LUIZ THIBÉRIO LIRA DINIZ RANGEL

Desenvolvimento da plataforma EGene para anotação funcional e integração com banco de dados: aplicação e validação em transcritos de *Eimeria* spp. de galinha doméstica

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

São Paulo
2011

LUIZ THIBÉRIO LIRA DINIZ RANGEL

Desenvolvimento da plataforma EGene para anotação funcional e integração com banco de dados: aplicação e validação em transcritos de *Eimeria* spp. de galinha doméstica

Dissertação apresentada ao Departamento de Parasitologia do 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.

Versão original

São Paulo
2011

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

Rangel, Luiz Thibério Lira Diniz.
Desenvolvimento da plataforma EGene para anotação funcional e integração com banco de dados: aplicação e validação em transcritos de *Eimeria* spp. de galinha doméstica / Luiz Thibério Lira Diniz Rangel. -- São Paulo, 2011.

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 of EGene platform for functional annotation and database integration: application and validation on transcript sequences of *Eimeria* spp. of domestic fowl.

Descritores: 1. Bioinformática 2. Anotação de sequências 3. Ortologia de sequências 4. Mapeamento de vias metabólicas 5. Transcriptoma 6. *Eimeria* spp. I. Gruber, Arthur II. Universidade de São Paulo. Instituto de Ciências Biomédicas. Programa de Pós-Biologia da Relação Patógeno-Hospedeiro III. Título.

ICB/SBIB0218/2011

UNIVERSIDADE DE SÃO PAULO

INSTITUTO DE CIÊNCIAS BIOMÉDICAS

Candidato(a): Luiz Thibério Lira Diniz Rangel.

Título da dissertação: Desenvolvimento da plataforma EGene para anotação funcional e integração com banco de dados: aplicação e validação em transcritos 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) 3091-8405
e-mail: cep@icb.usp.br

Comissão de Ética em Pesquisa

CERTIFICADO DE ISENÇÃO

Certificamos que o Protocolo CEP-ICB Nº **391/10** referente ao projeto intitulado: "***Desenvolvimento da plataforma EGene para anotação funcional e integração com banco de dados: aplicação e validação em transcritos de Eimeria spp. de galinha doméstica***" sob a responsabilidade de **Luiz Thibério L.D.Rangel,** foi analisado na presente data pela CEUA - COMISSÃO DE ÉTICA NO USO DE ANIMAIS 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, 21 de junho de 2010.

PROF. DR. WOTHAN TAVARES DE LIMA
Coordenador da CEUA - ICB/USP

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

AGRADECIMENTOS

Aos meus pais, mais que a qualquer um (por motivos óbvios).

À Tia Flávia, que tornou possível este mestrado (a casa e a cama foram os menos importantes). Sou muito grato ao Perboyre, que me acolheu de uma maneira que nunca poderei retribuir.

Ao Professor Arthur Gruber. Não só pela orientação, e sim por todos ensinamentos relacionados, ou não, à ciência e pela oportunidade. Também ao Professor Alan Durham, pela paciência e ajuda.

À Andressa Florêncio, que teve extrema importância durante todo este período, me apoiando e ajudando quando precisei, sempre com muito carinho.

À FAPESP (processo 2009/12643-8) e ao CNPq (processo 138100/2009-8), pelo suporte financeiro.

Não irei agradecer a nenhum amigo em particular, pois todos foram muito importantes, alguns mais e outros menos, e sei que terminaria esquecendo alguém. Quero agradecer a todo mundo que teve qualquer tipo de participação durante este período. Seja ajudando ou atrapalhando, o importante é participar!

"I may not have gone where I intended to go,
but I think I have ended up where
I needed to be."

Douglas Adams

RESUMO

Rangel LTLD. Desenvolvimento da plataforma EGene para anotação funcional e integração com banco de dados: aplicação e validação em transcritos de *Eimeria* spp. de galinha doméstica. [dissertação (Mestrado em Ciências)] - Instituto de Ciências Biomédicas, Universidade de São Paulo, São Paulo, 2011.

A coccidiose da galinha doméstica é uma doença de caráter mundial, causada por sete espécies de parasitas protozoários do gênero *Eimeria*. O genoma da espécie modelo, *Eimeria tenella*, apresenta uma complexidade de 55 MB, distribuída em 14 cromossomos. Relativamente poucos estudos foram realizados para desvendar a complexidade do transcriptoma de parasitas do gênero *Eimeria*. O nosso grupo gerou 45.000 leituras do tipo ORESTES (Open Reading Frame Expressed Sequence Tag) das três espécies mais importantes de *Eimeria*: *E. tenella*, *E. maxima* e *E. acervulina*. As leituras de cDNA, cobrindo vários estágios de desenvolvimento, foram montadas para constituir índices de genes, e submetidas a um amplo pipeline de anotação funcional utilizando-se o sistema EGene (Durham et al. – *Bioinformatics* 21: 2812-2813, 2005). No presente trabalho, relatamos o desenvolvimento de alguns componentes da plataforma EGene e sua aplicação na anotação funcional de transcritos reconstruídos de *Eimeria* spp. Componentes específicos para a análise de ortologia, mapeamento de vias metabólicas e integração de dados com o visualizador GBrowse foram desenvolvidos. Além disso, também descrevemos um componente que utiliza os arquivos de montagem de transcritos para a produção de perfis digitais de expressão. As análises de ortologia identificaram genes conservados em diferentes parasitas apicomplexas, bem como genes restritos ao gênero *Eimeria*. Perfis de expressão digital obtidos de contagens de leituras de transcritos montados foram submetidos a uma análise hierárquica de agrupamento. Árvores de distância demonstraram que oocistos não esporulados e esporoblásticos constituem um clado distinto em todas as espécies do parasita, com oocistos esporulados formando um ramo mais externo. Esse último estágio também mostrou uma relação próxima com esporozoítos, enquanto que merozoítos de primeira e segunda gerações são mais próximos entre si do que com esporozoítos. Os perfis foram inequivocamente associados com os distintos estágios de desenvolvimento e mostraram uma forte correlação com a ordem desses estágios no ciclo de vida dos parasitas. Finalmente, apresentamos o portal The *Eimeria* Transcript Database, um sítio web que fornece acesso público a todos os dados de sequenciamento, anotação e análises comparativas. Esperamos que este repositório possa se constituir numa ferramenta útil para a comunidade científica de *Eimeria*, ajudando a definir potenciais candidatos ao desenvolvimento de novas estratégias de controle da coccidiose da galinha doméstica.

Palavras-chave: Bioinformática. Anotação de sequências. Ortologia de sequências. Mapeamento de vias metabólicas. Transcriptoma. *Eimeria* spp.

ABSTRACT

Rangel LTLD. Development of EGene platform for functional annotation and database integration: application and validation on transcript sequences of *Eimeria* spp. of domestic fowl [master thesis (Sciences)] - Instituto de Ciências Biomédicas, Universidade de São Paulo, São Paulo, 2011.

Coccidiosis of the domestic fowl is a worldwide disease caused by seven species of protozoan parasites of the genus *Eimeria*. The genome of the model species, *Eimeria tenella*, presents a complexity of 55 MB distributed in 14 chromosomes. Relatively few studies have been undertaken to unravel the complexity of the transcriptome of *Eimeria* parasites. Our group has generated 45,000 open reading frame expressed sequence tag (ORESTES) of the three most important species: *E. tenella*, *E. maxima* and *E. acervulina*. The cDNA reads, covering several developmental stages, were assembled to constitute gene indices and submitted to a comprehensive functional annotation pipeline using EGene system (Durham et al. – *Bioinformatics* 21: 2812-2813, 2005). In the present work, we report the development of some components for EGene platform and their application in the functional annotation of reconstructed transcripts of *Eimeria* spp. Specific components for orthology analysis, metabolic pathway mapping and data integration with GBrowse have been developed. Also, we describe a component uses transcript assembly files for the generation of digital expression profiles. Orthology analyses have identified genes conserved across different apicomplexan parasites, as well as genes restricted to the genus *Eimeria*. Digital expression profiles obtained from read countings of the assembled transcripts were submitted to a hierarchical clustering analysis. Distance trees showed that unsporulated and sporoblastic oocysts constitute a distinct clade in all species, with sporulated oocysts forming a more external branch. This latter stage also shows a close relationship with sporozoites, whereas first and second generation merozoites are more closely related to each other than to sporozoites. The profiles were unambiguously associated with the distinct developmental stages and strongly correlated with the order of the stages in the parasite life cycle. Finally, we present The *Eimeria* Transcript Database, a website that provides open access to all sequencing data, annotation and comparative analysis. We expect this repository to represent a useful resource to the *Eimeria* scientific community, helping to define potential candidates for the development of new strategies to control coccidiosis of the domestic fowl.

Key words: Bioinformatics. Sequence annotation. Sequence orthology. Metabolic pathway mapping. Transcriptome. *Eimeria* spp.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O levantamento de evidências sobre a função de uma sequência e a identificação de suas características visa contextualizá-la biologicamente. Esta tarefa, contudo, vem tornando-se cada vez mais complexa, dado o enorme crescimento no volume de dados gerado pelas novas plataformas de sequenciamento. A grande quantidade de sequências geradas por estas novas tecnologias exige, portanto, o desenvolvimento de novas ferramentas para sua manipulação, que permitam a análise desses dados da maneira mais prática possível. Essa exigência por novas ferramentas tem levado à construção de sistemas computacionais que permitem o processamento dos dados de sequênciamento em um fluxo contínuo, onde os resultados de uma análise servem como entrada da próxima. Esta estrutura de dados fluindo por processos encadeados recebe o nome de pipeline. 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. Utilizando sistemas de pipelines é possível manipular de forma eficiente uma grande quantidade de dados, seja para o pré-processamento das sequências ou sua 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.

1.1 EGene – uma plataforma 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 o encadeamento de uma série de diferentes componentes de processamento, cuja ordem e composição é completamente customizável 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”, 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 simples, mesmo para programadores inexperientes. O EGene aceita vários formatos de entrada de seqüências, como FASTA, PHD (arquivo de gerado pelo programa Phred), 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

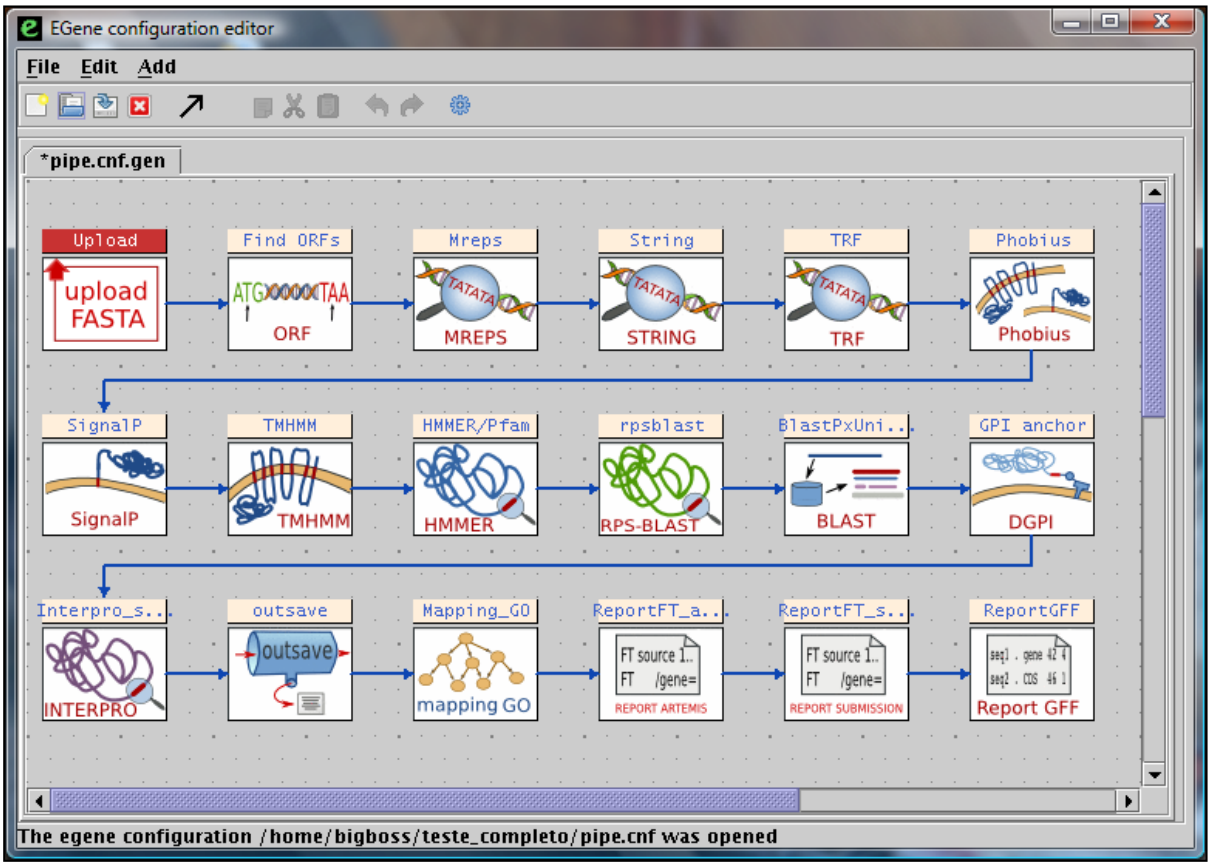
interconectá-los livremente, sem a necessidade de possuir conhecimentos de programação. O sistema pode, ainda, ser facilmente acoplado a um banco de dados relacional. Assim, pode atender as necessidades tanto de usuários avançados, como também usuários de pequenos laboratórios. Além disso, ele é totalmente 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 sequências 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 sequência e resultados obtidos após o 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 segue uma estrutura definida.

O EGene é fornecido juntamente com o CoEd (Figura 1), 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.

A versão 2 do sistema EGene, ainda não disponibilizada publicamente, possui uma ampla gama de componentes de anotação que inclui um programa para a determinação de ORFs com tradutor das sequências protéicas, componentes para sete preditores de genes (Genscan, GlimmerM, GlimmerHMM, Twinscan, Phat, ESTscan e SNAP) , três programas de busca de repetições seriadas (TRF, String e MREPs), tRNAscan-SE, mapeamento de cDNAs (Sim4 e Exonerate), busca de similaridade (Blast), busca de motivos protéicos (HMMER/Pfam, RPS-Blast, InterProScan), busca de domínios transmembranares (TMHMM) e de peptídeo sinal (SignalP) ou ambos (Phobius), busca de sítios de ancoragem GPI (DGPI), mapeamento de termos GO, e geradores de anotação no padrão Feature table (FT) e GFF3. Além disso, há um componente que agrega todas as evidências de anotação numa página web para cada uma das sequências anotadas de forma automática. Assim, partindo de sequências FASTA, o usuário obtém ao final a coleta de um grande conjunto de evidências, arquivos de anotação nos dois formatos mais comuns de intercâmbio de anotação (FT e GFF3) e um sítio web completo para a fácil inspeção dos dados.

Figura 1 - Captura de tela do editor gráfico (CoEd) do sistema EGene, mostrando um exemplo de pipeline com vários componentes de anotação automática (vide ícones na tela) utilizados para processamento de sequências de cDNA.



Fonte: Durham et al., 2005 imagem gerada por Rangel, 2011.

1.2 Anotação funcional de proteínas baseada em ortologia

Como definido por Fitch (1970), dois genes são considerados ortólogos se tiverem divergido a partir de um ancestral comum através de um evento de especiação. Caso a divergência dos genes tenha-se originado em um evento de duplicação, diz-se que os genes são parálogos. Sabe-se que genes ortólogos costumam manter suas funções originais, enquanto que genes parálogos tendem a divergir funcionalmente (Peterson et al., 2009). As diferenças funcionais observadas entre genes parálogos podem ocorrer em diversas taxas e por diferentes razões, entretanto, observa-se que genes relacionados através de um evento de duplicação recente mantêm suas características funcionais originais. Visando diferenciar os genes parálogos de acordo com o período em que houve a duplicação, utiliza-se o termo inparálogos para referir-se a genes cuja duplicação ocorreu depois do evento de especiação,

considerado recente, e outparálogos quando o evento de duplicação ocorreu antes da especiação utilizada como referência, considerada antiga (Remm et al. 2001).

Como discutido por Kristensen et al. (2011), o termo ortologia foi originalmente cunhado para referir-se a relações pareadas, mas na prática refere-se a grupos de genes em diversas espécies e é bastante utilizado em estudos evolutivos de famílias gênicas. Em genômica comparativa, os conjuntos de genes dos diferentes organismos devem ser estudados à luz de suas origens evolutivas. Assim, utilizam-se nesse tipo de estudo grupos de genes ortólogos cujas proteínas têm funções comuns, auxiliando a identificar em organismos pouco caracterizados, a função de seus homólogos respectivos. A anotação de proteínas a partir da inferência de relações de ortologia é bastante difundida, e existem diversas metodologias que permitem agrupar proteínas ortólogas. As duas principais metodologias utilizam árvores filogenéticas e alinhamentos locais recíprocos (Altenhoff e Dessimoz, 2009; Chen et al., 2007; Kristensen et al., 2011). A metodologia baseada em árvores filogenéticas compara a topologia do dendrograma obtido a partir de um conjunto de genes ortólogos, com a árvore filogenética das respectivas espécies. Como exemplos de ferramentas que utilizam esta abordagem podemos citar o RIO (Zmasek e Eddy, 2002) e Orthostrapper (Storm e Sonnhammer, 2002). O método baseado em alinhamentos recíprocos assume que proteínas ortólogas tendem a ser mais similares entre si do que com outras proteínas não ortólogas. As ferramentas que utilizam esta metodologia realizam numa primeira etapa uma busca de similaridade bidirecional entre proteínas de pares de organismos (por exemplo, AxB e BxA – AxC e CxA – BxC e CxB). Os resultados dos alinhamentos são então avaliados quanto aos scores. O melhor hit de cada alinhamento é comparado com o melhor hit do alinhamento recíproco, buscando-se os casos em que ocorrem coincidências. Por exemplo, dadas as proteínas A_1 e B_1, pertencentes aos conjuntos de proteínas dos organismos A e B, respectivamente, pode-se inferir uma relação de ortologia entre elas se as mesmas forem os melhores hits recíprocos em alinhamentos pareados, como segue:

a) B_1 é o melhor hit quando A_1 é alinhada contra a base de proteínas do organismo B;
b) A_1 é o melhor hit quando B_1 é alinhada contra a base de proteínas do organismo A;

O método dos melhores hits bidirecionais (best bidirectional hits), por ser mais rápido do que os métodos baseados em análises filogenéticas e ser mais facilmente automatizado, é o mais frequentemente utilizado. Exemplos de ferramentas que utilizam a abordagem de melhores hits recíprocos são o OrthoMCL (Li et al., 2007) e Inparanoid/MultiParanoid (Alexeyenko et al., 2006; Remm et al., 2001). Existem ainda outras abordagens, menos

difundidas, que utilizam perfis de HMM (modelos ocultos de Markov) (Ebersberger et al., 2009) e conservação de sintenia (Jun et al., 2009).

Entre os principais bancos de dados de proteínas ortólogas podemos citar, o COK/KOG (Tatusov et al., 1997, 2003), orthoMCL-DB (Chen et al., 2006), OMA (Altenhoff et al., 2010; Schneider et al., 2007), eggNOG (Jensen et al., 2008; Muller et al., 2010), KEGG ORTHOLOGY (Kanehisa e Goto, 2000) e InParanoid (Ostlund et al., 2009). Todas estas bases utilizam variações do método de melhores hits bidirecionais para definir os pares de ortólogos. Uma das primeiras bases de dados de ortologia desenvolvida foi o COG (Clusters of Orthologous Groups), a qual inicialmente foi construída a partir de proteínas de algumas poucas bactérias e arqueas, distribuídas em um total de 720 grupos ortólogos. A última versão do COG, atualizado pela última vez em 2003, possui 66 genomas bacterianos divididos em 5.666 grupos ortólogos. A base KOG (euKaryotic Orthologous Groups), por sua vez, compreende proteínas derivadas dos genomas de sete eucariotos: *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Homo sapiens*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe* e *Encephalitozoon cuniculi*. O banco KOG possui 60.759 proteínas distribuídas em 4.852 grupos ortólogos. Mesmo não sendo atualizadas há quase dez anos, as bases COG/KOG ainda são consideradas como referência pela comunidade (Altenhoff e Dessimoz, 2009). O banco de dados eggNOG (evolutionary genealogy of genes: Non-supervised Orthologous Groups) é uma versão estendida do COG/KOG, que abrange 630 espécies, contendo um total de 2.242.035 proteínas, distribuídas por 224.847 grupos ortólogos (Muller et al., 2010). Os grupos ortólogos contidos no eggNOG são, em sua maioria, extensões dos grupos existentes no COG/KOG, contendo proteínas oriundas de um conjunto muito maior de organismos. Além desses grupos derivados do COG/KOG, a base eggNOG contém ainda novos grupos ortólogos obtidos utilizando-se o protocolo de best bidirectional hit de forma automática (Jensen et al., 2008). A base orthoMCL-DB, por sua vez, tem como grande vantagem a utilização do algoritmo de agrupamento de Markov (Markov Cluster Algorithm, MCL) (Enright et al., 2002) para análise dos resultados dos alinhamentos locais realizados entre os conjuntos de proteínas de cada organismo. O MCL é baseado em probabilidade e na teoria dos grafos, e é utilizado para solucionar comparações entre os muitos ortólogos de vários organismos. O OMA (Orthologous MAtrix) é, provavelmente, a base de dados de grupos ortólogos mais abrangente existente, e possui mais de 1.000 organismos em sua última versão (maio de 2011). A base de dados InParanoid, por sua vez, é uma das que contém as relações de homologia mais acuradas (Altenhoff e

Dessimoz, 2009; Chen et al., 2007), mas abrange apenas comparações pareadas entre organismos. O KEGG ORTHOLOGY (KO) é uma base de dados de proteínas ortólogas do KEGG (Ogata et al., 1999). Entre as vantagens dessa base de ortologia, podemos citar que a construção de todos os grupos ortólogos é curada manualmente, e a base é integrada com uma série de outras bases e serviços do KEGG (BRITE, PATHWAYS, KAAS, ENZYME, etc.).

1.3 Mapeamento de proteínas em vias metabólicas

A anotação funcional de uma sequência pode ser abordada utilizando-se diversos aspectos que dependem da contextualização desejada. Estes aspectos podem ser bioquímicos, fisiológicos ou fenotípicos (Friedberg, 2006). O aspecto fisiológico implica na interação de proteínas em vias metabólicas. Existem vários bancos de dados de vias metabólicas, sendo os principais o MetaCyc (Caspi et al., 2006) e o KEGG PATHWAYS (Kanehisa e Goto, 2000; Ogata et al., 1999). A base MetaCyc possui um conjunto não redundante e manualmente curado de vias metabólicas descritas na literatura científica. Esta base contém mais de 1.790 vias metabólicas genéricas descritas em mais de 2.000 espécies, em sua maioria plantas e microrganismos. O banco de dados MetaCyc é ligado à base de dados BioCyc (Caspi et al., 2011), que é uma coleção de vias metabólicas de genomas específicos, para mais de 500 organismos. As vias metabólicas do BioCyc são preditas utilizando-se o Pathway Tool (Karp et al., 2002), o qual usa o MetaCyc como modelo para suas predições. O KEGG PATHWAYS é uma das bases de dados que compõem o KEGG, e possui mapas de vias metabólicas curadas e desenhadas manualmente. Suas vias metabólicas estão classificadas em seis categorias funcionais: metabolismo, processamento de informações genéticas, processamento de informações ambientais, processos celulares, sistemas orgânicos e doenças humanas, e ainda uma sétima categoria, relativa a desenvolvimento de drogas. O KEGG PATHWAYS possui, tanto conjuntos de vias metabólicas de organismos específicos, quanto catálogos de vias metabólicas genéricas, onde seus elementos estão associados às bases de dados KEGG ORTHOLOGY, KEGG ENZYME e KEGG REACTION. A base KEGG PATHWAYS é muito mais utilizada para o mapeamento de proteínas em vias metabólicas do que o banco MetaCyc (Mao et al., 2005; Meyer et al., 2003; Moriya et al., 2007; Schmid e Blaxter, 2008). A principal vantagem do KEGG PATHWAYS é sua integração nativa com os diversos outros serviços e bancos de dados disponibilizados pelo KEGG. Além desta base de dados, o KEGG

disponibiliza ainda o KAAS (KEGG Automatic Annotation Server), um sistema web para anotação automática de proteínas, classificando-as em grupos ortólogos e mapeando-as em vias metabólicas (Moriya et al., 2007).

1.4 *Eimeria* spp. de galinha doméstica

A coccidiose aviária é causada por parasitas do gênero *Eimeria*, e caracteriza-se por uma enterite que causa enormes prejuízos à indústria avícola. Os prejuízos causados podem ser diretos, como o menor ganho de peso em frangos, aumento da mortalidade e aumento de infecções secundárias; ou indiretos, como o custo da utilização de drogas anticoccidianas e/ou vacinas na prevenção da infecção (Shirley et al., 2005). Calcula-se que os gastos mundiais relacionados ao controle desta doença variem de 800 milhões (Allen e Fetterer, 2002) a 3 bilhões de dólares por ano (Shirley et al., 2004).

Os parasitas do gênero *Eimeria* pertencem à classe Coccidia, um grupo de protozoários cujo ciclo de vida pode ser monoxênico, heteroxênico facultativo e heteroxênico obrigatório, e são capazes de infectar uma vasta diversidade de organismos (Tenter et al., 2002). Esta classe inclui, juntamente com os membros do gênero *Eimeria*, os parasitas do gênero *Toxoplasma*, de grande importância médica, e do gênero *Neospora*, muito importante em medicina veterinária. A classe Coccidia encontra-se dentro do filo Apicomplexa, o qual contém mais de 5.000 espécies descritas, todos parasitas intracelulares obrigatórios. Além dos três gêneros já citados, este filo compreende ainda outros parasitas de grande importância médica e médico-veterinária, como *Babesia*, *Theileria*, *Plasmodium* e *Cryptosporidium*. As espécies do filo Apicomplexa caracterizam-se por possuir uma série de organelas que formam o complexo apical, uma estrutura supramolecular envolvida com a adesão do parasita à célula hospedeira, invasão e estabelecimento no vacúolo parasitóforo. Além do complexo apical, os organismos do filo Apicomplexa, com exceção dos gêneros *Cryptosporidium* e *Gregarina*, possuem uma organela denominada apicoplasto, cuja origem deriva de um evento de endossimbiose secundária (Fast et al., 2001; Toso e Omoto, 2007; Zhu et al., 2000).

O gênero *Eimeria* possui mais de 1.700 espécies descritas (Levine, 1988), constituindo um grupo de importantes patógenos de grande relevância em medicina veterinária. O ciclo de vida é monoxênico e os parasitas geralmente colonizam células epiteliais da mucosa intestinal. Esses parasitas podem ser observados em uma grande diversidade de hospedeiros,

como insetos, anelídeos, anfíbios, repteis, aves e mamíferos. Um total de sete espécies infectam a galinha doméstica: *E. acervulina, E. maxima, E. tenella, E. necatrix, E. brunetti, E. praecox* e *E. mitis*. *E. tenella* é a espécie que possui maior prevalência e virulência dentre as causadoras da coccidiose aviária. Além disso, é a única espécie capaz de ser cultivada *in vitro*, sendo portanto, considerada a espécie modelo para o estudo da coccidiose aviária. O genoma de *E. tenella* possui cerca de 55 milhões de pares de bases, distribuídos em 14 cromossomos, com um conteúdo GC de 53% (Chapman e Shirley, 2003; Shirley, 2000; Shirley e Harvey, 2000). O cromossomo 1 de *E. tenella* já foi sequenciado (Ling et al., 2007), e observou-se um curioso padrão onde regiões com alta densidade gênica estão associadas a longas regiões de repetições seriadas, e intercalam-se com regiões de baixa densidade gênica contendo poucas repetições observadas. O cromossomo 1 tem um conteúdo GC de 51,3% e, enquanto as regiões de baixa densidade gênica apresentam uma distribuição homogênea de bases GC, a região de alta densidade possui uma oscilação de distribuição destas bases, refletindo a distribuição de regiões codificantes (Ling et al., 2007). Uma versão rascunho do genoma de *E. tenella* também revelou o mesmo tipo de padrão de segmentação observada no cromossomo 1, sugerindo que essa característica se estenda ao longo de todo o genoma. Esses dados estão disponíveis no Instituto Sanger (Instituto Sanger, 2011). Um projeto para o sequenciamento do genoma de *E. maxima* também está em andamento (Malaysia Genome Institute, 2011).

Além dos dados referentes a sequências genômicas, vários projetos de sequenciamento de transcritos de *E. tenella* foram realizados, tendo sido depositados cerca de 40.000 leituras do tipo EST (Expressed Sequence Tags) em bancos de dados públicos (Li et at., 2003; Miska et al., 2004; Ng et al., 2002; Wan et al., 1999). A grande maioria destas leituras é referente aos estágios de esporozoítos e merozoítos de segunda geração. Nosso grupo também gerou cerca de 45.000 leituras ORESTES (Open Reading Frame Expressed Sequence Tags) de *E. tenella* e de mais duas importantes espécies causadoras da coccidiose aviária: *E. acervulina* e *E. maxima*. Ao nosso conjunto de cDNAs de *E. tenella* com cerca de 14.000 ORESTES, adicionamos os cerca de 40.000 ESTs disponíveis em bancos de dados públicos e realizamos uma montagem dos transcritos. No caso de *E. acervulina* e *E. maxima*, também foram realizadas as montagens dos cDNAs, mas somente das leituras ORESTES do nosso laboratório. Estes conjuntos de leituras foram processados e montados utilizando-se a plataforma EGene, tendo resultado na formação de 3.413 transcritos reconstruídos em *E. acervulina*, 3.426 em *E. maxima* e 8.700 em *E. tenella*. Estes transcritos foram submetidos a um pipeline de anotação automática com o sistema EGene 2. Após a utilização do protocolo

desenvolvido, verificamos que uma grande porção das sequências traduzidas a partir dos transcritos reconstruídos não apresentou similaridade significante com proteínas na base *nr*. Esses dados exemplificam um atual problema encontrado em praticamente todos os projetos de anotação de sequências proteicas: a falta de caracterização funcional da maioria das sequências proteicas geradas nos projetos de sequenciamento em larga escala. A plataforma EGene, mesmo já abrangendo uma grande quantidade de componentes para levantamento de evidências, algumas interpretações da caracterização funcional ainda estão ausentes, como identificação de grupos ortólogos e atuação em vias metabólicas. Outra característica ausente no sistema EGene a integração direta com um visualizador genômico como o GBrowse (Stein et al., 2002), que possibilite a curadoria das sequências anotadas de maneira mais dinâmica e integrada. Diante do exposto acima, decidimos aumentar a abrangência da plataforma EGene 2 através do desenvolvimento de componentes adicionais, e aplicar esses novos componentes para incrementar a anotaçãodos transcritos reconstruídos de *Eimeria* spp.

2 OBJETIVOS

2.1 Objetivos gerais

- Incrementar o escopo de aplicações da plataforma EGene e permitir a anotação funcional de genes com base em ortologia e mapeamento de vias metabólicas;
- Caracterizar o transcriptoma de *Eimeria* spp. de galinha doméstica quanto às suas características funcionais e relações de homologia com organismos do filo Apicomplexa.

2.2 Objetivos específicos

- Desenvolver componentes da plataforma EGene para a identificação de ortologia em diferentes bases (COG/KOG e eggNOG), e mapeamento de proteínas em vias metabólicas do KEGG;
- Desenvolver um componente para a integração de dados de anotação da plataforma EGene com o visualizador Genome Browser;
- Utilizar a plataforma EGene e os novos componentes para a anotação funcional de transcritos de *Eimeria acervulina*, *E. maxima* e *E. tenella*;
- Realizar estudos comparados entre o transcriptoma das três espécies de *Eimeria* e de organismos do filo Apicomplexa através de relações de ortologia.

3 MATERIAL E MÉTODOS

3.1 Desenvolvimento do componente para classificação em grupos ortólogos

O componente para classificação em grupos ortólogos, denominado `annotation_orthology.pl`, foi desenvolvido na linguagem Perl, seguindo os padrões da plataforma EGene. Foram implementados dois algoritmos de classificação de sequências protéicas em grupos ortólogos, baseados nos métodos descritos para as bases de ortologia COG/KOG e eggNOG. Ambas as implementações utilizam o programa BLAST como ferramenta inicial para a inferência de homologia entre a proteína consulta e o banco de ortologia. Utilizando-se as bases COG/KOG, os alinhamento são avaliados quanto ao bitscore, identidade e comprimento da região alinhada de todos os hits do BLAST, de acordo com parâmetros pré-estabelecidos. No caso da base eggNOG avalia-se o bitscore e o comprimento da região alinhada, apenas do primeiro hit do BLAST. Os grupos ortólogos identificados são armazenados no arquivo de anotação do EGene, em formato XML, para serem posteriormente recuperados pelos componentes de relatório e geradores de anotações. Ambos os bancos de dados descrevem funcionalmente seus grupos ortólogos e possuem 25 categorias funcionais amplas.

3.1.1 Classificação de sequências em grupos ortólogos da base KOG

Durante a classificação de proteínas nesta base, cada hit de BLAST é avaliado de acordo com os parâmetros mínimos estabelecidos: bitscore = 28, identidade = 20% e bloco de alinhamento com no mínimo 25 aminoácidos. Estes valores foram definidos empiricamente, através de múltiplos testes comparativos, visando obter a maior correspondência possível com as classificações realizadas pela ferramenta online do KOG (`kognitor`) (National Center for Biotechnology Information, 2011a). Para cada hit que obedece aos parâmetros determinados, mapeia-se o respectivo grupo ortólogo. O primeiro grupo ortológo, identificado em três hits oriundos de organismos distintos, é utilizado para classificar a proteína consulta.

3.1.2 Classificação utilizando a base eggNOG

Pelo fato de ser muito maior e mais representativa que o COG/KOG, o protocolo para classificação de proteínas em grupos ortólogos da base eggNOG avalia apenas o primeiro hit do BLAST. Desta forma, evitando classificações errôneas devido a alinhamentos aleatórios. Caso o hit observado possua um bitscore de pelo menos 60 e um bloco de alinhamento com pelo menos 30 resíduos de aminoácidos, e esteja mapeado em um grupo ortólogo, utiliza-se seu grupo ortólogo para classificar a proteína consulta.

3.1.3 Geração de páginas HTML e relatórios de anotação

Dentro da arquitetura da plataforma EGene, a análise e coleta de evidências são feitas por componentes distintos daqueles que geram relatórios para inspeção pelo usuário. Para permitir a inspeção dos resultados das inferências de ortologia foi desenvolvido o componente `report_orthology.pl`. Dado um conjunto de sequências protéicas submetidas a um pipeline, este componente produz um arquivo em formato HTML contendo uma classificação e quantificação dessas sequências dentro das 25 categorias funcionais. Além disto, este arquivo possui um gráfico do tipo pizza, apresentando a proporção relativa em que essas categorias foram observadas. Além de proporcionar uma visão geral dos resultados, este componente gera um arquivo HTML para cada categoria funcional, o qual lista todas as proteínas que foram classificadas. Esta tabela contém o grupo ortólogo identificado, a descrição deste grupo e o resultado do BLAST da proteína consulta contra o respectivo banco de dados.

A plataforma EGene possui três componentes (`annotation_gff.pl`, `annotation_feature_table_submission.pl` e `annotation_feature_table_artemis.pl`) para a geração de arquivos de anotação, nos formatos GFF3 e Feature table. Para esse último formato, existem dois componentes que produzem dados numa versão resumida, no padrão do GenBank, e outra estendida, que contém features adicionais que podem ser visualizados com o editor/visualizador Artemis (Carver et al., 2008; Rutherford et al. 2000). Para incorporar as evidências obtidas com o `annotation_orthology.pl`, foram acrescentadas rotinas nos três componentes geradores de arquivos de anotação. Esses

componentes passaram a extrair as informações inseridas pelo `annotation_orthology.pl` no XML, e as incorporam nos arquivos de saída em formatos GFF3 e Feature table.

3.2 Mapeamento de proteínas em vias metabólicas do KEGG

Para o mapeamento de proteínas em vias metabólicas do KEGG, foi desenvolvido o componente `annotation_pathways.pl`. Esse programa utiliza como entrada sequências proteicas, e realiza uma busca de similaridade com BLAST contra um banco de dados de proteínas construído dinamicamente a partir da base GENES do KEGG. Atualmente a base GENES do KEGG possui proteínas de cerca de 1.627 organismos, e para evitar associações entre organismos distantes (errôneas), além de diminuir o tempo de processamento computacional, utiliza-se apenas um subconjunto dessa base (Moriya et al., 2007). Esta base de dados, gerada dinamicamente no início do processamento, é composta pelas proteínas de organismos determinados pelo usuário, e permite inferir relações de ortologia. O mapeamento das proteínas em vias metabólicas pode ser dividido em duas etapas. Na primeira, após a busca de similaridade por BLAST da proteína consulta contra a base de dados construída dinamicamente, é feita uma seleção dos resultados positivos. Assim, são selecionados apenas os hits que apresentam um bitscore maior ou igual a 60 e que concomitantemente possuem um score maior ou igual a 85% do score obtido pelo primeiro hit da lista. Tomemos, por exemplo, os hits A, B, C e D, com bitscores de 70, 68, 60 e 55, e scores respectivos de 96, 87, 80 e 70. O hit D é descartado por apresentar um bitscore inferior a 60. O hit C é descartado por ter um score (80) inferior a 85% do score do primeiro hit A (0,85 x 96 = 81,6). No exemplo acima, somente os hits A e B teriam seus grupos ortólogos contabilizados para a classificação da proteína consulta. O grupo ortólogo que possuir a maior soma de *bit*scores é utilizado para classificar a proteína consulta. Na segunda etapa do processo, com o grupo ortólogo definido, utiliza-se seu identificador (número KO) para mapear as vias metabólicas em que suas proteínas atuam. O produto final desse processamento é, portanto, o mapeamento de cada proteína, classificada em um grupo ortólogo, nas respectivas vias metabólicas em que atuam. O resultado desse mapeamento é inserido no arquivo de anotação em formato XML do EGene.

Seguindo o mesmo procedimento empregado no componente de ortologia, foi desenvolvido o componente `report_pathways.pl`, específico para a geração de um

relatório de resultados em formato HTML (HyperText Markup Language). Assim, são geradas tabelas para cada uma das categorias de vias metabólicas do KEGG, num total de sete, e um resumo que demonstra a frequência com que cada uma das categorias foram observadas nos dados analisados. Essa quantificação é representada em uma tabela e um gráfico em formato de pizza. Além destas informações, caso o usuário solicite, o componente conecta-se aos servidores do KEGG através de sua API (Application Program Interface) (Kawashima et al., 2005) utilizando o módulo de `Perl SOAP::Lite`. A conexão com estes servidores permite o download da imagem de uma via metabólica, onde o grupo ortólogo especificado encontra-se destacado em relação aos demais, e de uma página HTML contendo um mapeamento da imagem (image map). Nesta imagem, cada um dos grupos ortólogos representados na figura possui um link para sua página de descrição no KEGG. Finalmente, os três componentes (`annotation_gff.pl`, `annotation_feature_table_submission.pl` e `annotation_feature_table_artemis.pl`) foram implementados com rotinas específicas para a geração dos resultados de mapeamento em vias metabólicas nos formatos GFF3 e Feature table.

3.3 Desenvolvimento de componente para integração do GBrowse

Visando permitir a visualização de resultados de anotação da plataforma EGene no GBrowse (Stein et al., 2002), foi desenvolvido o componente `report_gbrowse.pl`, também escrito na linguagem PERL. Este componente obtém as informações relativas às sequências processadas a partir do arquivo XML do EGene. Com base nessas informações, o componente gera um conjunto de arquivos de anotação no formato GFF3. Embora o EGene 2 tenha um componente para gerar arquivos de anotação no formato GFF3 (`report_gff.pl`), o `report_gbrowse.pl` permite gerar arquivos GFF3 mais adequados e otimizados para a visualização no GBrowse. Além dos arquivos em formato GFF3, o GBrowse requer mais dois arquivos para possibilitar a visualização dos dados: [1] um arquivo de configuração específico para os dados em questão, que especifica quais arquivos GFF3 são usados, a descrição dos dados para a página HTML e as regras para utilização e apresentação desses dados no GBrowse; [2] um arquivo de configuração central do GBrowse, devidamente editado, no qual é inserida a informação referente à especificação do arquivo de configuração específico e a URL na qual os dados devem estar disponíveis.

3.4 Sequências de cDNA de *Eimeria* spp.

Para validar os componentes `annotation_orthology.pl`, `annotation_pathways.pl` e `report_gbrowse.pl`, foram utilizados três conjuntos de proteínas traduzidas a partir dos transcritos reconstruídos de três espécies de *Eimeria*.. Um total de 3.233 proteínas de *E. acervulina*, 3.096 de *E. maxima* e 7.990 de *E. tenella.*

3.5 Criação de grupos ortólogos de proteínas de organismos do filo Apicomplexa

Visando a criação de um banco de proteínas ortólogas de membros do filo Apicomplexa, foram utilizados conjuntos de proteínas preditas a partir do genoma completo dos seguintes organismos: *Toxoplasma gondii*, *Plasmodium falciparum*, *Neospora caninum*, *Cryptosporidium parvum*, *Theileria annulata* e *Babesia bovis*. A Tabela 1 apresenta as informações relativas a esses conjuntos de proteínas. A esses conjuntos de proteínas de Apicomplexa adicionou-se ainda as sequências protéicas traduzidas a partir dos cDNAs reconstruídos de *Eimeria tenella*, *E. maxima* e *E. acervulina* (item 3.4). Estes conjuntos proteícos foram submetidos ao programa InParanoid para a criação de grupos ortólogos entre organismos pareados, utilizando buscas de similaridade bidirecionais com a ferramenta BLAST. As análises pareadas consistiram na comparação de todas as proteínas de um organismo contra elas mesmas, e contra as proteínas do outro organismo (por exemplo, A x A e B x B, seguido por A x B e B x A). Neste tipo de análise, os grupos ortólogos são inicialmente formados pelos pares de proteínas com maior score recíproco. Após a formação dos grupos ortólogos, inicia-se a inserção dos inparálogos, os quais são proteínas que obtiveram maior score recíproco com uma proteína do mesmo organismo, já pertencente a um grupo ortólogo, que com qualquer outra proteína. Uma vez obtidas todas as combinações pareadas possíveis, um total de 36, utilizou-se o programa MultiParanoid para combinar todos os grupos, formados entre pares de organismos, em grupos de ortólogos entre as nove espécies analisadas.

Tabela 1- Bancos de proteínas de organismos do filo Apicomplexa utilizados na análise de ortologia

Organismo	Base de dados	Versão	Quantidade de proteínas
Toxoplasma gondii	http://toxodb.org	6.0	7.993
Plasmodium falciparum	http://plasmodb.org	6.3	5.446
Neospora caninum	ftp://ftp.sanger.ac.uk	Set 2009	7.083
Cryptosporidium parvum	http://cryptodb.org	4.3	3.805
Theileria annulata	ftp://ftp.sanger.ac.uk	15 Jul 2005	3.795
Babesia bovis	http://www.ncbi.nlm.nih.gov	08 Ago 2007	3.703

Fonte: Rangel (2011)

3.6 The *Eimeria* Transcript Database

Todos os dados de anotação gerados foram integrados no portal The *Eimeria* Transcript Database (http://www.coccidia.icb.usp.br/eimeriatdb). Este sítio foi desenvolvido em HTML, incluindo ainda as linguagens de programação PHP e JavaScript. Entre outros serviços, foram implementados uma versão local do BLAST com interface web (instalada a partir do pacote wwwblast do BLAST (http://www.ncbi.nlm.nih.gov/staff/tao/URLAPI /wwwblast/); um banco de dados relacional utilizando o sistema MySQL, populado com as informações relativas às evidências funcionais dos cDNAs reconstruídos das três espécies de *Eimeria*; e um conjunto de páginas contendo links para cada evidência de anotação, além das anotações propriamente ditas em formato GFF3 e Feature table. O portal foi instalado em um servidor com sistema operacional Linux Ubuntu Server rodando o conjunto de serviços LAMP (Linux, Apache, MySQL e PHP).

3.7 Geração de perfis digitais de expressão a partir de dados de montagem

A plataforma EGene contém dois componentes, `assembly_phrap.pl` e `assembly_cap3.pl`, para a montagem de leituras de sequenciamento convencional, os quais utilizam como montadores os programas Phrap (Green, 1996) e CAP3 (Huang e Madan, 1999), respectivamente. Para se obter perfis digitais de expressão gênica a partir de arquivos de montagem de transcritos, foi adicionada a estes componentes uma rotina que permite criar

perfis de expressão digital a partir de arquivos de montagem em formato ace, gerados pelos montadores. Esses arquivos de saída são gerados no formato CSV (comma-separated values).

3.8 Agrupamento hierárquico a partir de perfis de expressão

O componente `assembly_cap3.pl`, modificado como descrito acima, foi utilizado para gerar perfis digitais de expressão contendo as contagens do número de leituras que compõe cada sequência montada em função de sua origem respectiva (estágio de desenvolvimento). Esses perfis de expressão foram submetidos a uma análise de agrupamento hierárquico com o programa Simcluster (Vencio et al., 2007), utilizando diferentes métricas de distância. Este programa normaliza os diferentes dados de enumeração, permitindo que experimentos/bibliotecas de diferentes tamanhos sejam utilizados e comparados. Os arquivos de saída, em formato Newick, foram editados com o programa FigTree (Rambaut, 2011).

4 RESULTADOS E DISCUSSÃO

4.1 Validação do componente de classificação de ortologia

Foi desenvolvido um programa de classificação de sequências proteicas em grupos ortólogos, `annotation_orthology.pl`, conforme descrito no item 3.1. O programa teve de ser inteiramente implementado para que fosse possível realizar buscas locais. No caso do banco de dados COG/KOG, não ocorrem atualizações da base de dados desde 2003, e a única ferramenta local de busca disponível publicamente (National Center for Biotechnology Information, 2011b), o programa `dignitor` (não publicado), apresenta uma série de problemas e limitações. Primeiramente, a ferramenta é mal documentada e funciona bem apenas contra bases relativamente pequenas, como o KOG. Ao empregá-la em bases maiores, como eggNOG, ocorrem erros de alocação de memória. Além disso, o algoritmo e modo de trabalho do programa empregado remotamente pelo servidor web não são plenamente documentados, e os resultados gerados divergem dos providos pelo `dignitor`. Em relação à base de dados eggNOG, os autores têm realizado atualizações regulares (Muller et al., 2010), porém não oferecem publicamente nenhuma ferramenta para uso local. Semelhantemente ao que ocorre com o KOG, o protocolo de classificação em grupos ortólogos é descrito de forma lacônica e incompleta (Jensen et al., 2008). Em função disso, nossa escolha foi a de desenvolver um programa inteiramente novo, usando uma metodologia mais próxima quanto possível das descrições disponíveis na literatura para as bases KOG e eggNOG. O componente `annotation_orthology.pl` foi testado utilizando-se como conjuntos de dados das proteínas traduzidas dos cDNAs reconstruídos de *E. tenella, E. maxima* e *E. acervulina* contra as bases KOG e eggNOG, conforme metodologia descrita nos itens 3.1. Foram feitas então algumas validações, comparando-se os resultados gerados pelo componente local com aqueles obtidos nos servidores web das bases KOG e eggNOG.

O primeiro teste consistiu na análise de 1.150 sequências de *E. maxima* contra a base KOG. Para isso, as sequências foram submetidas a classificações em três ferramentas: [1] o servidor web do KOG (`kognitor`); [2] o programa `dignitor`, disponível no sítio do NCBI para *download*; e [3] o componente desenvolvido neste trabalho, `annotation_orthology.pl`. A Tabela 2 mostra os resultados comparativos das três análises. Como referência da comparação, foram adotados os resultados obtidos através do servidor web do KOG, o `kognitor`. Assim, das 1.150 proteínas analisadas, 246 foram classificadas em grupos de ortologia pelo `kognitor`, enquanto que 904 não obtiveram

classificação. Das 246 proteínas com atribuição de ortologia, 227 foram classificadas nos mesmos grupos ortólogos pelo `annotation_orthology.pl`, enquanto que apenas 159 proteínas obtiveram a mesma classificação pelo programa `dignitor`. Enquanto o `annotation_orthology.pl` apresentou 39 discordâncias (19 falsos negativos, 10 falsos positivos e 10 classificações em grupos distintos) em relação ao servidor web do KOG, o `dignitor` divergiu em relação a 90 proteínas. Além disso, o `dignitor` contabilizou 87 falsos negativos, contra apenas 19 do componente do EGene 2.

Tabela 2 - Comparação entre as classificações realizadas pelo EGene e `dignitor`.

Resultados de referência		EGene 2		`dignitor`	
		+	-	+	-
`kognitor`	+	227	19	159	87
	-	10	884	3	901
Total		237	903	161	988

Fonte: Rangel (2011)

Concluindo, os resultados comparativos da Tabela 2, tendo o `kognitor` como referência, indicam que apesar do programa `dignitor` possuir uma especificidade muito alta (99%) sua sensibilidade é bastante baixa (64%). O componente que desenvolvemos, por sua vez, não possui uma especificidade tão alta quanto o `dignitor` (97%), mas sua sensibilidade é bem maior (92%). Além dos falsos positivos e negativos observados nas classificações do `annotation_orthology.pl`, observamos a classificação de dez proteínas em grupos ortólogos diferentes dos classificados pelo `kognitor`. A menor taxa de erro observada pelas classificações de nosso componente demonstra sua vantagem em relação ao `dignitor`. Mesmo tendo apresentado mais falsos positivos (e classificações errôneas), a diferença entre as quantidades de falsos negativos torna mais coerente o uso o `annotation_orthology.pl`. De qualquer forma, esses resultados são baseados no uso do `kognitor` como referência, o que não exclui a possibilidade de que o próprio `kognitor` tenha uma certa taxa de erro em relação ao que seriam os resultados teoricamente corretos.

No caso da base eggNOG, não foi possível desenvolver uma ferramenta para submissão automática de sequências ao servidor web desse banco, pois a sua interface é extremamente complexa e dificulta imensamente uma interação com um *script* de submissão automática. Os testes de validação foram, portanto, realizados através da submissão manual

de 220 proteínas de *E. maxima* ao servidor web do eggNOG. Desta maneira, foi obtida uma concordância em cerca de 95% das proteínas submetidas à classificação pelo componente do EGene 2 e pelo servidor web do eggNOG. Apenas uma proteína foi classificada em um grupo ortólogo pelo `annotation_orthology.pl` e não pelo servidor do eggNOG (falso positivo), enquanto que oito proteínas foram classificadas pelo servidor web e não pelo nosso componente (falsos negativos). O último tipo de erro observado foi o de classificações em grupos ortólogos diferentes, tendo ocorrido em quatro proteínas.

4.2 Aplicação do componente de classificação de ortologia em proteínas de *Eimeria* spp.

Foram utilizados os três conjuntos de proteínas derivadas dos cDNAs reconstruídos de *E. acervulina*, *E. maxima* e *E. tenella* (Tabela 3). Comparando-se os resultados obtidos com as bases KOG e eggNOG, observou-se resultados bastante semelhantes em relação à taxa de classificação das proteínas em grupos ortólogos. Os resultados obtidos para as três espécies de *Eimeria* usando a base KOG podem ser visualizados no The *Eimeria* Transcript Database (aba Annotations, subdivisão KOG).

Utilizando-se o componente para a classificação em grupos ortólogos na base KOG, conseguimos levantar evidências sobre as funções de 707, 604 e 1.517 proteínas de *E. acervulina*, *E. maxima* e *E. tenella*, respectivamente. Comparativamente, observamos que o `annotation_orthology.pl` conseguiu classificar funcionalmente 55-59% do total das proteínas com resultados significativos de BLAST contra a base nr. Embora a quantidade de proteínas classificadas por ortologia seja bastante inferior àquela observada por BLAST, é importante mencionar que esse tipo de análise é muito mais rigorosa e pressupõe um conceito evolutivo (o de que as proteínas compartilham função devido e ancestralidade comum), o qual não é considerado na busca de similaridade por BLAST. Outro aspecto importante é que das 2.559 proteínas de *E. tenella* com resultados positivos por BLAST contra nr, 914 apresentaram similaridade contra proteínas hipotéticas. O `annotation_orthology.pl`, por sua vez, classificou apenas 57 proteínas na classe “Function unknown”. Desta forma, ao subtrairmos estes resultados pouco conclusivos do total de proteínas caracterizadas, observamos valores muito próximos. Em *E. tenella*, por exemplo, 1.645 proteínas apresentaram resultados positivos contra proteínas de função conhecida por BLAST, e 1.460

foram classificadas em grupos funcionalmente caracterizados pelo `annotation_orthology.pl`.

Tabela 3 - Resultados de análise de ortologia na anotação de proteínas de *Eimeria* spp

Resultado	***E. acervulina***	***E. maxima***	***E. tenella***
# de transcritos	3.413	3.426	8.700
# de proteínas (>50 aa)	3.233	3.096	7.990
BLAST x *nr* positivo (e-value < 10^{-6})	1.235	1.091	2.559
KOG	707	604	1.517
eggNOG	707	623	1.487
Ortologia em Apicomplexa	745	614	1.236
Anotação incrementada por ortologia em Apicomplexa	47	35	85

Fonte: Rangel (2011)

Ao aplicarmos nosso novo componente para a classificação em grupos ortólogos da base eggNOG observamos que um total de 2.817 proteínas das três espécies de *Eimeria* foram classificadas em algum grupo ortólogo (Tabela 3), sendo 1.487 de *E. tenella*, 623 de *E. maxima* e 707 de *E. acervulina*. A análise realizada com a base eggNOG apresentou valores bastante similares aos obtidos com a base KOG, a despeito da primeira ser muito maior. Isso se explica por terem sido utilizados algoritmos de classificação distintos para cada base de ortologia. Pelo fato do algoritmo usado para a base eggNOG avaliar apenas o primeiro hit do BLAST, torna-se mais provável que não se observe um grupo ortólogo que seria identificado pelo protocolo implementado para as bases COG/KOG, o qual avalia todos os hits com características acima do seu limiar. A Figura 2 mostra um exemplo de resultado fornecido pelo componente de classificação em grupos ortólogos na base eggNOG, incluindo a quantificação das categorias funcionais observadas. O componente gera uma tabela com as contagens, além de uma representação em forma de gráfico de pizza das proporções encontradas. O conjunto de resultados em formato HTML para as três espécies de *Eimeria* pode ser visualizado no The *Eimeria* Transcript Database (aba Annotations, subdivisão eggNOG)

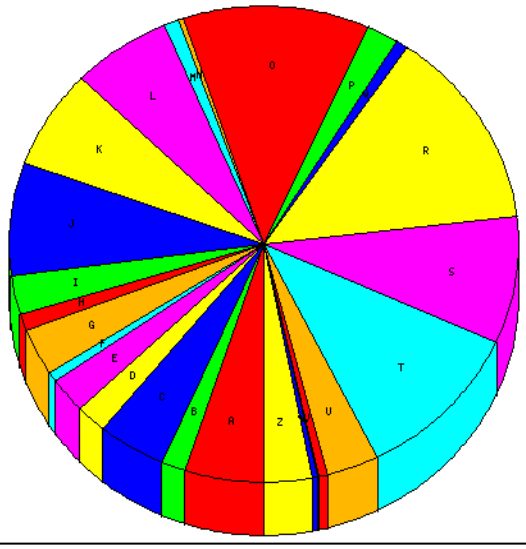
Figura 2 - Classificação por ortologia na base eggNOG de um conjunto de 3.096 sequências proteicas de *E.*

maxima utilizando o componente desenvolvido neste trabalho.

Quantification of functional classes

Class abbreviation	Functional category	# of counts	%
A	RNA processing and modification	73	5.03
B	Chromatin structure and dynamics	22	1.52
C	Energy production and conversion	63	4.34
D	Cell cycle control, cell division, chromosome partitioning	28	1.93
E	Amino acid transport and metabolism	35	2.41
F	Nucleotide transport and metabolism	10	0.69
G	Carbohydrate transport and metabolism	46	3.17
H	Coenzyme transport and metabolism	16	1.10
I	Lipid transport and metabolism	39	2.69
J	Translation, ribosomal structure and biogenesis	111	7.66
K	Transcription	97	6.69
L	Replication, recombination and repair	92	6.34
M	Cell wall/membrane/envelope biogenesis	14	0.97
N	Cell motility	5	0.34
O	Posttranslational modification, protein turnover, chaperones	172	11.86
P	Inorganic ion transport and metabolism	29	2.00
Q	Secondary metabolites biosynthesis, transport and catabolism	11	0.76
R	General function prediction only	197	13.59
S	Function unknown	126	8.69
T	Signal transduction mechanisms	157	10.83
U	Intracellular trafficking, secretion, and vesicular transport	49	3.38
V	Defense mechanisms	8	0.55
W	Extracellular structures	1	0.07
Y	Nuclear structure	5	0.34
Z	Cytoskeleton	44	3.03
Total		**1450**	**100.00**

Functional classes distribution



Fonte: Rangel (2011)

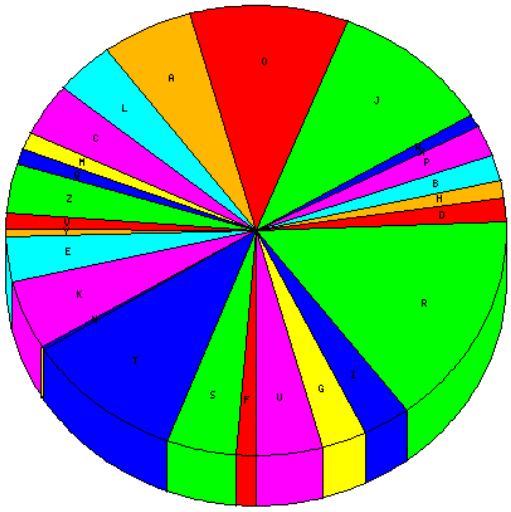
Figura 3 - Quantificação da classificação de 7.990 proteínas de *E. tenella* em grupos de ortologia da base

eggNOG, utilizando o protocolo de classificação da base COG/KOG, utilizando o componente desenvolvido neste trabalho.

kogNOG terms

kogNOG Code	Functional category	# of counts	%
A	RNA processing and modification	158	5.93
B	Chromatin structure and dynamics	50	1.88
C	Energy production and conversion	101	3.79
D	Cell cycle control, cell division, chromosome partitioning	46	1.73
E	Amino acid transport and metabolism	85	3.19
F	Nucleotide transport and metabolism	33	1.24
G	Carbohydrate transport and metabolism	78	2.93
H	Coenzyme transport and metabolism	32	1.20
I	Lipid transport and metabolism	82	3.08
J	Translation, ribosomal structure and biogenesis	278	10.44
K	Transcription	133	4.99
L	Replication, recombination and repair	102	3.83
M	Cell wall/membrane/envelope biogenesis	34	1.28
N	Cell motility	8	0.30
O	Posttranslational modification, protein turnover, chaperones	272	10.21
P	Inorganic ion transport and metabolism	59	2.22
Q	Secondary metabolites biosynthesis, transport and catabolism	29	1.09
R	General function prediction only	408	15.32
S	Function unknown	121	4.54
T	Signal transduction mechanisms	274	10.29
U	Intracellular trafficking, secretion, and vesicular transport	114	4.28
V	Defense mechanisms	30	1.13
W	Extracellular structures	25	0.94
Y	Nuclear structure	15	0.56
Z	Cytoskeleton	95	3.57
Total		**2663**	**100.00**

kogNOG category distribution



Fonte: Rangel (2011)

Ao aplicarmos o protocolo de classificação de proteínas em grupos ortólogos do COG/KOG, porém utilizando a base de dados do eggNOG, observamos a classificação de 2.326 proteínas de *E. tenella* em grupos de ortologia do eggNOG, onde apenas 124 foram classificadas em grupos sem categorização funcional. Esta combinação de análises levou a um ganho de mais de 30% de proteínas classificadas em grupos ortólogos, e pode ser observada na Figura 3 e no endereço http://www.coccidia.icb.usp.br/eimeriatdb/kognog/kogNOG.html. Este resultado, juntamente com o do KOG e do eggNOG, confirma o fato de que o uso da ortologia pode contribuir para a anotação de proteínas de organismos distantes daqueles considerados modelos, como é o caso dos parasitas do gênero *Eimeria*.

4.3 Validação do componente de mapeamento em vias metabólicas do KEGG

Visando o mapeamento de proteínas em vias metabólicas do KEGG, desenvolvemos o componente `annotation_pathways.pl`. Esse componente segue o modelo adotado pelo KAAS, ferramenta online do KEGG para a classificação de proteínas em grupos ortólogos (KO) e mapeamento em vias metabólicas. Por não haver nenhum software disponibilizado pela equipe do KEGG para realizar esta tarefa localmente, o componente para plataforma EGene 2 teve de ser totalmente desenvolvido com base nas descrições da literatura (Moriya et al., 2007). A validação do programa foi feita utilizando-se 7.990 sequências proteicas de *E. tenella* traduzidas a partir dos cDNAs reconstruídos. Utilizamos como referência os resultados obtidos no servidor do KAAS, submetendo o mesmo conjunto de sequências. Conforme apresentado na Tabela 4, 1.074 proteínas foram classificadas em KOs pelo KAAS, enquanto que 6.916 não tiveram nenhuma classificação. O componente `annotation_pathways.pl` classificou um total de 991 proteínas, sendo que desse total, 953 obtiveram as mesmas classificações do KAAS (verdadeiros-positívos), 9 não foram classificadas pelo KAAS (falsos-positivos) e 29 tiveram uma classificação diferente daquela determinada pelo KAAS. Essas diferenças de atribuições funcionais podem ser devidas a diferenças nas implementações dos algoritmos de classificação, mas também podem ser atribuídas a diferenças de versões entre o banco de dados que utilizamos, disponível no sítio FTP do KEGG, e a versão utilizado pelo servidor do KAAS. De fato, temos evidências que a base pública disponível para download não é a mesma que é usada pelo servidor web do KAAS.

Tabela 4 - Comparação entre os classificações realizadas pelo EGene e pelo KAAS

Resultados de referência		EGene 2	
		+	-
KAAS	+	953	92
	-	9	6.907
Total		962	6.999

Fonte: Rangel (2011)

4.4 Mapeamento de proteínas de *Eimeria* spp. em vias metabólicas do KEGG

Os três conjuntos proteicos traduzidos a partir dos cDNAs reconstruídos de *Eimeria* spp. foram submetidos ao componente `annotation_pathways.pl` para o mapeamento em vias metabólicas do KEGG. Os resultados obtidos para as três espécies de *Eimeria* podem ser visualizados no The *Eimeria* Transcript Database (aba Annotations, subdivisão KEGG). Das 1.089 proteínas de *E. tenella* classificadas em grupos ortólogos do KEGG (KO) pelo componente, 678 foram classificadas em grupos mapeados em vias metabólicas. As demais proteínas fazem parte de grupos ortólogos cujas funções são conhecidas, entretanto, as vias metabólicas nas quais atuam ainda não são definidas. Grande parte das vias metabólicas atribuídas às espécies de *Eimeria* (18,57% para *E. acervulina*, 23,43% para *E. maxima* e 25,20% para *E. tenella*) foram classificadas em na categoria "Processamento de informações genéticas" (Figura 4). Esses dados mostram-se consistentes por sabermos que os mecanismos relacionados à replicação, transcrição e tradução são bastante conservados ao longo da evolução e, por isso, é possível encontrar uma maior quantidade de ortólogos entre os organismos mais estudados e as eimerias. Em compensação, "Processamento de informações ambientais" é a classe com menor proporção de vias atribuídas às espécies do gênero (Figura 4), provavelmente porque elas possuem um nicho pouco comum entre os organismos mais bem caracterizados, o que dificulta a identificação dessas proteínas por homologia.

Para gerar arquivos em formato HTML a partir dos resultados de mapeamento de vias metabólicas da base KEGG, foi desenvolvido o `report_pathways.pl` (item 3.2). Os resultados são divididos em dois conjunto de arquivos de HTML: [1] uma lista de todas as categorias de vias metabólicas, mostrando as frequências com que foram mapeadas nas sequências proteicas e um gráfico de pizza apresentando suas distribuições (Figura 4); [2] páginas web detalhando cada uma das categorias funcionais do KEGG observadas durante o

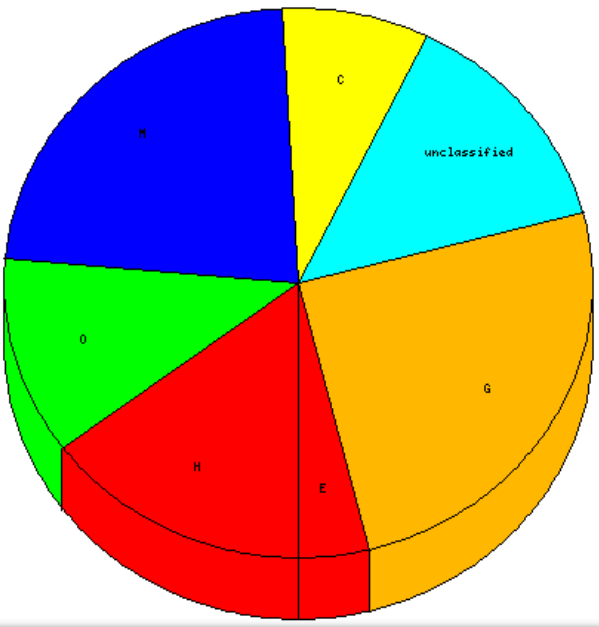
mapeamento. A primeira página web de resultados apresenta a quantificação das sete categorias funcionais de vias metabólicas do KEGG, organizadas como elementos em árvores expansíveis que, ao serem selecionadas, exibem as frequências das subcategorias das vias metabólicas. Quando o usuário seleciona uma subcategoria, uma nova página é aberta contendo uma tabela com informações referentes às proteínas dessa categoria. Na primeira coluna, o usuário encontra o nome da proteína consulta, o qual possui um link para o resultado de seu alinhamento contra a base de sequências derivada do GENES (item 3.2). A segunda coluna mostra o identificador KO (KEGG Orthology) do grupo ortólogo atribuído à proteína, com um link para a página que o descreve. A terceira coluna apresenta a descrição funcional do grupo ortólogo, e a quarta coluna contém o identificador E.C. (Enzyme Commission) com um link externo para a sua definição. Finalmente, a quinta coluna mostra os nomes das vias metabólicas em que a proteína foi mapeada. A lista de vias metabólicas está associada a links para uma terceira página gerada pelo componente, que contém uma figura da via metabólica onde o grupo ortólogo em questão está evidenciado (Figura 5). A imagem da via também está associada a um image map, no qual cada proteína, representada em um quadro com seu código E.C., possui um link que remete para a respectiva definição de seu grupo ortólogo.

Figura 4 - Classificação funcional das vias metabólicas da base KEGG para um conjunto de sequências proteicas de *E. tenella* utilizando o componente desenvolvido neste trabalho

Quantification of Metabolic Pathway Classes

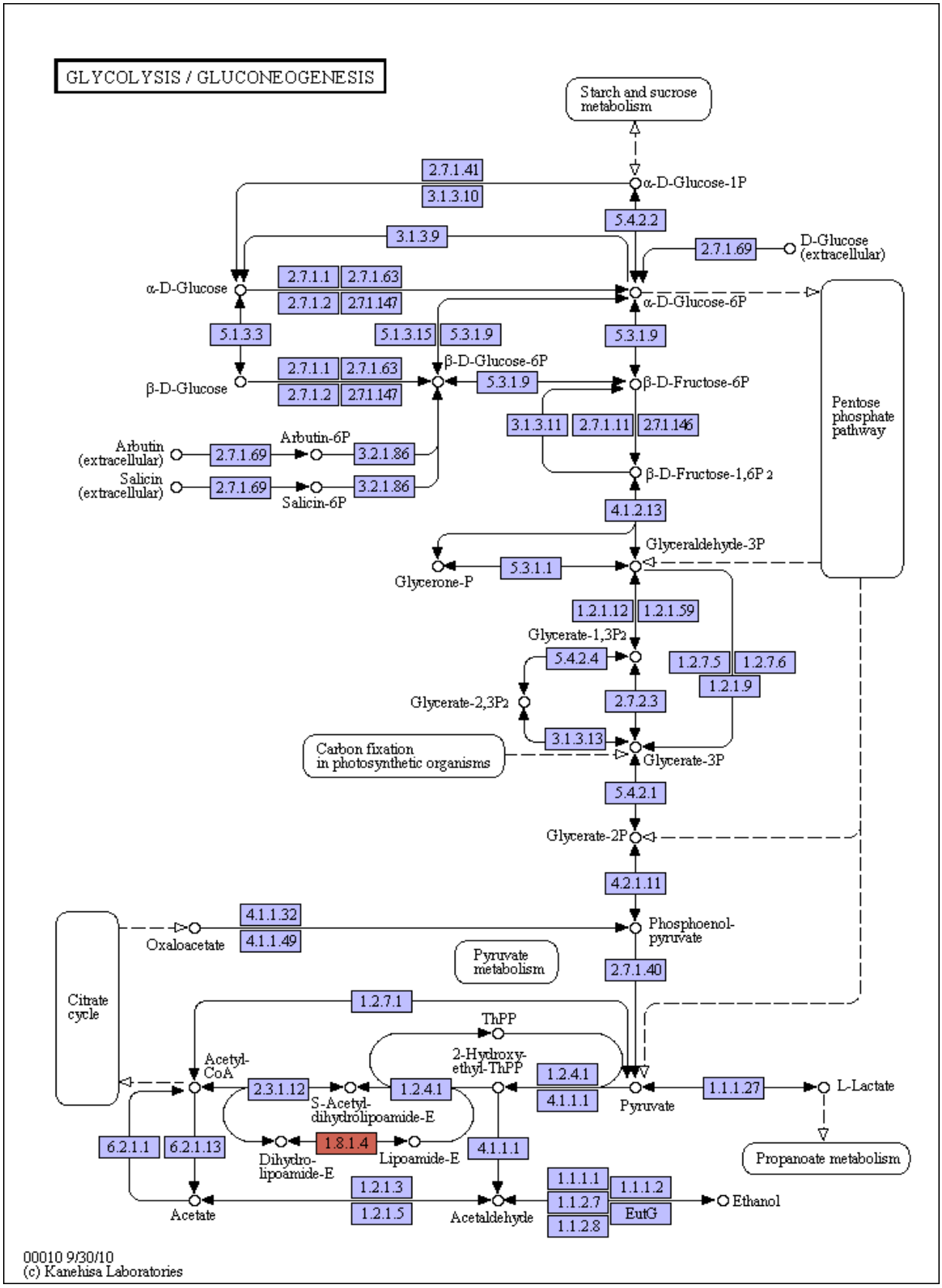
KEGG Code	KEGG Pathway Class	Absolute Frequency	%
M	⊟ Metabolism ▪ Carbohydrate (113) ▪ Energy (78) ▪ Lipid (55) ▪ Nucleotide (84) ▪ Amino Acid (65) ▪ Other Amino Acid (26) ▪ Glycan Biosynthesis and metebolism (11) ▪ Cofator/Vitamin (22) ▪ Terpenoid/PK (4) ▪ Other secondary metabolite (4) ▪ Xenobiotics Biodegradation and Metabolism (12)	474	22.71
G	⊞ Genetic Information Processing	526	25.20
E	⊞ Enviromental information processing	81	3.88
C	⊞ Cellular processing	169	8.10
O	⊞ Organismal systems	243	11.64
H	⊞ Human diseases	308	14.76
D	⊞ Drug development	0	0.00
Unclassified	Orthologous groups not mapped into metabolic pathways	286	13.70
Total		2087	100%

Pathway classes distribution



Fonte: Rangel (2011).

Figura 5 - Via metabólica identificada para a proteína Eten_0341 de *E. tenella*, que possui identificador EC 1.8.1.4, e foi classificada como uma diidrolipoamida desidrogenase.

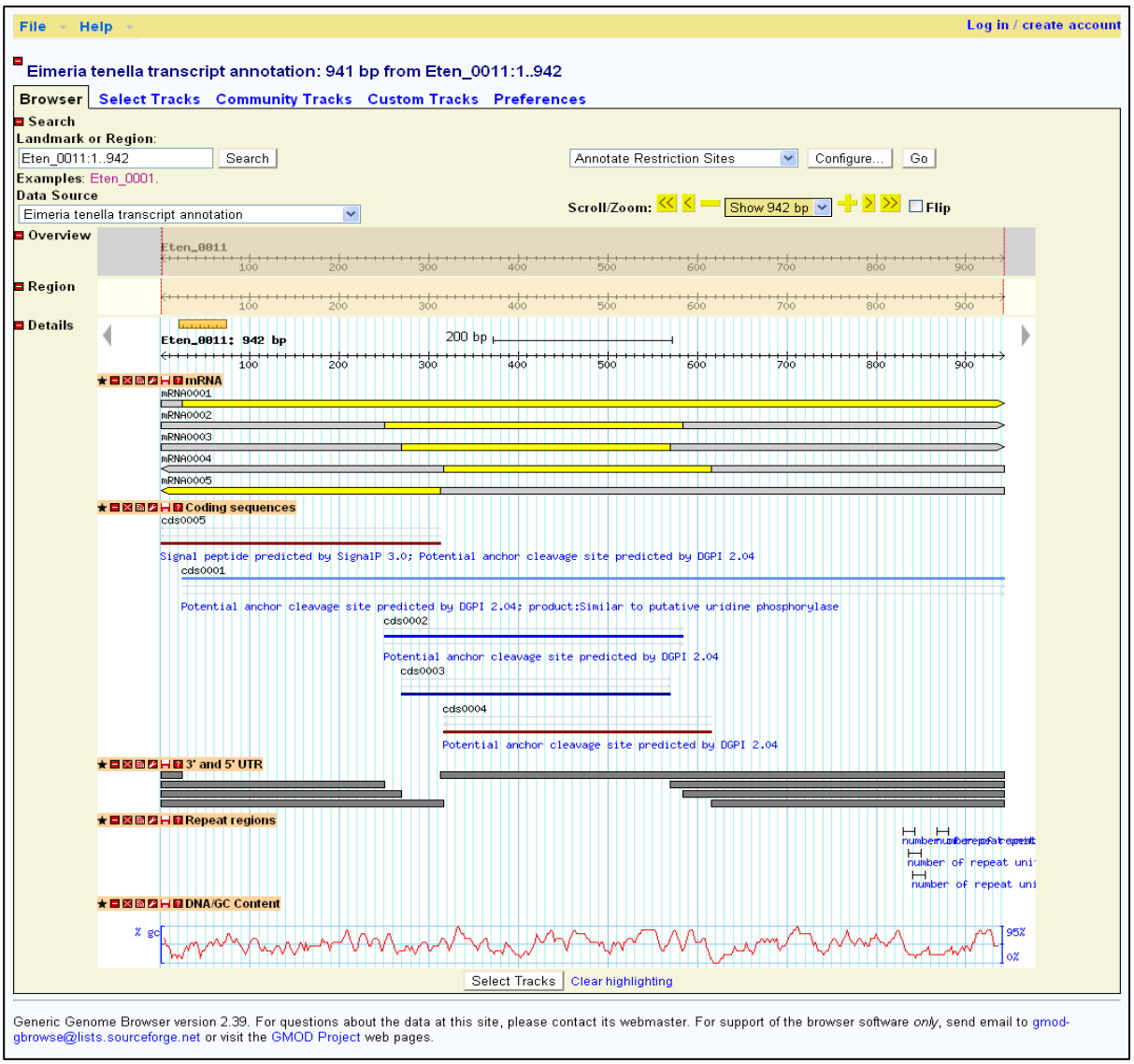


Fonte: Kanehisa e Goto (2000) gerado por Rangel (2011).

4.5 Integração de dados de anotação com o GBrowse

Para integrar os dados de anotação do EGene com o visualizador GBrowse, foi desenvolvido o componente `report_gbrowse.pl`. Na atual versão, o componente é mais adequado para a visualização de anotação de transcritos, e permite a representação gráfica de cinco elementos distintos: CDS (coding sequence – sequência codificadora), mRNA, 3' e 5' UTR (UTR - untranslated region – região não traduzida), regiões repetitivas e conteúdo GC. Cada um destes elementos é representado por uma *track* distinta, onde a CDS de um dado mRNA (ex. RNA0001) é identificada pelos sufixos numéricos (mRNA0001 e cds0001, respectivamente) (Figura 6). As *track*s utilizadas para identificar os mRNAs possuem uma identificação e uma divisão em duas regiões distintas, uma amarela, que representa a região codificante (CDS) e outra cinza, simbolizando as UTRs. Ao se selecionar um mRNA, uma nova janela é aberta, contendo o conjunto completo de informações sobre o mRNA, sua CDS e suas UTRs. As CDS são simbolizadas por uma *track* que representa a direção da transcrição possui uma identificação relativa ao seu mRNA de origem, e um conjunto de notas referentes às evidencias coletadas pelo EGene. Selecionando-se uma CDS específica, são visualizadas as suas informações respectivas, como referências cruzadas com bancos de dados externos (ex. KOG, eggNOG, InterPro e KEGG), descrição do produto, notas da anotação, sequência nucleotídica e os termos GO associados. As regiões 3' e 5' UTR, ao serem selecionadas, informam ao usuário sua posição e sequência. As regiões de repetições, ao serem selecionadas, disponibilizam sua sequência, o comprimento de sua unidade repetitiva, o número de cópias e o tipo da repetição. Por último, uma track representa a distribuição do conteúdo GC ao longo do cDNA (Figura 6). A utilização do GBrowse para visualizar a anotação das sequências permite uma análise mais dinâmica e integrada das informações, possibilitando a compreensão dos dados como um todo e como eles se relacionam.

Figura 6 - Visualização da anotação do cDNA Eten_0011, de *Eimeria tenella*, utilizando o GBrowse



Fonte: Stein et al., 2002 gerado por Rangel (2011).

4.6 Construções de grupos de proteínas ortólogas entre o filo Apicomplexa

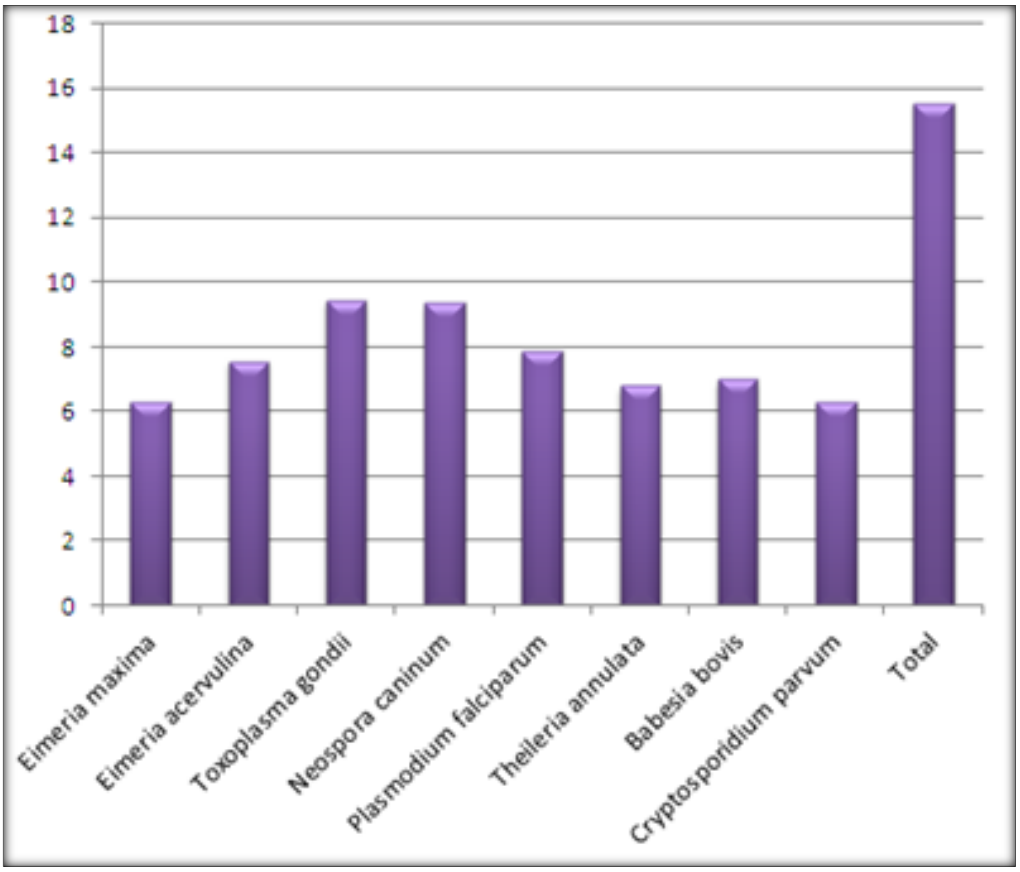
Utilizamos os programas InParanoid e MultiParanoid para construir um banco de dados de proteínas ortólogas de *Eimeria tenella*, *E. maxima*, *E. acervulina*, *Toxoplasma gondii*, *Plasmodium falciparum*, *Neospora caninum*, *Cryptosporidium parvum*, *Theileria annulata e Babesia bovis*, como explanado no item 3.5. Foram formados 7.489 grupos ortólogos, nos quais estão distribuídas 25.887 proteínas. Do total de grupos ortólogos, 1.299 possuem pelo menos uma das 2.595 proteínas de *Eimeria* spp. que foram agrupadas com seus ortólogos. Este agrupamento por ortologia permitiu classificar 1.236 proteínas de *E. tenella*

(15,53% do total de proteínas de espécie) (Figura 7), 745 de *E. acervulina* (23,14%) e 614 de *E. maxima* (19,96%). Analisando-se a totalidade dos grupos ortólogos de Apicomplexa, foram encontrados 175 grupos contendo simultaneamente proteínas das três espécies de *Eimeria* estudadas, com um total de 753 sequências proteicas. Ao se analisar estes 175 grupos, identificamos 76 que contêm exclusivamente proteínas de *E. tenella*, *E. maxima* e *E. acervulina*, e nenhuma sequência de outro Apicomplexa. Como o programa InParanoid é muito rigoroso na formação dos grupos ortólogos, é possível que algumas proteínas inseridas nesses grupos de fato possuam homólogos em outros organismos Apicomplexa, os quais não tenham sido agregados a esses grupos por não ter atingido os limiares exigidos pela ferramenta. Com o intuito de identificar proteínas absolutamente restritas ao gênero *Eimeria*, submetemos as proteínas desses grupos ortólogos a uma análise subsequente. Para isso, as proteínas destes 76 grupos foram submetidas a uma busca de similaridade, com BLAST, contra os conjuntos proteicos dos outros organismos Apicomplexa. Assim, um grupo ortólogo que contivesse qualquer proteína cujo alinhamento possuísse um bitscore maior que 70 e um e-value menor que 10^{-12} seria excluído do conjunto de grupos ortólogos restritos ao gênero *Eimeria*. Ao final deste processo, obtivemos 27 grupos ortólogos cujas proteínas não apresentam relações de homologia com proteínas de outros organismos Apicomplexa (34 proteínas de *E. acervulina*, 28 de *E. maxima* e 50 de *E. tenella*). O estudo destas proteínas revelou a presença de diversos genes de antígenos de superfície (SAG), os quais também estão presentes em grande quantidade em *Toxoplasma gondii*. Segundo Tabares et al. (2004), os SAGs de *T. gondii* não possuem homólogos em *E. tenella*, confirmando nossos resultados. Observamos, também, que este grupo ortólogo não é restrito a esta espécie, mas sim compartilhado com as outras duas espécies de *Eimeria* estudadas. Dos 27 grupos ortólogos comuns e específicos às três eimerias em questão, encontramos dois grupos contendo as proteínas de micronema EtMIC2 e EtMIC3 de *E. tenella*, e acreditamos que este resultado é a primeira descrição dessas proteínas em *E. acervulina*, e também que este grupo ortólogo deve conter proteínas de outras eimérias de galinha. Nossa análise identificou diversos grupos de genes conservados entre vários organismos do filo Apicomplexa, o que sugere certa importância dentro do taxon, e portanto, devem ser considerados para futuros estudos funcionais mais profundos, e até mesmo sua aplicação como alvo de drogas.

As baixas frequências de proteínas ortólogas das espécies de *Eimeria*, entre si e entre os demais organismos do filo Apicomplexa (Figura 7), podem ser explicadas pela diferente natureza e cobertura dos dados utilizados. Assim, as sequências proteicas dos outros

organismos apicomplexas são oriundas de predições realizadas a partir de seus genomas completos. No caso das espécies de *Eimeria*, foram utilizados os produtos proteicos traduzidos a partir dos cDNAs reconstruídos a partir de leituras do tipo ORESTES e/ou ESTs.

Figura 7 - Distribuição das proteínas de *Eimeria tenella* em grupos ortólogos compartilhados com outros organismos do filo Apicomplexa. Os valores se referem a porcentagens relativas ao total de proteínas de *E. tenella*.



Fonte: Rangel (2011)

A Tabela 5 apresenta os resultados completos dessa análise de ortologia/inparalogia dos organismos Apicomplexa, na forma de uma matriz de compartilhamento de grupos ortólogos entre organismos estudados, todos contra todos. Mesmo observando-se uma porcentagem relativamente baixa de proteínas de *E. tenella* com relações de ortologia com proteínas de outros apicomplexas, as porcentagens seguiram um padrão esperado pela proximidade evolutiva entre as espécies. Assim, *Toxoplasma gondii* e *Neospora caninum*, membros da classe Coccidia, assim como *E. tenella*, foram os organismos com maiores percentuais de proteínas ortólogas com esta espécie de *Eimeria*, 9,44% e 9,36%, respectivamente, seguindo-se *Plasmodium falciparum*, *Babesia bovis* e *Theileria annulata*.

Por último, encontramos *Cryptosporidium parvum*, a espécie mais divergente entre as estudadas do filo Apicomplexa, o que é compatível com o que se conhece a respeito das relações filogenéticas desse gênero (Kuo et al., 2008). Os dados da Tabela 6 mostram ainda a proximidade entre *T. gondii* e *N. caninum* (Tg x Nc – 78.44% e Nc x Tg – 89,83%) e entre *B. bovis* e *T. annulata* (Bv x Ta – 71,51% e Ta x Bv – 69,70%), além da alta divergência de *C. parvum* com todos os demais organismos. Estes fatos nos levam a crer que apesar da pequena quantidade de proteínas ortólogas encontradas entre as eimerias, nossos dados apresentam uma consistência considerável.

Tabela 5 - Matriz de compartilhamento de ortologia baseada nas análises de ortologia/inparalogia entre organismos do filo Apicomplexa. Os valores estão apresentados em porcentagens relativas ao total de proteínas de cada organismo.

	E. tenella	*E. maxima*	*E. acervulina*	*T. gondii*	*N. caninum*	*P. falciparum*	*B. bovis*	*T. annulata*	*C. parvum*
E. tenella	-	6,26	7,55	9,44	9,36	7,87	7,0	6,78	6,3
E. maxima	13,34	-	11,98	7,95	7,78	6,59	5,85	5,72	5,52
E. acervulina	16,27	12,03	-	9,28	9,0	7,24	6,62	6,43	6,03
T. gondii	10,95	4,89	5,87	-	78,44	31,2	26,1	25,56	23,72
N. caninum	12,28	5,44	6,42	89,83	-	34,91	29,06	28,49	26,39
P. falciparum	11,86	4,96	5,66	43,59	43,41	-	40,07	39,85	31,93
B. bovis	15,2	6,29	6,99	52,28	51,8	57,57	-	71,51	40,91
T. annulata	14,39	5,98	6,67	50,49	50,12	56,07	69,7	-	40,69
C. parvum	13,48	6,68	6,02	45,6	45,28	43,97	39,37	39,58	-

Fonte: Rangel (2011).

Os resultados da Tabela 5 indicam que as quantidades de proteínas que compartilham relações de ortologia entre os organismos testados são inversamente proporcionais à distância evolutiva destes organismos. Esse resultado é obviamente esperado, visto que a ortologia é um conceito baseado em relações de ancestralidade comum e divergência a partir de eventos de especiação.

Visando comprovar que as análises realizadas no presente trabalho estavam de fato refletindo relações de ortologia, decidimos gerar um dendrograma das espécies analisadas a partir dos dados de compartilhamento de grupos ortólogos, e comparar sua topologia com as

das árvores filogenéticas descritas na literatura. Assim, geramos um outro banco de dados de proteínas ortólogas, composto de proteínas preditas apenas a partir de genomas completamente sequenciados. Para tanto, utilizamos os genomas de *T. gondii, P. falciparum, N. caninum, C. parvum, T. annulata, B. bovis e Tetrahymena thermophila* (filo Ciliophora), além da versão do genoma de *E. tenella*, liberada em setembro de 2010 pelo Instituto Sanger (http://www.sanger.ac.uk/resources/downloads/protozoa/*Eimeria*-tenella.html), a qual contém 8.786 proteínas preditas. O genoma de *T. thermophila* foi obtido da base de dados RefSeq (NZ_AAGF00000000.3) e possui 24.725 proteínas preditas. Utilizando-se os programas InParanoid e MultiParanoid de maneira idêntica à análise anterior, foram gerados 7.016 grupos ortólogos contendo 31.237 proteínas. A Tabela 6 mostra as distâncias de Jaccard, calculadas para cada par de organismos.

Tabela 6 - Matriz de distância de Jaccard relativa aos grupos de proteínas ortólogas compartilhados pelos organismos Apicomplexa. As proteínas foram conceitualmente traduzidas a partir dos genomas completos dos organismos.

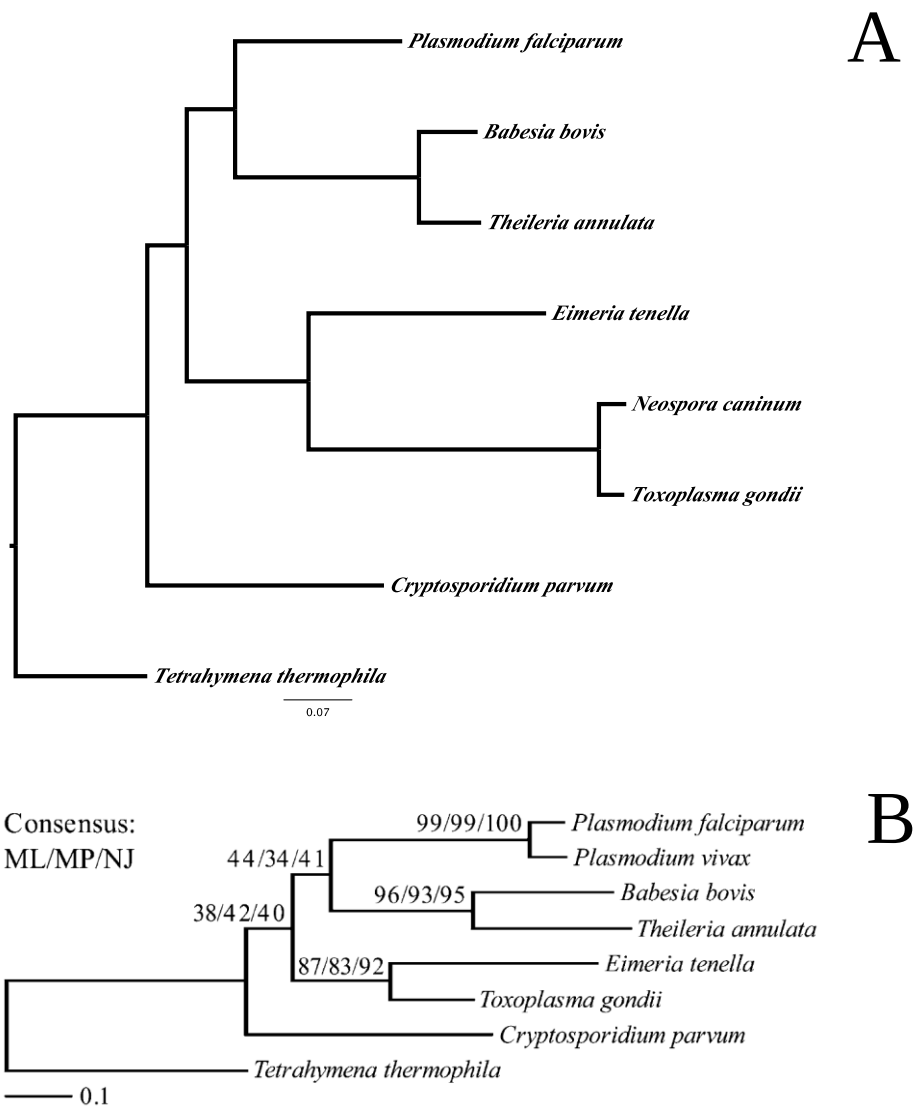
	C. parvum	*T. annulata*	*B. bovis*	*T. thermophila*	*T. gondii*	*E. tenella*	*N. caninum*	*P. falciparum*
C. parvum	-	-	-	-	-	-	-	-
T. annulata	0,56	-	-	-	-	-	-	-
B. bovis	0,57	0,12	-	-	-	-	-	-
T. thermophila	0,51	0,63	0,63	-	-	-	-	-
T. gondii	0,74	0,74	0,73	0,75	-	-	-	-
E. tenella	0,64	0,67	0,67	0,66	0,56	-	-	-
N. caninum	0,75	0,75	0,74	0,75	0,05	0,56	-	-
P. falciparum	0,50	0,42	0,41	0,51	0,67	0,59	0,67	-

Fonte: Rangel (2011)

Os dados da Tabela 6 foram utilizados como matriz de entrada de dados do programa PAUP e submetidos a uma análise Neighbor-Joining, resultando no dendrograma apresentado na Figura 8A. Assim como no trabalho de Kuo et al. (2008) *T. thermophila* foi utilizada como grupo externo na construção do dendrograma. Esse dendrograma (Figura 8A) apresenta uma topologia congruente com a árvore de espécies (Figura 8B) descrita por Kuo et al., (2008), a qual foi gerada a partir de uma análise filogenômica. A congruência da árvore que geramos a

partir de agrupamentos de proteínas ortólogas do filo Apicomplexa com aquela obtida por análises filogenômicas é uma forte indicação de que os programas InParanoid/MultiParanoid foram capazes de gerar grupos que de fato refletem relações de ortologia dos organismos do filo Apicomplexa.

Figura 8 - (A) Dendrograma de organismos do filo Apicomplexa calculado a partir do grau de compartilhamento de grupos ortólogos. (B) Árvore de espécies (*species tree*) de apicomplexas deduzida a partir de dados filogenômicos.



Fonte A: Rangel (2011).

Fonte B: Kuo et al. (2008)

4.7 Anotação transitiva de *Eimeria* spp. por associação de ortologia entre organismos Apicomplexa

Utilizando-se o agrupamento por ortologia com as bases KOG e eggNOG, conseguimos classificar apenas uma pequena parte das proteínas de *Eimeria* spp. Entretanto, ao cruzarmos os dados de ortologia de organismos Apicomplexa (todos contra todos) com a classificação de todas as proteínas desses organismos em grupos ortólogos do KOG, fomos capazes de incrementar de forma transitiva o número total de proteínas de Apicomplexa (incluindo as de *Eimeria*) classificadas nessa base.

As proteínas de cada organismo Apicomplexa foram mapeadas nos grupos de ortologia da base KOG (Tabela 7) usando-se ao componente `annotation_orthology.pl` desenvolvido nesse trabalho (item 4.2). Associando-se os resultados de ortologia dos organismos Apicomplexa na base KOG, com os resultados de agrupamento de ortologia entre as proteínas dos próprios organismos Apicomplexa (item 4.6), foi possível se fazer uma anotação transitiva, conforme o algoritmo que segue. Seja *A* uma proteína sem classificação KOG e *X* um grupo ortólogo de Apicomplexa, composto das proteínas *A*, *B* e *C*, se a proteína *C* tiver uma classificação KOG, essa classificação será incorporada de forma transitiva à proteína *A*. Para se implementar a anotação transitiva descrita acima, desenvolvemos um *script* que integra e cruza os dados de classificação de todas as proteínas na base KOG (item 4.1), com a lista de grupos ortólogos gerados a partir das proteínas dos diferentes organismos Apicomplexa (item 4.6). A partir desse cruzamento de dados, o programa realizou a anotação transitiva das proteínas derivadas dos cDNAs reconstruídos das três espécies de *Eimeria*.

Tabela 7 - Frequências absolutas de proteínas de diferentes organismos do filo Apicomplexa classificadas na base KOG, utilizando-se a classificação direta na própria base do KOG, ou a classificação transitiva através dos grupos ortólogos de Apicomplexa.

	Toxoplasma gondii	***Neospora caninum***	***Plasmodium falciparum***	***Babesia bovis***	***Theileria annulata***	***Cryptosporidium parvum***
Diretamente	3.589	3.511	2.532	1.999	2.009	2.156
Transitivamente	240	212	209	160	163	81

Fonte: Rangel (2011)

Com esta abordagem, fomos capazes de classificar em grupos de ortologia do KOG mais de 1.232 proteínas de organismos Apicomplexa que não haviam sido previamente classificadas nessa base (Tabela 7). Essa anotação transitiva permitiu ainda categorizar um total de 167 proteínas de *Eimeria* spp., sendo 85 de *E. tenella*, 35 de *E. maxima* e 47 de *E. acervulina*. Um exemplo consistente dessa abordagem é o grupo ortólogo "1.736" (Figura 8). Este grupo possui três proteínas de cada uma das espécies de *Eimeria* que trabalhamos, sendo duas delas classificadas no identificador KOG1164 como caseína quinase (serine/threonine/tyrosine protein kinase). A proteína Emax_0723, membro deste grupo, não havia obtido classificação na base KOG, mas, com esta metodologia, passou a incorporar de forma transitiva a classificação funcional do grupo KOG1164. Além da congruência entre as classificações do KOG, as outras proteínas deste grupo, Eten_2431 e Eace_0350, foram ambas anotadas como "Similar to casein kinase I beta isoform" utilizando-se um BLAST contra nr. Os dados de classificações em KOGs e das relações de ortologia entre as espécies do filo Apicomplexa já foram integradas com os dados de anotação gerados previamente pelo EGene 2 em nosso banco de dados. Os bancos das três espécies de *Eimeria* estão interligados para facilitar a comparação entre as diferentes classificações, como visualizado na Figura 9.

Figura 9 - Captura de tela do banco de dados do transcriptoma de *Eimeria* spp., referente à proteína Emax_0723 de *E. maxima*. As referências cruzadas com as proteínas Eten_2431 e Eace_0350, pertencentes ao mesmo grupo de ortologia, estão presentes através de links.

***Eimeria maxima* EGene annotations**

Sequences

- FASTA sequence: Emax_0723
- Emax_0723_ORFs: DNA, Protein

Annotation

- Without ORF selection: FT, extended FT, GFF3
- With ORF selection: FT, extended FT, GFF3
- View into GBrowse

Orthology relations of the selected ORF

Organism	Protein	Confidence value	Relation	KOG result
Eimeria acervulina	Eace_0350	1.0	Ortholog	KOG1164
Eimeria tenella	Eten_2431	1.0	Ortholog	KOG1164

Evidences

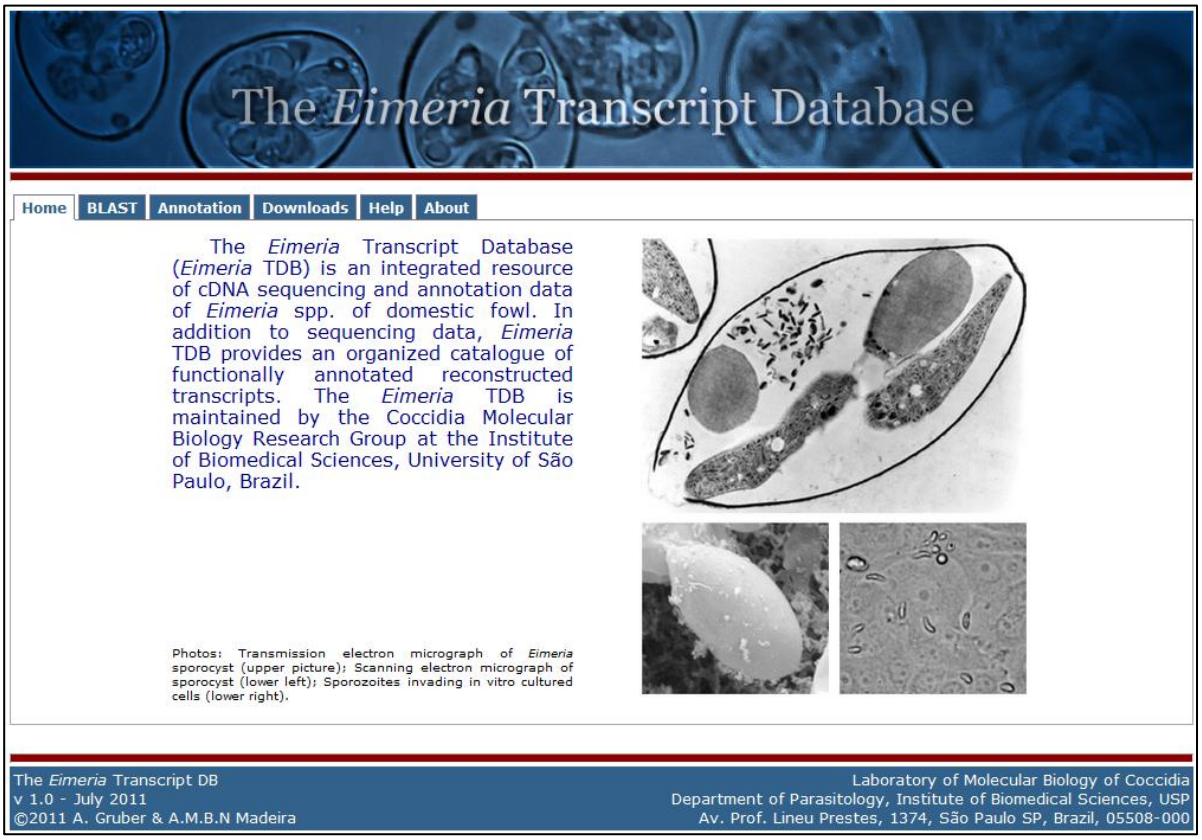
ORF	SignalP	TMHMM	Phobius	DGPI	BLAST	RPS-BLAST	InterProScan	GO terms	Orthology	KEGG
ORF1	No hits	1 potential transmembrane helix	1 potential transmembrane helix	No hits	No hits	No hits	No hits	No hits	eggNOG:No hits KOG:No hits	No hits
ORF2	No hits	No hits	No hits	No hits	casein kinase I, putative	No hits	No hits	No hits	eggNOG:No hits KOG:No hits	No hits

Fonte: Rangel (2011)

4.8 The *Eimeria* Transcript Database

Durante este trabalho desenvolvemos o portal The *Eimeria* Transcript Database, visando centralizar todos os dados que geramos em relação aos transcritos reconstruídos de *E. acervulina*, *E maxima* e *E. tenella*. Este portal possui seis abas para organizar seu conteúdo: Home, BLAST, Annotation, Downloads, Help e About (Figura 10).

Figura 10 - Captura de tela do sítio The *Eimeria* Transcript Database na aba inicial, Home. A figura apresenta a organização das suas seis abas.



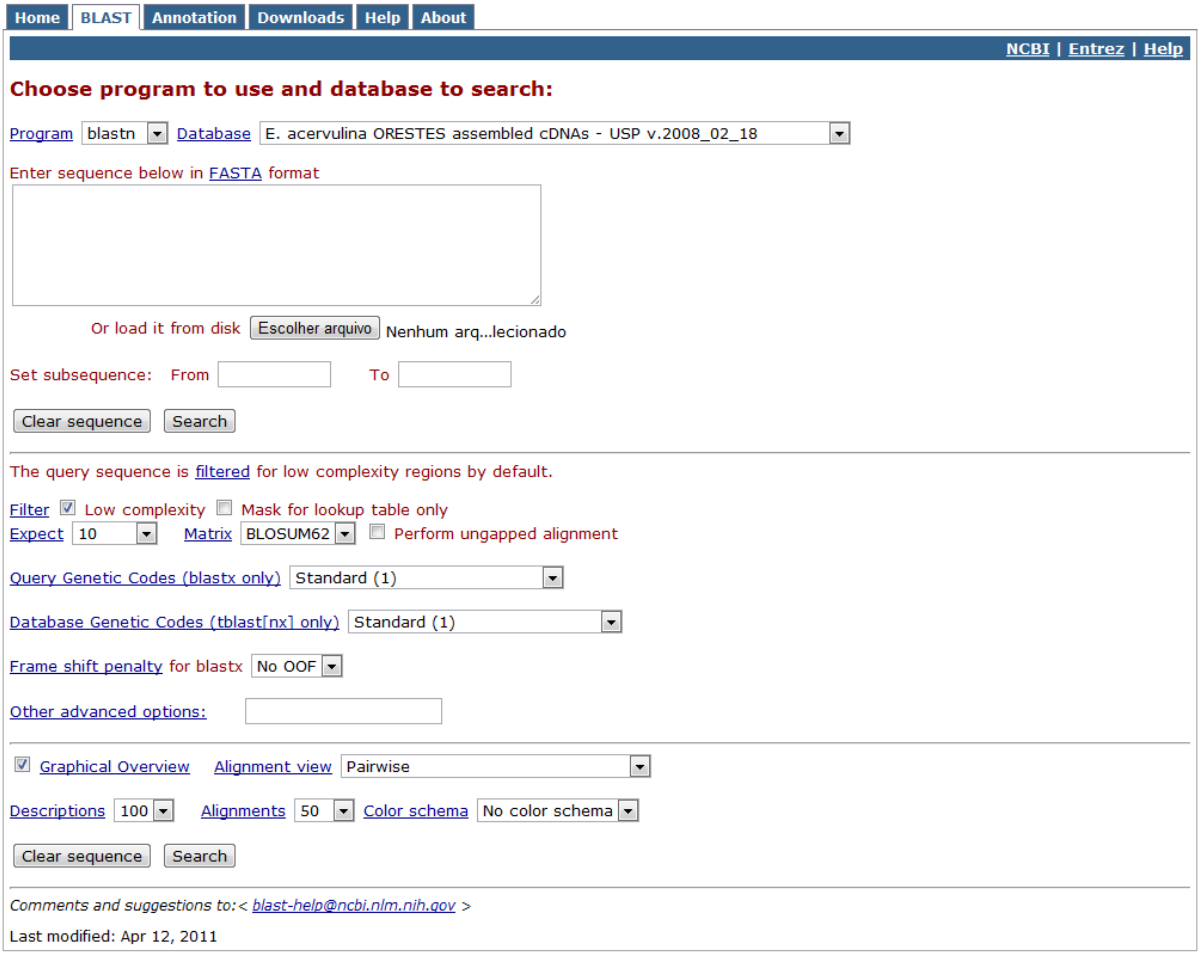
Fonte: Rangel (2011)

4.8.1 Serviço de BLAST local

O portal conta com um serviço BLAST, utilizando uma instalação local do mesmo, contra diversas bases de dados de *Eimeria* spp., que incluem dados genômicos, cDNAs e sequências mitocondriais (Figura 11). As sequências genômicas abrangem leituras de shotgun e diversas versões de montagens do genoma de *E. tenella*, disponibilizadas pelo Instituto Sanger ao longo dos anos. Os cDNAs de *E. tenella* foram montados através de uma combinação de leituras ORESTES e ESTs. No caso de *E. acervulina* e *E. maxima* a montagem dos cDNAs foi feita utilizando-se apenas leituras ORESTES. A tradução dos produtos dos cDNAs também são disponibilizadas como base de dados, incluindo ORFs maiores que 50 e 100 aminoácidos. Por fim, um banco de dados contendo os genomas mitocondriais das sete espécies de *Eimeria* que infectam galinha doméstica está disponível para buscas de similaridade. Todas as versões de BLAST estão disponíveis: blastn, blastp,

blastx, tblastn e tblastx. Após identificar-se um hit de cDNA o usuário pode usar seu identificador para inspecioná-lo em outros serviços disponíveis no portal.

Figura 11 - Página do serviço BLAST, exibindo todas as suas opções e parâmetros.



Fonte: Rangel (2011)

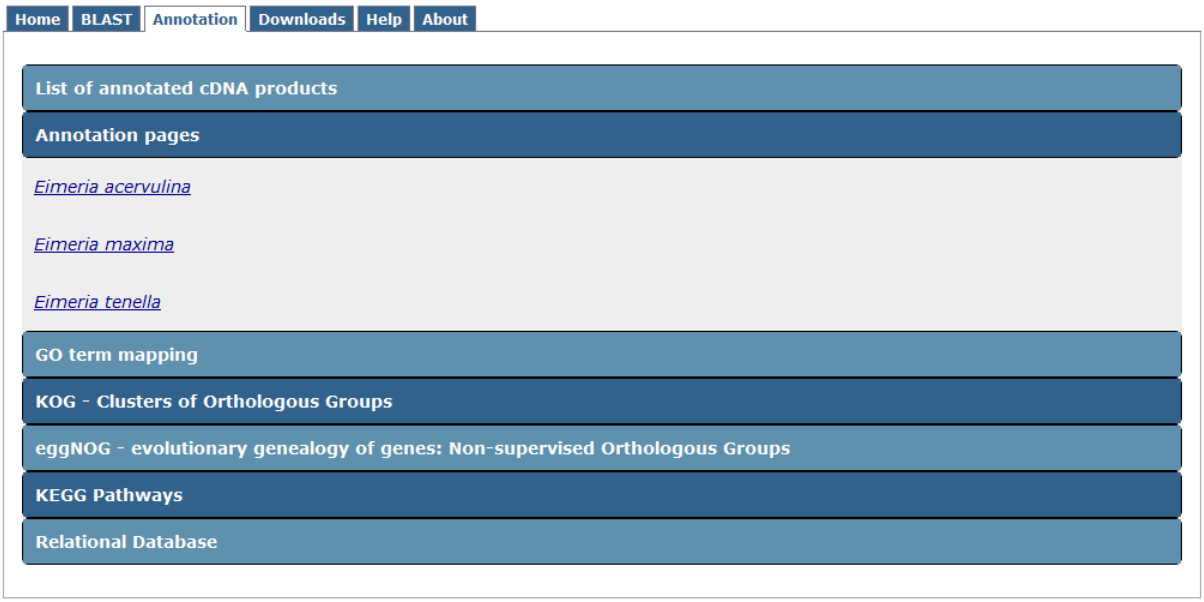
4.8.2 Anotação

Esta aba centraliza todos os dados de anotação funcional dos cDNAs reconstruídos (Figura 12), e possui sete subdivisões: List of annotated cDNA products, Annotation pages, GO term mapping, KOG, eggNOG, KEGG Pathways e Relational Database.

4.8.2.1 Lista dos produtos dos cDNAs anotados

Esta é uma lista dos cDNAs montados e seus respectivos produtos. Os produtos que apresentaram resultados positivos no BLAST encontram-se em uma lista separada dos que não possuem nenhum hit significativo (proteínas hipotéticas). Esta é a maneira mais rápida de saber se uma determinada proteína foi anotada em algum cDNA.

Figura 12 - Conteúdo da aba Annotation, com suas sete subdivisões e links separados para os dados de cada espécie.



Fonte: Rangel (2011)

4.8.2.2 Páginas de anotação

Caso o usuário tenha o identificador de um cDNA específico (ex. Emax_0723) através do BLAST ou buscando na lista de produtos dos cDNAs, esta sequência pode ser analisada de acordo com sua anotação completa. As páginas de anotação contém links para todos os cDNAs montados foram construídas com um componente específico (`report_html.pl`) da plataforma EGene, o qual gera automaticamente todo o conjunto de páginas HTML sem a necessidade de se desenvolver uma base de dados relacional. Os links são divididos em grupos de 100 sequências anotados. Assim, caso se queira, por exemplo, inspecionar o cDNA Emax_0723, deve-se primeiro selecionar a opção “Emax_0701..Emax_0800” para, em

seguida, poder selecionar o link da sequência Emax_0723, o qual permite acessar a página contendo todas as evidências e anotações dessa sequência (Figura 9). O portal oferece links para páginas de anotações das três espécies de *Eimeria* estudadas nesse trabalho. Cada página de anotação oferece um vasto conjunto de informações, como será descrito nos itens a seguir.

4.8.2.*2*.1 Sequências

Nesta seção encontra-se um link para a sequência do cDNA montado, em formato FASTA. As sequências nucleotídicas e protéicas das ORFs potenciais também estão disponíveis.

4.8.2.2.2 Anotação

Os arquivos de anotação estão disponíveis em três formatos distintos: Feature table, Feature table estendido e GFF3. Além destes arquivos, a anotação pode ser observada através do visualizador genômico GBrowse, que contém os dados de anotação das três espécies de *Eimeria*. A anotação pode ser obtida para todas as ORFs potenciais de um cDNA ou, alternativamente, somente para a ORF selecionada para aquele cDNA. O nosso pipeline de anotação automática, executado com o sistema EGene, ao invés de empregar preditores de genes, determinou todas as ORFs potenciais, codificando proteínas maiores do que 50 resíduos de aminoácidos. Em seguida, todas as sequências proteicas foram analisadas quanto à similaridade contra bancos de dados, presença de domínios e motivos proteicos, entre outras características. O EGene contém um componente que permite selecionar a ORF mais provável com base num sistema de decisão baseado em evidências. Este processo considera a presença de ortólogos, motivos proteicos, comprimento da região codificadora, etc. Concluindo, o usuário pode inspecionar a anotação relativa apenas à ORF automaticamente selecionada pelo EGene ou as anotações de todas as ORFs hipotéticas.

4.8.2.2.3 Relações de homologia da ORF selecionada

Incrementamos a anotação dos produtos do transcriptoma das três espécies de *Eimeria* desenvolvendo um banco de dados de proteínas ortólogas do filo Apicomplexa (item 4.6). A lista de proteínas homólogas ao produto selecionado pode ser observada na tabela "Orthology relations of the selected ORF". Esta tabela contém uma relação de proteínas ortólogas/parálogas, organismo de origem, o valor de confiança da relação, tipo de relacionamento (ortologia ou inparalogia). A última coluna contém, ainda, o identificador do KOG no qual a proteína foi classificada e um link para sua página de descrição.

4.8.2.2.4 Evidências observadas

A atribuição funcional realizada pelo EGene baseia-se na coleta e avaliação de diversas evidências funcionais da sequência anotada. No caso de um cDNA com mais de uma provável ORF, são coletadas evidências para todas as ORFs, e todas essas evidências são disponibilizadas para inspeção, mesmo que o EGene já tenha definido, automaticamente, qual a ORF selecionada. Portanto, caso o usuário discorde da seleção automática de uma determinada ORF, todos os resultados originais estão presentes, possibilitando uma prática curadoria manual. Todos os links para os resultados das evidências coletadas estão organizados em uma tabela, onde as ORFs estão separadas por linhas e as diferentes evidências em colunas (BLAST, RPS-BLAST, InterProScan, SignalP, Phobius, TMHMM, DGPI, termos GO mapeados através do InterPro, classificações em grupos ortólogos das bases KOG e eggNOG e mapeamento em vias metabólicas do KEGG).

4.8.3 Mapeamento de termos GO

O sítio web também disponibiliza os resultados do mapeamento e quantificação de todos os termos GO encontrados. Estes resultados encontram-se de duas formas distintas: árvores expansíveis e tabelas. Cada árvore expansível é composta por três árvores, correspondendo respectivamente aos três domínios da ontologia. Clicando nos símbolos de

“mais” e “menos” à esquerda, os ramos se expandem ou retraem, respectivamente. Selecionando o termo GO, a página é redirecionada para o browser AMIGO, contendo a descrição do termo em questão. À direita dos termos GO encontram-se links para todas as sequências que foram mapeadas neste termo. Ao lado da lista de sequências mapeadas para cada termo encontram-se novos links para suas sequências nucleotídicas e proteicas, além do conjunto de termos que qualificam a sequência. Vale ressaltar que esta visualização não é compatível com o MS Internet Explorer, já que o mesmo não suporta arquivos em formato XML. Neste caso, também disponibilizamos estes dados utilizando um HTML convencional com tabelas. Como alternativa para os usuários do MS Internet Explorer, existe o link “Table of ontologies”. Ao invés de apresentar as árvores expansíveis, o usuário é direcionado para uma tabela em formato HTML, que contém exatamente as mesmas informações das árvores expansíveis.

4.8.4 KOG - euKaryotic clusters of Orthologous Groups

Classificamos todos os produtos dos transcritos reconstruídos em grupos ortólogos da base KOG (item 4.2). Esta página contém uma tabela contendo as categorias funcionais do KOG e a respectiva frequência em que elas classificaram os produtos dos transcritos, juntamente com um gráfico, em formato de pizza, com as frequências relativas de cada categoria. Clicando-se na abreviação das categorias funcionais, a página é redirecionada para uma outra tabela, contendo todas as sequências classificadas nesta categoria. Juntamente com a lista de sequências, são apresentados, as categorias funcionais que qualificam esta sequência, a descrição do grupo ortólogo e seu identificador, além de um link para o resultado do BLAST da sequência contra a base KOG.

4.8.5 eggNOG - evolutionary genealogy of genes: Non-supervised Orthologous Groups

Semelhantemente aos resultados apresentados para a base KOG (item 4.8.4), também disponibilizamos a classificação dos produtos dos transcritos na base eggNOG (item 4.2).

4.8.6 Vias metabólicas do KEGG

Esta seção apresenta o resultado do mapeamento das proteínas traduzidas a partir dos produtos dos cDNAs em vias metabólicas do KEGG (item 4.4). Os resultados são apresentados numa tabela de links para cada categoria funcional de vias metabólicas (Figura 4). Além disso, para cada sequência mapeada, o usuário pode clicar num link e inspecionar uma figura da via metabólica na qual a proteína foi mapeada (Figura 5).

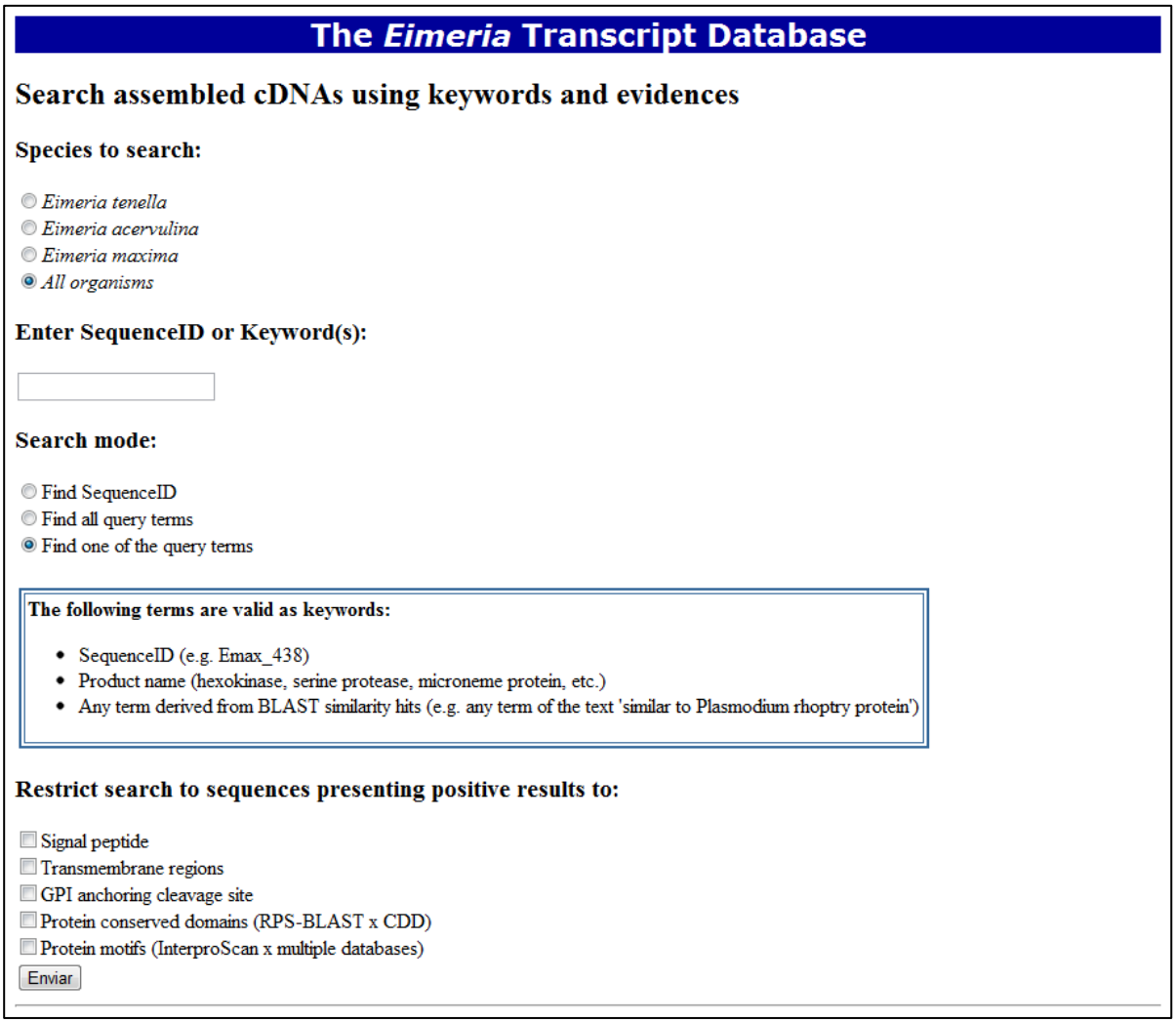
4.8.7 Banco de dados relacional

No portal também esta disponível uma base de dados relacional (Figura 13) construída com dados das três espécies de *Eimeria*, e as evidências funcionais coletadas pelos componentes listados no item 4.8.2.2.4. A base de dados pode ser utilizada como alternativa às páginas de anotação, mas apenas as evidências da ORF selecionada são apresentadas. Por isso, caso o usuário suspeite que houve algum erro durante a seleção da ORF correta, sugerimos que utilize as páginas de anotação convencionais. Esta base de dados ainda é experimental e não contém ainda as evidências relacionadas à classificação em grupos ortólogos ou mapeamento em vias metabólicas. Esses dados deverão ser adicionados no futuro. As buscas podem ser feitas utilizando-se vários filtros e parâmetros. O usuário o pode por exemplo selecionar qual espécie de *Eimeria* pretende consultar, ou ainda selecionar todas as espécies. Caso o usuário já possua o id da sequência que deseja visualizar (através do BLAST e/ou da lista de produtos), a busca pode ser feita diretamente pela inserção do *id* no campo “Enter SequenceID”. Caso o usuário insira um id que não esteja relacionado à espécie selecionada, nenhum resultado será retornado. Por exemplo, Eten_0009 é um id válido para uma sequência de *E. tenella.*

O usuário pode ainda inserir qualquer nome de produto, completo ou parcial, ou um termo derivado de uma anotação por BLAST. A busca pode ser restrita apenas às sequências cujo produto contenha todos os termos consultados. Por exemplo, caso o usuário busque por “serine protease” em todas as espécies, obterá 157 resultados relativos ao primeiro termo, “serine”, e 61 relativos ao segundo, “protease”. Caso o usuário restrinja a busca aos produtos com ambos os termos, o banco de dados retornará apenas sete sequências.

É possível ainda fazer a busca de forma que sequências relacionadas a qualquer termo da pesquisa sejam apresentadas. Utilizando-se o exemplo anterior, o usuário obtêm 211 resultados. Este número corresponde à soma dos 157 produtos encontrados utilizando o termo "serine" e dos 61 utilizando "protease", menos a interseção entre os dois grupos, sete sequências. As consultas podem ainda ser filtradas em função da existência de um ou mais tipos de evidências. Retornando ao exemplo anterior (busca em todas as espécies utilizando "serine protease"), se a busca for restrita apenas aos produtos que contenham evidências de motivos protéicos, a base de dados retornará apenas cinco sequências.

Figura 13 - Captura de tela da interface da base relacional do portal The *Eimeria* Transcript Database.



Fonte: Rangel (2011).

4.8.8 Download

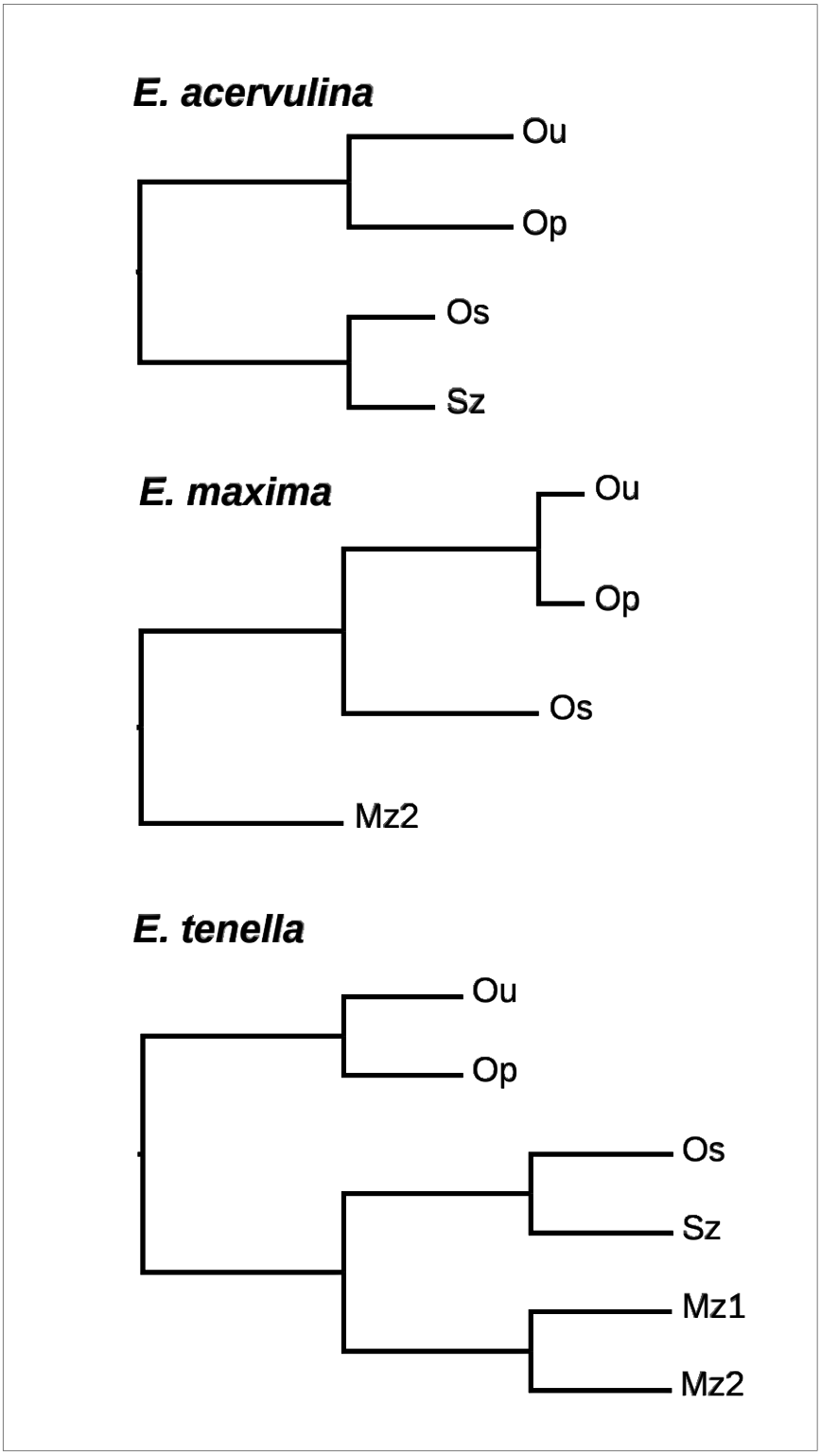
No portal também disponibilizamos para download as sequências em arquivos FASTA, compactados utilizando o programa gunzip. Estão disponíveis as sequências nucleotídicas dos cDNAs montados, assim como as sequências traduzidas dos produtos maiores que 50 ou 100 aminoácidos. Estes dados estão disponíveis para as três espécies de *Eimeria*.

4.9 Análise de expressão gênica com perfis digitais de expressão

Utilizamos o componente `assembly_cap3.pl` para extrair dados quantitativos de número de leituras por contig a partir dos arquivos de montagens em formato ace. Assim, os arquivos ace resultantes das montagens dos cDNAs das três espécies de *Eimeria* foram processados e arquivos em formato CSV foram gerados, listando para cada transcrito, o número de leituras de acordo com o estágio de desenvolvimento do parasita. Esses dados foram utilizados primeiramente para uma análise de agrupamento (clustering) hierárquico usando-se o programa Simcluster.

A Figura 14 apresenta os dendrogramas de distância observados para as três espécies de *Eimeria*. Conforme pode ser visto, os padrões de expressão observados estão claramente relacionados com a posição relativa de cada estágio de desenvolvimento no ciclo de vida dos parasitas. Assim, oocistos esporulados (Os), que contém no seu interior quatro esporocistos com dois esporozoítos cada um, apresentam um perfil de expressão mais próximo de formas esporozoítas (Sz) do que de oocistos não esporulados (Ou) ou em esporulação (Op). De forma semelhante, merozoítos de primeira (Mz1) e segunda geração (Mz2), em *E. tenella*, apresentam perfis de expressão mais próximos entre si do que com outros estágios do parasita. Para mais detalhes acerca dessa análise, favor consultar o manuscrito aceito para publicação (ANEXO A).

Figura 14 - Análise de agrupamento hierárquico utilizando perfis digitais de expressão de *Eimeria*. Foram utilizados dados de quantificação de leituras de EST/ORESTES de *E. acervulina*, *E. maxima* e *E. tenella*, derivadas de diferentes estágios de desenvolvimentos (Ou –oocistos não esporulados, Op – oocistos em fase de esporoblasto, Os – oocistos esporulados, Sz – esporozoítos, Mz1 – merozoítos de primeira geração, Mz2 – merozoítos de segunda geração). Método: distância de Pearson com agrupamento hierárquico completo.



Fonte: Rangel (2011).

5 CONCLUSÕES

- Foram desenvolvidos componentes de anotação automática para a plataforma EGene visando a classificação de proteínas em grupos ortólogos das bases KOG e eggNOG, o mapeamento de proteínas em vias metabólicas do KEGG, a geração de perfis digitais de expressão a partir de dados de montagem, e a integração dos resultados de anotação com o visualizador GBrowse.
- O desenvolvimento dos componentes de anotação descritos nesse trabalho, em conjunto com os demais componentes da plataforma EGene, permitiram anotar funcionalmente os transcriptomas de *E. acervulina*, *E. maxima* e *E. tenella*;
- Estabelecemos relações de ortologia entre proteínas de *Eimeria* spp. e organismos Apicomplexa;
- As análises de ortologia permitiram melhorar a classificação funcional de proteínas das eimerias em grupos do KOG através de anotação transitiva.
- Foram identificadas proteínas conservadas em organismos do filo Apicomplexa, sugerindo que possam ter algumas funções importantes e comuns aos diferentes organismos. Essas proteínas deverão ser consideradas para futuros estudos de atribuição funcional, bem como para o desenvolvimento de drogas.
- Foram identificadas proteínas restritas ao gênero *Eimeria*, as quais podem estar implicadas em mecanismos moleculares especificamente relacionados com o ciclo de vida desses parasitas.
- Perfis de expressão digital obtidos a partir de dados de montagem de transcritos revelaram um padrão altamente conservado nas três distintas espécies de *Eimeria*, uma associação com os distintos estágios de desenvolvimento analisados, e uma forte correlação com a ordem desses estágios no ciclo de vida do parasita.
- Foi criado o portal *The Eimeria Transcript Database*, que agrega informações relativas às sequências de transcritos de *Eimeria* spp. e as respectivas anotações funcionais. Essa base de dados pública poderá ajudar a comunidade científica que trabalha com *Eimeria* spp. a definir possíveis alvos moleculares para desenvolvimento de novas estratégias de controle da coccidioses da galinha doméstica.

REFERÊNCIAS*

Alexeyenko A, Tamas I, Liu G, Sonnhammer EL. Automatic clustering of orthologs and inparalogs shared by multiple proteomes. Bioinformatics. 2006 Jul; 22(14):e9-15

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.

Altenhoff AM, Schneider A, Gonnet GH, Dessimoz C. OMA 2011: orthology inference among 1000 complete genomes. Nucleic Acids Res. 2011 Jan; 39(Database issue):D289-94.

Altenhoff AM, Dessimoz C. Phylogenetic and functional assessment of orthologs inference projects and methods. PLoS Comput Biol. 2009 Jan; 5(1):e1000262.

Carver T, Berriman M, Tivey A, Patel C, Böhme U, Barrell BG, Parkhill J, Rajandream MA. Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. Bioinformatics. 2008 Dec; 24(23):2672-6.

Caspi R, Foerster H, Fulcher CA, Hopkinson R, Ingraham J, Kaipa P, Krummenacker M, Paley S, Pick J, Rhee SY, Tissier C, Zhang P, Karp PD. MetaCyc: a multiorganism database of metabolic pathways and enzymes. Nucleic Acids Res. 2006 Jan 1; 34(Database issue):D511-6.

Caspi R, Altman T, Dreher K, Fulcher CA, Subhraveti P, Keseler IM, Kothari A, Krummenacker M, Latendresse M, Mueller LA, Ong Q, Paley S, Pujar A, Shearer AG, Travers M, Weerasinghe D, Zhang P, Karp PD. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases.Nucleic Acids Res. 2011 Nov 18.

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.

Chen F, Mackey AJ, Stoeckert CJ Jr, Roos DS. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. Nucleic Acids Res. 2006 Jan 1;34(Database issue):D363-8.

Chen F, Mackey AJ, Vermunt JK, Roos DS. Assessing performance of orthology detection strategies applied to eukaryotic genomes. PLoS One. 2007 Apr 18;2(4):e383.

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. 21(12): 2812-3.

Ebersberger I, Strauss S, von Haeseler A. HaMStR: profile hidden markov model based search for orthologs in ESTs. BMC Evol Biol. 2009 Jul; 9:157

Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 2002 Apr; 30(7):1575-84.

*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]

Fast NM, Kissinger JC, Roos DS, Keeling PJ. Nuclear-encoded, plastid-targeted genes suggest a single common origin for apicomplexan and dinoflagellate plastids. Mol Biol Evol. 2001 Mar; 18(3):418-26.

Fitch WM. Distinguishing homologous from analogous proteins. Syst Zool. 1970 Jun; 19(2):99-113

Friedberg I. Automated protein function prediction--the genomic challenge. Brief Bioinform. 2006 Sep; 7(3):225-42.

Green P. PHRAP. 1996. Disponível em: http://www.genome.washington.edu/uwgc/analysistools/phrap.htm.

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

Instituto Sanger. Rascunho do genoma de *E. tenella* [ftp]. Disponível em: <ftp://ftp.sanger.ac.uk/pub/pathogens/Eimeria/tenella>. [2011 Nov 10].

Jensen LJ, Julien P, Kuhn M, von Mering C, Muller J, Doerks T, Bork P. eggNOG: automated construction and annotation of orthologous groups of genes. Nucleic Acids Res. 2008. 36(Database issue): D250-4.

Jun J, Mandoiu II, Nelson CE. Identification of mammalian orthologs using local synteny. BMC Genomics. 2009 Dec; 10:630.

Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000 Jan; 28(1):27-30.

Karp PD, Paley S, Romero P. The Pathway Tools software. Bioinformatics. 2002; 18 Suppl 1:S225-32.

Kawashima S, Katayama T, Sato Y, Kanehisa M. KEGG API: A web service using SOAP/WSDL to access the KEGG system. Genome Informatics 2005. 14: 673-674.

Kristensen DM, Wolf YI, Mushegian AR, Koonin EV. Computational methods for Gene Orthology inference. Brief Bioinform. 2011 Sep; 12(5):379-91.

Kuo CH, Wares JP, Kissinger JC. The Apicomplexan whole-genome phylogeny: an analysis of incongruence among gene trees. Mol Biol Evol. 2008 Dec; 25(12):2689-98.

Levine ND. Progress in taxonomy of the Apicomplexan protozoa. J Protozool. 1988 Nov; 35(4):518-20.

Li L, Stoeckert CJ Jr, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003 Sep; 13(9):2178-89.

Li L, Brunk BP, Kissinger JC, Pape D, Tang K, Cole RH, Martin J, Wylie T, Dante M, Fogarty SJ, Howe DK, Liberator P, Diaz C, Anderson J, White M, Jerome ME, Johnson EA, Radke JA, Stoeckert CJ Jr, Waterston RH, Clifton SW, Roos DS, Sibley LD. Gene discovery in the apicomplexa as revealed by EST sequencing and assembly of a comparative gene database. Genome Res. 2003 Mar; 13(3):443-54.

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.

Malaysia Genome Institute. Rascunho do genoma de *E. maxima* [página da internet]. Disponível em: <http://www.genomemalaysia.gov.my/emaxdb>. [2011 Nov 10]

Mao X, Cai T, Olyarchuk JG, Wei L. Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. Bioinformatics. 2005 Oct; 21(19):3787-93.

Meyer F, Goesmann A, McHardy AC, Bartels D, Bekel T, Clausen J, Kalinowski J, Linke B, Rupp O, Giegerich R, Pühler A. GenDB--an open source genome annotation system for prokaryote genomes. Nucleic Acids Res. 2003 Apr; 31(8):2187-95.

Miska KB, Fetterer RH, Barfield RC. Analysis of transcripts expressed by *Eimeria tenella* oocysts using subtractive hybridization methods. J Parasitol. 2004 Dec; 90(6):1245-52.

Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. 2007 Jul; 35(Web Server issue): W182-5.

Muller J, Szklarczyk D, Julien P, Letunic I, Roth A, Kuhn M, Powell S, von Mering C, Doerks T, Jensen LJ, Bork P. eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. Nucleic Acids Res. 2010 Jan; 38(Database issue):D190-5.

National Center for Biotechnology Information. KOGnitor [página da internet]. Disponível em: < http://www.ncbi.nlm.nih.gov/COG/grace/kognitor.html>. [2011 Nov 10]a.

National Center for Biotechnology Information. Disponibilização do dignitor [ftp]. Disponível em: < ftp://ftp.ncbi.nih.gov/pub/tatusov/dignitor/old/>. [2011 Nov 10]b.

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.

Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 1999 Jan; 27(1):29-34.

Ostlund G, Schmitt T, Forslund K, Köstler T, Messina DN, Roopra S, Frings O, Sonnhammer

EL. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Res. 2010 Jan; 38(Database issue):D196-203

Peterson ME, Chen F, Saven JG, Roos DS, Babbitt PC, Sali A. Evolutionary constraints on structural similarity in orthologs and paralogs. Protein Sci. 2009 Jun; 18(6):1306-15.

Rambault, A. Figtree [página da internet]. Disponível em: <http://tree.bio.ed.ac.uk/ software/figtree/>. [2011 Nov 10].

Remm M, Storm CE, Sonnhammer EL. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J Mol Biol. 2001 Dec; 314(5):1041-52.

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.

Schmid R, Blaxter ML. annot8r: GO, EC and KEGG annotation of EST datasets. BMC Bioinformatics. 2008 Apr; 9:180.

Schneider A, Dessimoz C, Gonnet GH. OMA Browser--exploring orthologous relations across 352 complete genomes. Bioinformatics. 2007 Aug; 23(16):2180-2.

Shirley MW, Harvey DA. *Eimeria tenella*: genetic recombination of markers for precocious development and arprinocid resistance. Appl Parasitol. 1996 Dec; 37(4):293-9.

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

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

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

Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A e Lewis S. The generic genome browser: a building block for a model organism system database. Genome Res. 2002 Oct; 12(10):1599-610

Storm CE, Sonnhammer EL. Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. Bioinformatics. 2002 Jan;18(1):92-9.

Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. Science. 1997 Oct; 278(5338): 631-7.

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA. The COG database: an updated version includes eukaryotes. BMC Bioinformatics. 2003 Sep;4:41.

Toso MA e Omoto CK. *Gregarina niphandrodes* may lack both a plastid genome and organelle. J Eukaryot Microbiol. 2007 Jan-Feb; 54(1):66-72.

Tenter AM, Barta JR, Beveridge I, Duszynski DW, Mehlhorn H, Morrison DA, Thompson RC e Conrad PA. The conceptual basis for a new classification of the coccidia. Int J Parasitol. 2002 May; 32(5):595-616.

Vêncio RZ, Varuzza L, de B Pereira CA, Brentani H, Shmulevich I. Simcluster: clustering enumeration gene expression data on the simplex space. BMC Bioinformatics. 2007 Jul; 8:246.

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.

Zhu G, Marchewka MJ e Keithly JS. *Cryptosporidium parvum* appears to lack a plastid genome. Microbiology. 2000 Feb; 146 (Pt 2):315-21.

Zmasek CM, Eddy SR. RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. BMC Bioinformatics. 2002 May; 3:14.

ANEXO A

A COMPARATIVE TRANSCRIPTOME ANALYSIS REVEALS EXPRESSION PROFILES CONSERVED ACROSS THREE *EIMERIA* SPP. OF DOMESTIC FOWL AND ASSOCIATED WITH MULTIPLE DEVELOPMENTAL STAGES ✯

Jeniffer Novaes [a,1], Luiz Thibério L.D. Rangel [a,1], Milene Ferro [a], Ricardo Y. Abe [b], Alessandra P.S. Manha [a], Joana C.M. de Mello [a], Leonardo Varuzza [b], Alan M. Durham [b], Alda Maria B.N. Madeira[a,*], Arthur Gruber [a,*]

[a] *Departamento de Parasitologia, Instituto de Ciências Biomédicas, Universidade de São Paulo, Av. Prof. Lineu Prestes, 1374, São Paulo SP, 05508-000, Brazil.*

[b] *Departamento de Ciências da Computação, Instituto de Matemática e Estatística, Universidade de São Paulo, Rua do Matão, 1010, Bloco C, São Paulo SP, 05508-000, Brazil.*

[1]Both authors contributed equally to the work

[*] Corresponding authors. Address: Departamento de Parasitologia, Instituto de Ciências Biomédicas, Universidade de São Paulo, Av. Prof. Lineu Prestes, 1374, São Paulo SP, 05508-000, Brazil. Tel.: +55 11 30917274; fax: +55 11 30917417.

E-mail addresses: argruber@usp.br (A. Gruber); albackx@usp.br (A.M.B.N. Madeira).

✯Note: All ORESTES sequence data reported in this paper are available in the GenBank™ database under the accession codes **EB741091** to **EB741366** and **JK217768** to **JK263396**. The sequences are also available at http://www.coccidia.icb.usp.br/eimeriatdb/.

Note: Supplementary data associated with this article.

ABSTRACT

Coccidiosis of the domestic fowl is a worldwide disease caused by seven species of protozoan parasites of the genus *Eimeria*. The genome of the model species, *Eimeria tenella*, presents a complexity of 55-60 MB distributed in 14 chromosomes. Relatively few studies have been undertaken to unravel the complexity of the transcriptome of *Eimeria* parasites. We report here the generation of more than 45,000 open reading frame expressed sequence tag (ORESTES) cDNA reads of *E. tenella*, *Eimeria maxima* and *Eimeria acervulina*, covering several developmental stages: unsporulated oocysts, sporoblastic oocysts, sporulated oocysts, sporozoites and second generation merozoites. All reads were assembled to constitute gene indices and submitted to a comprehensive functional annotation pipeline. In the case of *E. tenella*, we also incorporated publicly available ESTs to generate an integrated body of information. Orthology analyses have identified genes conserved across different apicomplexan parasites, as well as genes restricted to the genus *Eimeria*. Digital expression profiles obtained from ORESTES/EST countings, submitted to clustering analyses, revealed a high conservation pattern across the three *Eimeria* spp. Distance trees showed that unsporulated and sporoblastic oocysts constitute a distinct clade in all species, with sporulated oocysts forming a more external branch. This latter stage also shows a close relationship with sporozoites, whereas first and second generation merozoites are more closely related to each other than to sporozoites. The profiles were unambiguously associated with the distinct developmental stages and strongly correlated with the order of the stages in the parasite life cycle. Finally, we present The *Eimeria* Transcript Database (http://www.coccidia.icb.usp.br/eimeriatdb), a website that provides open access to all sequencing data, annotation and comparative analysis. We expect this repository to represent a useful resource to the *Eimeria* scientific community, helping to define potential candidates for the development of new strategies to control coccidiosis of the domestic fowl.

Keywords: *Eimeria*; Coccidia; Apicomplexa; Transcriptome analysis; Sequence annotation; Orthology; Gene expression profiles; Clustering analysis

1. Introduction

Coccidiosis of the domestic fowl is a worldwide disease caused by seven species of protozoan parasites of the genus *Eimeria*. Parasites are highly host-specific, present a

monoxenous life cycle and are transmitted via the oral-fecal route. High host population densities, associated with continuous physical contact with fecal droppings, are directly involved with the high morbidity of the disease. *Eimeria* parasites colonize intestinal cells and may lead to massive epithelial destruction. As a consequence, the host may present diarrhea, malabsorption and a decrease in weight gain. Coccidiosis is a major cause of economical losses in broiler chicken production (Allen and Fetterer, 2002), due to direct effects such as poor feed conversion or to indirect costs associated with control measures that include the preventive use of anticoccidial drugs and/or vaccines (Shirley et al., 2005).

Eimeria tenella, the most widely studied species, presents a genomic complexity of approximately 55-60 MB distributed in 14 chromosomes, with an estimated GC content of 53% (Shirley, 2000; Shirley and Harvey, 2000; Chapman and Shirley, 2003). The complete sequence of the chromosome 1 of *E. tenella* has been determined (Ling et al., 2007). This chromosome revealed, as a major feature, a segmentation pattern consisting of gene-rich regions associated with a high content of short tandemly repeated sequences, and regions with poor gene density and low repetitive content. A whole-genome sequencing (WGS) project of *E. tenella* is underway at the Sanger Institute (UK), and a draft sequence is publicly available (http://www.sanger.ac.uk/resources/downloads/protozoa/eimeria-tenella.html). In *Eimeria maxima*, a WGS project is also being carried out and assembled draft sequences are available at the *Emax*DB site in Malaysia (http://www.genomemalaysia.gov.my/emaxdb/).

In addition to genome sequencing, a good characterization of the parasite transcriptome may lead to a better understanding of the parasite biology and the development of new control strategies. Some reports in the literature have described the generation and analysis of *E. tenella* expressed sequence tags (ESTs) of second generation merozoites (Mz2) (Wan et al., 1999), sporozoites (Sz) (Ng et al., 2002) or both stages (Li et al., 2003). In terms of oocysts, Miska et al. (2004) and Han et al. (2010) reported the generation of a few hundred ESTs for unsporulated (Ou) and sporulated oocysts (Os). Finally, almost 10,000 ESTs have been generated by the Institute for Animal Health/Sanger Institute (UK) for in vitro-cultured first generation merozoites (Mz1), Ou and Sz (Rajandream et al., 2006 – unpublished data available on GenBank). Altogether, these works have generated circa 40,000 ESTs of *E. tenella*. In the case of *Eimeria acervulina*, the only available work (Miska et al., 2008) has described the generation of 1,847 ESTs derived from Mz2s and some schizonts. Similarly, in *E. maxima* there is only a single report describing the generation of 2,680 ESTs from merozoites (Schwarz et al., 2010). With the exception of Li et al. (2003), who performed a

comparative study involving several apicomplexan parasites and a large number of ESTs of *E. tenella*, most works were restricted to relatively small numbers of ESTs. Considering the status reported above, there is still a lack of large-scale cDNA sequencing of *Eimeria* parasites, especially of *E. acervulina* and *E. maxima* which, together with *E. tenella*, represent the most prevalent and economically relevant species for poultry production. Moreover, most of the researchers have focused on zoite stages rather than oocysts, probably because these infective stages may contain many potential targets to inhibit host cell adhesion and invasion by the parasites. Oocysts, on the other hand, are shed in the feces of infected birds and must undergo a sporulation process in the environment before they become infective. Possible control measures directed towards oocysts would benefit by better knowledge of the molecular aspects that underlie sporogony. In view of the currently available information on the *Eimeria* transcriptome, we aimed in the present work to fulfill some identified gaps by (i) sampling various developmental stages of *E. acervulina*, *E. maxima* and *E. tenella*; (ii) establishing gene indices; (iii) performing a comprehensive functional annotation; and (iv) providing a comparative analysis of the transcriptome in three *Eimeria* spp. Finally, to make this integrated body of information available to the scientific community, we present The *Eimeria* Transcript Database (http://www.coccidia.icb.usp.br/eimeriatdb), a web repository that provides open access to all sequencing data, annotation and comparative analysis.

2. MATERIALS AND METHODS

2.1. Parasite strains and propagation

The Houghton (H) strains of *E. acervulina*, *E. maxima* and *E. tenella* were used throughout this work. The parasites were propagated by oral infection of 3 to 4 week old male chicks (Bovans White) reared in a coccidia-free environment with ad libitum supply of filtered water and anticoccidial- and antibiotic-free feed, following standard protocols (Long et al., 1976; Shirley, 1995). Experimental procedures employing animals followed the institutional guidelines for the care and use of animals for research purposes, and the institutional animal ethics committee has formally approved the use of experimentation animals in this work. All parasite samples were submitted to a PCR species discrimination assay (Fernandez et al., 2003) to check the purity of the respective samples.

2.2. Developmental stages

We utilized five distinct developmental stages/phases of the parasites: Ou, sporulating oocysts (late sporoblast phase Op), Os, Sz and Mz2. Ou of *E. tenella* were collected directly from the ceca at 7 days p.i., and immediately purified through a chemical degradation of the tissue with sodium hypochlorite as described (Shirley, 1995). In the case of *E. acervulina* and *E. maxima*, oocysts were collected from fecal droppings at multiple 6 h intervals during p.i. periods of 100-120 h and 130-160 h, respectively. This procedure was aimed at minimizing the typical low synchronicity of the sporulation process observed in these species. Subsequent storage of these oocysts at 4 ^{o}C successfully arrested the progress of sporogony, as assessed by microscopic visualization. The oocysts were then purified according to Long et al. (1976) and sporulated by incubation in a 2.5% (w/v) potassium dichromate solution for 72 h at 28 ^{o}C under forced aeration. In order to produce a late sporoblast-predominant oocyst population, we first performed kinetic studies to determine the optimal sporulation time that produced a peak of sporoblast-phase oocysts. Thus, we collected oocyst samples from the sporulation suspension on an hourly basis and monitored the morphology of the oocysts under microscopy. Once the optimal sporulation times had been determined for the various species, large-scale experiments were then performed to produce Op. Oocysts were in vitro excysted and the released Sz purified with DE-52/nylon wool columns, following standard methods (Shirley, 1995; Chapman and Shirley, 2003). Mz2 of *E. maxima* and *E. tenella* were purified from small pieces of intestine/ceca from infected chickens as described (Shirley, 1995; Chapman and Shirley, 2003).

2.3. mRNA extraction

Parasite mRNA extraction was performed with RNase-free solutions and laboratory ware, using the µMACs mRNA isolation kit (Miltenyi Biotec GmbH, Bergisch Gladbach, Germany). A typical extraction employed a total of 1-2 x 10^{8} Ou, Op, Os, Sz or Mz2. Ou, Op and Os were cleaned with sodium hypochlorite solution (10-12% active chlorine) for 10 min at 4 ^{o}C, washed three times with deionized water and resuspended in 0.5 mL of a buffer solution (10 mM Tris–HCl, pH 8.0/50 mM EDTA, pH 8.0). The oocysts were fully disrupted by vortexing with a half volume of 425–600 mm acid-washed glass beads (Sigma-Aldrich Corp., St Louis, MO, USA). The lysate was then added to the Lysis Solution provided in the kit, and the mRNA extraction followed the manufacturer's protocol. In the case of Sz and

Mz2, cells were directly lysed with Lysis Solution and mRNA extraction followed the standard protocol of the kit. After purification, all mRNA samples were treated with 1 U of DNAse RQI RNAse-free (Promega Corporation, Madison, WI, USA) per 10 µl of the final mRNA eluate, according to instructions of the manufacturer. The mRNA concentration was measured by spectrophotometry (GeneQuant Pro™, GE Healthcare Biosciences, Pittsburgh PA, USA) and the samples were then aliquoted and stored at -80 °C until use. All samples were tested for genomic contamination using a PCR assay with single non-coding genomic sequence characterized amplified regions (SCAR) markers (Fernandez et al., 2003). RNA integrity was monitored by electrophoresis separation followed by mRNA visualization under UV light. We also checked the ability of the mRNA samples to generate long amplicons (>800 bp) by performing species-specific reverse transcription-PCR (RT-PCR) amplifications with different target genes (Supplementary Table S1).

2.4. Open reading frame EST (ORESTES) minilibrary construction and DNA sequencing

ORESTES minilibraries were constructed using a modification of the original protocol (Dias Neto et al., 1997, 2000), using one arbitrary primer per library. Briefly, 10 ng of DNase-treated polyA(+) RNA were subjected to reverse transcription at 37 °C for 90 min with 130 U of SuperScript II Reverse Transcriptase (Invitrogen Corporation, Carlsbad, CA, USA) and 22.5 pmols of 17- to 31-mer arbitrary primers in a final volume of 3.5 µL. After cDNA synthesis, the whole content of the single stranded cDNA was PCR amplified with a bead of the PuRE Taq Ready-to-GO PCR kit (GE Healthcare Biosciences) in a final volume of 25 µl. We generated ORESTES amplification profiles according to the protocol described by Verjovski-Almeida et al. (2003). Amplification profiles were electrophoretically separated on preparative 1% agarose gels and gel fragments containing bands in the range of 0.4 to 1.0 kb were selected and excised. The corresponding cDNAs were purified with spin-columns (GFX PCR DNA and Gel Band Purification Kit, GE Healthcare Biosciences) and cloned into a pGEM T-Easy vector (Promega Corporation). After transfection of *Escherichia coli* DH10B competent cells, we randomly selected 96 recombinant clones of each minilibrary for DNA sequencing. We performed all DNA sequencing reactions using the ABI PRISM Big Dye™ Terminator Cycle Sequencing kit (Applied Biosystems, Foster City CA, USA). All samples were run on an ABI PRISM 3700 Genetic Analyzer.

2.5. EST processing

We submitted all sequencing reads to a pre-processing pipeline created with EGene (Durham et al., 2005), a generic platform for pipeline construction. The configuration file of this pipeline, which corresponds to a detailed protocol of the method, is provided in Supplementary Data S1. Briefly, we submitted all sequences to quality evaluation, vector masking and end-trimming steps. All reads were also submitted to a series of filtering steps using the BLASTn program against databases of plastid, mitochondrial and ribosomal sequences. Additionally, we filtered out the reads against repetitive sequences and potential contaminants including bacterial, chicken and human genomes.

2.6. Third-party EST data

In the case of *E. tenella*, in addition to our ORESTES reads, we also used a set of approximately 40,000 ESTs available in public repositories (Supplementary Table S2). These sequences were derived from Os and Ou, Sz and Mz1 and Mz2. Whenever possible, we used trace files or standard chromatogram format (SCF) files; both formats which contain quality information. All sets of files were submitted to the same processing pipeline used for the ORESTES reads, as described in section 2.5. In the case of Mz1 reads, we incorporated an additional contaminant-filtering step against the bovine (*Bos taurus*) genome. This step was introduced due to the fact that this developmental stage has been in vitro cultured in MDBK cells, a bovine-derived cell line. All processed third-party data is available upon request.

2.7. Transcript reconstruction and annotation

We reconstructed the cDNAs using the CAP3 assembler (Huang and Madan, 1999) with parameters `p 90` and `y 250`. The assembled sequences were then submitted to an automatic annotation pipeline using the EGene 2 platform (AM Durham and A Gruber unpublished data, package available upon request). The configuration file of this pipeline is provided in Supplementary Data S2. First, ORFs were determined and coding sequences of at least 50 amino acid residues were translated. The protein sequences were then submitted to similarity searches using BLASTp (Altschul et al., 1997) against the non-redundant (`nr`) database. Hits were considered positive when presenting e-values below 10^{-6}. Additional analyses included conserved motif finding against the CDD database (Marchler-Bauer et al.,

2007) using RPS-BLAST (Marchler-Bauer et al., 2002), protein motif finding employing InterProScan (Mulder and Apweiler, 2007), signal peptide and transmembrane helix prediction with Phobius (Kall et al., 2004), and glycosylphosphatidylinositol (GPI) cleavage *site* (GPI-anchor) prediction with DGPI (Kronegg and Buloz – downloaded from http://129.194.185.165/dgpi/ on March 2008) The sequences were also submitted to orthology analyses using KOG (Tatusov et al., 2000, 2003), eggNOG (Muller et al., 2010) and KEGG Orthology (KO) (Aoki-Kinoshita and Kanehisa, 2007) databases. Using the identified KOs, we mapped the corresponding metabolic pathways. Identified InterPro entries were also used to map Gene Ontology (GO) terms. Finally, we generated annotation files in both Feature Table and GFF3 formats.

2.8. Orthology analysis and annotation enrichment

We incremented the annotation of transcriptome products of the three *Eimeria* spp. by performing an integrated orthology analysis with datasets of proteins predicted from the genomes of six apicomplexan parasites: *Toxoplasma gondii* ME49 (7,993 proteins, ToxoDB release 6.0, http://toxodb.org/toxo/), *Plasmodium falciparum* 3D7 (5,446 proteins, PlasmoDB, release 6.3, http://plasmodb.org/plasmo/), *Neospora caninum* NC_LIV (7,083 proteins, GeneDB, release September 2009, http://www.genedb.org/Homepage/Ncaninum), *Babesia bovis* T2Bo (3,703 proteins, GenBank accession number **AAXT00000000**), *Theileria annulata* Ankara clone C9 (3,795 proteins, GeneDB, release 15 July 2005, http://www.genedb.org/Homepage/Tannulata) and *Cryptosporidium parvum* Iowa II (3,805 proteins, CryptoDB, release 4.3, http://cryptodb.org/cryptodb/). Each proteome dataset was first submitted to functional classification using a local implementation of the KOG algorithm. We ran BLAST searches against the KOG database and used the following criteria to select positive hits: bitscore ≥ 28, identity percentage $\geq 20\%$ and minimum alignment block size ≥ 25. KOG IDs were mapped to all positive hits and the first KOG observed in three distinct organisms was ascribed to the query protein. To identify pairs of orthologous proteins across the different apicomplexans, we carried out an all-against-all comparison of the translated products of the nine organisms using InParanoid (Remm et al., 2001; Ostlund et al., 2010), in a total of 36 paired analyses. Next, we merged all pairwise ortholog clusters identified by InParanoid into multi-species clusters using MultiParanoid (Alexeyenko et al., 2006). Finally, we identified orthologous groups populated by proteins of the three *Eimeria* spp., as well as groups exclusively composed of *Eimeria*-derived proteins.

2.9. cDNA mapping

We mapped the reconstructed cDNA sequences onto the assembled genomes of *E. tenella* (downloaded from the Sanger Institute, release September 2010, ftp://ftp.sanger.ac.uk/pub/pathogens/Eimeria/tenella/) and *E. maxima* (downloaded in April 2011 from the Malaysia Genome Institute, 454 + Sanger contigs dataset, http://www.genomemalaysia.gov.my/emaxdb/). First, we used the DUST program (from the WU-BLAST 2.0 package, http://blast.wustl.edu/) to identify low-complexity regions in the genome sequences, followed by a soft-masking step (using lower-case characters) performed with a proprietary script. The transcript sequences were then mapped onto the genomes using BLAT (Kent, 2002) with the option `-mask=lower`, which prevents alignments from starting in masked regions, but permits them to extend through these low-complexity regions. In addition, we used a dataset of translated proteins of our assembled cDNA sequences to perform similarity searches against a database of predicted proteins in the *E. tenella* genome (8,786 protein sequences, release September 2010, downloaded from ftp://ftp.sanger.ac.uk/pub/pathogens/Eimeria/tenella/).

2.10. Gene expression profiling and hierarchical clustering

The assembly files produced by CAP3 for the transcript reconstruction step (Section 2.7) were utilized to generate digital expression profiles (Okubo et al., 1992; Audic and Claverie, 1997). We used proprietary scripts (available upon request) to convert the assembly files into comma-separated value (CSV) files containing countings of the number of reads composing each assembled sequence (contig) according to their respective source (developmental stage). The worksheet files were used as input for agglomerative hierarchical clustering analyses using the Simcluster package (Vencio et al., 2007). This program normalizes the different enumeration data, thus permitting libraries/experiments of different sample sizes to be used and compared. Tree files in Newick format were edited with FigTree (Rambaut, A., http://tree.bio.ed.ac.uk/software/figtree/, downloaded in March 2010).

2.11. The Eimeria Transcript Database

To make all data reported here publicly available, we constructed The *Eimeria*

Transcript Database (http://www.coccidia.icb.usp.br/eimeriatdb/), a comprehensive website comprising all collected evidence and automatic annotations of individual transcripts, integrated qualitative and quantitative analyses of orthology, GO and metabolic pathway mapping.

3. RESULTS AND DISCUSSION

3.1. Assessing the sporulation in multiple Eimeria spp.

There are only a few reports in the literature (Miska et al., 2004; Han et al., 2010) using Ou and Os to generate ESTs. In addition to these forms, we decided to assess the transcriptome of sporulating oocysts at the late sporoblast phase, a post-meiotic form (Wagenbach and Burns, 1969; Ferguson et al., 1978). Sporulation kinetics studies (described in Section 2.2) revealed a peak of late Op at 18 h for *E. tenella*. This period of time is longer than the 12 h period reported by Ryan et al. (2000) and Kinnaird et al. (2004), but is relatively close to the result of Wagenbach and Burns (1969). The apparent discrepancy in results can be ascribed to some potential variables, including the different parasite strains and/or quality of reagents used across the distinct laboratories. For *E. acervulina*, Wilson and Fairbairn (1961) observed 90% of the oocysts at the sporoblast phase after 10 h incubation at 30 ^{o}C. In our case, using incubation at 28 ^{o}C, we observed a final sporulation rate of 85-90%, with a peak of sporoblasts at 13 h. In *E. maxima*, Pittilo and Ball (1985) reported that within 21-30 h of incubation at 29 ^{o}C, the sporoblasts have been transformed into sporocysts containing sporozoites. Our results, using the same strain (H), are in good agreement, since we obtained a final sporulation rate of 75-80%, with a peak of sporoblasts at 23 h.

Regarding oocyst collection, *E. tenella* oocysts were obtained from the ceca in a single step and then immediately used for sporulation, a condition under which the process can be considered relatively synchronous (Wang et al., 1975; Ryan et al., 2000; Kinnaird et al., 2004). We observed overall sporulation rates of 80-90%, with a late-sporoblast peak containing circa 80-85% of the oocysts with the expected rounded-blast morphology, in agreement with what was reported by Ryan et al. (2000). In the case of *E. acervulina* and *E. maxima*, oocysts were collected from the feces during a period of time that comprises the peak of the patent period. Thus, sporulation cannot be considered synchronous in these species, as the oocysts shed at the beginning of the patent period might already have initiated

sporulation by the end of the collection period. Some authors have reported the use of sodium dithionite to produce an anaerobic condition that arrests sporulation (Wang, 1976; Ryan et al., 2000). In our case, we refrained from using dithionite to avoid any side effect that this molecule might exert in the metabolism of the parasites. Our pooled samples of oocysts of *E. acervulina* and *E. maxima* yielded peaks containing 65-70% Op. This reasonably high synchronicity stimulated us to use Op as distinct samples in our ORESTES sequencing project, thus representing an intermediate phase between Ou and Os.

3.2. Transcript sequencing and reconstruction

In this work, we constructed a total of 605 ORESTES cDNA minilibraries of *E. acervulina*, *E. maxima* and *E. tenella*. A total of 58,080 trace files were submitted to a pre-processing pipeline to remove low-quality reads and potential contaminants. In the end, we obtained 45,902 high quality ORESTES reads, covering various developmental stages of the parasite, including two invasive stages, Sz and Mz2s, and three phases of the oocyst stage: Ou, Op and Os. In order to obtain gene indices of the transcripts, we assembled the cDNA reads using a CAP3 assembler. At the time of analysis, there was no available EST data for *E. acervulina* and *E. maxima*, and to date relatively few reads have been publicly released for these species. Thus, cDNA assemblies for these species were performed solely with our locally generated data. Conversely, in the case of *E. tenella*, in addition to our ORESTES data, we combined our data with all publicly available EST sequences (Supplementary Tables S2 and S3). Table 1 shows the assembly results obtained for the three *Eimeria* spp. The total number of assembled cDNAs, including both contigs and singlets, was 3,413 for *E. acervulina*, 3,426 for *E. maxima* and 8,700 for *E. tenella*. The distribution of the contigs according to the number of reads is presented in Supplementary Fig. S1 for the three *Eimeria* spp. The charts show that most contigs are composed of a very small number of reads, whereas only a few contigs are populated by large amounts of reads. If one considers that the frequency of reads composing the contigs resembles the gene expression levels of the corresponding transcripts, this result may suggest that a small number of genes is highly expressed, while a large set of the remaining genes is expressed in tiny amounts. These findings stimulated us to check whether this kind of distribution is observed across specific developmental stages. Thus, we selected reads which originated from each of the six sampled stages of *E. tenella* and repeated the assembly step using these particular subsets. The same pattern of contig population has been observed in all stages (Supplementary Fig. S2), and

possibly reflects a typical gene expression profile in *Eimeria* parasites.

3.3. Transcriptome complexity

The cDNA assembly of *E. tenella* resulted in 8,700 distinct events (Table 1). This number does not necessarily reflect the real transcriptome complexity, as many of the assembled sequences may represent distinct non-overlapping regions of the same transcripts. Thus, the final number of unique transcripts covered by our data could be significantly lower. On the other hand, the data sampling is biased towards some developmental stages (oocysts and zoites), whereas some other stages such as micro- and macrogametocytes have not been sampled. Therefore, one should expect that some stage-specific expressed genes might not have been covered. Despite the aforementioned limitations, our estimated number of genes is fairly close to the complexity (8,786 proteins) of the predicted proteins translated from the *E. tenella* genome (Sanger Institute, release of September 2010). Assuming a homogeneous gene density across the different chromosomes and a genome complexity of 55 MB (Shirley, 2000), gene density can be estimated in 160 genes/MB (Table 2). This value is relatively similar to the gene density values observed for two other coccidian parasites: *T. gondii* (127 genes/MB) and *N. caninum* (114 genes/MB). Conversely, much higher gene densities have been observed in *P. falciparum* (237 genes/MB), and in the compact genomes of piroplasms and *Cryptosporidium* (423-474 genes/MB). Coccidian parasites present an oral-fecal monoxenous life cycle, whereas piroplasms such as *Theileria* and *Babesia* are obligatory heteroxenous hemoparasites. However, this apparent correlation between gene density and parasite biology is not fully supported, since *Cryptosporidium* presents an oral-fecal direct life cycle but shows a gene density similar to piroplasms. *Plasmodium*, on the other hand, is an obligatory heteroxenous hemoparasite but shows genome size and density values intermediate between coccidians and piroplasms. Ling et al. (2007) reported on *E. tenella* chromosome 1, a total of 216 genes distributed along 889,314 bp of assembled sequence (excluding sequence and physical gaps). The estimated gene density (242 genes/MB) is much higher than the values estimated with our reconstructed cDNA data or with the set of predicted proteins translated from the *E. tenella* genome. This striking discrepancy may suggest that the number of predicted genes in chromosome 1 could have been overestimated or, alternatively, that this chromosome might present a much higher gene density than the rest of the genome. In the case of *E. acervulina* and *E. maxima*, the cDNA assembly results indicate that we have covered only a fraction of the respective transcriptomes, since the total number of events

(contigs plus singlets) was approximately 3,400 sequences.

3.4. Functional annotation

We submitted all reconstructed cDNA sequences of the three *Eimeria* spp. to a comprehensive annotation pipeline (Section 2.7) and, whenever possible, classified the proteins into orthology groups and GO terms. The assembled transcript sequences, annotation data and supporting evidence are publicly available on The *Eimeria* Transcript Database site (http://www.coccidia.icb.usp.br/eimeriatdb/). We observed that more than 90% of the reconstructed cDNAs of the three *Eimeria* spp. presented a putative protein coding sequence (Table 3). However, only a relatively small subset (32-38%) of the predicted proteins presented positive BLAST results against the `nr` database (Table 3). These preliminary results clearly suggest that most of our predicted sequences either represented novel proteins or did not span conserved regions of currently known proteins. The high number of hypothetical proteins (62-68%) is not surprising, since more than 50% of ESTs from other well-studied apicomplexan parasites still remain unknown (Li et al., 2003; Cui et al., 2005; Dybas et al., 2008). Other studies employing ESTs from *Eimeria* spp. also showed high rates of uncharacterized proteins, varying from 47% to 73% (Wan et al., 1999; Ng et al., 2002; Miska et al., 2004, 2008; Schwarz et al., 2010).

When we performed an orthology analysis using the KOG database, we successfully classified 707 sequences of *E. acervulina*, 604 of *E. maxima* and 1,517 of *E. tenella* (Table 3). These slightly lower numbers, compared with BLAST against `nr` results, can be ascribed to the more complex set of requirements for a protein to be assigned to an orthologous group, and due the fact that the KOG database is composed of proteins derived from only seven eukaryotic organisms (Tatusov et al., 2003). In terms of functional category assignment, we observed a very similar pattern of sequence distribution across the three *Eimeria* spp. (Fig. 1). The most assigned categories for all species were O (posttranslation modification, protein turnover, chaperones), R (general prediction function only), T (signal transduction mechanisms) and A (RNA processing and modification). Despite the very different stage sampling of the cDNA reads used in each species (Supplementary Table S3), the overall functional category distribution is strikingly similar across the three *Eimeria* spp.. A possible explanation for this finding is that KOG is highly biased towards housekeeping proteins (Koonin, 2005). Alternatively, this result may also suggest that the vast majority of the genes

are not differentially expressed. Using eggNOG v2.0, a much more comprehensive database of orthologous groups (Muller et al., 2010), we obtained very similar results (not shown – data available on the project website).

We also classified the protein sets of the three *Eimeria* spp. using the three ontology domains of the GO database. We were able to categorize 644 reconstructed sequences of *E. acervulina*, 541 of *E. maxima* and 1,404 of *E. tenella* in GO terms of any ontology (Table 3). A detailed list of identified GO terms is available on the project website. An overview of the distribution of GO term assignments for the three ontology domains is presented in Fig. 2. Similar to what has been observed in KOG functional categorization, most of the GO assignments were quantitatively similar across the different *Eimeria* spp. In terms of biological process ontology, the most prevalent GO terms were related to primary metabolism, protein metabolism and modification, nucleoside/nucleotide and nucleic acid metabolism, and biosynthetic process.

Aiming to enrich the protein annotation, we also carried out a series of orthology analyses using protein datasets of six apicomplexan genomes, plus the translated products of the three *Eimeria* spp. First, we submitted all proteins to a functional classification using the KOG database and then we determined the orthology groups among all proteins. A KOG analysis of the apicomplexan organisms allowed us to functionally classify 44.9% to 56.7% of the proteins (data not shown). In total, we obtained 7,489 orthology groups of apicomplexans, comprising 25,988 proteins (56.3%) out of 46,144. A total of 745 proteins of *E. acervulina*, 614 of *E. maxima* and 1,236 of *E. tenella* were clustered into any of these orthology groups (Table 3). We also identified 174 proteins of *E. acervulina*, 161 of *E. maxima* and 180 of *E. tenella* that are not functionally characterized but are conserved across Apicomplexa (Table 3). Also, we obtained 175 orthology groups containing protein representatives of the three *Eimeria* spp. (Table 3, Supplementary Tables S4 and S5). From this set, we were able to identify 27 orthology groups devoid of representatives of any of the non-eimerian apicomplexans, that is, they were restricted to the *Eimeria* genus. This set comprised 34 proteins of *E. acervulina*, 28 of *E. maxima* and 50 of *E. tenella* (Table 3, Supplementary Tables S4 and S5). A manual inspection and curation of the data revealed the presence of numerous surface antigen gene (SAG) proteins, a set of GPI-anchored variant surface antigens that constitute an important family of differentially expressed proteins. *Toxoplasma gondii* also possesses several SAG families but no homologs to *E. tenella* SAGs have been found (Tabares et al., 2004). Our orthology analysis is in agreement with this result, however it

additionally suggests that SAG proteins are not restricted to *E. tenella* but are also present as orthologs in *E. acervulina* and *E. maxima*. We also found in this *Eimeria*-specific protein dataset two sequences homologous to microneme proteins EtMIC2 and EtMIC3. In agreement with this result, orthologs to EtMIC2 (Blake et al., unpublished data available on GenBank under the accession code **CBX60033**) and EtMIC3 (Labbe et al., 2005) have also been found in *E. maxima*. For *E. acervulina*, we believe our result represents the first description of these proteins, and may suggest that this class of proteins is most probably present in other chicken *Eimeria* spp.

3.5. Transcript mapping

Since we obtained a relatively low number of functionally ascribed genes, we decided to assess whether the reconstructed transcript sequences could be mapped onto the respective parasite genome. A high rate of successful mapping would rule out the hypothesis of these sequences being heterologous and, therefore, derived from putative unknown contaminants. Thus, reconstructed cDNA sequences were submitted to similarity searches against genome assemblies available for *E. maxima* and *E. tenella*. We observed that 91.0% of the assembled transcripts (including contigs and singlets) were successfully mapped onto the genome in *E. maxima* and 92.6% in *E. tenella* (Supplementary Table S6), with coverages of 90.4% and 93.5%, respectively. This result shows that the alignments were not restricted to small block sizes but rather extended across most of the sequence length. It also strongly indicates that, at least in *E. maxima* and *E. tenella*, a significant contamination of the sequences with exogenous sources can be discarded. The Sanger Institute has released (September 2010) a set of 8,786 predicted protein sequences translated from the *E. tenella* genome based on 1,000 gene models, BLAST similarity results and transcriptome analysis. Thus, we decided to compare the translated proteins of our set of *E. tenella* assembled transcripts with this predicted protein dataset, in order to assess the level of congruence between them. Assuming a minimum identity percentage of 90% and a minimum alignment block size of 40 residues, we observed that 1,877 out of 3,684 (50.95%) of our contig-derived protein sequences showed positive hits, whereas 818 out of 4,306 (19.0%) singlet-derived sequences gave positive results. This much lower alignment rate at the protein level, compared with the result obtained with nucleotide sequences, can be ascribed to multiple causes. First, a significant fraction of the reconstructed transcript sequences could be covering non-coding regions such as 5’ and 3’ untranslated regions (UTRs). A potential contamination of the cDNAs with

nuclear DNA could also be incriminated. However, at least in the case of ORESTES data, rigorous controls have been employed to check the purity of the mRNA (see Section 2.3). From the genomic side, a relatively low accuracy of the protein prediction of the *E. tenella* genome could be the cause of the observed results. Alternatively, partial genome coverage could represent an additional explanation. Finally, the existence of splicing variants could lead to misalignments between the reconstructed cDNAs and genome sequences. Most probably, the low alignment rate observed at the protein level was a consequence of the combined effect of the various aforementioned factors.

3.6. Hierarchical clustering analysis

In order to assess how the gene expression profiles correlate with the distinct developmental stages of the parasites, we performed agglomerative clustering analyses (Section 2.10). Fig. 3 shows a Pearson distance tree obtained for expression data of the three *Eimeria* spp. As can be seen, the three oocyst phases are closely related to each other, and Op show an expression pattern more similar to Ou than to Os oocysts in all species. Os were grouped with Sz in *E. acervulina* and *E. tenella*. In the case of *E. maxima*, in which the Sz stage has not been sampled, Os formed a more external branch to the clade of Ou and Op, followed by Mz2. In *E. tenella*, Mz1 and Mz2 formed a sister clade to the clade composed of Os and Sz, thus constituting a major clade of the four zoite stages (if one considers Os a 'zoite' stage). This result clearly indicates that zoite stages share gene expression patterns and that Mz1 and Mz2 are more closely related to each other than to Sz. To assess the robustness of this approach, we tested two other distance methods and obtained very similar tree topologies (Supplementary Fig. S3).

It is worth mentioning that despite using three distinct *Eimeria* spp., different stage samplings (Supplementary Table S3), cDNA construction methods (EST, ORESTES or both), and distance metrics, the observed dendrograms showed a remarkable topological congruence. The high reproducibility of the clustering analyses suggests that our digital expression profiles are in fact reflecting the differential gene expression patterns of Ou, Op and Os, even though the sporulation process is not synchronous in *E. acervulina* and *E. maxima*. In addition, this result indicates that a great predominance of some highly expressed genes could be masking the effects of the asynchronous nature of sporulation. The hierarchical clustering analyses indicate that Ou and Op present gene expression patterns that

are more closely related to each other than to Os. This result is in agreement with what has been reported in the literature for many genes. In fact, Ellis and Thurlby (1991) have shown by in vitro translation in *E. maxima* that there is a burst of transcription of many mRNAs in the early phases of sporulation, followed by an overall reduction of their levels. Similar transcription patterns have also been reported for individual genes of *E. tenella* (Herbert et al., 1992; Jean et al., 2000; Kinnaird et al., 2004). In all cases, the turning point of expression has been observed at around 12 h, when Op predominate. On the other hand, an opposite expression pattern have also been observed, with an increase in the expression levels of some mRNAs after the formation of sporoblasts. Such a pattern is exemplified by a sporozoite surface antigen (Brothers et al., 1988), HSP90 (Péroval et al., 2006) and several microneme proteins (Ryan et al., 2000).

Our hierarchical clustering analyses also show that the expression profiles from samples collected at three different instances of sporulation seem to reflect the transformation of an oocyst into a zoite-type stage. Thus, to the extent that the oocyst becomes fully sporulated, its gene expression pattern concomitantly resembles to that of a Sz. This result is in agreement with Lal et al. (2009), who observed in proteomic analysis that late oocysts share many more proteins with Sz than with early oocysts. Miska et al. (2004), using ESTs derived from subtractive cDNA libraries, reported similar results, with Sz and Os sharing many common transcripts, whereas Ou revealed many transcripts not shared by either Sz or merozoites.

Despite the relative similarity of the expression profiles of Os and Sz, our hierarchical clustering analyses still reveal a clear distance between both stages. The excystation is an energy-demanding process (Coombs et al., 1997) that is possibly associated with some level of regulation of the gene expression, thus explaining the distinct expression patterns observed between resting sporozoites within fully sporulated oocysts, and newly excysted sporozoites. In addition, some other metabolic alterations might enable the sporozoite to adhere to and subsequently invade host cells. Time course experiments during the excystation process could better clarify the role of specific genes in the conversion of a dormant parasite into an active host cell invader.

Another interesting result relates to Mz1 and Mz2, which seem to share a very similar gene expression pattern. Regardless of the distance method used, these stages always grouped in the same clade. A preliminary differential gene expression analysis (data not shown), based on digital profiles, revealed that Mz1 and Mz2 share many hypothetical and ribosomal

proteins. In addition, many differentially expressed genes have been identified between both stages. Thus, Mz1 presents a much higher expression and variety of ribosomal proteins than Mz2. On other hand, Mz2 expresses a large number of SAGs that are not detected in Mz1. Marked differences in polypeptide profiles between Mz1 and Mz2 have also been observed by Tomley (1994).

Summarising, the hierarchical clustering analyses revealed a high conservation pattern across the three *Eimeria* spp. Also, the expression profiles of the distinct developmental stages strongly correlate with the order of these stages in the life cycle of the parasites. This result indicates that changes in mRNA levels play an important role in regulating stage-specific functionality, similar to what has been found by Radke et al. (2005) in *T. gondii*.

3.7. The Eimeria Transcript Database

In order to make all data presented in this work publicly available, we decided to construct a website for easy data inspection and retrieval. The site contains all sequences for download, including unassembled ORESTES reads, reconstructed cDNA sequences and conceptually translated protein products. A BLAST engine permits users to make their own similarity searches against assembled and unassembled data of the present work. Also, integrated and quantitative analysis of GO, orthology and metabolic pathway mappings are available. Finally, we provide a specific page for each reconstructed transcript where a comprehensive annotation is available in Feature Table and GFF3 annotation formats. All evidence supporting the ascribed function can be easily inspected through appropriate links to outputs of the bioinformatics programs utilized in the analyses. Cross-references of *Eimeria* proteins successfully mapped to orthologous groups of apicomplexan organisms are also provided. Finally, a preliminary relational database was constructed, allowing users to make complex queries.

3.8. CONCLUSIONS

This work presents a broad and comparative analysis of several developmental stages for three of the most relevant *Eimeria* spp. of domestic fowl. Our generated cDNA sequencing data is complementary to the genome sequencing projects of *E. tenella* and *E. maxima*, and may be useful to validate current gene predictions, constitute gene model training sets and

define exon-intron boundaries. Our analyses have identified genes that are conserved across different apicomplexan parasites, suggesting that they may play some common and important roles in this group of parasites, and should be considered for future studies on functional assignment and potential application as drug targets. Digital expression profiles revealed a high conservation pattern across the three *Eimeria* spp. and an unambiguous correlation with multiple developmental stages. In conclusion, we believe that the gene survey study and the associated analyses reported here, together with the publicly released database, constitute an important information resource for the *Eimeria* scientific community, and might help to define potential candidates for the development of new strategies to control coccidiosis of the domestic fowl.

ACKNOWLEDGEMENTS

The authors are in debted to Martin W. Shirley (formerly at the Institute for Animal Health, UK) for kindly providing the H strains of *Eimeria acervulina*, *Eimeria maxima* and *Eimeria tenella*. We wish to thank our colleagues of the *Eimeria* Genome Consortium for publicly releasing unpublished sequencing data of *E. tenella* (Martin W. Shirley and Fiona M. Tomley – formerly at the Institute for Animal Health; Matt Berriman – Wellcome Trust Sanger Institute, UK) and *E. maxima* (Kiew-Lian Wan - Universiti Kebangsaan, Malaysia; Damer P. Blake - Royal Veterinary College, University of London, UK). We would also like to thank Katarzyna Miska (USDA, USA) for sending us trace files of *E. tenella* ESTs. Granja Kunitomo (Mogi das Cruzes, Brazil) generously supplied 1 day old chicks used in this work. The technical assistance of Luciana Terumi Nagao, Cleonice da Silva and Lívia Rodrigues da Silva is also acknowledged. This work was supported by FAPESP (São Paulo, Brazil - grant 03/14031-3). J.N. and A.P.S.M. received scholarships from the National Council for Scientific and Technological Development (CNPq, Brazil) and the work presented herein formed part of their Ph.D. theses. L.T.L.D.R. received a scholarship from FAPESP (2009/12643-8) and the work presented herein formed part of his M.Sc. thesis. A.G., A.M.B.N.M. and A.M.D. received Productivity Research fellowships from CNPq.

References

Alexeyenko, A., Tamas, I., Liu, G., Sonnhammer, E.L., 2006. Automatic clustering of orthologs and inparalogs shared by multiple proteomes. Bioinformatics 22, e9-15.

Allen, P.C., Fetterer, R.H., 2002. 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 15, 58-65.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25, 3389-3402.

Aoki-Kinoshita, K.F., Kanehisa, M., 2007. Gene annotation and pathway mapping in KEGG. Methods Mol Biol 396, 71-91.

Audic, S., Claverie, J.M., 1997. The significance of digital gene expression profiles. Genome Res 7, 986-995.

Brothers, V.M., Kuhn, I., Paul, L.S., Gabe, J.D., Andrews, W.H., Sias, S.R., McCaman, M.T., Dragon, E.A., Files, J.G., 1988. Characterization of a surface antigen of *Eimeria tenella* sporozoites and synthesis from a cloned cDNA in *Escherichia coli*. Mol Biochem Parasitol 28, 235-247.

Chapman, H.D., Shirley, M.W., 2003. The Houghton strain of *Eimeria tenella*: a review of the type strain selected for genome sequencing. Avian Pathol 32, 115-127.

Coombs, G.H., Denton, H., Brown, S.M., Thong, K.W., 1997. Biochemistry of the coccidia. Adv Parasitol 39, 141-226.

Cui, L., Fan, Q., Hu, Y., Karamycheva, S.A., Quackenbush, J., Khuntirat, B., Sattabongkot, J., Carlton, J.M., 2005. Gene discovery in *Plasmodium vivax* through sequencing of ESTs from mixed blood stages. Mol Biochem Parasitol 144, 1-9.

Dias Neto, E., Correa, R.G., Verjovski-Almeida, S., Briones, M.R., Nagai, M.A., da Silva, W., Jr., Zago, M.A., Bordin, S., Costa, F.F., Goldman, G.H., Carvalho, A.F., Matsukuma, A., Baia, G.S., Simpson, D.H., Brunstein, A., de Oliveira, P.S., Bucher, P., Jongeneel, C.V., O'Hare, M.J., Soares, F., Brentani, R.R., Reis, L.F., de Souza, S.J., Simpson, A.J., 2000. Shotgun sequencing of the human transcriptome with ORF expressed sequence tags. Proc Natl Acad Sci U S A 97, 3491-3496.

Dias Neto, E., Harrop, R., Correa-Oliveira, R., Wilson, R.A., Pena, S.D., Simpson, A.J., 1997. 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 186, 135-142.

Durham, A.M., Kashiwabara, A.Y., Matsunaga, F.T., Ahagon, P.H., Rainone, F., Varuzza, L., Gruber, A., 2005. EGene: a configurable pipeline generation system for automated sequence analysis. Bioinformatics 21, 2812-2813.

Dybas, J.M., Madrid-Aliste, C.J., Che, F.Y., Nieves, E., Rykunov, D., Angeletti, R.H., Weiss, L.M., Kim, K., Fiser, A., 2008. Computational analysis and experimental validation of gene predictions in *Toxoplasma gondii*. PLoS One 3, e3899.

Ellis, J., Thurlby, T., 1991. Changes in the messenger RNA population during sporulation of *Eimeria maxima*. Parasitology 102 Pt 1, 1-8.

Ferguson, D.J.P., Birchandersen, A., Hutchison, W.M., Siim, J.C., 1978. Light and Electron-Microscopy on Sporulation of Oocysts of *Eimeria brunetti* .1. Development of Zygote and Formation of Sporoblasts. Acta Pathologica Et Microbiologica Scandinavica Section B-Microbiology 86, 1-11.

Fernandez, S., Pagotto, A.H., Furtado, M.M., Katsuyama, A.M., Madeira, A.M., Gruber, A., 2003. A multiplex PCR assay for the simultaneous detection and discrimination of the seven *Eimeria* species that infect domestic fowl. Parasitology 127, 317-325.

Han, H.Y., Lin, J.J., Zhao, Q.P., Dong, H., Jiang, L.L., Xu, M.Q., Zhu, S.H., Huang, B., 2010. Identification of differentially expressed genes in early stages of *Eimeria tenella* by suppression subtractive hybridization and cDNA microarray. J Parasitol 96, 95-102.

Herbert, R.G., Pasternak, J.J., Fernando, M.A., 1992. Characterization of *Eimeria tenella* unsporulated oocyst-specific cDNA clones. J Parasitol 78, 1011-1018.

Huang, X., Madan, A., 1999. CAP3: A DNA sequence assembly program. Genome Res 9, 868-877.

Jean, L., Grosclaude, J., Labbe, M., Tomley, F., Pery, P., 2000. Differential localisation of an *Eimeria tenella* aspartyl proteinase during the infection process. Int J Parasitol. 30, 1099-1107.

Kall, L., Krogh, A., Sonnhammer, E.L., 2004. A combined transmembrane topology and signal peptide prediction method. J Mol Biol 338, 1027-1036.

Kent, W.J., 2002. BLAT--the BLAST-like alignment tool. Genome Res 12, 656-664.

Kinnaird, J.H., Bumstead, J.M., Mann, D.J., Ryan, R., Shirley, M.W., Shiels, B.R., Tomley, F.M., 2004. EtCRK2, a cyclin-dependent kinase gene expressed during the sexual and asexual phases of the *Eimeria tenella* life cycle. Int J Parasitol. 34, 683-692.

Koonin, E.V., 2005. Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet 39, 309-338.

Labbe, M., de Venevelles, P., Girard-Misguich, F., Bourdieu, C., Guillaume, A., Pery, P., 2005. *Eimeria tenella* microneme protein EtMIC3: identification, localisation and role in host cell infection. Mol Biochem Parasitol 140, 43-53.

Lal, K., Bromley, E., Oakes, R., Prieto, J.H., Sanderson, S.J., Kurian, D., Hunt, L., Yates, J.R., 3rd, Wastling, J.M., Sinden, R.E., Tomley, F.M., 2009. Proteomic comparison of four *Eimeria tenella* life-cycle stages: unsporulated oocyst, sporulated oocyst, sporozoite and second-generation merozoite. Proteomics 9, 4566-4576.

Li, L., Brunk, B.P., Kissinger, J.C., Pape, D., Tang, K., Cole, R.H., Martin, J., Wylie, T., Dante, M., Fogarty, S.J., Howe, D.K., Liberator, P., Diaz, C., Anderson, J., White, M., Jerome, M.E., Johnson, E.A., Radke, J.A., Stoeckert, C.J., Jr., Waterston, R.H., Clifton, S.W., Roos, D.S., Sibley, L.D., 2003. Gene discovery in the apicomplexa as revealed by EST sequencing and assembly of a comparative gene database. Genome Res 13, 443-454.

Ling, K.H., Rajandream, M.A., Rivailler, P., Ivens, A., Yap, S.J., Madeira, A.M., Mungall, K., Billington, K., Yee, W.Y., Bankier, A.T., Carroll, F., Durham, A.M., Peters, N., Loo, S.S., Isa, M.N., Novaes, J., Quail, M., Rosli, R., Nor Shamsudin, M., Sobreira, T.J., Tivey, A.R., Wai, S.F., White, S., Wu, X., Kerhornou, A., Blake, D., Mohamed, R., Shirley, M., Gruber, A., Berriman, M., Tomley, F., Dear, P.H., Wan, K.L., 2007. Sequencing and analysis of chromosome 1 of *Eimeria tenella* reveals a unique segmental organization. Genome Res 17, 311-319.

Long, P.L., Millard, B.J., Joyner, L.P., Norton, C.C., 1976. A guide to laboratory techniques used in the study and diagnosis of avian coccidiosis. Folia Vet Lat 6, 201-217.

Marchler-Bauer, A., Anderson, J.B., Derbyshire, M.K., DeWeese-Scott, C., Gonzales, N.R., Gwadz, M., Hao, L., He, S., Hurwitz, D.I., Jackson, J.D., Ke, Z., Krylov, D., Lanczycki, C.J., Liebert, C.A., Liu, C., Lu, F., Lu, S., Marchler, G.H., Mullokandov, M., Song, J.S., Thanki, N., Yamashita, R.A., Yin, J.J., Zhang, D., Bryant, S.H., 2007.

CDD: a conserved domain database for interactive domain family analysis. Nucleic Acids Res 35, D237-240.

Marchler-Bauer, A., Panchenko, A.R., Shoemaker, B.A., Thiessen, P.A., Geer, L.Y., Bryant, S.H., 2002. CDD: a database of conserved domain alignments with links to domain three-dimensional structure. Nucleic Acids Res 30, 281-283.

Miska, K.B., Fetterer, R.H., Barfield, R.C., 2004. Analysis of transcripts expressed by *Eimeria tenella* oocysts using subtractive hybridization methods. J Parasitol 90, 1245-1252.

Miska, K.B., Fetterer, R.H., Rosenberg, G.H., 2008. Analysis of transcripts from intracellular stages of *Eimeria acervulina* using expressed sequence tags. J Parasitol 94, 462-466.

Mulder, N., Apweiler, R., 2007. InterPro and InterProScan: tools for protein sequence classification and comparison. Methods Mol Biol 396, 59-70.

Muller, J., Szklarczyk, D., Julien, P., Letunic, I., Roth, A., Kuhn, M., Powell, S., von Mering, C., Doerks, T., Jensen, L.J., Bork, P., 2010. eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. Nucleic Acids Res 38, D190-195.

Ng, S.T., Sanusi Jangi, M., Shirley, M.W., Tomley, F.M., Wan, K.L., 2002. Comparative EST analyses provide insights into gene expression in two asexual developmental stages of *Eimeria tenella*. Exp Parasitol 101, 168-173.

Okubo, K., Hori, N., Matoba, R., Niiyama, T., Fukushima, A., Kojima, Y., Matsubara, K., 1992. Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. Nat Genet 2, 173-179.

Ostlund, G., Schmitt, T., Forslund, K., Kostler, T., Messina, D.N., Roopra, S., Frings, O., Sonnhammer, E.L., 2010. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Res 38, D196-203.

Péroval, M., Péry, P., Labbé, M., 2006. The heat shock protein 90 of *Eimeria tenella* is essential for invasion of host cell and schizont growth. Int J Parasitol. 36, 1205-1215.

Pittilo, R.M., Ball, S.J., 1985. Ultrastructural observations on the sporogony of *Eimeria maxima*. Int J Parasitol. 15, 617-620.

Radke, J.R., Behnke, M.S., Mackey, A.J., Radke, J.B., Roos, D.S., White, M.W., 2005. The

transcriptome of *Toxoplasma gondii*. BMC Biol 3, 26.

Remm, M., Storm, C.E., Sonnhammer, E.L., 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J Mol Biol 314, 1041-1052.

Ryan, R., Shirley, M., Tomley, F., 2000. Mapping and expression of microneme genes in *Eimeria tenella*. Int J Parasitol. 30, 1493-1499.

Schwarz, R.S., Fetterer, R.H., Rosenberg, G.H., Miska, K.B., 2010. Coccidian merozoite transcriptome analysis from *Eimeria maxima* in comparison to *Eimeria tenella* and *Eimeria acervulina*. J Parasitol 96, 49-57.

Shirley, M.W., 1995. *Eimeria* species and strains of chicken, in: Eckert, J., Braun, R., Shirley, M.W., Coudert, P. (Eds.), COST 89/820 Biotechnology: Guidelines on techniques in coccidiosis research. European Comission, Brussels, Luxembourg pp. 1-24.

Shirley, M.W., 2000. The genome of *Eimeria* spp., with special reference to *Eimeria tenella* - a coccidium from the chicken. Int J Parasitol. 30, 485-493.

Shirley, M.W., Harvey, D.A., 2000. A genetic linkage map of the apicomplexan protozoan parasite *Eimeria tenella*. Genome Res 10, 1587-1593.

Shirley, M.W., Smith, A.L., Tomley, F.M., 2005. The biology of avian *Eimeria* with an emphasis on their control by vaccination. Adv Parasitol 60, 285-330.

Tabares, E., Ferguson, D., Clark, J., Soon, P.E., Wan, K.L., Tomley, F., 2004. *Eimeria tenella* sporozoites and merozoites differentially express glycosylphosphatidylinositol-anchored variant surface proteins. Mol Biochem Parasitol 135, 123-132.

Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., Rao, B.S., Smirnov, S., Sverdlov, A.V., Vasudevan, S., Wolf, Y.I., Yin, J.J., Natale, D.A., 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4, 41.

Tatusov, R.L., Galperin, M.Y., Natale, D.A., Koonin, E.V., 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res 28, 33-36.

Tomley, F., 1994. Antigenic diversity of the asexual developmental stages of *Eimeria tenella*. Parasite Immunol 16, 407-413.

Vencio, R.Z., Varuzza, L., de, B.P.C.A., Brentani, H., Shmulevich, I., 2007. Simcluster: clustering enumeration gene expression data on the simplex space. BMC Bioinformatics 8, 246.

Verjovski-Almeida, S., DeMarco, R., Martins, E.A., Guimaraes, P.E., Ojopi, E.P., Paquola, A.C., Piazza, J.P., Nishiyama, M.Y., Jr., Kitajima, J.P., Adamson, R.E., Ashton, P.D., Bonaldo, M.F., Coulson, P.S., Dillon, G.P., Farias, L.P., Gregorio, S.P., Ho, P.L., Leite, R.A., Malaquias, L.C., Marques, R.C., Miyasato, P.A., Nascimento, A.L., Ohlweiler, F.P., Reis, E.M., Ribeiro, M.A., Sa, R.G., Stukart, G.C., Soares, M.B., Gargioni, C., Kawano, T., Rodrigues, V., Madeira, A.M., Wilson, R.A., Menck, C.F., Setubal, J.C., Leite, L.C., Dias-Neto, E., 2003. Transcriptome analysis of the acoelomate human parasite *Schistosoma mansoni*. Nat Genet 35, 148-157.

Wagenbach, G.E., Burns, W.C., 1969. Structure and Respiration of Sporulating *Eimeria stiedae* and *E. tenella* Oocysts. J Protozool 16, 257-263.

Wan, K.L., Chong, S.P., Ng, S.T., Shirley, M.W., Tomley, F.M., Jangi, M.S., 1999. A survey of genes in *Eimeria tenella* merozoites by EST sequencing. Int J Parasitol. 29, 1885-1892.

Wang, C.C., 1976. Inhibition of the respiration of *Eimeria tenella* by quinolone coccidiostats. Biochem Pharmacol 25, 343-349.

Wang, C.C., Stotish, R.L., Poe, M., 1975. Dihydrofolate reductase from *Eimeria tenella*: rationalization of chemotherapeutic efficacy of pyrimethamine. J Protozool 22, 564-568.

Wilson, P.A., Fairbairn, D., 1961. Biochemistry of Sporulation in Oocysts of *Eimeria acervulina*. J Protozool 8, 410-416.

Figure captions

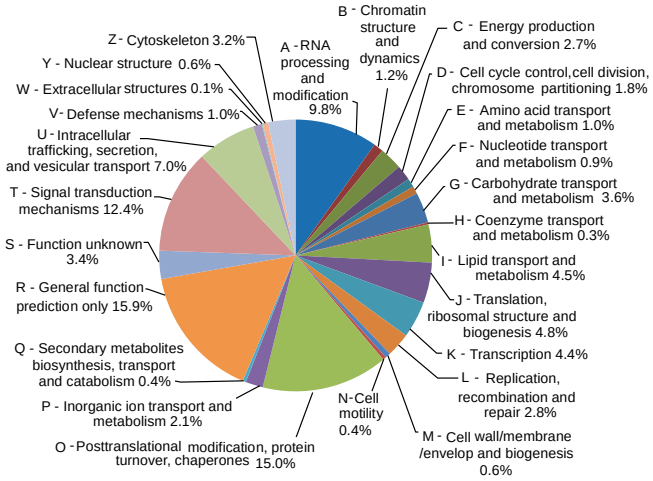
Fig. 1. Functional assignments to eukaryotic orthologous groups (KOG) categories. Protein products from assembled cDNAs of *Eimeria acervulina*, *Eimeria maxima* and *Eimeria tenella* were classified into 25 functional categories (A – W, Y, Z) of the KOG database. Percentages refer to the relative proportions of the categories.

Fig. 2. Comparative functional classification of three *Eimeria* spp. using Gene Ontology (GO). Distribution of GO term assignments to protein products derived from assembled cDNAs of *Eimeria acervulina*, *Eimeria maxima* and *Eimeria tenella*. The ontology domains (biological process, cellular component and molecular function) are presented in separate charts. Percentages refer to the number of proteins mapped to each GO term in regard to the total number of assembled cDNAs of the respective *Eimeria* sp.

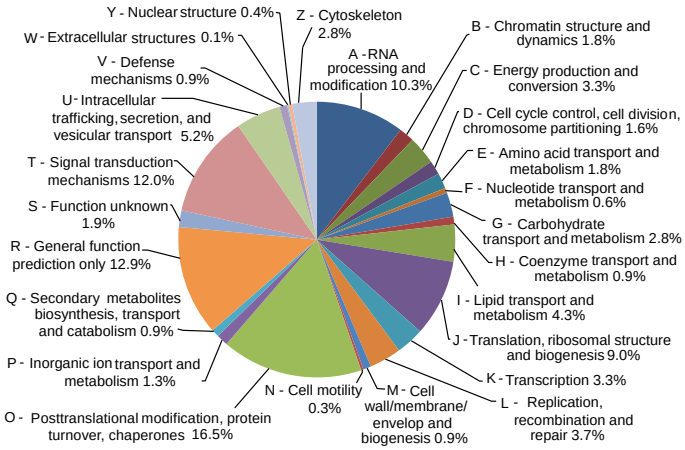
Fig. 3. Clustering analysis of *Eimeria* digital expression profiles. Transcript enumeration data based on expressed sequence tag/open reading frame EST (EST/ORESTES) countings of *Eimeria acervulina*, *Eimeria maxima* and *Eimeria tenella*, derived from different developmental stages (Ou, unsporulated oocysts; Op, sporoblast phase oocysts; Os, sporulated oocysts; Sz, sporozoites; Mz1, first generation merozoites; Mz2, second generation merozoites). The method used was Pearson distance with complete linkage agglomerative hierarchical clustering.

Figure 1

E. acervulina



E. maxima



E. tenella

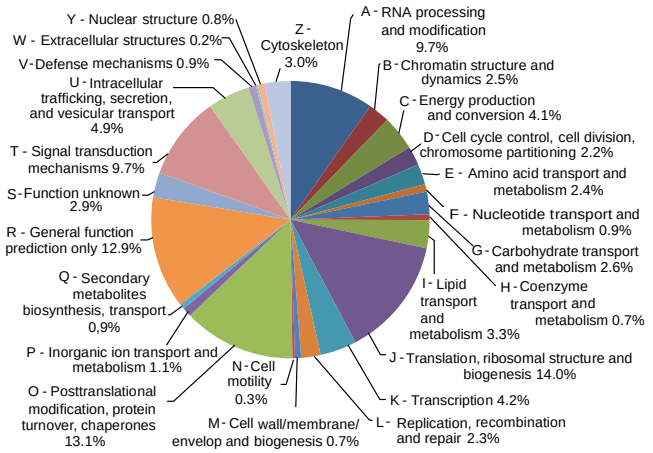


Figure 2

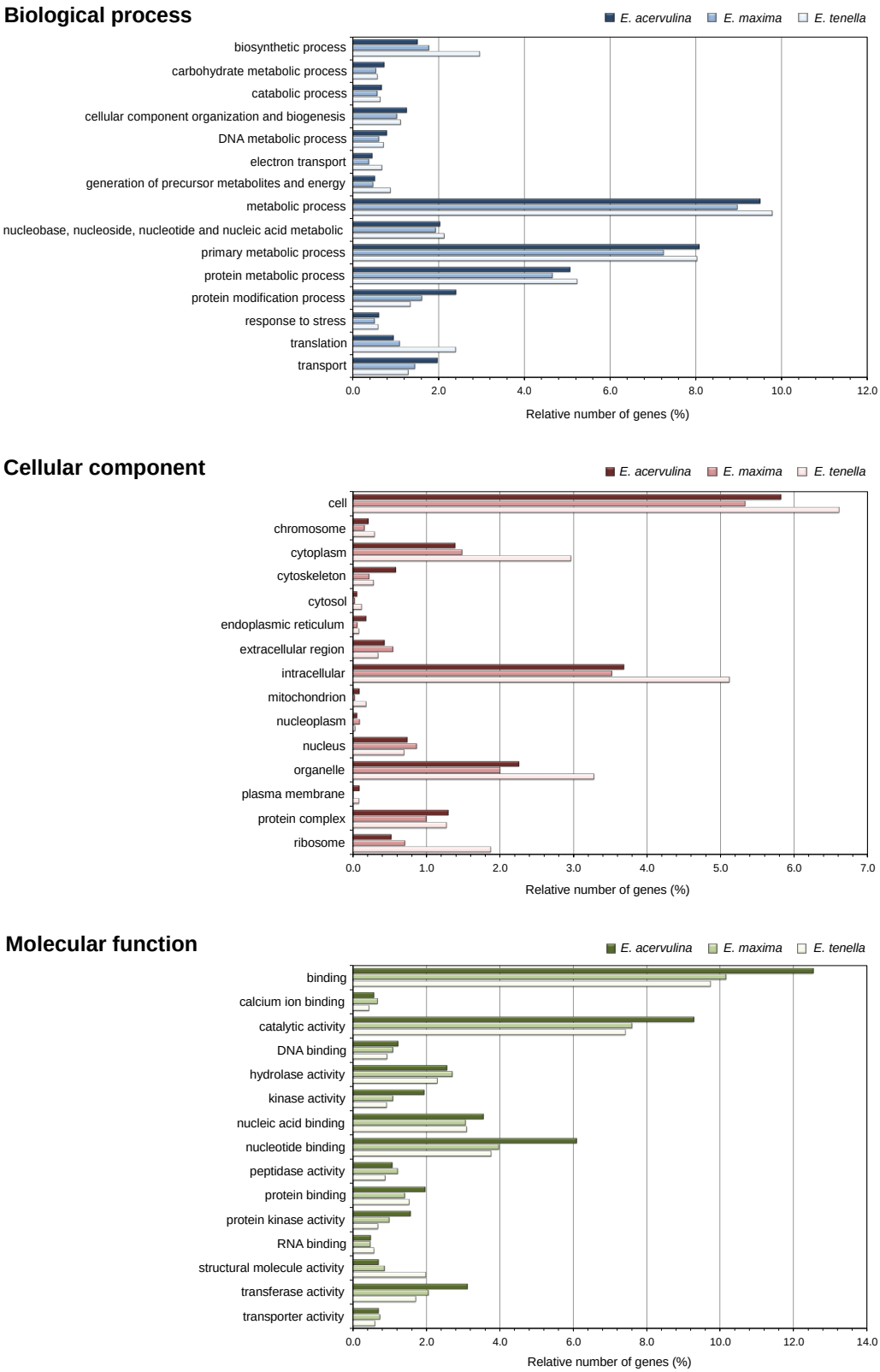
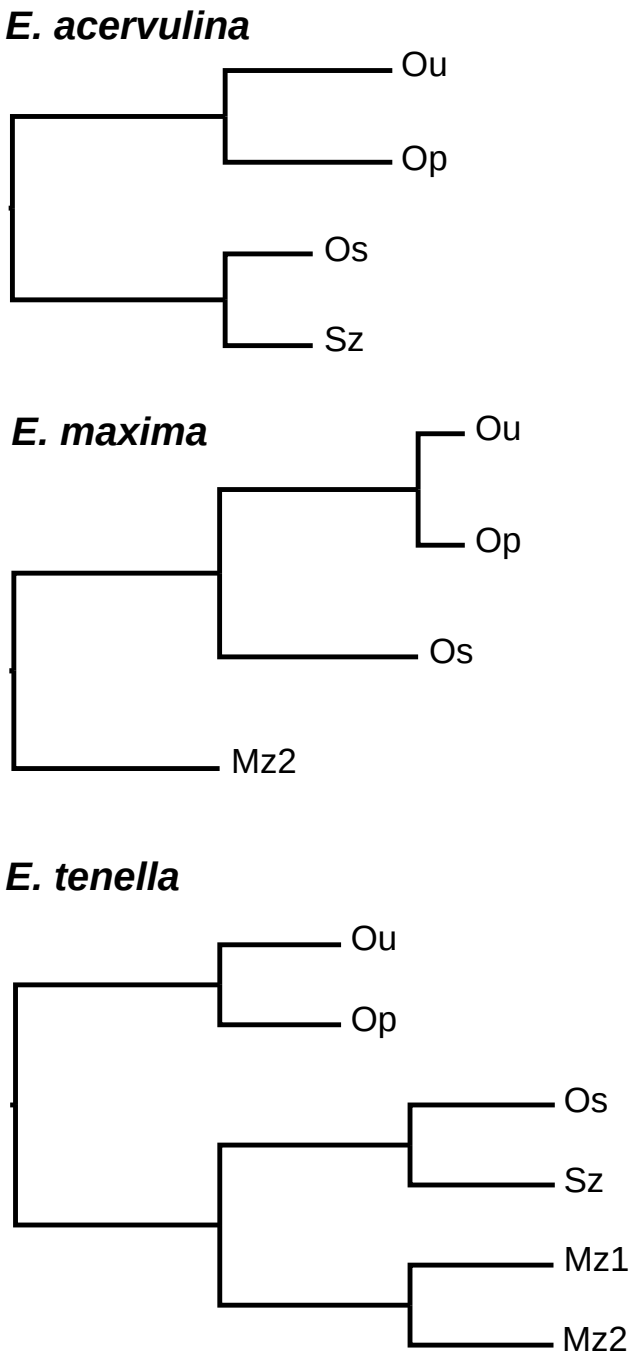

Biological process
E. acervulina
E. maxima
E. tenella
biosynthetic process
carbohydrate metabolic process
catabolic process
cellular component organization and biogenesis
DNA metabolic process
electron transport
generation of precursor metabolites and energy
metabolic process
nucleobase, nucleoside, nucleotide and nucleic acid metabolic
primary metabolic process
protein metabolic process
protein modification process
response to stress
translation
transport
0.0
2.0
4.0
6.0
8.0
10.0
12.0
Relative number of genes (%)
Cellular component
E. acervulina
E. maxima
E. tenella
cell
chromosome
cytoplasm
cytoskeleton
cytosol
endoplasmic reticulum
extracellular region
intracellular
mitochondrion
nucleoplasm
nucleus
organelle
plasma membrane
protein complex
ribosome
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
Relative number of genes (%)
Molecular function
E. acervulina
E. maxima
E. tenella
binding
calcium ion binding
catalytic activity
DNA binding
hydrolase activity
kinase activity
nucleic acid binding
nucleotide binding
peptidase activity
protein binding
protein kinase activity
RNA binding
structural molecule activity
transferase activity
transporter activity
0.0
2.0
4.0
6.0
8.0
10.0
12.0
14.0
Relative number of genes (%)

Figure 3


E. acervulina
Ou
Op
Os
Sz
E. maxima
Ou
Op
Os
Mz2
E. tenella
Ou
Op
Os
Sz
Mz1
Mz2

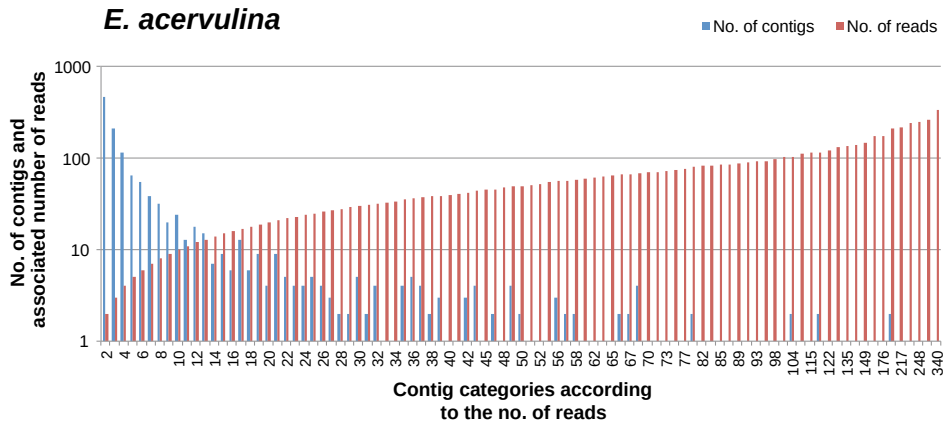
Supplementary Figure legends

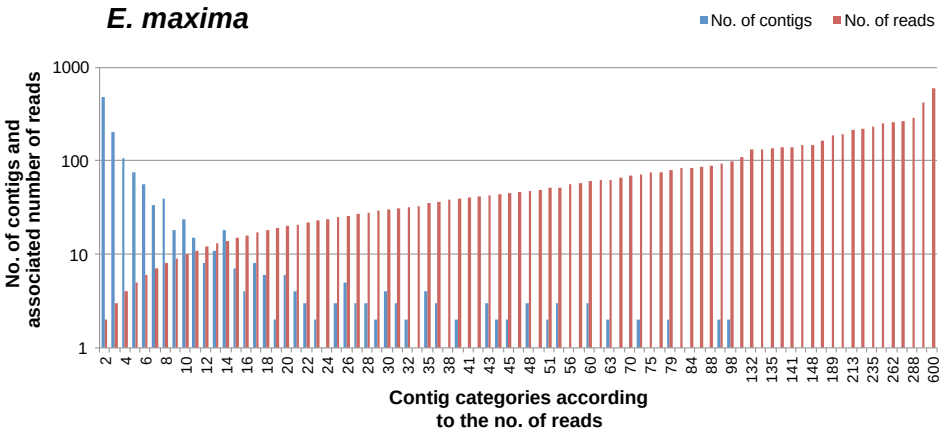
Supplementary Fig. S1. Global distribution of assembled cDNAs according to the number of reads. *Eimeria acervulina*, *E. maxima* and *E. tenella* cDNA reads were assembled with CAP3 and the resulting contigs were categorized in respect to the number (no.) of composing reads. Singlets were not considered in the analysis. Y-axis is in logarithmic scale.

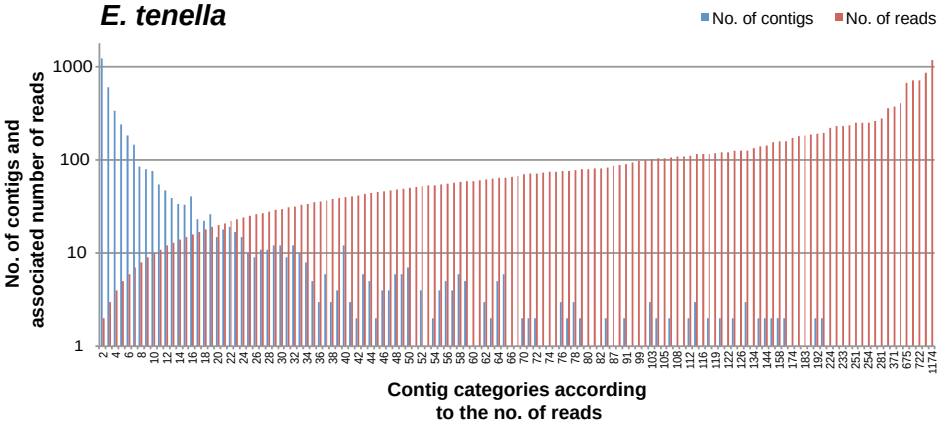
Supplementary Fig. S2. Stage-specific distribution of assembled cDNAs according to the number of reads. *Eimeria tenella* cDNA reads derived from unsporulated oocysts, sporoblast-phase oocysts, sporulated oocysts, sporozoites, 1st -generation merozoites and 2nd -generation merozoites were assembled with CAP3, and the resulting contigs of each stage were categorized in respect to the number (no.) of composing reads. Singlets were not considered in the analysis. Y-axis is in logarithmic scale.

Supplementary Fig. S3. Clustering analysis of *Eimeria* digital expression profiles. Transcript enumeration data based on expressed sequence tag/open reading frame ESTs (EST/ORESTES) countings of *E. acervulina*, *E. maxima* and *E. tenella*, derived from different developmental stages (Ou – unsporulated oocysts , Op – sporoblast-phase oocysts , Os – sporulated oocysts , Sz – sporozoites, Mz1 – 1st-generation merozoites, Mz2 - 2nd-generation merozoites). Methods: (A) Aitchison distance with complete linkage agglomerative hierarchical clustering; (B) Spearman distance with complete linkage agglomerative hierarchical clustering.

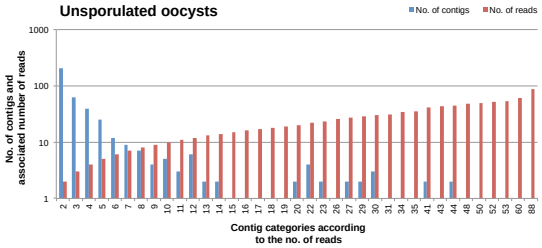
Supplementary Figure 1

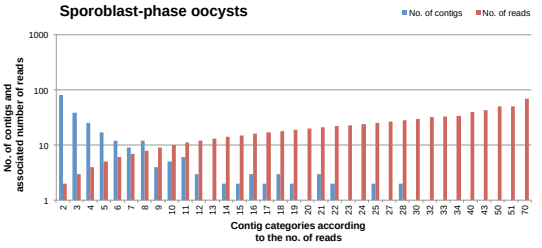


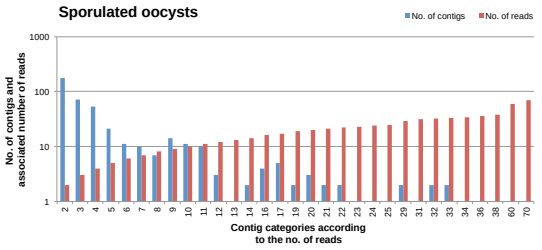


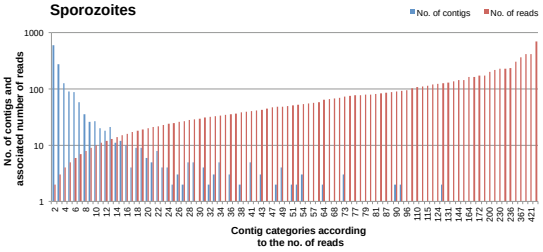


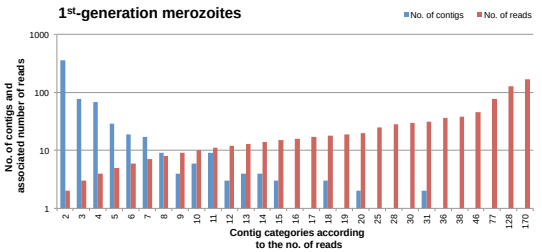
Supplementary Figure 2

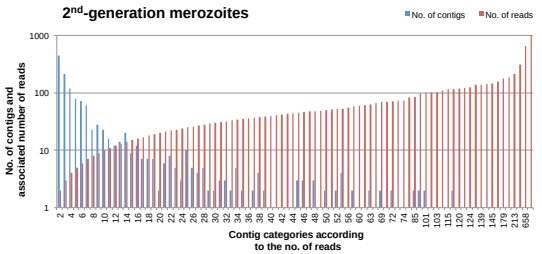












Supplementary Figure 3

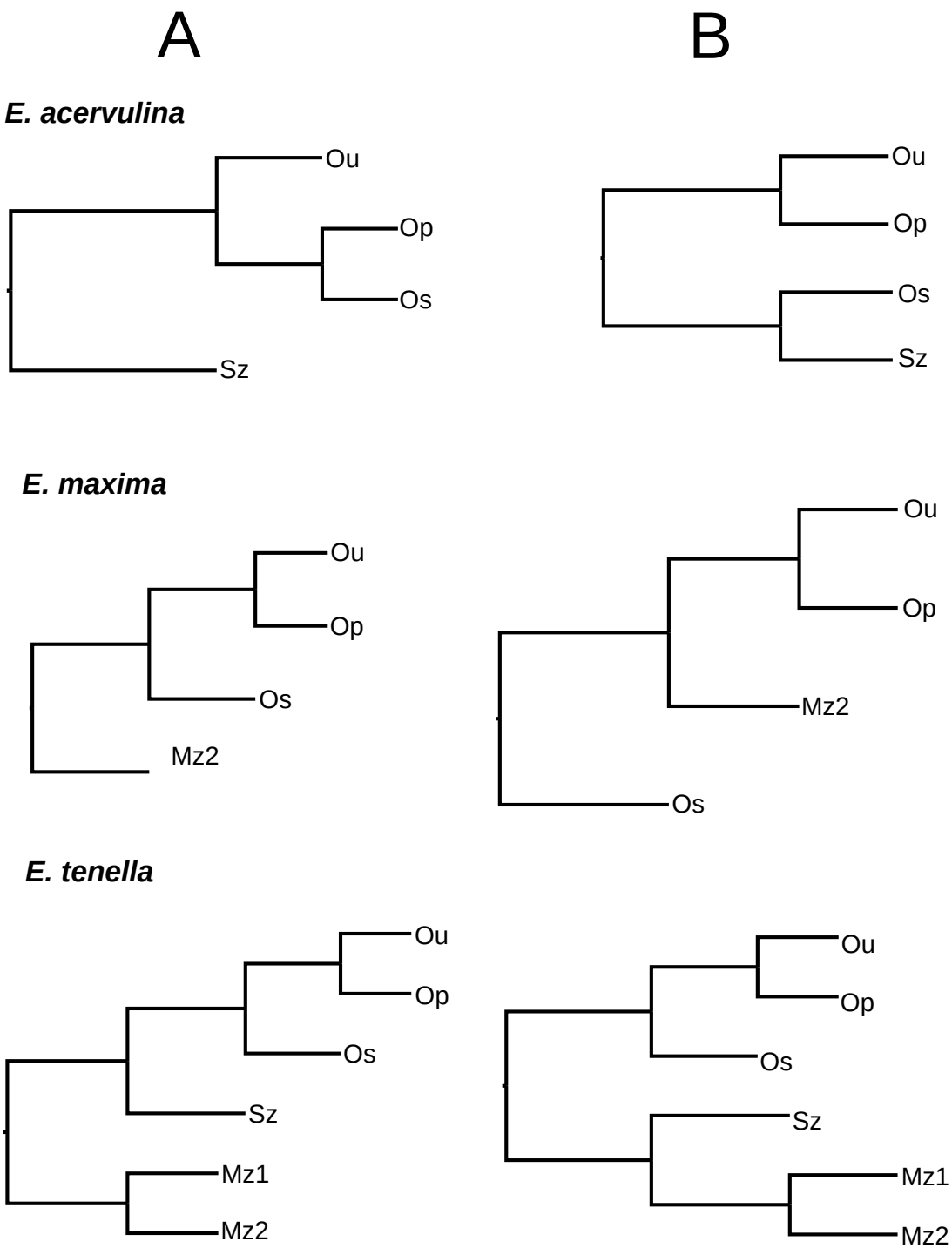


Table 1

Assembly results of cDNAs of *Eimeria acervulina*, *Eimeria maxima* and *Eimeria tenella* using open reading frame expressed sequence tags (ORESTES) and EST reads (only for *E. tenella*).

Assembly results	*E. acervulina*	*E. maxima*	*E. tenella*
No. of reads	16,289	15,490	48,361
Contigs	1,280	1,233	3,724
Total length (bp)	874,527	707,815	2,922,818
Average length (bp)	683	574	785
Singlets	2,133	2,193	4,976
Total length (bp)	758,709	704,429	1,900,014
Average length (bp)	356	321	382
Total no. of assembled cDNAs	3,413	3,426	8,700
Total length (bp)	1,633,236	1,412,244	4,822,832
Average length (bp)	479	412	554

Table 2

Genome complexity and predicted number of proteins of apicomplexan parasites.

Organism	Genome size (Mb)	No. of chromosomes	No. of proteins	Gene density (genes/MB)
Eimeria tenella H[a]	55	14	8,786	160
Toxoplasma gondii ME49[b]	63	14	7,993	127
Neospora caninum NCLiv[c]	62	14	7,083	114
Plasmodium falciparum 3D7[d]	23	14	5,446	237
Babesia bovis T2Bo[e]	8	4	3,703	463
Theileria annulata Ankara clone C9[f]	8	4	3,795	474
Cryptosporidium parvum Iowa II[g]	9	8	3,805	423

Sources:

[a]Sanger Institute – v. Sep 2010 - (ftp://ftp.sanger.ac.uk/pub/pathogens/Eimeria/tenella/)

[b]ToxoDB – v. 6.0 - (http://toxodb.org/toxo/)

[c]GeneDB – v. Sep 2009 - (http://www.genedb.org/Homepage/Ncaninum)

[d]PlasmoDB – v. 6.3 - (http://plasmodb.org/plasmo/)

[e]NCBI – v. 08 Aug 2007, accession number **AAXT00000000**

[f]GeneDB – v. 15 Jul 2005 - (http://www.genedb.org/Homepage/Tannulata)

[g]CryptoDB – v. 4.3 - (http://cryptodb.org/cryptodb/)

Table 3

Functional annotation of the assembled transcripts of *Eimeria acervulina*, *Eimeria maxima* and *Eimeria tenella*.

Feature	*E. acervulina*	*E. maxima*	*E. tenella*
Assembled transcripts	3,413	3,426	8,700
Translated products (≥50 aa)	3,233	3,096	7,990
BLAST (e-value ≤ 10^{-6})	1,235	1,091	2,559
KOG	707	604	1,517
Proteins with positive BLAST and/or KOG results	1,276	1,114	2,633
Proteins within Apicomplexa orthology groups	745	614	1,236
Hypothetical proteins conserved across Apicomplexa	174	161	180
Proteins within orthology groups common to three *Eimeria* species (175 groups)	240	215	298
Proteins within orthology groups restricted to *Eimeria* spp. (27 groups)	34	28	50
Gene Ontology (any)	644	541	1,404
Biological Process	398	337	946
Cell Component	203	182	555
Molecular Function	579	473	1,247

aa, amino acid; KOG, eukaryotic orthologous groups

Supplementary Table S1

Primer sequences used for *Eimeria* mRNA integrity tests with different genome targets.

Species/Target	Accession code	Primer sequences	Amplicon size (bp)
Eimeria acervulina			
α-tubulin	**AY488134**	GAGGTACAGGTTCAGGTTTGGG	818
		CTTCCTCCATACCTTCACCAAC	
Eimeria maxima			
TFP250	**AY239227**	GCTGCTCTGTAAACGCCACTTG	900
		CACTCGTCGATGTCGGTACACC	
Eimeria tenella			
EtMIC4	**AJ306453**	TATAGACGAGTGCCAAGACCCG	1007
		CCGTCACCTGAATAGCCAGCTA	

Supplementary Table S2

Source, library type and number of reads of the different sets of cDNA sequences used for transcript reconstruction of *Eimeria tenella*.

Source	Library type	Format	Repository	No. of input reads	No. of accepted reads
Sanger Institute, UK	EST	SCF and FASTA	Sanger [a] and NCBI [b]	9,778	5,939
Universiti Kebangsaan, Malaysia	EST	FASTA	NCBI [b]	1,051	1,028
Washington University, USA	EST	Trace files	WUSTL [c]	27,500	26,249
USDA, USA	EST	Trace files	USDA [d] /NCBI [b]	1,666	1,022
USP, Brazil	ORESTES	Trace files	USP [e]	17,568	14,123
Total	-	-	-	57,563	48,361

[a] http://www.sanger.ac.uk/resources/downloads/protozoa/eimeria-tenella.html

[b] http://www.ncbi.nlm.nih.gov/dbEST/

[c] http://genome.wustl.edu/data/est_projects

[d] Trace files kindly provided by Dr. K. Miska. FASTA sequences available at the NCBI.

[e] http://www.coccidia.icb.usp.br/eimeriatdb/

EST, expressed sequence tag; ORESTES, open reading frame EST

Supplementary Table S3

Distribution of high-quality open reading frame expressed sequence tags (ORESTES) and conventional EST reads of *Eimeria acervulina, Eimeria maxima* and *Eimeria tenella* used in this work.

Eimeria sp./ Developmental stage	Number of reads
E. acervulina (Ac)	
Unsporulated oocysts (Ou)	3,532
Sporulating oocysts (Op)	2,570
Sporulated oocysts (Os)	1,574
Sporozoites (Sz)	8,613
Total	16,289
E. maxima (Mx)	
Unsporulated oocysts (Ou)	3,118
Sporulating oocysts (Op)	3,215
Sporulated oocysts (Os)	8,921
2nd generation merozoites (Mz)	236
Total	15,490
E. tenella (Tn)[a]	
Unsporulated oocysts (Ou)	3,081 (1,429)
Sporulating oocysts (Op)	2,233 (2,233)
Sporulated oocysts (Os)	3,256 (2,914)
2nd generation merozoites (Mz)	16,589 (3,580)
Sporozoites (Sz)	19,043 (3,967)
1st generation merozoites (Mz)	4,150 (0)
Undefined stage	9 (0)
Total	48,361 (14,123)

[a]In *E. tenella* the numbers correspond to the sum of conventional ESTs plus ORESTES reads (indicated in parentheses).

Supplementary Table S6

Mapping results of reconstructed cDNAs onto the assembled genomes of *Eimeria maxima* and *Eimeria tenella*.

Mapping results	*E. maxima*	*E. tenella*
Contigs		
Aligned cDNAs	1,152 (93.43%)	3,677 (98.73%)
Mean coverage	90.84%	93.93%
Singlets		
Aligned cDNAs	1,966 (89.64%)	4,379 (88.0%)
Mean coverage	90.07%	93.21%
Total		
Aligned cDNAs	3,118 (91.0%)	8,056 (92.59%)
Mean coverage	90.35%	93.54%

Assunto: IJPara11_362R1
De: "Int J Parasitol" <editor@IJP.org.au>
Data: 18/10/11 22:06
Para: argruber@usp.br, arthur.gruber@gmail.com
CC: editor@IJP.org.au

Ms. Ref. No.: IJPara11_362R1 Title: A comparative transcriptome analysis reveals expression profiles conserved across three Eimeria spp. of domestic fowl and associated with multiple developmental stages International Journal for Parasitology

Dear Prof Gruber,

I am pleased to confirm that your paper "A comparative transcriptome analysis reveals expression profiles conserved across three Eimeria spp. of domestic fowl and associated with multiple developmental stages" has been accepted for publication in the International Journal for Parasitology.

For every issue of the IJP the Editor will endeavour to feature an image on the cover from, or relating to, an article in that issue. We invite all authors to submit images that would be suitable. To maximize their aesthetic qualities, these images may be stylized/modified versions of pictures or diagrams from the author's article.

The Editor will choose the image that is the most eye-catching and informative for each issue and complements those chosen for recent issues.

Thank you for submitting your work to this journal.
Yours sincerely,

Alex Loukas Editor-in-Chief International Journal for Parasitology
****
For further editorial assistance, please contact the International Journal for Parasitology
E-mail: editor@IJP.org.au.

TRACK YOUR ACCEPTED ARTICLE

Welcome! Login to get personalized options. New user? Register | Why register? Help

Your article's details and status are shown in the following table:

Article status

Article title:	A comparative transcriptome analysis reveals expression profiles conserved across three Eimeria spp. of domestic fowl and associated with multiple developmental stages
Reference:	PARA3338
Journal title:	International Journal for Parasitology
Corresponding author:	Prof. Arthur Gruber
First author:	Dr. Jeniffer Novaes
Received at Editorial Office:	14 Sep 2011
Article revised:	15 Oct 2011
Article accepted for publication:	19 Oct 2011
Received at Elsevier:	15 Nov 2011
Journal publishing agreement sent to author:	17 Nov 2011
Offprint order form sent to author:	17 Nov 2011
PDF offprint:	Yes
Expected dispatch of proofs:	2 Dec 2011
Journal publishing agreement returned:	18 Nov 2011
DOI information:	10.1016/j.ijpara.2011.10.008
Status comment:	At this moment it is not yet possible to give you information about the publication date. This depends on the number of articles lined up for publication in the journal. Citation information will be shown when available.

Track another article