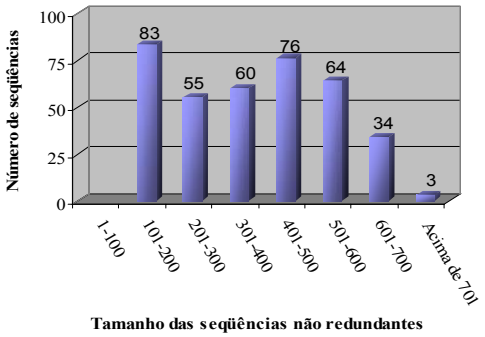


Figura 4.6: Distribuição dos tamanhos das seqüências não redundantes (GSS-nr).

Tabela 4.3: Resumo da quantidade dos cromatogramas submetidos e descartados, seqüências válidas pelo GARSA, únicas (*singlets*), grupos (*clusters*), seqüências não redundantes (GSS-nr) e o conteúdo G+C destas seqüências.

	Total
Seqüências submetidos	1720
Seqüências descartados	804 (46%*)
Seqüências válidas	915 (54%*)
Grupos *(Clusters)*	117
Seqüências em grupos	657
Seqüências únicas	258
GSS-nr	375
Conteúdo G+C	50%

* em relação ao número total de *reads* submetidos ao sistema GARSA

Não foi encontrada nenhuma seqüência de RNAt, mesmo utilizando os 4 modelos disponibilizados pelo programa tRNA-Scan, nas 375 seqüências GSS-nr do projeto.

Foram realizadas predições de genes utilizando um conjunto de 1.808 seqüências de genes hipotéticos de *T. cruzi* para a construção do modelo IMM através do programa *Glimmer3*. Foram encontrados 25 genes com tamanho médio de 219 pb (Tabela 4.4), entretanto ao realizar análises

de similaridade com as seqüências protéicas preditas pelo *Glimmer3*, todos os genes não apresentaram similaridade com nenhuma base consultada.

Tabela 4.4: Resultado da predição de genes pelo programa *Glimmer3*, *score*, conteúdo G+C dos genes, a fase de leitura, o início e fim dos genes (partindo da primeira base da seqüência analisada) e o tamanho de cada gene predito.

Seqüência	***Score***	***G+C***	***Fase***	***Início do gene***	***Final do Gene***	***Tamanho***
TGEG101003F12.g	6,24	54%	-3	281	48	233
TGEG101004B09.g	4,29	55%	-3	413	6	407
TGEG101004E10.g	5,55	47%	-2	418	161	257
TGEG101007C12.g	2,71	54%	3	72	374	302
TGEG101008B05.g	2,08	56%	-3	371	111	260
TGEG101008G07.g	7,15	31%	-1	240	82	158
TGEG101011B04.g	3,51	48%	-1	408	226	182
TGEG101012C12.g	11,56	36%	-1	204	22	182
TGEG101013E12.g	4,44	48%	-3	410	228	182
TGEG101014H06.g	4,86	46%	-3	449	189	260
TGEG101016H06.g	7,06	54%	-1	279	46	233
TGEG101018H06.g	7,75	49%	-1	342	175	167
TGEG101038A02.b	4,84	46%	1	205	363	158
TGEG101038A09.b	4,43	41%	1	124	417	293
TGEG101043B06.b	5,06	52%	3	222	389	167
TGEG101043D10.b	3,25	43%	2	119	412	293
TGEG101043H11.b	7,85	44%	2	80	244	164
TGEG101047H06.b	3,01	44%	3	126	326	200
TGEG102024E12.g	5,37	33%	-1	180	28	152
TGEG102034E10.b	2,72	47%	-1	402	139	263
TGEG102051A01.b	3,93	52%	-2	322	62	260
TGEG102052C02.b	4,85	47%	3	66	371	305
TGEG102053A06.b	2,8	46%	-1	396	154	242
TGEG102053B09.b	3,14	51%	-1	405	232	173
TGEG102053E07.b	2,23	41%	3	117	326	209

Foram realizadas 25 análises de similaridade com as 375 GSS-nr, utilizando diferentes combinações entre os programas do pacote *BLAST* ou *RPSBlast* e bases de dados, com ponto de corte *e-value* $\leq 1e10^{-5}$, totalizando 9.375 análises com 1.681 resultados com pelo menos uma entrada nas base de dados (Tabela 4.5).

Tabela 4.5: Resumo das 25 análises de similaridade realizadas utilizando o pacote BLAST.

Programas	***Base de dados***	***Seqüências com Hits***		***Seqüências sem Hits***	
blastx	euglenozoa-not-kineto-aa.fasta	1	0,26%	374	99,74%
blastn	euglenozoa-not-kineto-nt.fasta	3	0,80%	372	99,20%
blastn	lbraziliensis-nt.fasta	5	1,33%	370	98,67%
rpsblast	Smart	5	1,33%	370	98,67%
blastx	ecoli.aa	7	1,86%	368	98,14%
blastx	env_nr	8	2,14%	367	97,86%
blastx	kineto-non-tritryps.fasta	11	2,94%	364	97,06%
rpsblast	Cog	13	3,46%	362	96,54%
blastn	tbrucei-nt.fasta	17	4.53%	358	95,47%
rpsblast	Pfam	18	4,80%	357	95,20%
blastx	uniprot_sprot.fasta	18	4,80%	357	95,20%
rpsblast	Cdd	23	6,13%	352	93,87%
blastn	repbase1008.fasta	25	6,66%	350	93,34%
rpsblast	Kog	28	7,56%	347	92,44%
blastn	lmajor-nt.fasta	30	8,00%	345	92,00%
blastn	minicircles	34	9,00%	341	91,00%
blastx	Lmajor_NCBI.fasta	101	26,93%	274	73,67%
blastx	uniprot_trembl.fasta	103	27,46%	272	72,54%
blastn	refseq_genomic	143	38,13%	232	61,87%
blastx	Tbrucei_NCBI.fasta	152	40,53%	223	59,47%
blastx	uniref90.fasta	181	48,26%	194	51,74%
blastx	refseq_protein	183	48,80%	192	51,20%
blastn	tcruzi-nt.fasta	188	50,13%	187	49,87%
blastx	Tcruzi_NCBI.fasta	190	51,00%	185	49,00%
blastx	kinetoplastida-aa.fasta	194	52,00%	181	48,00%

Das 375 GSS-nr, 133 (35%) não obtiveram similaridade com nenhuma das bases de dados e 242 (65%) apresentaram pelo menos uma entrada (Figura 4.6a).

A busca de domínios e famílias de proteínas nas 375 seqüências com o programa *InteProScan*, utilizando versão 12.0 do banco de dados, encontrou pelo menos uma entrada em 39 seqüências (9%) (Figura 4.6b).

Algumas seqüências não tiveram nenhuma entrada utilizando os programas do pacote BLAST, porém foi encontrada entrada com o programa *InteProScan*, com isso o número de seqüências sem nenhuma entrada após as análises realizadas com os dois programas foi de 129 (34%) (Figura 4.6c).

Na tentativa de reduzir o número de seqüências sem entradas nas base de dados e procurar por homólogos distantes, as 133 seqüências que não apresentaram entradas com os programas (*blastn* e *blastx*) foram analisadas utilizando a rotina *psiblast* (ponto de corte do *e-value* $<10^{-2}$ e duas interações) contra as base de dados, refseq_protein e Uniref90. Com isso foi possível atribuir similaridade para 10 seqüências, reduzindo o número de seqüências sem nenhuma entrada encontrada para 119 (31%) (Figura 4.6d).

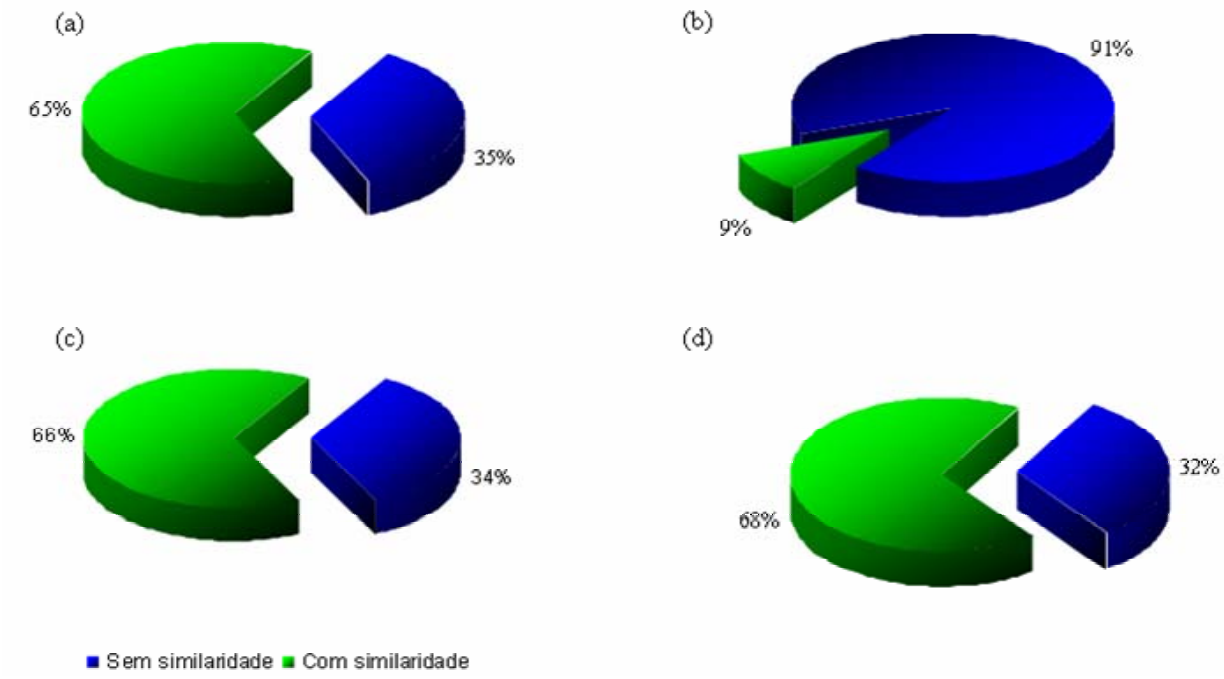
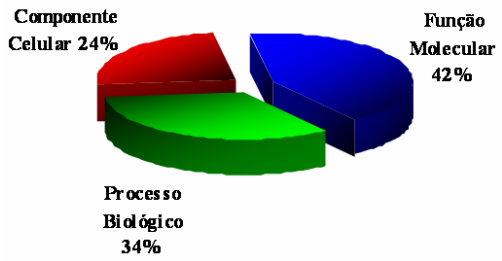


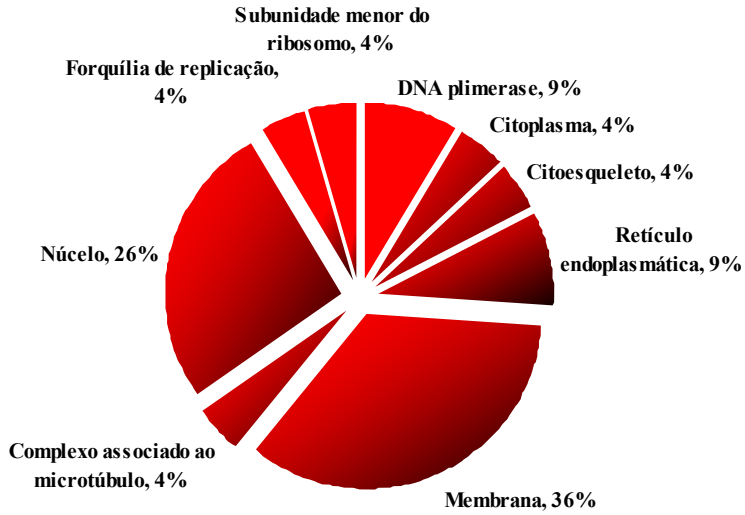
Figura 4.6: Porcentagem de seqüências não redundantes com pelo menos uma entrada significativo (*e-value* $< 10^{-5}$) e sem nenhuma entrada. (a) apenas com o pacote BLAST; (b) apenas com o programa *InteProScan*; (c) combinando ambos os programas antes das análises de *psibast*; (d) com as análises de *psibast*.

Foi também utilizado o programa HMMER na procura de homólogos distantes. Foram utilizados os resultados da saída do programa *psiblast* das 10 seqüências com entradas e conjuntos de seqüências de genes de interesse, como as trans-sialidase, RNA helicase ATP-dependent, alguns genes hipotéticos de *T. cruzi*. Entretanto, não foi encontrada nenhuma seqüência similar entre as 375 seqüências.

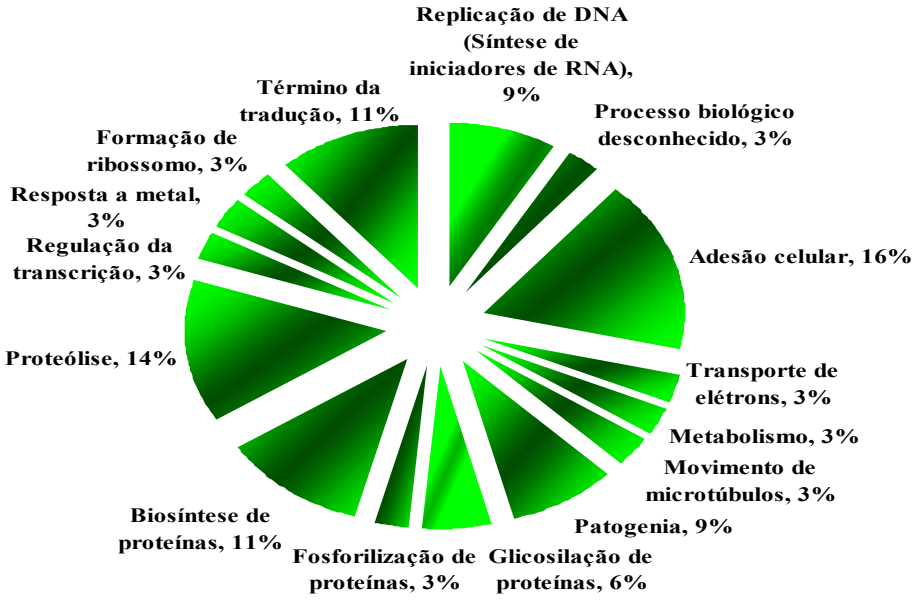
Com relação à classificação funcional das seqüências através da utilização do vocabulário controlado do Consórcio *Gene Ontology*, foi possível atribuir a função para 63 (17%) GSS-nr, pois foi encontrada pelo menos uma classificação (termo) em pelo menos uma categoria funcional (ontologia) para cada seqüência. Entre todos os termos encontrados, 42% estão associados à ontologia Função Molecular, 34% à Processo Biológico e 24% à Componente Celular (Figura 4.7).



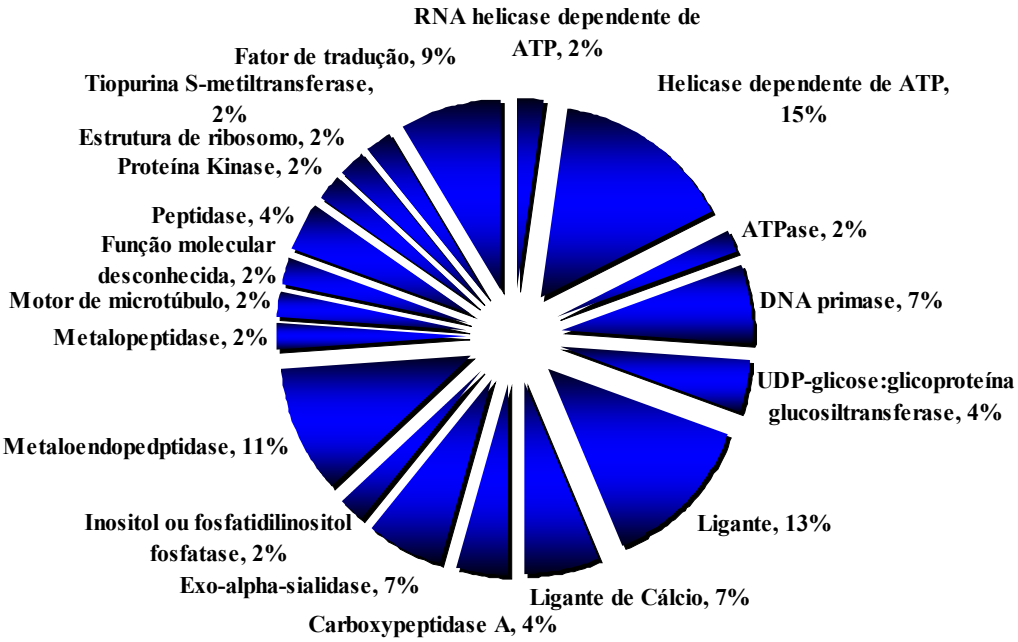
(a)



(b)



(c)



(d)

Figura 4.7: Resultados da anotação funcional pelo Consórcio *Gene Ontology*. (a) porcentagem de termos associados a cada categoria funcional (ontologia) da GO; (b) classificação para a categoria componente celular; (c) processo biológico; (d) função molecular.

Na categoria de componente celular, foram anotados em maior quantidade os termos associados à membrana celular (36%) e núcleo (26%), na categoria de processo biológico os

termos associados à adesão celular (16%) e proteólise (11%), e na categoria função molecular os termos Helicase dependente de ATP (15%), Ligante (13%) e Metalopeptidase (11%).

Após as análises de predição de gene, similaridade, homologia e anotação funcional, foram realizadas as anotações manuais de 242 (65%) GSS-nr, restando 133 (33%) que não puderam ser anotadas devido a dados pouco conclusivos.

Nesta etapa, foram avaliados os resultados de todos os programas de bioinformática simultaneamente através de uma interface gráfica do sistema GARSA. Estas seqüências foram divididas em codificantes com conteúdo G+C de 55% e não codificantes com conteúdo G+C de 35%.

Foi possível atribuir similaridade entre as nossas seqüências a pelo menos 23 genes com função conhecida, entre eles destacam-se RNA helicase ATP-Dependente, proteínas de superfície, como as trans-sialidase e GP63, serino petidases, além de várias seqüências apresentarem similaridade com kDNA, de *T. rangeli* e *T. cruzi*, contendo regiões conservadas e sítios replicativos. Entretanto o maior número de seqüências similares foi de proteínas hipotéticas. A tabela 4.6 apresenta as descrições e quantidade de seqüências para cada anotação. Na tabela A2 dos anexos estão apresentadas todas as anotações realizadas para cada uma das 242 seqüências não redundantes anotadas.

O organismo com o maior número de entradas foi o *T. cruzi* com 86.9% das seqüências similares pertencentes a esta espécie, em seguida seqüências do próprio *T. rangeli* (10,2%), sendo as seqüências de kDNA as principais responsáveis por este valor, na seqüência aparece o *T. brucei* com 2,5% e *L.major* com 0,4% das seqüências anotadas (Figura 4.8).

As seqüências TGEG101043D10.b e TGEG101013C01.g apresentaram 2 regiões codificantes, a primeira contém na região 5', entre os nucleotídeos 1 a 140, similaridade com o final do gene de serino peptidase Bem46-like (gi|70887002[1]) localizada no *locus* Tc00.1047053508461.260 e na porção 3', a partir da posição 391, obteve similaridade com o gene RNA helicase ATP-dependente (EAN99767[1]) localizado no *locus* Tc00.1047053508461.250, ambos localizados no genoma de *T. cruzi* (http://www.genedb.org/). Já a segunda, apresentou dois genes hipotéticos nas *frames* -1 (EAO00000[1]) e -3 (EAN98750[1]).

A seqüência TGEG101008G10.g foi a única em que pode ser determinado uma seqüência completa de um gene hipotético similar a *T. cruzi* (XP_815903[1]), entretanto a sua seqüência é curta.

[1] Código de acesso do GenBank: http://www.ncbi.nih.nlm.gov/entrez

Tabela 4.6: Lista com as descrições e quantidade de seqüências anotadas.

Descrição	***Qtd.***
Gene ribossomal 28S	1
Proteína ribossomal 40S	1
Kinesina-like	1
Antígeno TC40	1
Tiopurina S-metiltransferase	1
Tubulina-tirosina ligase-like	1
Caseina kinase	1
Proteína kinase	1
Serino peptidase	1
Ecotin	2
Proteína de ligação ao RNA	2
UDP-glicose:glicoproteina glicosiltransferase	2
Glicosiltransferase	2
Zinco metalopeptidase mitocondrial ATP-dependente	2
Proteína de transporte vesicular (CDC48 homólogo)	2
Gene ribossomal 18S	3
DNA primase (subunidade menor)	3
TC38	3
Trans-sialidase	3
Proteínas Hipotéticas Conservadas	4
Fator de liberação de cadeia 1	4
GP63	7
Serino peptidase Bem46-like	11
RNA helicase ATP-dependente	26
kDNA	34
Proteínas Hipotéticas	125
Total	**244**

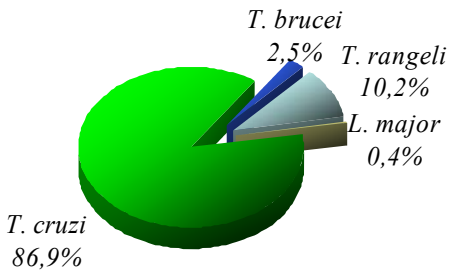


Figura 4.8: Porcentagem organismos com entradas significativas.

Ao comparar os resultados obtidos pelas análises de similaridade contra os genomas de *T. cruzi*, *T. brucei* e *L. major* (*TriTryps*), pode-se observar que 92 (24%) das seqüências não redundantes obtiveram seqüências similares nos 3 genomas. Em apenas 4 seqüências foi encontrada similaridade somente com *L. major* e *T. brucei*, e 4 seqüência apenas com *T. brucei*.

O grande número de seqüências com similaridade nos genomas dos *TriTryps* nos levou a buscar os possíveis genes ortólogos entre estes genomas e as seqüências de *T. rangeli* e os seus genes parálogos. Foi utilizado o programa *OthoMCL* para o confronto entre todos os genes dos genomas dos *TriTryps* e as regiões codificantes anotadas neste projeto para a formação de grupos de ortólogos, que são apresentados na tabela A3 em anexo.

Entre os resultados, destacamos os grupos de ortólogos: (1) ORTHOMCL6769 formado pelas seqüências TGEG102053A06.b de *T. rangeli*, gi|68125416| de *L. major* e gi|70801381| de *T. brucei*, correspondente ao gene Ecotin; (2) ORTHOMCL58 contendo todas as serino peptidases encontradas neste trabalho, além das serino peptidases de *T. cruzi*, *T. brucei* e *L. major*; (3) ORTHOMCL189 formado pelas seqüências de trans-sialidase de *T. rangeli*, (TGEG102052G11.b) e *T. cruzi*, além de um gene hipotético de *T. brucei*.

Também foi possível destacar 4 grupos de parálogos do gene RNA helicase ATP-dependente de *T. rangeli* (ORTHOMCL7914, ORTHOMCL190, ORTHOMCL7927, ORTHOMCL7924), 9 grupos de parálogos de Proteínas Hipotéticos em *T. rangeli* (ORTHOMCL7919, ORTHOMCL7926, ORTHOMCL145, ORTHOMCL7921, ORTHOMCL7917, ORTHOMCL7913, ORTHOMCL7925, ORTHOMCL7922, ORTHOMCL7918).

Um grupo de ortólogos foi formado com todas as serino peptidases, Bem46-like e serino peptidase, descritas neste trabalho (ORTHOMCL58), uma seqüência de Bem46-like de *T. cruzi* e *T. brucei*, além de um serino peptidase de *L.major*.

Para as análises filogenéticas, foi selecionada a região codificante (protéica) de uma GSS-nr que obteve entradas nos genomas dos *TriTryps* e seqüências de kinetoplastideos que não os *TriTryps*. Os alinhamentos foram construídos com o programa *Probcons* e as árvores *Neighbor-Joining* com modelo de distância-*p* (*Phylip*) e *Kimura-2-parameters* (*MEGA*) obtendo os mesmos resultados em ambos os programas (Figura 4.9).

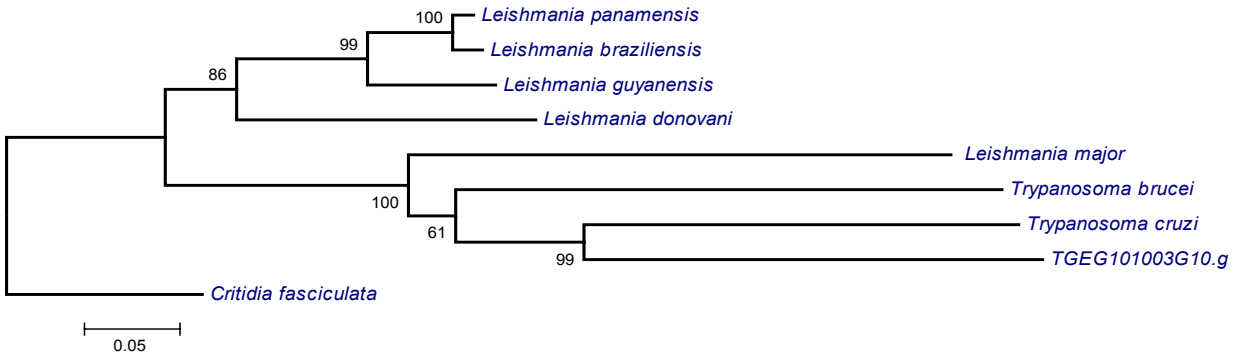


Figura 4.9: Árvore *Neighbor-Joining* (distância-*p*) do gene GP63, construída a partir do alinhamento com seqüência protéica TGEG101003G10.g e as melhores entradas obtidos nos genomas dos *TriTryps* e de outras espécies de Trypanosomatidae.

Todas as seqüências geradas neste projeto, e as respectivas anotações serão disponibilizadas através portal BiowebDB (http://www.biowebdb.org/) assim que este trabalho for publicado.

5 – DISCUSSÃO

5.1- Sistema de anotação

Há duas formas de se analisar um grande número de seqüências: (1) submetendo as seqüências para diversos programas em diferentes "*home pages*" via internet; (2) ou executando diversos programas em um servidor ou máquina localmente.

A primeira opção fornece formas simples e amigáveis de executar e analisar os dados obtidos, porém na maioria dos casos é possível executar apenas uma ou poucas seqüências por consulta, o que torna limitado este tipo de abordagem quando se tem um número elevado de seqüências.

Em contrapartida executando um programa localmente, é possível analisar milhares de seqüências simultaneamente, além de não depender da disponibilidade do serviço oferecido na *web*. Porém, a forma de execução e demonstração dos resultados obtidos nem sempre é clara ou simples, pois a maioria dos programas são executáveis em sistema operacional LINUX e através de comandos extensos, por exemplo, para executar o programa "*blastx*" do pacote *BLAST* contra uma base de dados de proteína o comando mais simples é "blastall –p blastn –d nr –i seq.fasta –e 1e10-5 –o saída.blast", onde –p indica o programa, -d a base de dados para a consulta, -i o arquivo com as seqüências para analisar, -e o ponto de corte do *e-value* e –o arquivo de saída do programa (Altschul *et al*., 1997). Isto requer o mínimo de habilidade com o sistema operacional e o conhecimento básico do funcionamento de cada programa, o que incorre em um aumento do tempo de análise ou o requerimento de recursos humanos especializados.

Assim o uso de ferramentas de informática para semi-automatizar o processo de anotação torna-se útil na análise de grande quantidade de dados biológicos. Há uma grande diversidade de formatos de dados gerados durante uma análise (Rice *et al.*, 2000), muitas vezes necessitando de adaptações pelos usuários para que a saída de um programa possa ser utilizada como entrada em outro programa. Para isso a montagem de um *pipeline* de programas de bioinformática torna-se uma estratégia vantajosa no processo de anotação (Baxevanis & Oullette, 2001), isto faz com que haja a mínima intervenção dos usuários, este não necessitam conhecer a forma de execução de cada programa e executar linhas de comandos em algum sistema operacional, cabendo a ele apenas avaliar a veracidade dos resultados fornecidos pelo sistema, analisá-los e efetuando a anotação das seqüências.

Portanto, o principal objetivo que levou ao desenvolvimento do sistema GARSA foi facilitar o processo de anotação de seqüências genômicas (GSS) e transcriptômicas (EST e ORESTES), através da semi-automatização das análises organizados em um *pipeline* de 21 programas amplamente utilizado pela comunidade científica (Tabela 3.2).

Este tipo de abordagem vem sendo amplamente utilizada e diversos programas já foram desenvolvidos com esta finalidade, por exemplo, *ESTAnnotator* (Hotz-Wagnblastt *et al.,* 2003), *SABIA* (Almeida *et al*., 2004), *ESTIMA* (Kumar *et al.*, 2004), *GATO* (Fujita *et al.*, 2005) e *GENDB* (Meyer *et al*., 2003), entre outros.

O GARSA utiliza para armazenamento das análises de todos os programas e anotações manuais dos usuários um banco de dados relacional *MySQL*, que vem sendo empregado em diversos sistemas de anotação e de análises comparativas (Almeida *et al*., 2004; Gene Ontology Consortium, 2006; http://www.esmbl.org/). Este banco de dados possui 22 tabelas relacionadas, entretanto ainda há a necessidade de normalizar este banco para seu melhor desempenho. Esta etapa de normalização já foi iniciada, tanto que alguns bancos de dados extras foram organizados para armazenar dados comuns aos projetos.

A maioria dos sistemas de anotação são eficientes em um tipo de seqüência transcrita ou genômica, como por exemplo, o *ESTAnnotator* (Hotz-Wagnblastt *et al.,* 2003) e *ESTIMA* (Kumar *et al.*, 2004) são programas que realizam análises de seqüências transcritas, não sendo eficientes nas análises de seqüências genômicas, já no caso do *SABIA* (Almeida *et al*., 2004) e *GENDB* (Meyer *et al.,* 2003), apenas são realizadas análises de dados genômicos. O programa *GATO* (Fujita *et al.*, 2005) pode ser utilizado para as análises de ambos os tipos de seqüências, entretanto, não é permitida a análise de ambos os dados simultaneamente. No sistema GARSA, podem ser analisadas seqüências genômicas e transcritas simultaneamente, pois o sistema permite a decisão do usuário de executar ou não certos programas no decorrer das análises.

Outra vantagem do GARSA é a utilização de seqüências em formato fasta, armazenadas localmente ou oriundas do GenBank, o que é realizado automaticamente (Dávila *et al*., 2005). Não se conhece um sistema de anotação que possui esta facilidade, além de utilizar a combinação desde tipo de seqüência com seqüências provindas de projetos de seqüenciamento.

A manipulação dos dados é feita através de ambientes *web*, devido a todas as interfaces serem desenvolvidas em CGI. Isto possibilita que os usuários acessem o sistema de qualquer computador com acesso à internet. Isto vem sendo empregado em diversos programas, como no *SABIA* (Almeida *et al*., 2004) e no *ESTAnnotator* (Hotz-Wagnblastt *et al.,* 2003), mostrando ser uma forma fácil para se ter acesso aos sistemas. Além disso, os usuário podem alterar parâmetros

de cada programa, uitlizando campos específicos em cada interface de execução, isto torna o sistema GARSA flexível o suficiente para a cobrir as necessidades de cada projeto.

O GARSA tem a capacidade de comportar diversos projetos e vários usuários em um mesmo servidor, onde usuários podem ter acesso a diversos projetos, com permissões distintas. Isto permite o sistema funcionar como um sistema de anotação e uma base de consulta dos dados anotados pela comunidade científica.

A visualização dos resultados durante a execução do *pipeline* é outra vantagem desse sistema. Na maioria dos outros sistemas o *pipeline* é configurado antes de sua execução e dos programas e bases de dados que serão utilizadas no decorrer das análises, entretanto o GARSA possibilita que o usuário execute os programas individualmente e seus resultados são armazenados no banco de dados e podem ser recuperados para a visualizados a qualquer momento pelo usuário, não havendo a necessidade de esperar pelo resultado para prosseguir a execução de outros programas, com exceção do *psiblast.*

Outra vantagem desse sistema é a possibilidade de realizar análises filogenéticas a partir dos resultados de similaridade do pacote BLAST utilizando o pacote *Phylip* (Felsenstein, 2005). Esta facilidade permite que o usuário faça suas análises filogenéticas preliminares diretamente no sistema de forma simples e intuitiva.

Uma das etapas finais do pipeline é a anotação funcional utilizando o vocabulário controlado do Consórcio *Gene Ontology* (Gene Ontology Consortium, 2006), entretanto o uso deste vocabulário para anotar um gene ou seqüência de forma manual é difícil pela quantidade de 1.618.739 possibilidade de anotações (Gene Ontology Consortium, 2006), assim há a necessidade do uso de abordagens para realizar esta tarefa manualmente (Khan *et al.*, 2003; Chalmel *et al.*, 2005; Jones *et al.*, 2005). Nós utilizamos a abordagem descrita por Jones *et al* (2005) que utiliza as anotações da GO de seqüências similares, encontradas pelo programa *BLAST*, para a definição da anotação mais provável. Entretanto esta abordagem deve ser melhorada, pois muitos resultados falso positivos ainda são encontrados e requerem análises manuais. Uma alternativa poderia ser o uso de alinhamentos múltiplos (Chalmel *et al.*, 2005) ou cadeias de Markov (Eddy, 1998).

5.2 – Análise das seqüências GSS de *T. rangeli* geradas no presente estudo

Após a eliminação de seqüências pequenas (< 100 pb) e de baixa qualidade (*Phred* ≤ 20), foram aceitas pelo sistema 915 (54%) seqüências com tamanhos variando entre 100 e 700 pb.

Estes valores estão próximos aos valores obtidos em outros projetos de GSS (Akopyants *et al*., 2001; Guerreiro *et al*., 2005). Dentre as seqüências descartadas, 698 foram devido à baixa qualidade, principalmente durante a primeira fase de seqüenciamento, justificando assim a segunda etapa de seqüenciamento e apenas 107 foram descartadas por possuírem tamanhos inferiores a 100 pb.

O total de bases seqüenciado aceita pelo GARSA foi de 359.864 pb ou 0,35 Mb. O tamanho do genoma de *T. rangeli* ainda não foi estimado, entretanto Toaldo *et al.* (2001) ao realizarem análises do cariótipo desse parasito, observaram uma variação bastante marcante na quantidade (16 a 24) e tamanho (390Kb a 3.130Kb) dos cromossomos entre as diferentes cepas utilizadas no estudo. Se considerarmos o tamanho do genoma de *T. rangeli*, semelhante a estimativa do genoma de este trabalho seqüenciou apenas 0,31% do genoma deste parasita. Entretanto este tamanho deve estar superestimado, pois como a cepa utilizada no seqüenciamento do *T. cruzi* trata-se de um cepa híbrida (Sturm *et al.,* 2003) e a estimativa inicial do genoma haplóide de *T. cruzi* é de 48MB (Cano *et al.*, 1995; Degrave *et al.,* 2001), podemos estimar que foi seqüenciado cerca de 0,77% do genoma de *T. rangeli.*

No presente estudo, grandes quantidades de seqüências foram redundantes, pois das 915 seqüências aceitas pelo sistema, 657 foram agrupadas em seqüências consensos (*clusters*), isto pode ter sido gerado pelo seqüenciamento de um clone mais de uma vez.

O conteúdo médio G+C destas seqüências não redundantes (GSS-nr) foi de 50%. Quando avaliamos apenas as regiões codificantes este conteúdo foi de aproximadamente 55%, próximo aos 53,4% encontrado em regiões codificantes no genoma de *T. cruzi* (El-Sayed *et al.,* 2005a) e aos 54% encontrado em transcritos de *T. rangeli* (Rodrigues, 2005) e diferente de *T. brucei* e *L. major* que respectivamente apresentam 50,9% e 62,5% (Berriman *et al*., 2005; Ivens *et al.,* 2005).

Foram preditos 25 genes pelo programa *Glimmer*, mas os *scores* destes genes são baixos e o tamanho médio de 219pb está abaixo dos observados em *T. cruzi*, e *T. brucei*, respectivamente de 1.151pb e 1.242pb (El-Sayed *et al*., 2005a; Berriman *et al*., 2005). Portanto, o encontro destes genes podem ser considerados falso positivos ou genes ainda não caracterizados, pois não foi verificada similaridade destes genes com nenhuma seqüência nos genomas dos *TriTryps* e outros genomas, assim mais análises com estes genes devem ser realizadas.

Não foi verificada a presença de nenhum RNA transportador (tRNA) entre as 375 GSS-nr, o que pode ser esperado, pois a quantidade destes genes em relação ao total de genes é relativamente baixa. Em *T. cruzi* foram identificados 115 genes relacionados a tRNA, entre 23.216 genes identificados (El-Sayed *et al*., 2005a). Da mesma forma em outros projetos que

utilizaram a abordagem GSS também não identificaram RNAt em suas análises, (Agüero *et al.,* 1996; El-Sayed *et al.,* 1997; Smith *et al*., 1998; Akopyants *et al.,* 2001) sugerindo a baixa freqüência destes.

A quantidade de seqüências sem similaridade em nenhuma das análises de similaridade e homologia (31%) foi abaixo da média de 60% obtida em outros trabalhos (Peterson *et al*., 1993; Agüero *et al.,* 1996; El-Sayed *et al.,* 1997; Cliften *et al.,* 2001; Laurentino *et al*., 2004). Mesmo após o uso de programas mais sensíveis, como HMMER e *psiblast*, a quantidade de seqüências sem entradas significativas não reduziu muito, demonstrando que estas podem ser exclusivas desta espécie ou ainda não foram descritos, conseqüentemente não estarão em nenhuma base de dados utilizadas neste trabalho.

Entre as seqüências com similaridade houve um grande número de seqüências hipotéticas ou hipotéticas conservadas, refletindo as anotações realizadas neste trabalho, onde 125 seqüências foram anotadas como proteínas hipotéticas e 4 como hipotéticas conservadas, isto representa 53% das anotações realizadas neste projeto.

O grande número de seqüências hipotéticas pode estar relacionado com a grande quantidade de proteínas hipotéticas anotadas nos genomas de *T. cruzi* (El-Sayed *et al*., 2005a), pois 86,9% das seqüências não redundantes foram similares a seqüências do genoma desta espécie (Figura 4.8). Portanto estas seqüências hipotéticas merecem análises mais acuradas, pois segundo Galperin & Koonin (2004) estes genes são importantes alvos de análises experimentais para a caracterização de funções específicas de cada espécie.

A anotação funcional automática, baseada no vocabulário controlado do Consórcio *Gene Ontology*, foi possível de ser realizada em 63 seqüências, número bastante baixo em relação à quantidade de seqüências anotadas. Entretanto, outros trabalhos também têm apresentado baixo número de seqüências anotadas automaticamente utilizando este vocabulário, podendo ainda conter algumas anotações falsas (Jones *et al.,* 2005).

O vocabulário da GO está dividido em 3 principais categorias: (1) Função Molecular, (2) Componente Celular e (3) Processo Biológico. Segundo Ashburner *et al.* (2000), a função molecular é definida como a atividade bioquímica de um produto gênico, o componente celular é o local da célula em que o produto do gene está ativo e o processo biológico é metabolismo em que estes genes estão inseridos.

As anotações automáticas utilizando os termos da GO foram bastante similares às anotações manuais realizadas, onde a maior parte das anotações dentro da categoria de função molecular está relacionada com RNA helicase, serino peptidases e proteínas ligantes (Figura

4.7c). Isto refletiu também na categoria componente celular, com grande quantidade de anotações associadas à membrana e núcleo. Já na categoria de Processo biológico, apesar de prevalecer às anotações de adesão celular e proteólise, houve um número considerável de seqüências anotadas apenas dentro do processo patogenia, reflexo da quantidade de proteínas de superfície GP63 e peptidases associadas a processos patogênicos (Burleigh & Woolsey, 2002; El-Sayed *et al*., 2005a).

Entretanto, houve um número considerável de redundância nas anotações, isto é, algumas seqüências tiveram as mesmas seqüências similares em diversas bases de dados, resultando o elevado número de grupos de parálogos encontrados (Tabela A3). Por exemplo, o grupo ORTHOMCL145 é formado apenas por seqüências de proteínas hipotéticas de *T. rangeli*, o que segundo Koonin (2005) nos levaria a classificar como genes parálogos, entretanto, todos as seqüências formadas neste grupo apresentaram as mesmas seqüências similares no genoma de *T. cruzi*.

Este fato levanta duas hipóteses: (1) erros no momento do agrupamento das seqüências, onde os parâmetros utilizados não foram suficientes para agrupar todas as seqüências redundantes em um mesmo grupo ou (2) regiões repetitivas nas regiões 5' e/ou 3' das seqüências, que dificultam a etapa de agrupamento.

O genoma de *T. cruzi* contém aproximadamente 50% de seqüências repetidas, como retrotransposons, regiões teloméricas e genes de famílias de proteínas de superfície, como as pertencentes ao grupo das metaloproteases, gp63 e Zinco metalopeptidase mitocondrial ATP-dependente, as trans-sialidades, mucinas, entre outras. Além destas, outros genes são encontrados em grande número de cópias neste genoma, como RNA helicase, glycosiltransferases, algumas proteínas Kinases (El-Sayed *et al*., 2005a).

Neste trabalho a maioria das seqüências encontradas pertence às seqüências com grande número de cópias de genes por genomas, como as moléculas de superfície celular, gp63 e trans-sialidase, RNA helicase, algumas proteínas hipotéticas (Tabela 4.4). Desta forma, isto pode explicar a redundância das seqüências, pois as famílias gênicas nos genomas dos *TriTryps* são encontradas em regiões teloméricas (El-Sayed *et al*., 2005b). Este fato ocorreu na montagem do genoma de *T. cruzi*, algumas seqüências que seriam trans-sialidase não foram agrupadas com as devidas regiões repetitivas (El-Sayed *et al*., 2005a).

A superfamília das trans-silidases (TS) em *T. cruzi* é composta por 1.430 genes, sendo 693 pseudo-genes, entretanto há a formação de dois grupos, um composto por TS ativas e outro por TS inativas. As TS ativas possuem a capacidade de transferir o ácido siálico de

glicoconjugados da superfície celular do hospedeiro para moléculas de mucinas do parasito (Frasch, 2000). O mesmo autor afirma que as TS são divididas em 4 grupos, de acordo com a similaridade entre as seqüências, função e peso molecular de suas proteínas. No primeiro grupo estão as trans-sialidase de *T. cruzi* (TcTS) e as de *T. rangeli* (TrSial).

TrSial são diferentes das TcTS, enquanto que as TcTS possuem a capacidade de transferir o ácido siálico da superfície celular do hospedeiro para o parasito, a TrSial perdeu a capacidade de transferir para a superfície do parasito, apenas retirando da superfície celular do seu hospedeiro, tendo um papel ainda desconhecido nos hospedeiros vertebrados (Añez-Rojas *et al.,* 2005). Rodrigues (2005) encontrou em formas tripomastigotas de *T. rangeli* um grande número de trans-sialidase em suas ORESTES, cerca de 15% das seqüências. Neste trabalho encontramos 3 seqüências de TS (Tabela 4.4), sendo que na seqüência TGEG101001B01.g foi encontrado o domínio de trans-sialidase pelo *CDD* e pelo *InterProScan*, bem como um homólogo desta enzima nos genomas de *T. cruzi* através da formação de um grupo de ortólogo ORTHOMCL189 (Tabela A3), demonstrando mais uma vez que esta enzima merece novos estudos na tentativa de identificar sua função no ciclo biológico deste parasito.

Além das trans-sialidases, outras proteínas de superfície foram encontradas neste trabalho em grande quantidade como as GP63, o que era esperado, pois no genoma de *T. cruzi* foram encontrados mais de 400 genes compondo esta família gênica (El-Sayed *et al*., 2005a). Também houve a formação de alguns grupos de ortólogos entre seqüências de GP63 de *T. rangeli* descritas neste trabalho com genes de *T. cruzi* (Tabela A3), sugerindo a similaridade entre os constituintes da superfície destes parasitos.

Grande parte das seqüências anotadas são de genes encontrados no kDNA, isto pode sugerir que uma considerável parte dos clones obtidos provém de ligação desta região. Entre estes genes destacamos RNA helicase ATP dependente que fazem parte do processo de edição do RNAm na mitocôndria (Stuart *et al*., 2005). Foram encontrados os domínios funcionais desta enzima, DEAD/DEAH box helicase, em pelo menos uma base de domínios em 3 seqüências, corroborando com os dados propostos por Stuart *et al*. (2005), bem como os domínios apenas descritos como ATP-dependent RNA helicase domain, em mais 3 seqüências.

Podemos perceber que houve a formação de 4 grupos de parálogos referente a este gene, sugerindo uma possível família gênica, bem como a formação de 2 grupos de ortólogos. O primeiro grupo (ORTHOMCL4183) formado pelas seqüências TGEG101043F05.b de *T. rangeli*, gi|68128080 de *L. major*, gi|70801072 de *T. brucei* e gi|70874282 de *T. cruzi* e o segundo (ORTHOMCL686) formado pelas seqüências TGEG101003H06.g de *T. rangeli*, gi|70905988 de

L. major, gi|70831538 de *T. brucei* e gi|70887001 de *T. cruzi*. Este resultado vem ao encontro com o que foi encontrado nos genomas dos *TriTryp*s, onde as RNA helicase são compostas por famílias gênicas (El-Sayed *et al*., 2005a) e mostrando que há homologia entre os genomas dos *TriTryps* e *T. rangeli*.

Dentre as peptidases, a encontrada em maior número foi a serino peptidase Bem46-*like* e uma serino peptidase. Ao realizar uma busca das serino peptidases, em específico as Bem46-*like,* no GeneDB (http://www.genedb.org/, acessado em 04/04/06), percebemos que estas peptidases são uma família gênica presente em *T. cruzi*. No nosso trabalho foram identificadas 11 seqüências codificantes desta proteína, entretanto, foi formado apenas um grupo de ortólogo contendo todas estas proteínas, e seus respectivos homólogos em *T. cruzi*, *T. brucei* e *L. major* (ORTHOMCL58 (Tabela A3)).

O gene da Ecotina (TGEG102053A06.b), identificada pela primeira vez neste trabalho em *T. rangeli*, está relacionada com a inibição de algumas serino peptidases, mais especificamente as tripsina/quimiotripsina serino peptidases, e principalmente em bactérias (Erpel *et al*. 1992; Pál *et al*. 1994). Esta proteína Ecotina também foi encontrada nos genomas dos *TriTryps,* entretanto estes genomas não possuem genes codificantes para estes tipos de serino peptidase. Desta forma, a presença deste gene pode desempenhar um papel essencial na interação parasito/hospedeiro, uma vez que estas peptidases são abundantes em humanos e insetos (Ivens *et al.,* 2005). O mais notável é a formação de um grupo de ortólogo (ORTHOMCL6769, Tabela A3) entre os genes de Ecotina de *T. rangeli*, *T. brucei* e *L. major*, sem a presença de *T. cruzi*, podendo ser um alvo de estudo para entender as diferenças moleculares envolvidas no ciclo biológico no hospedeiro vertebrado entre *T. rangeli* e *T. cruzi*.

Na seqüência TGEG101043D10.b podemos observar a organização de dois genes no genoma de *T. rangeli*, uma serino peptidase Bem46-like e uma RNA helicase ATP dependente. Nesta seqüência foram encontradas a porção final do gene Bem46-like na porção 5', um espaço de 251pb até o inicio do gene RNA helicase, esta ordem também foi encontrada nos genomas de *T. cruzi* e *T. brucei* (http://www.genedb.org/, acessado em 04/04/06), apenas com um região intergênica de 618 pb e 395 pb, respectivamente. Este resultado vai ao encontro com os resultados obtidos por El-Sayed *et al.* (2005b), onde foi observado uma conservação acentuada na sintenia entre os genomas dos *TriTryps*, como anteriormente proposto por Ghedin *et al.* (2004). Portanto, muitos estudos podem ser realizados com o intuito de identificar genes, especialmente em *T. rangeli,* baseando-se na organização dos genes em *T. cruzi* ou *T. brucei*, como proposto por Guhl (2001).

Vários autores vêm discutindo a posição taxonômica do *T. rangeli*, alguns classificando este parasito como pertencente à secção Stercoraria (Stevens & Gibson, 1999; Briones, 1999; Hughes & Piontkivska, 2003), portanto mais relacionado filogeneticamente com *T. cruzi* enquanto outros o classificam dentro da secção Salivaria (Amorin *et al*., 1993, Henriksson *et al.,* 1996), filogeneticamente mais próximo a *T. brucei*. Entretanto, a maioria destes trabalhos utilizaram marcadores filogenéticos clássicos, como o gene ribossomal 18S, porém quando a quantidade de organismos é pequena, árvores com este marcador tornam-se pouco confiáveis pelo fato de haver um desequilíbrio das taxas evolutivas entre os clados (Stevens & Gibson, 1999).

Para realizar análises filogenéticas conclusivas e sólidas a escolha do grupo externo é um ponto crucial (Stevens & Gibson, 1999). Moreira *et al*. (2004) utilizando como grupo externo seqüências ambientais, realizaram a reconstrução filogenética da Ordem Kinetoplastida, mostrando que estas amostras podem servir como bons grupos externos para avaliar a relações filogenéticas dos tripanosomatídeos. Além disso, Stevens & Gibson (1999) mostram a formação de um grupo, chamado "clado aquático", compreendendo os tripanossomos aquáticos.

Portanto, procuramos utilizar diferentes seqüências para realizar nossas árvores filogenéticas e procurar ajudar na classificação do *T. rangeli*. Pode-se observar que utilizando genes de superfície como GP63 e como grupo externo *C. fasciculata*, houve o agrupamento de *T. cruzi* e *T. rangeli*, entretanto a distinção deste grupo com *T. brucei* não foi bem estabelecida (Figura 4.9). De acordo com Baldauf (2003) utilizar um grupo de genes ordenados de forma concatenada para realizar análises filogenéticas pode ser uma boa alternativa para resolver filogenias controversas, assim novas análises com este gene e outros já identificados, de forma concatenada, devem ser realizadas a fim de compreender a posição taxonômica de *T. rangeli.*

Até o presente momento nenhum trabalho com a finalidade de seqüenciar o genoma de *T. rangeli* foi desenvolvido, portanto este trabalho pode ser considerado como o primeiro com o objetivo de explorar em maior escala o genoma desta espécie.

Este trabalho, juntamente com os trabalhos de transcriptoma desenvolvidos (Snoeijer *et al*., 2004; Rodrigues, 2005; Stoco *et al.*, 2005) e em desenvolvimento na Universidade Federal de Santa Catarina (UFSC) em colaboração com o IOC/FIOCRUZ, serão de grande importância no entendimento da biologia e evolução do *T. rangeli*.

Novas bibliotecas genômicas da cepa SC58 de *T. rangeli* estão sendo geradas com objetivo de obter um maior número de seqüências não redundantes do genoma do parasito para identificar possíveis genes metabólicos conservados em algumas espécies da Ordem Kinetoplastida. Além

disso, através da plataforma GARSA os dados obtidos das bibliotecas de GSS com as de EST e ORESTES serão analisados em conjunto, permitindo a exploração em larga escala do genoma desta espécie.

Também se encontra em curso o trabalho para finalizar a segunda versão do sistema GARSA, com a incorporação de novas ferramentas e melhorias no seu *pipeline*, além de tornar-se uma versão para análises de seqüências protéicas e SNPs (*Single Nucleotide Polymorphisn*).

6 – CONCLUSÕES

1 – O sistema de anotação GARSA provou ser uma excelente alternativa para análises de grande quantidade de seqüências genômicas, apresentando-se como uma solução intuitiva, escalonável, adaptável e semi-automatizada para a análise de seqüências.

2 – A estratégia de seqüenciamento por GSS demonstrou-se bastante eficiente na obtenção de seqüências codificantes do genoma do *T. rangeli*, principalmente de famílias gênicas, bem como na identificação das regiões intergências.

3 – A maioria (86,4%) das seqüências não redundantes do genoma do *T. rangeli* geradas no presente projeto foram similares a seqüências de *T. cruzi* e uma pequena parcela (2,5%) foi similar à *T. brucei*.

4 – Foram encontradas muitas seqüências relacionadas a moléculas de superfície deste parasito como a trans-sialidade e a ecotina, as quais apresentam grande relevância para o estudo da biologia deste parasito.

5 – A utilização dos programas para a detecção de homologias distantes, como *HMMER* e *psiblast* demonstraram ser bastante útil na identificação de genes em seqüências que não apresentaram similaridade nas análises com os programas convencionais.

6 – O encontro de genes completos neste trabalho sem similaridade com genes de outros organismos sugere que estes genes podem ser específicos de *T. rangeli*.

7 – A busca de novos genes em *T. rangeli* pode ser baseada na organização dos genes em *T. cruzi,* pois a identificação de seqüências sintênicas neste trabalho está de acordo com trabalhos de análise de sintenia entre os genomas dos *TriTryps*.

8 – A identificação de possíveis genes ortólogos com os *TriTryps* sugere um nível de conservação bastante grande entre os genomas destas espécies com o *T. rangeli.*

9 – Estes resultados são os primeiros dados genômicos do organismo *Trypanosoma rangeli*, tendo uma importância fundamental para o caráter evolutivo e biológico desta espécie.

7 – REFERÊNCIAS BIBLIOGRÁFICAS

Acosta L, Romanha AJ, Cosenza H, Krettli AU. Trypanosomatid isolates from Honduras: differentiation between *Trypanosoma cruzi* and *Trypanosoma rangeli*. *Am. J. Trop. Med. Hyg*. 1991; 44(6): 676 – 683.

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, *et al*.. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* 1991; 252: 1651-1656.

Afchain D, Le Ray D, Fruit J, Capron A. Antigenic make-up of *Trypanosoma cruzi* culture forms: Identification of a specific component. *J. Parasitol.* 1979; 65: 507-514.

Agüero F, Verdún RE, Frasch ACC, Sanchez DO. A Random Sequencing Approach for the Analysis of the *Trypanosoma cruzi* Genome: General Structure, Large Gene and Repetitive DNA Families, and Gene Discovery. *Genome Res.* 1996; 10:1996–2005

Akopyants NS, Clifton SW, Martin J, Pape D, Wylie T, Li L, *et al*.. A survey of the *Leishmania major* Friedlin strain V1 genome by shotgun sequencing: a resource for DNA microarrays and expression profiling. *Mol. Biochem. Parasitol.* 2001; 113: 337–340.

Almeida LG, Paixao R, Souza RC, Costa GC, Barrientos FJ, Santos MT, *et al*.. A System for Automated Bacterial (genome) Integrated Annotation--SABIA. *Bioinformatics* 2004; 20:2832-2833.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J. Mol. Biol*. 1990; 215:403-410.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 1997; 25:3389-3402.

Amorin MI, Momen H, Traub-Cseko YM. *Trypanosoma rangeli*: sequence analysis of β-tubulin gene suggests closer relationship to *Trypanosoma brucei* than to *Trypanosoma cruzi*. *Acta Tropica* 1993; 53: 99 – 105.

Añez-Rojas N, Peralta A, Crisante G, Rojas A, Añez N, Ramirez JL, Chiurillo MA. *Trypannosoma rangeli* express a genes of the group II trans-sialidase superfamily. *Mol. Biochem. Parasitol*. 2005; 142: 133 – 136.

Aziz A e Blaxter M. Hox gene evolution in nematodes: novelty conserved. *Curr. Opin. Genet. Dev.* 2003; 13: 593-598.

Badger JH, Olsen GJ. CRITICA: Coding Region Identification Tool Invoking Comparative Analysis. *Mol. Biol. Evol.* 1999; 16(4): 512-514.

Balakirev ES, Ayala FJ. PSEUDOGENES: Are They "Junk" or Functional DNA? *Annu. Rev. Genet.* 2003; 37:123–151

Baldauf SL. The deep roots of Eukaryotes. *Science* 2003; 300: 1703 – 1706.

Barker WC, George DG, Mewes HW, Pfeiffer F, Tsugita A. The PIR-International databases. *Nucleic Acids Res.* 1993; 21(13):3089 – 3092.

Barrett MP, Burchmore RJS, Stich A, Lazzari JO, Frasch AC, Cazzulo JJ, Krishna S. The trypanosomiases. *The Lancet* 2003; 362: 1469-1480.

Baxevanis AD, Oullette BFF. Bioinformatics: A pratical guide to the analysis of genes and proteins. 2 ed. Washignton (DC) USA: Wiley-Interscience; 2001

Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H, Bartholomeu DC, *et al*. The genome of the African Trypanosome *Trypanosoma brucei*. *Science* 2005; 309: 416 – 422.

Borodovsky M, McIninch J. Recognition of genes in DNA sequence with ambiguities. *Biosystems.* 1993; 30:161-171.

Brazilian National Genome Project Consortium. The complete genome sequence of Chromobacterium violaceum reveals remarkable and exploitable bacterial adaptability. *PNAS* 2003; 100(20): 11660 – 11665.

Brener Z, Andrade ZA, Barral-Netto, M. *Trypanosoma cruzi* e Doença de Chagas. 2 ed. Rio de Janeiro: Guanabara Koogan, 2000.

Briones MRS, Souto RP, Stolf BS, Zingales B. The evolution of two *Trypanosoma cruzi* subgroups inferred from rRNA genes can be correlated with the interchange of American mammalian faunas in the Cenozoic and has implications to pathogenicity and host specificity. *Mol. Biochem. Parasitol.* 1999; 104: 219 – 232.

Burleigh BA, Woolsey AM. Cell signalling and *Trypanosoma cruzi* invasion. *Cell Microbiol*. 2002; 4(11): 701 – 711.

Cai W, Pei J, Grishin NV. Reconstruction of ancestral protein sequences and its applications. *BMC Evol. Biol.* 2004, 4:33.

Campbell DA, Thomas S, Sturm NR. Transcription in kinetoplastid protozoa. Why be normal?. *Microbes Infect.* 2003; 5: 1231–1240

Cano MI, Gruber A, Vazquez M, Cortes A, Levin MJ, Gonzalez A, *et al.* Molecular karyotype of clone CL Brener chosen for the *Trypanosoma cruzi* genome project. *Mol. Biochem. Parasitol.* 1995; 71: 273-278.

Carlton JMR, Muller R, Yowell CA, Fluegge MR, Sturrock KA, Pritt JR, *et al.*. Profiling the malaria genome: a gene survey of three species of malaria parasite with comparison to other apicomplexan species. *Mol. Biochem. Parasitol.* 2001; 118: 201–210.

Chalmel F, Lardenois A, Thompson JD, Muller J, Sahel JA, Léveillard T, Poch O. GOAnno: GO annotation based on multiple alignment. *Bioinformatics* 2005; 21(9): 2095 – 2096.

Chiurillo MA, Crisante G, Rojas A, Peralta A, Dias M, Guevara P, *et al*. Detection of *Trypanosoma cruzi* and *Trypanosoma rangeli* Infection by Duplex PCR Assay Based on Telomeric Sequences. *Clin. Diagn. Lab. Immunol.* 2003; 10(5): 775 – 779.

Cliften PF, Hillier LW, Fulton L, Graves T, Miner T, Gish WG, *et al*. Surveying Saccharomyces Genomes to Identify Functional Elements by Comparative DNA Sequence Analysis. *Genome Res.* 2001; 11: 1175 – 1186.

Coural JR, Fernandes O, Arboledaz M, Barrett TV, Carraral N, Degrave W, Campbell DA. Human infection by *Trypanosoma range* in the Brazilian Amazon. *Trans. R. Soc. Trop. Med. Hyg.* 1996; 90: 278-279

Crooks GE, Hon G, Chandonia J, Brenner SE. WebLogo: A Sequence Logo Generator. *Genome Res.* 2004; 14: 1180 – 1190.

D' Alessandro, A. Biology of *Trypanosoma (Herpetosoma) rangeli* Tejera, 1920. In: Lumsden, WHR, Evans, DA. Biology of the kinetoplastida. Londres: London Academic, 327-403; 1976. 3v.

D' Alessandro, A, Saraiva, NG. *Trypanosoma rangeli*. In: Protozoal Diseases, HM Gilles editor. Arnold.1999.

Dávila AMR. Bioinformatics in developing countries: how many tools are need?. *Inabis* 2002.

Davila AM, Majiwa PA, Grisard EC, Aksoy S, Melville SE. Comparative genomics to uncover the secrets of tsetse and livestock-infective trypanosomes. *Trends Parasitol*. 2003a; 19(10):436-439.

Dávila AMR, Guerreiro LTA, Souza SS, Hall N. Exploring the genome of *Trypanosoma vivax*: towards a comparative genomics approach. In: XXX Annual Meeting on Basic Research in Chagas Disease - XIX Meeting of the Brazilian Society of Protozoology, 2003, Caxambú, MG, *Brasil. Rer. Ins. Med. Trop.* São Paulo 2003b; 45: 01 – 228.

Dávila AM, Lorenzini DM, Mendes PN, Satake TS, Sousa GR, Campos LM, *et al.* GARSA: genomic analysis resources for sequence annotation. *Bioinformatics* 2005; 21(23):4302 – 4303.

De Souza W. Basic cell biology of *Trypanosoma cruzi. Curr. Pharm. Des.* 2002; 8: 269 – 285.

Degrave W, Melville S, Ivens A, Aslett. Parasite genome initiatives. *Int. J. Parasitol.* 2001; 31: 532 – 536.

Delcher Al, Harmon D, Kasif, White O, Salzberg. Improved microbial gene identification with Glimmer. *Nucleic Acids Res.* 1999; 27(23): 4636 – 4641.

Dessen P, Zagulski M, Gromadka R, Plattner H, Kissmehl R, Meyer E, *et al*.. Paramecium genome survey: a pilot project. *Trends Genet.* 2001; 17: 306-308.

Dias-Neto E, Correa RG, Verjovski-Almeida S, Briones MRS, Nagaid MA, Silva W, *et al*. Shotgun sequencing of the human transcriptome with ORF expressed sequence tags. *PNAS* 1999; 97(7): 3491 – 3496

Do CB, Mahabhashyam MSP, Brudno M, Batzoglou S. ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Res.* 2005; 15: 330-340.

Eddy SR. Profile Hidden Markov models. *Bioinformatics* 1998; 14(9): 755 – 163.

Eddy SR. What is a hidden markov model? *Nat. Biotechnol.* 2004. 22(10): 1315 – 1316.

Eddy S. HMMER - profile hidden Markov models for biological sequence analysis Version 2.3.2. http://hmmer.wustl.edu/. Howard Hughes Medical Institute and Dept. of Genetics Washington University School of Medicine, St. Louis, USA, 2003.

Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 2004, 5:113.

Eger-Mangrich I, de Oliveira MA, Grisard EC, De Souza W, Steindel M. Interaction of *Trypanosoma rangeli* Tejera, 1920 with different cell lines in vitro. *Parasitol Res.* 2001; 87(6):505 – 509.

El-Sayed NM, Donelson JE. A survey of the *Trypanosoma brucei rhodesiense* genome using shotgun sequencing. *Mol. Biochem. Parasitol.* 1997; 84: 167-178.

El-Sayed NM, Hegdea P, Quackenbusha J, Melvilleb SE, Donelsonc JE. The African trypanosome genome. *Int. J. Parasitol.* 2000; 30: 329-345.

El-Sayed NM, Myler PJ, Bartholomeu DC, Nilsson D, Aggarwal G, Tran A, *et al.* The Genome Sequence of *Trypanosoma cruzi*, Etiologic Agent of Chagas Disease. *Science* 2005a; 309: 409 – 415.

El-Sayed NM, Myler PJ, Blandin G, Berriman M, Crabtree J, Aggarwal G, *et al.* Comparative Genomics of Trypanosomatid Parasitic Protozoa. *Science* 2005b; 309: 404 – 409.

Erpel T, Hwang P, Craik CS, Terick RJF, McGrath ME. Physical Map Location of the New *Escherichia coli* Gene eco, Encoding the Serine Protease Inhibitor Ecotin. *J. Bacteriol.* 1992; 174(5): 1704.

Ewing B, Hillier L, Wendl MC, Green P. Base-Calling of Automated Sequencer Traces Using Phred. I. Accuracy Assessment. *Genome Res.* 1998a; 8:175–185.

Ewing B, Green P. Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities. *Genome Res.* 1998b; 8:186–194.

Felsenstein J. PHYLIP (Phylogenetic Inference Package), version 3.65. http://evolution.genetics.washington.edu/phylip.html. Departament of Genetics, University of Washington, Seatle, USA. 2005.

Ferrari I, Lorenzi H, Santos MR, Brandariz S, Requena JM, Schijman A, *et al*.. Towards the physical map of the *Trypanosoma cruzi* nuclear genome: construction of YAC and BAC libraries of the reference clone T. cruzi CL-Brener. *Mem. Inst. Oswaldo Cruz* 1997; 92: 843-852.

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, *et al*.. Whole-genome random sequencing and assembly of *Haemophilus influenzae*. *Science* 1995; 269(5223):496-512.

Frash AC. Functional diversity in the trans-sialidase and mucin families in *Trypanosoma cruzi*. *Parasitol. Today* 2000; 16(7): 282 – 286;

Fujita A, Massirer KB, Durham AM, Ferreira CE, Sogayar MC. The GATO gene annotation tool for research laboratories. *Braz. J. Med. Biol. Res.* 2005; 38(11): 1571 – 1574.

Galperin MY, Koonin EV. 'Conserved hypothetical' proteins: prioritization of targets for experimental study. *Nucleic Acids Res.* 2004; 32(18): 5452 – 5463.

Gene Ontology Consortium. The Gene Ontology (GO) project in 2006. *Nucleic Acids Res.* 2006; 34: 322 – 326.

Ghedin E, Wang S, Foster JM, Slatko BE. First sequenced genome of a parasitic nematode. *Trends Parasitol.* 2004a; 20(4): 151 – 153.

Ghedin E, Bringaud F, Peterson J, Myler P, Berriman M, Ivens A, *et al*.. Gene synteny and evolution of genome architecture in trypanosomatids. *Mol. Biochem. Parasitol.* 2004b; 134(2): 183 – 191.

Gregory SG, Sekhon M, Schein J, Zhao S, Osoegawak K, Scott CA, *et al*.. A physical map of the mouse genome. *Nature* 2002; 418: 743 – 752.

Grisard EC, Campbell DA, Romanha AJ. Mini-exon gene sequence polymorphism among *Trypanosoma rangeli* strains isolated from distinct geographical regions. *Parasitology* 1999; 118: 375-382.

Grisard EC, Steindel M. *Trypanosoma (Herpetosoma) rangeli*. In: Neves DP. Parasitologia Humana, 11 ed. Porto Alegre: Atheneu, 2004.

Guerreiro LT, Souza SS, Wagner G, Souza EA, Mendes PN, Campos LM, *et al*. Exploring the Genome of *Trypanosoma vivax* through GSS and In Silico Comparative Analysis. *OMICS* 2005; 9(1): 116 – 128.

Guhl F, Jaramillo C, Carranza JC, Vallejo GA. Molecular Characterization and Diagnosis of *Trypanosoma cruzi* and *T. rangeli. Arch. Med. Res.* 2002; 33: 362–370.

Guhl F, Vallejo GA. *Trypanosoma (Herpetosoma) rangeli* Tejera, 1920 – An updated review. *Men. Inst. Oswaldo Cruz.* 2003; 98(4): 435 – 442.

Guo F, Ou H, Zhang C. ZCURVE: a new system for recognizing proteincoding genes in bacterial and archaeal genomes. *Nucleic Acids Res.* 2003; 31(6): 1780 – 1789.

Hammer J, Schneider M. Genomics Algebra: A new, integrating data model, language, and tool for processing and querying genomic information. *Proccedings of the CIDR Conference.* 2003

Hanke J, Sanchez DO, Henriksson J, Aslund L, Pettersson U, Frasch AC, Hoheisel JD. Mapping the *Trypanosoma cruzi* genome: analyses of representative cosmid libraries. *Biotechniques* 1996; 21:686-690.

Hannaert V, Saavedra E, Duffieux F, Szikora J, Rigden DJ, Michels PAM, Opperdoes FR. Plant-like traits associated with metabolism of Trypanosoma parasites. *PNAS* 2003; 100(3): 1067 – 1071.

Hardison CR. Comparative Genomics. *PLoS Biol.* 2003; 1(2): 156 – 160.

Henriksson J, Solari A, Rydaker M, Sousa OE, Pettersson U. Karyotype variability in *Trypanosoma rangeli. Parasitology* 1996; 112: 385 – 391.

Hoare CA. The classification of mammalian trypanosomes. Oxford: Blackwell Scientific Publications. 1972.

Hotz-Wagenblatt A, Hankeln T, Ernst P, Glatting K, Schmidt ER, Suhai S. ESTAnnotator: a tool for high throughput EST annotation. *Nucleic Acids Res.* 2003; 31: 3761 – 3719.

Huang X, Madan A. CAP3: A DNA Sequence Assembly Program. *Genome Res.* 1999; 9:868–877.

Hughes AL, Piontkivska H. Molecular phylogenetics of Trypanosomatidae: contrasting results from 18S rRNA and protein phylogenies. *Kinetoplastid Biol. Dis.* 2003; 2: 15 – 25.

Hughey R, Karplus K, Krogh A. SAM Sequence Alignment and Modeling Software System. Version 3.0. http://www.cse.ucsc.edu/research/compbio/sam.html. Computer Engineering, University of California, Santa Cruz, USA. 1996

Ivens AC, Peacock CS, Worthey EA, Murphy L, Aggarwal G, Berriman M, *et al*.. The Genome of the Kinetoplastid Parasite, *Leishmania major. Science* 2005; 309: 436 – 442.

Janssen CS, Barrett MP, Lawson D, Quail MA, Harris D, Bowman S, *et al*.. Gene discovery in *Plasmodium chabaudi* by genome survey sequencing. *Mol. Biochem. Parasitol.* 2001; 113: 251–260.

Jones CE, Baumann U, Brown AL. Automated methods of predicting the function of biological sequences using GO and BLAST. *BMC Bioinformatics* 2005, 6:272

Karplus K, Barrett C, Hughey R. Hidden markov models for detecting remote proteins homologies. *Bioinformatics* 1998; 14(10): 846 – 856.

Khan S, Situ G, Decker K, Schmidt CJ. GoFigure: Automated Gene Ontology™ Annotation. *Bioinformatics* 2003; 19(18): 2484 – 2485.

Koonin EV. Orthologs, Paralogs and Evolutionary Genomics. *Annu. Rev. Genet.* 2005; 39:309–338.

Kumar S, Tamura K, Nei M. MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief Bioinform.* 2004; 5(2):150 – 163.

Kunisawa T, Nakamura M, Watanabe H, Otsuka J, Tsugita A, *et al.*. *Escherichia coli* K12 genomic database. *Protein Seq. Data Anal.* 1990; 3(2):157-162.

Laurentino EC, Ruiz JC, Fazelinia G, Myler PJ, Degrave W, Alves-Ferreira M, *et al*.. A survey of *Leishmania braziliensis* genome by shotgun sequencing. *Mol. Biochem. Parasitol.* 2004; 137: 81 – 86.

Leech V, Quail MA, Melville SE. Separation, difestion, and cloning of intact parasites chromosomes embedded in agarose. *Methods Mol. Biol.* 2004; 270: 335 – 352.

Lewis SE, Searle SMJ, Harris N, Gibson M, Iyer V, Richter J, *et al.*. Apollo: a sequence annotation editor. *Genome Biology* 2002, 3(12): 0082.1–0082.14

Li L, Stoeckert CJ, Roos DS. Genomes OrthoMCL: Identification of Ortholog Groups for Eukaryotic. *Genome Res.* 2003; 13: 2178 – 2189.

Lopez R, Silventoinen V, Robinson S, Kibria A, Gish W. WU-Blast2 server at the European Bioinformatics Institute. *Nucleic Acids Res.* 2003; 31(13): 3795 – 3798.

Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* 1997; 35(5): 955 – 964.

Machado PE, Eger-Mangrich I, Rosa G, Koerich LB, Grisard EC, Steindel M. Differential susceptibility of triatomines of the genus Rhodnius to *Trypanosoma rangeli* strains from different geographical origins. *Int. J. Parasitol.* 2002; 31: 632 – 634.

Meirelles RMS, Henrique-Pons A, Soares MJ, Steindel M. Penetration of the salivary glands of *Rhodnius domesticus* Neiva & Pinto, 1923 (Hemiptera: Reduviidae) by *Trypanosoma rangeli* Tejera, 1920 (Protozoa: Kinetoplastida). *Parasitol Res*. 2005; 97(4): 259 – 269.

Merino EF, Fernandez-Becerra C, Madeira AMBM, Machado AL, Durham A, Gruber A, *et al.*. Pilot survey of expressed sequence tags (ESTs) from the asexual blood stages of *Plasmodium vivax* in human patients. *Malar. J.* 2003; 2:21

Meyer F, Goesmann A, McHardy AC, Bartels D, Bekel T, Clausen J, *et al.*. GenDB--an open source genome annotation system for prokaryote genomes. *Nucleic Acids Res.* 2003; 31(8):2187 – 2195.

Michels PAM., Hannaert V, Bringaud F. Metabolci aspects of glycosomes in trypanosomatidae – new data and views. *Parasitol. Today* 2000; 16(11): 482 – 489.

Moreira D, López-Garcia P, Vickerman K. An updated view of kinetoplastid phylogeny using environmental sequences and a closer outgroup: proposal for a new classification of the class Kinetoplastea. *Int. J. Syst. Evol. Microbiol.* 2004; 54, 1861–1875.

Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, *et al*.. InterPro, progress and status in 2005. *Nucleic Acids Res.* 2005; 33: D201–D205.

Murthy VK, Dibbern KM, Campbell DA. PCR amplification of mini-exon genes differentiates *Trypanosoma cruzi* from *Trypanosomas rangeli*. *Mol. Cell. Probes* 1992; 6: 237 – 243.

Nilsen TW. trans-Splicing An update. *Mol. Biochem. Parasitol.* 1995; 73: 1-6.

Notredame C, Higgins DG, Heringa J. T-Coffee: A Novel Method for Fast and Accurate Multiple Sequence Alignment. *J. Mol. Biol.* 2000; 302: 205 – 217.

Osorio Y, Travi BL, Palma GI, Saravia NG. Infectivity of *Trypanosoma rangeli* in a promonocytic mammalian cell line. *J. Parasitology* 1995: 81(5): 687-693

Pál G, Sprengelb G, Patthy A, Gráf L. Alteration of the specificity of ecotin, an *E. coli* serine proteinase inhibitor, by site directed mutagenesis. *FEBS* 1994; 342: 57-60.

Pappas GJ, Benabdellah K, Zingales B, Gonzalez A. Expressed sequence tags from the plant trypanosomatid *Phytomonas serpens*. *Mol. Biochem. Parasitol.* 2005; 142: 149 -157.

Parkinson J, Guiliano DB, Blaxter M. Making sense of EST sequences by CLOBBing them. *BMC Bioinformatics* 2002; 25(3): 1-8.

Passagia MP, Zaha A. Técnicas de biologia molecular. In: Zaha A 2003 Biologia Molecular Básica. 3 ed. Porto Alegre: Mercado Aberto, 2003.

Pearson WR, Lipman SJ. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci.* 1988; 85: 2444 – 2448.

Peterson SN, Schramm N, Hu PC, Bott KF, Hutchison CA. A random sequence approach for placing markers on the physical map of *Mycoplasma genitalium. Nuclei Acids Res.* 1991; 19: 6027-6031.

Porcel BM, Tran A, Tammi M, Nyarady Z, Rydåker M, Urmenyi TP, *et al.*. *Trypanosoma cruzi* Gene Survey of the Pathogenic Protozoan. *Genome Res.* 2000; 10: 1103-1107

Puiu D, Enomoto S, Buck GA, Abrahamsen MS, Kissinger JC. CryptoDB: the Cryptosporidium genome resource. *Nucleic Acids Res.* 2004; 32: D329 – D331.

Rice P, Longden I, Bleasby A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* 2000; 16: 276 – 277.

Roberts RJ. Identifying Protein Function-A Call for Community Action. *PLoS Biol.* 2004: 2(3): 293 – 294.

Rodrigues JB. Geração e análise de etiquetas de seqüências transcritas de *Trypanosoma rangeli* utilizando a técnica de ORESTES. Florianópolis, 2005. Tese de Mestrado.

Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B. Artemis: sequence visualization and annotation. *Bioinformatics* 2000; 16:944-945

Saas J, Ziegelbauer K, von Haeseler A, Fast B, Boshart M. A Developmentally Regulated Aconitase Related to Iron-regulatory Protein-1 Is Localized in the Cytoplasm and in the Mitochondrion of *Trypanosoma brucei. J. Biol. Chem.* 2000; 275(4): 2745 – 2755.

Sambrook J, Russel, DW. Molecular Cloning: A Laboratory Manual. 3 ed. New York: Cold Spring Harbor Laboratory Press. 2001, 2v.

Santos IFK, Pereira MEA. Lectins discriminate between pathogenic and nonpathogenic south american trypanosomes. *Am. J. Trop. Med. Hyg.* 1984; 35(5): 839 – 844.

Schaeffer SW, Bernhardt MJ, Anderson WW. Evolutionary rearrangement of the amylase genomic regions between Drosophila melanogaster and Drosophila pseudoobscura. *J. Hered.* 2003; 94: 464-471.

Silva AM. Riscos de transmissão da doença de Chagas por transfusão sanguínea no Estado de Santa Catarina, Brasil. São Paulo; 2002. Tese de Doutorado.

Simpson AJ, Reinach FC, Arruda P, Abreu FA, Acencio M, Alvarenga R, *et al*.. The genome sequence of the plant pathogen Xylella fastidiosa. The *Xylella fastidiosa* Consortium of the Organization for Nucleotide Sequencing and Analysis. *Nature* 2000a; 406: 151 – 157.

Simpson L, Thiemann OH, Savill NJ, Alfonzo JF, Maslov DA. Evolution of RNA editing in trypanosome mitochondria. *PNAS* 2000b; 97(13): 6986 – 6993.

Smith MW, Aley SB, Sogin M, Gillin FD, Evans GA. Sequence survey of the *Giardia lamblia* genome. *Mol. Biochem. Parasitol.* 1998; 95: 267 – 280.

Snoeijer CQ, Picchi GF, Dambrós BP, Steindel M, Goldenberg S, Fragoso SP, *et al*.. *Trypanosoma rangeli* Transcriptome Project: Generation and analysis of expressed sequence tags. *Kinetoplastid Biol. Dis.* 2004; 3: 1 – 4.

Souza SJ, Camargo AA, Briones MRS, Costa FF, Nagai MA, Verjovski-Almeida S, *et al*.. Identification of human chromosome 22 transcribed sequences with ORF expressed sequence tags. *PNAS* 2002; 97(23): 12690 – 12693.

Stanke M, Schoeffmann O, Morgenstern B, Waack S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* 2006; 7:62.

Stein L. Genome annotation: from sequence to biology. *Nature Reviews* 2001; 2: 493 – 503.

Steindel M, Carvalho Pinto JC, Toma HK, Mangia RHR, Ribeiro-Rodrigues R, Romanha AJ. *Trypanosoma rangeli* (Tejera, 1920) isolated from a sylvatic rodent (Echimys dasythrix) in Santa Catarina Island, Santa Catarina State: First report of this Trypanosome in Southern Brazil. *Mem. Inst. Oswaldo Cruz* 1991; 86: 73-79.

Steindel M, Dias Neto E, Pinto CJ, Grisard EC, Menezes CL, Murta SM, et al.. Randomly amplified polymorphic DNA (RAPD) and isoenzyme analysis of *Trypanosoma rangeli* strains. *J Eukaryot Microbiol.* 1994; 41(3):261 – 267.

Stevens JR, Gibson W. The molecular evolution of trypanosomes. *Parasitol. Today* 1999; 15(11): 432 – 437.

Stoco PH, Rodrigues JB, Rotava G, Wagner G, Snoeijer CQ, Pacheco LK, *et al.*. Anotação parcial do transcriptoma de *Trypansoma rangeli* utilizando a plataforma GARSA (Genome Analysis Resources of sequence annotation). Resumo apresentado no XIX Congresso Brasileirdo de Parasitologia. 2005 nov. Porto Alegre.

Stuart KD, Schnaufer A, Ernst NL, Panigrahi AK. Complex management: RNA editing in trypanosomes. *TRENDS in Bioche. Sciences* 2005; 30(2): 97 – 105.

Sturm NR, Vargas NS, Westenberger SJ, Zingales B, Campbell DA. Evidence for multiple hybrid groups in *Trypanosoma cruzi*. *Int. J. Parasitol*. 2003; 33: 269 – 279.

Suzek BE, Ermolaeva MD, Schreiber M, Salzberg SL. A probabilistic method for identifying start codons in bacterila genomes. *Bioinformatics* 2001; 17(12): 11230 – 1130.

Tanaka T. Differential diagnosis of *Trypanosoma cruzi* and *T. rangeli* infection by PCR of Cysteine Proteinase Genes. *Kansenshogaku Zasshi* 1997; 27(9): 903 – 909.

Tatusov RL, Galperin MY, Natale DA, Koonin EV. The COG database: a tool for genome-scale analysis of protein functions na evolution. *Nucleic Acids Res.* 2000; 28(1): 33 – 36.

Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, *et al*.. The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* 2001; 29(1): 22 – 28.

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 2003; 4:41.

Tech M, Merkl R. YACOP: Enhanced gene prediction obtained by a combination of existing methods. *In Sil. Biol.* 2003; 3:441-51.

Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 1994; 22(22): 4673 – 4680.

Toaldo GB, Steindel M, Sousa MA, Tavares CC. Molecular Karyotype and Chromosomal Localization of Genes Encoding b-tubulin, Cysteine Proteinase, hsp 70 and Actin in *Trypanosoma rangeli. Mem. Inst. Oswaldo Cruz* 2001; 96(1): 113 – 121.

Vallejo GA, Macedo AM, Chiari E, Pena SDJ. Kinetoplast DNA from *Trypanosoma rangeli* contains two distinct classes of minicircle with different size and molecular organization. *Mol. Biochem. Parasitol.* 1994; 67: 245–253.

Vallejo GA, Guhl F, Carranza JC, Lozano LE, Sanchez JL, Jaramillo JC, *et al.*. kDNA markers define two major *Trypanosoma rangeli* lineages in Latino-America. *Acta Tropica* 2002; 81: 77 – 82.

Vargas N, SoutoRP, Carranza JC, Vallejo GA, Zingales B. Amplification of a Specific Repetitive DNA Sequence for *Trypanosoma rangeli* identification and Its Potential Application in Epidemiological Investigations. *Exp. Parasitol.* 2000; 96: 147 – 159.

Vasconcelos ATR, Ferreira HB, Bizarro CV, Bonatto SL, Carvalho MO, Pinto PM, *et al.*. Swine and Poultry Pathogens: the Complete Genome Sequences of Two Strains of *Mycoplasma hyopneumoniae* and a Strain of *Mycoplasma synoviae*. *J. Bacteriol.* 2005; 187(6): 5568 – 5577.

Venter JC, Remington K, Heidelberg JF, Halpern AL , Rusch D, Eisen JA, *et al*.. Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science* 2004; 304: 66 – 74.

Vickerman K. The diversity of the kinetoplastid flagellates. In: Lumsden WH, Evans DA. Biology of the Kinetoplastida. Vol 1. London: Academic Press Inc. 1976

Volker C, Brown JR. Bioinformatics and the discovery of novel anti-microbial targets. *Curr. Drug. Targets Disord.* 2002; 2(4): 279 – 290;

Wilkinson SR, Obado SO, Mauricio IL, Kelly JM. *Trypanosoma cruzi* expresses a plant-like ascorbatedependent hemoperoxidase localized to the endoplasmic reticulum. *PNAS* 2002; 99(21): 13453 – 13458.

8 - ANEXOS

Tabela A1: Tabela contendo o nome e o número de campos das tabelas que constituem o banco de dados dos projetos GARSA, bem como a descrição da função de cada.

Tabela	***N°. de Campos***	***Função***
Anotation	4	Armazena as notas de cada usuário para cada seqüência não redundante.
Blast_Hit	22	Armazena todos os resultados de similaridade encontrados pelos algoritmos BLAST e RPSBlast.
Blast_Search	4	Relaciona o programa e a base de dados utilizado nas análises de similaridade
CDS	23	Armazena todas as anotações realizadas pelos usuários para uma determinada região da seqüência não redundante, bem como as seqüências nucleotídicas e protéicas desta região. Uma seqüência pode conter mais de uma anotação.
Cluster_Fasta	4	Armazena a seqüência final de cada seqüência não redundante realizada pelo programa CAP3
Clustering	5	Contem os dados de cada clusterização realizadas pelo CAP3
Clusters	7	Contem os dados de cada seqüência não redundante. Relacionada com a tabela Cluster_Fasta
Codon_Usage	6	Armazena os resultados do programa *cusp*
Contaminant	3	Contem a seqüência fasta do contaminante selecionado pelo administrador
GO_Hit	9	Armazena todos os códigos de acesso dos termos encontrados pela anotação funcional da *Gene Ontology* automática.
GO_Search	5	Relaciona as anotações automáticas da GO com os bancos utilizados para esta busca
HMM_Hit	13	Armazena os resultados de busca de homologia realizada pelo HMMER
HMM_Search	5	Relaciona o modelo utilizado para cada busca de homologia realizado pelo HMMER.
Interpro_Hit	11	Armazena os resultados das análises realizadas pelo programa InterProScan
Library	5	Armazena as informações das bibliotecas criadas pelo administrador, como o nome de vetor, iniciador usado.
ORF_Predict	11	Armazena os resultados dos programas de predição de genes.
ORF_Search	4	Contém as informações necessárias para diferenciar os programas utilizados para gene encontrado e armazenado na tabela ORF_Search.
Primer	3	Armazena as seqüências dos iniciador, caso haja a necessidade.
Reads	23	Armazena todas seqüências (*reads*) inseridos no sistema, bem como as análises de qualidade, tamanho e as seqüências já limpas destes *reads*.
Reads_Cluster	2	Relaciona os *reads* que compõe um cluster.
tRNA	16	Armazena os resultados do programa tRNA-Scan
Vector	3	Armazena a seqüência de vetor selecionada pelo administrador do projeto na hora da criação da biblioteca

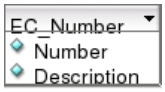


Figura A2: Figura mostrando à única tabela do banco de dados EC_database. Este banco é chamando pelo sistema GARSA apenas quando o usuário fará à anotação final de uma região codificante, quando se trata de uma enzima. Este banco contem os números de EC e as respectivas descrições das subclasses de enzimas.

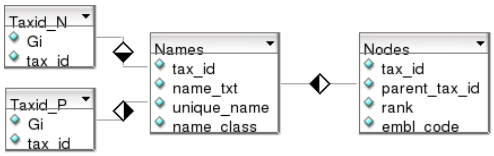


Figura A3: Nesta figura está representado o banco de dados Taxonomy. Este banco foi desenvolvido para armazenar a relação entre cada número de acesso do GenBank (GI) de cada seqüência e o respectivo código taxonômico, nas tabelas Taxid_N (nucleotídeos) e Taxid_P (proteínas). Também são armazenados os nomes dos grupos taxonômicos de cada código na tabela Names e a relação entre os códigos (Nodes). Este banco será consultado para carregar o campo Taxid na tabela Blast_Hit nos bancos dos projetos e ao utilizar a interface "*Taxonomy"* no sistema GARSA.

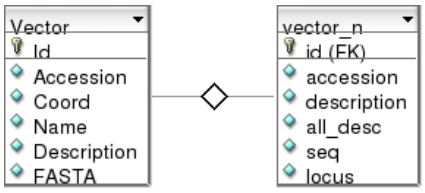


Figura A4: Esquema do banco UniVec, que armazena todas às seqüências de vetores de clonagem na tabela Vector. Esta tabela é consultada pelo sistema no momento em que é definido o vetor para ser utilizado pelo programa *Crossmatch* e durante a etapa de retirada das regiões de vetores dos *reads* realizada pelo mesmo programa. O administrador de cada projeto pode inserir novas seqüências neste banco por meio de uma interface gráfica ("*Insert New Vector Sequence*").

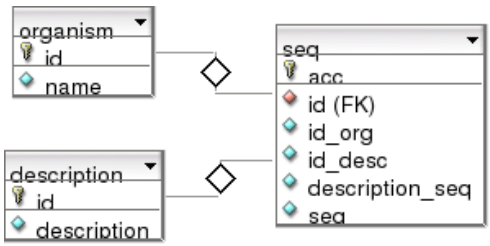


Figura A5: Figura mostrando o esquema do banco contaminant. Este banco armazena na tabela "seq" algumas seqüências ribossomais e mitocôndriais de organismos modelos que são utilizadas pelo sistema para excluir seqüências conhecidas. Esta etapa somente será realizada se o administrador do projeto selecionar uma destas seqüências.

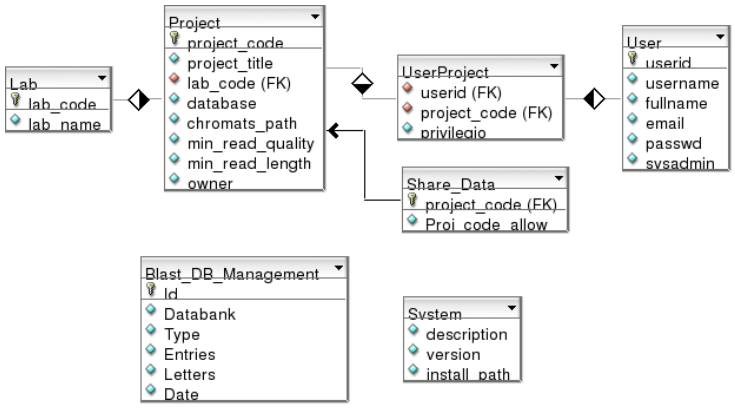


Figura A6: Esquema do banco de dados administrativo seqonsql. Como o GARSA é um programa mutli-projeto e multi-usuário, fez-se necessário à montagem de um banco de dados que armazenam todas as informações dos projetos (Tabela Project), usuários do sistema (Tabela User), privilégios de cada usuário e os projetos que estes possuem acesso (Tabela UserProject). A tabela Share_Data é responsável por armazenar os relacionamentos entre os projetos. A tabela Blast_DB_Management contém informações sobre cada base de dados utilizada pelo BLAST e RPSBlast.

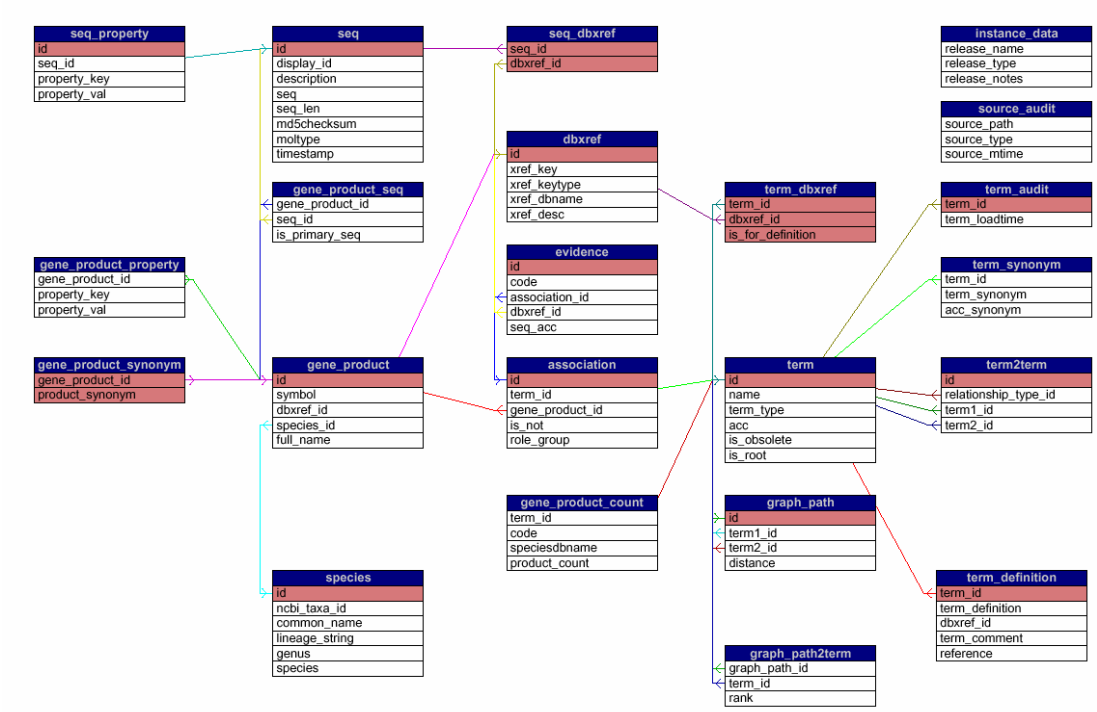
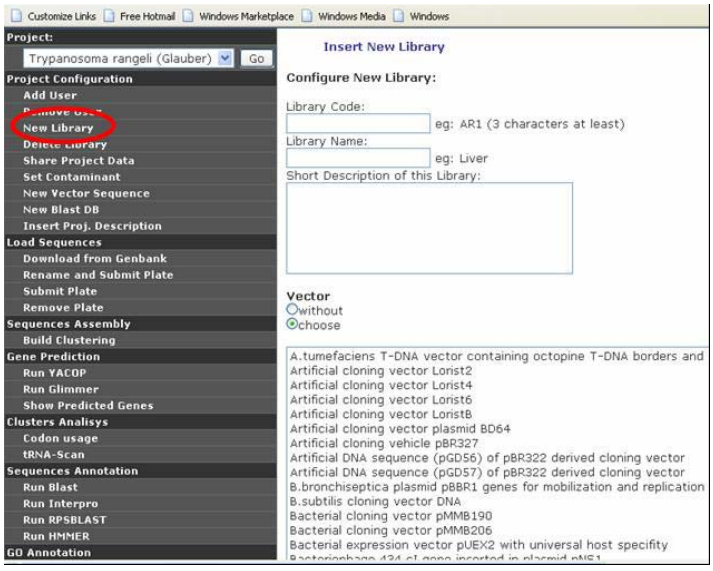


Figura A7: Este esquema mostra a estrutura e organização do banco MySQL do Consórcio Gene Ontology (Fonte: http://www.geneontology.org/dev/doc/diagrams.html, acessado 06/02/06). Este banco é consultado pelo sistema GARSA durante a anotação funcional para a busca das ontologias e termos encontrados pela anotação semi-automática, sendo os códigos de acesso dos termos encontrados por este método armazenados na tabela GO_Hit do banco de cada projeto.

(a)



(b)

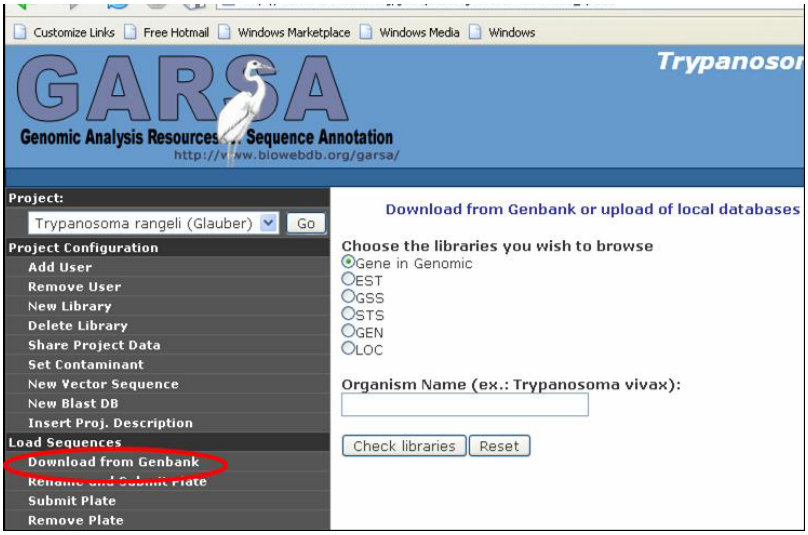


Figura A8: Interfaces CGI do sistema GARSA. (a) "*Insert New Library*", para a configuração das bibliotecas; (b) "*Download from GenBank*", para carregar seqüências diretamente do GenBank ou seqüências em formato fasta armazenadas em arquivos locais.

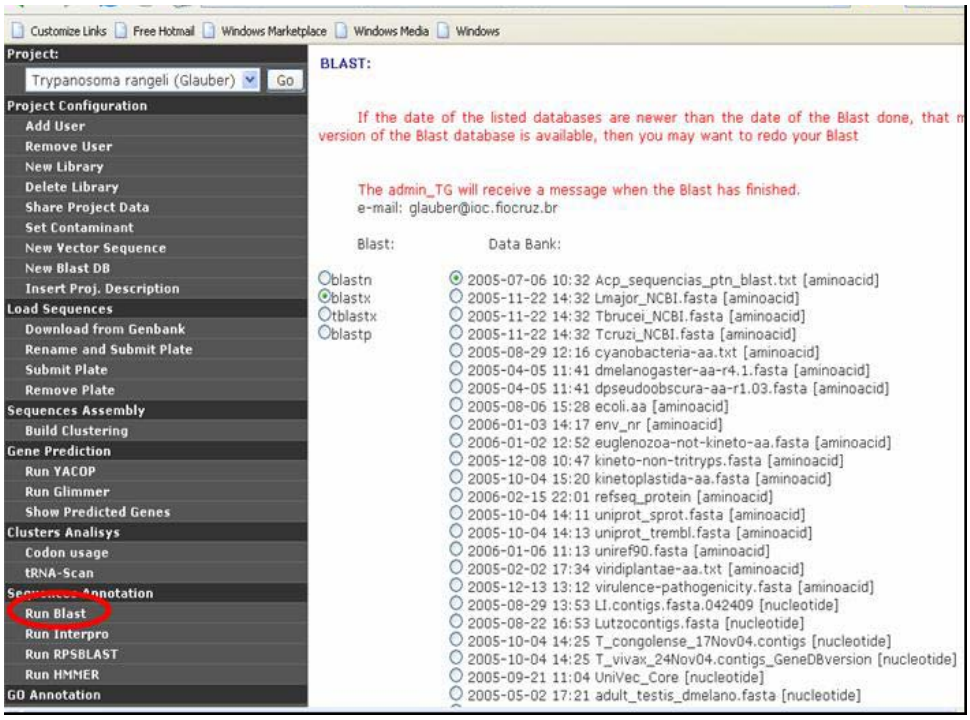


Figura A9: Parte da interface gráfica para a execução dos programa do pacote BLAST. Este formato se repete para os demais programas do *pipeline* GARSA.

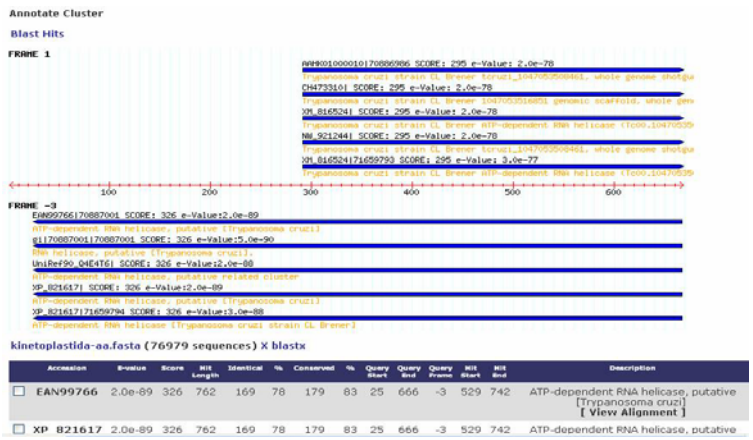


Figura A10: Parte da interface onde são apresentados os resultados dos programas executados pelo sistema GARSA. Nesta figura estão representados as análises realizadas pelo programa BLAST, tanta em forma de gráficos quanto em tabelas.

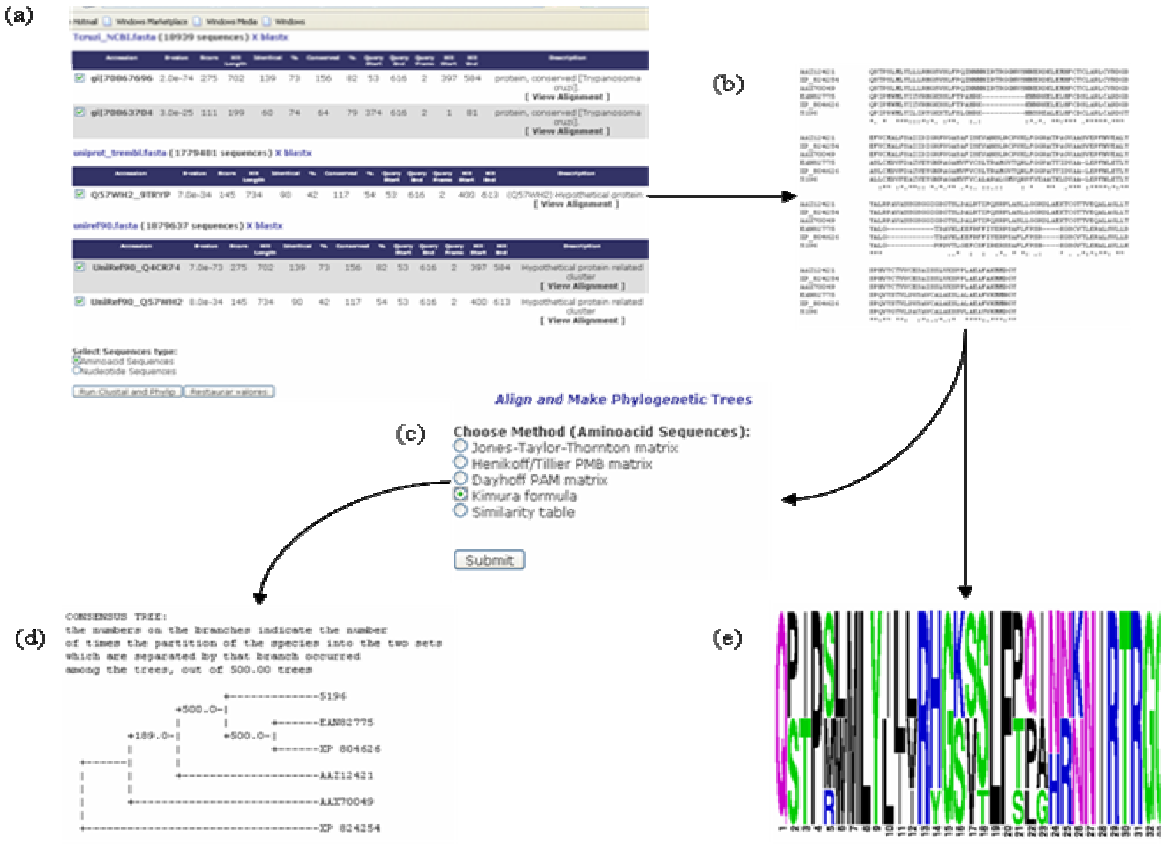


Figura A11: Figura mostrando o *pipeline* das análises filogenéticas através do sistema GARSA. (a) mostra os resultados das busca de similaridade através do programa BLAST para uma

determinada seqüência não redundante, onde existe um campo para a seleção da seqüência similar, e logo abaixo das tabelas são apresentados as opções de seqüências e os botões de execução da análise filogenética; (b) mostra um alinhamento múltiplo (*ClustalW*) gerado a partir das seqüências selecionadas e a seqüência não redundante em questão; (c) parte da interface gráfica onde são selecionados o modelo filogenético para executar o método de distancia (*Phylip*); (d) a árvore consenso gerada pelo programa *Phylip;* (e) o gráfico do alinhamento múltiplo gerado pelo programa *WebLogo*.

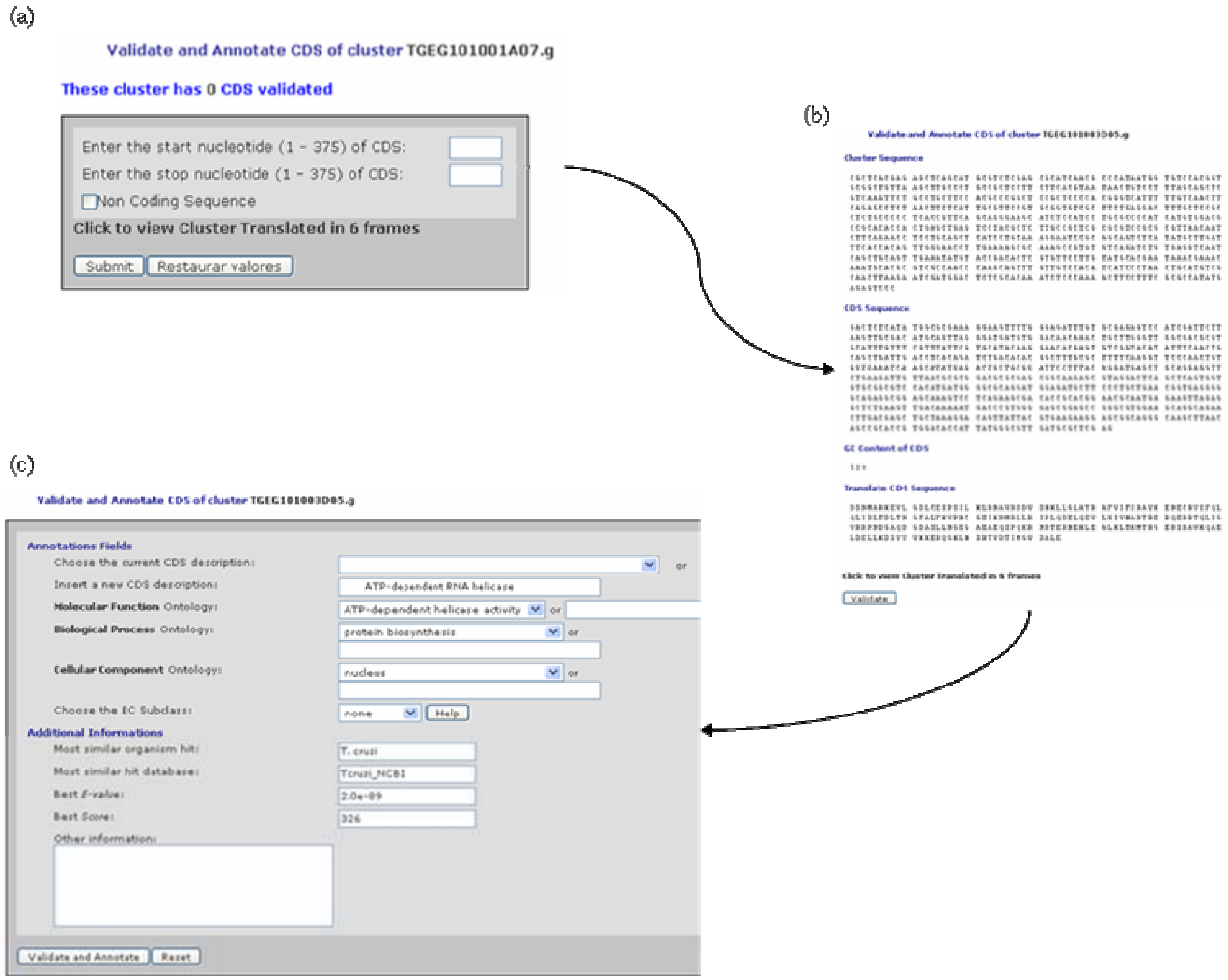


Figura A12: Figura mostrando o *pipeline* para as anotações das regiões de cada seqüência não redundante. Em (a) o campos para o usuário determinar a região codificante; em (b) o GARSA mostra as seqüências nucleotídicas e protéicas da região determinada pelo usuário e em (c) os campos para a anotação manual da seqüência, contendo os campos para a anotação da GO, onde são fornecido as anotações automáticas para cada clasificação funcional.

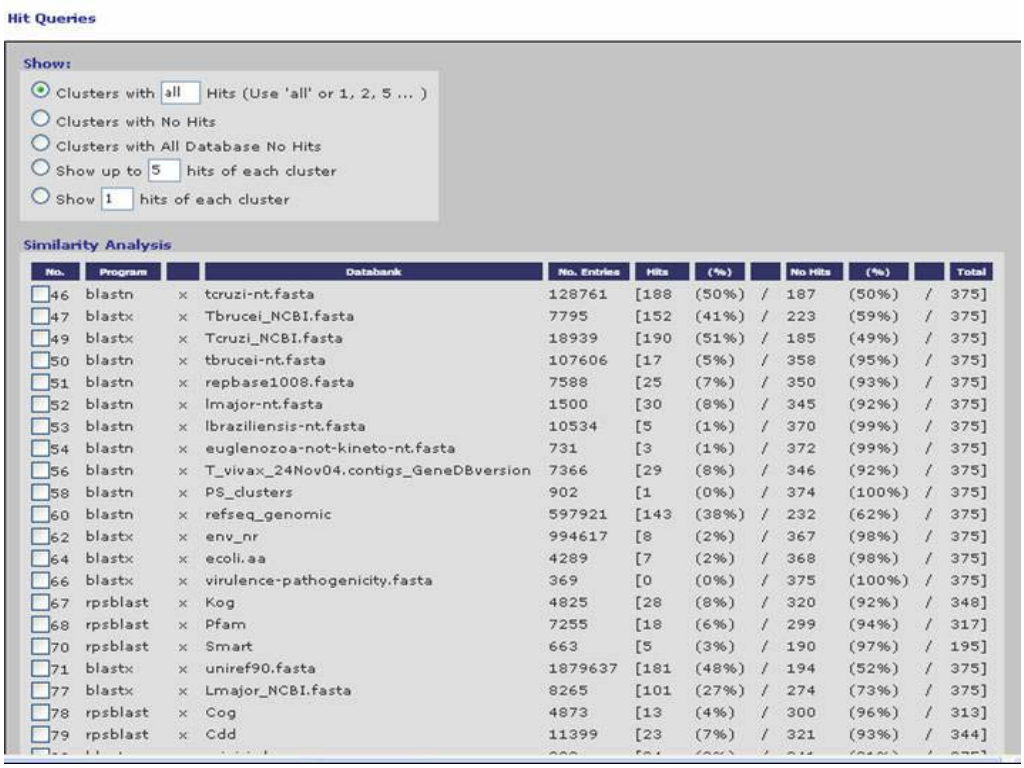


Figura A13: Interface gráfica para a consultada dos resultados encontrados pelo programa BLAST, relacionando o algoritmo e a base de dados utilizado em cada busca.

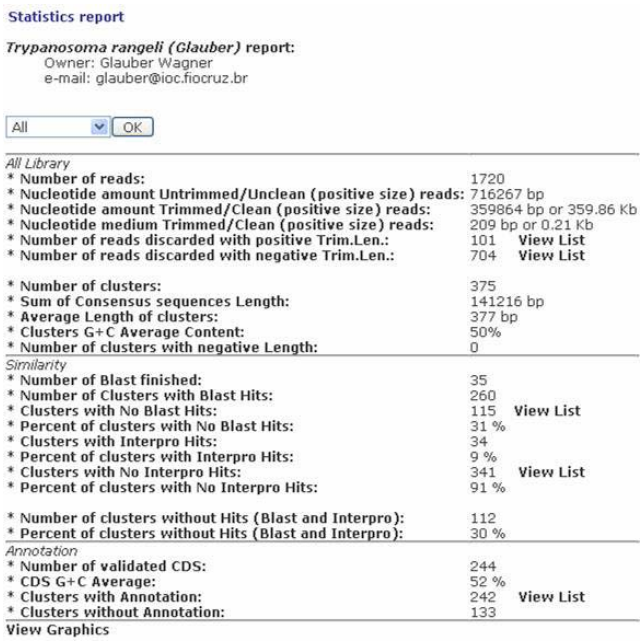


Figura A14: Esta é a interface onde estão apresentados os dados estatísticos de cada projeto.

Tabela A2: Tabela mostrando as seqüências anotadas de cada cluster, conteúdo G+C, tamanho da região anotada (quando seqüência codificante), organismo similar, *e-value*, *score*, base de dados, domínios encontrados pelo CDD / Pfam / Smart e InterProScan, grupo de ortólogos e as seqüências encontradas pelo *psiblast*.

Cluster	*G+C*	*Aa.*	*Descrição*	*Org.*	*Código de Acesso*	*e-value*	*Score*	*Base de dados*	*CDD / Pfam / Smart*	*KOG / COG*	*Interpro*
TGEG101003D05.g	53%	214	ATP-dependent RNA helicase	*T. cruzi*	EAN99766	2,0e-88	326	kinetoplastida-aa			
TGEG101004C10.g	31%		kinetoplasto DNA (minicircle replication)	*T. rangeli*	L28039	2,0e-40	161	minicircle	ATPase domain	ATPase domain	
TGEG101003D07.g	54%	217	Hypothetical Protein	*T. cruzi*	Q4E577	3,0e-70	266	Uniref90			
TGEG101004E05.g	30%		kinetoplasto minicircle DNA	*T. cruzi*	AAHK01021033	5,0e-23	111	tcruzi-nt.fasta			
TGEG101004H07.g	30%		kinetoplasto minicircle DNA	*T. rangeli*	L28039	2,0e-23	103	minicircle			
TGEG101003H12.g	57%	204	Hypothetical Protein	*T. cruzi*	XP_812744	5,0e-61	235	refseq_protein			
TGEG101038G02.b	51%	192	Hypothetical Protein	*T. cruzi*	Q4E5C8	3,0e-73	276	Uniref90	Leucine-rich repeats (LRRs)		
TGEG101047C02.b	32%		kinetoplasto minicircle DNA	*T. rangeli*	L28039	1,0e-36	147	minicircle			
TGEG101038A09.b	58%	78	ATP-dependent RNA helicase	*T. cruzi*	EAN99766	4,0e-30	127	kinetoplastida-aa	DEAD/DEAH box helicase	ATP-dependent RNA helicase	
TGEG101003C04.g	55%	138	Hypothetical Protein, Conserved	*T. cruzi*	XP_802527	9,0e-69	261	refseq_protein			conserved domain WD40
TGEG101004B09.g	56%	130	Hypothetical Protein	*T. cruzi*	AAHK01025937	9,0e-06	94	tcruzi-nt.fasta			
TGEG101003B10.g	49%	195	Hypothetical Protein	*T. cruzi*	EAN90857	2,0e-71	266	kinetoplastida-aa			
TGEG101012G08.g	55%	213	Hypothetical Protein	*T. cruzi*	EAN87977	1,0e-103	172	kinetoplastida-aa			
TGEG101048H11.b	59%	162	Hypothetical Protein	*T. cruzi*	gi|70877562	4,0e-56	213	Tcruzi_NCBI			
TGEG101038C09.b	64%	200	Hypothetical Protein	*T. cruzi*	EAN96718	2,0e-34	153	kinetoplastida-aa			
TGEG101010C10.g	54%	175	Thiopurine S-methyltransferase	*T. cruzi*	EAN83555	2,0e-56	216	kinetoplastida-aa			
TGEG101003B03.g	49%		18S ribosomal RNA gene	*T. cruzi*	AF359471	1,0e-126	454	tcruzi-nt.fasta			
TGEG101004H10.g	55%	77	Hypothetical Protein	*T. cruzi*	XP_816085	4,0e-32	139	refseq_protein	Dpy-30 motif		
TGEG101008G10.g (C)	53%	180	Hypothetical Protein	*T. cruzi*	XP_815903	2,0e-23	110	refseq_protein			
TGEG102054E04.b	54%	188	Hypothetical Protein	*T. cruzi*	Q4CR74	7,0e-73	275	Uniref90			
TGEG101011D06.g	61%	105	Hypothetical Protein	*T. cruzi*	Q4CV50	8,0e-36	151	Uniref90			

TGEG101004E10.g	54%	21	Hypothetical Protein	*T. cruzi*	AAHK01000072	1,0e-20	103	tcruzi-nt.fasta			
TGEG101010G04.g	60%	164	casein kinase	*T. cruzi*	AAHK01023451	4,0e-93	336	Tcruzi_NCBI	Serine/Threonine protein kinases domain		
TGEG101010B05.g	41%	150	peptide chain release factor 1	*T. cruzi*	XP_804953	6,0e-57	222	refseq_protein			
TGEG101010B09.g	67%	172	Hypothetical Protein	*T. cruzi*	EAN85385	3,0e-41	166	kinetoplastida-aa		WH2 domain	
TGEG101016H06.g	67%	95	Hypothetical Protein	*T. cruzi*	Q581W9_9TRYP	3,0e-24	112	uniprot_trembl		SWAP RNAm splicing regulator	
TGEG101004F10.g	61%	195	Hypothetical Protein	*T. cruzi*	Q4CWT0	2,0e-93	343	Uniref90			
TGEG101038H11.b	56%	84	Hypothetical Protein	*T. cruzi*	XP_816925	9,0e-19	95	refseq_protein			
TGEG101013C12.g	54%	119	Hypothetical Protein	*T. cruzi*	XP_813651	3,0e-49	196	refseq_protein			
TGEG101008H11.g	60%	183	Hypothetical Protein	*T. cruzi*	EAN95075	1,0e-43	174	kinetoplastida-aa			
TGEG101043B09.b	59%	167	Hypothetical Protein	*T. cruzi*	EAN92051	1,0e-48	190	kinetoplastida-aa			
TGEG101010A10.g	55%	80	TC38	*T. cruzi*	Q9GV75_TRYCR	1,0e-21	104	uniprot_trembl			
TGEG101016H11.g	32%		kinetoplasto minicircle DNA	*T. rangeli*	L28039	3,0e-40	159	minicircle			
TGEG101011G11.g	61%	169	Hypothetical Protein	*T. cruzi*	XP_814577	1,0e-57	230	refseq_protein			zinc metallopeptidase putative
TGEG101010A02.g	54%	180	ATP-dependent RNA helicase	*T. cruzi*	EAN99766	3,0e-70	226	kinetoplastida-aa		ATP-dependent RNA helicase	
TGEG101043D12.b	55%	150	Hypothetical Protein	*T. cruzi*	Q4D157	3,0e-36	152	Uniref90			
TGEG101003A08.g	54%	190	ATP-dependent RNA helicase	*T. cruzi*	EAN99766	9.0e-73	277	kinetoplastida-aa		ATP-dependent RNA helicase	
TGEG101043G05.b	57%	182	surface protease GP63	*T. cruzi*	Q4DTZ3	5,0e-51	202	Uniref90	Leishmanolysin		
TGEG101004C05.g	32%		kinetoplasto minicircle DNA	*T. cruzi*	M18815	2,0e-22	100	minicircle			
TGEG101043C12.b	53%	79	DNA primase small subunit	*T. cruzi*	Q4DH01	7,0e-37	155	Uniref90		Eukaryotic-type DNA primase, catalytic (small) subunit	
TGEG101007C12.g	59%	123	Hypothetical Protein	*T. cruzi*	XP_818569	1,0e-37	157	refseq_protein			

TGEG102053F12.b	56%	89	Hypothetical Protein	*T. cruzi*	gi|70872244	6,0e-14	73	Tcruzi_NCBI		
TGEG102053F08.b	53%	192	Hypothetical Protein	*T. cruzi*	EAN87977	2,0e-88	323	kinetoplastida-aa		
TGEG101013C01.g	47%	78	Hypothetical Protein	*T. cruzi*	EAO00000	3,0e-24	109	kinetoplastida-aa		
	50%	110	Hypothetical Protein	*T. cruzi*	EAN98750	3,0e-22	102	kinetoplastida-aa		
TGEG101010H05.g	55%	33	Hypothetical Protein	*T. cruzi*	EAN89903	9.0e-10	60	kinetoplastida-aa		
TGEG102051B06.b	43%	185	Hypothetical Protein	*T. cruzi*	XP_814126	1,0e-64	247	refseq_protein		
TGEG101003C03.g	51%	54	Hypothetical Protein	*T. brucei*	EAN78637	4,0e-06	48	kinetoplastida-aa		
TGEG101011D11.g	56%	96	RNA-binding protein	*T. brucei*	EAN91986	8,0e-06	49	kinetoplastida-aa		
TGEG101004D06.g	57%	44	Hypothetical Protein	*T. cruzi*	EAN95135	2,0e-06	51	kinetoplastida-aa		
TGEG101012G09.g	32%		kinetoplasto minicircle DNA	*T. cruzi*	M18815	2,0e-22	100	minicircle		
TGEG101018H06.g	53%	83	Hypothetical Protein	*T. cruzi*	AAP12841	8,0e-09	59	kinetoplastida-aa		
TGEG101043E02.b	49%	116	Hypothetical Protein	*T. cruzi*	EAN89902	9,0e-43	170	kinetoplastida-aa		
TGEG101010G02.g	50%	199	Hypothetical Protein	*T. cruzi*	XP_812708	7,0e-70	265	refseq_protein		
TGEG101038D02.b	56%	90	Hypothetical Protein	*T. cruzi*	Q4CTF1	9,0e-33	140	Uniref90		
TGEG101011E03.g	55%	69	ATP-dependent RNA helicase	*T. cruzi*	EAN99766	5,0e-23	108	kinetoplastida-aa		
TGEG102054F01.b	62%	158	40S ribosomal protein S2	*T. cruzi*	Q4DIZ9	2,0e-63	243	Uniref90		
TGEG101011E12.g	30%		kinetoplasto minicircle DNA	*T. cruzi*	M18815	2,0e-22	100	minicircle		
TGEG101011E10.g	59%	79	surface protease GP63	*T. cruzi*	EAN82680	3,0e-22	102	kinetoplastida-aa		
TGEG102023E07.g	58%	107	ATP-dependent RNA helicase	*T. cruzi*	EAN99766	6,0e-38	157	kinetoplastida-aa		
TGEG101014H08.g	50%	47	Hypothetical Protein	*T. cruzi*	EAO00000	2,0e-12	69	kinetoplastida-aa		
TGEG102034D01.b	58%	67	ATP-dependent RNA helicase	*T. brucei*	EAN99766	1,0e-09	64	kinetoplastida-aa		
TGEG101043F05.b	53%	206	ATP-dependent RNA helicase	*T. cruzi*	EAN82784	5,0e-95	348	kinetoplastida-aa	DEXH-box helicases domain	ATP-dependent RNA helicase
TGEG101011A06.g	59%	52	Hypothetical Protein	*T. cruzi*	EAN81081	2,0e-22	102	kinetoplastida-aa		
TGEG101043B04.b	49%	47	Bem46-like serine peptidase	*T. cruzi*	Q4E4T5	3,0e-09	63	Uniref90		
TGEG101047B06.b	55%	34	Hypothetical Protein	*T. cruzi*	gi|70831978	2,0e-15	77	Tcruzi_NCBI		
TGEG101047D01.b	66%	78	Hypothetical Protein	*T. cruzi*	EAN82171	4,0e-30	127	kinetoplastida-aa		
TGEG101013H11.g	64%	200	Hypothetical Protein	*T. cruzi*	Q57U44	2,0e-24	114	Uniref90	Atrophin-1	
TGEG101038G11.b	66%	60	Hypothetical Protein	*T. cruzi*	EAN86202	3,0e-10	63	kinetoplastida-aa		
TGEG101043D10.b	48%	46	Bem46-like serine peptidase	*T. cruzi*	gi|70887002	8,0e-10	61	Tcruzi_NCBI		
	59%	38	ATP-dependent RNA helicase	*T. cruzi*	EAN99767	1,0e-07	52	kinetoplastida-aa		
TGEG102053E07.b	48%	44	Bem46-like serine	*T. cruzi*	gi|70887002	2,0e-08	55	Tcruzi_NCBI		

			peptidase								
TGEG101011F12.g	56%	69	Hypothetical Protein	*T. cruzi*	EAN83775	8,0e-07	52	kinetoplastida-aa			
TGEG101048G01.b	64%	144	Hypothetical Protein	*T. cruzi*	XP_815139	9.0e-20	97	refseq_protein			
TGEG101011F10.g	53%	169	ATP-dependent RNA helicase	*T. cruzi*	EAN87989	6,0e-69	261	kinetoplastida-aa	HrpA-like helicases	DEAD/DEAH box helicase	
TGEG101047F10.b	66%	145	Hypothetical Protein	*T. cruzi*	EAN84413	1,0e-15	80	kinetoplastida-aa			
TGEG101018F03.g	58%	110	ATP-dependent RNA helicase	*T. cruzi*	EAN99766	3,0e-34	145	kinetoplastida-aa			
TGEG101038D09.b	60%	62	Hypothetical Protein	*T. cruzi*	EAN84732	2,0e-21	99	kinetoplastida-aa	T-cell surface antigen CD2 protein	GTPase effector FRL	
TGEG101012E02.g	56%	102	Hypothetical Protein	*T. cruzi*	EAN99964	5,0e-26	114	kinetoplastida-aa			
TGEG101047D09.b	57%	145	Hypothetical Protein	*T. cruzi*	EAN87111	7.0e-29	124	kinetoplastida-aa			
TGEG101001D09.g	51%	76	Hypothetical Protein	*T. cruzi*	EAN96954	4,0e-15	79	kinetoplastida-aa			
TGEG102054C12.b	58%	69	Tubulin-tyrosine ligase-like protein	*T. cruzi*	Q4DRI1	4,0e-10	65	Uniref90			
TGEG101004G07.g	56%	136	Hypothetical Protein	*T. cruzi*	Q4E577	3,0e-28	125	Uniref90			
TGEG101001B01.g	56%	155	trans-sialidase	*T. cruzi*	XP_809726	8,0e-18	92	refseq_protein	Sialidase		TCSIALIDASE
TGEG101018F10.g	58%	95	ATP-dependent RNA helicase	*T. cruzi*	EAN99766	5,0e-30	127	kinetoplastida-aa			
TGEG101005H08.g	31%		kinetoplasto minicircle DNA	*T. rangeli*	M19178	5,0e-17	82	minicircle			
TGEG102054F04.b	64%	102	Hypothetical Protein	*T. cruzi*	XP_816548	2,0e-41	169	refseq_protein			
TGEG102034E05.b	34%		kinetoplasto minicircle DNA	*T. cruzi*	M18815	4,0e-17	82	minicircle			
TGEG101007B05.g	32%		kinetoplasto minicircle DNA	*T. rangeli*	AJ747972	1,0e-06	46	minicircle			
TGEG102023G04.g	65%	55	Kinesin-like protein	*T. cruzi*	XP_806544	1,0e-13	77	refseq_protein			
TGEG101003H04.g	58%	135	ATP-dependent RNA helicase	*T. cruzi*	EAN99766	3.0e-41	169	kinetoplastida-aa			
TGEG101011D05.g	32%		kinetoplasto minicircle DNA	*T. rangeli*	L28039	1,0e-10	60	minicircle			
TGEG102053B10.b	48%	47	Bem46-like serine peptidase	*T. cruzi*	Q4E4T5	4,0e-09	62	Uniref90			
TGEG101012E08.g	56%	81	ATP-dependent RNA helicase	*T. cruzi*	EAN99766	2,0e-28	126	kinetoplastida-aa			
TGEG101005D04.g	54%	84	Hypothetical Protein	*T. cruzi*	Q4E577	2,0e-16	86	Uniref90			
TGEG101012E06.g	64%	72	Hypothetical Protein	*T. cruzi*	EAN89090	3,0e-10	63	kinetoplastida-aa			
TGEG102053A08.b	48%	44	Bem46-like serine peptidase	*T. cruzi*	EAN99767	5,0e-08	55	kinetoplastida-aa			

TGEG101005H01.g	58%	40	Hypothetical Protein	*T. cruzi*	Q4D157	3,0e-13	76	Uniref90
TGEG102022F03.g	55%	109	ATP-dependent RNA helicase	*T. cruzi*	EAN99766	2,0e-30	132	kinetoplastida-aa
TGEG101011G05.g	32%		kinetoplasto minicircle DNA	*T. rangeli*	M18815	9,0e-22	98	minicircle
TGEG102052E07.b	32%		kinetoplasto minicircle DNA	*T. rangeli*	L28039	1,0e-17	84	minicircle
TGEG102021H07.g	29%		kinetoplasto minicircle DNA	*T. rangeli*	L28039	8,0e-29	121	minicircle
TGEG102051B04.b	52%	166	Hypothetical Protein	*T. cruzi*	EAN82775	3,0e-39	159	kinetoplastida-aa
TGEG101038G12.b	58%	45	Hypothetical Protein	*T. cruzi*	Q4CXH8	3,0e-09	62	Uniref90
TGEG102023H08.g	51%	75	ATP-dependent RNA helicase	*T. cruzi*	EAN87989	2,0e-32	135	kinetoplastida-aa
TGEG102034B03.b	54%	65	Hypothetical Protein	*T. cruzi*	EAN84732	1,0e-21	100	kinetoplastida-aa
TGEG102051B03.b	47%	29	Bem46-like serine peptidase	*T. cruzi*	EAN99767	6,0e-06	47	kinetoplastida-aa
TGEG101003F02.g	58%	151	Hypothetical Protein	*T. cruzi*	EAN86426	6,0e-23	105	kinetoplastida-aa
TGEG102003D04.g	58%	60	ATP-dependent RNA helicase	*T. cruzi*	EAN99766	5,0e-11	69	kinetoplastida-aa
TGEG101018D01.g	56%	106	Hypothetical Protein	*T. cruzi*	EAN99964	3,0e-27	118	kinetoplastida-aa
TGEG102005G07.g	63%	73	Hypothetical Protein	*T. cruzi*	XP_818238	3,0e-18	92	refseq_protein
TGEG102054A10.b	59%	49	Hypothetical Protein	*T. cruzi*	XP_816548	3,0e-03	42	refseq_protein
TGEG102051D04.b	49%	40	Bem46-like serine peptidase	*T. cruzi*	EAN99767	3,0e-09	59	kinetoplastida-aa
TGEG102023D03.g	56%	123	ATP-dependent RNA helicase	*T. cruzi*	EAN99766	6,0e-34	144	kinetoplastida-aa
TGEG101016A12.g	53%	76	surface protease GP63	*T. cruzi*	Q4D292	2,0e-07	57	Uniref90
TGEG101004G05.g	58%	27	Hypothetical Protein	*T. cruzi*	XP_808109	2,0e-06	53	refseq_protein
TGEG102003F02.g	34%		kinetoplasto minicircle DNA	*T. rangeli*	L28039	1.0e-17	84	minicircle
TGEG101007B09.g	58%	42	Hypothetical Protein	*T. cruzi*	Q4CZX2	3,0e-11	69	Uniref90
TGEG101047A05.b	54%	117	Hypothetical Protein	*T. cruzi*	gi|70887204	9,0e-38	151	Tcruzi_NCBI
TGEG102053H12.b	55%	37	surface protease GP63	*T. cruzi*	gi|70883803	5,0e-06	45	Tcruzi_NCBI
TGEG102021A10.g	50%	47	Hypothetical Protein	*T. cruzi*	EAN90857	2,0e-13	72	kinetoplastida-aa
TGEG101014D05.g	62%	97	Hypothetical Protein	*T. cruzi*	EAN88296	1,0e-20	96	kinetoplastida-aa
TGEG102052B09.b	64%	74	Hypothetical Protein	*T. cruzi*	EAN86202	5,0e-12	69	kinetoplastida-aa
TGEG102021B05.g	31%		kinetoplasto minicircle DNA	*T. rangeli*	L28039	5,0e-39	155	minicircle
TGEG101015H07.g	54%	38	Hypothetical Protein	*T. cruzi*	EAN81081	3,0e-08	55	kinetoplastida-aa
TGEG101001C01.g	47%	134	Hypothetical Protein	*T. cruzi*	gi|70883473	5,0e-21	97	Tcruzi_NCBI
TGEG101004B07.g	56%	72	ATP-dependent RNA	*T. cruzi*	EAN99766	2,0e-45	188	kinetoplastida-aa

			helicase							
TGEG102052E04.b	44%	141	Bem46-like serine peptidase	*T. cruzi*	EAN99767	6,0e-06	48	kinetoplastida-aa		
TGEG101038E11.b	47%	44	Bem46-like serine peptidase	*T. cruzi*	EAN99767	9,0e-09	57	kinetoplastida-aa		
TGEG101012G10.g	51%	54	Hypothetical Protein	*T. brucei*	gi|70833133	5,0e-07	48	Tbrucei_NCBI		
TGEG101003H07.g	55%	54	Hypothetical Protein	*T. cruzi*	EAN87977	4.0e-23	104	kinetoplastida-aa		
TGEG101015G09.g	55%	58	Hypothetical Protein	*L. major*	CT005257	2,0e-06	52	lmajor-nt		
TGEG101005C08.g	35%	52	peptide chain release factor 1	*T. cruzi*	EAN83102	9.0e-20	93	kinetoplastida-aa		
TGEG102022D08.g	59%	51	Hypothetical Protein	*T. cruzi*	EAN82776	5.0e-23	104	kinetoplastida-aa		
TGEG102052E01.b	63%	90	Hypothetical Protein	*T. cruzi*	XP_819916	2,0e-16	87	refseq_protein		
TGEG102021D02.g	37%		kinetoplasto minicircle DNA	*T. rangeli*	L28039	7,0e-24	103	minicircle		
TGEG102053A05.b	53%	121	vesicular transport protein (CDC48 homologue)	*T. cruzi*	EAN94730	2.0e-58	221	kinetoplastida-aa	ATPase doamin	
TGEG102021F02.g	49%	68	Hypothetical Protein	*T. cruzi*	EAN99964	2.0e-06	49	kinetoplastida-aa		
TGEG102051C04.b	48%	59	Hypothetical Protein	*T. cruzi*	XM_809033	2,0e-11	70	refseq_protein		
TGEG101003A09.g	53%	72	Hypothetical Protein	*T. cruzi*	Q4DJA2	6,0e-21	101	Uniref90		
TGEG101048G11.b	54%	63	DNA primase small subunit	*T. cruzi*	Q4DH01	8,0e-18	91	Uniref90		
TGEG101018B06.g	57%	111	ATP-dependent RNA helicase	*T. cruzi*	EAN99766	9,0e-38	154	kinetoplastida-aa		
TGEG101005G01.g	31%		kinetoplasto minicircle DNA	*T. rangeli*	L28039	2,0e-11	62	minicircle		
TGEG101047H06.b	50%	40	Bem46-like serine peptidase	*T. cruzi*	EAN99767	6,0e-09	57	kinetoplastida-aa		
TGEG102021A12.g	51%	95	mitochondrial ATP-dependent zinc metallopeptidase	*T. cruzi*	Q4DAV1	1,0e-07	57	Uniref90		
TGEG102021G07.g	51%	64	Hypothetical Protein	*T. cruzi*	EAN99964	1,0e-23	107	kinetoplastida-aa		
TGEG102033E03.b	61%	127	Hypothetical Protein	*T. cruzi*	EAN96086	4,0e-38	155	kinetoplastida-aa		
TGEG102023F07.g	34%		kinetoplasto minicircle DNA	*T. rangeli*	L28039	1,0e-13	70	minicircle		
TGEG101001D01.g	52%	140	mitochondrial ATP-dependent zinc metallopeptidase	*T. cruzi*	XP_821024	8,0e-70	263	refseq_protein	ATP-dependent Zn proteases domain	ATPAse associated
TGEG101004D03.g	55%	40	Hypothetical Protein	*T. cruzi*	EAN83555	2,0e-08	56	kinetoplastida-aa		
TGEG101018C10.g	55%	117	Hypothetical Protein	*T. cruzi*	EAN85821	3,0e-25	112	kinetoplastida-aa		
TGEG102005D04.g	59%	75	Hypothetical Protein	*T. cruzi*	EAN99964	3,0e-11	65	kinetoplastida-aa		
TGEG102053F09.b	39%		kinetoplasto minicircle DNA	*T. rangeli*	L28039	1,0e-16	78	minicircle		

TGEG102021H08.g	53%	38	Hypothetical Protein	*T. cruzi*	EAN96954	2.0e-09	59	kinetoplastida-aa	
TGEG101014G08.g	57%	61	Hypothetical Protein	*T. cruzi*	EAN81081	2.0e-21	99	kinetoplastida-aa	
TGEG101012G04.g	52%	40	Hypothetical Protein	*T. cruzi*	EAN85821	4,0e-12	68	kinetoplastida-aa	
TGEG101043D08.b	51%	49	Hypothetical Protein	*T. cruzi*	EAN82775	7,0e-07	50	kinetoplastida-aa	
TGEG102052G08.b	62%	48	Ecotin	*T. brucei*	AAX79124	2,0e-08	57	kinetoplastida-aa	Ecotin, Protease Inhibitor
TGEG102021G03.g	40%	63	peptide chain release factor 1	*T. cruzi*	EAN83102	4,0e-25	112	kinetoplastida-aa	
TGEG101003F12.g	63%	55	Hypothetical Protein	*T. cruzi*	EAN82171	8,0e-20	94	kinetoplastida-aa	
TGEG101001D06.g	57%	170	Hypothetical Protein	*T. cruzi*	EAN92000	3,0e-65	264	kinetoplastida-aa	
TGEG101004G09.g	53%	70	TC38	*T. cruzi*	Q9GV75	2,0e-14	81	Uniref90	
TGEG101011A12.g	55%	166	protein kinase	*T. cruzi*	EAN97908	2,0e-28	123	kinetoplastida-aa	
TGEG101018B09.g	54%	62	Hypothetical Protein	*T. cruzi*	EAN99964	8,0e-14	74	kinetoplastida-aa	
TGEG101038C05.b	57%	118	Hypothetical Protein	*T. cruzi*	Q4DXR2	2.0e-55	155	Uniref90	
TGEG101015E06.g	54%	185	Tc40 antigen	*T. cruzi*	Q26872_TRYCR	9,0e-63	240	uniprot_trembl	
TGEG102003G05.g	61%	35	Hypothetical Protein	*T. cruzi*	EAN99964	5,0e-08	54	kinetoplastida-aa	
TGEG101003G10.g	57%	113	surface protease GP63	*T. cruzi*	Q4D1E6	1,0e-27	123	Uniref90	
TGEG102053H08.b	55%	71	ATP-dependent RNA helicase	*T. cruzi*	EAN82784	1,0e-25	112	kinetoplastida-aa	
TGEG102022A05.g	52%		18S ribosomal RNA gene	*T. cruzi*	AF359471	2,0e-85	317	tcruzi-nt.fasta	
TGEG101010F12.g	30%		kinetoplasto minicircle DNA	*T. rangeli*	L28039	4,0e-10	58	minicircle	
TGEG102052F04.b	60%	157	surface protease GP63	*T. cruzi*	Q4DF53	4,0e-45	182	Uniref90	Leishmanolysin
TGEG101038F10.b	52%	118	Hypothetical Protein	*T. cruzi*	EAN99965	4,0e-34	140	kinetoplastida-aa	
TGEG102053C06.b	50%	42	Hypothetical Protein	*T. cruzi*	AAX79031	4.0e-07	54	kinetoplastida-aa	
TGEG102021H06.g	36%		kinetoplasto minicircle DNA	*T. cruzi*	M19191	4,0e-15	74	minicircle	
TGEG102051H08.b	52%	194	Hypothetical Protein, Conserved	*T. cruzi*	XP_818805	4,0e-62	239	refseq_protein	
TGEG101003E12.g	33%		kinetoplasto minicircle DNA	*T. cruzi*	M18815	4,0e-20	92	minicircle	
TGEG102053D05.b	52%	95	Hypothetical Protein	*T. cruzi*	EAN99965	5,0e-27	117	kinetoplastida-aa	
TGEG101005A06.g	34%		kinetoplasto minicircle DNA	*T. rangeli*	L28039	7,0e-41	161	minicircle	
TGEG102022C07.g	55%	171	UDP-glucose:glycoprotein glucosyltransferase	*T. cruzi*	Q86G51	6,0e-52	204	Uniref90	
TGEG102024D02.g	55%	45	ATP-dependent RNA helicase	*T. cruzi*	EAN99766	8,0e-24	106	kinetoplastida-aa	
TGEG102024F03.g	54%	95	Hypothetical Protein	*T. cruzi*	EAN85821	1,0e-18	90	kinetoplastida-aa	

TGEG102033B09.b	51%	58	Hypothetical Protein	*T. cruzi*	EAN99965	2.0e-14	75	kinetoplastida-aa
TGEG101013B07.g	50%	68	Hypothetical Protein	*T. cruzi*	EAN90857	2,0e-22	102	kinetoplastida-aa
TGEG102023F05.g	35%		kinetoplasto minicircle DNA	*T. rangeli*	L28039	3,0e-20	92	minicircle
TGEG102033E12.b	33%		kinetoplasto minicircle DNA	*T. rangeli*	L28039	3,0e-10	58	minicircle
TGEG101003H06.g	59%	45	Hypothetical Protein	*T. cruzi*	EAN96718	2,0e-16	82	kinetoplastida-aa
TGEG102033A01.b	54%	56	DNA primase small subunit	*T. cruzi*	EAN91801	3,0e-25	113	kinetoplastida-aa
TGEG101047C06.b	57%	89	Hypothetical Protein	*T. cruzi*	EAN83556	2,0e-32	135	kinetoplastida-aa
TGEG102054A03.b	47%	47	serine peptidase	*T. cruzi*	gi|70887002	5,0e-06	45	Tcruzi_NCBI
TGEG102021E10.g	56%	53	Hypothetical Protein	*T. cruzi*	EAN96718	8,0e-16	81	kinetoplastida-aa
TGEG102024H05.g	62%	50	Hypothetical Protein, Conserved	*T. cruzi*	Q4CWT0	2,0e-32	140	Uniref90
TGEG101011F04.g	56%	64	Hypothetical Protein	*T. cruzi*	EAN95789	1,0e-16	84	kinetoplastida-aa
TGEG101043B03.b	52%	67	Hypothetical Protein	*T. cruzi*	AAX79031	3,0e-06	48	kinetoplastida-aa
TGEG102005B05.g	53%	36	Hypothetical Protein	*T. cruzi*	EAN99964	2,0e-07	52	kinetoplastida-aa
TGEG101003G02.g	56%	85	Hypothetical Protein	*T. cruzi*	EAN96718	3,0e-25	111	kinetoplastida-aa
TGEG101001H04.g	49%	112	Hypothetical Protein	*T. cruzi*	EAN90857	1,0e-41	166	kinetoplastida-aa
TGEG101016A09.g	54%	36	Hypothetical Protein	*T. cruzi*	EAN84732	5,0e-14	74	kinetoplastida-aa
TGEG102003C01.g	55%	138	Hypothetical Protein, Conserved	*T. cruzi*	XP_802527	2,0e-63	242	refseq_protein
TGEG102051C07.b	51%	193	UDP-glucose:glycoprotein glucosyltransferase	*T. cruzi*	XP_821190	2,0e-92	339	refseq_protein
TGEG102021C04.g	46%	122	peptide chain release factor 1	*T. cruzi*	EAN83102	5,0e-36	148	kinetoplastida-aa
TGEG101003C11.g	57%	54	TC38	*T. cruzi*	Q9GV75	5,0e-08	59	Uniref90
TGEG101015B08.g	51%	53	Hypothetical Protein	*T. cruzi*	EAN99964	2,0e-11	66	kinetoplastida-aa
TGEG101001C08.g	61%	38	Hypothetical Protein	*T. cruzi*	EAN92726	3,0e-17	85	kinetoplastida-aa
TGEG102022A06.g	46%	141	Hypothetical Protein	*T. cruzi*	EAN89156	6,0e-27	115	Tcruzi_NCBI
TGEG101010E01.g	30%		kinetoplasto minicircle DNA	*T. rangeli*	M18815	9,0e-22	98	minicircle
TGEG102052G11.b	60%	162	trans-sialidase	*T. cruzi*	XP_802237	2,0e-35	149	refseq_protein
TGEG102052A03.b	51%	70	Hypothetical Protein	*T. cruzi*	EAN99965	3,0e-18	88	kinetoplastida-aa
TGEG101008D07.g	56%	54	Hypothetical Protein	*T. cruzi*	EAN77561	8,0e-16	80	kinetoplastida-aa
TGEG101038B12.b	55%	47	surface protease GP63	*T. cruzi*	XP_813009	2,0e-10	67	refseq_protein
TGEG102034C05.b	35%		kinetoplasto minicircle DNA	*T. rangeli*	AJ748059	1,0e-17	88	minicircle
TGEG102051A03.b	38%		kinetoplasto minicircle DNA	*T. cruzi*	AF188228	8,0e-10	56	minicircle
TGEG102022A04.g	55%	93	glycosyltransferase	*T. cruzi*	EAN86423	2,0e-27	118	kinetoplastida-aa

TGEG102053C10.b	56%	91	Hypothetical Protein	*T. cruzi*	gi|70868802	2,0e-34	140	Tcruzi_NCBI
TGEG101010D02.g	51%		18S ribosomal RNA gene	*T. cruzi*	AF359471	5,0e-87	323	tcruzi-nt.fasta
TGEG102024F02.g	33%		kinetoplasto minicircle DNA	*T. rangeli*	L28039	2,0e-35	143	minicircle
TGEG102021D11.g	34%		kinetoplasto minicircle DNA	*T. rangeli*	L28039	5,0e-15	74	minicircle
TGEG102052G06.b	55%		28S ribosomal RNA	*T. cruzi*	L22334	2,0e-92	341	tcruzi-nt.fasta
TGEG102034E02.b	57%	78	Hypothetical Protein	*T. cruzi*	EAN99964	2,0e-07	52	kinetoplastida-aa
TGEG102023A02.g	56%	36	ATP-dependent RNA helicase	*T. cruzi*	EAN77043	2,0e-06	49	kinetoplastida-aa
TGEG102034F07.b	52%	36	Hypothetical Protein	*T. cruzi*	EAN99964	2,0e-08	56	kinetoplastida-aa
TGEG102052A04.b	54%	148	glycosyltransferase	*T. cruzi*	EAN86423	2,0e-60	229	kinetoplastida-aa
TGEG102033G07.b	48%	47	Bem46-like serine peptidase	*T. cruzi*	EAN99767	4,0e-09	59	kinetoplastida-aa
TGEG102054F03.b	60%	86	Hypothetical Protein	*T. cruzi*	EAN90893	2,0e-19	92	kinetoplastida-aa
TGEG101001A01.g	59%	168	trans-sialidase	*T. cruzi*	XP_809726	6,0e-18	92	refseq_protein
TGEG101004F07.g	57%	80	Hypothetical Protein	*T. cruzi*	EAN99964	2,0e-21	99	kinetoplastida-aa
TGEG102053A06.b	61%	45	Ecotin	*T. brucei*	AAX79124	1,0e-08	57	kinetoplastida-aa
TGEG101012F06.g	57%	135	ATP-dependent RNA helicase	*T. cruzi*	EAN99766	4,0e-43	171	kinetoplastida-aa
TGEG101003A01.g	55%	77	Hypothetical Protein	*T. cruzi*	EAN91800	4,0e-28	122	kinetoplastida-aa
TGEG101001D12.g	56%	88	ATP-dependent RNA helicase	*T. cruzi*	EAN99766	2,0e-49	192	kinetoplastida-aa
TGEG101016G08.g	45%	43	Hypothetical Protein	*T. cruzi*	EAN89728	1,0e-17	86	kinetoplastida-aa
TGEG101004F02.g	51%	48	Hypothetical Protein	*T. cruzi*	EAN92436	5,0e-07	51	kinetoplastida-aa
TGEG102022C11.g	57%	49	Hypothetical Protein	*T. cruzi*	EAN99964	7,0e-07	50	kinetoplastida-aa
TGEG101011G08.g	58%	54	ATP-dependent RNA helicase	*T. cruzi*	EAN99766	4,0e-18	88	kinetoplastida-aa
TGEG102052G03.b	41%		kinetoplasto minicircle DNA	*T. rangeli*	L28039	8,0e-25	105	kinetoplastida-aa
TGEG101007H12.g	57%	53	RNA-binding protein	*T. cruzi*	Q9GV75	2,0e-08	60	Uniref90
TGEG102051F05.b	40%		kinetoplasto minicircle DNA	*T. cruzi*	AY169986	9,0e-06	44	minicircle
TGEG101008A02.g	44%	57	Hypothetical Protein	*T. cruzi*	EAO00000	4,0e-07	51	kinetoplastida-aa
TGEG101038C02.b	58%	24	Hypothetical Protein	*T. brucei*	gi|70831978	4,0e-07	44	Tbrucei_NCBI
TGEG102023A05.g	57%	85	vesicular transport protein (CDC48 homologue)	*T. cruzi*	Q4DQD6	3,0e-30	132	Uniref90
TGEG101003B09.g	54%	38	Hypothetical Protein	*T. cruzi*	EAN99964	2,0e-07	52	kinetoplastida-aa
TGEG101003C01.g	61%	42	Hypothetical Protein	*T. cruzi*	EAN81081	7,0e-15	77	kinetoplastida-aa
TGEG101015D07.g	60%	66	Hypothetical Protein	*T. cruzi*	XP_816986	1,0e-22	106	refseq_protein
TGEG102034G08.b	64%	35	Hypothetical Protein	*T. cruzi*	XP_808053	3,0e-37	155	refseq_protein

Tabela A3: Tabela contendo os grupos de ortólogos formado pela comparação entre as 244 seqüências anotadas neste projeto e os genomas dos TriTryps. Na primeira coluna temos código de cada grupo de ortólogo, em seguida o organismo, código de acesso (para *T. rangeli* referente ao código da seqüência neste projeto e para os demais é o código de acesso do GenBank) e a descrição de cada gene.

Grupo	***Organismo***	***Código de acesso***	***Descrição***
ORTHOMCL4178	*Trypanosoma rangeli*	TGEG102051B06.b	Proteína Hipotética
	Leishmania major	gi|68126970	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70868165	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70879072	hypothetical protein, conserved
ORTHOMCL7914	*Trypanosoma rangeli*	TGEG102023A02.g	RNA helicase ATP-dependente
	Trypanosoma rangeli	TGEG102024D02.g	RNA helicase ATP-dependente
ORTHOMCL4183	*Trypanosoma rangeli*	TGEG101043F05.b	RNA helicase ATP-dependente
	Leishmania major	gi|68128080	ATP-dependent RNA helicase, putative
	Trypanosoma brucei	gi|70801072	ATP-dependent RNA helicase, putative
	Trypanosoma cruzi	gi|70874282	ATP-dependent RNA helicase, putative
ORTHOMCL685	*Trypanosoma rangeli*	TGEG101003A09.g	Proteína Hipotética
	Leishmania major	gi|68224099	hypothetical protein, conserved
	Trypanosoma brucei	gi|70802415	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70879429	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70883621	hypothetical protein, conserved
ORTHOMCL4190	*Trypanosoma rangeli*	TGEG101008G10.g	Proteína Hipotética
	Trypanosoma brucei	gi|70834503	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70879247	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70880993	hypothetical protein, conserved
ORTHOMCL7919	*Trypanosoma rangeli*	TGEG101038F10.b	Proteína Hipotética
	Trypanosoma rangeli	TGEG101047A05.b	Proteína Hipotética
ORTHOMCL143	*Trypanosoma rangeli*	TGEG102054F01.b	Proteína 40S ribosomal S2
	Leishmania major	gi|68126038	40S ribosomal protein S2
	Leishmania major	gi|68128430	40S ribosomal protein S2
	Trypanosoma brucei	gi|70833323	40S ribosomal protein S2, putative
	Trypanosoma brucei	gi|70833333	40S ribosomal protein S2, putative
	Trypanosoma cruzi	gi|70869314	40S ribosomal protein S2, putative
	Trypanosoma cruzi	gi|70876647	40S ribosomal protein S2, putative
	Trypanosoma cruzi	gi|70879313	40S ribosomal protein S2, putative
	Trypanosoma cruzi	gi|70879320	40S ribosomal protein S2, putative
ORTHOMCL7923	*Trypanosoma rangeli*	TGEG101004D03.g	Proteína Hipotética
	Trypanosoma cruzi	gi|70868801	hypothetical protein Tc00.1047053503651.30
ORTHOMCL190	*Trypanosoma rangeli*	TGEG101003H04.g	RNA helicase ATP-dependente
	Trypanosoma rangeli	TGEG101012F06.g	RNA helicase ATP-dependente
	Trypanosoma rangeli	TGEG101012E08.g	RNA helicase ATP-dependente
	Trypanosoma rangeli	TGEG101018F03.g	RNA helicase ATP-dependente
	Trypanosoma rangeli	TGEG101018F10.g	RNA helicase ATP-dependente
	Trypanosoma rangeli	TGEG101018B06.g	RNA helicase ATP-dependente
	Trypanosoma rangeli	TGEG102023D03.g	RNA helicase ATP-dependente
	Trypanosoma rangeli	TGEG102023E07.g	RNA helicase ATP-dependente

ORTHOMCL144	*Trypanosoma rangeli*	TGEG101010G04.g	caseina kinase
	Trypanosoma brucei	gi|70801288	casein kinase, putative
	Trypanosoma cruzi	gi|70865911	casein kinase, delta isoform, putative
	Trypanosoma cruzi	gi|70867272	casein kinase, putative
	Trypanosoma cruzi	gi|70867273	casein kinase, putative
	Trypanosoma cruzi	gi|70868437	casein kinase, putative
	Trypanosoma cruzi	gi|70880947	casein kinase, putative
	Trypanosoma cruzi	gi|70880948	casein kinase, putative
	Leishmania major	gi|70905689	casein kinase, putative
ORTHOMCL7926	*Trypanosoma rangeli*	TGEG101003H06.g	Proteína Hipotética
	Trypanosoma rangeli	TGEG102021E10.g	Proteína Hipotética
ORTHOMCL265	*Trypanosoma rangeli*	TGEG101004G09.g	TC38
	Trypanosoma rangeli	TGEG101007H12.g	Proteína de ligadção ao RNA
	Trypanosoma rangeli	TGEG101010A10.g	TC38
	Leishmania major	gi|68126034	hypothetical protein, conserved
	Trypanosoma brucei	gi|70833332	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70869315	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70879319	hypothetical protein, conserved
ORTHOMCL7916	*Trypanosoma rangeli*	TGEG101047F10.b	Proteína Hipotética
	Trypanosoma cruzi	gi|70875556	hypothetical protein Tc00.1047053510823.90
ORTHOMCL4176	*Trypanosoma rangeli*	TGEG102054E04.b	Proteína Hipotética
	Leishmania major	gi|68125357	hypothetical protein, conserved
	Trypanosoma brucei	gi|70802516	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70867696	hypothetical protein, conserved
ORTHOMCL669	*Trypanosoma rangeli*	TGEG102054F04.b	Proteína Hipotética
	Leishmania major	gi|68223795	hypothetical protein, conserved
	Trypanosoma brucei	gi|70802848	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70881684	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70885238	hypothetical protein, conserved
ORTHOMCL4188	*Trypanosoma rangeli*	TGEG101011F12.g	Proteína Hipotética
	Trypanosoma cruzi	gi|70868480	hypothetical protein Tc00.1047053509281.30
	Trypanosoma cruzi	gi|70869108	hypothetical protein Tc00.1047053508087.10
	Trypanosoma cruzi	gi|70871846	hypothetical protein Tc00.1047053511873.5
ORTHOMCL145	*Trypanosoma rangeli*	TGEG101004F07.g	Proteína Hipotética
	Trypanosoma rangeli	TGEG101004G07.g	Proteína Hipotética
	Trypanosoma rangeli	TGEG101005D04.g	Proteína Hipotética
	Trypanosoma rangeli	TGEG101012E02.g	Proteína Hipotética
	Trypanosoma rangeli	TGEG101015B08.g	Proteína Hipotética
	Trypanosoma rangeli	TGEG101018B09.g	Proteína Hipotética
	Trypanosoma rangeli	TGEG101018D01.g	Proteína Hipotética
	Trypanosoma rangeli	TGEG102005B05.g	Proteína Hipotética
	Trypanosoma rangeli	TGEG102005D04.g	Proteína Hipotética
ORTHOMCL378	*Trypanosoma rangeli*	TGEG101003B10.g	Proteína Hipotética
	Trypanosoma rangeli	TGEG101010G02.g	Proteína Hipotética
	Leishmania major	gi|68224011	hypothetical protein, conserved
	Trypanosoma brucei	gi|70831964	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70872774	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70877522	hypothetical protein, conserved
ORTHOMCL7921	*Trypanosoma rangeli*	TGEG101011A06.g	Proteína Hipotética
	Trypanosoma rangeli	TGEG102022D08.g	Proteína Hipotética

ORTHOMCL4189	*Trypanosoma rangeli*	TGEG101010B09.g	Proteína Hipotética
	Trypanosoma brucei	gi|70834415	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70871194	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70872176	hypothetical protein, conserved
ORTHOMCL6773	*Trypanosoma rangeli*	TGEG101016H06.g	Proteína Hipotética
	Leishmania major	gi|68126404	hypothetical protein, conserved
	Trypanosoma brucei	gi|70803076	hypothetical protein, conserved
ORTHOMCL680	*Trypanosoma rangeli*	TGEG101011D06.g	Proteína Hipotética
	Leishmania major	gi|68127689	hypothetical protein, conserved
	Trypanosoma brucei	gi|70801827	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70869623	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70877947	hypothetical protein, conserved
ORTHOMCL377	*Trypanosoma rangeli*	TGEG101004F10.g	Proteína Hipotética
	Leishmania major	gi|68128177	hypothetical protein, conserved
	Trypanosoma brucei	gi|70803654	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70870368	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70878353	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70879888	hypothetical protein, conserved
ORTHOMCL7917	*Trypanosoma rangeli*	TGEG101043B03.b	Proteína Hipotética
	Trypanosoma rangeli	TGEG102053C06.b	Proteína Hipotética
ORTHOMCL673	*Trypanosoma rangeli*	TGEG101047B06.b	Proteína Hipotética
	Leishmania major	gi|68223994	hypothetical protein, conserved
	Trypanosoma brucei	gi|70831978	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70864947	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70872118	hypothetical protein, conserved
ORTHOMCL688	*Trypanosoma rangeli*	TGEG101001C01.g	Proteína Hipotética
	Trypanosoma rangeli	TGEG102022A06.g	Proteína Hipotética
	Trypanosoma cruzi	gi|70872569	hypothetical protein Tc00.1047053508939.60
	Trypanosoma cruzi	gi|70875622	hypothetical protein Tc00.1047053508979.90
	Trypanosoma cruzi	gi|70883473	hypothetical protein Tc00.1047053506409.220
ORTHOMCL376	*Trypanosoma rangeli*	TGEG101013C01.g	Proteína Hipotética
	Trypanosoma rangeli	TGEG101013C01.g	Proteína Hipotética
	Leishmania major	gi|68125888	hypothetical protein, conserved
	Trypanosoma brucei	gi|70833246	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70885952	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70887240	hypothetical protein, conserved
ORTHOMCL4182	*Trypanosoma rangeli*	TGEG101043C12.b	DNA primase (subunidade menor)
	Leishmania major	gi|68224076	DNA primase small subunit, putative
	Trypanosoma brucei	gi|70802390	DNA primase small subunit, putative
	Trypanosoma cruzi	gi|70878555	DNA primase small subunit, putative
ORTHOMCL4177	*Trypanosoma rangeli*	TGEG102053A05.b	Proteína de transporte vesicular (CDC48 homólogo)
	Leishmania major	gi|70800011	vesicular transport protein (CDC48 homolog), putative
	Trypanosoma brucei	gi|70833641	vesicular transport protein (CDC48 homologue), putative
	Trypanosoma cruzi	gi|70881720	vesicular transport protein (CDC48 homologue), putative
ORTHOMCL687	*Trypanosoma rangeli*	TGEG101003H12.g	Proteína Hipotética
	Leishmania major	gi|68129547	hypothetical protein, conserved
	Trypanosoma brucei	gi|70800789	hypothetical protein, conserved

	Trypanosoma cruzi	gi|70871724	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70877562	hypothetical protein, conserved
ORTHOMCL6775	*Trypanosoma rangeli*	TGEG101013H11.g	Proteína Hipotética
	Leishmania major	gi|68128171	hypothetical protein, conserved
	Trypanosoma brucei	gi|70803656	hypothetical protein, conserved
ORTHOMCL4179	*Trypanosoma rangeli*	TGEG102051C07.b	UDP-glicose:glicoproteina glicosiltransferase
	Leishmania major	gi|70799752	hypothetical protein, conserved
	Trypanosoma brucei	gi|70800587	UDP-glucose:glycoprotein glucosyltransferase, putative
	Trypanosoma cruzi	gi|70886562	UDP-glucose:glycoprotein glucosyltransferase
ORTHOMCL7915	*Trypanosoma rangeli*	TGEG102023A05.g	Proteína de transporte vesicular (CDC48 homólogo)
	Trypanosoma cruzi	gi|70878144	vesicular transport protein (CDC48 homologue), putative
ORTHOMCL6769	*Trypanosoma rangeli*	TGEG102053A06.b	Ecotina
	Leishmania major	gi|68125416	ecotin, putative
	Trypanosoma brucei	gi|70801381	ecotin, putative
ORTHOMCL670	*Trypanosoma rangeli*	TGEG102052A04.b	Glicosiltransferase
	Leishmania major	gi|68125950	glycosyltransferase, putative
	Trypanosoma brucei	gi|70833190	glycosyltransferase, putative
	Trypanosoma cruzi	gi|70872446	glycosyltransferase, putative
	Trypanosoma cruzi	gi|70879573	glycosyltransferase, putative
ORTHOMCL681	*Trypanosoma rangeli*	TGEG101011G11.g	Proteína Hipotética
	Leishmania major	gi|68223745	hypothetical protein, conserved
	Trypanosoma brucei	gi|70802889	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70877208	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70879562	hypothetical protein, conserved
ORTHOMCL7913	*Trypanosoma rangeli*	TGEG102033B09.b	Proteína Hipotética
	Trypanosoma rangeli	TGEG102052A03.b	Proteína Hipotética
ORTHOMCL689	*Trypanosoma rangeli*	TGEG101001D06.g	Proteína Hipotética
	Leishmania major	gi|68125106	hypothetical protein, conserved
	Trypanosoma brucei	gi|70833886	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70878774	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70878868	hypothetical protein, conserved
ORTHOMCL6770	*Trypanosoma rangeli*	TGEG102051H08.b	Proteína Hipotética Conserveda
	Leishmania major	gi|68126243	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70884076	hypothetical protein, conserved
ORTHOMCL7927	*Trypanosoma rangeli*	TGEG101001D12.g	RNA helicase ATP-dependente
	Trypanosoma rangeli	TGEG102022F03.g	RNA helicase ATP-dependente
ORTHOMCL690	*Trypanosoma rangeli*	TGEG101001D01.g	Zinco metalopeptidase mitocondrial ATP-dependente
	Leishmania major	gi|68129273	metallo-peptidase, Clan MA(E), Family M41; mitochondrial ATP-dependent zinc metallopeptidase, putative
	Trypanosoma brucei	gi|70801013	mitochondrial ATP-dependent zinc metallopeptidase, putative
	Trypanosoma cruzi	gi|70876163	mitochondrial ATP-dependent zinc metallopeptidase, putative
	Trypanosoma cruzi	gi|70886390	mitochondrial ATP-dependent zinc metallopeptidase, putative
ORTHOMCL4180	*Trypanosoma rangeli*	TGEG102033E03.b	Proteína Hipotética
	Leishmania major	gi|68126779	hypothetical protein, conserved

	Trypanosoma brucei	gi|70803487	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70883159	hypothetical protein, conserved
ORTHOMCL6778	*Trypanosoma rangeli*	TGEG101003F02.g	Proteína Hipotética
	Trypanosoma cruzi	gi|70868948	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70872449	hypothetical protein, conserved
ORTHOMCL6776	*Trypanosoma rangeli*	TGEG101012E06.g	Proteína Hipotética
	Trypanosoma brucei	gi|70801383	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70875549	hypothetical protein, conserved
ORTHOMCL4185	*Trypanosoma rangeli*	TGEG101038G02.b	Proteína Hipotética
	Leishmania major	gi|68129424	hypothetical protein, conserved
	Trypanosoma brucei	gi|70800896	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70887204	hypothetical protein, conserved
ORTHOMCL58	*Trypanosoma rangeli*	TGEG101038E11.b	Serino peptidase Bem46-like
	Trypanosoma rangeli	TGEG101043B04.b	Serino peptidase Bem46-like
	Trypanosoma rangeli	TGEG101043D10.b	Serino peptidase Bem46-like
	Trypanosoma rangeli	TGEG101047H06.b	Serino peptidase Bem46-like
	Trypanosoma rangeli	TGEG102033G07.b	Serino peptidase Bem46-like
	Trypanosoma rangeli	TGEG102051B03.b	Serino peptidase Bem46-like
	Trypanosoma rangeli	TGEG102051D04.b	Serino peptidase Bem46-like
	Trypanosoma rangeli	TGEG102052E04.b	Serino peptidase Bem46-like
	Trypanosoma rangeli	TGEG102053A08.b	Serino peptidase Bem46-like
	Trypanosoma rangeli	TGEG102053B10.b	Serino peptidase Bem46-like
	Trypanosoma rangeli	TGEG102053E07.b	Serino peptidase Bem46-like
	Trypanosoma rangeli	TGEG102054A03.b	serino peptidase
	Trypanosoma brucei	gi|70831539	Bem46-like serine peptidase
	Trypanosoma cruzi	gi|70887002	Bem46-like serine peptidase, putative
	Leishmania major	gi|70905987	Serine peptidase, Clan SC, Family S09X
ORTHOMCL4191	*Trypanosoma rangeli*	TGEG101004D06.g	Proteína Hipotética
	Trypanosoma brucei	gi|70834139	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70866762	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70882149	hypothetical protein, conserved
ORTHOMCL4193	*Trypanosoma rangeli*	TGEG101003D07.g	Proteína Hipotética
	Leishmania major	gi|68129423	hypothetical protein, conserved
	Trypanosoma brucei	gi|70800897	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70887203	hypothetical protein, conserved
ORTHOMCL7925	*Trypanosoma rangeli*	TGEG101003C03.g	Proteína Hipotética
	Trypanosoma rangeli	TGEG101012G10.g	Proteína Hipotética
ORTHOMCL684	*Trypanosoma rangeli*	TGEG101008H11.g	Proteína Hipotética
	Leishmania major	gi|68223916	hypothetical protein, conserved
	Trypanosoma brucei	gi|70832057	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70878283	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70882086	hypothetical protein, conserved
ORTHOMCL672	*Trypanosoma rangeli*	TGEG101047D09.b	Proteína Hipotética
	Leishmania major	gi|68128581	hypothetical protein, conserved
	Trypanosoma brucei	gi|70834951	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70873269	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70882844	hypothetical protein, conserved
ORTHOMCL671	*Trypanosoma rangeli*	TGEG101048G1.b	Proteína Hipotética
	Leishmania major	gi|68128188	hypothetical protein, conserved
	Trypanosoma brucei	gi|70801133	protein phosphatase 2C, putative

	Trypanosoma cruzi	gi|70880172	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70882565	hypothetical protein, conserved
ORTHOMCL683	*Trypanosoma rangeli*	TGEG101010B05.g	Fator de liberação de cadeia 1
	Leishmania major	gi|68126969	peptide chain release factor-like protein
	Trypanosoma brucei	gi|70800233	peptide chain release factor 1, putative
	Trypanosoma cruzi	gi|70868167	peptide chain release factor 1, putative
	Trypanosoma cruzi	gi|70879071	peptide chain release factor 1, putative
ORTHOMCL191	*Trypanosoma rangeli*	TGEG101003C04.g	Proteína Hipotética, Conserved
	Trypanosoma rangeli	TGEG102003C01.g	Proteína Hipotética, Conserved
	Leishmania major	gi|68125358	hypothetical protein, conserved
	Trypanosoma brucei	gi|70802515	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70863705	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70867697	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70868355	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70874728	hypothetical protein, conserved
ORTHOMCL675	*Trypanosoma rangeli*	TGEG101038D09.b	Proteína Hipotética
	Leishmania major	gi|68223915	hypothetical protein, conserved
	Trypanosoma brucei	gi|70832058	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70878282	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70882085	hypothetical protein, conserved
ORTHOMCL4187	*Trypanosoma rangeli*	TGEG101011D11.g	Proteína de ligação ao RNA
	Leishmania major	gi|68124283	hypothetical protein, conserved
	Trypanosoma brucei	gi|70802966	RNA-binding protein, putative
	Trypanosoma cruzi	gi|70878759	hypothetical protein, conserved
ORTHOMCL6774	*Trypanosoma rangeli*	TGEG101014D05.g	Proteína Hipotética
	Trypanosoma brucei	gi|70832011	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70874637	hypothetical protein, conserved
ORTHOMCL676	*Trypanosoma rangeli*	TGEG101038G12.g	Proteína Hipotética
	Leishmania major	gi|68223845	hypothetical protein, conserved
	Trypanosoma brucei	gi|70832129	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70869171	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70870691	hypothetical protein, conserved
ORTHOMCL32	*Trypanosoma rangeli*	TGEG101011E10.g	GP63
	Trypanosoma rangeli	TGEG102052F04.b	GP63
	Trypanosoma cruzi	gi|70863199	surface protease GP63, putative
	Trypanosoma cruzi	gi|70863504	surface protease GP63, putative
	Trypanosoma cruzi	gi|70865500	surface protease GP63, putative
	Trypanosoma cruzi	gi|70866659	surface protease GP63, putative
	Trypanosoma cruzi	gi|70866802	surface protease GP63, putative
	Trypanosoma cruzi	gi|70867581	surface protease GP63, putative
	Trypanosoma cruzi	gi|70869009	surface protease GP63, putative
	Trypanosoma cruzi	gi|70869059	surface protease GP63, putative
	Trypanosoma cruzi	gi|70869060	surface protease GP63, putative
	Trypanosoma cruzi	gi|70869890	surface protease GP63, putative
	Trypanosoma cruzi	gi|70870106	surface protease GP63, putative
	Trypanosoma cruzi	gi|70871140	surface protease GP63, putative
	Trypanosoma cruzi	gi|70871182	surface protease GP63, putative
	Trypanosoma cruzi	gi|70873060	surface protease GP63, putative
	Trypanosoma cruzi	gi|70873286	surface protease GP63, putative
	Trypanosoma cruzi	gi|70874619	surface protease GP63, putative

	Trypanosoma cruzi	gi|70874680	surface protease GP63, putative
	Trypanosoma cruzi	gi|70874681	surface protease GP63, putative
	Trypanosoma cruzi	gi|70874748	surface protease GP63, putative
	Trypanosoma cruzi	gi|70874816	surface protease GP63, putative
	Trypanosoma cruzi	gi|70875419	surface protease GP63, putative
	Trypanosoma cruzi	gi|70875420	surface protease GP63, putative
	Trypanosoma cruzi	gi|70877460	surface protease GP63, putative
	Trypanosoma cruzi	gi|70877461	surface protease GP63, putative
	Trypanosoma cruzi	gi|70877852	surface protease GP63, putative
	Trypanosoma cruzi	gi|70882940	surface protease GP63, putative
ORTHOMCL6772	*Trypanosoma rangeli*	TGEG101038C09.b	Proteína Hipotética
	Trypanosoma cruzi	gi|70883828	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70885499	hypothetical protein, conserved
ORTHOMCL375	*Trypanosoma rangeli*	TGEG102054C12.b	Tubulin-tyrosine ligase-like protein
	Leishmania major	gi|68124805	tubulin-tyrsoine ligase-like protein
	Trypanosoma brucei	gi|70834140	tubulin-tyrsoine ligase-like protein, putative
	Trypanosoma cruzi	gi|70866322	tubulin-tyrsoine ligase-like protein, putative
	Trypanosoma cruzi	gi|70866761	tubulin-tyrosine ligase-like protein, putative
	Trypanosoma cruzi	gi|70882132	tubulin-tyrosine ligase-like protein, putative
ORTHOMCL10	*Trypanosoma rangeli*	TGEG101043G05.b	GP63
	Leishmania major	gi|68128257	GP63-like protein, leishmanolysin-like protein; metallo-peptidase, Clan MA(M), Family M8
	Trypanosoma cruzi	gi|70863010	surface protease GP63, putative
	Trypanosoma cruzi	gi|70864551	trans-sialidase, putative
	Trypanosoma cruzi	gi|70864604	surface protease GP63, putative
	Trypanosoma cruzi	gi|70865154	surface protease GP63, putative
	Trypanosoma cruzi	gi|70865982	surface protease GP63, putative
	Trypanosoma cruzi	gi|70866141	surface protease GP63, putative
	Trypanosoma cruzi	gi|70866270	trans-sialidase, putative
	Trypanosoma cruzi	gi|70866457	surface protease GP63, putative
	Trypanosoma cruzi	gi|70866517	surface protease GP63, putative
	Trypanosoma cruzi	gi|70866726	surface protease GP63, putative
	Trypanosoma cruzi	gi|70866729	surface protease GP63, putative
	Trypanosoma cruzi	gi|70867025	surface protease GP63, putative
	Trypanosoma cruzi	gi|70868514	surface protease GP63, putative
	Trypanosoma cruzi	gi|70869474	surface protease GP63, putative
	Trypanosoma cruzi	gi|70871138	surface protease GP63, putative
	Trypanosoma cruzi	gi|70871528	surface protease GP63, putative
	Trypanosoma cruzi	gi|70871884	surface protease GP63, putative
	Trypanosoma cruzi	gi|70872690	surface protease GP63, putative
	Trypanosoma cruzi	gi|70872716	surface protease GP63, putative
	Trypanosoma cruzi	gi|70872736	trans-sialidase, putative
	Trypanosoma cruzi	gi|70872832	surface protease GP63, putative
	Trypanosoma cruzi	gi|70873067	surface protease GP63, putative
	Trypanosoma cruzi	gi|70874024	surface protease GP63, putative
	Trypanosoma cruzi	gi|70874636	surface protease GP63, putative
	Trypanosoma cruzi	gi|70874879	surface protease GP63, putative
	Trypanosoma cruzi	gi|70875365	surface protease GP63, putative
	Trypanosoma cruzi	gi|70875399	trans-sialidase, putative
	Trypanosoma cruzi	gi|70875506	surface protease GP63, putative

	Trypanosoma cruzi	gi|70876068	surface protease GP63, putative
	Trypanosoma cruzi	gi|70876842	hypothetical protein Tc00.1047053503909.40
	Trypanosoma cruzi	gi|70876908	surface protease GP63, putative
	Trypanosoma cruzi	gi|70878398	surface protease GP63, putative
	Trypanosoma cruzi	gi|70878401	surface protease GP63, putative
	Trypanosoma cruzi	gi|70878805	surface protease GP63, putative
	Trypanosoma cruzi	gi|70880187	surface protease GP63, putative
	Trypanosoma cruzi	gi|70880517	surface protease GP63, putative
	Trypanosoma cruzi	gi|70880729	surface protease GP63, putative
	Trypanosoma cruzi	gi|70880892	surface protease GP63, putative
	Trypanosoma cruzi	gi|70880954	surface protease GP63, putative
	Trypanosoma cruzi	gi|70881426	surface protease GP63, putative
	Trypanosoma cruzi	gi|70883063	surface protease GP63, putative
	Trypanosoma cruzi	gi|70883626	surface protease GP63, putative
	Trypanosoma cruzi	gi|70883736	surface protease GP63, putative
	Trypanosoma cruzi	gi|70883780	surface protease GP63, putative
	Trypanosoma cruzi	gi|70883803	surface protease GP63, putative
	Trypanosoma cruzi	gi|70883806	surface protease GP63, putative
	Trypanosoma cruzi	gi|70883808	surface protease GP63, putative
	Trypanosoma cruzi	gi|70883970	surface protease GP63, putative
	Trypanosoma cruzi	gi|70885339	surface protease GP63, putative
	Trypanosoma cruzi	gi|70885514	surface protease GP63, putative
	Trypanosoma cruzi	gi|70885830	surface protease GP63, putative
	Trypanosoma cruzi	gi|70885839	surface protease GP63, putative
	Trypanosoma cruzi	gi|70885931	surface protease GP63, putative
	Trypanosoma cruzi	gi|70886685	surface protease GP63, putative
	Trypanosoma cruzi	gi|70886788	surface protease GP63, putative
	Trypanosoma cruzi	gi|70886801	surface protease GP63, putative
	Trypanosoma cruzi	gi|70886824	surface protease GP63, putative
ORTHOMCL4192	*Trypanosoma rangeli*	TGEG101004H10.g	Proteína Hipotética
	Leishmania major	gi|70800104	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70872245	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70881188	hypothetical protein, conserved
ORTHOMCL679	*Trypanosoma rangeli*	TGEG101012G08.g	Proteína Hipotética
	Trypanosoma rangeli	TGEG102053F08.b	Proteína Hipotética
	Leishmania major	gi|68127068	hypothetical protein, conserved
	Trypanosoma brucei	gi|70800375	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70874268	hypothetical protein, conserved
ORTHOMCL189	*Trypanosoma rangeli*	TGEG102052G11.b	trans-sialidase
	Trypanosoma brucei	gi|70801252	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70860135	trans-sialidase, putative
	Trypanosoma cruzi	gi|70868353	trans-sialidase, putative
	Trypanosoma cruzi	gi|70869039	trans-sialidase, putative
	Trypanosoma cruzi	gi|70873178	trans-sialidase, putative
	Trypanosoma cruzi	gi|70875649	trans-sialidase, putative
	Trypanosoma cruzi	gi|70883646	trans-sialidase, putative
ORTHOMCL6771	*Trypanosoma rangeli*	TGEG102023G04.g	Kinesina-like
	Trypanosoma cruzi	gi|70865969	myosin heavy chain, putative
	Trypanosoma cruzi	gi|70866515	hypothetical protein, conserved
ORTHOMCL4181	*Trypanosoma rangeli*	TGEG101043B09.b	Proteína Hipotética

	Trypanosoma brucei	gi|70801828	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70869621	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70878828	hypothetical protein, conserved
ORTHOMCL677	*Trypanosoma rangeli*	TGEG101016G08.g	Proteína Hipotética
	Leishmania major	gi|68130024	hypothetical protein, conserved
	Trypanosoma brucei	gi|70834430	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70876256	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70880903	hypothetical protein, conserved
ORTHOMCL7924	*Trypanosoma rangeli*	TGEG101004B07.g	RNA helicase ATP-dependente
	Trypanosoma rangeli	TGEG101011E03.g	RNA helicase ATP-dependente
ORTHOMCL674	*Trypanosoma rangeli*	TGEG101043E02.g	Proteína Hipotética
	Leishmania major	gi|68128154	hypothetical protein, conserved
	Trypanosoma brucei	gi|70833091	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70876454	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70878528	hypothetical protein, conserved
ORTHOMCL7918	*Trypanosoma rangeli*	TGEG101038G11.b	Proteína Hipotética
	Trypanosoma rangeli	TGEG102052B09.b	Proteína Hipotética
ORTHOMCL6777	*Trypanosoma rangeli*	TGEG101003G02.g	Proteína Hipotética
	Trypanosoma rangeli	TGEG101007C12.g	Proteína Hipotética
	Trypanosoma brucei	gi|70801770	hypothetical protein, conserved
ORTHOMCL678	*Trypanosoma rangeli*	TGEG101015E06.g	Tc40 antigeno
	Leishmania major	gi|68125306	immunodominant antigen, putative; tc40 antigen-like
	Trypanosoma brucei	gi|70802557	immunodominant antigen, putative
	Trypanosoma cruzi	gi|70873865	immunodominant antigen, putative
	Trypanosoma cruzi	gi|70874943	immunodominant antigen, putative
ORTHOMCL4186	*Trypanosoma rangeli*	TGEG101013C12.g	Proteína Hipotética
	Leishmania major	gi|68224075	hypothetical protein, conserved
	Trypanosoma brucei	gi|70802389	hypothetical protein, conserved
	Trypanosoma cruzi	gi|70878554	hypothetical protein, conserved
ORTHOMCL686	*Trypanosoma rangeli*	TGEG101003H06.g	RNA helicase ATP-dependente
	Trypanosoma rangeli	gi|10101017	ATP-dependent RNA helicase
	Trypanosoma brucei	gi|70831538	ATP-dependent DEAD/H RNA helicase, putative
	Trypanosoma cruzi	gi|70887001	ATP-dependent RNA helicase, putative
	Leishmania major	gi|70905988	ATP-dependent RNA helicase, putative
ORTHOMCL7920	*Trypanosoma rangeli*	TGEG101016A12.g	GP63
	Trypanosoma rangeli	TGEG101018H06.g	Proteína Hipotética
ORTHOMCL682	*Trypanosoma rangeli*	TGEG101011A12.g	proteina kinase
	Leishmania major	gi|68124126	protein kinase, putative
	Trypanosoma brucei	gi|70802681	protein kinase, putative
	Trypanosoma cruzi	gi|70885075	protein kinase, putative
	Trypanosoma cruzi	gi|70886333	protein kinase, putative
ORTHOMCL4184	*Trypanosoma rangeli*	TGEG101038D02.b	Proteína Hipotética
	Trypanosoma rangeli	TGEG101047C06.b	Proteína Hipotética
	Trypanosoma rangeli	TGEG102053F12.b	Proteína Hipotética
	Trypanosoma cruzi	gi|70868802	hypothetical protein Tc00.1047053503651.40