MILENE FERRO

**Desenvolvimento e validação de protocolos
para a anotação automática de seqüências
ORESTES de *Eimeria* spp.
de galinha doméstica**

Dissertação apresentada ao Instituto de
Ciências Biomédicas da Universidade de São
Paulo, para obtenção do Título de Mestre em
Ciências.

São Paulo

2008

MILENE FERRO

**Desenvolvimento e validação de protocolos
para a anotação automática de seqüências
ORESTES de *Eimeria* spp.
de galinha doméstica**

Dissertação apresentada ao Instituto de
Ciências Biomédicas da Universidade de São
Paulo, para obtenção do Título de Mestre em
Ciências.

Área de Concentração: Biologia da Relação
Patógeno-Hospedeiro.

Orientador: Prof. Dr. Arthur Gruber.

São Paulo

2008

DADOS DE CATALOGAÇÃO NA PUBLICAÇÃO (CIP)
Serviço de Biblioteca e Informação Biomédica do
Instituto de Ciências Biomédicas da Universidade de São Paulo

© reprodução total

Ferro, Milene

Desenvolvimento e validação de protocolos para a anotação automática de seqüências ORESTES de *Eimeria* spp. de galinha doméstica / Milene Ferro. -- São Paulo, 2008.

Orientador: Arthur Gruber.

Dissertação (Mestrado) – Universidade de São Paulo. Instituto de Ciências Biomédicas. Departamento de Parasitologia. Área de concentração: Biologia da Relação Patógeno-Hospedeiro. Linha de pesquisa: Biologia molecular de *Eimeria*.

Versão do título para o inglês: Development and validation of protocols for automated annotation of ORESTES sequences of *Eimeria* spp. of domestic fowl.

Descritores: 1. Anotação de seqüências biológicas 2. Bioinformática 3. Processo encadeado 4. Seqüência de DNA 5. Etiquetas de seqüências expressas 6. *Eimeria* spp. I. Gruber, Arthur II. Universidade de São Paulo. Instituto de Ciências Biomédicas. Programa de Pós Graduação em Biologia da Relação Patógeno-Hospedeiro III. Título.

ICB/SBIB178/2008

UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE CIÊNCIAS BIOMÉDICAS

Candidato(a): Milene Ferro.

Título da dissertação: Desenvolvimento e validação de protocolos para a
anotação automática de seqüências ORESTES de *Eimeria*
spp. de galinha doméstica.

Orientador(a): Arthur Gruber.

A Comissão Julgadora dos trabalhos de Defesa da **Dissertação de Mestrado**,
em sessão pública realizada a/...../.....,

☐ **Aprovado(a)**

☐ **Reprovado(a)**

Examinador(a): Assinatura:
Nome:
Instituição:

Examinador(a): Assinatura:
Nome:
Instituição:

Presidente: Assinatura:
Nome:
Instituição:

UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE CIÊNCIAS BIOMÉDICAS

Cidade Universitária "Armando de Salles Oliveira"
Av. Prof. Lineu Prestes, 2415 – CEP. 05508-000 São Paulo, SP – Brasil
Telefone : (55) (11) 3091-7733 - telefax : (55) (11) 3091-7438
e-mail: cep@icb.usp.br

Comissão de Ética em Pesquisa

CERTIFICADO DE ISENÇÃO

Certificamos que o Protocolo CEP-ICB N° 176, referente ao projeto intitulado “*Desenvolvimento e validação de protocolos de anotação automática de seqüências Orestes de Eimeria spp. de galinha doméstica*” sob a responsabilidade de **Milene Ferro**, foi analisado na presente data pela CEEA - COMISSÃO DE ÉTICA EM EXPERIMENTAÇÃO ANIMAL e pela CEPSh - COMISSÃO DE ÉTICA EM PESQUISA COM SERES HUMANOS, tendo sido deliberado que o referido projeto não envolve manipulação animal ou humana que justifique uma aprovação quanto aos princípios éticos exigidos por ambas as Comissões.

São Paulo, 17 de outubro de 2006.

PROF. DR. UBIRATAN FABRES MACHADO
Coordenador da CEEA - ICB/USP

PROF. DR. PAOLO M.A. ZANOTTO
Vice-Coordenador da CEPSh - ICB/USP

Dedico este trabalho aos meus familiares e esposo,
pelo incentivo e apoio em todas as etapas.

AGRADECIMENTOS

Ao meu orientador, Prof. Dr. Arthur Gruber, pelos ensinamentos teóricos e práticos ao longo de 5 anos de trabalho. De todos os anos ficaram conhecimentos valiosos.

Ao meu co-orientador, Prof. Dr. Alan Mitchell Durham, pela maneira fácil de ensinar conceitos computacionais de Linux e Perl.

A todos os membros atuais do laboratório, professora Alda, Jennifer, Ursula, Alessandra, Márcio e Samuel pela colaboração durante todo esse trabalho.

Ao estudante Ricardo Yamamoto Abe, em especial, pela paciência que teve durante toda a minha pós-graduação, me auxiliando tanto na teoria quanto na prática computacional.

À minha professora, Silvia, que ainda no ensino fundamental, me apresentou a Biologia e fez nascer em mim a paixão por essa área.

Aos professores Maria Doralice Cella e Maurício Bacci Junior, da Universidade Estadual Paulista (UNESP) de Rio Claro, que muito antes de iniciar os estudos da graduação, me concederam o privilégio de entrar em contato com o mundo maravilhoso da Bioinformática e da Biologia Celular e Molecular, através de aulas e estágio.

Aos professores de graduação do Mackenzie, em especial Profa. Maria Beatriz Riveron Acosta e Prof. Roberto Marcos Kalili, por acreditarem desde o início da graduação em meu potencial.

A cada um dos meus familiares, irmãos, tios, tias, primos e primas, por se preocuparem e por acompanharem mesmo que distante todo esse trajeto até o final.

De forma carinhosa e especial, a minha mãe Vera Ligia Godoy Ferro e minha avó materna Thereza Cazonatto Godoy, por serem responsáveis por toda minha formação dentro e fora das instituições de ensino. Agradeço por vocês existirem.

Ao meu marido Erik, por todo carinho, dedicação, respeito e atenção. Obrigada por me fazer sentir amada e por estar ao meu lado todos os dias. Obrigada por você existir.

Ao Instituto de Ciências Biomédicas da Universidade de São Paulo, pelo apoio e formação durante a pós-graduação.

À Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) pelo suporte financeiro e pelas bolsas de Iniciação Científica (2005 a 2006) e Mestrado (2006-2008).

RESUMO

Ferro M. Desenvolvimento e validação de protocolos para a anotação automática de seqüências ORESTES de *Eimeria* spp. de galinha doméstica [dissertação]. São Paulo: Instituto de Ciências Biomédicas da Universidade de São Paulo; 2008.

A coccidiose aviária é uma doença entérica causada por protozoários parasitas do gênero *Eimeria*. Visando uma maior compreensão dos mecanismos moleculares envolvidos na regulação do ciclo de vida dos parasitas, foram geradas 15.000 seqüências expressas (ORESTES) para cada uma das três espécies mais importantes: *E. tenella*, *E. maxima* e *E. acervulina*. Essas seqüências foram submetidas a um pipeline de pré-processamento utilizando o EGene, plataforma previamente desenvolvida pelo nosso grupo para a construção de processamentos encadeados (*pipelines*). No caso de *E. tenella*, utilizamos ainda um conjunto de seqüências de ESTs previamente depositadas em bancos internacionais. As leituras de cada espécie foram montadas com o programa CAP3 e as seqüências consenso resultantes foram usadas para uma anotação automática. Para a anotação, foram desenvolvidos 28 componentes para o sistema EGene, abrangendo a busca de ORFs, possíveis regiões codificadoras através de preditores de genes, genes de tRNA, regiões de repetições seriadas, busca de similaridade, mapeamento de cDNAs em genomas, motivos protéicos, regiões transmembranares, e sinais de peptídeo sinal e ancoragem por GPI. Também foram criados componentes específicos para a geração de relatórios de anotação nos formatos *feature table* e GFF3, bem como para a criação automática de páginas *web* contendo *links* para todas as evidências coletadas e para mapeamento de termos de ontologia gênica. A partir desses componentes, foram construídos *pipelines* de anotação automática das seqüências-consenso obtidas a partir da montagem dos ESTs/ORESTES de *Eimeria* spp. A anotação consistiu na identificação dos genes e atribuição da função dos respectivos produtos protéicos, baseando-se em um conjunto de evidências. As seqüências também foram classificadas e quantificadas utilizando-se um vocabulário controlado de termos de ontologia gênica (GO).

Palavras-chave: Anotação de seqüências biológicas. Bioinformática. Processamento encadeado. Seqüências de DNA. Etiquetas de seqüências expressas. *Eimeria* spp.

ABSTRACT

Ferro M. Development and validation of protocols for automated sequences annotation of ORESTES *Eimeria* spp. of domestic fowl [dissertation]. São Paulo: Instituto de Ciências Biomédicas da Universidade de São Paulo; 2008.

Avian coccidiosis is an enteric disease caused by protozoan parasites of the genus *Eimeria*. Aiming at obtaining a better understanding of the molecular mechanisms that regulate the life cycle of the parasites, our group generated 15,000 expressed sequences (ORESTES) for each one of the three most important species: *E. tenella*, *E. maxima* and *E. acervulina*. These reads were submitted to a pre-processing pipeline using EGene, a platform for pipeline construction previously described by our group. In the case of *E. tenella*, we also employed a set of conventional ESTs deposited by other groups on public databases. Reads from each *Eimeria* species were assembled with CAP3 and the resulting consensus sequences were used for the annotation process. We developed in total 28 annotation components for EGene system, comprising ORF finding, gene prediction, tRNA gene finding, tandem repeat finding, similarity searching, protein motif searching, and for the identification of transmembrane regions, signal peptide and GPI anchoring sequences. Also, we developed specific components for the automated generation of annotation reports on feature table and GFF3 formats, and web pages displaying GO term mapping and specific links to all collected evidences. Using these components, we constructed pipelines for the automated annotation of the *Eimeria* spp. EST/ORESTES assembled sequences. The annotation process consisted in the identification of genes and the corresponding protein function based on a set of evidences. The sequences were also mapped and quantified using a controlled vocabulary of gene ontology (GO) terms.

Keywords: Sequence annotation. Bioinformatics. Pipeline. DNA sequences. Expressed sequence tags. *Eimeria* spp.

LISTA DE FIGURAS

- Figura 1** - Ciclo de vida da *Eimeria* spp., o protozoário que causa a coccidiose aviária. Arte Gráfica: Helton Barreiro. 23
- Figura 2** - Esquema ilustrativo de um *pipeline*, onde cada ícone representa um programa que segue um fluxo de processamento pré-determinado pelo usuário. 32
- Figura 3** - Captura de tela do editor gráfico (CoEd) do sistema EGene, mostrando um exemplo de *pipeline*, com alguns componentes de processamento de seqüências. Os ícones representam os respectivos componentes do *pipeline*. 34
- Figura 4** - Tela do editor gráfico (CoEd) do sistema EGene mostrando o *pipeline* com vários componentes de anotação automática (vide ícones na tela) utilizados para processamento de seqüências de cDNA. 45
- Figura 5** - Exemplo de tela do programa Artemis, mostrando todas as ORFs (quadros horizontais em azul) e as respectivas anotações geradas após o processamento de um *pipeline* de anotação. Um total de oito ORFs podem ser visualizadas para esta seqüência de cDNA. 66
- Figura 6** - Exemplo de tela do programa Artemis, mostrando a anotação da seqüência mostrada na Figura 5, após o processamento e seleção da ORF pelo componente `report_orf.pl`. Notar que a seqüência agora apresenta uma única ORF (quadro horizontal em azul). O quadro de coloração azul clara refere-se a uma seqüência de uma etiqueta de SAGE alinhada por BLAST à seqüência do cDNA. 67
- Figura 7** - Exemplo de saída HTML gerada pelo componente `mapping_GO.pl` para um conjunto de seqüências de *Eimeria maxima*. 69
- Figura 8** - Exemplo de saída em formato HTML listando as seqüências positivas para um termo GO específico. Note que há *links* para as seqüências nucleotídicas e protéicas respectivamente. *Links* para outros termos GO associados a cada uma das seqüências também estão disponíveis. 70
- Figura 9** - Saída *web* no formato XML gerada pelo programa `mapping_GO.pl` para um conjunto de seqüências de *Eimeria maxima*. 71

- Figura 10** - Tela do editor gráfico (CoEd) do sistema EGene mostrando o *pipeline* com os componentes de anotação automática (vide ícones na tela) utilizados para processamento das seqüências de *Eimeria* spp. **80**
- Figura 11** - Uma anotação completa em formato *feature table* que inclui evidências do BLAST, RPS-BLAST (contra banco CDD) e InterProScan com os termos de GO. **85**
- Figura 12** - Tela capturada de interface *web* de uma página de anotação produzida automaticamente pelo componente `report_html.pl`. Os *links* para as anotações nos diferentes formatos, bem como para todas as evidências coletadas para cada ORF específica, estão mostrados. **94**

LISTA DE TABELAS

Tabela 1	- Descrição e características de todos os componentes de anotação gerados para o sistema EGene.	38
Tabela 2	- Características de origem, tipo de seqüências e quantidade de leituras utilizadas para pré-processamento para as seqüências de <i>E. tenella</i> .	41
Tabela 3	- Opções de código genético para diferentes organismos para o componente <code>annotation_orf.pl</code> .	50
Tabela 4	- Parâmetros de configuração do componente <code>annotation_orf.pl</code> para a busca e tradução conceitual de ORFs.	51
Tabela 5	- Parâmetros de configuração do componente <code>annotation_trna.pl</code> para a busca de genes de tRNA.	52
Tabela 6	- Parâmetros de configuração do componente <code>annotation_mreps.pl</code> para a busca de repetições seriadas.	53
Tabela 7	- Parâmetros de configuração do componente <code>annotation_string.pl</code> para a busca de repetições seriadas.	53
Tabela 8	- Parâmetros de configuração do componente <code>annotation_trf.pl</code> para a busca de repetições seriadas.	54
Tabela 9	- Parâmetros de configuração dos componentes preditores de genes.	54
Tabela 10	- Parâmetros de configuração do componente <code>annotation_sim4.pl</code> para o mapeamento de cDNA em DNA genômico.	55
Tabela 11	- Parâmetros de configuração do componente <code>annotation_exonerate.pl</code> para o mapeamento de cDNA em DNA genômico.	55
Tabela 12	- Parâmetros de configuração do componente <code>annotation_blast.pl</code> para busca de similaridade entre seqüências.	56
Tabela 13	- Opções do pacote <code>blastall</code> para buscas de similaridade.	56
Tabela 14	- Parâmetros de configuração do componente <code>annotation_hmmer.pl</code> para a busca de motivos protéicos contra a base de dados Pfam.	58
Tabela 15	- Parâmetros de configuração do componente <code>annotation_rpsblast.pl</code> para a busca de similaridade para proteínas.	58

Tabela 16 -	Parâmetros de configuração do componente <code>annotation_interpro.pl</code> .	59
Tabela 17 -	Parâmetros de configuração do componente <code>annotation_tmhmm.pl</code> .	60
Tabela 18 -	Parâmetros de configuração do componente <code>annotation_signalp.pl</code> .	61
Tabela 19 -	Parâmetros de configuração do componente <code>annotation_phobius.pl</code> .	61
Tabela 20 -	Parâmetros de configuração do componente <code>annotation_dgpi.pl</code> .	62
Tabela 21 -	Parâmetros de configuração do componente <code>report_orf.pl</code> .	63
Tabela 22 -	Caso 1: Resultados hipotéticos de uma anotação de uma sequência, mostrando as evidências e a escolha do programa <code>report_orf.pl</code> baseada em evidências.	64
Tabela 23 -	Caso 2: Resultados hipotéticos de uma anotação de uma sequência, mostrando as evidências e a escolha do programa <code>report_orf.pl</code> baseada em evidências.	64
Tabela 24 -	Parâmetros de configuração do componente <code>mapping_GO.pl</code> .	67
Tabela 25 -	Parâmetros de configuração dos componentes <code>report_feature_table_artemis.pl</code> e <code>report_feature_table_submission.pl</code> .	71
Tabela 26 -	Parâmetros de configuração do componente <code>report_gff.pl</code> .	72
Tabela 27 -	Parâmetros de configuração do componente <code>report_html.pl</code> .	73
Tabela 28 -	Pré-processamento das leituras ORESTES de <i>E. maxima</i> .	74
Tabela 29 -	Pré-processamento das leituras ORESTES de <i>E. acervulina</i> .	75
Tabela 30 -	Pré-processamento das leituras ORESTES de <i>E. tenella</i> .	75
Tabela 31 -	Dados comparativos de montagem de cDNAs utilizando-se duas versões do programa CAP3 e diferentes combinações dos parâmetros <code>p</code> e <code>y</code> .	76
Tabela 32 -	Resultados da montagem de transcritos das espécies de <i>Eimeria</i> com o programa CAP3.	77

Tabela 33 - Resultados de anotação de ORFs de <i>E. maxima</i> obtidos usando-se valores de 100, 75 e 50 para o parâmetro de tamanho mínimo da ORF (em resíduos de aminoácidos).	79
Tabela 34 - Frequências de termos de ontologia gênica de Processo Biológico das seqüências protéicas codificadas pelos cDNAs reconstruídos de <i>E. maxima</i> .	81
Tabela 35 - Frequência de termos de ontologia gênica de Componente Celular das seqüências protéicas codificadas pelos cDNAs reconstruídos de <i>E. maxima</i> .	82
Tabela 36 - Frequência de termos de ontologia gênica de Função Molecular das seqüências protéicas codificadas pelos cDNAs reconstruídos de <i>E. maxima</i> .	83
Tabela 37 - Frequência de termos de ontologia gênica de Processo Biológico das seqüências protéicas codificadas pelos cDNAs reconstruídos de <i>E. acervulina</i> .	86
Tabela 38 - Frequência de termos de ontologia gênica de Componente Celular das seqüências protéicas codificadas pelos cDNAs reconstruídos de <i>E. acervulina</i> .	87
Tabela 39 - Frequência de termos de ontologia gênica de Função Molecular das seqüências protéicas codificadas pelos cDNAs reconstruídos de <i>E. acervulina</i> .	88
Tabela 40 - Frequência de termos de ontologia gênica de Processo Biológico das seqüências protéicas codificadas pelos cDNAs reconstruídos de <i>E. tenella</i> .	89
Tabela 41 - Frequência de termos de ontologia gênica de Componente Celular das seqüências protéicas codificadas pelos cDNAs reconstruídos de <i>E. tenella</i> .	90
Tabela 42 - Frequência de termos de ontologia gênica de Função Molecular das seqüências protéicas codificadas pelos cDNAs reconstruídos de <i>E. tenella</i> .	91
Tabela 43 - Comparação das características de diferentes tipos de sistemas de anotação de seqüências em relação ao EGene.	99

LISTA DE ABREVIATURAS E SIGLAS

BLAST - *Basic Local Alignment Search Tool* (Ferramenta de busca de alinhamento local básica)

BBSRC - *Biotechnology and Biological Sciences Research* (Comitê de Pesquisa em Ciências Biotecnológicas e Biológicas - Reino Unido)

CAP3 - *Contig Assembly Program* (Programa de Montagem de Contigs)

DNA - ácido desoxirribonucleico

DAG - Grafo Acíclico Direto

cDNA - *complementary DNA* (DNA complementar)

DDBJ - *DNA Data Bank of Japan* (banco de dados de DNA do Japão)

EC - *Enzyme Commission* (Comissão de enzimas)

EMBL - *European Molecular Biology Laboratory* (Laboratório Europeu de Biologia Molecular)

EST - *Expressed Sequence Tags* (etiquetas de seqüências expressas)

FT - *feature table* (tabela de características)

G+C - guanina + timina

GFF - *Generic Feature Format* (Formato Genérico de Características)

GO - *Gene Ontology* (Ontologia Gênica)

HTML - *HyperText Markup Language* (Linguagem de Marcação de Hipertexto)

NCBI - *National Center for Biotechnology Information* (Centro Nacional de Informação Biotecnológica - EUA)

ORF - *Open Reading Frame* (fase aberta de leitura)

ORESTES - ORF ESTs (*Open Reading Frame Expressed Sequence Tags*)

PC - *personal computer* (Computador Pessoal)

PCR - *polymerase chain reaction* (reação em cadeia da polimerase)

RNA - ácido ribonucléico

RGB - *Red-Green-Blue* (Modelo aditivo de representar cores)

SAGE - *Serial Analysis of Gene Expression* (Análise Serial de Expressão Gênica)

spp. - espécie

SO - *Sequence Ontology* (Ontologia de Sequências)

tRNA - RNA transportador

TRF - *Tandem Repeats Finder* (achador de repetições seriadas)

URL - *Uniform Resource Locator* (Localizador Uniforme de Recursos)

USDA - *United States Department of Agriculture* (Departamento de Agricultura dos Estados Unidos)

XML - *eXtensible Markup Language* (Linguagem de Marcação Estendida)

LISTA DE SÍMBOLOS

°C - grau Celsius

GB – Gigabytes

pb - pares de bases

Mpb - mega pares de bases

SUMÁRIO

1 INTRODUÇÃO	22
1.1 Coccidiose aviária e <i>Eimeria</i> spp.....	22
1.2 Biologia dos parasitas.....	22
1.3 Características diferenciais das espécies	24
1.4 Controle da coccidiose aviária.....	25
1.5 Caracterização do genoma e seqüenciamento	26
1.6 Projetos ESTs e determinação do transcriptoma	27
1.7 Agrupamento de seqüências (<i>clustering</i>).....	28
1.8 Anotação de seqüências.....	28
1.9 Ontologia gênica	30
1.10 Sistemas de anotação automática.....	30
1.11 Sistemas de <i>pipelines</i>	31
1.12 EGene – sistema de construção de <i>pipelines</i> automáticos.....	32
 2 OBJETIVOS	 35
2.1 Objetivo geral.....	35
2.2 Objetivos específicos.....	35
 3 MATERIAIS E MÉTODOS	 37
3.1 Desenvolvimento de componentes de anotação automática	37
3.2 Conjuntos de dados	41
3.3 Pré-processamento das seqüências	42
3.4 Montagem das seqüências ESTs/ORESTES (<i>clustering</i>).....	44
3.5 Elaboração de protocolos e construção de <i>pipelines</i> de anotação	45
3.6 Execução dos <i>pipelines</i> de anotação	48
3.7 Curadoria manual da anotação automática.....	48
 4 RESULTADOS	 50
4.1 Desenvolvimento de componentes de anotação automática para o sistema EGene ...	50
4.1.1 Busca de ORFs (<i>Open Reading Frames</i>).....	50
4.1.2 Busca de genes de tRNA.....	51
4.1.3 Busca de regiões repetitivas seriadas com o programa <i>mreps</i>	52
4.1.4 Busca de regiões repetitivas seriadas com o programa String	53

4.1.5 Busca de regiões repetitivas seriadas usando TRF.....	53
4.1.6 Predição de genes.....	54
4.1.7 Mapeamento de cDNAs em genomas com os programas Sim4 e Exonerate	55
4.1.8 Busca de similaridade com o programa BLAST	56
4.1.9 Busca de motivos protéicos usando HMMER/Pfam	58
4.1.10 Busca de motivos protéicos com o programa RPS-BLAST.....	59
4.1.11 Busca de motivos protéicos com o programa InterProScan.....	59
4.1.12 Busca de domínios transmembranares com o programa TMHMM	61
4.1.13 Busca de peptídeo sinal com o programa SignalP	62
4.1.14 Busca de domínios transmembranares e peptídeo sinal com programa Phobius .	62
4.1.15 Busca de sítios de ancoragem GPI com o programa DGPI	63
4.1.16 Seleção de ORF	63
4.1.17 Mapeamento de termos de ontologia gênica (GO)	67
4.1.18 Geração de relatório no formato <i>Feature Table</i>	71
4.1.19 Geração de relatório no formato GFF3.....	72
4.1.20 Geração de páginas <i>web</i>	73
4.3 Pré-processamento das seqüências de <i>E. maxima</i> , <i>E. acervulina</i> e <i>E. tenella</i>	75
4.4 Reconstrução dos transcritos.....	77
4.5 Anotação automática das seqüências reconstruídas de cDNAs de <i>Eimeria</i> spp.	79
4.6. Construção de páginas <i>web</i>	93
 5 DISCUSSÃO	 96
5.1 Anotação automática.....	96
5.2 EGene e outras plataformas de anotação automática.....	98
5.3 Sistema EGene: características e perspectivas	103
5.4 Anotação de <i>Eimeria</i>	105
 6 CONCLUSÕES.....	 109
 REFERÊNCIAS BIBLIOGRÁFICAS	 111
 GLOSSÁRIO	 120

INTRODUÇÃO

1 INTRODUÇÃO

1.1 Coccidiose aviária e *Eimeria* spp.

A coccidiose aviária é uma doença entérica causada por parasitas protozoários da classe Coccidia, gênero *Eimeria*, e constitui-se em uma doença de alto impacto econômico devido aos grandes prejuízos gerados na indústria avícola. Calcula-se que os gastos mundiais com a doença, incluindo os prejuízos zootécnicos e a utilização de medicamentos com fim preventivo e curativo, cheguem a 800 milhões de dólares anuais (Allen e Fetterer, 2002).

Na galinha doméstica, a coccidiose atinge principalmente frangos de corte e matrizes reprodutoras, provocando enterite, redução na absorção de nutrientes, perda de peso, aumento de conversão alimentar e susceptibilidade a infecções por outros agentes (Kawazoe, 2000; Shirley *et al.*, 2005; Allen e Fetterer, 2002). Esse quadro faz com que a coccidiose seja uma das doenças infecciosas de maior relevância econômica na avicultura industrial, mesmo contando com um número considerável de drogas anticoccidianas e vacinas que podem ser utilizadas na sua prevenção (Shirley, 1997).

1.2 Biologia dos parasitas

Os parasitas protozoários do gênero *Eimeria* pertencem ao Filo Apicomplexa, Classe Coccidia. Parasitas deste filo caracterizam-se por apresentar um conjunto de organelas designado de complexo apical, que possui estruturas relacionadas com a penetração do parasita na célula do hospedeiro, além de estar envolvido com funções secretoras. Além dos protozoários do gênero *Eimeria*, o filo Apicomplexa possui também alguns dos mais conhecidos e importantes patógenos humanos, como *Plasmodium* spp., agente causador da malária, *Toxoplasma* spp., responsável pela toxoplasmose, *Cryptosporidium* spp. e *Cyclospora* spp., que causam enterites severas.

A coccidiose é um termo genérico que designa as infecções causadas pelos protozoários da classe Coccidia. No caso da galinha doméstica, podem ocorrer infecções por *Cryptosporidium* spp. e *Eimeria* spp., sendo este último gênero o de maior relevância em termos de patogenicidade e dos prejuízos econômicos que acarreta. Um total de sete espécies de *Eimeria* podem causar a coccidiose da galinha doméstica: *E. acervulina*, *E. maxima*, *E. tenella*, *E. necatrix*, *E. brunetti*, *E. praecox* e *E. mitis*. Os protozoários do gênero *Eimeria* apresentam um ciclo de vida direto (monoxênico), com vários ciclos intestinais endógenos envolvendo estágios assexuais que se multiplicam por fissão múltipla (merogonia ou

esquizogonia), seguidos de um ciclo sexual (gamogonia ou gametogonia) que resulta na formação de um oocisto não esporulado (Current *et al.*, 1990). Um esquema ilustrando o ciclo de vida da *Eimeria* spp. pode ser visualizado na Figura 1.

Figura 1 - Ciclo de vida da *Eimeria* spp., o protozoário que causa a coccidiose aviária. Arte Gráfica: Helton Barreiro.

Uma ave suscetível se infecta pela ingestão de um oocisto esporulado (Figura 1, etapa 1). No interior da moela o oocisto é triturado mecanicamente e sofre a ruptura física da sua parede. No intestino delgado, na presença de tripsina e sais biliares, ocorre o processo de excitação, no qual os esporozoítos saem ativamente dos esporocistos (Figura 1, etapa 2). Em seguida, os esporozoítos aderem e penetram nas células da mucosa intestinal (Figura 1, etapa 3), diferenciando-se em formas merontes ou esquizontes. Os esquizontes sofrem uma reprodução assexuada denominada merogonia ou esquizogonia (Figura 1, etapa 4), formando-se um conjunto de merozoítos. Quando o esquizonte amadurece, ocorre a ruptura das células do hospedeiro (Figura 1, etapa 5) e a liberação das formas merozoítas. Os merozoítos invadem então um novo conjunto de enterócitos (Figura 1, etapa 6), resultando em uma esquizogonia de segunda geração. Esse processo se repete até quatro vezes, dependendo da espécie, e, finalmente, ocorre uma nova diferenciação, gerando-se macrogametócitos (Figura

1, etapa 7) e microgametócitos (Figura 1, etapa 8), os quais produzem, respectivamente, os macro- e microgametas. Uma nova ruptura de células do hospedeiro libera os microgametas, que possuem um par de flagelos e são, portanto, móveis. Os microgametas fertilizam os macrogametas (Figura 1, etapa 9), resultando na formação de um zigoto, único estágio diplóide do parasita. O zigoto é coberto por camadas multilamelares que irão constituir a parede de oocisto e, uma vez maduro, rompe a célula do hospedeiro (Figura 1, etapa 10) e é liberado no ambiente através das fezes como oocisto não esporulado. No ambiente, sob condições adequadas de umidade, temperatura (entre 26-28 °C) e oxigenação, o oocisto sofre uma reprodução assexuada onde ocorre uma mitose seguida de meiose. Nesse processo, conhecido como esporulação ou esporogonia, forma-se um esporoblasto após a meiose (Figura 1, etapa 11) e, finalmente, um oocisto plenamente esporulado contendo no seu interior quatro esporocistos com dois esporozoítos em cada um (Figura 1, etapa 12). O oocisto esporulado é liberado com as fezes e representa o único estágio infectante do parasita. Esse oocisto poderá então infectar uma nova ave suscetível, fechando o ciclo de vida do parasita.

1.3 Características diferenciais das espécies

Os protozoários do gênero *Eimeria* causadores da coccidiose aviária, são parasitas espécie-específicos, ou seja, infectam apenas uma única espécie de hospedeiro. Entretanto, mais de uma espécie de *Eimeria* pode infectar um mesmo hospedeiro (Ruff, 1999). As sete espécies de *Eimeria* que infectam a galinha doméstica apresentam graus variados de patogenicidade, sendo *E. mitis* e *E. praecox* consideradas não patogênicas ou de baixa patogenicidade. Por outro lado, *E. acervulina* e *E. maxima* apresentam média patogenicidade, enquanto que *E. brunetti*, *E. necatrix* e *E. tenella*, são consideradas altamente patogênicas, podendo levar à morte das aves quando em alto grau de infecção. *E. tenella* apresenta a maior prevalência e virulência entre as sete espécies. Além disso, é a única espécie que permite o cultivo *in vitro*, sendo, portanto, a espécie mais estudada e utilizada como modelo para o entendimento da doença (Chapman e Shirley, 2003).

Na coccidiose aviária, cada espécie de *Eimeria* coloniza uma região específica do intestino, e causa lesões de características diferentes (Joyner e Long, 1984). Assim, os aspectos de cada lesão, bem como seus diferentes sítios de localização, como, por exemplo, intestino delgado para as espécies *E. maxima* e *E. acervulina*, e ceco para *E. tenella*, servem como parâmetro para a diferenciação das espécies. Outro fator importante na diferenciação é o formato e o tamanho dos oocistos. Além disso, é importante considerar também na diferenciação de espécies, o período de pré-patência, que varia entre 4 a 5 dias pós-infecção

(Allen e Fetterer, 2002), e o tempo mínimo de esporulação, que varia entre 15 a 30 horas nas diferentes espécies (Long e Reid, 1982).

O diagnóstico diferencial das espécies de *Eimeria* de galinha pode ser feito mediante a análise do conjunto de características biológicas descritas acima. Mais recentemente, tornou-se possível o diagnóstico molecular utilizando-se como alvos de amplificação a sequência ribossômica ITS1 (Schnitzler *et al.*, 1998, 1999) ou sequências anônimas de SCARs (Sequence-Characterized Amplified Regions), desenvolvidas pelo nosso grupo (Fernandez *et al.*, 2003).

1.4 Controle da coccidiose aviária

O controle da coccidiose aviária tem sido feito através de duas abordagens principais: o uso de drogas anticoccidianas na ração e a vacinação com parasitas vivos. Há uma série de drogas anticoccidianas disponíveis no mercado, que administradas em doses subótimas na ração, conseguem reduzir a carga parasitária e promover a formação de imunidade, prevenindo a ocorrência de surtos clínicos. Contudo, o uso continuado dessas drogas por longos períodos tem como consequência o surgimento esporádico de linhagens de parasitas resistentes, levando a surtos clínicos. Além disso, as agências regulatórias, especialmente na Europa, têm banido várias drogas ao longo dos anos devido ao risco potencial de seus resíduos permanecerem na carcaça, podendo, assim, exercer algum efeito tóxico no homem. Os limites máximos aceitáveis de resíduos também têm sido gradativamente reduzidos, tornando o uso de drogas cada vez mais limitado e de manejo mais complexo. Finalmente, alguns países estão prevendo o banimento completo do uso de drogas em rações animais dentro de alguns anos. Esse quadro tem desestimulado o desenvolvimento de novas drogas por parte da indústria farmacêutica, sendo que a semduramicina foi a última droga nova a ser introduzida no mercado, em 1995. Uma das alternativas encontradas para minimizar os efeitos da resistência, é o uso rotativo de vários anticoccidianos, combinando diferentes compostos químicos (Allen e Fetterer, 2002).

Outra alternativa para a prevenção da coccidiose aviária é o uso de vacinas, que podem conter cepas vivas virulentas em baixa dose, ou, alternativamente, cepas vivas atenuadas. No caso das vacinas vivas virulentas de baixa dose, ocorre o desenvolvimento de imunidade e de quadros sub-clínicos. Contudo, é importante ressaltar que pelo fato de não haver proteção cruzada entre as distintas espécies, essa abordagem requer o uso de vacinas multivalentes, encarecendo o produto. No caso das vacinas vivas atenuadas, utilizam-se cepas selecionadas para precocidade, isto é, cepas continuamente selecionadas para períodos de pré-

patência curtos. Essas cepas têm uma redução do número de esquizogonias, resultando em um potencial reprodutivo muito inferior e, portanto, em muito menor desenvolvimento de lesões na mucosa intestinal. Como, entretanto, ocorre uma infecção propriamente dita, desenvolve-se uma imunidade adequada. A desvantagem desse tipo de vacina é que como o potencial reprodutivo é muito menor (cerca de 5% em comparação com linhagens selvagens), o custo de produção para o fabricante é muito maior, resultando em um custo final para o produtor de aves bastante superior. Existem ainda abordagens de controle da coccidiose que preconizam o uso de vacinas por alguns períodos, intercalados pelo uso de drogas anticoccidianas (Chapman *et al.*, 2002). O princípio por trás dessa abordagem é que as drogas seriam substituídas após algum tempo de uso, quando linhagens resistentes poderiam estar surgindo. O uso então de vacinas vivas causaria uma repopulação das granjas com linhagens vacinais suscetíveis a drogas. No período seguinte de uso de drogas, as linhagens predominantes de *Eimeria* seriam novamente sensíveis aos anticoccidianos.

1.5 Caracterização do genoma e seqüenciamento

Em relação à caracterização do genoma, Cornelissen *et al.* (1984) estudaram a ploidia e conteúdo de DNA de *E. tenella*, observando que vários estágios apresentam complexidade genômica ao redor de $6,7 \times 10^7$ pb, e que todos os estágios são haplóides, exceto o zigoto. O material genômico haplóide está organizado em 14 cromossomos que variam de tamanho entre 1,0 a 6,0 Mpb (Shirley *et al.*, 1990; Shirley, 1994), e o conteúdo G+C é de cerca de 53% (Shirley, 2000). O cromossomo 1 de *E. tenella* foi totalmente seqüenciado (Ling *et al.*, 2006) e as análises de conteúdo e organização mostraram que o cromossomo tem cerca de 49,7% de conteúdo A+T e há longas repetições da trinca CAG e do heptâmero AGGGTTT, sendo que esse heptâmero tem sido identificado com uma unidade repetitiva telomérica em *Plasmodium*.

O genoma de *E. tenella* da cepa H também foi seqüenciado com uma cobertura de cinco vezes, através de leituras do tipo *shotgun* (Chapman e Shirley, 2003). Esta iniciativa foi patrocinada pelo BBSRC (*Biotechnology and Biological Sciences Research*) e realizada pelo *Wellcome Trust Sanger Institute*, Cambridge, Inglaterra, em colaboração com o *Institute for Animal Health*, Compton, Inglaterra. Este projeto foi inserido dentro das atividades do Consórcio Internacional do Genoma de *Eimeria tenella*, do qual o nosso grupo também faz parte (http://www.sanger.ac.uk/Projects/E_tenella/consortium.shtml; Shirley *et al.*, 2004).

Além do genoma nuclear, há uma única mitocôndria de formato tubular contendo um genoma com unidades repetitivas de cerca de 6 kb (Romano *et al.* – dados não publicados). Estudos em *E. tenella* mostraram também a presença de uma segunda organela com genoma,

o apicoplasto que se localiza próximo ao núcleo e possui quatro membranas concêntricas. O genoma dessa organela contém cerca de 35 kb e caracteriza-se por ser muito rico em conteúdo A+T (Cai *et al.*, 2003).

1.6 Projetos ESTs e determinação do transcriptoma

Os ESTs (*Expressed Sequence Tags*) são seqüências curtas, com média de tamanho entre 100-700 bp, obtidas a partir de bibliotecas de cDNA de um tecido ou de um estágio evolutivo de um organismo, e que permitem a caracterização de um subconjunto de genes expressos (Adams *et al.*, 1991). Vários projetos de seqüenciamento de ESTs têm sido feitos em parasitas como *Leishmania*, *Schistosoma* e *Trypanosoma cruzi* (Porcel *et al.*, 2000). Dentre os membros do filo Apicomplexa, foram feitos projetos de ESTs em *Plasmodium*, *Toxoplasma gondii* (Ajioka, 1998; Ajioka *et al.*, 1998) e *Eimeria tenella* (Wan *et al.*, 1999; Shirley, 2000). Além do seqüenciamento do genoma, até o momento, foram depositadas no Genbank cerca de 39.000 seqüências ESTs, derivadas principalmente de esporozoítos e merozoítos de segunda geração de *E. tenella*. Além da Washington University/Merck, que produziram a maioria destas ESTs, algumas ESTs derivados também foram descritos por Wan *et al.*, 1999 e Ng *et al.*, 2002. Dos 500 ESTs seqüenciados por Wan *et al.* (1999), 47,7% apresentaram similaridade com seqüências depositadas em banco de dados, incluindo proteínas ribossômicas, enzimas metabólicas e proteínas com outras funções. Dentre estes 47,7% ESTs, 14,3% representaram genes já conhecidos de *E. tenella*, os 52,3% restantes não apresentaram nenhuma similaridade com os bancos de dados existentes (Wan *et al.*, 1999).

A determinação da seqüência dos genes expressos, além de permitir conhecer o transcriptoma dos parasitas, pode também proporcionar grandes avanços no desenvolvimento de novas estratégias de controle do parasita, seja através do desenho de drogas anticoccidianas mais eficientes, dirigidas contra alvos moleculares específicos, seja na obtenção de vacinas de proteínas recombinantes, ou ainda vacinas de DNA. O seqüenciamento de ESTs em diferentes estágios de desenvolvimento é uma abordagem importante para se obter perfis de expressão de genes, e para a identificação de genes regulados diferencialmente (Manger *et al.*, 1998). A técnica ORESTES (ORF ESTs) é uma metodologia alternativa para a síntese de cDNA, envolvendo uma amplificação por PCR de baixa especificidade, com *primers* arbitrários (Dias Neto *et al.*, 2000). Como consequência, esta técnica resulta numa maior cobertura na região central da seqüência dos mRNAs, a qual contém a maior parte da porção codificadora dos genes. Assim, esta técnica é muito útil para projetos de seqüenciamento cujo objetivo maior seja o de realizar um levantamento dos genes transcritos.

A técnica de geração de seqüências ORESTES tem sido amplamente utilizada em diferentes projetos de transcriptomas. Como exemplos, podemos citar o projeto transcriptoma de *Trypanosoma rangeli* (Snoeijer *et al.*, 2004), abelha (Nunes *et al.*, 2004), *Schistosoma mansoni* (Verjovski-Almeida *et al.*, 2003), tecidos de câncer humano (de Souza *et al.*, 2000; Camargo *et al.*, 2001), entre outros. O que podemos considerar, é que ambas as técnicas se complementam e são importantes para o mapeamento de genes ativos em diferentes estágios de desenvolvimento de um organismo. O nosso grupo completou recentemente o seqüenciamento de ORESTES das três espécies mais importantes de *Eimeria*: *E. tenella*, *E. acervulina* e *E. maxima*, tendo sido obtidas cerca de 15.000 leituras para cada uma delas (<http://www.coccidia.icb.usp.br/eimeria/>).

1.7 Agrupamento de seqüências (*clustering*)

A informação contida em múltiplos ESTs apresenta um caráter fragmentado e redundante. Assim, visando organizar melhor a informação potencialmente gerada pelos ESTs, preconiza-se agrupar as seqüências que compartilham trechos em comum (*clusters*) e posteriormente proceder à montagem e conseqüente obtenção das seqüências consenso. Esse processo de agrupamento de ESTs e posterior montagem denomina-se reconstrução de transcritos (Gruber, 2007). Há abordagens nas quais as etapas de agrupamento (*clustering*) e montagem são feitas de forma separada, como no pacote StackPACK (Miller *et al.*, 1999; Burke *et al.*, 1999) e no TIGR Gene Indices Clustering Tool (TGICL - Pertea *et al.*, 2003; Lee *et al.*, 2005). Nesses casos, o agrupamento envolve uma comparação “todos-contra-todos” sob baixa estringência, seguida de uma montagem na qual as leituras com sobreposição de cada grupo são montadas de forma a gerar uma seqüência consenso. A vantagem de se fazer um agrupamento prévio das seqüências com um algoritmo rápido, é que se economiza tempo de processamento na etapa de montagem, a qual é lenta e complexa. Contudo, em conjuntos de dados da ordem de dezenas de milhares de ESTs, a montagem pode ser feita diretamente, utilizando-se programas como CAP3 (Huang e Madan, 1999) ou Phrap (Green, 1996).

1.8 Anotação de seqüências

A anotação pode ser compreendida como características que descrevem e qualificam uma seqüência biológica. A anotação de seqüências, segundo Stein (2001), consiste em um processo múltiplo, pelo qual uma ou mais seqüências brutas de DNA ou de aminoácidos são analisadas com a finalidade de se atribuir características, contextualizando-se estas seqüências

do ponto de vista biológico conforme suas funções. Assim, na anotação de seqüências acrescentam-se informações à seqüência nucleotídica, tais como localização de genes, presença de regiões repetitivas, identificação de genes de tRNA, rRNA e codificadores de proteínas, etc. Algumas dessas descrições podem ser deduzidas a partir de dados de similaridade com seqüências biológicas cuja função já seja conhecida. Embora a anotação geralmente esteja relacionada a seqüências genômicas, qualquer seqüência biológica pode ser anotada, incluindo seqüências de cDNAs e proteínas (Eibeck *et al.*, 2005; Berriman e Harris, 2004).

Para que as seqüências anotadas possam ser submetidas aos bancos de dados internacionais DDBJ/EMBL/GenBank, precisam estar no formato denominado *feature table*, cuja especificação encontra-se disponível para consulta em <http://www.ncbi.nlm.nih.gov/colab/FT/>. O formato *feature table* caracteriza-se por apresentar uma chave principal (*Feature Key*) que contém a localização (*Location*) e os qualificadores (*Qualifiers*) específicos que descrevem características da seqüência que está sendo anotada. Segundo a definição, a *Feature Key* refere-se a uma palavra única ou abreviação indicando um grupo funcional, já em *Location* têm-se as instruções de localização da *feature* (coordenadas) e os *qualifiers* trazem informações auxiliares sobre a *feature*. Um exemplo prático pode ser visualizado abaixo:

Key	Location/Qualifiers
CDS	23..400 /product="alcohol dehydrogenase" /gene="adhI"

Neste exemplo, o *Feature Key* CDS apresenta uma seqüência codificadora da proteína “alcohol dehydrogenase”, começando na coordenada 23 e terminado na coordenada 400 da seqüência nucleotídica, sendo que o nome do gene que codifica essa proteína é o “adhI”.

Outro formato de anotação de seqüências amplamente utilizado é o GFF3 (*Generic Feature Format*). O GFF3 é um formato de intercâmbio de anotações que consiste em um arquivo texto contendo nove colunas separadas entre si por tabulações (Stein, 2007). Tanto anotações no formato *feature table* quanto no GFF3, podem ser visualizadas e editadas através do programa Artemis (Rutherford *et al.*, 2000).

1.9 Ontologia gênica

O consórcio de Ontologia Gênica (GO) foi criado para desenvolver ontologias de genes, isto é, conjuntos de vocabulários controlados, dinâmicos e padronizados, contendo termos e suas interrelações. Três ontologias foram estabelecidas: processo biológico, componente celular e função molecular (Eibeck e Lewis, 2004). O uso de um vocabulário comum de termos é importante, pois genes de diferentes espécies podem ser comparados baseados nas anotações de GO. As ontologias estão representadas como grafos acíclicos diretos (*DAGs*), e os termos GO apresentam relacionamentos dos tipos *is a* (é um) ou *part of* (é parte de). É possível realizar classificações funcionais de seqüências anotadas utilizando-se subconjuntos de ontologias, denominados *GO Slim*, permitindo assim uma categorização e quantificação rápida e eficiente. Há uma série de modelos disponíveis de arquivos *GO Slim* para vários organismos (ftp://ftp.geneontology.org/pub/go/GO_slims). Cada entrada de GO apresenta um único identificador numérico com o formato GO:nnnnnnnn (onde n corresponde a um número de 0 a 9), seguido de um nome do termo como, por exemplo, GO:0006096 : *glycolysis*, que está inserido na ontologia *biological process*. O GO permite ainda um mapeamento direto com outras bases de dados, como por exemplo, o Swiss-Prot, Interpro e EC (*Enzyme Comission*). Por outro lado, o GO não cobre nomes de genes e seus produtos, processos únicos de organismos mutantes ou alterados (ex. Câncer), interações proteína-proteína, aspectos anatômicos ou histológicos, estrutura dos genes em termos de introns e éxons, etc.

A Ontologia de Seqüências (SO), por sua vez, utiliza o mesmo conceito de ontologia aplicado à descrição de seqüências de DNA. Atualmente, os termos controlados de SO podem ser utilizados para se gerar uma anotação no formato GFF3. Essa descrição das seqüências (anotação) pode ser feita no formato GFF3, sendo que a diferença entre a descrição usando *Feature Table* (FT) e SO é que esta última especifica sub-classes, ou seja, relacionamentos que existem entre os termos, enquanto que no FT isso não ocorre (Eilbeck *et al.*, 2004; Eilbeck e Lewis, 2004).

1.10 Sistemas de anotação automática

O processo de anotação de grandes projetos de seqüenciamento, dado o grande volume de informações, é geralmente feito de forma automatizada. Contudo, anotações mais confiáveis ainda requerem uma análise manual, feita em cima dos dados gerados automaticamente. Anotações que foram inspecionadas e editadas manualmente são ditas

curadas, pois sofrem uma curadoria. Nesse processo, a atribuição de função é particularmente importante, assim como a designação do possível produto. O processo manual de atribuição de função é baseado na coleta e análise de evidências. Assim, para que um produto protéico seja identificado como tendo uma determinada função, vários tipos de evidências podem ser levadas em conta. Entre essas evidências, pode-se utilizar, por exemplo, a similaridade com proteínas de função conhecida, a presença de motivos protéicos relacionados com essa função, e algumas outras características que eventualmente podem sugerir localização celular como regiões transmembranares, peptídeo sinal, sítio de ancoragem GPI, etc.

Atualmente, existe uma série de plataformas para a anotação de seqüências, dentre elas podemos citar Genescript (Hudek *et al.*, 2003), GARSa (Dávila *et al.*, 2005), annot8r (Schimid e Blaxter, 2008), EST Express (Smith *et al.*, 2008) e ESTExplorer (Nagaraj *et al.*, 2007). Embora essas ferramentas sejam poderosos sistemas para a anotação automática de seqüências, apresentam uma série de limitações, como a funcionalidade específica para seqüências ESTs ou para seqüências genômicas, a falta de modularidade tornando mais complexa a construção e manipulação de *pipelines*, além de um conjunto de módulos para a anotação bastante restrito, sendo que o desenvolvimento de novos componentes exige um grande conhecimento computacional. Todas as ferramentas descritas funcionam como coletoras de evidências, mais do que como anotadores propriamente ditos, pois os resultados obtidos para diferentes componentes apresentam-se de maneira isolada, não estando integrados em relatórios de anotação de seqüências. É importante ressaltar, ainda, que nem todos os sistemas de anotação automática de seqüências, desenvolvidos em grandes centros de seqüenciamento, são públicos. Além disso, alguns são de difícil adequação a laboratórios de pequeno porte e possuem ainda componentes específicos para o organismo estudado.

1.11 Sistemas de *pipelines*

Com o crescente aumento da quantidade de seqüências biológicas geradas, houve a necessidade de se construir sistemas computacionais que permitissem o processamento de grandes lotes de seqüências. Cada processamento é executado por um componente, em um fluxo contínuo e ordenado, onde os dados de saída de um componente servem como entrada para o processamento do próximo. Esse tipo de estrutura computacional recebe o nome de *pipeline*. Não há uma única palavra na língua portuguesa que exprima o sentido computacional de *pipeline*, mas poderíamos dizer que se trata de um fluxo encadeado de processos. Os sistemas de *pipelines* assemelham-se a linhas de montagem, como, por exemplo, de automóveis, onde em cada etapa de produção, acrescenta-se uma nova peça ou

acessório até a etapa final da montagem onde o carro estará pronto para ser distribuído. A Figura 2 mostra um modelo ilustrativo do que seria um *pipeline*.

Figura 2 - Esquema ilustrativo de um *pipeline*, onde cada ícone representa um programa que segue um fluxo de processamento pré-determinado pelo usuário.

Com sistemas de *pipelines* é possível processar de forma eficiente uma quantidade grande de seqüências, seja para o processamento pré-anotação ou para a anotação propriamente dita. Os resultados obtidos podem estar integrados, de forma a facilitar a visualização e a análise de dados gerados em cada programa do *pipeline*.

1.12 EGene – sistema de construção de *pipelines* automáticos

A plataforma EGene (Durham *et al.*, 2005) foi desenvolvida pelo nosso grupo e se caracteriza por ser um sistema integrado e customizável para a construção de *pipelines*. O sistema EGene permite encadear uma série de componentes diferentes de processamento, em uma ordem e composição totalmente definidas pelo usuário. Os componentes funcionam como módulos que tanto podem exercer uma função como um programa independente (*standalone*), como também podem funcionar como uma “casca” ou *shell*, interagindo com programas de terceiros. Novos componentes podem ser facilmente criados, pois o EGene provê um padrão simples para a geração de componentes, tornando essa tarefa fácil, mesmo para programadores iniciantes. O EGene aceita vários formatos de entrada de seqüências, como FASTA, PHD (arquivo de gerado pelo programa Phred - Ewing e Green, 1998; Ewing *et al.*, 1998), cromatogramas e XML. O formato básico de saída do sistema é o XML, mas o EGene permite salvar relatórios em uma série de outros formatos como XML, FASTA e PHD. Como os componentes utilizam um único modelo de dados para a entrada e saída, o

usuário pode interconectá-los livremente, sem a necessidade de possuir conhecimentos de programação. O sistema pode, ainda, ser facilmente acoplado a um banco de dados relacional. Assim, pode atender as necessidades tanto de usuários avançados, como também usuários de pequenos laboratórios. Além disso, ele é totalmente genérico, ao contrário da maioria dos sistemas de anotação existentes, permitindo seu uso em qualquer projeto de pequena ou larga escala, de seqüências de ESTs e genomas.

No sistema EGene, os *pipelines* processam seqüências biológicas como DNA, RNA ou proteínas, e quem assegura a abstração dessa representação no sistema é o componente denominado `SequenceObject.pm`, que é um módulo em Perl que reporta todas as manipulações realizadas com a seqüência e resultados obtidos após processamento. Tudo o que é processado pelo *pipeline*, inclusive a forma de processamento, é relatado por esse componente. Além disso, esse módulo facilita a construção de componentes para o sistema EGene, pois seguem uma estrutura definida.

O EGene é fornecido com o CoEd (Figura 3), um editor gráfico de configuração escrito em Java, que facilita a visualização da tarefa de construção de *pipelines*. Esse editor, além de permitir a visualização dos dados, também auxilia o usuário a configurar cada um dos componentes do *pipeline*, uma vez que os nomes de parâmetros, valores `default` e os argumentos obrigatórios são indicados pela ferramenta através de janelas contendo formulários específicos.

O sistema EGene foi originalmente concebido para processamentos de pré-anotação de seqüências, o que inclui avaliação de qualidade, mascaramento de vetores, aparamento de pontas, filtragem de contaminantes e montagem de seqüências, entre outras funções. Visando tornar o EGene uma plataforma robusta, tanto para pré-processamento como para a anotação automática de seqüências, decidimos desenvolver uma série de componentes de anotação automática de seqüências, os quais serão descritos neste trabalho. Também descreveremos a elaboração de protocolos de anotação de seqüências através da construção de *pipelines*, e a posterior validação de todo o sistema utilizando seqüências ESTs e ORESTES de três espécies de *Eimeria* spp.

Figura 3 - Captura de tela do editor gráfico (CoEd) do sistema EGene, mostrando um exemplo de *pipeline*, com alguns componentes de processamento de seqüências. Os ícones representam os respectivos componentes do *pipeline*.

OBJETIVOS

2 OBJETIVOS

2.1 Objetivo geral

- Desenvolver um conjunto de componentes para a plataforma EGene, visando a elaboração de protocolos de anotação automática e sua validação com seqüências de cDNAs de *Eimeria* spp.

2.2 Objetivos específicos

- Desenvolver um conjunto de componentes para o sistema EGene de geração de *pipelines* que permitam a anotação automática de seqüências biológicas;
- Realizar o pré-processamento de seqüências de cDNAs de *E. maxima*, *E. acervulina* e *E. tenella* e posterior reconstrução dos transcritos;
- Elaborar protocolos de anotação automática e implementá-los através da construção de *pipelines* do sistema EGene;
- Aplicar e validar os protocolos de anotação automática para os cDNAs de *E. máxima*, *E. acervulina* e *E. tenella*;
- Realizar a curadoria manual da anotação automática de genes e utilizar os resultados para melhorar os protocolos de anotação automática;
- Publicar os resultados da anotação automática na *web*, para consulta e análise dos resultados obtidos.

MATERIAIS E MÉTODOS

3 MATERIAIS E MÉTODOS

3.1 Desenvolvimento de componentes de anotação automática

Foram desenvolvidos 28 componentes de anotação automática de seqüências, conforme apresentado na Tabela 1. Deste total, 22 componentes estão acoplados a programas de terceiros, e funcionam como um “shell” ou casca. Os demais componentes não dependem de programas externos e, portanto, rodam no sistema EGene de forma autônoma. O desenvolvimento dos componentes que utilizam programas de terceiros consistiu nos seguintes passos: (a) pesquisa bibliográfica sobre o funcionamento, parâmetros e arquivos de entrada e saída de cada programa de terceiro; (b) instalação e execução do programa no servidor local; (c) definição dos parâmetros obrigatórios e default do programa; (d) desenvolvimento do componente do EGene; (e) teste de execução do componente sob diferentes condições de seqüências e parâmetros; (f) análise manual dos resultados gerados. Todos os componentes foram desenvolvidos utilizando-se a linguagem de programação Perl, e executados em servidores PC Dual Xeon HT 2,4 GHz, com 3GB de memória RAM, rodando o sistema operacional Linux da distribuição Debian e, posteriormente, Ubuntu Server versão 7.10.

Em paralelo ao desenvolvimento dos componentes de anotação automática, o modelo de dados do sistema EGene foi incrementado para poder lidar com as informações adicionais compreendidas pelas anotações. Para isso, o componente denominado *SequenceObject.pm*, foi alterado e adicionado de todas as funções necessárias para o armazenamento e processamento das anotações. Esse trabalho foi desenvolvido por Ricardo Yamamoto Abe, estudante de graduação do Instituto de Matemática e Estatística (IME) da USP, com a orientação do Prof. Alan M. Durham, do Departamento de Computação do IME.

Conforme pode ser visto na Tabela 1, foram desenvolvidos componentes para as principais funções de anotação, incluindo a busca e tradução de fases abertas de leitura (ORFs), predição de genes, busca de genes de tRNA, busca de similaridade, busca de repetições seriadas e busca de motivos protéicos, entre outras. Além disso, foram também desenvolvidos componentes para a geração de relatórios de anotação em formatos *feature table* e GFF3, além de um componente para mapeamento de quantificação de seqüências por termos de ontologia gênica (GO) e de geração automática de páginas *web* para apresentação dos resultados.

Tabela 1 - Descrição e características de todos os componentes de anotação gerados para o sistema EGene.

Componente	Programa associado	Função	Referência(s)
Busca de ORFs			
annotation_orf.pl	Standalone	Busca de fases de leitura abertas e sua tradução conceitual em seqüências protéicas	-
Busca de genes de tRNA			
annotation_trna.pl	tRNAScan-SE	Busca de genes de tRNA usando modelos probabilísticos	Lowe e Eddy, 1997
Busca de regiões repetitivas seriadas			
annotation_mreps.pl	MREPS	Busca <i>ab initio</i> de regiões repetitivas seriadas	Kolpakov <i>et al.</i> , 2003
annotation_trf.pl	TRF	Busca <i>ab initio</i> de regiões repetitivas seriadas	Benson, 1999
annotation_string.pl	String	Busca <i>ab initio</i> de regiões repetitivas seriadas	Parisi <i>et al.</i> , 2003
Predição de genes			
annotation_estscan.pl	ESTScan	Predição de genes para cDNAs	Iseli <i>et al.</i> , 1999
annotation_genscan.pl	GENSCAN	Predição de genes para genomas	Burge e Karlin, 1997
annotation_glimmer.pl	GlimmerM	Predição de genes para genomas	Pertea e Salzberg, 2002
annotation_glimmerhmm.pl	GlimmerHMM	Predição de genes para genomas	Majoros <i>et al.</i> , 2004
annotation_snap.pl	SNAP	Predição de genes para genomas	Korf, 2004
annotation_phat.pl	Phat	Predição de genes para genomas	Cawley <i>et al.</i> , 2001
annotation_twinscan.pl	Twinscan	Predição de genes para genomas	Korf <i>et al.</i> , 2001

(continua)

Tabela 1 (continuação)

Componente	Programa associado	Função	Referência(s)
Mapeamento de cDNA em genoma			
annotation_sim4.pl	Sim4	Mapeamento de cDNAs em seqüências genômicas usando o modelo éxon-íntron	Florea <i>et al.</i> , 1998
annotation_exonerate.pl	Exonerate	Alinhamento pareado de seqüências, incluindo o uso do modelo éxon-íntron	Slater e Birney, 2005
Busca de similaridade			
annotation_blast.pl	BLAST	Busca de similaridade entre seqüências nucleotídicas ou protéicas	Altschul <i>et al.</i> , 1997
Busca de peptídeo sinal e regiões transmembranares			
annotation_signalp.pl	SignalP	Predição de sítios de clivagem de peptídeo sinal em seqüências protéicas	Bendtsen <i>et al.</i> , 2004
annotation_tmhmm.pl	TMHMM	Predição de hélices transmembranares e regiões intra- e extracelulares em seqüências protéicas	Krogh <i>et al.</i> , 2001
annotation_phobius.pl	Phobius	Predição de hélices transmembranares e de sítios de clivagem de peptídeo sinal em seqüências protéicas	Käll <i>et al.</i> , 2004
Busca de motivos protéicos			
annotation_hmmer.pl	HMMER	Busca de motivos protéicos utilizando bancos com perfis de modelos ocultos de Markov (HMMs) como o Pfam	Eddy, 1998; Bateman <i>et al.</i> , 1999
annotation_interpro.pl	InterProScan	Busca de motivos protéicos usando uma combinação de bases de dados obtidas por diferentes métodos, incluindo as bases PROSITE, PRINTS, Pfam, ProDom e SMART, entre outras	Zdobnov e Apweiler, 2001; Quevillon <i>et al.</i> , 2005
annotation_rpsblast.pl	RPS-BLAST	Busca de domínios conservados usando bases de dados compostas de matrizes de pontuação posição-específica (PSSMs)	Marchler-Bauer <i>et al.</i> , 2002

(continua)

Tabela 1 (Continuação)

Componente	Programa associado	Função	Referência(s)
Busca de sítio de ancoragem GPI			
annotation_dgpi.pl	DGPI	Busca de regiões canônicas de clivagem para ligação em âncoras de GPI	Kronegg e Buloz, 1999
Mapeamento de termos GO			
mapping_GO.pl	map2slim.pl	Mapeamento de termos GO para cada produto de gene anotado. Gera um arquivo denominado <i>Gene Annotation Format</i> , a partir do qual se quantifica o número de seqüências que contém os termos GO definidos no subconjunto de ontologias (arquivo <i>GO slim</i>)	Chris Mungall, não publicado
Seletor de ORFs			
report_ORF.pl	<i>Standalone</i>	Seleção da ORF mais provável baseando-se em evidências de anotação e critérios configuráveis pelo usuário	-
Geração de relatórios de saída			
report_gff.pl	<i>Standalone</i>	Geração de anotação no padrão GFF3 (<i>Generic Feature Format 3</i>)	-
report_feature_table_submission.pl	<i>Standalone</i>	Geração de anotação no formato <i>Feature Table</i> para submissão de seqüências para bases de dados públicas	-
report_feature_table_artemis.pl	<i>Standalone</i>	Geração de anotação no formato <i>Feature Table</i> com etiquetas adicionais para visualização e edição no programa Artemis	-
report_html.pl	<i>Standalone</i>	Geração de páginas <i>web</i> para apresentação dos resultados da anotação de todas as seqüências processadas	-

(conclusão)

3.2 Conjuntos de dados

Foram utilizadas seqüências de cDNAs de três espécies de *Eimeria* de galinha doméstica: *E. acervulina*, *E. maxima* e *E. tenella*. No caso das duas primeiras espécies, todas as leituras consistiram em seqüências do tipo ORESTES, geradas pelo nosso laboratório dentro do Projeto ORESTES de *Eimeria*, coordenado pela Prof^a. Alda M. B. Madeira. No total, foram empregadas 19.488 leituras de *E. acervulina*, obtidas de quatro estágios diferentes de desenvolvimento (oocistos não esporulados, oocistos em fase de esporoblasto, oocistos esporulados e esporozoítos). Para *E. maxima*, foram empregadas 21.013 de seqüências de oocistos não esporulados, oocistos em fase de esporoblasto, oocistos esporulados e merozoítos de segunda geração.

No caso de *E. tenella*, foram utilizadas 17.568 leituras do tipo ORESTES geradas em nosso laboratório para os seguintes estágios: oocistos não esporulados, oocistos em fase de esporoblasto, oocistos esporulados, esporozoítos e merozoítos de segunda geração. Além disso, também utilizamos vários conjuntos de leituras obtidas por outros grupos, conforme apresentado na Tabela 2. Como pode ser visto nessa Tabela, os dados de outros grupos consistiram em seqüências em formato FASTA, bem como cromatogramas e arquivos SCF de experimento (contendo dados de qualidade).

Tabela 2 - Características de origem, tipo de seqüências e quantidade de leituras utilizadas para pré-processamento para as seqüências de *E. tenella*.

Origem	Tipo de seqüência	Nº de leituras	Formato	Fonte
SVRI ¹ e CAU ² , China	ESTs convencionais	2.914	FASTA	NCBI ³
Sanger Institute, Reino Unido	ESTs convencionais	9.778	FASTA e SCF	Sanger e NCBI ³
Universiti Kebangsaan Malaysia	ESTs convencionais	1.051	FASTA	NCBI ³
Washington University, EUA	ESTs convencionais	27.500	Cromatogramas	WUSTL ⁴
USDA ⁵ , EUA	ESTs convencionais	1.666	Cromatogramas	USDA ⁵
USP ⁶ , Brasil	ORESTES	17.568	Cromatogramas	USP ⁶
Total	-	60.477	-	-

¹ Shanghai Veterinary Research Institute, Xangai, China

² College of Veterinary Medicine, China Agricultural University, Beijing, China

³ National Center of Biotechnology Information (<http://www.ncbi.nlm.nih.gov/>), Bethesda, EUA

⁴ Washington University (<http://www.washington.edu/>), Seattle, EUA

⁵ United States Department of Agriculture (<http://www.usda.gov/wps/portal/usdahome>), Baltimore, EUA

⁶ Universidade de São Paulo, São Paulo, Brasil

A nomenclatura de todas as leituras foi padronizada de forma a conter um sufixo de identificação de espécie (Tn) e de estágio evolutivo (Ou – oocisto não esporulado, Op – oocisto em fase de esporoblasto, Os – oocisto esporulado, Sz – esporozoíto, lz – merozoíto de

primeira geração e Mz – merozoíto de segunda geração). A seguir foi acrescentada a nomenclatura original, seguida de um identificador do laboratório gerador do dado. Por exemplo, a leitura ETwisMER-10a04.plk gerada no Sanger Insitute, passou a ser Tn1z-10a04-sanger.plk.

3.3 Pré-processamento das seqüências

Todas as seqüências utilizadas no presente trabalho foram submetidas a um *pipeline* de pré-processamento visando a avaliação de qualidade, mascaramento de vetores e *primers*, aparamento de pontas de baixa qualidade, filtragem por tamanho, filtragem de contaminantes, etc. Sempre que possível, foram utilizados os dados de qualidade. Assim, cromatogramas foram inseridos diretamente no *pipeline* para avaliação pelo programa Phred (Ewing e Green, 1998; Ewing *et al.*, 1998). No caso de arquivos de experimento SCF, foi escrito um *script* de conversão para formato PHD (formato do programa Phred), contendo os dados da seqüência e os respectivos valores de qualidade. Quando valores de qualidade não estavam disponíveis, como no caso de seqüências no formato FASTA, foram criados arquivos PHD com qualidade 13 para todas as bases. Essa qualidade, relativamente baixa, é bastante neutra para os programas de montagem. Por exemplo, se houver discrepâncias de bases em posições com qualidade 13, as mesmas não serão consideradas como discrepâncias de alta qualidade. O *pipeline* de processamento pré-anotação foi construído e executado na plataforma EGene, e consistiu nas seguintes etapas:

- a) **inserção (*upload*) dos dados no *pipeline*.** No caso de dados no formato de cromatogramas, foi feita a atribuição de bases (“base calling”) e avaliação da qualidade com o componente `upload_traces.pl` do EGene, acoplado ao programa Phred. No caso de dados convertido previamente para o formato PHD (vide acima), utilizou-se o componente `upload_phd_dir.pl`.
- b) **mascaramento das seqüências dos *primers* de ORESTES.** Cada biblioteca de ORESTES foi construída com um ou mais *primers* arbitrários e um total de 96 clones foram seqüenciados em uma microplaca. As seqüências derivadas de cada placa foram submetidas a um mascaramento contra o(s) *primer*(s) utilizado(s) na construção respectiva. Para isso, utilizou-se o componente `mask_cross_match.pl` do EGene e o programa Cross_match. Essa etapa não foi realizada para os dados externos ao nosso laboratório.

- c) **maskamento de bases de vetor.** Todas as seqüências nucleotídicas foram mascaradas contra a seqüência do vetor pGEM T-Easy (Promega) no caso dos ORESTES produzidos localmente, e com a base de vetores UniVec para todos os demais dados. Utilizou-se o componente do EGene `mask_cross_match.pl` acoplado ao programa `Cross_match` (Phil Green, não publicado), do pacote Phred/Phrap/Consed.
- d) **filtragem por qualidade.** Foi utilizado o componente `filter_quality.pl`. Usando-se uma janela deslizante, para ser aceita, uma leitura deveria apresentar uma seqüência contínua de 100 pb com pelo menos 85% das bases com qualidade Phred igual ou maior do que 13.
- e) **aparamento das pontas.** Após a filtragem por qualidade, as seqüências das pontas tiveram as suas bases de qualidade ruim aparadas. Para isso, usou-se o componente `trimming.pl`. Na primeira etapa usou-se uma janela deslizante de 30 pb. As regiões de seqüência, para serem aceitas, tinham que apresentar 80% das bases dentro dessa janela com índice Phred igual ou maior do que 10.
- f) **filtragem por tamanho.** Após o processamento e aparamento de pontas, todas as seqüências foram submetidas ao programa `filter_size.pl` do EGene. Somente foram aceitas seqüências acima de 70 pb.
- g) **filtragem de seqüências mitocondriais.** Foi feita a filtragem de todas as seqüências através do componente `filter_cross_match.pl`, que utiliza o programa `Cross_match` para fazer buscas de similaridade. Todas as seqüências foram comparadas aos genomas mitocondriais de todas as espécies de *Eimeria* (determinadas previamente em nosso laboratório). Como critério de positividade, a seqüência tinha que apresentar alinhamento usando-se os parâmetros de `Cross_match` “-minmatch 35 -penalty -1 -minscore 30”.
- h) **filtragem de seqüências de plastídeo.** Foi feita a filtragem de todas as seqüências com o componente `filter_cross_match.pl`, que utiliza o programa `Cross_match` para fazer buscas de similaridade. Como base de dados, foi usada a seqüência nucleotídica do genoma de apicoplasto de *E. tenella* (código de acesso AY217738). Foram consideradas positivas as seqüências que apresentaram alinhamento usando-se os parâmetros de `Cross_match` “-minmatch 35 -penalty -1 -minscore 30”.
- i) **filtragem de seqüências ribossômicas.** Foi feita a filtragem de todas as seqüências com o componente `filter_blast.pl`, que se acopla ao programa BLAST. Como base de dados, foi usada uma combinação de seqüências ribossômicas de organismos

Alveolata. Para uma sequência ser considerada positiva, tinha que atender um conjunto de três critérios: apresentar um bloco de alinhamento com *e-value* com valor máximo de 10^{-20} , um tamanho de bloco de no mínimo 90 pb, com porcentagem de identidade de no mínimo 88% dentro do bloco.

- j) **filtragem de seqüências repetitivas.** Nesta etapa também se utilizou o componente `filter_cross_match.pl`, e os mesmos critérios de estringência adotados no item h. Como base de dados, empregou-se o Repbase (<http://www.girinst.org/repbase/index.html>), uma base ampla de elementos repetitivos de eucariotos.
- k) **filtragem de seqüências bacterianas.** Utilizou-se o componente `filter_blast.pl` e os mesmos critérios de estringência adotados no item i. Como base de seqüências, utilizou-se os genomas de *Escherichia coli* (NC_000913.1), *Xanthomonas campestris* pv. *campestris* (NC_003902.1), *Xanthomonas axonopodis* pv. *citri* (NC_003919.1) e *Xylella fastidiosa* 9a5c (NC_002488.1). A *E. coli* foi usada pelo fato de ter sido a bactéria empregada em todas as etapas de clonagem e propagação dos clones recombinantes. As demais bactérias foram incluídas pelo fato de seu DNA ter sido amplamente manipulado dentro do laboratório, devido aos diferentes projetos de seqüenciamento de genomas do qual o nosso grupo fez parte.
- l) **filtragem de seqüências de galinha doméstica.** Utilizou-se o componente `filter_blast.pl` e os mesmos critérios de estringência adotados no item i, com exceção de um valor de *e-value* 10^{-25} , mais estrigente. Como base de seqüências utilizou-se as seqüências cromossômicas do genoma de *Gallus gallus*.
- m) **filtragem de seqüências humanas.** Utilizou-se o componente `filter_blast.pl` e os mesmos critérios de estringência adotados no item i, com exceção de um valor de *e-value* 10^{-25} , mais estrigente. Como base de seqüências utilizou-se as seqüências cromossômicas do genoma humano montado, disponível em <http://genome.ucsc.edu/> (“Golden Path”). Esta etapa de filtragem contra o genoma humano foi utilizada apenas para os dados ORESTES pelo fato dessa técnica ser baseada em amplificação arbitrária de DNA e, portanto, poderia em teoria amplificar o DNA dos manipuladores.

3.4 Montagem das seqüências ESTs/ORESTES (*clustering*)

Para iniciar a anotação dos cDNAs, foi necessário antes realizar a reconstrução de transcritos que corresponde ao agrupamento e montagem dos ESTs/ORESTES. Embora exista

uma série de abordagens e diferentes pacotes para realizar esse agrupamento, decidimos utilizar o programa CAP3 (Huang e Madan, 1999). Visando utilizar a estringência mais adequada, foram testadas diferentes combinações dos parâmetros p (limiar de porcentagem de identidade de sobreposição) e y (faixa de corte de pontas), assim como versões do programa. Como conjunto de dados, foram utilizadas as seqüências de *E. tenella* (Tabela 2) previamente processadas (ver item 3.3).

3.5 Elaboração de protocolos e construção de *pipelines* de anotação

Para a elaboração de protocolos de anotação automática, foram utilizados os componentes de anotação desenvolvidos no presente trabalho (ver item 3.1 e Tabela 1). Um exemplo de *pipeline* de anotação construído no editor de configurações CoEd pode ser visto na Figura 4.

Figura 4 - Tela do editor gráfico (CoEd) do sistema EGene mostrando um *pipeline* com vários componentes de anotação automática (vide ícones na tela) utilizados para processamento de seqüências de cDNA.

Foram testados vários protocolos de anotação, com diferentes programas e parâmetros. O protocolo final utilizado para a anotação automática das seqüências de *Eimeria* spp. está descrito a seguir:

- **Determinação das fases de leitura abertas (ORFs) e tradução protéica.** Nessa etapa utilizou-se o componente `annotation_orf.pl` do EGene. Foram escolhidas as fases de leitura abertas contendo pelo menos 50 resíduos de aminoácidos, sem requerimento de códon de iniciação. Todas as ORFs foram traduzidas conceitualmente utilizando-se o código genético universal e as seqüências nucleotídicas e protéicas correspondentes armazenadas em relatórios.

- **Busca de similaridade.** Para as buscas de similaridade foi utilizado o componente `annotation_blast.pl` do EGene, acoplado ao programa BLAST (Altschul *et al.*, 1997). Como base de dados utilizou-se o Uniref (*UniProt Reference Clusters* – Suzek *et al.*, 2007), que combina seqüências proximamente relacionadas em um única entrada, a fim de acelerar as buscas de similaridade. Foi usada a base de dados UniRef 90, a qual é construída de tal forma que cada grupo (*cluster*) seja composto por seqüências que apresentam pelo menos 90% de similaridade entre si. Foram armazenados os resultados das buscas em arquivos texto. Além disso, os quatro melhores alinhamentos com *e-values* inferiores 1e-06 e mínima porcentagem de identidade igual a 55% foram anotados.

- **Busca de domínios conservados.** As seqüências protéicas geradas pelo componente `annotation_orf.pl` foram submetidas a uma busca de domínios conservados contra a base de dados CDD (Marchler-Bauer *et al.*, 2007). Para isso, utilizou-se o componente `annotation_rpsblast.pl` acoplado ao programa RPS-BLAST (Marchler-Bauer *et al.*, 2002). O RPS-BLAST é um programa que compara uma seqüência protéica contra uma base de dados de matrizes de pontuação posição-específica (*Position Specific Scoring Matrices* – PSSMs). Os resultados completos foram armazenados em arquivos texto e ainda, os alinhamentos com *e-values* inferiores a 10^{-6} foram anotados.

- **Busca de domínios protéicos.** As seqüências protéicas foram submetidas a uma busca de motivos protéicos utilizando-se o componente do EGene `annotation_interpro.pl` acoplado ao programa InterproScan (Mulder e Apweiler, 2007). Todos os resultados foram armazenados em arquivos texto, assim como em arquivos HTML, os quais podem ser visualizados através de qualquer navegador *web*.

- **Busca de seqüências de peptídeo-sinal.** Para essa busca foram utilizados os componentes `annotation_signalP.pl` e `annotation_phobius.pl`, acoplados respectivamente aos programas SignalP (Bendtsen *et al.*, 2004) e Phobius (Käll *et al.*, 2004).

- **Busca de domínios transmembranares.** Nessa busca foram utilizados os componentes `annotation_tmhmm.pl` e `annotation_phobius.pl`, acoplados respectivamente aos programas TMHMM (Krogh *et al.*, 2001) e Phobius.

- **Busca de seqüências de sítios de âncoras GPI.** Nessa etapa utilizou-se o componente `annotation_dgpi.pl` acoplado ao programa DGPI (Kronegg e Buloz, 1999).

- **Seleção das ORFs anotadas.** Foi utilizado o componente `report_ORF.pl` para selecionar as ORFs com maior número de evidências. O conjunto de evidências a ser considerado para a seleção das ORF é livremente definido pelo usuário e nesse caso incluiu resultados de similaridade por BLAST, considerando resultados com melhor *e-value*, presença de domínios conservados detectados pelo RPS-BLAST e motivos protéicos encontrados pelo InterproScan. Quando ocorre um empate do número de evidências, o componente seleciona a ORF de maior comprimento.

- **Mapeamento dos termos GO (*Gene Ontology*).** Uma vez concluído o *pipeline* de anotação automática, foi usado o componente do EGene `mapping_GO.pl` para realizar o mapeamento e quantificação dos termos GO nas três ontologias gênicas para cada seqüência anotada. O mapeamento foi originalmente realizado pelo próprio programa InterproScan, que utiliza uma tabela de conversão denominada `interpro2go`, e guarda os termos encontrados nos resultados da saída. O componente do EGene `mapping_GO.pl`, gera a partir desses termos um arquivo denominado *Gene Association File*, o qual é utilizado pelo *script* `map2slim.pl` para quantificar o número de seqüências com termos GO dentro do subconjunto de ontologias *GO Slim*. O componente gera no final um conjunto de arquivos HTML contendo uma tabela dos termos de *GO Slim*, o número de seqüências positivas para cada termo GO, e os *links* para as respectivas seqüências nucleotídica e protéica. Também é gerada uma versão XML dessa saída, mas, nesse caso as ontologias são representadas de forma hierárquica, com uma árvore na qual os *links* podem ser expandidos ou colapsados.

- **Geração de anotação automática em formato *Feature Table*.** Foram utilizados os componentes do EGene `report_feature_table_submission.pl` e `report_feature_table_artemis.pl` para gerar arquivos de anotação completa em formato *feature table* no padrão de submissão de seqüências para bancos internacionais, e na versão ampliada para visualização e edição no Artemis. Essa última versão contém algumas etiquetas que somente são validas dentro do programa Artemis, como, por exemplo, a visualização de cor nos quadros que representam as ORFs.

- **Geração de anotação automática em formato *Generic Feature Format*.** Foi utilizado o componente do EGene `report_gff.pl` para gerar arquivos de anotação

completa em formato GFF3, que está se tornando um formato universal de intercâmbio de anotações genômicas.

- **Geração de páginas web com resultados da anotação.** Foi utilizado o componente do EGene `report_html.pl` para gerar dinamicamente páginas web para a visualização e consulta dos resultados de anotação para todas as seqüências processadas para cada espécie de *Eimeria* spp.

3.6 Execução dos *pipelines* de anotação

Uma vez construído o *pipeline* de anotação automática no sistema EGene, o mesmo foi executado em três servidores simultâneos, dividindo-se os conjunto de seqüências em tarefas de 100 seqüências cada uma. Assim, no caso de qualquer problema de *software* ou *hardware*, a perda máxima de dados não seria superior a 100 anotações e estas poderiam ser facilmente localizadas e, utilizando-se o mesmo *pipeline*, processadas novamente. Para realizar essa divisão de tarefas, foi criado um *script* em Perl no qual o usuário especifica a quantidade de seqüências para a divisão, e o número e nome dos servidores. O programa monta então uma estrutura de diretórios e arquivos, os quais são enviados a cada um dos servidores de processamento. Assim, após o processamento de cada lote de seqüências, foi gerada uma saída no formato XML com os resultados de todos componentes processados para cada seqüência. Posteriormente, os arquivos XML gerados para cada grupo de 100 seqüências foram concatenados com um segundo *script*, e o XML final foi utilizado para gerar os relatórios nos formatos *Feature Table* e GFF3.

3.7 Curadoria manual da anotação automática

Após a construção do protocolo de anotação, implementação na forma de um *pipeline* no sistema EGene, e execução para cada um dos conjuntos de cDNA reconstruídos de *Eimeria maxima*, *Eimeria acervulina* e *Eimeria tenella*, procedeu-se a um processo de curadoria das anotações. Esse processo foi realizado de forma iterativa para conjuntos contendo algumas dezenas de seqüências, de forma que os problemas ou erros detectados eram corrigidos nos componentes de geração de anotação, um novo *pipeline* era executado, e as anotações novamente checadas de forma manual. Esse processo foi repetido por várias vezes até se concluir que as anotações automáticas estavam confiáveis.

RESULTADOS

4 RESULTADOS

4.1 Desenvolvimento de componentes de anotação automática para o sistema EGene

Conforme mostrado na Tabela 1, apresentada no item 3.1, foi desenvolvido um conjunto de componentes de anotação para o sistema EGene, os quais serão detalhadamente descritos nos itens a seguir.

4.1.1 Busca de ORFs (*Open Reading Frames*)

O componente `annotation_orf.pl` foi desenvolvido para a busca de fases abertas de leitura (ORFs), sendo que a construção desse componente teve como base a ferramenta *ORF Finder*, no site do NCBI (<http://www.ncbi.nlm.nih.gov/gorf/gorf.html>). Dada uma sequência de DNA, o componente encontra todas as possíveis ORFs, conforme um tamanho mínimo (em resíduos de aminoácidos) definido pelo usuário e faz a tradução conceitual das ORFs utilizando a tabela de código genético especificada. As opções de código genético, bem como o número que deverá ser usado no campo para a escolha da tabela, estão apresentados abaixo (Tabela 3), e seguem o padrão das tabelas de códigos genéticos do NCBI (*National Center for Biotechnology Information*).

Tabela 3 - Opções de código genético para diferentes organismos para o componente `annotation_orf.pl`.

Identificação no parâmetro	Código genético
1 ou standard	<i>Standard</i>
2 ou vertebrate	<i>Vertebrate Mitochondrial</i>
3 ou yeast	<i>Yeast Mitochondrial</i>
4 ou mold	<i>Mold, Protozoan and Coelenterate Mitochondria</i>
5 ou invertebrate	<i>Invertebrate Mitochondrial</i>
6 ou ciliate	<i>Ciliate, Dasycladacean and Hexamita Nuclear</i>
9 ou echinoderm	<i>Echinoderm Mitochondrial</i>
10 ou euplotid	<i>Euplotid Nuclear</i>
11 ou bacterial	<i>Bacterial Code</i>
12 ou alternative yeast	<i>Alternative Yeast Nuclear</i>
13 ou ascidian	<i>Ascidian Mitochondrial</i>
14 ou flatworm	<i>Flatworm Mitochondrial</i>
15 ou blepharisma	<i>Blepharisma Nuclear</i>
16 ou chlorophycean	<i>Chlorophycean Mitochondrial</i>
21 ou trematode	<i>Trematode Mitochondrial</i>
22 ou scenedesmus	<i>Scenedesmus obliquus Mitochondrial</i>

As seqüências nucleotídicas das ORFs e suas traduções são armazenadas em arquivos texto no formato FASTA, incluindo ainda informações de coordenadas da seqüência, fase de leitura e tamanho. O usuário pode especificar ainda o uso ou não de um códon de iniciação alternativo, que varia para cada tabela de código genética de cada organismo. O componente também possui parâmetros que definem a obrigatoriedade ou não do uso do códon de iniciação, sendo que o códon ATG (metionina) é definido como valor `default`. Além disso, É importante lembrar que a não obrigatoriedade de um códon de iniciação na seqüência, é particularmente importante para encontrar ORFs em *clusters* de ESTs que não necessariamente cobrem a porção 5' terminal dos respectivos transcritos. Essa característica também é útil quando se deseja encontrar as ORFs correspondentes a éxons em seres eucariotos. A Tabela 4 apresenta os parâmetros existentes para esse componente, que podem ser alterados pelo usuário, bem como seus respectivos significados e valores `default`.

Tabela 4 - Parâmetros de configuração do componente `annotation_orf.pl` para a busca e tradução conceitual de ORFs.

Parâmetro	Descrição	Valor default
<code>minimum_size</code>	Tamanho mínimo em aminoácidos para a busca da ORF	100
<code>genetic_code</code>	Tabela de código genético usada na tradução	<code>standard</code>
<code>use_alternative_start_codon</code>	Uso de códon de iniciação alternativo para cada código genético	<code>false</code>
<code>obligatory_start_codon</code>	Uso de códon de iniciação obrigatório	ATG (met)
<code>translation_fasta_file</code>	Arquivo de saída que armazena as ORFs (nt e aa)	<code>orf.txt</code>
<code>annotation_color</code>	Especifica padrão de cor para visualização em editor	

O parâmetro `annotation_color` especifica um modelo aditivo de cores do tipo RGB (*red-green-blue*), onde podemos realizar combinações numéricas de 0 a 255 para cada uma das cores primárias, permitindo o uso de um total de 16 milhões de cores possíveis para a visualização de *features* obtidas e que podem ser abertas e visualizadas em um editor e anotador como o Artemis (Rutherford *et al.*, 2000).

4.1.2 Busca de genes de tRNA

Para realizar a busca por genes de tRNA foi desenvolvido o componente `annotation_trna.pl` que utiliza o programa tRNAscan-SE (Lowe e Eddy, 1997). Esse

aplicativo detecta genes de RNA transportador em seqüências genômicas, usando três métodos que combinam velocidade, sensibilidade e seletividade. O programa une a especificidade do pacote de predição probabilística de RNA (Eddy e Durbin, 1994), com a velocidade e sensibilidade do tRNAscan 1.3 (Fichant e Burks, 1991), adicionado de uma implementação do algoritmo descrito por Pavesi *et al.* (1994) para pesquisa por promotores de tRNA. Os parâmetros existentes para o componente `annotation_trna.pl`, bem como seus respectivos significados e valores `default`, estão apresentados na Tabela 5.

Tabela 5 - Parâmetros de configuração do componente `annotation_trna.pl` para a busca de genes de tRNA.

Parâmetro	Descrição	Valor default
<code>tRNA_search_model</code>	Modelo para a busca de tRNA	-G
<code>secondary_structure_output_file</code>	Arquivo de estruturas secundárias dos tRNAs	
<code>standard_output_file</code>	Arquivo com resultados padrão do programa	
<code>statistics_output_file</code>	Arquivo com resultados estatísticos do programa	
<code>report_file</code>	Arquivo com a lista de tRNA encontrados	<code>trna.txt</code>
<code>annotation_color</code>	Padrão de cor para visualização em editor	

O programa tRNAscan-SE apresenta diferentes opções de modelos para a busca de tRNAs, incluindo um modelo para arqueos, eubactérias, organelas e um modelo geral de busca (-G), usada como `default` pelo componente.

4.1.3 Busca de regiões repetitivas seriadas com o programa *mreps*

Uma repetição seriada no DNA, corresponde a duas ou mais cópias contínuas de um padrão de nucleotídeos. O componente `annotation_mreps.pl` funciona acoplado ao programa *mreps* (Kolpakov *et al.*, 2003). O programa *mreps* realiza a busca de seqüências repetitivas em duas fases. Primeiro, ele coleciona seqüências repetitivas através de um algoritmo combinatorial eficiente e, posteriormente, essas seqüências servem como base para aplicar um tratamento heurístico a fim de se obter repetições biologicamente relevantes (Kolpakov *et al.*, 2003). A Tabela 6 lista e descreve os parâmetros de configuração utilizados pelo componente `annotation_mreps.pl`, e que podem ser alterados pelo usuário.

Tabela 6 - Parâmetros de configuração do componente `annotation_mreps.pl` para a busca de repetições seriadas.

Parâmetro	Descrição	Valor default
Resolution	Resolução (nível de erro)	
start_search_at	Posição inicial da sequência	
end_search_at	Posição final da sequência	
Exponent	Reporta as repetições com expoente menor do que o especificado	
Allowsmall	Especifica uma busca aleatória	<i>false</i>
Minsize	Tamanho mínimo da repetição	
Maxsize	Tamanho máximo da repetição	
Minperiod	Período mínimo da repetição	
Maxperiod	Período máximo da repetição	
sliding_window	Processa o resultado em janelas deslizantes	
output_file	Arquivo de saída que armazena os resultados	<code>mreps.txt</code>
annotation_color	Padrão de cor para visualização em editor	

4.1.4 Busca de regiões repetitivas seriadas com o programa String

O componente `annotation_string.pl` trabalha acoplado ao programa String (Parisi *et al.*, 2003) para a busca de regiões repetitivas seriadas. Os parâmetros de configuração do componente estão descritos na Tabela 7.

Tabela 7 - Parâmetros de configuração do componente `annotation_string.pl` para a busca de repetições seriadas.

Parâmetro	Descrição	Valor default
minimum_score	Valor mínimo de escore para a busca	
report_file	Arquivo de saída que armazena os resultados	<code>string.txt</code>
command_line_options	Parâmetros adicionais do String na linha de comando	
annotation_color	Padrão de cor para visualização em editor	

4.1.5 Busca de regiões repetitivas seriadas usando TRF

O programa Tandem Repeats Finder (Benson, 1999), também realiza a busca por repetições seriadas em sequências. Para realizar essa busca, o TRF executa duas etapas, sendo a primeira para a detecção *ab initio* dos candidatos através de critérios estatísticos, e uma segunda, na qual são feitas as análises que geram os alinhamentos. A Tabela 8 apresenta os parâmetros do componente `annotation_trf.pl`.

Tabela 8 - Parâmetros de configuração do componente `annotation_trf.pl` para a busca de repetições seriadas.

Parâmetro	Descrição	Valor default
<code>matching_weight</code>	Penalidade para <i>match</i>	
<code>mismatch_penalty</code>	Penalidade para <i>mismatch</i>	
<code>indel_penalty</code>	Penalidade para inserção/deleção	
<code>match_probability</code>	Probabilidade de <i>match</i>	
<code>indel_probability</code>	Probabilidade de inserção/deleção	
<code>minimum_alignment_score</code>	Escore mínimo do alinhamento	
<code>maximum_period_size</code>	Tamanho máximo do período	
<code>report_file</code>	Arquivo de saída que armazena os resultados	<code>trf.txt</code>
<code>annotation_color</code>	Padrão de cor para visualização em editor	

4.1.6 Predição de genes

Foram desenvolvidos sete componentes para interação com programas preditores de genes (vide Tabela 1). Procurou-se cobrir uma gama relativamente ampla de preditores que utilizam diferentes métodos de busca de regiões codificadoras. A Tabela 9 apresenta os componentes e respectivos parâmetros de configuração.

Tabela 9 - Parâmetros de configuração dos componentes preditores de genes.

Parâmetros comuns	Descrição	Valor default
ESTScan, Genscan, GlimmerM, GlimmerHMM, Phat, SNAP e Twinscan		
<code>training_set_location</code>	Arquivo ou diretório de treinamento	
<code>command_line_options</code>	Parâmetros adicionais de linha de comando de cada programa	
<code>output_file</code>	Arquivo de saída que armazena os resultados	
<code>Annotation_color</code>	Padrão de cor para visualização em editor	
Parâmetros específicos		
Twinscan		
<code>Informant</code>	Arquivo do informante	
<code>conservation_model_file</code>	Arquivo com modelo de conservação	

4.1.7 Mapeamento de cDNAs em genomas com os programas Sim4 e Exonerate

A ferramenta Sim4 (Florea *et al.*, 1998) realiza o mapeamento de seqüências de cDNA em DNA genômico, através da busca de similaridade entre o genoma e os cDNAs. O programa Exonerate (Slater e Birney, 2005), por sua vez, é uma ferramenta genérica de comparação de seqüências, que utiliza para sua execução a biblioteca de programação dinâmica C4. Essa biblioteca gera modelos heurísticos de alinhamentos via programação dinâmica, o que torna esse aplicativo mais rápido e genérico em relação à ferramenta Sim4. Foram desenvolvidos dois componentes para mapeamento de cDNA em genomas, o `annotation_sim4.pl` e `annotation_exonerate.pl`, cujos parâmetros de configuração estão apresentados nas Tabela 10 e 11.

Tabela 10 - Parâmetros de configuração do componente `annotation_sim4.pl` para o mapeamento de cDNA em DNA genômico.

Parâmetro	Descrição	Valor default
Database	Base de dados para a busca de similaridade (genoma)	
remove_polyA_tails	Opção de remoção de cauda poliA	true
command_line_options	Parâmetros adicionais da linha de comando do Sim4	
output_file	Arquivo de saída que armazena os resultados	sim4.txt
annotation_color	Padrão de cor para visualização em editor	

Tabela 11 - Parâmetros de configuração do componente `annotation_exonerate.pl` para o mapeamento de cDNA em DNA genômico.

Parâmetro	Descrição	Valor default
Database	Base de dados para a busca de similaridade (genoma)	
minimum_score	Escore mínimo para gerar alinhamento	
command_line_options	Parâmetros adicionais da linha de comando do Exonerate	
output_file	Arquivo de saída que armazena os resultados	exonerate.txt
annotation_color	Padrão de cor para visualização em editor	

É importante ressaltar que o componente `annotation_exonerate.pl` utiliza como modelo de alinhamento do Exonerate a opção `est2genome`, que permite o mapeamento de cDNAs e inclui a possível existência de íntrons.

4.1.8 Busca de similaridade com o programa BLAST

Para a utilização do pacote BLAST (Altschul *et al.*, 1997), visando o alinhamento local de seqüências de DNA e proteínas, foi desenvolvido o componente `annotation_blast.pl`. Os parâmetros de configuração desse componente estão apresentados na Tabela 12.

Tabela 12 - Parâmetros de configuração do componente `annotation_blast.pl` para busca de similaridade entre seqüências.

Parâmetro	Descrição	Valor default
Database	Base de dados para execução	
Program	Tipo de programa (blastn, blastp, etc.)	
window_size	Tamanho de janela (pb) para processamento	
window_overlap	Tamanho da sobreposição (pb) entre as janelas	
maximum_Evalue	Valor máximo de E_value para mostrar alinhamento	1e-06
minimum_block_size	Tamanho mínimo do bloco para mostrar alinhamento	
minimum_identity_percent	Porcentagem mínima de identidade para alinhamento	
minimum_similarity_percent	Porcentagem mínima de similaridade para alinhamento	
number_of_top_scoring_matches	Número de blocos de alinhamento para armazenamento	4
output_file	Arquivo de saída que armazena os resultados	blast.txt
command_line_options	Parâmetros adicionais da linha de comando do Blast	
output_dir_HTML	Diretório de armazenamento de arquivos HTML de saída	blast_dir
annotation_color	Padrão de cor para visualização em editor	
input_from_evidence	Define se anotação se refere ao DNA ou aos produtos protéicos	false

No parâmetro `program`, o usuário especifica qual programa do pacote `blastall` deseja utilizar para a busca de similaridade. Há cinco programas diferentes disponíveis, sendo que as descrições de cada programa estão apresentadas na Tabela 13.

Tabela 13 - Opções do pacote `blastall` para buscas de similaridade.

Programa	Descrição
<code>blastp</code>	Compara uma sequência protéica contra um banco de sequências protéicas
<code>blastn</code>	Compara uma sequência nucleotídica contra um banco de sequências nucleotídicas
<code>blastx</code>	Compara uma sequência nucleotídica traduzida contra um banco de sequências protéicas
<code>tblastn</code>	Compara uma sequência protéica contra um banco de sequências nucleotídicas traduzidas
<code>tblastx</code>	Compara uma sequência nucleotídica traduzida contra um banco de sequências nucleotídicas traduzidas

As buscas de similaridade com BLAST não devem ser feitas com fragmentos longos demais por varias razões, incluindo a alta demanda de recursos computacionais. Além disso, como o BLAST não reporta todos os blocos de alinhamento, mas apenas os melhores, se a sequência for muito longa, poderá gerar um número muito grande de *hits* em alguns trechos (sem *hits* positivos), impedindo a visualização de blocos de similaridade de maior *E-value* em outros trechos. Para lidar com esses problemas, o componente `annotation_blast.pl` permite dividir a sequência em trechos de tamanho definido, definido pelo parâmetro `window_size`, e cada um desses fragmentos é então submetido a uma busca de similaridade. Para que nenhum bloco de alinhamento da sequência seja eventualmente quebrado, o usuário pode definir um valor de sobreposição em pares de bases entre as “janelas”, através do parâmetro `window_overlap`.

Outro aspecto importante diz respeito aos critérios para se definir se um resultado de alinhamento de BLAST deve ou não ser considerado significativo. Assim, o componente `annotation_blast.pl` permite definir o tamanho mínimo do bloco de alinhamento (parâmetro `minimum_block_size`), as porcentagens mínimas de identidade (parâmetro `minimum_identity_percent`) e similaridade (parâmetro `minimum_similarity_percent`) e o valor máximo do *E-value* (parâmetro `maximum_Evalue`). Caso o alinhamento obtido esteja dentro dos critérios estabelecidos pelo usuário, o mesmo será reportado, caso contrário não será considerado. Além desses critérios, o usuário também pode especificar através do parâmetro `number_of_tops_scoring_matches` quantos resultados positivos deverão ser reportados. Assim, por exemplo, se esse valor for quatro, isso significa que da lista de *hits* gerada pelo BLAST, apenas os quatro melhores *hits* (de acordo com o *E-value*) serão apresentados. Caso nenhum valor seja especificado, será mostrado apenas o melhor resultado.

O parâmetro `input_from_evidence` está presente em todos os componentes que utilizam como entrada sequências de proteínas. Esse parâmetro pode receber as opções `true` ou `false`. Se a opção `true` for utilizada, o componente será executado apenas sobre as

evidências, ou seja, usando as seqüências protéicas traduzidas a partir das ORFs geradas, o que implica que o usuário tenha que utilizar previamente o componente `annotation_orf.pl` (descrito no item 4.1.1).

4.1.9 Busca de motivos protéicos usando HMMER/Pfam

O HMMER (Eddy, 1998) é um pacote que implementa a construção e utilização de bases de dados de perfis de modelos ocultos de Markov (HMMs). Assim, seqüências de proteínas relacionadas são submetidas a um alinhamento múltiplo e esse alinhamento é empregado para se construir um perfil probabilístico baseado em HMMs. Essa é a abordagem utilizada na construção da base de dados Pfam (Bateman *et al.*, 1999). Essa base é uma biblioteca de alinhamentos múltiplos de seqüências e de seus perfis de HMM. O componente `annotation_hmmer.pl` interage com o programa `hmmpfam` para realizar buscas de motivos protéicos contra a base Pfam. Os parâmetros de configuração presentes nesse componente estão mostrados na Tabela 14.

Tabela 14 - Parâmetros de configuração do componente `annotation_hmmer.pl` para a busca de motivos protéicos contra a base de dados Pfam.

Parâmetro	Descrição	Valor default
<code>HMM_file</code>	Arquivo da base de dados de HMMs	
<code>evaluate_cutoff</code>	Valor máximo de <code>E_value</code> para corte	
<code>number_of_tops_scoring_matches</code>	Número de resultados positivos a serem armazenados	4
<code>command_line_options</code>	Parâmetros adicionais do <code>hmmpfam</code>	
<code>output_file</code>	Arquivo de saída que armazenaarmazena os resultados	<code>hmmer.txt</code>
<code>input_from_evidence</code>	Define se anotação se refere ao DNA ou aos produtos protéicos	<code>true</code>

O parâmetro `HMM_file` corresponde ao arquivo que contém os perfis de HMM. No caso da base Pfam, existem dois tipos de arquivos: o `Pfam_ls`, que corresponde aos modelos obtidos a partir de alinhamentos globais/completos do domínio ou família, e o `Pfam_fs` que contém os modelos relativos a alinhamentos locais ou fragmentados.

4.1.10 Busca de motivos protéicos com o programa RPS-BLAST

O RPS_BLAST (Marchler-Bauer *et al.*, 2002) é um programa que executa buscas de similaridade por posições específicas reversas, uma variante do algoritmo PSI-BLAST. Assim, as buscas são realizadas utilizando-se uma sequência protéica como consulta, contra matrizes de pontuação posição-específica, derivadas de alinhamentos da base de dados CDD (*Conserved Domains Databases*, <http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml>). O componente `annotation_rpsblast.pl`, trabalha de forma integrada ao RPS-BLAST, e seus parâmetros de configuração estão apresentados na Tabela 15.

Tabela 15 - Parâmetros de configuração do componente `annotation_rpsblast.pl` para a busca de similaridade para proteínas.

Parâmetro	Descrição	Valor default
Database	Nome da base de dados	
maximum_Evalue	Valor máximo de E_value para alinhamentos	
number_of_top_scoring_matches	Número de resultados positivos a serem armazenados	4
command_line_options	Parâmetros adicionais da linha de comando do RPS_BLAST	
output_file	Arquivo de saída que armazena os resultados	rpsblast.txt
input_from_evidence	Define se anotação se refere ao DNA ou aos produtos protéicos	true

4.1.11 Busca de motivos protéicos com o programa InterProScan

O InterProScan (Zdobnov e Apweiler, 2001; Quevillon *et al.*, 2005) é uma ferramenta poderosa para a detecção de motivos protéicos, através da busca integrada em bases de dados como PROSITE, PRINTS, Pfam, ProDom, SMART, TIGRFAMS, PIR superfamily, SUPERFAMILY, Gene3D e PANTHER. Essas bases compreendem motivos e famílias protéicas determinados por diferentes métodos, que incluem o uso de expressões regulares, perfis baseados em matrizes de escore posição-específica e modelos ocultos de Markov (HMMs), entre outros. O componente `annotation_interpro.pl` trabalha de forma acoplada ao InterProScan e seus parâmetros de configuração estão apresentados na Tabela 16.

Tabela 16 - Parâmetros de configuração do componente `annotation_interpro.pl`.

Parâmetro	Descrição	Valor default
<code>run_program</code>	Modo de execução do InterPro	<code>standalone</code>
<code>URL</code>	Endereço da página do programa InterProScan	
<code>tmp_directory</code>	Caminho do diretório temporário para apagar arquivos temporários gerados após a execução	
<code>run_BlastProDom</code>	Utilizar base BlastProDom	<code>yes</code>
<code>run_HMMTigr</code>	Utilizar base HMMTigr	<code>yes</code>
<code>run_TMHMM</code>	Utilizar aplicativo TMHMM	<code>yes</code>
<code>run_FPrintScan</code>	Utilizar base FprintScan	<code>yes</code>
<code>run_ProfileScan</code>	Utilizar base ProfileScan	<code>yes</code>
<code>run_HMMPanther</code>	Utilizar base HMMPanther	<code>yes</code>
<code>run_HMMPIR</code>	Utilizar base HMMPIR	<code>yes</code>
<code>run_ScanRegExp</code>	Utilizar base ScanRegExp	<code>yes</code>
<code>run_Gene3D</code>	Utilizar base Gene3D	<code>yes</code>
<code>run_HMMPfam</code>	Utilizar base HMMPfam	<code>yes</code>
<code>run_SuperFamily</code>	Utilizar base SuperFamily	<code>yes</code>
<code>run_HMMSmart</code>	Utilizar base HMMSmart	<code>yes</code>
<code>run_SignalPHMM</code>	Utilizar aplicativo SignalP HMM	<code>yes</code>
<code>sequence_type</code>	Tipo de sequência para execução	<code>protein</code>
<code>translation_table_code</code>	Tabela de código genético para tradução	<code>standard</code>
<code>minimum_ORF_size</code>	Tamanho mínimo para busca da ORF	<code>100</code>
<code>output_file</code>	Arquivo de saída que armazena os resultados	<code>interpro.txt</code>
<code>output_dir_HTML</code>	Diretório com resultados HTML para cada sequência	<code>interpro_dir</code>
<code>images_dir</code>	Diretório de imagens para composição do HTML	
<code>input_from_evidence</code>	Define se anotação se refere ao DNA ou aos produtos protéicos	<code>true</code>

O componente `annotation_interpro.pl` trabalha de duas maneiras. Se o usuário especificar o modo de execução através de um servidor remoto (parâmetro `run_program = web`), o componente utiliza um servidor *web*, cujo endereço é definido pelo parâmetro `URL`. Esse modo de execução é particularmente útil para usuários que não dispõem de recursos computacionais suficientes para instalar todas as bases de dados e executar o programa. Alternativamente, o usuário pode executar o programa localmente (parâmetro `run_program = standalone`).

O usuário pode ainda especificar que bases de dados utilizar em suas buscas, assim como se deseja empregar os aplicativos TMHMM e SignalP HMMs. O campo `sequence_type` é obrigatório, e define o tipo de sequência de entrada, sequência nucleotídica (`nucleotide`) ou de aminoácidos (`protein`). Caso a opção tenha sido por sequência nucleotídica, o usuário pode especificar qual tabela de código genético (ver Tabela

3) deseja utilizar para a tradução das ORFs através do parâmetro `translation_table_code`, o valor `default` para esse campo é 0 (tabela `standard`).

O parâmetro `minimum_ORF_size` permite ao usuário especificar o tamanho mínimo das seqüências protéicas a serem traduzidas a partir das ORFs, sendo o valor `default` 100. Em relação ao campo `output_file`, o usuário deve especificar o nome do arquivo que armazenará os resultados, que nesse caso é um relatório da própria saída gerada pelo programa no formato *raw*, um formato texto separado por tabulações. O InterProScan também pode gerar arquivos nos formatos XML, ebixml, HTML, txt e GFF. O componente gera saídas nos formatos *raw* e HTML, permitindo a fácil visualização pelo usuário através de um navegador *web*. O parâmetro `output_dir_HTML` especifica o diretório de armazenamento dos arquivos HTML e imagens.

4.1.12 Busca de domínios transmembranares com o programa TMHMM

O programa TMHMM (Krogh *et al.*, 2001) é responsável pela predição de hélices transmembranares e regiões intra e extracelulares em seqüências protéicas, usando Modelos Ocultos de Markov (HMMs). O componente `annotation_tmhmm.pl` se acopla ao TMHMM e seus parâmetros de configuração estão apresentados na Tabela 17.

Tabela 17 - Parâmetros de configuração do componente `annotation_tmhmm.pl`.

Parâmetro	Descrição	Valor default
<code>output_file</code>	Arquivo de saída que armazena os resultados	<code>tmhmm.txt</code>
<code>graphics_output_dir</code>	Diretório de armazenamento de saídas gráficas	<code>tmhmm_dir</code>
<code>converter_eps2png</code>	Conversor de arquivos eps para png	<code>true</code>
<code>input_from_evidence</code>	Define se anotação se refere ao DNA ou aos produtos protéicos	<code>true</code>

No parâmetro `eps2png` o usuário tem a opção de escolher se deseja que os arquivos gerados pelo programa TMHMM sejam convertidos do formato *eps* (*encapsulated PostScript*) para *png* (*Portable Network Graphics*). Essa conversão é importante para a visualização dos resultados em um navegador de internet, pois arquivos com extensão *eps*, produzidos nativamente pelo TMHMM, necessitam de um visualizador específico como o Ghostview.

4.1.13 Busca de peptídeo sinal com o programa SignalP

O componente `annotation_signalp.pl` realiza a busca de peptídeo sinal usando o programa SignalP (Bendtsen *et al.*, 2004). Esse programa realiza a predição da localização de sítios de clivagem de peptídeo sinal em seqüências protéicas, baseado na utilização combinada de redes neurais artificiais e Modelos Ocultos de Markov (HMMs). Os parâmetros de configuração desse componente estão apresentados na Tabela 18, bem como a descrição de cada parâmetro.

Tabela 18 - Parâmetros de configuração do componente `annotation_signalp.pl`.

Parâmetro	Descrição	Valor default
<code>model</code>	Modelo de organismo para a busca	
<code>output_file</code>	Arquivo de saída que armazena os resultados	<code>signalp.txt</code>
<code>graphics_output_dir</code>	Diretório de armazenamento de saídas gráficas	<code>signalp_dir</code>
<code>converter_eps2png</code>	Conversor de arquivos eps para png	<code>true</code>
<code>input_from_evidence</code>	Define se anotação se refere ao DNA ou aos produtos protéicos	<code>true</code>

4.1.14 Busca de domínios transmembranares e peptídeo sinal com o programa Phobius

O Phobius (Käll *et al.*, 2004) é uma ferramenta responsável pela predição de hélices transmembranares e de sítios de clivagem de peptídeo sinal em seqüências protéicas. Ele é baseado em Modelos Ocultos de Markov (HMMs) que modelam as diferentes regiões da seqüência para o peptídeo sinal e as regiões transmembranares em uma série de estados interconectados. Segundo os autores, o uso do Phobius reduz consideravelmente os erros de detecção de peptídeo sinal e domínios transmembranares em relação ao SignalP e TMHMM, respectivamente. Os parâmetros de configuração do componente `annotation_phobius.pl` estão disponíveis na Tabela 19.

Tabela 19 - Parâmetros de configuração do componente `annotation_phobius.pl`.

Parâmetro	Descrição	Valor default
<code>output_file</code>	Arquivo de saída que guarda os resultados	<code>phobius.txt</code>
<code>graphics_output_dir</code>	Diretório de armazenamento de saídas gráficas	<code>phobius_dir</code>
<code>input_from_evidence</code>	Define se anotação se refere ao DNA ou aos produtos protéicos	<code>true</code>

4.1.15 Busca de sítios de ancoragem GPI com o programa DGPI

O componente `annotation_dgpi.pl` trabalha acoplado ao DGPI (Kronegg e Buloz, 1999), um programa para a busca de sítios de clivagem para a ligação de proteínas em âncoras de GPI (glicofosfatidil-inositol). Os parâmetros de configuração desse componente estão apresentados na Tabela 20.

Tabela 20 - Parâmetros de configuração do componente `annotation_dgpi.pl`.

Parâmetro	Descrição	Valor default
<code>output_file</code>	Arquivo de saída que armazena os resultados	<code>dgpi.txt</code>
<code>HTML_output_dir</code>	Diretório de armazenamento de saída HTML	<code>dgpi_dir</code>
<code>command_line_options</code>	Parâmetros adicionais da linha de comando do programa DGPI	
<code>input_from_evidence</code>	Define se anotação se refere ao DNA ou aos produtos protéicos	<code>true</code>

4.1.16 Seleção de ORF

Um dos problemas enfrentados por qualquer sistema de anotação automática envolve a decisão de que ORFs encontradas para uma dada sequência são de fato verdadeiras. Tomando-se como exemplo um cDNA, somente uma ORF verdadeira deve existir. Contudo, dependendo do tamanho da sequência, ao se atribuir um valor mínimo de, por exemplo, 100 resíduos de aminoácidos, mais de uma única ORF pode atender este critério. Em função do exposto acima, decidimos criar um novo componente no sistema EGene, especificamente para a anotação de cDNAs (ESTs ou transcritos inteiros). Uma maneira convencional de se lidar o problema da escolha da melhor ORF em sequências de cDNA, tem sido a utilização de preditores de genes. Nesse caso, utiliza-se um conjunto de treinamento, geralmente constituído de algumas centenas de genes bem conhecidos. Uma vez treinado o programa, as informações colhidas a partir do conjunto de treinamento são utilizadas para fazer a predição de que regiões de uma sequência seriam potencialmente codificadoras. Um exemplo de programa preditor de genes para ESTs é o ESTScan. Entre as desvantagens e/ou limitações desta abordagem podemos citar que nem sempre a constituição de um conjunto de treinamento é simples, especialmente quando pouco se conhece sobre o organismo em estudo, como é o caso do parasita *Eimeria* spp. Outra limitação é o fato que a taxa de acerto dos melhores preditores, mesmo utilizando-se um bom conjunto de treinamento, é baixa.

Finalmente, os resultados de diferentes programas de predição de genes podem variar bastante, uma vez que os mesmos utilizam diferentes heurísticas para realizar o processo de predição.

Dessa forma, o componente, denominado `report_orf.pl`, parte da premissa que a escolha da ORF verdadeira é na verdade uma hipótese e, como tal, deve-se basear em evidências. O componente deve checar quais e quantas evidências dão suporte a cada uma das ORFs encontradas. Assim, utilizando um critério (configurável pelo usuário) para hierarquizar e quantificar estas evidências, o programa seleciona então a ORF com maior grau de suporte. Obviamente, dada sua forma de funcionamento, este programa não se aplica à anotação de genomas, nem de cDNAs originados de mRNAs policistrônicos.

Após uma série de análises, concluímos que o critério de seleção seria baseado na existência do maior número de evidências, como similaridades significativas por BLAST, domínios conservados pelo RPS-BLAST e busca de motivos protéicos detectados pelo InterProScan e/ou pelo programa HMMER/Pfam. No caso de duas ORFs apresentarem similaridades significativas por BLAST, o critério de escolha é a que apresenta menor *e-value*. Já no caso de duas ORFs apresentarem ambas as mesmas evidências, ou ainda, nenhuma apresentar, o componente seleciona a ORF de maior comprimento em aminoácidos. Embora esses critérios sejam discutíveis, o componente foi desenvolvido de maneira a se adequar aos critérios de qualquer usuário. Assim, os parâmetros de configuração apresentados na Tabela 21, permitem personalizar o componente de acordo com a necessidade e critérios específicos de cada usuário.

Tabela 21 - Parâmetros de configuração do componente `report_orf.pl`.

Parâmetro	Descrição	Valor default
<code>selection</code>	Escolha dos componentes para critério de seleção de ORF	
<code>output_file</code>	Arquivo XML de saída com os resultados	

A Tabela 21 mostra um parâmetro denominado `selection`, no qual o usuário especifica os critérios de seleção para a escolha da ORF. Esse campo deve ser preenchido com o nome de cada componente, seguido pela respectiva prioridade. Por exemplo, podemos utilizar:

```
selection= annotation_blast:1; annotation_rpsblast:2;
annotation_interpro:3; annotation_hmmer:4
```

Utilizando os critérios de seleção de ORFs do exemplo acima, podemos observar possíveis resultados em seqüências hipotéticas nas Tabelas 22 e 23. No primeiro caso (Tabela 22), a seqüência apresenta três ORFs distintas. A análise pelo componente de seleção de ORFs indicou que as ORFs 2 e 3 apresentaram resultado positivo de BLAST, sendo o *e-value* menor na ORF3. Dentro do critério de prioridade, o BLAST já é suficiente para definir a ORF verdadeira e, portanto, a ORF3 foi a escolhida.

Tabela 22 - Caso 1: Resultados hipotéticos de uma anotação de uma seqüência, mostrando as evidências e a escolha do programa `report_orf.pl` baseada em evidências.

	ORF1	ORF2	ORF3	Resultado
BLAST	<i>No hits</i>	1e-05	1e-45	ORF2/ORF3
InterPro	-	+	+	ORF2/ORF3
RPS-BLAST	-	+	+	ORF2/ORF3
Seleção por anotação		ORF3		
Tamanho da ORF (a.a.)	100	350	250	

No caso 2, apresentado na Tabela 23, podemos notar que não há evidências pelo programa BLAST, pois as três ORFs não apresentaram *hits*. Contudo, as ORFs 1 e 3 apresentaram evidências positivas para motivos protéicos pelo programa InterProScan e domínios conservados, pelo programa RPS-BLAST. Uma vez que houve empate para as duas ORFs, o programa selecionou a de tamanho maior, ou seja, a ORF1.

Tabela 23 - Caso 2: Resultados hipotéticos de uma anotação de uma seqüência, mostrando as evidências e a escolha do programa `report_orf.pl` baseada em evidências.

	ORF1	ORF2	ORF3	Resultado
BLAST	<i>No hits</i>	<i>No hits</i>	<i>No hits</i>	-
InterPro	+	-	+	ORF1/ORF3
RPS-BLAST	+	-	+	-
Seleção por anotação				ORF1/ORF3
Tamanho da ORF (a.a.)	350	200	150	
Seleção por tamanho				ORF1

Nas Figuras 5 e 6, podemos observar um exemplo real do funcionamento do seletor de ORFs, utilizando-se para a visualização o editor de anotações Artemis.

Figura 5 - Exemplo de tela do programa Artemis, mostrando todas as ORFs (quadros horizontais em azul) e as respectivas anotações geradas após o processamento de um *pipeline* de anotação. Um total de oito ORFs podem ser visualizadas para esta sequência de cDNA.

No exemplo apresentado na Figura 5, uma sequência real de cDNA de *E. tenella* foi submetida um *pipeline* completo de anotação, tendo apresentado um total de oito ORFs (Figura 5). Após ter sido processada pelo seletor de ORFs (componente `report_orf.pl`), a ORF 3 foi selecionada. As demais anotações foram expurgadas da anotação. Uma evidência adicional que suporta pelo menos a orientação dessa escolha é a existência de uma etiqueta de SAGE, a qual pode ser visualizada na Figura 6 como um quadro azul claro na região 3' à juzante do término da codificação. Essa anotação também demonstra que as evidências não relacionadas diretamente a ORFs não são afetadas pelo processo de seleção

Figura 6 - Exemplo de tela do programa Artemis, mostrando a anotação da sequência mostrada na Figura 5, após o processamento e seleção da ORF pelo componente `report_orf.pl`. Notar que a sequência agora apresenta uma única ORF (quadro horizontal em azul). O quadro de coloração azul clara refere-se a uma sequência de uma etiqueta de SAGE alinhada por BLAST à sequência do cDNA.

4.1.17 Mapeamento de termos de ontologia gênica (GO)

O componente denominado `mapping_GO.pl` foi desenvolvido para realizar a classificação das sequências anotadas dentro de termos de ontologia gênica (GO) e quantificar esses mapeamentos. Conforme descrito no item 1.10, o consórcio de GO foi criado para desenvolver um conjunto de termos padronizados para produtos gênicos e estabelecer suas interrelações. O GO compreende no total um conjunto de três ontologias (processo biológico, função molecular e componente celular). O componente `mapping_GO.pl` utiliza o relatório de saída gerado pelo `annotation_interpro.pl`, onde cada motivo protéico encontrado é mapeado para um termo de ontologia gênica. Em alguns casos, um motivo pode apresentar um único termo GO e, em outros, pode ser mapeado em termos GO das três ontologias. Juntamente com os dados obtidos a partir do componente `annotation_blast.pl`, gera-se

um arquivo de formato tabular denominado *Gene Association File* (GAF). Esse arquivo contém informações a respeito do nome da sequência, códigos de acesso dos bancos Interpro e GO, e códigos de evidência experimental, entre outros. A partir desse arquivo GAF, o *script* `map2slim` (Chris Mungall, não publicado; <http://search.cpan.org/~cmungall/go-perl/>), realiza a classificação e quantificação das sequências de acordo com as categorias definidas pelos termos GO de um arquivo *GO slim*. O arquivo *GO slim* contém uma versão simplificada do GO e representa um subconjunto de ontologias. A Tabela 24 apresenta os parâmetros de configuração do componente `mapping_GO.pl`.

Tabela 24 - Parâmetros de configuração do componente `mapping_GO.pl`.

Parâmetro	Descrição	Valor default
<code>taxon_gene_id</code>	ID de taxonomia para o organismo estudado	
<code>DB_id</code>	Identificador da base de dados das sequências em estudo	
<code>output_file</code>	Arquivo de saída (GAF)	
<code>assignment</code>	Nome do conjunto de sequências em estudo	
<code>run_map2slim</code>	Opção de uso do programa <code>map2slim</code> para mapeamento	<code>true</code>
<code>GO_slim_file</code>	Arquivo GO slim	
<code>ontology_GO</code>	Arquivo com ontologia completa	
<code>output_file_XML</code>	Arquivo de saída XML com resultados do mapeamento	
<code>output_file_HTML</code>	Arquivo de saída HTML com resultados do mapeamento	
<code>report_dir</code>	Diretório de armazenamento de arquivos FASTA e HTML de todas as sequências	<code>GO_dir</code>

Podemos notar na Tabela 24, que nos campos `DB_id` e `assignment` o usuário deve especificar o nome para o conjunto de sequências que está analisando, por exemplo, o nome da espécie seguido do sufixo DB. O parâmetro `run_map2slim` aceita as opções `true` ou `false`. No primeiro caso, o *script* é executado, realizando então a categorização e quantificação dos termos GO encontrados nas sequências dentro da ontologia do arquivo GO Slim. Ao final, criam-se duas saídas, em formato HTML e XML. Os parâmetros `output_file_HTML` e `output_file_XML` designam os nomes desses respectivos arquivos. Como a saída HTML contém *links* para as sequências nucleotídicas e protéicas, o parâmetro `report_dir` indica o nome do diretório que irá armazenar os subdiretórios HTML e FASTA, contendo os arquivos em formato HTML com os termos GO, e os respectivos arquivos em formato FASTA das sequências processadas. Se for escolhida a opção `false`, o programa `map2slim` não é executado e, nesse caso, o componente gera apenas o arquivo *Gene Association File*, mas nenhum arquivo HTML e XML é gerado.

É importante lembrar que os parâmetros `GO_slim_file` e `ontology_GO`, deverão especificar os arquivos contendo a versão reduzida da ontologia gênica (*GO Slim*) e a ontologia completa, respectivamente. Em ambos os casos, o componente trabalha com os arquivos de ontologias nos formatos GO, antigo e em desuso, e OBO, atualmente em desenvolvimento e compatível com o programa OBO_Edit.

Um exemplo de saída HTML do componente `mapping_GO.pl` pode ser visto na Figura 7. Nesse tipo de saída, os termos GO são listados, na coluna da esquerda. Para cada termo GO é indicado o número total de seqüências classificadas neste termo (coluna da direita) e um *link* para as seqüências respectivas (coluna central).

GO Term - Biological Process	sequences	# of counts
GO:0008150: biological process	sequences	51 (89.47%)
GO:0008152: metabolic process	sequences	14 (24.56%)
GO:0044238: primary metabolic process	sequences	9 (15.79%)
GO:0019538: protein metabolic process	sequences	7 (12.28%)
GO:0006810: transport	sequences	4 (7.02%)
GO:0016043: cellular component organization and biogenesis	sequences	3 (5.26%)
GO:0009605: response to external stimulus	sequences	2 (3.51%)
GO:0009058: biosynthetic process	sequences	2 (3.51%)
GO:0006996: organelle organization and biogenesis	sequences	2 (3.51%)
GO:0006950: response to stress	sequences	2 (3.51%)
GO:0015031: protein transport	sequences	1 (1.75%)
GO:0009056: catabolic process	sequences	1 (1.75%)
GO:0007010: cytoskeleton organization and biogenesis	sequences	1 (1.75%)
GO:0006519: amino acid and derivative metabolic process	sequences	1 (1.75%)
GO:0006464: protein modification process	sequences	1 (1.75%)
GO:0006412: translation	sequences	1 (1.75%)
GO:0005975: carbohydrate metabolic process	sequences	1 (1.75%)
GO:0050789: regulation of biological process	sequences	0 (0.00%)
GO:0044403: symbiosis, encompassing mutualism through parasitism	sequences	0 (0.00%)
GO:0040029: regulation of gene expression, epigenetic	sequences	0 (0.00%)
GO:0040007: growth	sequences	0 (0.00%)
GO:0030154: cell differentiation	sequences	0 (0.00%)
GO:0019748: secondary metabolic process	sequences	0 (0.00%)
GO:0019725: cell homeostasis	sequences	0 (0.00%)
GO:0016265: death	sequences	0 (0.00%)
GO:0016049: cell growth	sequences	0 (0.00%)

Figura 7 - Exemplo de saída HTML gerada pelo componente `mapping_GO.pl` para um conjunto de seqüências de *Eimeria maxima*.

Ao se clicar em algum *link* de seqüências, abre-se uma página contendo uma lista de todas as seqüências classificadas dentro daquele termo GO específico (Figura 8). Nessa página o usuário pode clicar nos *links* respectivos para visualizar as seqüências nucleotídica e protéica em formato FASTA ou, ainda, para obter todos os demais termos GO que também foram associados a esta seqüência.

Figura 8 - Exemplo de saída em formato HTML listando as seqüências positivas para um termo GO específico. Note que há *links* para as seqüências nucleotídicas e protéicas respectivamente. *Links* para outros termos GO associados a cada uma das seqüências também estão disponíveis.

Uma alternativa à saída em formato HTML é o arquivo em formato XML, que oferece os mesmos resultados na forma de uma árvore expansível (Figura 9). A informação nesse caso é rigorosamente a mesma que aquela apresentada em formato HTML (Figura 7), mas a apresentação dos termos é feita de maneira hierárquica para as três ontologias. De qualquer forma, ao clicar nos *links* das seqüências, o usuário é direcionado para exatamente a mesma página exemplificada na Figura 8. Como desvantagem, este formato pode ser visualizado nos programas Mozilla Firefox e Opera, mas não é compatível com o programa Internet Explorer.

Figura 9 - Saída web no formato XML gerada pelo programa `mapping_GO.pl` para um conjunto de seqüências de *Eimeria maxima*.

4.1.18 Geração de relatório no formato *Feature Table*

Foram criados dois componentes para a geração de anotações no formato *feature table*, o padrão utilizado pelas bases GenBank, EMBL e DDBJ (<http://www.ncbi.nlm.nih.gov/collab/FT>). Nesse formato, as evidências coletadas a partir dos diferentes componentes são integradas em um só documento, em que as anotações referentes a produtos protéicos são apresentadas no contexto de uma CDS (seqüência codificadora de proteína). O componente `report_feature_table_submission.pl` é responsável pela geração de *features tables* para a submissão para bancos internacionais. Como o usuário pode eventualmente querer visualizar e editar manualmente a anotação em um programa como o Artemis, foi gerado um segundo componente, denominado `report_feature_table_artemis.pl`, o qual gera um relatório no formato *feature table* contendo algumas etiquetas extras, contendo informações mais detalhadas, importantes no processo de curadoria manual. Os dois geradores de anotação no formato *feature table* são capazes de extrair as coordenadas nucleotídicas de cada ORF, e buscar todas as

evidências/inferências contidas nos intervalos respectivos (inicial e final). A Tabela 25 apresenta os parâmetros de configuração dos dois componentes geradores de *feature table*.

Tabela 25 - Parâmetros de configuração dos componentes `report_feature_table_artemis.pl` e `report_feature_table_submission.pl`.

Parâmetro	Descrição	Valor default
<code>components_used</code>	Componentes para os quais serão gerados relatórios	<code>.*</code>
<code>locus_tag_prefix</code>	Prefixo a ser usado na nomenclatura dos <i>loci</i>	
<code>submission_format</code>	Opção de formato do <i>feature table</i>	<code>genbank</code>
<code>output_all</code>	Gera arquivo com a sequência e anotação em formato <i>feature table</i>	<code>true</code>
<code>output_sequence</code>	Gera apenas arquivo da sequência em formato <i>feature table</i>	<code>true</code>
<code>output_feature_table</code>	Gera apenas arquivo em formato <i>feature table</i>	<code>true</code>
<code>output_dir_name</code>	Diretório com relatórios <i>feature table</i> para cada sequência	

O parâmetro `components_used` permite definir para quais componentes executados no *pipeline* deverão ser gerados relatórios *feature table*. O valor default `.*` refere-se a uma expressão regular em Perl e significa que todos os componentes do *pipeline* serão considerados. O campo `submission_format` especifica qual tipo de formato de anotação será gerada. As opções disponíveis são `genbank`, `ddbj` e `embl`, que correspondem respectivamente aos três principais bancos internacionais de seqüências e consideram algumas particularidades específicas de seus formatos. O parâmetro `locus_tag_prefix` permite especificar um prefixo que deverá ser adicionado para nomear todos os supostos *loci*. Se por exemplo, for especificado `Eten`, os *loci* serão automaticamente nomeados `Eten_orf1`, `Eten_orf2`, e assim por diante. Por outro lado, se o usuário não especificar um valor para este parâmetro, e o nome da sequência anotada for, por exemplo, `Etenella.fasta`, o componente denominará automaticamente as ORFs de `Etenella_orf1`, `Etenella_orf2`, `Etenella_orf3`, etc.

4.1.19 Geração de relatório no formato GFF3

O GFF3 é um formato delimitado por tabulações, que descreve as características (*features*) de uma sequência, seja ela genômica ou de cDNA. Todos os campos de um arquivo de formato GFF3 são bem definidos (<http://song.sourceforge.net/gff3.shtml>). Como esse formato tem se tornado um padrão de intercâmbio de informações de anotação, foi

desenvolvido um componente denominado `report_gff.pl`, o qual gera um relatório de saída no formato GFF3 para cada sequência processada no *pipeline* de anotação automática. Os parâmetros de configuração do componente podem ser visualizados na Tabela 26.

Tabela 26 - Parâmetros de configuração do componente `report_gff.pl`.

Parâmetro	Descrição	Valor default
<code>gff_dir</code>	Diretório com arquivos no formato GFF3 de cada sequência	<code>gff3_dir</code>
<code>source</code>	Nome do organismo ou espécie em estudo	
<code>type_of_sequence</code>	Tipo da sequência anotada (genômica ou cDNA)	

Na Tabela 26, podemos notar que o parâmetro `type_of_sequence` refere-se ao tipo de sequência que está sendo anotada, que pode ser genômica (opção `genomic`) ou cDNA (cDNA). O componente foi testado em um conjunto de sequências e os arquivos de anotação gerados foram então submetidos a um programa validador de GFF3, disponível na internet (http://dev.wormbase.org/db/validate_gff3/validate_gff3_online). Todas as incompatibilidades de formato detectadas foram sanadas, de forma que no final conseguiu-se gerar uma anotação totalmente aprovada pelo validador.

4.1.20 Geração de páginas *web*

Como o sistema EGene ainda não está acoplado a uma base de dados relacional, e visando uma apresentação dos resultados em interface *web*, foi desenvolvido o componente `report_html.pl`. Este componente é responsável pela geração dinâmica de páginas *web* para a apresentação e consulta básica da anotação gerada para todas as sequências processadas no *pipeline*. Esse componente foi utilizado para criar um conjunto de páginas para cada uma das espécies de *Eimeria* spp. (<http://www.coccidia.icb.usp.br/eimeria/annotation/>). A Tabela 27 apresenta os parâmetros de configuração do componente. O parâmetro `organism_name` refere-se ao nome do organismo ou espécie em estudo, como, por exemplo, *E. tenella*. O `HTML_index_file` corresponde ao nome do arquivo da página principal que será carregada na *web*. O parâmetro `background_image_file` indica o arquivo que contém uma imagem que será utilizada como plano de fundo para as páginas geradas. Caso esse parâmetro não seja especificado, não será utilizada nenhuma imagem. No parâmetro `component_name_list` devem ser listados os nomes de todos os componentes utilizados no *pipeline* de anotação,

separados por vírgula. Atualmente, o componente `report_html.pl` aceita extrair e apresentar em *web* as evidências dos seguintes programas: `annotation_signalp.pl`, `annotation_tmhmm.pl`, `annotation_phobius.pl`, `annotation_dgpi.pl`, `annotation_blast.pl`, `annotation_rpsblast.pl` e `annotation_interpro.pl`. No futuro, entretanto, se novos componentes forem criados, será necessário apenas adicionar no `report_html.pl` a função de geração dos *links* específicos das novas saídas. Os demais parâmetros apresentados na Tabela 27 referem-se aos nomes de diretórios que armazenarão as diferentes anotações e seqüências.

Tabela 27 - Parâmetros de configuração do componente `report_html.pl`.

Parâmetro	Descrição	Valor default
<code>organism_name</code>	Nome do organismo ou espécie em estudo	
<code>HTML_index_file</code>	Arquivo da página principal	<code>index.php</code>
<code>background_image_file</code>	Arquivo da imagem do plano de fundo	
<code>component_name_list</code>	Lista dos nomes dos componentes executados no <i>pipeline</i>	
<code>FT_artemis_dir</code>	Diretório com arquivos no formato <i>feature table</i> para o Artemis	
<code>FT_submission_dir</code>	Diretório com arquivos no formato <i>feature table</i> para submissão	
<code>FT_artemis_selected_dir</code>	Diretório com arquivos no formato <i>feature table</i> para o Artemis, após a seleção da ORF com o programa <code>report_orf.pl</code>	
<code>FT_submission_selected_dir</code>	Diretório com arquivos no formato <i>feature table</i> para submissão, após a seleção da ORF com o programa <code>report_orf.pl</code>	
<code>GFF_dir</code>	Diretório com arquivos GFF3	
<code>GFF_selected_dir</code>	Diretório com arquivos GFF3 após a seleção da ORF	
<code>fasta_dir</code>	Diretório com arquivos FASTA para cada seqüência processada	
<code>aa_fasta_dir</code>	Diretório com arquivos FASTA contendo as ORFs em aminoácidos para cada seqüência	
<code>nt_fasta_dir</code>	Diretório com arquivos FASTA contendo as ORFs em nucleotídeos para cada seqüência	
<code>xml_programs_dir</code>	Diretórios com arquivos XML para cada componente, que armazena valores de cada <i>link</i> da página.	

4.2 Validação e depuração dos componentes desenvolvidos

Durante o desenvolvimento de cada componente, foram realizados vários testes de execução dos programas e análise dos dados gerados. Cada componente foi validado e depurado, com o intuito de verificar se não apresentava erros de execução. O processo de validação consistiu na elaboração de *pipelines* contendo múltiplas etapas aos quais foram submetidas seqüências conhecidas, cujos resultados de anotação já eram previstos. Os resultados foram analisados em vários níveis. Primeiramente, verificou-se se cada um dos componentes executava sem erros. Muito freqüentemente erros de código resultavam em erros de execução, os quais freqüentemente abortavam o processo inteiro do *pipeline*. Esses erros foram sendo corrigidos um a um, componente por componente. Como alguns desses erros dependiam de uma combinação de fatores como, por exemplo, algum tipo de característica de uma seqüência específica, somente o processamento de um grande número de seqüências pôde garantir que pelo menos a grande maioria das falhas tenha sido eliminada. Neste sentido, foram processadas quase 7.000 seqüências de cDNAs reconstruídos de *E. maxima* e *E. acervulina*, inicialmente para validação e, posteriormente, para se obter a anotação propriamente dita. Posteriormente, foram submetidas para testes e anotação, mais de 11.000 seqüências de *E. tenella*. No final, podemos concluir que o sistema encontra-se bastante robusto.

Outro aspecto analisado foi o da consistência da anotação em termos de conseguir extrair adequadamente todos os dados relevantes a partir dos resultados dos diferentes programas empregados. Além disso, foi verificado se a anotação gerada de forma automática estava correta em termos de conteúdo e de formato, isto é, se os arquivos de relatório nos formatos *features table* e GFF3 estavam totalmente dentro das especificações. Finalmente, os arquivos de anotação de ambos os formatos foram testados no anotador Artemis e os resultados inspecionados.

4.3 Pré-processamento das seqüências de *E. maxima*, *E. acervulina* e *E. tenella*

Todas as leituras ORESTES de *E. maxima*, *E. acervulina* e *E. tenella* geradas em nosso laboratório foram submetidas a um *pipeline* de pré-processamento conforme descrito no item 3.2. As Tabelas 28 e 29 mostram os resultados obtidos para *E. maxima* e *E. acervulina*, respectivamente.

Tabela 28 - Pré-processamento das leituras ORESTES de *E. maxima*.

Discriminação	Nº de seqüências	
Filtro de qualidade	3.118	(14,84%)
Filtro de tamanho	184	(0,88%)
Mitocôndria	5	(0,02%)
Plastídeo	30	(0,14%)
Seqüências repetitivas	50	(0,24%)
Ribossomo	2.056	(9,78%)
Bactéria	1	(0,00%)
Galinha	38	(0,18%)
Humano	39	(0,19%)
Seqüências aceitas	15.490	(73,72%)
Total	21.013	(100,00%)

Como podemos observar na Tabela 28, após o descarte de todas as seqüências de baixa qualidade, tamanho pequeno, ou positivas para algum dos filtros de contaminantes, restaram 15.490 seqüências de *E. maxima*, o que corresponde a cerca de 74% do total de leituras produzidas no projeto. No caso de *E. acervulina* (Tabela 29), sobraram 16.289 seqüências, o que corresponde a aproximadamente 84% do total de seqüências. Para *E. tenella* (Tabela 30), obteve-se um total de 14.123 seqüências boas, representando 80,39% do total. Em todos os três casos, a contaminação foi relativamente baixa, sendo as seqüências de origem ribossômica as de maior ocorrência.

Tabela 29 - Pré-processamento das leituras ORESTES de *E. acervulina*.

Discriminação	Nº de seqüências	
Filtro de qualidade	2.084	(10,69%)
Filtro de tamanho	132	(0,68%)
Mitocôndria	25	(0,13%)
Plastídeo	0	(0,00%)
Seqüências repetitivas	29	(0,15%)
Ribossomo	921	(4,73%)
Bactéria	0	(0,00%)
Galinha	0	(0,00%)
Humano	8	(0,04%)
Seqüências aceitas	16.289	(84,58%)
Total	19.488	(100,00%)

Tabela 30 - Pré-processamento das leituras ORESTES de *E. tenella*.

Discriminação	Nº de seqüências	
Filtro de qualidade	1.916	(10,91%)
Filtro de tamanho	164	(0,93%)
Mitocôndria	92	(0,52%)
Plastídeo	17	(0,10%)
Seqüências repetitivas	10	(0,06%)
Ribossomo	1.209	(6,88%)
Bactéria	0	(0,0%)
Galinha	5	(0,03%)
Humano	32	(0,18%)
Seqüências aceitas	14.123	(80,39%)
Total	17.568	(100,00%)

No caso de *E. tenella*, vários outros grupos internacionais haviam seqüenciado ESTs e tornado estas seqüências disponíveis (ver Tabela 1). A fim de utilizar estas seqüências e inclui-las no processamento de reconstrução de transcritos juntamente com os ORESTES gerados pelo nosso grupo, submetemos essas seqüências a um pipeline semelhante ao empregado para as nossas leituras. Utilizamos inicialmente 42.909 seqüências de outros grupos, obtendo no final 39.764 seqüências limpas as quais, juntamente com as 14.123 leituras ORESTES, totalizaram 53.887 seqüências.

4.4 Reconstrução dos transcritos

Um aspecto importante em relação à reconstrução de transcritos diz respeito à estringência da montagem. Um agrupamento sob baixa estringência pode resultar na junção de seqüências de parálogos ou, pior ainda, de seqüências não relacionadas. Por outro lado, uma estringência alta demais pode vir a diferenciar corretamente os parálogos, porém pode gerar maior fragmentação da informação, resultando em seqüências consenso mais curtas. De fato, não há uma receita ou conjunto de parâmetros universal. A estringência apropriada é muito difícil de ser estimada *a priori*, uma vez que famílias gênicas podem apresentar divergências distintas entre os seus membros devido a diferentes tempos de paralogia. Além disso, as taxas de substituição podem variar entre diferentes famílias.

Em função do exposto acima, decidimos então fazer ensaios preliminares de reconstrução dos transcritos usando-se seqüências de *E. tenella*, e o programa CAP3 com duas versões diferentes e vários parâmetros de estringência. Os ensaios foram executados alterando-se os parâmetros p (limiar de porcentagem de identidade de sobreposição) e o parâmetro y (faixa de corte de pontas). A Tabela 31 apresenta os dados comparativos do número e tamanho médio de *contigs* (seqüências formadas por várias leituras sobrepostas) e *singlets* (compostas de seqüências únicas).

Tabela 31 - Dados comparativos de montagem de cDNAs utilizando-se duas versões do programa CAP3 e diferentes combinações dos parâmetros p e y .

Parâmetros/versão	Nº de contigs	Tam. médio (pb)	Nº de singlets
p80 y100 v122107	4283	798	7430
p85 y100 v122107	4282	795	7601
p90 y100 v122107	4311	788	7997
p95 y100 v122107	4321	777	9252
p90 y 250 v122107	4326	790	7210
p90 y 400 v122107	4318	788	7056
p90 y250 v041505	4315	791	7342

Conforme pode ser visto na Tabela 31, o número de *contigs* e *singlets*, e seus respectivos tamanhos médios estão apresentados, sendo que a combinação de programa e parâmetros utilizada na montagem final dos cDNAs está marcada em negrito. Nota-se que o aumento do valor de p gera um maior número de *singlets* não incorporados em *contigs*. Contudo, o número total de *contigs* e seu tamanho médio são pouco afetados. Por outro lado, o parâmetro y também reduz o número de *singlets*, mas pode levar a um grau excessivo de aparamento de pontas das leituras. As diferenças observadas entre as versões dos programas foram pouco significativas. Finalmente, todas as montagens foram analisadas também quanto à distribuição das seqüências em função do tamanho (dados não mostrados). No final, decidimos empregar os valores de p 90 e y 250 para a reconstrução final dos transcritos. Acreditamos que este conjunto de parâmetros apresenta um bom compromisso entre número de *singlets* e um nível razoavelmente alto de estringência. A montagem final dos transcritos para as três espécies foi feita com o programa CAP3 e os resultados estão apresentados na Tabela 32.

Tabela 32 - Resultados da montagem de transcritos das espécies de *Eimeria* com o programa CAP3.

Espécie	Nº de sequências inicial	Nº de <i>contigs</i>	Nº de <i>singlets</i>	Total de eventos
<i>E. acervulina</i>	16.289	1.233	2.193	3.426
<i>E. maxima</i>	15.490	1.280	2.133	3.413
<i>E. tenella</i>	53.887	4.323	7.207	11.530

4.5 Anotação automática das seqüências reconstruídas de cDNAs de *Eimeria* spp.

Uma vez concluídas as validações de cada um dos componentes, conforme descrito no item 4.2, definimos um *pipeline* final que foi empregado para a anotação das seqüências montadas a partir de ESTs e/ou ORESTES das três espécies de *Eimeria*: *E. acervulina*, *E. maxima* e *E. tenella*. A Figura 10 mostra uma captura de tela obtida do configurador gráfico do EGene, com todas as etapas realizadas:

- *Upload* das seqüências em formato FASTA;
- Busca de ORFs e tradução conceitual;
- Busca de peptídeo sinal com o programa SignalP;
- Busca de domínios transmembranares com o programa TMHMM;
- Busca de domínios transmembranares e peptídeo sinal com o programa Phobius;
- Busca de sítios de ancoragem GPI com o programa DGPI;
- Busca de similaridade por BLASTP das seqüências protéicas contra Uniref 90;
- Busca de domínios protéicos com RPS-BLAST das seqüências protéicas contra a base CDD;
- Busca de domínios protéicos com o programa InterProScan das seqüências protéicas contra múltiplas bases de dados do Interpro;
- Seleção de ORFs únicas para cada seqüência com base em evidências de anotação;
- Geração de relatório em formato *feature table* estendido com etiquetas do programa Artemis;

- Geração de relatório em formato *feature table* para fins de submissão a bancos públicos;
- Geração de relatório em formato GFF3 para intercâmbio de anotação;
- Mapeamento e classificação das seqüências em termos de ontologia gênica;
- Geração dinâmica de páginas *web*.

Figura 10 - Tela do editor gráfico (CoEd) do sistema EGene mostrando o *pipeline* com os componentes de anotação automática (vide ícones na tela) utilizados para processamento das seqüências de *Eimeria* spp.

A primeira anotação foi feita automaticamente para o conjunto de 3.426 seqüências montadas de cDNA de *E. maxima*. Inicialmente, foram anotadas apenas as ORFs codificando para pelo menos 100 resíduos de aminoácidos. Contudo, utilizando-se esse valor, observou-se que somente 1.873 seqüências (54,7%) de um total de 3.426 eventos montados atenderam esse critério e foram anotados. Visando se obter um maior número de seqüências anotadas, submetemos as seqüências não anotadas a um novo *pipeline*, agora alterando inicialmente o valor mínimo do tamanho da ORF para 75 resíduos de aminoácidos. Esse processo foi

posteriormente repetido para as não anotadas usando-se o corte de 50 resíduos de aminoácidos. Conforme pode ser visto na Tabela 33, a redução do tamanho de corte do comprimento das ORF permitiu aumentar muito o número de seqüências anotadas.

Tabela 33 – Resultados de anotação de ORFs de *E. maxima* obtidos usando-se valores de 100, 75 e 50 para o parâmetro de tamanho mínimo da ORF (em resíduos de aminoácidos).

Tamanho mínimo da ORF (aa)	Nº de ORFs anotadas (%)	
100	1.873	(54,67%)
75	692	(20,19%)
50	531	(15,49%)
Total de ORFs anotadas	3.096	(90,35%)
Total de ORFs processadas	3.426	(100,00%)

Analisando-se a Tabela 33, podemos notar que após o processamento com tamanho mínimo de ORF de 100, 75 e 50 aminoácidos, foi possível se obter um total de 3.096 ORFs anotadas (90,35%) em relação ao total de 3.426 *clusters* utilizados na entrada. Essa abordagem de se realizar *pipelines* sucessivos com valores mínimos de ORF decrescentes de 100 até 50 resíduos de aminoácidos permitiu, otimizar a anotação. A razão de não se empregar em um único ciclo de anotação o valor de 50, se deve ao fato de que isso acarretaria um custo computacional enorme. De fato, a utilização de um valor baixo de comprimento de ORFs resultaria em um número muito grande ORFs. Da maneira como o sistema de anotação foi concebido, todas essas ORFs seriam submetidas aos componentes de anotação, aumentando significativamente o processamento. Ao se fazer o processamento das seqüências em várias etapas, conseguiu-se aumentar de um lado a quantidade de potenciais genes anotados e, por outro, não aumentar dramaticamente os recursos computacionais necessários.

Do total de 3.096 ORFs anotadas de *E. maxima*, 2.326 apresentaram produtos protéicos sem similaridades significativas ($E\text{-value} < 10^{-6}$) contra a base de dados UniRef90. Um total de 770 produtos protéicos apresentaram similaridades contra proteínas de função conhecida ou hipotéticas conservadas. Uma tabela contendo a lista total de proteínas anotadas pode ser consultada no endereço http://www.coccidia.icb.usp.br/eimeria/products/Emax_products.html.

Em relação ao mapeamento de termos GO, podemos observar nas Tabelas 34, 35 e 36 as freqüências de seqüências que mapearam para a ontologia processo biológico, componente celular e função molecular, respectivamente.

Tabela 34 - Frequências de termos de ontologia gênica de Processo Biológico das seqüências protéicas codificadas pelos cDNAs reconstruídos de *E. maxima*.

Termo GO - Processo Biológico	Frequência
GO:0008150: biological process	3096 (100.00%)
GO:0008152: metabolic process	271 (8.75%)
GO:0044238: primary metabolic process	219 (7.07%)
GO:0019538: protein metabolic process	139 (4.49%)
GO:0006139: nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	61 (1.97%)
GO:0009058: biosynthetic process	55 (1.78%)
GO:0006464: protein modification process	49 (1.58%)
GO:0006810: transport	47 (1.52%)
GO:0016043: cellular component organization and biogenesis	33 (1.07%)
GO:0006412: translation	33 (1.07%)
GO:0006259: DNA metabolic process	19 (0.61%)
GO:0006091: generation of precursor metabolites and energy	18 (0.58%)
GO:0006950: response to stress	16 (0.52%)
GO:0005975: carbohydrate metabolic process	16 (0.52%)
GO:0009056: catabolic process	15 (0.48%)
GO:0006350: transcription	15 (0.48%)
GO:0006811: ion transport	13 (0.42%)
GO:0006118: electron transport	13 (0.42%)
GO:0006996: organelle organization and biogenesis	12 (0.39%)
GO:0015031: protein transport	11 (0.36%)
GO:0009719: response to endogenous stimulus	10 (0.32%)
GO:0007165: signal transduction	10 (0.32%)
GO:0007154: cell communication	10 (0.32%)
GO:0006519: amino acid and derivative metabolic process	9 (0.29%)
GO:0006629: lipid metabolic process	8 (0.26%)
GO:0009605: response to external stimulus	7 (0.23%)
GO:0019725: cell homeostasis	5 (0.16%)
GO:0044403: symbiosis, encompassing mutualism through parasitism	3 (0.10%)
GO:0009628: response to abiotic stimulus	3 (0.10%)
GO:0007010: cytoskeleton organization and biogenesis	3 (0.10%)
GO:0016032: viral reproduction	2 (0.06%)
GO:0050789: regulation of biological process	1 (0.03%)
GO:0007049: cell cycle	1 (0.03%)
GO:0000003: reproduction	1 (0.03%)
GO:0040029: regulation of gene expression, epigenetic	0 (0.00%)
GO:0040007: growth	0 (0.00%)
GO:0030154: cell differentiation	0 (0.00%)
GO:0019748: secondary metabolic process	0 (0.00%)
GO:0016265: death	0 (0.00%)
GO:0016049: cell growth	0 (0.00%)
GO:0009790: embryonic development	0 (0.00%)
GO:0009653: anatomical structure morphogenesis	0 (0.00%)
GO:0009607: response to biotic stimulus	0 (0.00%)
GO:0008371: obsolete biological process	0 (0.00%)
GO:0008283: cell proliferation	0 (0.00%)
GO:0008219: cell death	0 (0.00%)
GO:0008037: cell recognition	0 (0.00%)
GO:0007610: behavior	0 (0.00%)
GO:0007275: multicellular organismal development	0 (0.00%)
GO:0007267: cell-cell signaling	0 (0.00%)
GO:0007028: cytoplasm organization and biogenesis	0 (0.00%)
GO:0007005: mitochondrion organization and biogenesis	0 (0.00%)

Tabela 35 - Frequência de termos de ontologia gênica de Componente Celular das seqüências protéicas codificadas pelos cDNAs reconstruídos de *E. maxima*.

Termo GO - Componente Celular	Frequência
GO:0005575: cellular_component	3096 (100.00%)
GO:0005623: cell	166 (5.36%)
GO:0005622: intracellular	111 (3.59%)
GO:0043226: organelle	64 (2.07%)
GO:0005737: cytoplasm	47 (1.52%)
GO:0043234: protein complex	30 (0.97%)
GO:0005634: nucleus	28 (0.90%)
GO:0005840: ribosome	22 (0.71%)
GO:0005576: extracellular region	15 (0.48%)
GO:0005856: cytoskeleton	7 (0.23%)
GO:0005694: chromosome	5 (0.16%)
GO:0016023: cytoplasmic membrane-bound vesicle	3 (0.10%)
GO:0005794: Golgi apparatus	3 (0.10%)
GO:0005783: endoplasmic reticulum	3 (0.10%)
GO:0005654: nucleoplasm	3 (0.10%)
GO:0005635: nuclear envelope	2 (0.06%)
GO:0005829: cytosol	1 (0.03%)
GO:0005739: mitochondrion	1 (0.03%)
GO:0030313: cell envelope	0 (0.00%)
GO:0030312: external encapsulating structure	0 (0.00%)
GO:0009579: thylakoid	0 (0.00%)
GO:0009536: plastid	0 (0.00%)
GO:0008370: obsolete cellular component	0 (0.00%)
GO:0005941: unlocalized protein complex	0 (0.00%)
GO:0005929: cilium	0 (0.00%)
GO:0005886: plasma membrane	0 (0.00%)
GO:0005815: microtubule organizing center	0 (0.00%)
GO:0005811: lipid particle	0 (0.00%)
GO:0005777: peroxisome	0 (0.00%)
GO:0005773: vacuole	0 (0.00%)
GO:0005768: endosome	0 (0.00%)
GO:0005764: lysosome	0 (0.00%)
GO:0005730: nucleolus	0 (0.00%)
GO:0005618: cell wall	0 (0.00%)
GO:0005615: extracellular space	0 (0.00%)
GO:0005578: proteinaceous extracellular matrix	0 (0.00%)
GO:0000229: cytoplasmic chromosome	0 (0.00%)
GO:0000228: nuclear chromosome	0 (0.00%)

Tabela 36 - Frequência de termos de ontologia gênica de Função Molecular das seqüências protéicas codificadas pelos cDNAs reconstruídos de *E. maxima*.

Termo GO - Função Molecular	Frequência
GO:0003674: molecular_function	3096 (100.00%)
GO:0005488: binding	297 (9.59%)
GO:0003824: catalytic activity	231 (7.46%)
GO:0000166: nucleotide binding	122 (3.94%)
GO:0003676: nucleic acid binding	95 (3.07%)
GO:0016787: hydrolase activity	89 (2.87%)
GO:0016740: transferase activity	66 (2.13%)
GO:0005515: protein binding	43 (1.39%)
GO:0008233: peptidase activity	37 (1.20%)
GO:0003677: DNA binding	35 (1.13%)
GO:0016301: kinase activity	34 (1.10%)
GO:0004672: protein kinase activity	31 (1.00%)
GO:0005215: transporter activity	26 (0.84%)
GO:0005198: structural molecule activity	26 (0.84%)
GO:0005509: calcium ion binding	21 (0.68%)
GO:0003723: RNA binding	15 (0.48%)
GO:0030234: enzyme regulator activity	12 (0.39%)
GO:0004721: phosphoprotein phosphatase activity	8 (0.26%)
GO:0030528: transcription regulator activity	7 (0.23%)
GO:0005102: receptor binding	7 (0.23%)
GO:0004518: nuclease activity	7 (0.23%)
GO:0009055: electron carrier activity	6 (0.19%)
GO:0004872: receptor activity	5 (0.16%)
GO:0004871: signal transducer activity	5 (0.16%)
GO:0003700: transcription factor activity	5 (0.16%)
GO:0045182: translation regulator activity	4 (0.13%)
GO:0008135: translation factor activity, nucleic acid binding	4 (0.13%)
GO:0003774: motor activity	4 (0.13%)
GO:0030246: carbohydrate binding	2 (0.06%)
GO:0008289: lipid binding	2 (0.06%)
GO:0008092: cytoskeletal protein binding	2 (0.06%)
GO:0003779: actin binding	2 (0.06%)
GO:0016209: antioxidant activity	1 (0.03%)
GO:0005216: ion channel activity	1 (0.03%)
GO:0045735: nutrient reservoir activity	0 (0.00%)
GO:0032791: lead ion binding	0 (0.00%)
GO:0031386: protein tag	0 (0.00%)
GO:0030533: triplet codon-amino acid adaptor activity	0 (0.00%)
GO:0030188: chaperone regulator activity	0 (0.00%)
GO:0019825: oxygen binding	0 (0.00%)
GO:0008369: obsolete molecular function	0 (0.00%)
GO:0005326: neurotransmitter transporter activity	0 (0.00%)
GO:0003682: chromatin binding	0 (0.00%)

Um exemplo de anotação completa em formato *feature table* para um cDNA reconstruído para *E. maxima*, está apresentado na Figura 11. A anotação mostra um conjunto de evidências de similaridade encontradas com o programa BLAST (ex. */inference="similar*

to AA sequence:INSD:Q7PT18”), domínios conservados encontrados pelo RPS-BLAST (ex. *db_xref*=”CDD:63838”) e motivos protéicos com o InterProScan (ex. *inference*=”protein motif:InterPro:IPR012295”), além dos termos GO associados a estes [ex. *GO_function*=binding (GO:0005488)]. Podemos observar com esse exemplo que toda a anotação é centrada em nucleotídeos, sendo que as evidências protéicas estão associadas dentro desse fragmento de DNA (CDS).

Figura 11 - Uma anotação completa em formato *feature table* que inclui evidências do BLAST, RPS-BLAST (contra banco CDD) e InterProScan com os termos de GO.

No caso de *E. acervulina*, usou-se o conjunto de 3.413 seqüências montadas de cDNAs a partir das seqüências de ORESTES. Desse conjunto inicial, 3.233 ORFs foram anotadas (94,72%), sendo que deste total 2.391 produtos protéicos não apresentaram similaridades significativas ($E\text{-value} < 10^{-6}$) e 842 produtos protéicos apresentaram similaridade contra proteínas de função conhecida ou hipotéticas conservadas. Uma tabela contendo a lista total de proteínas anotadas para *E. acervulina* está disponível para consulta no endereço http://www.coccidia.icb.usp.br/eimeria/products/Eace_products.html. Quanto ao mapeamento de termos de ontologia gênica, podemos observar nas Tabelas 37, 38 e 39 a distribuição das seqüências mapeadas nas três ontologias.

Finalmente, foi realizada a anotação final para *E. tenella* utilizando-se o conjunto de 11.530 seqüências de cDNAs, montadas a partir de ESTs e ORESTES (item 4.4 e Tabela 31). Um total de 10.318 seqüências foram anotadas, o que corresponde a 89,48% do conjunto total. Deste total 7.928 produtos protéicos não apresentaram similaridades significativas ($E\text{-value} < 10^{-6}$) e 2.390 tiveram seus produtos protéicos similares a proteínas de função conhecida ou hipotéticas conservadas. Uma lista das proteínas anotadas de *E. tenella* está disponível no endereço http://www.coccidia.icb.usp.br/eimeria/products/Eten_products.html. As Tabelas 40, 41 e 42, apresentam a distribuição das seqüências mapeadas nas três ontologias.

Tabela 37 - Frequência de termos de ontologia gênica de Processo Biológico das seqüências protéicas codificadas pelos cDNAs reconstruídos de *E. acervulina*

Termo GO - Processo Biológico	Frequência
GO:0008150: biological_process	3233 (100.00%)
GO:0008152: metabolic process	298 (9.22%)
GO:0044238: primary metabolic process	256 (7.92%)
GO:0019538: protein metabolic process	159 (4.92%)
GO:0006464: protein modification process	78 (2.41%)
GO:0006139: nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	68 (2.10%)
GO:0006810: transport	66 (2.04%)
GO:0009058: biosynthetic process	49 (1.52%)
GO:0016043: cellular component organization and biogenesis	40 (1.24%)
GO:0006412: translation	30 (0.93%)
GO:0006259: DNA metabolic process	27 (0.84%)
GO:0007165: signal transduction	25 (0.77%)
GO:0007154: cell communication	25 (0.77%)
GO:0006950: response to stress	24 (0.74%)
GO:0005975: carbohydrate metabolic process	22 (0.68%)
GO:0006996: organelle organization and biogenesis	20 (0.62%)
GO:0009056: catabolic process	19 (0.59%)
GO:0006091: generation of precursor metabolites and energy	18 (0.56%)
GO:0006118: electron transport	15 (0.46%)
GO:0015031: protein transport	14 (0.43%)
GO:0007010: cytoskeleton organization and biogenesis	12 (0.37%)
GO:0006811: ion transport	11 (0.34%)
GO:0006629: lipid metabolic process	11 (0.34%)
GO:0050789: regulation of biological process	10 (0.31%)
GO:0006350: transcription	10 (0.31%)
GO:0009605: response to external stimulus	9 (0.28%)
GO:0006519: amino acid and derivative metabolic process	9 (0.28%)
GO:0009719: response to endogenous stimulus	8 (0.25%)
GO:0009628: response to abiotic stimulus	6 (0.19%)
GO:0019725: cell homeostasis	4 (0.12%)
GO:0044403: symbiosis, encompassing mutualism through parasitism	3 (0.09%)
GO:0007049: cell cycle	2 (0.06%)
GO:0009607: response to biotic stimulus	1 (0.03%)
GO:0007005: mitochondrion organization and biogenesis	1 (0.03%)
GO:0040029: regulation of gene expression, epigenetic	0 (0.00%)
GO:0040007: growth	0 (0.00%)
GO:0030154: cell differentiation	0 (0.00%)
GO:0019748: secondary metabolic process	0 (0.00%)
GO:0016265: death	0 (0.00%)
GO:0016049: cell growth	0 (0.00%)
GO:0016032: viral reproduction	0 (0.00%)
GO:0009790: embryonic development	0 (0.00%)
GO:0009653: anatomical structure morphogenesis	0 (0.00%)
GO:0008371: obsolete biological process	0 (0.00%)
GO:0008283: cell proliferation	0 (0.00%)
GO:0008219: cell death	0 (0.00%)
GO:0008037: cell recognition	0 (0.00%)
GO:0007610: behavior	0 (0.00%)
GO:0007275: multicellular organismal development	0 (0.00%)
GO:0007267: cell-cell signaling	0 (0.00%)
GO:0007028: cytoplasm organization and biogenesis	0 (0.00%)
GO:0000003: reproduction	0 (0.00%)

Tabela 38 - Frequência de termos de ontologia gênica de Componente Celular das seqüências protéicas codificadas pelos cDNAs reconstruídos de *E. acervulina*.

Termo GO - Componente Celular	Frequência
GO:0005575: cellular_component	3233 (100.00%)
GO:0005623: cell	187 (5.78%)
GO:0005622: intracellular	118 (3.65%)
GO:0043226: organelle	73 (2.26%)
GO:0005737: cytoplasm	45 (1.39%)
GO:0043234: protein complex	39 (1.21%)
GO:0005634: nucleus	25 (0.77%)
GO:0005856: cytoskeleton	19 (0.59%)
GO:0005840: ribosome	17 (0.53%)
GO:0005576: extracellular region	16 (0.49%)
GO:0005694: chromosome	7 (0.22%)
GO:0005783: endoplasmic reticulum	5 (0.15%)
GO:0005886: plasma membrane	3 (0.09%)
GO:0005739: mitochondrion	3 (0.09%)
GO:0005829: cytosol	1 (0.03%)
GO:0005654: nucleoplasm	1 (0.03%)
GO:0005635: nuclear envelope	1 (0.03%)
GO:0030313: cell envelope	0 (0.00%)
GO:0030312: external encapsulating structure	0 (0.00%)
GO:0016023: cytoplasmic membrane-bound vesicle	0 (0.00%)
GO:0009579: thylakoid	0 (0.00%)
GO:0009536: plastid	0 (0.00%)
GO:0008370: obsolete cellular component	0 (0.00%)
GO:0005941: unlocalized protein complex	0 (0.00%)
GO:0005929: cilium	0 (0.00%)
GO:0005815: microtubule organizing center	0 (0.00%)
GO:0005811: lipid particle	0 (0.00%)
GO:0005794: Golgi apparatus	0 (0.00%)
GO:0005777: peroxisome	0 (0.00%)
GO:0005773: vacuole	0 (0.00%)
GO:0005768: endosome	0 (0.00%)
GO:0005764: lysosome	0 (0.00%)
GO:0005730: nucleolus	0 (0.00%)
GO:0005618: cell wall	0 (0.00%)
GO:0005615: extracellular space	0 (0.00%)
GO:0005578: proteinaceous extracellular matrix	0 (0.00%)
GO:0000229: cytoplasmic chromosome	0 (0.00%)
GO:0000228: nuclear chromosome	0 (0.00%)

Tabela 39 - Frequência de termos de ontologia gênica de Função Molecular das seqüências protéicas codificadas pelos cDNAs reconstruídos de *E. acervulina*.

Termo GO - Função Molecular	Frequência
GO:0003674: molecular_function	3233 (100.00%)
GO:0005488: binding	389 (12.03%)
GO:0003824: catalytic activity	293 (9.06%)
GO:0000166: nucleotide binding	200 (6.19%)
GO:0003676: nucleic acid binding	116 (3.59%)
GO:0016740: transferase activity	103 (3.19%)
GO:0016787: hydrolase activity	86 (2.66%)
GO:0005515: protein binding	65 (2.01%)
GO:0016301: kinase activity	63 (1.95%)
GO:0004672: protein kinase activity	52 (1.61%)
GO:0003677: DNA binding	41 (1.27%)
GO:0008233: peptidase activity	34 (1.05%)
GO:0005215: transporter activity	25 (0.77%)
GO:0005198: structural molecule activity	23 (0.71%)
GO:0005509: calcium ion binding	19 (0.59%)
GO:0003723: RNA binding	16 (0.49%)
GO:0030234: enzyme regulator activity	13 (0.40%)
GO:0005102: receptor binding	10 (0.31%)
GO:0003774: motor activity	10 (0.31%)
GO:0009055: electron carrier activity	8 (0.25%)
GO:0045182: translation regulator activity	7 (0.22%)
GO:0030528: transcription regulator activity	7 (0.22%)
GO:0008135: translation factor activity, nucleic acid binding	7 (0.22%)
GO:0003700: transcription factor activity	6 (0.19%)
GO:0016209: antioxidant activity	4 (0.12%)
GO:0008092: cytoskeletal protein binding	4 (0.12%)
GO:0004721: phosphoprotein phosphatase activity	4 (0.12%)
GO:0003779: actin binding	4 (0.12%)
GO:0008289: lipid binding	3 (0.09%)
GO:0004872: receptor activity	3 (0.09%)
GO:0004871: signal transducer activity	3 (0.09%)
GO:0005326: neurotransmitter transporter activity	2 (0.06%)
GO:0004518: nuclease activity	2 (0.06%)
GO:0030246: carbohydrate binding	1 (0.03%)
GO:0005216: ion channel activity	1 (0.03%)
GO:0003682: chromatin binding	1 (0.03%)
GO:0045735: nutrient reservoir activity	0 (0.00%)
GO:0032791: lead ion binding	0 (0.00%)
GO:0031386: protein tag	0 (0.00%)
GO:0030533: triplet codon-amino acid adaptor activity	0 (0.00%)
GO:0030188: chaperone regulator activity	0 (0.00%)
GO:0019825: oxygen binding	0 (0.00%)
GO:0008369: obsolete molecular function	0 (0.00%)

Tabela 40 - Frequência de termos de ontologia gênica de Processo Biológico das seqüências protéicas codificadas pelos cDNAs reconstruídos de *E. tenella*.

Termo GO - Processo Biológico	Frequência
GO:0008150: biological_process	10318 (100.00%)
GO:0008152: metabolic process	938 (9.09%)
GO:0044238: primary metabolic process	783 (7.59%)
GO:0019538: protein metabolic process	522 (5.06%)
GO:0009058: biosynthetic process	300 (2.91%)
GO:0006412: translation	240 (2.33%)
GO:0006139: nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	193 (1.87%)
GO:0006810: transport	136 (1.32%)
GO:0006464: protein modification process	122 (1.18%)
GO:0016043: cellular component organization and biogenesis	116 (1.12%)
GO:0006091: generation of precursor metabolites and energy	78 (0.76%)
GO:0009056: catabolic process	69 (0.67%)
GO:0005975: carbohydrate metabolic process	65 (0.63%)
GO:0006950: response to stress	64 (0.62%)
GO:0006259: DNA metabolic process	63 (0.61%)
GO:0007154: cell communication	58 (0.56%)
GO:0006118: electron transport	58 (0.56%)
GO:0007165: signal transduction	57 (0.55%)
GO:0006996: organelle organization and biogenesis	55 (0.53%)
GO:0006519: amino acid and derivative metabolic process	52 (0.50%)
GO:0050789: regulation of biological process	40 (0.39%)
GO:0015031: protein transport	33 (0.32%)
GO:0006350: transcription	30 (0.29%)
GO:0006811: ion transport	29 (0.28%)
GO:0009628: response to abiotic stimulus	24 (0.23%)
GO:0009719: response to endogenous stimulus	23 (0.22%)
GO:0006629: lipid metabolic process	23 (0.22%)
GO:0007010: cytoskeleton organization and biogenesis	22 (0.21%)
GO:0019725: cell homeostasis	15 (0.15%)
GO:0009605: response to external stimulus	15 (0.15%)
GO:0044403: symbiosis, encompassing mutualism through parasitism	10 (0.10%)
GO:0040029: regulation of gene expression, epigenetic	3 (0.03%)
GO:0009607: response to biotic stimulus	3 (0.03%)
GO:0030154: cell differentiation	2 (0.02%)
GO:0016265: death	2 (0.02%)
GO:0008219: cell death	2 (0.02%)
GO:0007275: multicellular organismal development	2 (0.02%)
GO:0007049: cell cycle	2 (0.02%)
GO:0019748: secondary metabolic process	1 (0.01%)
GO:0007005: mitochondrion organization and biogenesis	1 (0.01%)
GO:0040007: growth	0 (0.00%)
GO:0016049: cell growth	0 (0.00%)
GO:0016032: viral reproduction	0 (0.00%)
GO:0009790: embryonic development	0 (0.00%)
GO:0009653: anatomical structure morphogenesis	0 (0.00%)
GO:0008371: obsolete biological process	0 (0.00%)
GO:0008283: cell proliferation	0 (0.00%)
GO:0008037: cell recognition	0 (0.00%)
GO:0007610: behavior	0 (0.00%)
GO:0007267: cell-cell signaling	0 (0.00%)

Tabela 41 - Frequência de termos de ontologia gênica de Componente Celular das seqüências protéicas codificadas pelos cDNAs reconstruídos de *E. tenella*.

Termo GO - Componente Celular	Frequência
GO:0005575: cellular_component	10318 (100.00%)
GO:0005623: cell	665 (6.45%)
GO:0005622: intracellular	513 (4.97%)
GO:0043226: organelle	331 (3.21%)
GO:0005737: cytoplasm	297 (2.88%)
GO:0005840: ribosome	186 (1.80%)
GO:0043234: protein complex	144 (1.40%)
GO:0005634: nucleus	71 (0.69%)
GO:0005576: extracellular region	34 (0.33%)
GO:0005856: cytoskeleton	30 (0.29%)
GO:0005694: chromosome	24 (0.23%)
GO:0005739: mitochondrion	17 (0.16%)
GO:0005886: plasma membrane	13 (0.13%)
GO:0005829: cytosol	12 (0.12%)
GO:0005783: endoplasmic reticulum	11 (0.11%)
GO:0005794: Golgi apparatus	4 (0.04%)
GO:0005654: nucleoplasm	4 (0.04%)
GO:0005635: nuclear envelope	4 (0.04%)
GO:0005777: peroxisome	3 (0.03%)
GO:0005773: vacuole	3 (0.03%)
GO:0016023: cytoplasmic membrane-bound vesicle	2 (0.02%)
GO:0005764: lysosome	2 (0.02%)
GO:0005815: microtubule organizing center	1 (0.01%)
GO:0005730: nucleolus	1 (0.01%)
GO:0005615: extracellular space	1 (0.01%)
GO:0005578: proteinaceous extracellular matrix	1 (0.01%)
GO:0030313: cell envelope	0 (0.00%)
GO:0030312: external encapsulating structure	0 (0.00%)
GO:0009579: thylakoid	0 (0.00%)
GO:0009536: plastid	0 (0.00%)
GO:0008370: obsolete cellular component	0 (0.00%)
GO:0005941: unlocalized protein complex	0 (0.00%)
GO:0005929: cilium	0 (0.00%)
GO:0005811: lipid particle	0 (0.00%)
GO:0005768: endosome	0 (0.00%)
GO:0005618: cell wall	0 (0.00%)
GO:0000229: cytoplasmic chromosome	0 (0.00%)
GO:0000228: nuclear chromosome	0 (0.00%)

Tabela 42 - Frequência de termos de ontologia gênica de Função Molecular das seqüências protéicas codificadas pelos cDNAs reconstruídos de *E. tenella*.

Termo GO - Função Molecular	Frequência
GO:0003674: molecular_function	10318 (100.00%)
GO:0005488: binding	947 (9.18%)
GO:0003824: catalytic activity	707 (6.85%)
GO:0000166: nucleotide binding	349 (3.38%)
GO:0003676: nucleic acid binding	304 (2.95%)
GO:0016787: hydrolase activity	229 (2.22%)
GO:0005198: structural molecule activity	208 (2.02%)
GO:0005515: protein binding	164 (1.59%)
GO:0016740: transferase activity	151 (1.46%)
GO:0008233: peptidase activity	97 (0.94%)
GO:0003677: DNA binding	88 (0.85%)
GO:0016301: kinase activity	81 (0.79%)
GO:0005215: transporter activity	61 (0.59%)
GO:0004672: protein kinase activity	60 (0.58%)
GO:0003723: RNA binding	60 (0.58%)
GO:0005509: calcium ion binding	44 (0.43%)
GO:0030234: enzyme regulator activity	32 (0.31%)
GO:0045182: translation regulator activity	26 (0.25%)
GO:0009055: electron carrier activity	26 (0.25%)
GO:0008135: translation factor activity, nucleic acid binding	26 (0.25%)
GO:0004871: signal transducer activity	22 (0.21%)
GO:0004872: receptor activity	21 (0.20%)
GO:0030528: transcription regulator activity	19 (0.18%)
GO:0005102: receptor binding	19 (0.18%)
GO:0003774: motor activity	17 (0.16%)
GO:0004721: phosphoprotein phosphatase activity	11 (0.11%)
GO:0008092: cytoskeletal protein binding	10 (0.10%)
GO:0004518: nuclease activity	10 (0.10%)
GO:0003779: actin binding	10 (0.10%)
GO:0003700: transcription factor activity	10 (0.10%)
GO:0016209: antioxidant activity	8 (0.08%)
GO:0003682: chromatin binding	8 (0.08%)
GO:0005216: ion channel activity	6 (0.06%)
GO:0008289: lipid binding	5 (0.05%)
GO:0030246: carbohydrate binding	2 (0.02%)
GO:0030188: chaperone regulator activity	1 (0.01%)
GO:0005326: neurotransmitter transporter activity	1 (0.01%)
GO:0045735: nutrient reservoir activity	0 (0.00%)
GO:0032791: lead ion binding	0 (0.00%)
GO:0031386: protein tag	0 (0.00%)
GO:0030533: triplet codon-amino acid adaptor activity	0 (0.00%)
GO:0019825: oxygen binding	0 (0.00%)
GO:0008369: obsolete molecular function	0 (0.00%)

4.6 Construção de páginas web

Utilizando-se o componente `report_html.pl`, foram construídas de forma automática páginas web contendo os resultados de anotação de cada um dos transcritos reconstruídos de *Eimeria* spp. Como o sistema de anotação ainda não está integrado a uma base de dados relacional, o usuário deve selecionar o transcrito desejado através de páginas índices. Uma vez selecionado o transcrito, abre-se uma página (Figura 12) na qual são apresentados *links* para as seqüências nucleotídicas e protéicas em formato FASTA, as anotações nos formatos *feature table* (nas versões para Artemis e para submissão a bancos internacionais) e GFF. Essas anotações estão disponíveis com e sem a seleção da ORF verdadeira (vide item 4.1.16). Além disso, a página contém *links* para todas as evidências coletadas, especificamente para cada uma das ORFs encontradas. Esse tipo de visualização é fundamental para permitir o processo de curadoria de anotação. Assim, o usuário pode abrir e editar a anotação automática no Artemis e basear sua curadoria a partir do conjunto de evidências coletadas, cujos *links* estão todos agrupados em uma única página web. O endereço de acesso às páginas índices de cada espécie de *Eimeria* é <http://www.coccidia.icb.usp.br/eimeria/annotation/>.

Para integrar todas as informações do projeto, criamos uma página central no endereço <http://www.coccidia.icb.usp.br/eimeria/services/>. Essa página apresenta *links* às páginas índices mencionadas acima, bem como a páginas contendo os mapeamentos de termos GO (ver item 4.1.17) e às listas de nomes de produtos, extraídas a partir dos primeiros *hits* de BLAST de cada produto gênico. Além disso, está também disponível uma versão de BLAST local na qual o usuário pode fazer buscas de similaridade contra as seqüências montadas de cDNAs e os produtos protéicos de todas as três espécies de *Eimeria*, além de diferentes versões de montagens do genoma de *E. tenella*, feitas pelo Instituto Sanger. Assim, um usuário externo pode utilizar uma seqüência consulta para fazer buscas de similaridade contra nossos bancos, identificar a seqüência montada de interesse e a partir daí navegar pelas respectivas páginas de anotação. Alternativamente, o usuário pode entrar nas páginas de mapeamento de termos GO, selecionar a seqüência de interesse de acordo com o(s) termo(s) GO e consultar as páginas de anotação. Finalmente, o usuário pode ainda consultar as páginas de listas de produtos, procurar se há seqüências codificando para um produto específico e dirigir-se à respectiva página de anotação. Essa estrutura de apresentação de dados, embora ainda não integrada com uma base relacional de dados, já oferece a usuários externos uma maneira rápida e eficaz de obter informações de seqüências específicas. No futuro pretende-se completar a construção de uma base relacional de dados que permitirá a realização de

consultas complexas e relacionais, e a geração dinâmica de páginas *web* contendo *links* para as anotações e evidências respectivas. Esse trabalho está sendo conduzido pelo Prof. Alan M. Durham (IME-USP).

Emaxima EGene annotations						
Sequences						
<ul style="list-style-type: none"> FASTA sequence: Emax_2903 Without ORF selection: FI, extended FI, GFF3 With ORF selection: FI, extended FI, GFF3 						
Sequences						
<ul style="list-style-type: none"> Emax_2903_ORFs: DNA, Protein 						
Evidences						
Program	ORF1	ORF2	ORF3	ORF4	ORF5	ORF6
SignalP	No hits	No hits	No hits	No hits	No hits	No hits
TMHMM	No hits	No hits	No hits	2 probable transmembrane helices	No hits	2 probable transmembrane helices
Phobius	No hits	No hits	No hits	1 probable transmembrane helices	No hits	1 probable transmembrane helices
DGPI	No hits	Potential anchor cleavage site	No hits	No hits	No hits	Potential anchor cleavage site
BLAST	No hits	No hits	No hits	No hits	No hits	No hits
RPS-BLAST	No hits	No hits	No hits	No hits	No hits	No hits
InterProScan	No hits	No hits	IPR001778	No hits	IPR001007	No hits
GO terms	No hits	No hits	type I hypersensitivity (GO:0016068)	No hits	No hits	No hits

Figura 12 - Tela capturada de interface web de uma página de anotação produzida automaticamente pelo componente `report_html.pl`. Os *links* para as anotações nos diferentes formatos, bem como para todas as evidências coletadas para cada ORF específica, estão mostrados.

DISCUSSÃO

5 DISCUSSÃO

5.1 Anotação automática

O processo de desenvolvimento de um sistema de anotação automática envolveu várias etapas, desde o levantamento dos padrões de anotação existentes até a construção de componentes capazes de gerar anotações em conformidade total com estes padrões. Durante este desenvolvimento, ficou evidente que os padrões de anotação atualmente existentes são extremamente limitados, e apenas permitem a descrição adequada de um pequeno conjunto de características. Assim, a primeira limitação diz respeito ao fato que os bancos internacionais aceitam anotações centradas exclusivamente nos ácidos nucleicos (DNA e RNA), e não nas proteínas. Embora a descrição de uma proteína faça sentido no contexto de se representar um produto codificado por uma CDS (região codificadora), torna-se impossível definir as características inerentes a esta proteína dentro de um intervalo de coordenadas referente ao ácido nucleico. Assim, se, por exemplo, se uma dada sequência possui uma CDS entre as coordenadas 328 a 1680, o produto protéico deverá ter 451 resíduos de aminoácidos. Essa descrição pode ser facilmente anotada, uma vez que diz respeito a um intervalo de coordenadas na sequência nucleotídica (328..1680) e uma *feature key* (CDS). Contudo, qualquer característica relativa à sequência protéica terá de ser descrita na forma de uma evidência ou de uma nota. No primeiro caso, o formato de descrição é muito restrito. Por exemplo, a detecção de uma potencial região transmembrantar pelo programa Phobius seria anotada como `/inference="protein motif:Phobius:1.01"`. Essa descrição, que segue o padrão *feature table* não esclarece se a região encontrada é de fato transmembrantar ou se se trata de um peptídeo sinal (também detectável pelo Phobius). Além disso, não é possível definir o intervalo de coordenadas protéicas nas quais essa característica foi encontrada. A única maneira de fazê-lo, que está implementada no EGene, é gerar uma nota na qual se explicita melhor essa descrição:

```
/note="1 probable transmembrane helix predicted by Phobius"
```

Limitações semelhantes ocorrem em relação às demais características que podem ser encontradas em proteínas. Não existe, por exemplo, na especificação de *feature table*, nenhum etiqueta ou qualificador para definir os termos GO encontrados para um produto protéico. Por este motivo, decidimos criar dois tipos de saída em formato *feature table*: uma compatível com os bancos internacionais, gerada pelo componente `report_feature_table_submission.pl` e outra, contendo etiquetas adicionais e

compatível com o programa Artemis, gerada pelo componente `report_feature_table_artemis.pl`. No primeiro caso é feita uma anotação bastante limitada e curta, enquanto no segundo todas as evidências coletadas são reportadas, incluindo os termos GO, servindo-se melhor para uma curadoria manual. Apesar de os bancos internacionais não aceitarem a submissão de uma anotação tão detalhada, as traduções protéicas desses bancos possuem muito frequentemente tais anotações. Isso é feito de forma automática por *scripts* dos próprios bancos. Muito recentemente esses bancos passaram a descrever as evidências referências cruzadas como uma lista no cabeçalho da anotação, o que certamente é muito ruim, visto que não se encontra no corpo da anotação *feature table*. O que se pode concluir a partir dessas limitações e das mudanças frequentes que temos observado na forma de anotação desses bancos, é que o formato *feature table*, embora muito limitado, tem sido ainda usado por uma questão histórica e de compatibilidade. Ao invés de romper com esse padrão obsoleto, os bancos internacionais têm preferido alterar de forma paliativa o *feature table* e restringir a submissão de todo um conjunto de descrições por parte dos usuários. Para piorar esse quadro, os programas de visualização e edição de anotações, embora menos restritivos, também são altamente centrados na anotação de ácidos nucleicos, e não nos produtos protéicos.

Uma alternativa ao padrão *feature table* que tem ganhado espaço como formato de intercâmbio de anotações é o GFF3. Comparado ao *feature table*, o GFF3 é compatível com as normas do consórcio de ontologia de seqüências (SO) e permite uma descrição um pouco mais flexível. Contudo, mesmo este formato é totalmente centrado na anotação de ácidos nucleicos, não permitindo a descrição de intervalos de coordenadas relativas às seqüência dos produtos protéicos e, apesar de representar um formato de intercâmbio de seqüências, não é aceito para fins de submissão em bancos de dados internacionais.

A solução das limitações descritas acima pode ser a utilização de um formato padronizado de XML mais rico e que possibilite a inclusão de novas etiquetas na medida em que se julgue necessário adicioná-las. Uma proposta nesta direção é o GAME XML (*Genome Annotation Markup Elements XML* - <http://xml.coverpages.org/game.html>), um formato que é aceito pelo programa Apollo (Lewis *et al.*, 2002), um visualizador e editor de anotações. Contudo, a proposta ainda carece de uma abordagem mais centrada em proteínas, e a interface do Apollo é bem menos amigável do que a do programa Artemis. Infelizmente, daquilo que é do nosso conhecimento, não existe ainda um padrão de anotação que permita descrever de forma detalhada as características dos produtos protéicos e nem visualizá-las em alguma ferramenta com interface gráfica. O que temos visto em projetos de diferentes centros de seqüenciamento é que cada grupo desenvolve seu esquema de base de dados relacional e a

maioria das ferramentas para curadoria local. Por ocasião da publicação do artigo, a sequência é convertida para *feature table* e submetida aos bancos internacionais, e a maioria das descrições são perdidas. A contrapartida é que os grupos tornam geralmente disponíveis interfaces *web* para que usuários externos possam acessar e consultar as bases relacionais associadas.

Concluindo, o processo de desenvolvimento de ferramentas de anotação automática é extremamente difícil devido ao caráter ainda precário dos atuais padrões impostos pelos bancos de dados internacionais. Esse caráter precário se reflete tanto nas limitações descritivas como também nas freqüentes alterações que esses padrões sofrem. Assim, um *software* que gere uma anotação em formato compatível no presente, pode em questão de semanas ter de ser modificado para se adequar a novos padrões. Ainda assim, o sistema EGene tem uma concepção que minoriza esse problema, pois todas as evidências coletadas são armazenadas no formato XML proprietário. Dessa maneira, uma alteração no formato de anotação implica na alteração de apenas um ou dois componentes, aqueles geradores dos relatórios em formato *feature table* e GFF. Além disso, se um novo formato de anotação vier a se tornar o novo padrão, teremos apenas que criar um novo componente que converta os dados armazenados no XML para esse novo padrão, não invalidando todos os componentes de coleta de evidências.

5.2 EGene e outras plataformas de anotação automática

Ao lado do sistema EGene, várias outras plataformas de anotação automática já foram descritas na literatura. A Tabela 43 apresenta uma lista das principais plataformas e respectivas características. Pode-se notar que apenas os sistemas Genescript (Hudek *et al.*, 2003) e GARSA (Dávila *et al.*, 2005) podem realizar a anotação tanto em seqüências genômicas e ESTs, assim como o EGene. O Genescript possui módulos para BLAST, RepeatMasker e os preditores Genscan, HMMgene, GRAIL-EXP e MZEF. O GARSA, por sua vez, contém, entre outros, componentes para Phred/Phrap, CAP3, Glimmer, BLAST, busca de motivos protéicos com InterPro, ClustalW (para alinhamento múltiplo de seqüências) e programas de análise filogenética como o Phylip e WebLogo. Mesmo sendo ferramentas poderosas para a coleta de evidências, e funcionarem através de *pipelines*, ambos os pacotes apresentam várias limitações. Primeiramente, há disponibilidade de módulos ainda é relativamente pequena, seja para pré-processamento, seja para a anotação propriamente dita. O EGene apresenta uma grande quantidade de componentes já desenvolvidos e descritos, totalizando atualmente cerca de 50 componentes. Além disso, o sistema conta com o CoEd,

um editor gráfico para a construção e edição de *pipelines*. Em segundo lugar, os programas apresentam os resultados da anotação na forma de uma interface *web*, facilitando a visualização das evidências coletadas. Contudo, nenhum dos dois programas gera uma anotação integrada, isto é, quando se geram arquivos de anotação, os mesmos representam a anotação individual de cada programa. No caso do Genescript, por exemplo, o pacote pode gerar diversos formatos de saída de dados como HTML, PDF, FASTA e GFF2, mas de forma individualizada, conforme pode-se observar nos exemplos disponíveis no sítio do programa (<http://www.tcag.ca/genescript/cftr-example-css/html/info.html>). Essa limitação é particularmente importante quando se anota proteínas, uma vez que as descrições das evidências coletadas para um produto protéico têm que ser anotadas no contexto da CDS que o codifica. O EGene consegue realizar essa integração e gerar num único arquivo de anotação as descrições referentes às características do genoma (ex. genes de tRNA, repetições seriadas, predição de genes, etc.), como também dos produtos protéicos.

Quando analisamos outras ferramentas de anotação como annot8r (Schmid e Blaxter, 2008), ESTAnnotator (Wagenblatt *et al.*, 2003), EST Express (Smith *et al.*, 2008) e ESTExplorer (Nagaraj *et al.*, 2007), podemos perceber que estas plataformas são específicas para pré-processamento e anotação de seqüências de ESTs, diferentemente do sistema EGene. Podemos citar algumas limitações também presentes em cada sistema. No caso do software annot8r, foram criados módulos para a busca de termos GO, classificação enzimática (*Enzyme Comission*) e mapeamento dos produtos dentro de vias metabólicas da base KEGG (*Kyoto Encyclopedia of Genes and Genomes*). Os dados podem ser armazenados em uma base relacional, mas não há módulos para pré-processamento e montagem de seqüências, nem para uso de qualquer outro programa, o que implica na utilização de outro sistema para realizar tarefas complementares como, por exemplo, busca de motivos protéicos. Além disso, a plataforma não parece ser suficientemente modular para permitir a implementação simples de novos componentes, como no sistema EGene.

No caso da plataforma EST Express (Smith *et al.*, 2008), há módulos para programas de pré-processamento, porém para a coleta de evidências se restringe a componentes para BLAST e GO. Além disso, a arquitetura do software também é bastante restrita e pouco modular. A ferramenta ESTExplorer (Nagaraj *et al.*, 2007) realiza a montagem e a anotação de ESTs utilizando módulos para BLAST, BLAST2GO, ESTScan e InterPro, sendo que seus módulos foram elaborados utilizando tecnologias de código aberto como Zope e Python, que são linguagens de programação que exigem um conhecimento computacional maior para a construção de novos módulos. Podemos notar também, que algumas plataformas como Genescript e EST2uni apresentam uma abordagem diferente para a busca de possíveis regiões

codificadoras (ORFs), pois realizam essa tarefa utilizando preditores de genes como Genscan (Genescript) e ESTScan (EST2uni), sendo que esta abordagem apresenta algumas limitações em relação ao conjunto de treinamento e a taxa de acerto dos preditores. Já o sistema EGene, apresenta um módulo concebido especificamente para a busca de ORFs (`annotation_orf.pl`), baseado na presença de códons de terminação que podem ou não apresentar códon de iniciação e que traduz conceitualmente as seqüências protéicas conforme a tabela de código genético de cada organismo. Todas as seqüências traduzidas são submetidas aos programas de anotação e as evidências de cada uma delas são armazenadas. Finalmente, utilizando-se o componente `report_orf.pl`, pode-se fazer a escolha automática da ORF verdadeira de cada cDNA com base no conjunto de evidências de todas as ORFs. Embora essa abordagem torne o processamento bem mais lento por requerer que as análises sejam feitas em cima de todas as ORFs e não apenas de uma ORF por cDNA, o critério de escolha da melhor ORF é baseado em critérios biológicos. A despeito de utilizar essa abordagem, o EGene permite ainda a utilização convencional de preditores de genes e, para isso, possui módulos para sete programas (ESTScan, Genscan, GlimmerM, GlimmerHMM, SNAP, Phat e Twinscan) que podem ser utilizados paralelamente para a confirmação das possíveis regiões codificadoras encontradas pelo `annotation_orf.pl`.

Concluindo, as plataformas de anotação automática reportadas na literatura constituem de forma geral pacotes prontos contendo um conjunto pré-definido de componentes, sem documentação apropriada e com baixa modularidade. Assim, esses sistemas constituem-se muito mais, em coletores de evidências que, em alguns casos, armazenam os resultados em uma base relacional. Nenhum desses sistemas representa uma plataforma de coleta, interpretação e seleção de evidências, e geração de uma anotação integrada em formatos aceitos por bases de dados públicos. Em termos de configuração e execução, os sistemas são muito mais rígidos do que o EGene, geralmente apresentando *pipelines* prontos que rodam via linha de comando. O EGene, ao contrário, além de permitir a construção de *pipelines* via um arquivo texto e execução na linha de comando, possui um editor gráfico de configuração de *pipelines* (CoEd), o que torna muito fácil e flexível ao usuário a construção de novos *pipelines* personalizados. Além disso, o sistema EGene já apresenta módulos prontos para um grande conjunto de programas, diferentemente das demais plataformas, incluindo redundância de funções. Por exemplo, o usuário pode escolher executar uma montagem com Phrap ou com CAP3, alinhar cDNAs em genomas com Exonerate ou Sim4, buscar motivos protéicos com o HMMER, InterProScan ou RPS-BLAST, buscar regiões transmembranares com o Phobius ou o TMHMM, etc. A construção de novos componentes para programas para os quais ainda não existam módulos também é relativamente simples e amplamente documentada. Finalmente,

nosso sistema se diferencia por realmente gerar uma anotação automática propriamente dita, armazenando todas as evidências coletadas em formatos FASTA, HTML, PNG e texto, e integrando os resultados na forma de relatórios nos formatos padrão de anotação: GFF3 e *Feature Table*.

Tabela 43 - Comparação das características de diferentes tipos de sistemas de anotação de seqüências em relação ao EGene.

Nome do sistema	Descrição	Características	Requerimento para instalação	Referência
EGene	Plataforma de construção de pipelines para pré-processamento e anotação de seqüências genômicas e de ESTs	Saídas HTML, gráficas, tabular, XML, FASTA, GFF3 e <i>Feature Table</i> .	Perl, Bioperl e programas de terceiros	Durham <i>et al.</i> , 2005
Annot8r	Ferramenta para anotação de ESTs utilizando dados de GO, EC e KEGG.	Resultados são apresentados em arquivo texto, separados por vírgulas.	Perl, Bioperl e programas de terceiros.	Schmid e Blaxter, 2008
ESTAnnotator	Ferramenta para anotação de ESTs.	Saídas em formato XML e HTML.	W3H, W2H (HUSAR).	Wagenblatt <i>et al.</i> , 2003
ESTExplorer	Ferramenta para montagem e anotação de ESTs.	Saída com um sumário contendo <i>links</i> para visualização na <i>web</i> .	Zope (v2.8.1), Python (v2.4.3) e MySQL (v4.1.10a)	Nagaraj <i>et al.</i> , 2007
EST Express	Ferramenta de pré-processamento e anotação (BLAST e GO) para ESTs.	Saídas em HTML e tabelas relacionais.	PHP, MySQL e programas de terceiros	Smith <i>et al.</i> , 2008
EST2uni	Software de pré-processamento, clusterização e anotação de ESTs.	Apresenta os dados através de um banco de dados em MySQL, com interface <i>web</i> .	Perl, Apache, MySQL, PHP e Bioperl	Forment <i>et al.</i> , 2008
Genescript	Ferramenta de anotação para seqüências de DNA.	Saídas em formato GFF2, FASTA e HTML.	Perl, Bioperl e programas de terceiros.	Hudek <i>et al.</i> , 2003
GARSA	Ferramenta para pré-processamento e anotação de dados genômicos e ESTs utilizando BLAST, InterPro, CDD e programas de filogenia.	Saídas em formato tabular.	Perl, Bioperl, CGI, Apache, MySQL e programas de terceiros.	Dávila <i>et al.</i> , 2005

5.3 Sistema EGene: características e perspectivas

O processamento de *pipelines* de anotação requer o uso de um conjunto bastante amplo de programas para cada uma das seqüências anotadas, sendo que alguns deles exigem um grande poder computacional. Assim, buscas de similaridade por BLAST contra bases de dados grandes como nr e UniRef podem ser relativamente demoradas. Da mesma forma, a busca de motivos protéicos com o InterProScan contra uma combinação de bases de dados pode requerer vários minutos de processamento pleno. Embora esses dois programas permitam uma execução paralelizada, vários outros programas utilizados para a coleta de evidências ainda não foram concebidos para uma execução paralelizada. Assim, é muito importante que um *pipeline* de anotação seja realizado de maneira a utilizar o máximo de recursos computacionais disponíveis. *Pipes* são implementações do sistema operacional UNIX/Linux que permitem que os dados de saída de um programa sejam diretamente transmitidos como entrada para o processo seguinte e assim por diante, formando um sistema encadeado denominado *pipeline*. A plataforma EGene faz uso intensivo de *pipes* do sistema operacional e, com isso, consegue executar múltiplos processos de forma simultânea. Enquanto, por exemplo, o componente `annotation_orf.pl` está carregando e traduzindo uma seqüência, o componente seguinte está processando a seqüência anterior. Quando o *pipeline* está em plena execução, ele corresponde a uma linha de produção na qual cada componente e seu programa acoplado estão processando alguma seqüência. Essa forma de execução permite utilizar praticamente a totalidade dos recursos do servidor. Se, de um lado, essa característica permite o máximo aproveitamento dos recursos computacionais disponíveis, por outro lado traz como contrapartida potenciais problemas. Assim, todos os componentes do computador passam a trabalhar em sua capacidade máxima, especialmente os processadores e discos rígidos. Isso pode resultar em um incremento considerável da temperatura interna dos componentes. Em nossa experiência, utilizando o *software* SMART (disponível em praticamente todos os discos modernos), que monitora vários parâmetros de funcionamento dos discos rígidos, pudemos notar que a temperatura média de um disco ligado, sem processamento, é de cerca de 30-35°C. Durante a execução de um *pipeline* de milhares de seqüências, que pode demorar algumas semanas de processamento intensivo, temperaturas próximas de 60°C podem ser atingidas, reduzindo enormemente a vida útil dos discos e a confiabilidade da mídia magnética para retenção dos dados. De fato, chegamos a enfrentar algumas panes de disco nos numerosos *pipelines* executados ao longo do trabalho. Para evitar esse tipo de transtorno, recomendamos aos usuários do sistema EGene instalar ventoinhas de ventilação individuais para cada disco, utilizar gabinetes grandes e, se

necessário, instalar uma ou mais ventoinhas adicionais no gabinete para melhorar a ventilação e assim aumentar a eficiência da dissipação de calor.

Outra implicação no uso de encadeamento de processos através de *pipes* do sistema é que se houver uma falha em qualquer um dos componentes, todo o *pipeline* será atingido e o processamento será inteiramente abortado. Isso significa que não deve executar *pipelines* com conjuntos muito grandes de seqüências, visto que qualquer pane levará à perda de todo o processamento até então realizado. Visando contornar este potencial problema, criamos um *script* que divide automaticamente o conjunto de seqüências em grupos de tamanho pré-definido pelo usuário (ex. 100 seqüências). Assim qualquer problema de *software* ou *hardware* leva no máximo à perda de apenas 100 seqüências do lote e não do processamento integral. Posteriormente, os resultados gerados para cada grupo processado são agrupados em saídas únicas através da concatenação dos arquivos XML. O sistema EGene tem se mostrado robusto e flexível para o desenvolvimento de novos componentes de anotação, além de apresentar uma anotação totalmente integrada e nos padrões exigidos atualmente. Considerando-se o volume de seqüências processado no presente trabalho, foram anotados mais de 18.000 cDNAs das três espécies de *Eimeria*. Em nossa experiência, a maioria dos problemas de *software* enfrentados se deveram a problemas de configuração dos vários programas de terceiros utilizados pelo sistema, bem como a erros na configuração dos parâmetros do próprio EGene.

Embora já possamos assumir que o sistema EGene esteja num grau de desenvolvimento bastante maduro em termos de recursos e estabilidade, ainda existem várias limitações que nos permitem planejar futuros aprimoramentos. Assim, apesar do sistema ser modular e permitir a construção de *pipelines* com uma composição e ordem de componentes bastante flexível, o processo de geração de relatórios ainda é muito rígido. Por exemplo, se depois de rodarmos um *pipeline* de anotação com vários componentes decidirmos que mais um componente deverá ser acrescido, o próprio XML do primeiro *pipeline* poderá ser usado no segundo para rodar o componente faltante. Contudo, se a decisão envolver a re-execução de algum componente previamente executado, com outros parâmetros, o gerador de relatórios de anotação (em formatos *feature table* ou GFF3) não permitirá distinguir a anotação feita pelo mesmo componente em duas instâncias distintas. Para resolver isso, no futuro, deverá ser criado um sistema de indexação de etapas de processamento, de tal forma que o usuário possa definir qual dos processos executados pelo componente deverá ser usado para fins de anotação. Outra possibilidade é de se poder expurgar etapas específicas de processamento do XML.

Outro aspecto limitante é o fato que a atual versão do EGene ainda não está acoplada a

um sistema de gerenciamento de banco de dados e, portanto, não permite realizar buscas relacionais. Estamos no momento construindo uma base de dados relacional específica para as necessidades do projeto da *Eimeria*. Usuários externos, podem, entretanto, utilizar o componente `report_html.pl` para gerar automaticamente um conjunto de páginas *web* que permite visualizar as anotações de todas as seqüências em um contexto no qual todas as evidências estão disponíveis na forma de *links* numa mesma página. Embora essa interface não permita buscas relacionais, o usuário pode consultar facilmente a anotação de qualquer seqüência e visualizar todas as evidências disponíveis. Estamos atualmente discutindo a possibilidade de criar um esquema de base de dados relacional genérica para ser utilizado pelo EGene. Contudo, esquemas muito abrangentes como o GUS (The Genome Unified Schema - vide <http://www.gusdb.org/>) são muito complexos e tornam seu uso limitado a grande centros de pesquisa. Por outro lado, esquemas simples podem ser extremamente limitados para serem usados por vários grupos com requerimentos diferentes. Do balanço entre essas duas possibilidades deverá sair a decisão de que tipo de esquema será mais adequado ao EGene.

Finalmente, uma característica ainda não implementada no EGene, e que deverá ser de enorme valia, é a possibilidade de se editar diretamente o XML usando uma interface gráfica de usuário. Na atual implementação, as anotações geradas pelo EGene no formato *feature table* e GFF3 podem ser editadas no programa Artemis (Rutherford *et al.*, 2000) ou, alternativamente, os arquivos GFF3 também podem ser visualizados e anotados no programa Apollo (Lewis *et al.*, 2000), outro *software* para edição de anotações. O problema dessa abordagem é qualquer edição de anotação pode somente ser salva nos respectivos arquivos *feature table* ou GFF3, permanecendo o XML original inalterado. Isso significa que qualquer *script* que utilize o XML para inserção dos dados em uma base de dados estará inserindo a versão original não submetida à curadoria. Estamos considerando a possibilidade de criar uma ferramenta gráfica de visualização e edição de anotações que permita ler e gravar as anotações diretamente no formato XML do EGene.

5.4 Anotação de *Eimeria*

Os componentes de anotação do sistema EGene foram sendo desenvolvidos e testados utilizando-se as seqüências de cDNA do Projeto ORESTES *Eimeria*. Como o conjunto de componentes só ficou pronto recentemente, os dados de anotação de *Eimeria* spp. ainda são preliminares. Os resultados gerados deverão passar por uma inspeção mais cuidadosa e, eventualmente, novas análises serão realizadas. Com a construção do banco de dados relacional, perguntas complexas poderão ser feitas.

Um exemplo de aplicação dos resultados de anotação gerados para as três espécies de *Eimeria*, refere-se à identificação de possíveis antígenos candidatos para compor uma vacina recombinante. Possíveis antígenos imunizantes podem estar representados por proteínas extracelulares, ancoradas na membrana do parasita. Assim, a busca de potenciais antígenos candidatos poderá ser feita buscando-se proteínas contendo uma sequência de peptídeo sinal na porção amino-terminal, para atravessar a membrana, e um sítio de ancoragem via GPI (glicofosfatidilinositol) na extremidade carboxi-terminal. Essas características somente podem ser identificadas em sequências protéicas completas, o que não é necessariamente o caso neste trabalho. De fato, na espécie *E. tenella*, que apresenta um conjunto maior de cDNAs reconstruídos, um número razoável de transcritos apresentaram ORFs completas. No caso de *E. maxima* e *E. acervulina*, o nosso conjunto de leituras é o único atualmente disponível no mundo, e o fato de não cobrir por completo a grande maioria das ORFs se explica por dois motivos. Primeiramente, o número de genes estimado para este parasita é de cerca de 8.000 e, portanto, 15.000 ORESTES para cada espécie, resultando em cerca de 3.400 sequências montadas, não representam nem sequer metade do transcriptoma do organismo. Em segundo lugar, sequências ORESTES sabidamente cobrem preferencialmente a região central dos transcritos (Souza *et al.*, 2000). Enquanto esta característica é positiva para se obter uma boa anotação, visto ser esta a região que contém a maior parte da codificação protéica, por outro lado ela prejudica a cobertura das extremidades dos transcritos e, portanto, das regiões amino- e carboxi-terminais das respectivas proteínas. Em *E. tenella*, a complementaridade de cobertura de ESTs convencionais cobrindo as regiões terminais, e ORESTES cobrindo a região central, garante uma distribuição mais homogênea das leituras ao longo de todo o comprimento das ORFs. Portanto, este estudo deverá ser feito inicialmente em *E. tenella*.

Outro tipo de análise que está sendo realizada em paralelo refere-se à expressão gênica nos diferentes estágios do parasita. No caso de *E. tenella*, entre dados do nosso laboratório e de outros grupos, estão representados vários estágios de desenvolvimento do parasita: oocistos não esporulados, oocistos em fase de esporoblasto, oocistos esporulados, esporozoítos, merozoítos de primeira geração e merozoítos de segunda geração. Analisando-se a composição de leituras componentes dos *contigs* da montagem dos ESTs/ORESTES, é possível se estimar a frequência de leituras por gene de cada um dos estágios. Essa abordagem é conhecida na literatura como análise de *Northern* eletrônico (Schmitt *et al.*, 1999), e baseia-se no número de ESTs identificados em um dado gene para cada tecido/estágio para estimar a expressão gênica diferencial. Visando esse objetivo, foi desenvolvido um *script* que determina as contagens das leituras presentes nos *contigs*, e gera um arquivo em formato CSV (*comma-separated values*), o qual pode ser importado por qualquer planilha. Utilizando-se

uma ferramenta estatística (Kemp - L. Varuzza, A. Gruber e Pereira, C.A.B., manuscrito em preparação), foi possível identificar-se genes diferencialmente expressos nos diferentes estágios de cada espécie. Os resultados deverão ser comparados com dados de SAGE de formas esporozoítas e merozoítas de *E. tenella*, já disponíveis em nosso laboratório. Pretende-se ainda utilizar as seqüências reconstruídas, sem resultados de positivos de BLAST, para se realizar buscas de similaridade utilizando-se o programa tblastx contra bases de dados de ESTs de outros organismos do filo Apicomplexa. Com isso espera-se validar a identificação de genes cuja função seja desconhecida, mas, que por serem conservados em organismos do mesmo filo, devem ser reais e exercer importantes funções nos parasitas.

CONCLUSÕES

6 CONCLUSÕES

- Foram desenvolvidos 28 componentes de anotação automática para o sistema EGene;
- Todos os componentes foram validados e depurados utilizando-se seqüências de *E. maxima* e *E. acervulina* e *E. tenella*;
- Foram padronizados protocolos de anotação através do uso de *pipelines* de anotação automática;
- Foram criados componentes para a geração de relatórios de anotação automática nos formatos *Feature Table* e GFF3;
- Foi desenvolvido um componente para a geração de páginas *web* de forma dinâmica, mostrando as evidências obtidas, bem como os resultados da anotação automática;
- O sistema EGene demonstrou ser capaz de realizar de forma eficiente uma anotação automática “ácido nucléico-dirigida”, integrando as descrições das características protéicas no contexto das ORFs codificadas pelas respectivas seqüências de DNA;
- O sistema EGene demonstrou ser uma plataforma altamente flexível e modular que permite realizar de forma integrada e personalizada o pré-processamento de seqüências de DNA e a anotação automática de seqüências de DNA e proteínas, gerando anotações em formatos internacionalmente aceitos para submissão de seqüências e intercâmbio de anotações;
- Foram processadas, montadas e anotadas seqüências de cDNA das três espécies mais relevantes de *Eimeria* de galinha doméstica: *E. acervulina*, *E. maxima* e *E. tenella*;
- Foi criado um sítio *web* público contendo os índices de genes das três espécies de *Eimeria*.

REFERÊNCIAS BIBLIOGRÁFICAS

REFERÊNCIAS BIBLIOGRÁFICAS*

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 Jun 21;252(5013):1651-6.

Ajioka JW. *Toxoplasma gondii*: ESTs and gene discovery. Int J Parasitol. 1998 Jul; 28(7):1025-31. Review.

Ajioka JW, Boothroyd JC, Brunk BP, Hehl A, Hillier L, Manger ID, Marra M, Overton GC, Roos DS, Wan KL, Waterston R, Sibley LD. Gene discovery by EST sequencing in *Toxoplasma gondii* reveals sequences restricted to the Apicomplexa. Genome Res. 1998 Jan;8(1):18-28.

Allen PC, Fetterer RH. Recent advances in biology and immunobiology of *Eimeria* species and in diagnosis and control of infection with these coccidian parasites of poultry. Clin Microbiol Rev. 2002 Jan;15(1):58-65.

Altschul SF, Madden TL, Schäffer 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 Sep 1;25(17):3389-402.

An Introduction to the Gene Ontology. Disponível em: <http://www.geneontology.org/GO.doc.shtml> [2007 abr 25].

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000 May;25(1):25-9.

Bateman A, Birney E, Durbin R, Eddy SR, Finn RD, Sonnhammer EL. Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. Nucleic Acids Res. 1999 Jan 1;27(1):260-2.

Bendtsen JD, Nielsen H, von Heijne G, Brunak S. Improved prediction of signal peptides: SignalP 3.0. J Mol Biol. 2004 Jul 16;340(4):783-95.

* De acordo com:

International Committee of Medical Journal Editors. Uniform requirements for manuscripts submitted to Biomedical Journal: sample references. Available from: <http://www.icmje.org> [2007 May 22].

Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 1999 Jan 15;27(2):573-80.

Berriman M, Harris M. Annotation of parasite genomes. *Methods Mol Biol.* 2004;270:17-44.

Blake JA, Harris MA. The Gene Ontology (GO) project: structured vocabularies for molecular biology and their application to genome and expression analysis. *Curr Protoc Bioinformatics.* 2008 Sep;Chapter 7:Unit 7.2.

Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. *J Mol Biol.* 1997 Apr 25;268(1):78-94.

Cai X, Fuller AL, McDougald LR, Zhu G. Apicoplast genome of the coccidian *Eimeria tenella*. *Gene.* 2003 Dec 4;321:39-46.

Camargo AA, Samaia HP, Dias-Neto E, Simão DF, Migotto IA, Briones MR, Costa FF, Nagai MA, Verjovski-Almeida S, Zago MA, Andrade LE, Carrer H, El-Dorry HF, Espreafico EM, Habr-Gama A, Giannella-Neto D, Goldman GH, Gruber A, Hackel C, Kimura ET, Maciel RM, Marie SK, Martins EA, Nobrega MP, Paco-Larson ML, Pardini MI, Pereira GG, Pesquero JB, Rodrigues V, Rogatto SR, da Silva ID, Sogayar MC, Sonati MF, Tajara EH, Valentini SR, Alberto FL, Amaral ME, Aneas I, Arnaldi LA, de Assis AM, Bengtson MH, Bergamo NA, Bombonato V, de Camargo ME, Canevari RA, Carraro DM, Cerutti JM, Correa ML, Correa RF, Costa MC, Curcio C, Hokama PO, Ferreira AJ, Furuzawa GK, Gushiken T, Ho PL, Kimura E, Krieger JE, Leite LC, Majumder P, Marins M, Marques ER, Melo AS, Melo MB, Mestriner CA, Miracca EC, Miranda DC, Nascimento AL, Nobrega FG, Ojopi EP, Pandolfi JR, Pessoa LG, Prevedel AC, Rahal P, Rainho CA, Reis EM, Ribeiro ML, da Ros N, de Sa RG, Sales MM, Sant'anna SC, dos Santos ML, da Silva AM, da Silva NP, Silva WA Jr, da Silveira RA, Sousa JF, Stecconi D, Tsukumo F, Valente V, Soares F, Moreira ES, Nunes DN, Correa RG, Zalcberg H, Carvalho AF, Reis LF, Brentani RR, Simpson AJ, de Souza SJ. The contribution of 700,000 ORF sequence tags to the definition of the human transcriptome. *Proc Natl Acad Sci U S A.* 2001 Oct 9;98(21):12103-8.

Cawley SE, Wirth AI, Speed TP. Phat--a gene finding program for *Plasmodium falciparum*. *Mol Biochem Parasitol.* 2001 Dec;118(2):167-74.

Chapman HD, Shirley MW. The Houghton strain of *Eimeria tenella*: a review of the type strain selected for genome sequencing. *Avian Pathol.* 2003 Apr;32(2):115-27.

Chapman HD, Cherry TE, Danforth HD, Richards G, Shirley MW, Williams RB. Sustainable coccidiosis control in poultry production: the role of live vaccines. *Int J Parasitol.* 2002 May;32(5):617-29. Review.

Cornelissen AW, Overdulve JP, van der Ploeg M. Determination of nuclear DNA of five eucoccidian parasites, *Isospora (Toxoplasma) gondii*, *Sarcocystis cruzi*, *Eimeria tenella*, *E. acervulina* and *Plasmodium berghei*, with special reference to gamontogenesis and meiosis in *I. (T.) gondii*. Parasitology. 1984 Jun;88 (Pt 3):531-53.

Current WL, Upton SJ, Long PL. Taxonomy and life cycles. In: Long PL. Coccidiosis of man and domestic animals. Boca Raton: CRC Press; 1990. p. 1-16.

de Souza SJ, Camargo AA, Briones MR, Costa FF, Nagai MA, Verjovski-Almeida S, Zago MA, Andrade LE, Carrer H, El-Dorri HF, Espreafico EM, Habr-Gama A, Giannella-Neto D, Goldman GH, Gruber A, Hackel C, Kimura ET, Maciel RM, Marie SK, Martins EA, Nobrega MP, Paco-Larson ML, Pardini MI, Pereira GG, Pesquero JB, Rodrigues V, Rogatto SR, da Silva ID, Sogayar MC, de Fátima Sonati M, Tajara EH, Valentini SR, Acencio M, Alberto FL, Amaral ME, Aneas I, Bengtson MH, Carraro DM, Carvalho AF, Carvalho LH, Cerutti JM, Corrêa ML, Costa MC, Curcio C, Gushiken T, Ho PL, Kimura E, Leite LC, Maia G, Majumder P, Marins M, Matsukuma A, Melo AS, Mestriner CA, Miracca EC, Miranda DC, Nascimento AN, Nóbrega FG, Ojopi EP, Pandolfi JR, Pessoa LG, Rahal P, Rainho CA, da Rós N, de Sá RG, Sales MM, da Silva NP, Silva TC, da Silva W Jr, Simão DF, Sousa JF, Stecconi D, Tsukumo F, Valente V, Zalcbeg H, Brentani RR, Reis FL, Dias-Neto E, Simpson AJ. Identification of human chromosome 22 transcribed sequences with ORF expressed sequence tags. Proc Natl Acad Sci U S A. 2000 Nov 7;97(23):12690-3.

Dias Neto E, Harrop R, Correa-Oliveira R, Wilson RA, Pena SD, Simpson AJ. Minilibraries constructed from cDNA generated by arbitrarily primed RT-PCR: an alternative to normalized libraries for the generation of ESTs from nanogram quantities of mRNA. Gene. 1997 Feb 20;186(1):135-42.

Dias Neto E, Correa RG, Verjovski-Almeida S, Briones MR, Nagai MA, da Silva W Jr, Zago MA, Bordin S, Costa FF, Goldman GH, Carvalho AF, Matsukuma A, Baia GS, Simpson DH, Brunstein A, de Oliveira PS, Bucher P, Jongeneel CV, O'Hare MJ, Soares F, Brentani RR, Reis LF, de Souza SJ, Simpson AJ. Shotgun sequencing of the human transcriptome with ORF expressed sequence tags. Proc Natl Acad Sci U S A. 2000 Mar 28;97(7):3491-6.

DDBJ/EMBL/GenBank Feature Table documentation. Disponível em: http://www.ebi.ac.uk/embl/Documentation/FT_definitions/feature_table.html [2006 Mar 25].

Durham AM, Kashiwabara AY, Matsunaga FT, Ahagon PH, Rainone F, Varuzza L, Gruber A. EGene: a configurable pipeline generation system for automated sequence analysis. Bioinformatics. 2005 Jun 15;21(12):2812-3. Epub 2005 Apr 6.

Eilbeck K, Lewis SE. Sequence ontology annotation guide. Comp Funct Genomics. 2004;5(8):642-7.

Eilbeck K, Lewis SE, Mungall CJ, Yandell M, Stein L, Durbin R, Ashburner M. The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol.* 2005;6(5):R44. Epub 2005 Apr 29.

Fankhauser N, Mäser P. Identification of GPI anchor attachment signals by a Kohonen self-organizing map. *Bioinformatics.* 2005 May 1;21(9):1846-52. Epub 2005 Feb 2.

Fernandez S, Costa AC, Katsuyama AM, Madeira AM, Gruber A. A survey of the nter- and intraspecific RAPD markers of *Eimeria* spp. of the domestic fowl and the development of reliable diagnostic tools. *Parasitol Res.* 2003 Apr;89(6):437-45. Epub 2002 Dec 24.

Fernandez S, Pagotto AH, Furtado MM, Katsuyama AM, Madeira AM, Gruber A. A multiplex PCR assay for the simultaneous detection and discrimination of the seven *Eimeria* species that infect domestic fowl. *Parasitology.* 2003 Oct;127(Pt 4):317-25.

Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.* 1998 Sep;8(9):967-74.

Gene Ontology Consortium. The Gene Ontology (GO) project in 2006. *Nucleic Acids Res.* 2006 Jan 1;34(Database issue):D322-6.

GO Annotation Guide. Disponível em: <http://www.geneontology.org/GO.annotation.shtml/> [2007 abr 25].

Green P. PHRAP. 1996. Disponível em: <http://www.genome.washington.edu/uwgc/analysisistools/phrap.htm> [2007 abr 25].

Gruber A. Expressed Sequence Tags. In: Dear PH, editor. *Bioinformatics.* USA, Scion Publishing; 2007. p. 141-167.

Huang X, Madan A. CAP3: A DNA sequence assembly program. *Genome Res.* 1999 Sep;9(9):868-77.

Hudek AK, Cheung J, Boright AP, Scherer SW. Genescript: DNA sequence annotation pipeline. *Bioinformatics.* 2003 Jun 12;19(9):1177-8.

Iseli C, Jongeneel CV, Bucher P. ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *Proc Int Conf Intell Syst Mol Biol.* 1999:138-48.

Joyner LP, Long PL. The specific characters of the *Eimeria*, with special reference to the coccidia of the fowl. *Avian Pathol.* 1974 Jul;3(3):145-57.

Käll L, Krogh A, Sonnhammer EL. A combined transmembrane topology and signal peptide prediction method. *J Mol Biol.* 2004 May 14;338(5):1027-36.

Kolpakov R, Bana G, Kucherov G. mreps: Efficient and flexible detection of tandem repeats in DNA. *Nucleic Acids Res.* 2003 Jul 1;31(13):3672-8.

Korf I, Flicek P, Duan D, Brent MR. Integrating genomic homology into gene structure prediction. *Bioinformatics.* 2001;17 Suppl 1:S140-8.

Korf I. Gene finding in novel genomes. *BMC Bioinformatics.* 2004 May 14;5:59.

Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol.* 2001 Jan 19;305(3):567-80.

Lewis SE, Searle SM, Harris N, Gibson M, Lyer V, Richter J, Wiel C, Bayraktaroglu L, Birney E, Crosby MA, Kaminker JS, Matthews BB, Prochnik SE, Smithy CD, Tupy JL, Rubin GM, Misra S, Mungall CJ, Clamp ME. Apollo: a sequence annotation editor. *Genome Biol.* 2002;3(12):RESEARCH0082. Epub 2002 Dec 23. Review.

Ling KH, Rajandream MA, Rivailler P, Ivens A, Yap SJ, Madeira AM, Mungall K, Billington K, Yee WY, Bankier AT, Carroll F, Durham AM, Peters N, Loo SS, Isa MN, Novaes J, Quail M, Rosli R, Nor Shamsudin M, Sobreira TJ, Tivey AR, Wai SF, White S, Wu X, Kerhornou A, Blake D, Mohamed R, Shirley M, Gruber A, Berriman M, Tomley F, Dear PH, Wan KL. Sequencing and analysis of chromosome 1 of *Eimeria tenella* reveals a unique segmental organization. *Genome Res.* 2007 Mar;17(3):311-9. Epub 2007 Feb 6.

Long PL, Millard BJ, Joyner LP, Norton CC. A guide to laboratory techniques used in the study and diagnosis of avian coccidiosis. *Folia Vet Lat.* 1976 Jul-Sep;6(3):201-17.

Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* 1997 Mar 1;25(5):955-64.

Majoros WH, Pertea M, Salzberg SL. TigrScan and GlimmerHMM: two open source *ab initio* eukaryotic gene-finders. *Bioinformatics.* 2004 Nov 1;20(16):2878-9. Epub 2004 May 14.

Manger ID, Hehl A, Parmley S, Sibley LD, Marra M, Hillier L, Waterston R, Boothroyd JC. Expressed sequence tag analysis of the bradyzoite stage of *Toxoplasma gondii*: identification

of developmentally regulated genes. *Infect Immun*. 1998 Apr;66(4):1632-7.

Marchler-Bauer A, Panchenko AR, Shoemaker BA, Thiessen PA, Geer LY, Bryant SH. CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res*. 2002 Jan 1;30(1):281-3.

Miller RT, Christoffels AG, Gopalakrishnan C, Burke J, Ptitsyn AA, Broveak TR, Hide WA. A comprehensive approach to clustering of expressed human gene sequence: the sequence tag alignment and consensus knowledge base. *Genome Res*. 1999 Nov;9(11):1143-55.

Mungall CJ, Misra S, Berman BP, Carlson J, Frise E, Harris N, Marshall B, Shu S, Kaminker JS, Prochnik SE, Smith CD, Smith E, Tupy JL, Wiel C, Rubin GM, Lewis SE. An integrated computational pipeline and database to support whole-genome sequence annotation. *Genome Biol*. 2002;3(12):RESEARCH0081. Epub 2002 Dec 23.

Ng ST, Sanusi Jangi M, Shirley MW, Tomley FM, Wan KL. Comparative EST analyses provide insights into gene expression in two asexual developmental stages of *Eimeria tenella*. *Exp Parasitol*. 2002 Jun-Jul;101(2-3):168-73.

Nunes FM, Valente V, Sousa JF, Cunha MA, Pinheiro DG, Maia RM, Araujo DD, Costa MC, Martins WK, Carvalho AF, Monesi N, Nascimento AM, Peixoto PM, Silva MF, Ramos RG, Reis LF, Dias-Neto E, Souza SJ, Simpson AJ, Zago MA, Soares AE, Bitondi MM, Espreadico EM, Espindola FS, Paco-Larson ML, Simoes ZL, Hartfelder K, Silva WA Jr. The use of Open Reading frame ESTs (ORESTES) for analysis of the honey bee transcriptome. *BMC Genomics*. 2004 Nov 3;5(1):84.

Parisi V, De Fonzo V, Aluffi-Pentini F. STRING: finding tandem repeats in DNA sequences. *Bioinformatics*. 2003 Sep 22;19(14):1733-8.

Pertea M, Salzberg SL. Using GlimmerM to find genes in eukaryotic genomes. *Curr Protoc Bioinformatics*. 2002 Nov;Chapter 4:Unit 4.4.

Pertea G, Huang X, Liang F, Antonescu V, Sultana R, Karamycheva S, Lee Y, White J, Cheung F, Parvizi B, Tsai J, Quackenbush J. TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics*. 2003 Mar 22;19(5):651-2.

Porcel BM, Tran AN, Tammi M, Nyarady Z, Rydåker M, Urmenyi TP, Rondinelli E, Pettersson U, Andersson B, Aslund L. Gene survey of the pathogenic protozoan *Trypanosoma cruzi*. *Genome Res*. 2000 Aug;10(8):1103-7.

Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R. InterProScan: protein domains identifier. *Nucleic Acids Res*. 2005 Jul 1;33(Web Server issue):W116-20.

Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B. Artemis: sequence visualization and annotation. *Bioinformatics*. 2000 Oct;16(10):944-5.

Ruff MD. Important parasites in poultry production systems. *Vet Parasitol*. 1999 Aug 1;84(3-4):337-47.

Shirley MW, Jeffers TK, Long PL. Studies to determine the taxonomic status of *Eimeria mitis*, Tyzzer 1929 and *E. mivati*, Edgar and Seibold 1964. *Parasitology*. 1983 Oct;87 (Pt 2):185-98.

Shirley MW, Kemp DJ, Pallister J, Prowse SJ. A molecular karyotype of *Eimeria tenella* as revealed by contour-clamped homogeneous electric field gel electrophoresis. *Mol Biochem Parasitol*. 1990 Jan 15;38(2):169-73.

Shirley MW. The genome of *Eimeria tenella*: further studies on its molecular organisation. *Parasitol Res*. 1994;80(5):366-73.

Shirley MW. *Eimeria* spp. from the chicken: occurrence, identification and genetics. *Acta Vet Hung*. 1997;45(3):331-47.

Shirley MW. The genome of *Eimeria* spp., with special reference to *Eimeria tenella* - a coccidium from the chicken. *Int J Parasitol*. 2000 Apr 10;30(4):485-93.

Shirley MW, Ivens A, Gruber A, Madeira AM, Wan KL, Dear PH, Tomley FM. The *Eimeria* genome projects: a sequence of events. *Trends Parasitol*. 2004 May;20(5):199-201.

Shirley MW, Smith AL, Tomley FM. The biology of avian *Eimeria* with an emphasis on their control by vaccination. *Adv Parasitol*. 2005;60:285-330. Review.

Schmitt AO, Specht T, Beckmann G, Dahl E, Pilarsky CP, Hinzmann B, Rosenthal A. Exhaustive mining of EST libraries for genes differentially expressed in normal and tumour tissues. *Nucleic Acids Res*. 1999 Nov 1;27(21):4251-60.

Schnitzler BE, Thebo PL, Mattsson JG, Tomley FM, Shirley MW. Development of a diagnostic PCR assay for the detection and discrimination of four pathogenic *Eimeria* species of the chicken. *Avian Pathol*. 1998 Oct;27(5):490-7.

Schnitzler BE, Thebo PL, Tomley FM, Ugglä A, Shirley MW. PCR identification of chicken *Eimeria*: a simplified read-out. *Avian Pathol*. 1999 Feb;28(1):89-93.

Slater GS, Birney E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*. 2005 Feb 15;6:31.

Snoeijer CQ, Picchi GF, Dambrós BP, Steindel M, Goldenberg S, Fragoso SP, Lorenzini DM, Grisard EC. *Trypanosoma rangeli* Transcriptome Project: Generation and analysis of expressed sequence tags. *Kinetoplastid Biol Dis*. 2004 May 13;3(1):1.

Stein L. Genome annotation: from sequence to biology. *Nat Rev Genet*. 2001 Jul;2(7):493-503.

Stein L. Generic Feature Format version 3. 2007. Disponível em: <http://www.sequenceontology.org/gff3.shtml> [2007 Set 05].

Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*. 2007 May 15;23(10):1282-8. Epub 2007 Mar 22.

Verjovski-Almeida S, DeMarco R, Martins EA, Guimarães PE, Ojopi EP, Paquola AC, Piazza JP, Nishiyama MY Jr, Kitajima JP, Adamson RE, Ashton PD, Bonaldo MF, Coulson PS, Dillon GP, Farias LP, Gregorio SP, Ho PL, Leite RA, Malaquias LC, Marques RC, Miyasato PA, Nascimento AL, Ohlweiler FP, Reis EM, Ribeiro MA, Sá RG, Stukart GC, Soares MB, Gargioni C, Kawano T, Rodrigues V, Madeira AM, Wilson RA, Menck CF, Setubal JC, Leite LC, Dias-Neto E. Transcriptome analysis of the acoelomate human parasite *Schistosoma mansoni*. *Nat Genet*. 2003 Oct;35(2):148-57. Epub 2003 Sep 14.

Wan KL, Chong SP, Ng ST, Shirley MW, Tomley FM, Jangi MS. A survey of genes in *Eimeria tenella* merozoites by EST sequencing. *Int J Parasitol*. 1999 Dec;29(12):1885-92.

Zdobnov EM, Apweiler R. InterProScan--an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*. 2001 Sep;17(9):847-8.

GLOSSÁRIO

GLOSSÁRIO

Artemis

Ferramenta gráfica desenvolvida em Java no Sanger Institute (Rutherford *et al.*, 2000) para a visualização e edição de anotação de seqüências nucleotídicas. O programa utiliza os formatos *feature table* e GFF3 para a anotação.

BLAST

Basic Local Alignment Search Tools - *software* desenvolvido por Altschul *et al.* (1997) para buscas de similaridade de seqüências de DNA e proteínas através de alinhamentos locais.

CDD (*Conserved Domain Database*)

Base de dados (Marchler-Bauer *et al.*, 2007) que contém uma coleção de matrizes de escores de posição específica (PSSMs) derivadas a partir de alinhamentos múltiplos de seqüências protéicas, em regiões de domínios evolutivamente conservados.

Contigs

Seqüências resultantes da montagem de leituras de DNA contendo sobreposição entre si e que a partir de seu consenso, formam uma região “contígua”. No contexto de ESTs, são as seqüências que foram reconstruídas a partir de várias leituras, ao contrário dos *singlets*, seqüências derivadas a partir de apenas uma leitura.

DDBJ (*DNA Data Bank of Japan*)

Banco público de dados de seqüências de DNA do Japão.

DGPI

Software utilizado para a busca de sítios de ancoragem através de GPI em seqüências de proteínas (Kronegg & Buloz, 1999).

EMBL (*European Molecular Biology Laboratory*)

Originalmente o nome de um laboratório sediado em Heidelberg, na Alemanha. Atualmente, o laboratório europeu é o EBI (*European Bioinformatics Institute* - <http://www.ebi.ac.uk/>), sediado no Genome Campus em Hinxton, Inglaterra. A velha denominação é, contudo, ainda utilizada para designar um banco público de seqüências de DNA e proteínas, a versão européia do NCBI/GenBank americano.

ESTscan

Software que realiza predição de genes em seqüências de DNA (Iseli *et al.*, 1999).

ESTs (*Expressed Sequence Tags*)

São etiquetas de seqüências expressas, ou seja, seqüências obtidas a partir de bibliotecas de cDNAs de um tecido ou de um estágio evolutivo de um organismo, e que permitem caracterizar um subconjunto de genes expressos. São caracterizadas por serem resultantes de uma única leitura de DNA (*single-pass*).

Exonerate

Software para buscas de similaridade, o qual pode ser utilizado para o mapeamento de seqüências de cDNAs em DNA genômico (Slater e Birney, 2005). Entre os vários modelos disponíveis, o programa apresenta um modelamento da intercalação de introns e a presença de sítios aceptores e doadores de *splicing*.

Feature Table

Formato padrão de anotação de seqüências em bancos de dados internacionais como GenBank, DDBJ e EMBL (<http://www.ncbi.nlm.nih.gov/collab/FT/>).

GenBank

Banco de dados público de seqüências nucleotídicas e suas traduções protéicas. É hospedado no NCBI (*National Center for Biotechnology Information* - <http://www.ncbi.nlm.nih.gov>) em Bethesda, Estados Unidos.

Gene Ontology (GO)

Ontologia gênica. Refere-se a um projeto dinâmico envolvendo um consórcio de laboratórios (<http://www.geneontology.org/>) que desenvolvem um vocabulário controlado de termos para produtos protéicos e suas interrelações. Os termos GO estão distribuídos em três ontologias: processo biológico, componente celular e função molecular (Blake e Harris, 2008; Ashburner *et al.*, 2000).

GENSCAN

Software preditor de genes presentes em uma seqüência de DNA (Burge e Karlin, 1997).

GlimmerM

Software preditor de genes presentes em uma seqüência de DNA (Pertea e Salzberg, 2002).

GlimmerHMM

Software preditor de genes presentes em uma seqüência de DNA (Majoros *et al.*, 2004).

HMMER/Pfam

Software (Eddy, 1998) que realiza a busca de motivos protéicos contra um conjunto de perfis de modelos ocultos de Markov (HMMs) na base Pfam (<http://pfam.sanger.ac.uk>; Bateman *et al.*, 1999).

HTML (*HyperText Markup Language*)

É um tipo de linguagem de hipertexto, utilizada para produzir páginas da *Web*.

InterProScan

Software (Zdobnov e Apweiler, 2001; Quevillon *et al.*, 2005) que realiza buscas de domínios protéicos, através da identificação de assinaturas a partir do InterPro (<http://www.ebi.ac.uk/interpro/>), uma coleção contém as seguintes bases de dados: Pfam, PROSITE, PRINTS, ProDom, SMART, TIGRFAMs, PIRSF, SUPERFAMILY, Gene3D, e PANTHER.

MREPS

Software responsável pela identificação *ab initio* de regiões repetitivas seriadas em uma sequência de DNA (Kolpakov *et al.*, 2003).

ORESTES (*Open Reading frame ESTs*)

Sequências obtidas a partir de bibliotecas de cDNAs de um tecido ou de um estágio evolutivo de um organismo, utilizando-se a técnica de RT-PCR com a utilização de *primers* arbitrários. A técnica resulta em fragmentos que cobrem preferencialmente a região central dos transcritos, que é justamente a região que contém a ORF, daí o nome ORF ESTs (Dias Neto *et al.*, 1997; Souza *et al.*, 2000).

ORF (*Open Reading Frame*)

Quadro aberto de leitura. Refere-se a uma região de uma sequência nucleotídica livre de códons de terminação e, portanto, potencialmente codificadora de proteína.

Peptídeo-sinal

Sequência sinalizadora canônica, presente na porção amino-terminal de uma proteína, que possui pelo menos um aminoácido carregado positivamente e uma região altamente hidrofóbica que forma o centro da sequência sinalizadora. Apresenta os sinais que irão dirigir a proteína sintetizada para seu destino final e sofre uma clivagem após o transporte da proteína.

Período pré-patente

Corresponde ao espaço de tempo que decorre entre a infecção do hospedeiro por um parasita e a primeira eliminação de estágios evolutivos do mesmo. Especificamente para *Eimeria*, é o período que decorre entre a ingestão de um oocisto pelo hospedeiro e a eliminação dos primeiros oocistos nas fezes.

Pipeline

Fluxo de processos encadeados através das suas saídas padrão, de forma que a saída de um processo é utilizada como entrada do processo seguinte (Fonte: <http://www.babylon.com/>).

Phat

Software preditor de genes presentes em uma sequência de DNA (Cawley *et al.*, 2001).

Phobius

Software para a busca de regiões transmembranares e de peptídeo sinal em sequências protéicas (Käll *et al.*, 2004).

Predição *ab initio*

Identificação de genes ou outras estruturas com uma abordagem na qual se utiliza um modelo estatístico e um conjunto de regras pré-estabelecidas com base no uso de um conjunto de treinamento. Neste método, não há um conhecimento prévio da localização da estrutura na sequência.

PSSM (*Position Specific Scoring Matrices*)

Matrizes de escores posição-específicos. Cada matriz é uma tabela que representa a variação encontrada nos caracteres de um alinhamento múltiplo de sequências, para um conjunto de sequências definidas. As colunas da tabela correspondem às colunas de um alinhamento, e as linhas representam a frequência com que os caracteres da sequência aparecem em cada coluna. Os valores correspondem a escores *log-odd*, os quais são calculados como log na base 2 da frequência de substituição observada numa dada posição dividida pela frequência de substituição esperada naquela posição. Por exemplo, um escore positivo (>1) sugere que a substituição é favorecida no domínio conservado, ao passo que um escore negativo (<1) sugere que a substituição não é favorecida (Fonte: PSSM Viewer Help – <http://www.ncbi.nlm.nih.gov>).

Reconstrução de transcritos

Processo que compreende a geração de sequências reconstruídas a partir de dados fragmentados de leituras de ESTs. O processo pode ser realizado em duas etapas: o agrupamento de sequências de cDNA que compartilham similaridade entre si (também chamada de *clustering*); a montagem dos fragmentos de cada grupo com programas como CAP3 ou Phrap. Em alguns casos, especialmente quando o número total de leituras é relativamente baixo, o processo pode ser feito diretamente através da montagem.

Regiões repetitivas seriadas (*tandem repeats*)

Duas ou mais cópias contínuas de um padrão de sequência que constitui a unidade repetitiva. O tamanho da unidade repetitiva também é chamada de período. Ex: CAGCAGCAGCAG (unidade repetitiva = CAG, período = 3, número de cópias = 4)

Região transmembranar

Região de uma proteína constituída de uma α -hélice predominantemente hidrofóbica, contendo cerca de 20 a 30 resíduos de aminoácidos, e que atravessa a bicamada fosfolipídica de uma membrana. Proteínas transmembranares apresentam três tipos de regiões: citossólica, extracelular e transmembranar.

RPS-BLAST

Software (Marchler-Bauer *et al.*, 2002) para buscas de domínios conservados utilizando uma sequência protéica consulta contra uma base de dados de PSSMs (matrizes de escores posição-específica).

SignalP

Software para a busca de regiões de peptídeo sinal em sequências de proteínas (Bendtsen *et al.*, 2004).

Singlets

São sequências resultantes de um processo de montagem que não apresentaram sobreposição com nenhuma outra sequência em um processo de montagem. No contexto de ESTs, são os *clusters* (grupos) compostos por apenas uma leitura.

Sítio para âncora GPI

Proteínas secretadas possuem na extremidade amino-terminal uma sequência de aminoácidos (peptídeo sinal) que sinaliza o destino desta proteína para dentro do retículo endoplasmático. Na extremidade carboxi-terminal existe um sítio composto de aminoácidos hidrofóbicos (sítio de ancoragem GPI) que é reconhecido e clivado, e, posteriormente, a extremidade da proteína é ligada covalentemente a um glicolípido do tipo GPI (glicofosfatidilinositol). Após o transporte da proteína pela via de secreção e transferência através de vesículas do aparelho de Golgi, a mesma é alocada na parte externa da membrana plasmática (Fankhauser e Mäser, 2005).

SIM4

Software (Florea *et al.*, 1998) para mapeamento de cDNAs em DNAs genômicos através de análise de similaridade e modelamento da intercalação de introns e a presença de sítios aceptores e doadores de *splicing*.

SNAP

Software preditor de genes presentes em uma sequência de DNA (Korf, 2004).

STRING

Software que realiza a busca *ab initio* de regiões repetitivas seriadas em sequências de DNA (Parisi *et al.*, 2003).

TMHMM

Software baseado no uso de modelos ocultos de Markov (HMM) para a busca de domínios transmembranares em sequências protéicas (Krogh *et al.*, 2001).

TRF (*Tandem Repeats Finder*)

Software para a busca *ab initio* de regiões repetitivas seriadas em seqüências de DNA (Benson, 1999).

tRNAscan-SE

Software para a busca *ab initio* de genes de tRNA em seqüências nucleotídicas (Lowe e Eddy, 1997).

Twinscan

Software preditor de genes presentes em uma seqüência de DNA, que inclui nos seus modelos probabilísticos conhecimentos prévios da homologia existente entre dois genomas relacionados. O programa foi projetado especificamente para a análise seqüências genômicas contendo um número desconhecido de genes (Korf *et al.*, 2001).

XML (eXtensible Markup Language)

Formato padrão de linguagem capaz de descrever diversos tipos de dados com o uso de etiquetas (*tags*) e uma estrutura hierárquica de dados. Seu propósito principal é a facilidade de compartilhamento de informações através da Internet.