UNIVERSIDADE FEDERAL DE PELOTAS
Centro de Desenvolvimento Tecnológico – CDTec
Curso de Graduação em Biotecnologia

Trabalho Acadêmico de Conclusão de Curso

CestodaDB: Desenvolvimento de um banco de dados para proteínas de cestódeos

Frederico Schmitt Kremer

Pelotas, 2013

Frederico Schmitt Kremer

CestodaDB:

Desenvolvimento de um banco de dados para proteínas de cestódeos

Trabalho de Conclusão de Curso apresentado ao Curso de Bacharelado em Biotecnologia do Centro de Desenvolvimento Tecnológico da Universidade Federal de Pelotas, como requisito parcial à obtenção do título de Bacharel em Biotecnologia.

Orientador: Luciano da Silva Pinto

Pelotas, 2013

Dados de catalogação na fonte:
Ubirajara Buddin Cruz – CRB-10/901

Biblioteca de Ciência & Tecnologia - UFPel

K92d Kremer, Frederico Schmitt
Desenvolvimento de um banco de dados para proteínas de cestódeos / Frederico Schmitt Kremer. – 34f. : il. – Trabalho de conclusão de curso (Graduação em Biotecnologia). Universidade Federal de Pelotas. Centro de Desenvolvimento Tecnológico. Pelotas, 2013. – Orientador Luciano da Silva Pinto.

1.Biotecnologia. 2.Bioinformática. 3.Cestoda. 4.Taenia. 5. Echinococcus. I.Pinto, Luciano da Silva. II.Título.

CDD: 572.330285

Frederico Schmitt Kremer

CestodaDB: Desenvolvimento de um banco de dados para proteínas de cestódeos

Trabalho de conclusão de curso aprovado, como requisito parcial, para a obtenção do grau de Bacharel em Biotecnologia, Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas.

Data da Defesa: 16/12/2013

Banca examinadora:

Prof. Dr. Luciano da Silva Pinto (Orientador)
Doutor em Biotecnologia pela Universidade Federal de Pelotas

Prof. Dr. Alan John Alexander McBride
Doutor em Biologia Molecular Aplicada pela Universidade de Surrey

MSc. Marcus Redü Eslabão
Mestre em Biotecnologia pela Universidade Federal de Pelotas.

Agradecimentos

Aos meus pais, Roberto e Sandra, e ao meu irmão Oscar, por todo o incentivo e apoio que recebi durante a minha vida.

Aos meus amigos “dos tempos do CEFET-RS”, Bruno “Bolaxa”, Cássia, Felipe, Gabriel, Luiz, Rai e Thais, por todos os bons momentos, gordices, mesas de RPG e maratonas de filmes/jogos.

Ao meu orientador, professor Luciano da Silva Pinto, por todo o apoio e auxílio dado durante a minha iniciação científica e na execução deste trabalho.

Á professora Luciana Bicca Dode, pelo apoio e por todos os ensinamentos, tanto na pesquisa, quanto na extensão.

Ao doutorando em biotecnologia e amigo Marcus Redü Eslabão, pela orientação durante o meu estágio no laboratório de bioinformática, pelas aulas de programação e pelos churrascos e sanduiches de bacon.

Aos meus colegas da ATBiotec2013, especialmente à Gabriella Borba, Lívia, Luiza, Vinícius e Rafael Woloski, por todos os bons momentos que passamos nestes 4 anos.

A todos os professores do curso de graduação em Biotecnologia, por tudo que me foi ensinado nestes quatro anos.

Muito obrigado por tudo!

Resumo

KREMER, Frederico Schmitt. **CestodaDB**: Desenvolvimento de um banco de dados para proteínas de cestódeos. 2013. Trabalho de Conclusão de Curso (Bacharelado em Biotecnologia) – Curso de Bacharelado em Biotecnologia, Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas, Pelotas, 2013.

A classe de Cestoda é constituída por animais com comportamento parasitário, sendo esses organismos agentes etiológicos de um grande número de doenças em diferentes espécies de vertebrados, incluindo o ser humano. Teníase e cisticercose (causada por *Taenia* spp.), Equinococose (causada por *Echinococcus* spp.) E sparganosis (causada por *Spirometra* spp.), São alguns exemplos de doenças zoonóticas causadas por estes organismos. Para ajudar no desenvolvimento de novas abordagens biotecnológicas, a bioinformática pode fornecer ferramentas de gerenciamento de dados, tais como bancos de dados, para armazenamento, busca e visualizar informações moleculares. O presente trabalho teve como objetivo o desenvolvimento de um banco de dados on-line, chamado CestodaDB, para proteínas de cestóides. O conjunto de dados é constituído por 3878 proteínas retiradas de UniProt, físico-quimicamente, funcional e estruturalmente anotado por diferentes ferramentas de bioinformática, incluindo análises pelo Pfam, PRODOM, CATH, SignalP, TMHMM e TargetP. Conectando diferentes tipos de dados, os pesquisadores podem usar o CestodaDB para identificar novos alvos para vacinas, terapias e sistemas de diagnóstico. O presente banco de dados pode ser acessado pelo endereço htttp://labbioinfo.ufpel.edu.br/cestodadb/.

Palavras-chave: cestoda; taenia; echinococcus; bioinformática;

Abstract

KREMER, Frederico Schmitt. **CestodaDB**: developing a database for Cestoda proteins. 2013. **Final Year's Project** (Bacharelado em Biotecnologia) – Curso de Bacharelado em Biotecnologia, Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas, Pelotas, 2013.

The Cestoda class is composed by parasitic flatworms, which are etiological agents of a large number of diseases in different species of vertebrates, including humans. Taeniasis and cysticercosis (caused by *Taenia* spp.), echinococcosis (caused by *Echinococcus* spp.) and sparganosis (caused by *Spirometra* spp.), are some examples of zoonotic diseases caused by these organisms. To help in the development of new biotechnological approaches, bioinformatics can provide data-managing tools, such as databases, for storing, searching and visualizing molecular informations. The present work was to develop an online database, called CestodaDB, for proteins from cestodes. The dataset contains 3878 proteins taken Uniprot, their physic-chemical, functional and structural annotated by different bioinformatics tools, including Pfam, PRODOM, CATH, SignalP, TMHMM e TargetP. Set together different data, allow researchers to use CestodaDB to mine new targets for vaccines, therapeutics and diagnosis systems. The presente database can be accessed through the URL htttp://labbioinfo.ufpel.edu.br/cestodadb/.

Key-Words: cestoda; taenia; echinococcus; bioinformatics;

Sumário

1. Introdução

No presente trabalho de conclusão de curso de Graduação em Biotecnologia é apresentado o desenvolvimento de um banco de dados *online* para sequencias e informações físico-químicas e estruturais de proteínas de parasitas da classes Cestoda, denominado CestodaDB. Inicialmente, é apresentada uma revisão de literatura sobre a biologia destes organismos e sua relevância para a medicina humana, animal e para biotecnologia, além de também apresentar abordagens de bioinformática aplicáveis ao estudo dos mesmos. Posteriormente, é demonstrada a implementação do banco de dados, incluindo o processo de desenvolvimento da estrutura base para seu funcionamento, o processo de coleta de dados em fontes públicas e como foi realizado o cruzamento dos dados. Uma análise dos dados e metadados gerados neste trabalho é apresentada no final do mesmo.

2. Revisão Bibliográfica

2.1 A classe *Cestoda*

A classe Cestoda, pertencente ao filo Platyhelminthes, é constituída por organismos de natureza parasitária como corpos alongados, segmentados e que apresentam um órgão de fixação na extremidade (escólex) com diferentes ferramentas de adesão dependendo da espécie. Também conhecidos como tênias ou cestoides, os constituintes desta classes são parasitas obrigatórios do sistema gastrointestinal de um amplo espectro de espécies animais, sendo desprovidos de sistema digestório e apresentando um eficiente sistema reprodutivo (Rey, 2008).

2.1.1 Anatomia

Apesar corpo dos cestoides pode variar em tamanho de poucos milímetros à alguns metros dependendo da espécie, a tua organização anatômica é pouco variável. Em organismos são observadas três estruturas principais: escólex, colo e estróbio (Tortora et al, 2007).

Escólex: Consiste em uma dilatação localizada na extremidade mais delgada do corpo do animal. Nesta encontram-se as estruturas de fixação ao aparelho intestinal. Por sua localização anatômica é considerada a "cabeça", apesar da ausência de estruturas de ingestão de alimentos que caracterizam este estrutura nos demais animais (Cimerman et al, 2010).

Colo*:* É uma região delgada localizada abaixo do escólex, caracterizada pela presença de tecidos em contínuo crescimento, responsáveis pelo alongamento contínuo destes animais (Cimerman et al, 2010).

Estróbio: Região do corpo constituída por segmentos denominados proglótes ou proglótides, que são produzidas pela proliferação colo. As proglótes são constituídas por órgãos reprodutivos masculinos e

femininos, o que torna os cestódeos animais hermafroditas (Cimerman et al, 2010).

2.1.2 Nutrição

A natureza estritamente parasitária dos cestódeos se reflete nas adaptações fisiológicas relacionadas ao processo de nutrição, assim como ocorre em outros platelmintos. Nesta espécies, a ausência de um sistema digestório completo é compensada pela absorção dos nutrientes através do tegumento, uma estrutura localizada superfície do seus corpos, através de diferentes mecanismos de transporte, como transporte ativo, difusão facilitada e pinocitose (Halton, 1997).

2.1.3 Ciclo vital

O oócitos dos cestódeos são formados pela fecundação dos óvulos pelos espermatozoides localizados nos receptáculos seminais. Estes então são cercados por células vitelinas que secretam substâncias para a formação da primeira casca ovular, enquanto o oócito inicia um processo de divisão desigual para a formação do embrião e do vitelo. Os ovos maduros, quando ingeridos pelo seus hospedeiros, eclodem em função de processos mecânicos, como a mastigação, ou químicos, como a degradação das camadas externas pela ação de enzimas proteolíticas. Após a eclosão, pequenas esferas contendo ganchos, denominadas *oncosferas* (do grego *onko*: gancho) são liberadas, iniciando o desenvolvimento larval destes organismos. Dependendo da espécies, mais de uma espécie de hospedeiro pode estar envolvido no ciclo vital. (Rey, 2008).

2.1.4 Sistemática

A classe Cestoda é constituída pelas Cestodaria e Eucestoda. A sub-classe *Cestodaria* agrega as ordens *Amphilinidea*, constituída por 4 espécies divididas em 4 gêneros, e *Gyrocotylidea*, constituída por 4 espécies pertencentes ao gênero *Gyrocotyle*. Já a sub-classe Eucestoda é composta por 17 ordens que agregam 276 gêneros e 1086 espécies, incluindo as espécies pertencentes ao gêneros *Taenia* e *Echinococcus* (NCBI-Taxonomy, 2013).

2.2 Doenças humanas e veterinárias causadas por cestódeos

Os cestódeos são parasitas obrigatórios diversas espécies de vertebrados, dentre eles o ser humano. Dentre as espécies de relevância médica e veterinária inclui-se as pertencentes aos gêneros *Taenia*, *Echinococcus* e *Spirometra* (Cimerman, 2010).

O gênero *Taenia* é composto por mais de 40 espécies, sendo *T. saginata*, *T. solium* e *T. asiatica* as de maior ocorrência no ser humano (Hoberg, 2006). A infecção por estas espécies pode se dar de duas formas: ingestão de larvas (cisticercos) ou ovos. A infecção por larvas acarreta a teníases, caracterizada pela fixação do parasita na parede do trato intestinal do hospedeiro, uma doença normalmente assintomática (Yanagida, Sako, Nakao, Nakaya, & Ito, 2012). Já a infecção ingestão dos ovos acarreta a cisticercose, caracterizada pelo migração do parasita para tecidos como o muscular e nervoso. A infecção no sistema nervoso, a chamada neurocisticercose, pode acarretar quadros de epilepsia e demência, podendo necessitar de intervenção cirúrgica (Del Brutto, 2012). O diagnóstico pode ser realizado através de anamnese, análise coprológica (em caso de teníase) e tomografia computadorizada ou ressonância magnética (em caso de cisticercose/neurocisticercose) (RAOUL et al, 2013), e novas abordagens baseadas em detecção por imuno-ensaios e detecção de DNA também estão em desenvolvimento (RODRIGUES et al, 2012). Em caso de infecção, o fármaco Albendazol é a primeira escolha para o tratamento, se mostrando eficiente tanto no tratamento da teníase quando da cisticercose (SOTELO, 2011)

Infecções por espécies do gênero *Echinococcus*, sobretudo pelas espécies *E. granulosus*, *E. multilocularis*, *E. vogeli* e *E. oligarthrus* também possuem um considerável relevância médica e veterinária, mostrando-se doenças emergentes e de grande impacto em certas regiões do continente europeu, asiático e sul-americano (Moro & Schantz, 2009). O quadro clínico da doença acarretada por estes parasitas, denominada Hidatidose, pode variar de acordo com a espécie, apesar de terem em o desenvolvimento de estruturas globulares semelhantes a tumores em tecidos e órgãos do hospedeiro (Thompson,

2008)(Cimerman et al 2010). O diagnóstico pode ser realizado através de anamnese, ultrassom e análises molecular (TORGERSON, 2009) e o tratamento pode ser farmacológico, através do uso de benzemidazóis, ou cirúrgico, para a remoção dos cistos hidáticos (KERN, 2010).

A Esparganose, acarretada pelos parasitas do gênero *Spirometra*, também possui considerável impacto na saúde pública em países asiático, aonde apresenta maior prevalência (Anantaphruti et al, 2011). A migração do parasita pelo corpo do hospedeiro, característica desta doença, pode acarreta diversos dados, incluindo cegueira, paralisia e até mesmo o óbito (Li et al., 2011). O diagnóstico pode ser realizado através de anamnese, técnicas de tomografia e análise sorológica, e o tratamento, em muitos casos, necessita de intervenção cirúrgica, sobretudo em caso de esparganose cerebral (KIM, 1996).

Como observado em outros grupos de parasitas, as infecções por estes e outros helmintos está diretamente relacionada às características sanitárias das regiões aonde se fazem presentes (Flisser, Rodríguez-Canul, & Willingham, 2006; Walker & Zunt, 2005).

2.3 Bioinformática

O advento de técnicas de sequenciamento de DNA e proteínas acarretou um crescimento exponencial no volume de informações moleculares disponíveis em bancos de dados públicos para um grande número de organismos. Neste contexto, a necessidade de se desenvolver ferramentas específicas para o gerenciamento destas informações acarretou no surgimento e desenvolvimento da bioinformática (Lesk, 2005).

2.3.1 Conceito

A bioinformática consiste na aplicação de abordagens e ferramentas derivadas da ciência da informação e computação no gerenciamento, processamento e análise de dados biológicos, sobretudo, mas não restrito a, aqueles derivados da biologia molecular e bioquímica. Sequências de nucleotídeos e proteínas,

estruturas de biomoléculas, organização de genes, rede de interação entras processos biológicos e análises filogenéticas são exemplos de dados que podem ser tratados a partir destas abordagens (Pevzner & Shamir, 2011).

2.3.2 Surgimento

O uso de ferramentas de informática para a resolução de problemas de biologia molecular iniciou-se na década 1960 com os trabalhos de Margaret Dayhoff (1925-1983). Visando entender as relações existentes entre as sequencias de diferentes proteínas, Dayhoff desenvolveu diferentes métodos computacionais de análise de sequência, incluindo a matriz de substituição PAM (*Point Accepted Mutation*) e o método de *máxima parcimônia* para filogenia molecular. Além disso, criou também o código de uma letra para aminoácidos e o organizou o *Atlas of Protein Sequence and Structure*, considerando o primeiro banco de dados biológico (Hunt, 1983).

2.3.3 Surgimento do Protein Data Bank

O banco de dados de Dayhoff consistiu em uma compilação de informações realizada de forma manual e disponibilizado através de volumes impressos. A necessidade por plataformas informatizadas para armazenar estes e outros tipos de dados biológico, ao menos que possibilitassem o acesso remoto, resultou no desenvolvimento do *Brookhaven RAster Display* (BRAD), que posteriormente evoluiria para o *Protein Data Bank* (PDB), atualmente tido o principal banco de dados para estruturas de proteínas obtidas por métodos experimentais (Berman, 2008).

2.3.4 O sequenciamento de DNA

Na década de 1970, abordagens visando o sequenciamento de moléculas de DNA foram desenvolvidas. A primeira técnica de sequenciamento, baseada em degradação química, foi desenvolvida por Maxam & Gilbert (Maxam & Gilbert, 1977). No mesmo ano, Frederick Sanger e sua equipe desenvolveram o *método de terminação de cadeia* (Sanger *et al*, 1977). Os primeiros sequenciadores

automáticos baseados no método de Sanger foram desenvolvidos em 1987 (Griffin & Griffin, 1993) e foram amplamente utilizados para sequenciamentos de genes e genomas de diferentes organismos até o final da década de 1990 e início dos anos 2000.

Posteriormente, novas tecnologias, capazes de realizarem o sequenciamento de forma mais rápida, foram desenvolvidas. As abordagens de sequenciamento de DNA que surgiram a partir do começo dos anos 2000 foram denominadas “Next-Generation Sequencing (NGS)” ou “Sequenciamento de Nova Geração” e se caracterização pelo grande volume de dados gerados durante o processo, o que acarretou em um aumento no volume de informação em bancos de dados públicos (Zhang, 2011).

2.3.5 O Alinhamento de Sequências

A disponibilidade de novas sequencias biológicas tornou necessário também o desenvolvimento de métodos de análise e comparação específicos para este propósito. Ainda em 1970, Neddleman e Wunsch desenvolveram um algoritmo capazes de realizar o alinhamento global de duas sequencias de proteínas a partir de uma matriz que correlaciona a probabilidade de mutação entre os diferentes aminoácidos (matriz de substituição) (Needleman & Wunsch, 1970). Posteriormente, em 1981, Smith e Waterman propuseram um novo algoritmo, capaz de realizar o alinhamento local de sequências de proteínas e nucleotídeos. O algoritmo de alinhamento local de Smith-Waterman realiza o alinhamento apenas nas regiões em que apresenta similaridade, o que otimiza o score na função em caso de uma grande distância evolutiva entre as sequencias (Smith & Waterman, 1981).

Os algoritmos de Neddleman-Wusch e de Smith-Waterman serviram de base para outras ferramentas de bioinformática, como os pacotes de programas FASTA (Pearson et al, 1998) e BLAST (Altschul et al, 1990) para busca em bancos de dados de sequências, sendo estas ferramentas incorporadas em praticamente todos os bancos de dados biológicos de sequências de Nucleotídeos e Proteínas, como o Genbank (Benson et al, 2013) e o Uniprot

(Aptweiler et al, 2004) (respectivamente). Além disso, estes algoritmos também serviram de base para os softwares de alinhamento múltiplo MUSCLE (Edgar, 2004) e CLUSTAL (Higgins & Sharp, 1988).

2.4 Ferramentas de Bioinformática para análise de proteínas

- Descritores físico-químicos: Parâmetros intrínsecos à sequência de cada proteína calculados a partir de informações previamente conhecidas das propriedades de cada aminoácido, como massa e carga elétrica. Massa molecular, perfil hidropático/hidrofóbico, ponto isoelétrico e área acessível por solventes são exemplos de descritores que podem ser determinados através de softwares como o EMBOSS (RICE et al, 2000) e algoritmos como o de Kyte e Doolittle (Kyte & Doolittle, 1982).

- Preditores de localização sub-celular: Utilizam ferramentas de classificação (Ex: Maquinas Vetoriais de Suporte) e reconhecimento de padrões (Ex: Modelos Ocultos de Markov, Redes Neurais) para predizer a localização celular de uma determinada proteína a partir de sua sequência, tomando como base análises prévias de proteínas já conhecidas (Imai & Nakai, 2010). Dentre as ferramentas mais utilizadas inclui-se o PSORT (Yu et al, 2010) e o TargetP (Emanuelsson, 2000).

- Preditores de Hélices Transmembrânicas: Identificam padrões topológicos de hélices transmembrânica em sequências de proteínas através de diferentes abordagens, incluindo análises estatísticas, como no caso do HMMTOP (Tusnády & Simon, 2001), e de modelos ocultos de Markov, como no caso do TMHMM (Krogh et al, 2001).

- Preditores de peptídeo sinal: Identificam sequências similares à peptídeos sinais e seus respectivos sítios de clivagem a partir de ferramentas de alinhamento e reconhecimento de padrões. O SignalP (Petersen et al, 2011) é uma das ferramentas mais utilizadas para este propósito.

2.5 Bancos de Dados Online em Bioinformática

2.5.1 Bancos de dados Biológicos

O crescimento no volume de informações disponíveis sobre biomoléculas, como ácidos nucleicos e proteínas, tornou necessário o desenvolvimento de ferramentas capazes de gerenciar estas informações (Wah, 2009). Os bancos de dados consistem em sistemas de gerenciamento de informações capazes de realizar operações de inserção, deleção, atualização e busca através de linguagens e comandos intrínsecos a cada modelo (Date, 2005). Dentre as diferentes categorias de bancos de dados biológicos disponíveis para acesso público destacam-se os bancos para sequências de nucleotídeos (Ex: Genbank), sequências de proteínas (Ex: Uniprot).

2.5.2 Bancos de dados de sequências de nucleotídeos

- Genbank*:* O Genbank é, atualmente, o principal banco de dados para sequências de nucleotídeo, contendo informações de genomas inteiros, fragmentos de cromossomos, genes, mRNAs e ESTs de diferentes organismos. Gerenciando pelo *National Center for Biotechnology* Information (NCBI), possui conexões com outras bases de dados este mesmo órgão, como o *Bioproject*, *Refseq* e *Pubmed*. Além disso, o padrão de arquivos nativo do Genbank, o *Genbank File Format*, (*.gb, *.gbk), é um dos formatos mais utilizados para o armazenamento de dados de sequências biológicas, assim como FASTA e o EMBL (Benson *et al*, 2013).

- ENSEMBL*:* Banco de dados curado para informações de genomas de organismos eucariotos, fornecendo dados de genes, transcritos (incluindo splicing alternativo), RNA não codificantes, expressão gênica diferencial, dentre outros, apresentando conexões de dados com outras fontes. Além da busca pelo seu website, este também oferece uma API (*Application*

Programming Interface) que permite a busca interativa pelo usuário através de *scripts* escritos em linguagem Perl (Hubbard, 2002)

- GeneDB: Banco de dados para genomas de organismos eucariotos e procariotos mantido pelo Instituto Sanger. De foram similar ao ENSEMBL e ao Genbank, o GeneDB também é curado e conectadas com diferentes bancos de dados para cada gene (Hertz-Fowler et al., 2004).

2.5.3 Bancos de dados de Proteínas

Uniprot: o Universal Protein Knowledgebase é um banco de dados europeu para informações de proteínas, contendo dados se sequências, anotações estruturais e ferramentas de busca. O seu dataset é disponibilizado em duas versões: Uniprot-Swissprot (curado manualmente e não-redundante) e Uniprot-trEMBL (não-curado e redundante) (Aptweiler et al, 2004).

Pfam: O Pfam é um banco de dados para famílias de proteínas, relacionando sequências, funções e localização sub-celular através de padrões conservados observados por alinhamento múltiplo e HMMs (*Hidden Markov Models – Modelos Ocultos de Markov*)(Bateman et al., 2000).

PRODOM: Banco de dados para domínios conservados entre proteínas gerado a partir da análise do conjunto de dados do Uniprot (Servant et al., 2002).

CATH e SCOP: Bancos de dados para classificação estrutural de proteínas a partir de dados de estruturas tridimensionais derivados do Protein Data Bank e sequências (Murzin, Brenner, Hubbard, & Chothia, 1995; Orengo et al., 1997).

Gene Ontology: Iniciativa de padronização da representação de genes entre diferentes bancos de dados, consistindo em um banco de classificação de genes quanto à função desempenhada em diferentes organismos (Ashburner et al., 2000).

2.6 Bancos de Dados para Parasitologia

2.6.1 Bancos de dados para protozoários

RMgmDB (Rodent Malaria genetic modified DataBase): Banco de dados para informações de mutantes genéticas e informações fenotípicas, correlacionadas com dados de biologia de sistemas (Khan, Kroeze, Franke-Fayard, & Janse, 2013)

LeishCys: Banco de dados para rotas metabólicas de *Leishmania major* (Saunders et al., 2012)

MalAvi: Bancos de dados para informações de distribuição geográfica e hospedeiros de Malaria para diferentes linhagens (Bensch, Hellgren, & Pérez-Tris, 2009).

PlasmoDB: Banco de dados para informações de genômica funcional de *Plasmodium* spp., incluindo anotação gene por gene, transcritos, genética de populações e evolução (The Plasmodium Genome Database Collaborative, 2001).

FULL-malaria: Banco de dados para sequências de *Plasmodium falciparum* derivadas de sequenciamento de cDNA (Watanabe, Sasaki, Suzuki, & Sugano, 2001)

2.6.2 Bancos de Dados para Helmintos

AceDB: Banco de dados desenvolvido no Instituto Sanger para o gerenciamento de dados genômicos de *Caenorhabditis elegans*, tendo posteriormente seu código-fonte liberado para ser usado como base para o desenvolvimento de outros bancos biológicos (Walsh, Anderson, & Cartinhour, 1998).

WormBase: Banco de dados para *C. elegans* e organismos relacionados (outros nematódeos), relacionando dados de genomas inteiros, transcritos, proteínas, expressão gênica e filogenias. Contêm informações sobre Foi desenvolvido com base no AceDB (Stein, Sternberg, Durbin, Thierry-Mieg, & Spieth, 2001).

HSD: Banco de dados para proteínas secretadas de nematódeos e trematódeos preditas a partir da análise de ESTs (Garg & Ranganathan, 2012).

2.7 Desenvolvimento de novos bancos de dados

Atualmente o número de ferramentas para gerenciamento de informações moleculares em parasitologia é consideravelmente limitado, sobretudo no que diz respeito ao parasitas pluricelulares. Desta forma, a criação de novos bancos de dados para informações destes organismos pode auxiliar futuros estudos em parasitologia molecular, otimizando a busca por informações.

3. Objetivos

3.1 Objetivos Gerais

- Desenvolvimento de um banco de dados para informações de proteínas de organismos da classe Cestoda.

3.2 Objetivos Específicos

- Recuperar informações disponíveis em bancos de dados públicos referentes a proteínas de organismos da classe cestoda.
- Recuperar informações disponíveis em bancos de dados públicos de padrões estruturais conservados.
- Realizar a anotação estrutural e físico-química das proteínas de encontradas.
- Cruzamento dos dados adquiridos.
- Construção de uma interface web para acesso as informações cruzadas e com ferramentas para análise dos dados.

4. Metodologia

4.1 Busca e indexação das sequências de proteínas

As sequencias de proteínas utilizadas para a construção do CestodaDB são derivadas do Uniprot, sendo feita a busca através das palavras-chave ‘taxonomy:Cestoda [6199]’, aonde ‘6199’ indica o código referente ao táxon Cestoda no banco em questão. As sequencias resultantes foram importadas em formato FASTA e indexadas em um banco de dados MongoDB, que armazena os dados em uma estrutural não-relacional (NoSQL) de documentos (Mongodb, 2013).

4.2 Busca dos dados para anotação das sequências

Os bancos de dados *PFam*, *Gene Ontology* e *CATH* foram importados em formato FASTA de seus respectivos repositórios. O banco de dados PRODOM foi importado em formato SRS e convertido para formato FASTA a partir de um script escrito em linguagem Python (Python, 2013).

4.3 Anotação das sequencias

A anotação das sequencias foi feita a partir da ferramenta BLAST a partir dos bancos de dados previamente baixados. Para a realização desta etapa, os bancos de dados em formato FASTA foram formatados para o formato nativo do BLAST através da ferramenta makeblastdb presente no pacote blast+2.2.5. Os dados resultantes de cada anotação foram inseridos no banco de dados MongoDB aonde estavam armazenadas as informações de cada sequência.

4.4 Anotação de peptídeo sinal, hélice transmembrânica e localização sub-celular

A análise preditiva de peptídeos, hélices transmembrânicas e localização sub-celular das proteínas foi realizada, respectivamente, com auxílio das ferramentas

SignalP (Petersen et al, 2011), TMHMM (Krogh et al, 2001) e TargetP (Emanuelsson et al, 2000). Todas as ferramentas foram executadas com suas configurações padrões com exceção da alteração para o modo “eucarioto” no caso do SignalP e TargetP.

4.5 Anotação de características físico-químicas

O cálculo da massa molecular de cada proteínas foi realizado com base na massa parcial de cada resíduo de aminoácido a partir de dados provenientes do Expasy (Expasy, 2013). A predição do ponto isoelétrico foi feita com base em uma implementação em python do *algoritmo naive de Kozlowski* (originalmente escrito em c++) (Isoeletricpoint, 2013). O Cálculo do valor GRAVY, que representa a hidropatia de uma proteína a partir de sua sequência de aminoácidos, foi feito com base nos valores de índice de hidropatia inicialmente propostos por Kyte & Doolittle (Kyte & Doolittle, 1982). Os valores calculados para cada proteínas foram indexados no banco de dados MongoDB.

4.6 Adição dos dados de ESTs

As sequencias de ESTs (Expressed Sequence Tag) de organismos da classe Cestoda foram buscada no banco de dados de ETSs do NCBI através das palavras-chave “cestoda[organism]” e baixadas em formato FASTA. O arquivo resultante foi convertido em um banco de dados BLAST pela ferramenta makeblastdb. Após isso, as sequencias de proteínas presentes no banco de dados em construção foram alinhadas com o banco de ESTs, filtrado para o organismo em questão, pela ferramenta BLASTx e os dados de cada *hit* foram inseridos no banco de dados do CestodaDB, sendo conectado com os dados de proteína pelo Uniprot ID da sequência de maior similaridade (mas apenas em caso de concordância entre o organismo de origem da EST e da Proteína).

4.7 Criação do Back-End e Front-End

O Back-end do CestodaDB foi desenvolvi em Linguagem Python 2.7 (PYTHON, 2013) e consiste em *scripts* CGI (*Common Gateway Interface*) que se conectam

com um banco de dados não relacional MongoDB através da biblioteca PyMongo (Pymongo, 2013). Os scripts CGI recebem requisições dos usuários através de formulários POST e rodam através do modulo *mod_python* instalado no servidor *Web HTTP Apache 2* que, por sua vez, está instalado em um Servidor Dell *Poweredge Xeon* com oito núcleos de 3,2 Ghz e 24 Gb de RAM rodando *Ubuntu Server 11*. O Front-end do CestodaDB foi desenvolvido em HTML/CSS. Uma representação do fluxo de informação envolvido durante a interação do usuário pode ser vista na Figura 1. A organização interna do banco de dados está representada na Figura 2.

Figura 1. Representação do fluxo de informação durante uma consulta no banco de dados. (I) Interação com site. (II) Envio das requisições para o servidor via formulário POST. (III) Passagem das requisições para scripts CGI escritos em linguagem Python através do modulo *mod_python* rodando em um *webserver Apache HTTP*. (IV) Consulta no banco de dados *mongo* feita pelos *scripts* CGI. (V) Retorno dos dados solicitados. (VI e V) envio da resposta gerada pelo CGI a partir dos dados resultantes da busca. (VII e VIII) Apresentação dos dados na forma de uma página resposta.

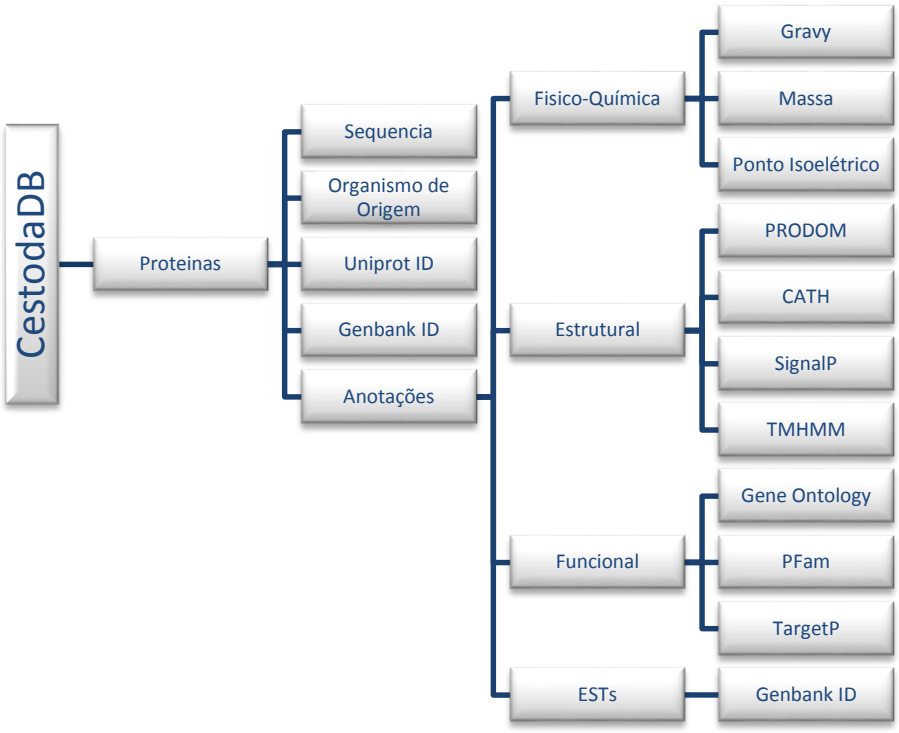

Figura 2. Representação hierárquica dos dados presentes no *CestodaDB*.

6. Resultados

O dataset final do *CestodaDB* é constituído por 3878 proteínas derivadas diretamente do Uniprot, sendo 61 provenientes de sua base de dados revisada (Uniprot-Swissprot) e 3817 de sua base não revisada (Uniprot-trEMBL). Estas proteínas estão divididas entre 353 espécies, por sua vez englobando 115 gêneros diferentes. Uma distribuição do número de proteínas por gênero de organismos está apresentada na figura 3.

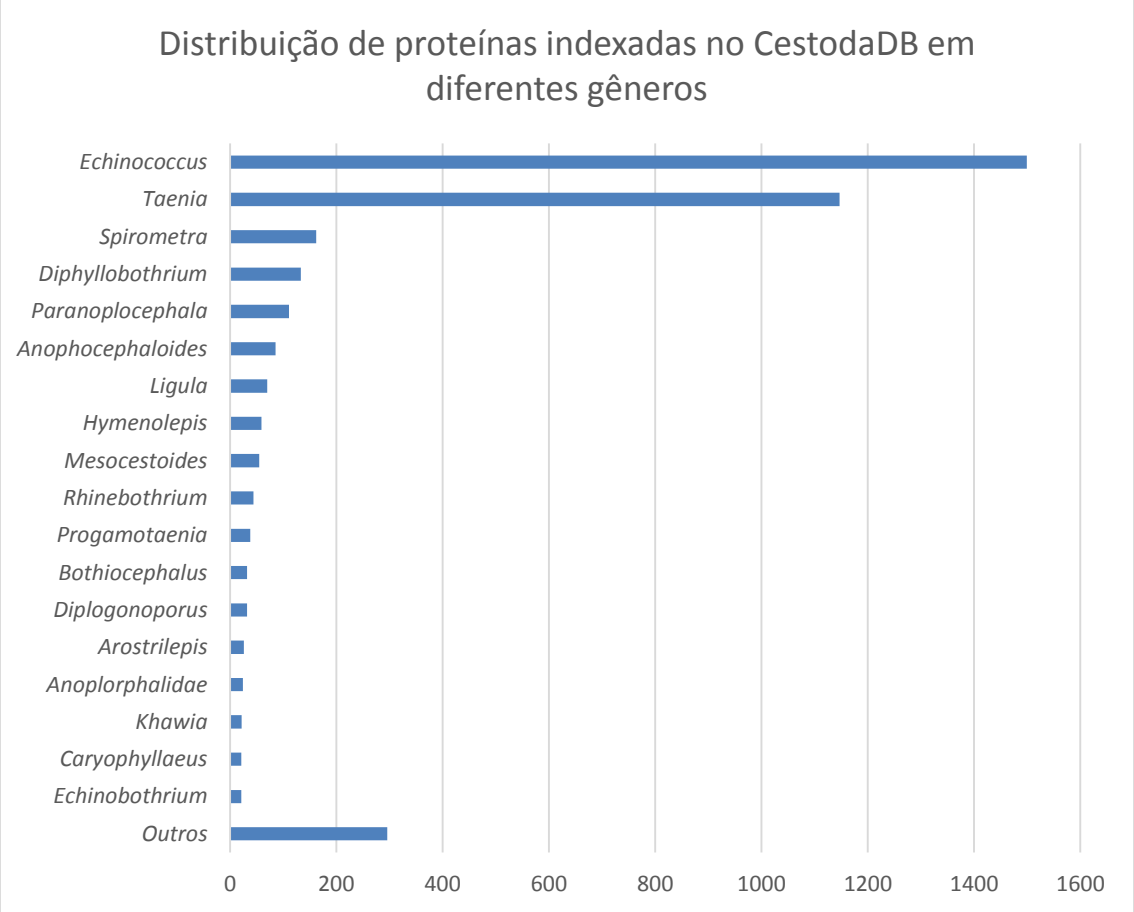

Figura 3. Distribuição das proteínas presentes no CestodaDB em diferentes gêneros de organismos.

O Limitado número de proteínas disponíveis em bancos públicos está diretamente relaciona as poucas informações genômicas sobre estes organismos em relação a outros grupos de parasitas. O sequenciamento de novos genomas, como os das espécies *Taenia solium*, *Echinococcus multilocularis*, *Echinococcus granulosus* e *Hymenolepis microstoma* (Tsai et al.,

2013; Zheng et al., 2013) promoverá um aumento significativo deste volume de informações, auxiliando na construção de plataformas similares a bancos já estabelecidos para outros grupos de parasitas, como o WormBase (Stein et al., 2001). Além disso, informações geradas por análises transcriptômicas, como RNA-Seq, também poderão auxiliar no maior entendimento dos transcritos e produtos expressos nestes organismos (Yang et al., 2012).

A análise estrutural pela ferramenta SignalP demonstrou que 10,6% destas proteínas apresentam, possivelmente, uma sequência de peptídeo sinal. Já a análise pela TMHMM demonstrou que, aproximadamente, 2,88% das proteínas estão localizadas no interior da membrana celular, 67,82% apresentam hélices transmembrânicas, 29,29% apresentam localização citoplasmática ou extracelular. Por fim, a análise pelo TargetP demonstrou que 68,15% das proteínas apresentam padrões relacionados à vias de secreção, enquanto 24,67% podem ser direcionadas para diferentes compartimentos celulares e 1,62% é de natureza mitocondrial.

Informações referentes a topologia destas proteínas, como a presença de peptídeos sinais e hélices transmembrânica, em conjunto com os dados de localização celular, podem auxiliar na identificação de novos alvos vacinais e de diagnóstico (Gan et al., 2010; Huang et al., 2009). Além disso, a combinação destes com as informações de ESTs pode permitir uma maior entendimento da fase do desenvolvimento em que as proteínas são expressas.

Das 95.797 sequencias de ESTs obtidas do banco de dados do NCBI, 5.912 (aproximadamente 6.17%) apresentaram similaridade com sequências de proteínas do banco de dados. Por sua vez, 1.478 das proteínas apresentaram similaridade com sequências de ESTs, sendo 983 pertencentes a organismos do gênero *Echinococcus*, 337 ao gênero *Taenia*, 147 ao gênero *Spirometra*, 5 ao gênero *Moniezia*, 5 ao gênero *Diphyllobothrium* e 1 ao gênero *Calliobothrium*.

Análises complementares das sequências de ESTs que não apresentaram similaridade com as sequências de proteínas presentem no banco de dados podem ser realizadas futuramente visando a identificação de novos genes.

Dentre estas análises inclui-se a abordagens de montagem, predição de CDSs e anotação funcional, de forma similar as desenvolvidas previamente para *Taenia solium* (Lundström et al., 2010), *Taenia asiatica* (Huang et al., 2009), *Echinococcus granulosus* (Gan et al., 2010) e nematódeos e trematódeos (Garg & Ranganathan, 2012).

O CestodaDB pode ser acessado através do endereço http://labbioinfo.ufpel.edu.br/cestodadb/, sendo em sua página inicial (Figura 4) disponibilizado uma ferramenta de busca por palavra-chave, permitindo a restrição por organismo, nome da proteína ou código de entrada no Uniprot. Uma página exemplo gerada para a apresentação de informações para a proteína *paramiosina* de *Taenia saginata* (Uniprot: Q8T305) está demonstrada na Figura 5.

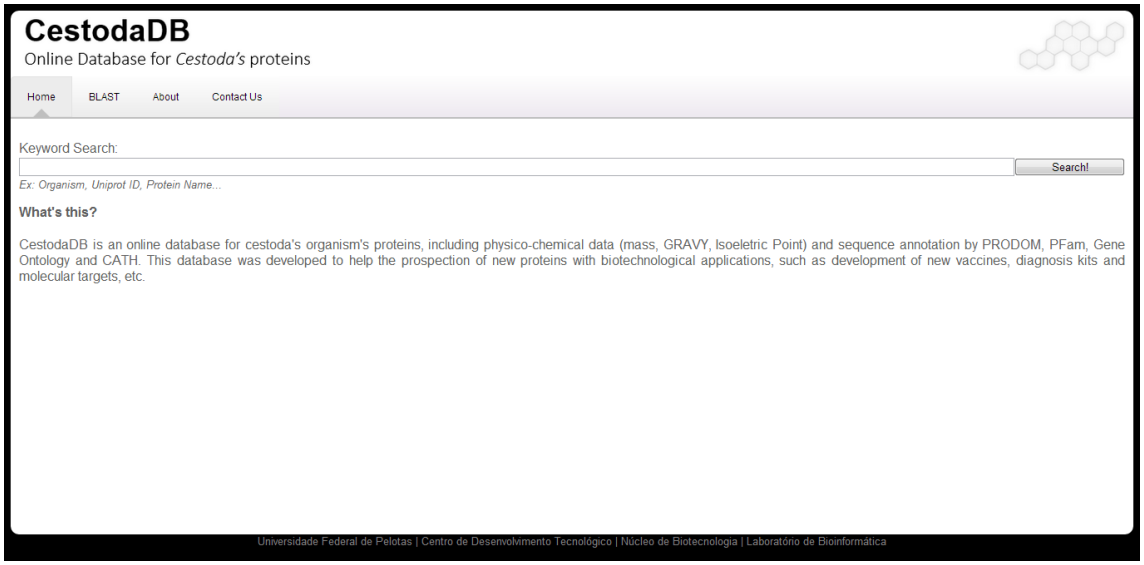

Figura 4. Página inicial do CestodaDB (disponível em htttp://labbionfo.ufpel.edu.br/CestodaDB/).

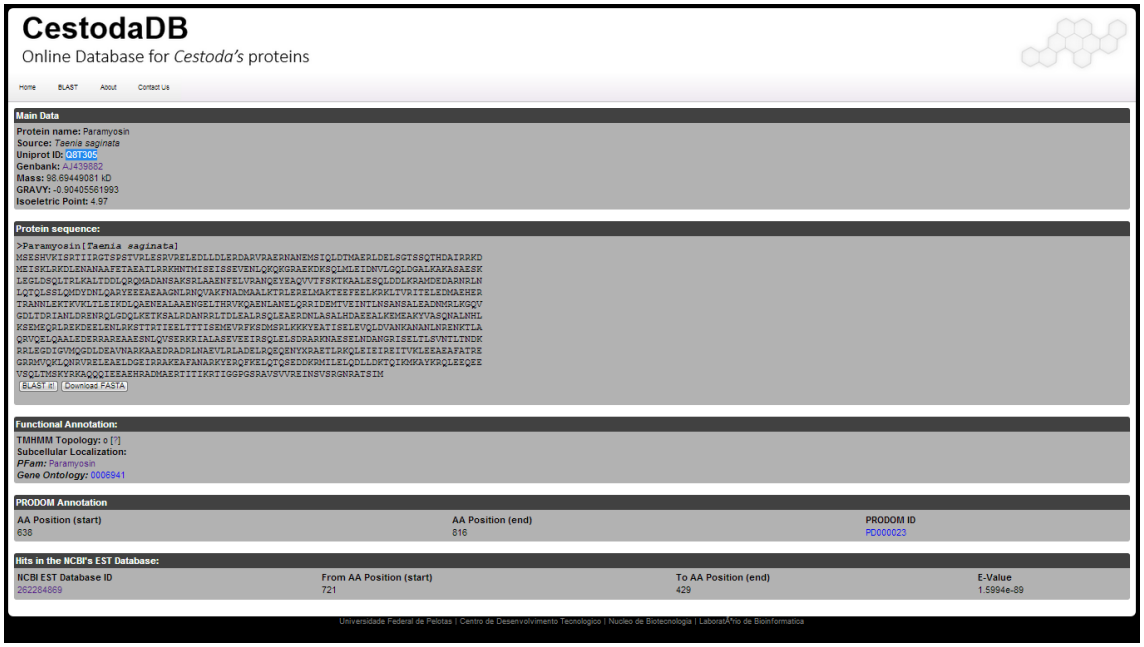

Figura 5. Página contendo as informações armazenadas no CestodaDB referentes a proteína *Paramiosina* de Taenia saginata (Uniprot: Q8T305).

7. Conclusão

No presente trabalho foi apresentando o desenvolvimento do CestodaDB, um banco de dados público para informações de proteínas de organismos da classe Cestoda. Apesar do número de proteínas destes organismos ser ainda limitado, a criação de uma plataforma direcionada para o armazenamento e recuperação destas informações poderá servir de base para futuros estudos sobre estes organismos. Além disso, a plataforma é escalável e novas informações poderão ser adicionadas à medida que novos dados sejam disponibilizados em bancos de dados públicos.

Futuramente, a adição de novas proteínas, assim como novas abordagens de anotação funcional e estrutural e a adição de novas ferramentas de busca, também poderá auxiliar no estudo da bioquímica e biologia molecular destes organismos, possibilitando também a identificação de proteínas de interesse biotecnológico, como alvos terapêuticos, vacinais e marcadores bioquímicos e moleculares para diagnóstico.

A atualização do banco de dados deverá ser realizada conforme novos dados genômicos e proteômicos destes organismos forem gerados e disponibilizados em outras bases de informação biológica.

8. Referências

ALTSCHUL, S. F.; GISH, W.; MYERS, E.; LIPMAN, D. Basic Local Alignment Search Tool. **Journal of Molecular Biology**, v. 215, n. 3, p. 403-410, 1990.

ANANTAPHRUTI, M. T.; NAWA, Y.; VANVANITCHAI, Y. Human sparganosis in Thailand: on overview. **Acta tropica**, v. 118, n. 3, p. 171-176, 2011.

ANSORGE, W. J. Next-generation DNA sequencing technologies. **Next Biotecnology**, v. 25, n. 4, p. 195-203, 2009.

APACHE HTTP Server. Disponivel em: <http://httpd.apache.org/>. Acessado em: 15 set. 2013

APWEILER, R; BAIROCH, A.; WU, C. H.; BARKER, W.; BOECKMANN, B.; FERRO, S.; GASTEIGER, E.; HUANG, H.; LOPEZ, R.; MAGRANE, M.; MARTIN, M.; NATALE, D.; O'DONOVAN, C.; REDASCHI, N.; YEAH, L. Uniprot: The universal protein knowledgebase. **Nucleic Acids Research**, v. 32, n. 1, p. 115-119, 2004.

ASYBURNER, M.; BALL, C. A.; BLAKE, J. A.; BOTSTEIN, D.; BUTLER, H.; Fubiology. **Nature genetics**, v. 25, p. 25-29, 2000.

BATEMAN, A.; BIRNEY, E.; DURBIN, R.; EDDY, S. R.; HOWE, K. L.; SONNHAMMER, E. L. The PFam protein families database. **Nucleic Acids Research**, v. 28, p. 263-266, 2000.

BENSCH, S.; HELLGREN, O.; PÉREZ-TRIS, J. MalAvi: a public database for malaria parasites and related haemosporidians in avian hosts based on mitochondrial cytochrome b lineages. **Molecular ecology resources**, v. 9, 2009.

BENSON, D.; CLARK, K.; KRASCH-MIZRACHI, I.; LIPMAN, D.; OSTELL, J.; SAYER, E. Genbank. Nucleic Acids Research (online), 2013.

BERMAN, H.; The Protein Data Bank: a historical persepective. **Acta Crystallographica (Section A)**, v. 64, p. 88-95, 2008.

CIMERMAN, B.; CIMERMAN, S. Parasitologia Humana e Seus Fundamentos Gerais - 2ª Ed. São Paulo: Atheneu, 2010.

DATE, C.J. Database in Depth. Cebastopol: O'Reilly, 2005.

DEL BRUTTO, O. H. Neurocysticercosis: a review. **The Scientific World Journal (online)**, 2012.

EDIGAR, R.; MUSCLE: Multiple sequence alignment with high accuracy and high throughput. **Nucleic Acids Research**, v. 32, n. 5, p. 1792-1797, 2004.

EMANUELSSON, O.; NIELSEN, H.; BRUNAK, S.; VON HEIJNE, G. Predicting subcellular localization of proteins based on theid n-terminal aminoacids sequence. **Journal of Molecular Biology**, v. 300, p. 1005-1016, 2000.

EXPASY. Disponível em: <http://www.expasy.org>. Acessado em: 10 out. 2013.

FISSER, A.; RODRÍGUEZ-CANUL, R.; WILLINGHAN, A. L.; Control of the taeniaosis/cysticercosis complex: future developments. **Veterinary parasitology**, v. 139, n. 4, 2006.

GAN, W.; ZHAO, G.; XU, H.; WU, W.; DU, W.; HUANG, J.; HU, X. Reverse Vaccinology approach indetify an Echinococcus granulosus tegumental membrane protein enolase as vaccine candidate. **Parasitology Research**, v. 106, n. 4, 2010.

GARG, G.; RANGANATHAN, S. Helminth secretome database (HSD): a collection of helmith excretory, secretory proteins predicted from expressed sequence tags (ESTs). **BMC genomics**, v. 13, 2012.

GRIFFIN, H. G.; GRIFFIN, A. M. DNA sequencing: Recent innovations and future trends. **Applied Biochemistry and Biotechnology**, v. 38, 147-159, 1993.

HALTON, David. Nutritional Adaptations to Parasistism within the Platyhelminthes. **International Journal of Parasitology**, v. 27, n. 6, p. 693-704, 1997.

HERTZ-FOWLER, C.; PEACOCK, C. S.; WOOD, V.; ASLETT, M.; KERHORNOU, A.; MOONEY, P.; BARREL, B. GeneDB: a resoure for prokaryotic and eukaryotic organisms. **Nucleic Acids Research**, v. 32, 2004.

HIGGINS, D.; SHARP, P. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. **Gene**, v. 73, n. 1, p. 237-244, 1998.

HOBERG, E. P; Phylogeny of Taenia: Species definitions and origins of human parasites**. Parasitology International**, v. 55, 2006.

HUANG, J.; HUANG, Y.; WU, X.; DU, W.; YU, X.; HU, X. Identification, expression, characterization and immunolocalization of lactata dehydrogenase from Taenia asiatica. **Parasitology Research**, v. 104, n. 2, p. 287-293, 2009.

HUBBARD, T. The ensembl genome database project. **Nucleic Acids Research**, v. 30, p. 38-41, 2002.

IMAI, K.; NAKAI, K. Prediction of subcellular localization of proteins: where to proceed? **Proteomics**, v. 10, n. 22, p. 3970-3983, 2010.

ISOELECTRICPOINT. Disponível em: <http://isoelectric.ovh.org/>. Acessado em: 3 out. 2013.

HUNT, L. T. Margaret O. Dayhoff 1925-1983. **DNA**, v. 2, p. 97-98, 1983.

KELLEY, S. Getting started with AceDB. **Briefings in Bioinformatics**, v. 1, n. 2, p. 131-137, 2000.

KERN, P. Clinical features and treatment of alveolar echinococcosis. **Current Opinion in Infectious Diseases**, v. 23, n. 5, p. 505-512, 2010.

KIM, D. G. PAEK, S. H.; CHANG, K. H.; WANG, K. C.; JUNG, H. W.; KIM, H. J.; CHI, J. G.; CHOI, K. S.; HAN, D. H. Cerebral sparganosis: clinical manifestations, treatment, and outcome. **Journal of Neurosurgery**, v. 85, n. 6, p. 1066-1077, 1996.

KHAN, S. M.; KROEZE, H.; FRANKE-KAYARD, B.; JANSE, C. J. Standardization in generating reporting genetically modified rodent malaria parasites: the RMgbDB database. **Methods in Molecular Biology**, n. 923, p. 139-150, 2013.

KROGH, A.; LARSON, B.; VON HEIJINE, G.; SONNHAMMER, E. L. L. Predicting transmembrane protein topology with hidden markov model: application to complete genomes. **Journal of Molecular Biology**, v. 305, p. 567-580, 2001.

KYTE, J.; DOOLITTLE, R. F. A simple method for displaying the hydropathic character of a protein. **Journal of Molecular Biology**, v. 157, n. 1, p. 105-132, 1982.

LESK, A. M. Introduction to Bioinformatics. Oxford: Oxford University Press, 2005.

LI, M. W.; SONG, H. Q.; LI, C.; LIN, H. Y.; XIE, W. T.; LIN, R. Q.; ZHU, X. Q. Sparganosis in mainland China**. International journal of infectous diseases (IJID),** v. 15, n. 3, p. 154-156, 2011.

LUNDSTROM, J.; SALAZAR-ANTON, F.; SHERWOOD, E.; ANDERSSON, B.; LIDH, J. analyses of expressed sequence tag library from Taenia solium, cysticerca. **PLoS neglected tropical diseases**, v. 4, n. 12, 2010.

MARKELL, E. K.; JOHN, D. T.; KROTOSKI, W. A. Markell e Vogue Parasitologia Médica - Oitava Edição. São Paulo: Guanabara Koogan, 2003.

MAXAM, A. M.; GILBERT, W. A new method for Sequencing DNA. **Proceedings of the National Academy of Sciences**, n. 74, v. 2, p; 560-564, 1977.

MODPYTHON. Disponível em: <http://modpython.org>. Acessado em: 10 jan. 2013.

MONGODB. Disponível em: <http://www.mongodb.org>. Acessado em: 10 jan. 2013.

MORO, P.; SCHANTZ, P. M.; Echinococcosis: a review. **International journal of infectious diseases (IJID)**, v. 13, n. 2, p. 125-133, 2009.

MURZIN, A. G.; BRENNER, S. E.; HUBBARD, T.; CHOTHIA, C.; SCOP: a structural classification of proteins database for the investigation of sequences and structures. **Journal of Molecular Biology**, v. 247, p. 536-540, 1995.

NCBI - EST. Disponível em: <http://www.ncbi.nih.gov/ncest/>. Acessado em: 8 out. 2013.

NCBI - Genbank. Disponível em: <http://www.ncbi.nih.gov/genbank/>. Acessado em: 8 out. 2013.

NCBI - Taxonomy. Disponível em: <http://www.ncbi.nih.gov/Taxonomy/>. Acessado em: 8 de out. 2013.

NEEDLEMAN, S.; WUHSCH, C. A general method applicable to the search for similarities in the amino acid sequence of two proteins. **Journal of Molecular Biology**, v. 48, n. 3, p. 443-453, 1970.

NEVES, D. P.; DE MELO, A. L.; LINARDI, P. M.; VITOR, R. W. A. Parasitologia Humana - 11ª Edição. São Paulo: Atheneu, 2008.

ORENGO, C. A.; MICHIE, A. D.; JONES, S.; JONES, D. T.; SWINDELLS, M. B.; THORNTON, J. M. CATH: a hierachic classification of protein domain structures. **Structure**, v. 5, p. 1093-1108, 1993.

PEARSON, W.; LIPMAN, D. Improved tools for biological sequence comparison. **Proceedings of the Nacional Academy of Science**, v. 85, n. 8, p. 2444-2448, 1988.

PETERSEN, T. N.; BRUNAL, S.; VON HEIKNE, G.; NIELSEN, H.; Signalp 4.0: discrimination signal peptides for transmembrane regions. **Nature Methods**, v. 8, p. 785-786, 2011.

PEVZNER, P.; SHAMIR, R. Bioinformatics for Biologists. Cambridge: Cambridge University Press, 2011.

PYTHON. Disponível em: <http://www.python.org>. Acessado em: 10 jan. 2013.

RAOUL, F.; LI, T.; YAKO, Y.; CHEN, X.; LONG, C.; YANAGIDA, T.; WU, Y.; NAKAO, M.; OKAMOTO, M.; CRAIG, P. S.; GIRAUDOUX, P. Advanves in diagnosis and spatial analysis of cysticercosis and Taeniasis. **Parasitology**, n. 140, p. 1578-1588, 2013.

REY, L. Parasitologia: parasitos e doenças parasitárias do homem nos trópicos ocidentais - 4ª Ed. Rio de Janeiro: Guanabara Koogan, 2008.

RICE, P.; LONGDEN, I.; BLEASBY A. EMBOSS: the European Molecular Biology Open Software Suite. **Trends in Genetics**, v. 16, n. 6, p. 276-277, 2000.

RODRIGUES, P.; WIKINS, P.; DORNY, P. Immunological and molecular diagnosis of cysticercosis. **Pathogens and Global Health**, v. 106, n. 5, p. 286-298, 2012.

SANGER, F.; NICKLEN, S.; COULSON, A. R. DNA sequencing with chain-terminationg inhibitors. **Proceedings of the National Academy of Science**, v. 74, n. 12, p. 5463-5467, 1997.

SAUNDERS, E. C.; MACRASE, J. I.; NADERER, T.; NG, M.; MCCONVILLE, M. J.; LIKIC, V. A. LeishCyc: a guide to build a metabolic pathway database and visualization of metabolomic data. **Methods in molecular biology**, v. 881, 505-529, 2012.

SERVANT, F.; BRU, C.; CARRERE, S.; COURCELLE, E.; GOUZY, J.; PEYRUC, D.; KAHN, D. ProDom: automated clustering of homologous domais. **Briefings in bioinformatics**, v. 3, p. 246-251, 2002.

SMITH, L. M.; SANDERS, J. Z.; KAISER, R. J.; HUGHES, P.; DODD, C.; CONNELL, C. R.; HENER, C.; KENT, S. B.; HOOD, L. E. Fluorescence detection in automated DNA sequence analysis. **Nature**, v. 321, n. 6071, p. 675-679, 1986.

SMITH, T. F.; WATERMAN, M. S. Identification of common molecular subsequences. **Journal of Molecular Biology**, v. 18, n. 1, 38-46, 1981.

SOTELO, J. Clinical Manifestations, Diagnosis, and Treatment of Neurocysticercosis. **Current Neurology and Neuroscience Reports**, v. 11, p. 529, 535, 2011.

STEIN, L.; STERNBERG, P; DURBIN, R.; THIERRY-MIEG, J.; SPIETH, J. WormBase: network access to the genome and biology Carnorhabditis elegans. **Nucleic Acids Research**, v. 29, n. 1, p. 82-86, 2001.

THOMPSON, R. C. A. The taxonomy, phylogeny, and transmission of Echinococcus. **Experimental parasitology**, v. 119, n. 4, p. 439-446, 2008.

The Plasmodion Genomes Database Collaborative. PlasmoDB: An integrative database of the plasmodium falciparum genome. **Nucleic Acids Research**, v. 29, n. 1, p. 66-69, 2001.

TORGERSON, P. R.; DEPLAZES, P. Echinococcosis: diagnosis and diagnostic interpretation in population studies. Trend in Prasitology, v. 25, n. 4, 2009.

TORTORA, G. J.; FUNKE, B. R.; CASE, C. L. Microbiologia - 8ª Edição. Porto Alegre: Artmed, 2007.

TSAI, I. J.; ZARAWIECKI, M.; HOLROYD, N.; GARCIARRUBIO, A.; SANCHEZ-FLORES, A.; BROOKS, K. L.; SCIUTTO, E. The genome of four tapeworm species reveal adaptations to parasitism. **Nature**, v. 296, n. 7443, p. 57-63, 2013.

TUSNÁDY, G. E.; SIMON, I. The HMMTOP transmembrane topology prediction server. **Bioinformatics**, v. 17, n. 9, 2001.

WAH, B. W. Wiley Encyclopedia of Computer Science and Engineering. New Jersey: Wiley, 2009.

WALKER, M. D.; ZUNT, J. R. Neuroparasitic infections: cestoda, trematoda and protozoans.**Seminars in neurology**, v. 25, n. 3, p. 262-277, 2005.

WALSH, S.; ANDERSON, M.; CARTINHOUR, S. W. ACEDB: a database for genome information. **Methods of biochemical analysis**, v. 39, p. 299-318, 1998.

WATANABE, J.; SASAKI, M.; SUZUKI, Y.; SUGANO, S. FULL-malaria: a database for a full-length enriched cDNA library from human malaria parasite, Plasmodium falciparum. **Nucleic acids research**, v. 29, n. 1, p. 70–71, 2001.

YANAGIDA, T.; SAKO, Y.; NAKAO, M.; NAKAYA, K.; ITO, A. Taeniasis and cysticercosis due to Taenia solium in Japan. **Parasites & vectors**, v. 5, n. 18, 2012.

YANG, D.; FU, Y.; WU, X.; XIE, Y; NIE, H.; CHEN, L.; YANG, G. Annotation of the transcriptome from Taenia pisiformis and its comparative analysis with three Taeniidae species, **PLoS one**, v. 7, n. 4, 2012.

YU, N. Y.; WAGNER, J. R.; LAIR, M. R.; MELLI, G.; REY, S.; LO, R.; DAO, P.; SAHINALP, S. C.; ESTER, M.; FOSTER, L. J.; BRINKMAN, F. S. PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. **Bioinformatics** v. 26, n. 13, p. 1608-1615, 2010.

ZHANG, J.; CHIDINI, R.; BADR, A.; ZHANG, G. The impact of next-generation sequencing no genomics. **Journal of Genetics and Genomics**, v. 38, p. 95-104, 2011.

ZHENG, H.; ZHANG, W.; ZHANG, L; ZHANG, Z.; LI, J.; LU, G.; WANG, S. The genome of the hydatid tapeworm Echinococcus granulosus. **Nature genetics**, v. 45, n. 10, p. 1168–1175, 2013.