UNIVERSIDADE FEDERAL DE SANTA CATARINA

NESTOR CUBAS WENDT

CONSTRUÇÃO DE UM BANCO DE DADOS E SISTEMA DE ANÁLISE SOBRE RESPOSTAS BIOLÓGICAS DE FAUNA AQUÁTICA EXPOSTA A DIFERENTES XENOBIÓTICOS

Florianópolis

2015

NESTOR CUBAS WENDT

CONSTRUÇÃO DE UM BANCO DE DADOS E SISTEMA DE ANÁLISE SOBRE RESPOSTAS BIOLÓGICAS DE FAUNA AQUÁTICA EXPOSTA A DIFERENTES XENOBIÓTICOS

Trabalho de Conclusão de Curso submetido à Universidade Federal de Santa Catarina como parte dos requisitos necessários para a obtenção do Grau de Bacharel em Ciências Biológicas. Sob orientação do Professor Doutor Afonso Celso Dias Bainy e coorientação do Doutor Guilherme de Toledo e Silva.

Florianópolis

2015

CONSTRUÇÃO DE UM BANCO DE DADOS E SISTEMA DE ANÁLISE SOBRE RESPOSTAS BIOLÓGICAS DE FAUNA AQUÁTICA EXPOSTA A DIFERENTES XENOBIÓTICOS

Este Trabalho de Conclusão de Curso foi julgado como adequado como parte dos requisitos necessários para obtenção do título de Bacharel em Ciências Biológicas.

Florianópolis, 26 de fevereiro de 2015.

_____

Professor Afonso Celso Dias Bainy, Dr.
Presidente da Banca – Orientador

_____

Professora Patricia Hermes Stoco, Dra.
Membro Titular

_____

Professor Guilherme Maciel Razzera, Dr.
Membro Titular

RESUMO

Atualmente a maioria das áreas costeiras mundiais apresentam danos causados por poluentes, impactando atividades como pesca e aquicultura. A exposição de organismos vivos a xenobióticos pode causar alterações em vários níveis de organização biológica, desde o nível molecular até o nível tecidual, que podem ser utilizadas como biomarcadores. Sequenciamentos de transcritomas são úteis para a ecotoxicologia pois ajudam os pesquisadores a entender os efeitos a nível molecular das respostas à exposição dos organismos vivos aos xenobióticos presentes no ambiente. O desenvolvimento das técnicas de sequenciamento gerou um grande fluxo de novos dados biológicos que precisam ser estudados. Neste trabalho, um sistema de análise de sequenciamentos foi desenvolvido, em conjunto com um banco de dados. A partir deste sistema, experimentos de peixes expostos a atrazina, cobre e fenantreno foram comparados. Seis programas de montagem de sequências foram comparados, e o programa CAP3 obteve os melhores resultados. Diversos *scripts* em Python foram desenvolvidos para automatizar e facilitar a análise dos

sequenciamentos, dentre eles estão análises de termos *Gene Ontology* e vias metabólicas. Um esquema de tabelas relacionais para o MySQL foi criado. Este passou por normalização, evitando redundância nos dados. A exposição de peixes das espécies *Prochilodus lineatus* e *Poecilia vivipara* a atrazina, cobre e fenantreno gerou respostas semelhantes, tais como, ativação de genes envolvidos na imunidade inata, do sistema complemento e do sistema de coagulação sanguínea. O sistema de análise e o banco de dados desenvolvido devem auxiliar as análises de sequenciamentos no LABCAI.

Palavras-chave: bioinformática; banco de dados biológico; hibridização subtrativa supressiva; sistema de análise integrada.

ABSTRACT

Presently most of the coastal areas already show damage caused by pollutants, impacting activities such as fishing and aquaculture. The exposure of living organisms to xenobiotics can cause changes at various levels of biological organization, from molecular to tissue level, which can be used as biomarkers. Transcriptome sequencing is useful for ecotoxicology helping researchers to understand the effects at molecular level of the organisms exposed to xenobiotics in the environment. The development of sequencing methods created a large flow of new biological data that need to be analyzed. In this work, a sequencing analysis system was developed together with a database. Furthermore, experiments of fish exposed to atrazine, copper and phenanthrene were compared. Six sequence assembly softwares were compared, and the software CAP3 got the best results. Several Python scripts were developed to automate and facilitate the sequencing analysis, including Gene Ontology term and metabolic pathway analysis. A MySQL scheme for relational tables was created. It went through standardization, avoiding redundancy in the data. The exposure of *Prochilodus lineatus* and *Poecilia vivipara* to atrazine, copper and

phenanthrene yield similar responses, with activation of genes involved in innate immunity, complement system and clotting. The analysis system and database developed should help sequencing analysis in LABCAI.

Keywords: bioinformatics; biological database; subtractive suppression hybridization; integrated analysis system.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

DBMS *Database Management System* – Sistemas de gerenciamento de bancos de dados

GO *Gene Ontology* – Ontologia Gênica

HPA Hidrocarboneto policíclico aromático

HTML *HyperText Markup Language* – Linguagem de Marcação de Hipertexto

LABCAI Laboratório de Biomarcadores de Contaminação Aquática e Imunoquímica

NCBI *National Center for Biotechnology Information* – Centro Nacional de Informação Biotecnológica

NR *Non redundant protein database* – Banco de dados de proteínas não redundantes

ORF *Open reading frame* – Fase aberta de leitura

PDB *Protein Data Bank* – Banco de Dados de Proteínas

SSH *Supression subtractive hybridization* – Hibridização subtrativa supressiva

UFSC Universidade Federal de Santa Catarina

UNIPROT *Universal Protein Resource* – Repositório Universal de Proteínas

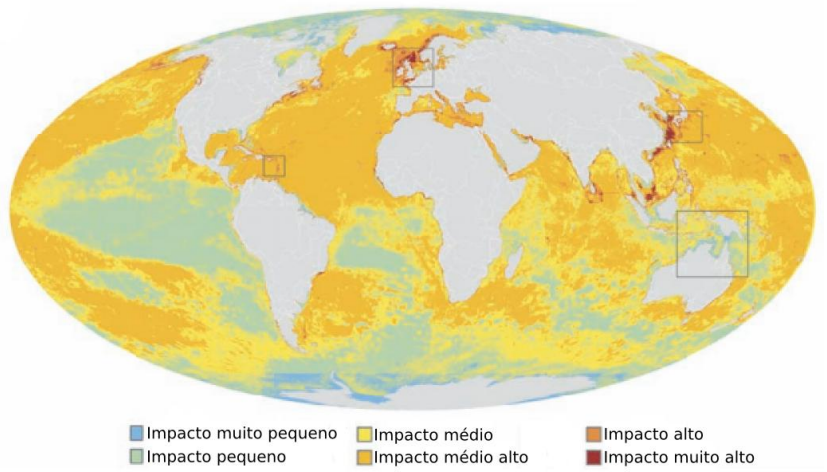
SUMÁRIO

1. INTRODUÇÃO

1.1 Contaminação de ecossistemas aquáticos

Muitas atividades humanas, como pesca e aquicultura, dependem de ambientes costeiros e marinhos, entretanto, desde as primeiras civilizações humanas, esses ambientes são afetados direta ou indiretamente por poluentes (Figura 1) (UNESCO, 2011; ISLAM e TANAKA, 2004). Hoje, a maioria das áreas costeiras mundiais já apresentam algum dano causado por poluentes (TANAKA, 2004), principalmente por compostos que geram grande preocupação como óleos, metais pesados, patógenos, herbicidas, hormônios, antibióticos, etc (WILLIAMS, 1996). Em 2012, mais de 2,8 bilhões de pessoas viviam a menos de 100 quilômetros da costa, e a rápida urbanização deve aumentar esse número nos próximos anos (UNESCO, 2011). A consequência dessa situação é que grande parte da contaminação tem origem em áreas populosas e de grande importância

econômica, o que acaba afetando a qualidade de vida nestes locais (MOORE et al., 2014). Do ponto de vista econômico, estima-se que os ecossistemas aquáticos conservados, que atuam na regulação climática, recreação e servem de fonte de alimentos, água, etc, valham mais de 33 mil dólares



por hectare por ano (DE GROOT et al., 2012).

Figura 1 – Mapa mundial dos impactos de atividades humanas em ambientes costeiros e marinhos.

Fonte: HALPERN et al., 2004

Estes contaminantes podem gerar efeitos tóxicos em organismos sensíveis, eventualmente causando impacto sobre as respectivas populações, através do aumento da mortalidade e de efeitos indiretos que também podem ocorrer dependendo da duração e intensidade da exposição. A disponibilidade de alimentos, as taxas de decomposição, os níveis de oxigênio, etc, podem ser influenciadas por contaminantes (FLEEGER; CARMAN; NISBET, 2003) e assim causar mudanças no funcionamento do ecossistema. Várias alterações na estrutura dos ecossistemas aquáticos já foram documentadas em razão da poluição aquática (AUSTEN; WARWICK; CARMEN ROSADO, 1989; SMITH; TILMAN; NEKOLA, 1999; WU, 1999; TANAKA, 2004; ZOU et al., 2011; IANNELLI et al., 2012).

1.1.1 Contaminantes: Atrazina, cobre e fenantreno

Neste trabalho serão avaliadas respostas de fauna marinha exposta aos contaminantes atrazina, cobre e fenantreno.

Atrazina é o segundo herbicida mais utilizado nos EUA (KIELY et al., 2004). Ele funciona inibindo a fotossíntese e é utilizado para controlar ervas daninhas de folhas largas. Nos últimos anos, o risco ambiental apresentado pela atrazina em ambientes aquáticos está sendo reavaliado pela Agência de Proteção ao Meio Ambiente dos EUA (U.S. EPA, 2003; U.S. EPA, 2007).

O Cobre é de modo geral tóxico para organismos aquáticos (TAYLOR et al., 1996; CLEARWATER et al., 2002). Na exposição aguda, há danos diretos à órgãos, como as brânquias que causam problemas respiratórios. Em exposição crônica, entretanto, sabe-se que o cobre gera mudanças neurológicas e endócrinas (LINDER, 1991).

O Fenantreno é um hidrocarboneto aromático policíclico (HAP) composto de três anéis de benzeno. É encontrada no alcatrão do tabaco, e também é uma substância conhecidamente irritante e fotossensibilizante da pele. HPAs estão altamente presentes no meio ambiente devido a sua ocorrência em petróleo, carvão, etc (HARDIN et al., 1992; MUDZINSKI et al., 1993; COOKE e DENIS, 1988).

1.2 Respostas biológicas e biomarcadores

A exposição dos organismos vivos a xenobióticos pode causar alterações em muitos processos fisiológicos (HANDY E DEPLEDGE, 1999). Essas alterações podem ser divididas em duas categorias: as que ajudam o organismo e se proteger dos efeitos tóxicos do xenobiótico e as que não o protegem. Um alteração protetora, por exemplo, é a indução de metalotioneínas, que diminuem a biodisponibilidade de metais no organismo (WALKER et al., 2001). Por outro lado,

muitos químicos acabam se ligando ao DNA e formam adutos que podem levar a mutações (WALKER et al., 2001). Alterações fisiológicas bem documentadas incluem distúrbios respiratórios, cardiovasculares, osmorregulatórios, neurológicos e endócrinos (HANDY E DEPLEDGE, 1999). Geralmente, as primeiras respostas observadas em animais expostos são à nível molecular, através do aumento ou diminuição da expressão de diversos genes (BRULLE et al., 2008). Como os xenobióticos interagem com as várias moléculas presentes nas células, acredita-se que a compreensão da ecotoxicologia à nível molecular é um ponto chave para um melhor entendimento dos efeitos biológicos desses compostos (FOWLER, 2005).

Biomarcadores são alterações à nível molecular, celular ou fisiológico que revelam efeitos causados por poluentes (WALKER et al., 2001). Através da sua utilização é possível identificar organismos que foram expostos a contaminantes, como também a magnitude dessa exposição

(CAJARAVILLE et al., 2000). A grande vantagem da utilização dos biomarcadores como meio de monitoramento ambiental está no potencial que eles possuem em predizer mudanças nos níveis populacionais ou ecossistêmicos (CAJARAVILLE et al., 2000). Assim, ao se identificar alguma contaminação, há tempo para realização de estratégias de biorremediação ou ainda de evitar que algum dano permanente ocorra no ambiente.

1.3 Sequenciamentos de DNA

O sequenciamento de DNA é o processo de determinação da ordem de nucleotídeos em uma molécula de DNA. Várias técnicas de sequenciamentos foram desenvolvidas, sendo notório o método conhecido como método dideoxi ou de terminação de cadeia (SANGER; NICKLEN; COULSON, 1977), desenvolvido por Frederick Sanger e colaboradores em 1977, rendendo o prêmio Nobel de Química em 1980. O sequenciamento baseado no método de Sanger foi

muito utilizado durante décadas, mas para sequenciamentos em larga escala vêm sendo substituído pelos métodos de sequenciamento de nova geração (SHENDURE e JI, 2008).

Sendo o DNA a molécula que contém as parte das instruções sobre o funcionamento e desenvolvimento dos seres vivos, estudá-la pode trazer inúmeros benefícios à sociedade (COLLINS et al., 2003). Com o sequenciamento do genoma humano, por exemplo, pesquisadores foram capazes de identificar genes relacionados à diversas doenças e de desenvolver fármacos específicos à determinadas proteínas alvo. Na agricultura, foi possível melhorar culturas e identificar biopesticidas (COLLINS, et al., 2003).

O dogma central da biologia molecular, enunciado pela primeira vez por Crick em 1958 (CRICK et al., 1970), é um conceito que trata do fluxo de informações genéticas através de moléculas biológicas. Segundo ele, muitas informações necessárias para síntese de RNAs e proteínas encontram-se no DNA. Para se produzir

uma proteína, entretanto, é necessário que o DNA seja convertido em RNA mensageiro (RNAm), molécula responsável por levar a informação aos ribossomos, onde ocorre a síntese de proteínas. Além disso, para que as células tenham a capacidade de se multiplicar, é necessário que o DNA possa ser duplicado.

O conjunto de moléculas de RNAm em uma célula, em determinado estágio de desenvolvimento ou condição fisiológica é chamado de transcriptoma (WANG; GERSTEIN; SNYDER, 2009). Especificamente para a ecotoxicologia, estudos de transcriptômica podem ajudar pesquisadores a entender os efeitos e respostas moleculares que xenobióticos, sozinhos ou em conjunto com outros estressores, geram nos organismos vivos e consequentemente no meio ambiente (SCHIRMER et al., 2010).

1.4 Bioinformática

Com o desenvolvimento das técnicas de sequenciamento em larga escala e popularização de projetos genomas e transcriptomas, criou-se um grande fluxo de novos dados biológicos que precisam ser estudados (VERLI, 2014). A tendência deste fluxo é de continuar crescendo (Figura 2), levando as ciências biológicas ao grupo de ciências que lidam com grandes conjuntos de dados, como a astronomia e física (MARX, 2013). A complexidade dos dados biológicos cria vários desafios aos pesquisadores. Enquanto que na física as informações são bem estruturadas e anotadas, na área biológica elas são bastante difíceis de se organizar. Além das sequências, seja de nucleotídeos ou aminoácidos, cientistas devem levantar informações sobre outros componentes celulares e condições ambientais. Muitos desses fatores são pouco conhecidos e com isso as análises se tornam bastante difíceis. Enquanto novos dados biológicos forem gerados, precisaremos de avanços em *hardware*, *software* e

novas estratégias de análise e armazenamento de dados.

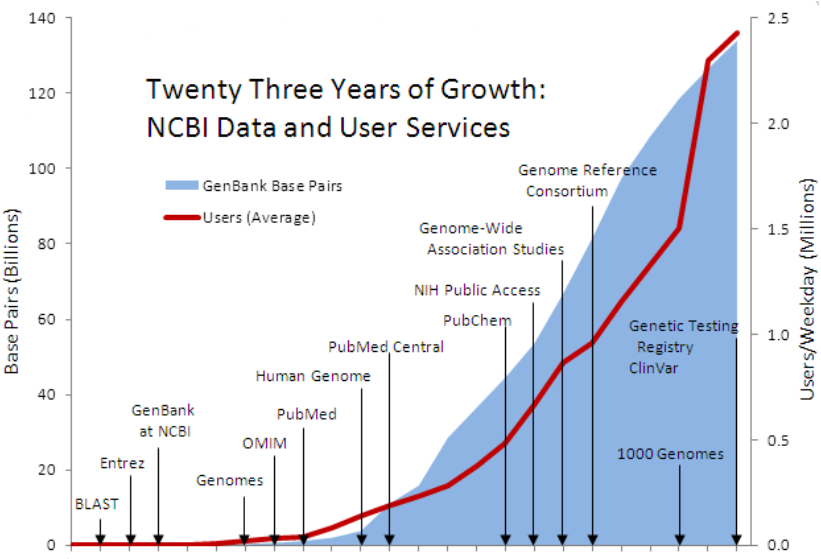


Figura 2 – Crescimento do número de usuários e sequências depositadas no National Center for Biotechnology Information (NCBI).

Fonte: http://www.nlm.nih.gov/about/2014CJ.html

Diversos métodos computacionais são intrínsecos às ciências biológicas (COLLINS et al., 2003). Com o aumento da informação disponível aos pesquisadores a partir de novas técnicas ou

bancos de dados tornou necessário a aplicação de modelagem matemática, simulação *in silico* para estudo de sistemas biológicos e vários outros métodos (WELCH et al., 2014).

A bioinformática é comumente definida como a disciplina que faz utilização de métodos computacionais na análise e armazenamento de dados biológicos. Teve origem na década de 1960, com a utilização de programas para visualização de estruturas tridimensionais de proteínas (VERLI, 2014), e foi impulsionada pela necessidade de criação de grandes bancos de dados para armazenar e disponibilizar sequências de DNA e proteínas, a partir da popularidade do sequenciamento de DNA. O avanço e barateamento dos computadores também tornaram os pesquisadores capazes de abordar problemas cada vez mais complexos. Hoje, a bioinformática é uma área interdisciplinar, abrangendo muitas disciplinas exatas como a estatística e a matemática e tem como objetivo aumentar o conhecimento de processos biológicas, como a

identificação de genes em um genoma ou a caracterização da estrutura tridimensional de uma proteína. Apesar de sua importância, ainda há necessidade de pessoas bem treinadas em bioinformática (WELCH et al., 2014).

1.4.1 Cromatogramas e montagem de sequências consenso

Ao final de um sequenciamento pelo método de Sanger, os pesquisadores obtêm um arquivo chamado cromatograma, que contém as leituras dos dideoxinucleotídeos fluorescentes (Figura 3). A partir dele, é possível determinar a sequência (*reads*) e a qualidade dos nucleotídeos sequenciados através da análise dos picos de leitura do cromatograma. Programas conhecidos como nomeadores de base são capazes realizar essa tarefa. Dentre eles, o Phred (EWING et al, 1998; EWING e GREEN, 1998) é bastante utilizado por também ser capaz de realizar o *trimming* (remoção de bases de baixa qualidade a partir das

extremidades dos *reads*) automaticamente com a leitura dos cromatogramas.

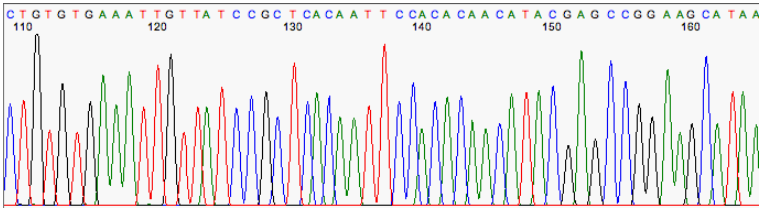


Figura 3 – Cromatograma obtido de um sequenciador que utiliza o método de Sanger. Os picos obtidos a partir de diferentes comprimentos de onda para cada base são utilizados para determinar a sequência dos nucleotídeos.

Fonte: http://www.biol.unt.edu/~jajohnson/Chromatogram_Interpretation

Até o momento, os métodos de sequenciamento são incapazes ler uma grande sequência de DNA, como um cromossomo, de ponta a ponta. Por isso técnicas como a de Sanger e outros sequenciadores de nova geração fragmentam o DNA em fragmentos menores. Desta forma, após os sequenciamentos, é necessário que

o DNA seja reconstruído através de um processo conhecido como montagem. Os programas Phrap (EWING et al, 1998; EWING e GREEN, 1998), CAP3 (HUANG; MADAN, 1999) e MIRA (CHEVREUX; WETTER; SUHAI, 1999) são capazes de comparar os *reads* provenientes da metodologia de Sanger e produzir uma sequência consenso (*contig*) (Figura 4). O método mais utilizado para produzir *contigs* em neste tipo de sequências é o de *overlap-consensus,* no qual primeiramente todos os *reads* são comparados entre si. Na segunda etapa é realizada a determinação da posição relativa dos *reads* ao longo da sequência e a terceira consiste na escolha do mais provável nucleotídeo para cada posição no *contig*, levando em conta os valores de qualidade (LI et al., 2012).

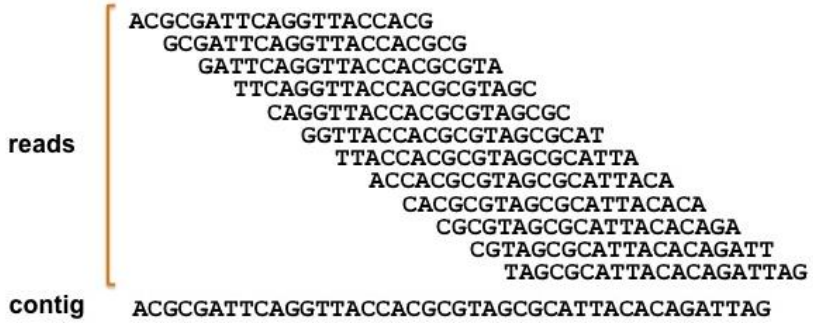


Figura 4 – Montagem de fragmentos sobrepostos em uma sequência consenso.
Fonte: http://contig.wordpress.com/2010/02/09/how-newbler-works/

1.4.2 Bancos públicos de dados biológicos

Os bancos de dados biológicos são repositórios de informações coletadas em experimentos científicos, literatura e análises computacionais. Os primeiro bancos surgiram no início da década de 1970, onde pesquisadores, influenciados pelo crescimento das ciências da computação e acúmulo de dados biológicos, começaram a agrupar resultados em portais de livre acesso (COORAY, 2012). O primeiro banco de

dados biológico surgiu em 1971 e foi chamado de *Protein Data Bank* (PDB) (BERNSTEIN et al., 1977), tendo como objetivo a disponibilização de informações sobre a estrutura de várias proteínas. A partir daí, bancos de dados com os mais variados objetivos foram criados, como aqueles focados em genomas, vias metabólicas ou famílias de proteínas.

Hoje, dentre os bancos biológicos públicos de sequências proteicas, podemos citar o Swiss-prot, mantido desde 2002 pelo *Universal Protein Resource* (UNIPROT) (UNIPROT CONSORTIUM, 2008), que contém somente sequências de alta qualidade e manualmente curadas. Outro repositório importante é o banco não redundante de proteínas (NR) mantido pelo NCBI, que congrega vários bancos de sequências proteicas mantidos por diferentes organizações, formando uma base central (http://www.ncbi.nlm.nih.gov/protein). Também existem bancos de dados focados na disponibilização de informações sobre famílias de proteínas, como o Pfam (FINN et al., 2014), ambos

disponibilizando modelos ocultos de Markov (HMM) para análise. Além deles, o PRINTS (ATTWOOD et al., 1994) é um banco de *fingerprints* de proteínas, ou seja, grupos de motivos conservados que podem ser utilizados para identificação de determinada família.

A identificação das vias metabólicas envolvidas em uma determinada resposta biológica é uma das maneiras de se estudar o fenômeno. Vários projetos se propõem a fornecer informações sobre vias metabólicas, entre eles, o Reactome (JOSHI-TOPE et al., 2005) é apresentado como um repositório de anotações manualmente curadas sobre vias metabólicas.

A iniciativa *Gene Ontology* (GO) provê um banco de dados que tem como objetivo fornecer um vocabulário preciso e controlado para as funções dos genes em todos os organismos (ASHBURNER et al., 2000). Esses termos são divididos em um dos seguintes domínios: processo biológico, função molecular ou componente celular. Processo biológico se refere ao objetivo biológico para o qual

o produto gênico em questão contribui. Termos como '*translation'* ou '*cell growth and maintenance'* fazem parte deste domínio. Função molecular está relacionada com a atividade bioquímica do produto gênico e termos como '*enzyme*' ou '*transporter*' estão nesta categoria. Componente celular é definido como o local onde o produto gênico está ativo, por exemplo, '*nuclear membrane*' ou '*proteosome*'. É importante notar que cada domínio possui termos em muitos níveis hierárquicos e a anotação de determinado produto gênico em um nível depende da quantidade de informação que existe sobre o mesmo (Figura 5).

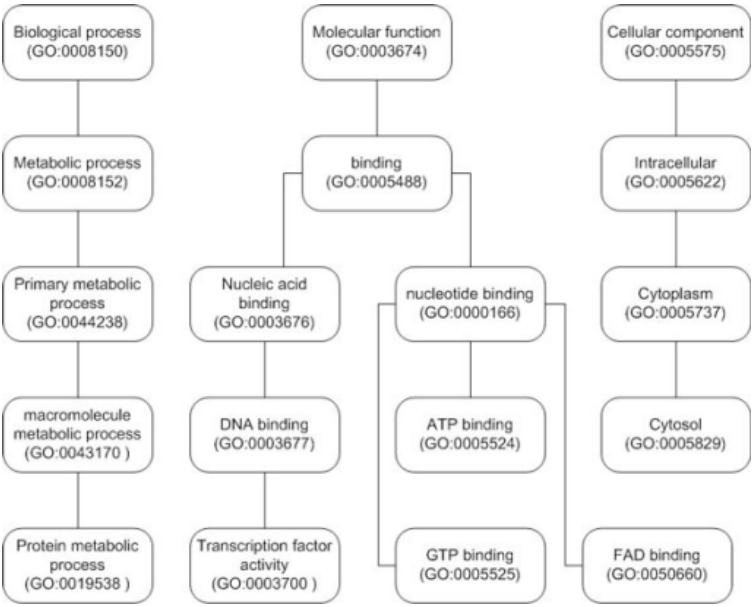


Figura 5 – Exemplo de hierarquia dos termos GO. Fonte: http://openi.nlm.nih.gov/detailedresult.php?img=3009494_1471-2105-11-S1-S23-2&req=4

1.4.3 Alinhamentos

Ao isolar uma sequência de DNA ou proteína, comumente os pesquisadores tentam identificar o gene ou função molecular pelo qual ela é responsável. Como sequências similares

geralmente possuem funções bastante parecidas, é possível inferir a função de determinada sequência ao calcular a similaridade entre esta e outra sequência já conhecida depositada em um banco de dados. Entre as várias técnicas existentes de obtenção dos níveis de similaridade entre sequências, os alinhamentos são bastante utilizados, pois através deles pode-se comparar uma sequência alvo com milhares de dados presentes nos vários bancos de dados de forma rápida (VERLI, 2014). Se o alinhamento de determinada sequência for estatisticamente significante, obtêm-se a hipótese de que a sequência alvo possui a mesma função da sequência alinhada.

Atualmente existem vários algoritmos capazes de alinhar duas sequências. Entre os alinhamentos simples, temos os globais e os locais (Figura 6). Nos alinhamentos globais a sequência alvo é alinhada completamente, geralmente havendo adição de vários *gaps* (espaçamentos). Os alinhamentos globais geram melhores resultados

em sequências de tamanhos semelhantes com alguma similaridade. Nos alinhamentos locais, somente os trechos com grande similaridade são alinhados. Neles os espaçamentos são menos comuns, tornando esse tipo de alinhamento mais apropriado para sequências de tamanhos diferentes ou que possuem somente um trecho conservado. Uma das ferramentas mais utilizadas, hoje, na análise de sequências é o programa BLAST (ALTSCHUL et al., 1997), que realiza o alinhamento local de sequências.

Figura 6 – Comparação entre um alinhamento global (acima) e um alinhamento local (abaixo).
Fonte: http://rosalind.info/problems/swat/

Além dos alinhamentos simples, existem algoritmos capazes de realizar alinhamentos

múltiplos, ou seja, de várias sequências ao mesmo tempo (Figura 7) (EDGAR, 2006). Do ponto de vista computacional, esse tipo de alinhamento apresenta um problema muito maior. Os alinhamentos múltiplos são geralmente utilizados na identificação de regiões conservadas entre grupos de sequências relacionadas evolutivamente. Comparado aos alinhamentos simples, os múltiplos são mais precisos na detecção de sequências homólogas distantes, graças a seu modelo matemático. O programa HMMER3 (EDDY, 2010) é utilizado na identificação de sequências homólogas de proteínas, através de alinhamentos múltiplos e HMM.

Figura 7 – Alinhamento múltiplo das sequências de várias proteínas. Regiões coloridas indicam conservação.

Fonte: http://www.bioinfo.ifm.liu.se/edu/TFTB29/HT 2013/assignment3.html

1.4.4 Sistemas de gerenciamento de bancos de dados (DBMS)

Os DBMS são programas utilizados para criar, gerenciar e manter um banco de dados, sendo utilizados em várias disciplinas há décadas (NELSON; REISINGER; HENRY, 2003). Eles são responsáveis por manter a integridade dos dados, ou seja, mantê-los continuamente acessíveis e organizados aos usuários, e também por manter a segurança dos dados, através de *backups* e

sistemas de permissões para acesso. Existem vários tipos de DMBS, sendo o relacional o mais utilizado. Nele, as informações são normalizadas em tabelas de forma que elas podem ser relacionadas entre si (Figura 8) sem a necessidade de reestruturação do banco. Hoje existem várias ferramentas disponíveis para a construção de bancos de dados, entretanto, somente o uso dessas ferramentas não garante uma boa estrutura para um banco de dados biológicos (NELSON; REISINGER; HENRY, 2003). Um bom esquema relacional deve ser capaz de prover a informação de forma fácil e intuitiva para seu público-alvo.

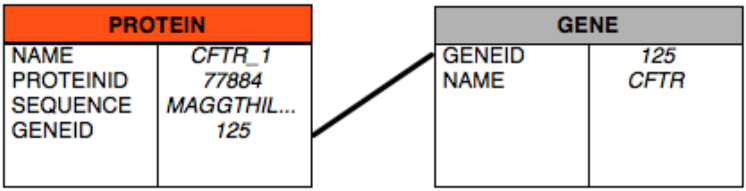


Figura 8 – Modelo contendo duas tabelas relacionadas entre si pela variável GENEID.
Fonte: http://www.gmod.org/wiki/Overview#Relational_Databases

Assim, levando em conta a importância da compreensão dos impactos causados por xenobióticos nos ecossistemas costeiros e o rápido acúmulo de informações moleculares nas ciências biológicas, entende-se que o desenvolvimento de um sistema de análise e banco de dados para a ecotoxicologia é uma iniciativa interessante o grupo de pesquisas do LABCAI.

2. OBJETIVOS

2.1 Objetivo Geral

Construir e implementar um sistema de análise e banco de dados de sequências obtidas no LABCAI, UFSC, referentes a experimentos de ecotoxicologia molecular.

2.2 Objetivos Específicos

- Organizar dados de sequenciamento de cDNA de várias espécies de fauna aquática expostas a diferentes xenobióticos, provenientes de experimentos *in situ* e *ex situ* gerados no LABCAI.

- Determinar a melhor maneira de realizar a montagem das sequências consenso.

- Importar as sequências e anotações em um banco de dados MySQL.

- Automatizar o processo de análise, anotação e depósito das informações no banco de dados.

- Comparar os resultados das análises em busca de padrões de respostas biológicas.

3. MATERIAIS E MÉTODOS

3.1 Sequenciamentos realizados no LABCAI

Foram utilizados sequenciamentos realizados no LABCAI, desde 2008, obtidos a partir de experimentos de Hibridização Subtrativa Supressiva (SSH) com amostras de RNA extraídas de animais aquáticos em experimentos de exposição *in situ* e *ex situ* à diferentes xenobióticos (TOLEDO-SILVA, 2009; MATTOS, 2010; MOSER, 2011; LÜCHMANN, 2012; PIAZZA, 2012). Os experimentos estão listados na Tabela 1.

Tabela 1 – Lista dos experimentos de SSH realizados no LABCAI, desde 2008, utilizados nos testes realizados e na criação do banco de dados SQL.

Espécie	Xenobiótico	Tempo de exposição (h)	Tecido
Poecilia vivipara	Esgoto sanitário	24	Fígado
Poecilia vivipara	Atrazina	24	Fígado

Espécie	Xenobiótico	Tempo de exposição (h)	Tecido
Poecilia vivipara	Cobre	24	Fígado
Poecilia vivipara	Fenantreno	24	Fígado
Poecilia vivipara	Óleo diesel	24	Fígado
Crassostrea gigas	Esgoto sanitário	24	Glândula digestiva
Crassostrea gigas	Esgoto sanitário	24	Brânquias
Crassostrea brasiliana	Óleo diesel	24	Glândula digestiva
Prochilodus lineatus	Atrazina	24	Fígado
Prochilodus lineatus	Cobre	24	Fígado
Prochilodus lineatus	Fenantreno	24	Fígado
Litopenaeus vannamei	Permetrina	96	Brânquias

3.2 Programas e bancos de dados utilizados

3.2.1 Programas de bioinformática

Diversos programas de bioinformática foram utilizados no presente trabalho (Tabela 2). Seis programas de montagem foram testados, assim como três programas referentes ao tratamento de sequências brutas e outros três para análises de similaridade.

Tabela 2 – Lista dos programas de bioinformática utilizados neste trabalho.

Programa	Versão	Função	Referência
Phred	0.071220.c	Nomeador de base e qualidade	EWING et al, 1998; EWING e GREEN, 1998
Crossmatch	1.090518	Retirada de vetores	EWING et al, 1998; EWING e GREEN, 1998
phd2fasta	0.130911	Conversão do formato phd para fasta	EWING et al, 1998; EWING e GREEN, 1998
CAP3	12/21/07	Montagem	HUANG; MADAN, 1999
Phrap[1]	1.090518	Montagem	EWING et al, 1998; EWING e GREEN, 1998

Programa	Versão	Função	Referência
MIRA[1]	4.0.2	Montagem	CHEVREUX; WETTER; SUHAI, 1999
iAssembler[1]	1.3.2	Montagem	ZHENG et al., 2011
Trans-AbySS[1]	1.5.2	Montagem	ROBERTSON et al., 2010
Velvet[1]	1.2.09	Montagem	ZERBINO; BIRNEY, 2008
BLAST+	2.2.26	Busca por similaridade	CAHAMACHO et al., 2009
InterProScan 5	5.8-49.0	Múltiplas análises de similaridade	ZDOBNOV; APWEILER, 2001
HMMER 3[2]	3.1b1	Busca por similaridade	EDDY, 2010

[1] Programas somente utilizados nos testes.

[2] Programas utilizados nas análises do InterProScan 5.

3.2.2 Bancos de informações biológicas

Onze bancos de dados foram utilizados para realização da anotação das sequências dos experimentos. Nove deles tem como foco sequências proteicas, e os restantes tratam de vias metabólicas e terminologia (Tabela 3).

Tabela 3 – Lista dos bancos de dados utilizados na anotação das sequências.

Nome	Escopo	Referência
NR	Conjunto de bancos de proteínas	http://www.ncbi.nlm.nih.gov/protein
Swiss-prot	Proteínas manualmente anotadas	UNIPROT CONSORTIUM, 2008
TIGRFAM[1]	Famílias de proteínas	HAFT; SELENGUT; WHITE, 2003
ProDom[1]	Famílias de proteínas geradas automaticamente do UniProtKB	SERVANT et al., 2002
Pfam[1]	Famílias de proteínas	FINN et al., 2014

Nome	Escopo	Referência
	representadas em alinhamentos múltiplos	
Prosite[1]	Documentação de domínios de proteínas, famílias e sítios funcionais	FALQUET et al., 2002
SMART[1]	Domínios de proteínas	LETUNIC et al., 2004
PRINTS[1]	Compêndio de *fingerprints* de proteínas	ATTWOOD et al., 1994
SuperFamily[1]	Conjunto de anotações estruturais e funcionais para proteínas	MADERA et al., 2004
Gene Ontology	Padronização do vocabulário de anotação	ASHBURNER et al., 2000
Reactome	Vias metabólicas	JOSHI-TOPE et al., 2005

[1] Bancos utilizados nas análises do InterProScan 5.

3.3 Montagem e anotação das sequências

3.3.1 Nomeação de base, *trimming*, retirada de vetores e conversão para fasta

Através do programa Phred, os *reads* e os seus respectivos valores de qualidade foram gerados a partir dos cromatogramas. Além disso, também foi realizado o *trimming* dos *reads* com valor de Phred mínimo de 20. Os arquivos em formato phd foram convertidos para o formato fasta pelo programa phd2fasta. Os vetores utilizados nos procedimentos de clonagem foram mascarados através do programa cross_match.

3.3.2 Avaliação de programas de montagem

As sequências de todos os experimentos foram submetidas à montagem através de seis programas (Phrap, CAP3, MIRA, iAssembler, Trans-AbySS e Velvet). As montagens foram avaliadas

através de *scripts*[1] quanto ao número *contigs* e *reads* não montados (*singlets*), tamanho N50[2] dos *contigs* e qualidade média dos *contigs*.

3.3.3 Anotação

As sequências foram alinhadas através do programa BLASTx utilizando os bancos de dados NR e Swiss-prot e a similaridade foi presumida para os alinhamentos com valor de *e-value* menor que 10^{-5}. A partir dos resultados do BLASTx, a cobertura do *subject*[3] que alinhou e da sequência em DNA do *query*[4] que alinhou foram obtidas através de um *script*. As sequências também foram alinhadas utilizando dados dos bancos TIGFAM, ProDom, Pfam, Prosite, SMART, PRINTS e SuperFamily

[1] Conjunto de instruções para serem realizadas automaticamente pelo computador.

[2] É uma estatística similar a média que mede tamanho de um conjunto de sequências.

[3] Sequência pertencente ao banco de dados, em análises do programa BLAST.

[4] Sequência enviada para análise através do programa BLAST.

através do programa InterProScan 5 (ZDOBNOV; APWEILER, 2001). A partir dos resultados das buscas no Swiss-prot, as respectivas anotações GO foram obtidas, sendo o mesmo realizado para os termos do banco Reactome.

Um *script* foi desenvolvido para a análise do termos GO Slim[5] de cada experimento. Através dele os termos foram obtidos, contabilizados e plotados automaticamente nas três diferentes categorias GO: função molecular, componente celular e processo biológico. Um *script* semelhante também foi desenvolvido para os termos Reactome.

Ao final de cada análise, o *script* desenvolvido durante este projeto gera um relatório contendo os principais resultados das análises, como estatísticas de montagem e anotação, histogramas, tabelas e gráficos referentes aos termos GO e Reactome

[5] Um conjunto de termos GO gerais, utilizados para se obter uma visão geral dos termos de um conjunto de genes.

3.4 MySQL

Um esquema de tabelas para o banco de dados MySQL foi desenvolvido buscando uma estrutura intuitiva para pesquisadores sem conhecimento de SQL. Esforços foram feitos para proteger a integridade dos dados, evitando a duplicação e inconsistências, através da normalização dos dados.

A automatização dos processos de montagem, anotação e depósito das informações no banco de dados MySQL foi feita através de *scripts*, escritos na linguagem de programação Python e estão listados na Tabela 4. Os *scripts* foram desenvolvidos para fácil utilização de qualquer pessoa com mínima experiência em linha de comando.

Como as buscas em um banco de dados MySQL requerem um código que poucos pesquisadores estão familiarizados, um *script* foi desenvolvido para a realização dos principais tipos de buscas, de forma simples, gerando um arquivo

HTML de fácil visualização. A Figura 9 esquematiza o processo de análise.

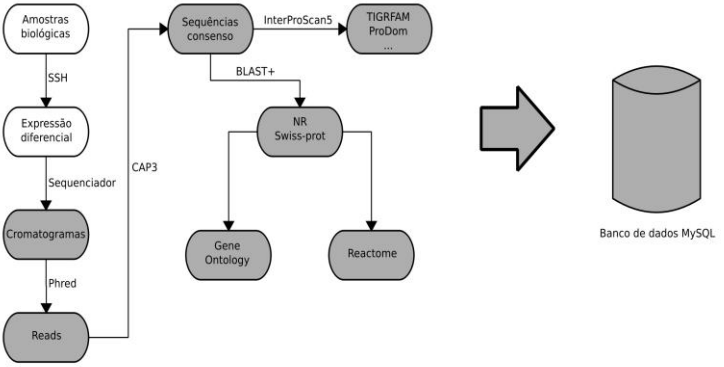


Figura 9 – Diagrama de trabalho para as análises dos sequenciamentos realizados no LABCAI, UFSC. As etapas em cinza estão contidas no sistema de análise.

3.5 Comparação das respostas biológicas dos peixes *Prochilodus lineatus* e *Poecilia vivipara* expostos a atrazina, cobre e fenantreno

Os experimentos utilizando as espécies de peixes *Prochilodus linetaus* e *Poecilia vivipara*, expostos a atrazina, cobre e fenantreno, foram comparados através dos resultados das análises

feitas pelo sistema desenvolvido. Além disso, as proteínas homólogas alinhadas no banco nr foram classificadas quanto ao seu processo biológico através de informações encontradas na literatura.

4. RESULTADOS E DISCUSSÃO

4.1 Escolha do programa de montagem

A partir das sequências montadas pelos seis programas testados, utilizando os 12 experimentos listados na Tabela 1, quatro gráficos foram montados. Para o número médio de sequências montadas (Figura 10), o Phrap obteve o melhor desempenho, gerando em média 54,79 sequências, enquanto que o MIRA teve o pior desempenho, com 23,38 sequências em média.

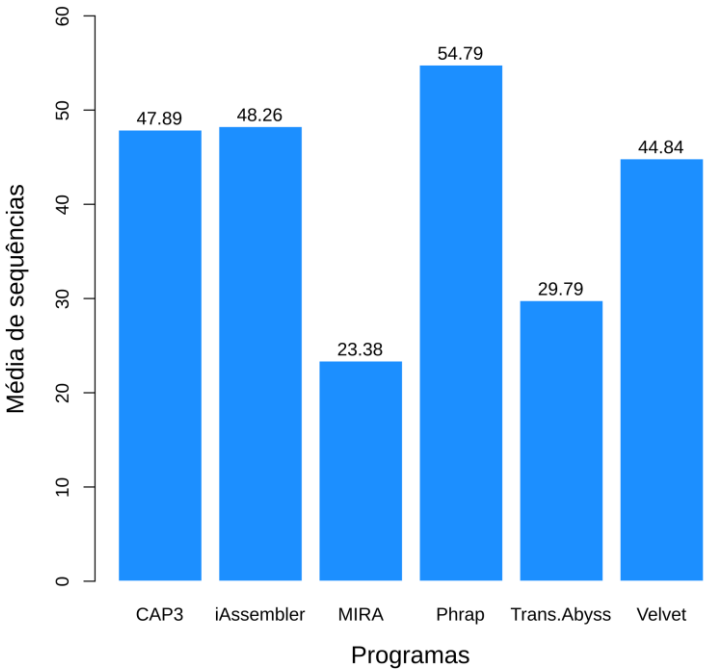


Figura 10 – Número médio de sequências obtidas através de seis programas de montagem.

Com relação ao tamanho das sequências montadas (Figura 11), avaliadas pelo tamanho do contig N50, o programas CAP3, iAssembler, MIRA e Phrap tiveram desempenhos bastante similares, com tamanho de contig N50 médio de

aproximadamente 450 nucleotídeos. O programa Velvet teve um desempenho bastante insatisfatório neste teste.

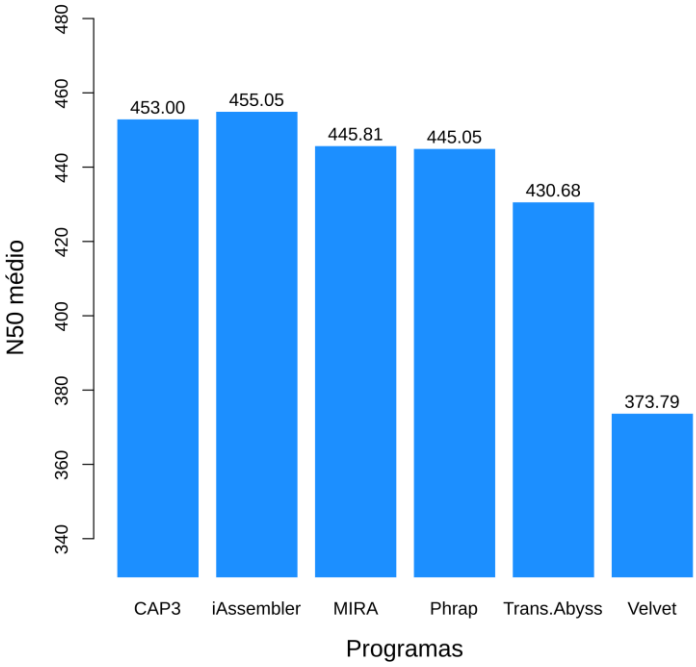


Figura 11 – N50 médio de sequências obtidas através de seis programas de montagem.

Visando avaliar a qualidade do ponto de vista biológico das montagens, o número de *hits* únicos

no banco de dados Swiss-prot foi calculado (Figura 12). Nesta análise o programa Phrap teve o pior desempenho, com somente 36,80% de suas sequências alinhando com um gene único.

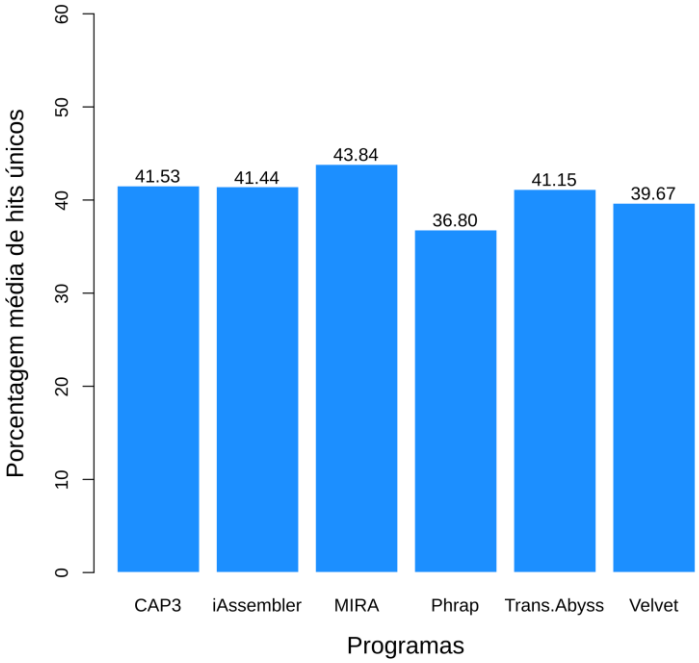


Figura 12 – Porcentagem média de hits únicos no banco de dados Swiss-prot para as montagens de seis programas de montagem.

Por fim, a qualidade média dos *contigs* foi calculada para os programas que geram o arquivo de qualidade (Figura 13). Neste teste o programa CAP3 se saiu consideravelmente melhor que os demais (MIRA e Phrap).

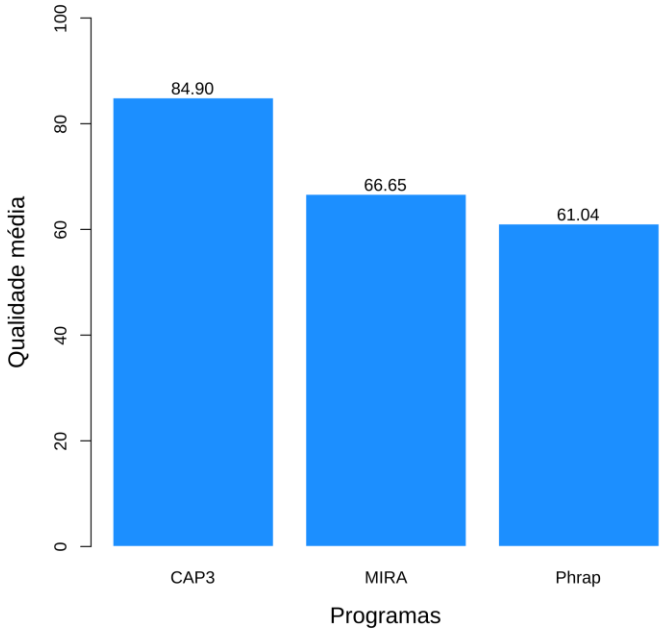


Figura 13 – Qualidade média dos contigs obtidos através de três programas de montagem. Os programas iAssembler, Trans-ABySS e Velvet não geram um arquivo contendo os valores de

qualidade das sequências e por isso não foram incluídos nesta análise.

Os programas CAP3, iAssembler, MIRA e Phrap utilizam algoritmos do tipo sobreposição-consenso, em que as sequências são geradas a partir da sobreposição dos *reads*. Esse tipo de algoritmo é bastante utilizado na montagem de sequências Sanger. Os programas Trans-ABySS e Velvet utilizam algoritmos baseados em grafos de Brujin[6], que é menos custoso computacionalmente, mas necessário para análises de sequenciamentos de alto rendimento. As análises aqui feitas mostram que, para *reads* Sanger, os algoritmos de sobreposição-consenso geram resultados melhores.

O programa iAssember utiliza um sistema de análise interativo, usando os programas MIRA, CAP3 e mega-blast, além de três níveis de correção de erros de montagem (ZHENG et al.,

[6] Representam a sobreposição entre sequências.

2011). Por isso, as sequências geradas pelo iAssembler possuem menos erros de montagem (ZHENG et al., 2011).

A baixa média de sequências geradas pelo programa MIRA é resultado do fato deste programa não gerar *singlets*[7], ao contrário de programas como CAP3 e Phrap. A superioridade na qualidade média das sequências geradas pelo CAP3 é consequência da remoção de regiões de baixa qualidade nas extremidades 3' e 5' realizadas pelo programa (HUANG; MADAN, 1999).

Lang e colaboradores, em 2000, realizaram uma comparação entre os programas CAP3 e Phrap e concluíram que CAP3 contém montagens mais fiéis, pois o Phrap tem baixa sensibilidade na montagem e acaba sacrificando a fidelidade das sequências consenso. Por fim, os programas CAP3, iAssembler e Phrap realizaram montagens semelhantes, com a maior diferença sendo a

[7] *Reads* que não tiveram similaridade com nenhum outro e que, portanto, não foram montados em *contigs*.

qualidade média significativamente maior para o CAP3.

4.2 *Scripts* desenvolvidos para análise

Visando automatizar e padronizar análises posteriores no LABCAI, foram desenvolvidos vários *scripts* em Python (Tabela 4).

Tabela 4 – Lista dos scripts desenvolvidos neste trabalho.

Nome	Função
blast-extra.py	Calcula variáveis extras para o arquivo de saída do BLAST+
uniprot2go.py	Busca anotações Gene Ontology para termos de acesso UniProtKB
uniprot2reactome.py	Busca anotações Reactome para termos de acesso UniProtKB
go.py	Obtêm, contabiliza e produz um gráfico com os termos GO Slim mais comuns para determinado experimento
reactome.py	Contabiliza e produz uma tabela com os termos Reactome mais comuns para determinado experimento

Nome	Função
report.py	Produz um relatório em HTML contendo vários resultados da análise em questão
sanger.bash	Executa todas as etapas de análise automaticamente
sanger2sql.py	Importa os resultados das análises para o banco MySQL
search-sql.py	Realiza buscas pré-definidas no banco de dados e produz uma tabela HTML com os resultados

O *script* blast-extra é utilizado para calcular duas variáveis extras no arquivo de saída do BLASTx: a porcentagem da proteína alvo que foi alinhada e a sequência em DNA enviada que alinhou. Os *scripts* uniprot2go e uniprot2reactome servem para buscar anotações dos respectivos bancos a partir de códigos de acesso do Swiss-prot.

O *script* go contabiliza os termos GO das análises, gerando três gráficos com os termos mais comuns para as categorias de termos GO: função

molecular, componente celular e processo biológico (Figura 14)

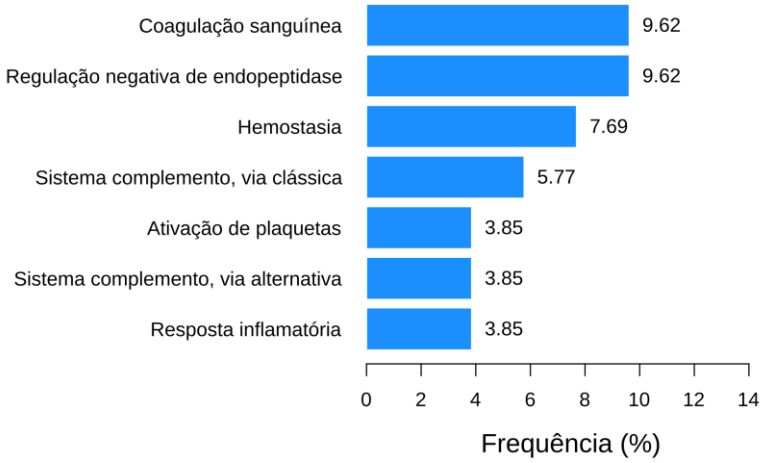


Figura 14 – Exemplo de gráfico para os termos GO da categoria processo biológico, gerado pelo script goslim.

O *script* reactome produz uma tabela HTML com as vias metabólicas Reactome mais comuns nos resultados das análises Tabela 5.

Tabela 5 – Exemplo de tabela para os termos gerado pelo script reactome.

Nome	Termo	Frequência (%)
Via comum de coagulação	REACT_327369	10,53
Via intrínseca de coagulação	REACT_289856	10,53
Degranulação de plaquetas	REACT_318	10,53

O *script* em bash chamado sanger, por sua vez, é responsável por automatizar todo o processo de análise em um só comando, enquanto que o *script* report é produz um relatório sintetizando todos os resultados da análise Figura 15.

Prochilodus lineatus - Atrazina

Dados do experimento:

Espécie: *Prochilodus lineatus*
Contaminante: Atrazina
Tempo de exposição: 24h
Tecido: Fígado

Estatísticas da análise

Contigs e singlets: 66
N50: 427
Hits NR: 35
Hits InterProScan5: 225
Termos GO: 249
Termos Reactome: 47

Resultados NR:

Codigo de acesso	Descricao	E-value	Porcentagem da proteina alinhada	Especie	Contaminante	Biblioteca	
AGO58845.1	fibrinogen alpha chain, partial [Carassius auratus]	2e-28	49.55	Prochilodus lineatus	Atrazina	forward	ACAATA
XP_007238158.1	PREDICTED: complement C3-like [Astyanax mexicanus]	1e-63	35.37	Prochilodus lineatus	Atrazina	forward	ACTATG

Figura 15 – Exemplo de relatório com os resultados da análise produzido pelo script report.

Para inserir os resultados no banco de dados MySQL foi criado o *script* sanger2sql. Por último, o *script* search-sql realiza buscas pré-determinadas no banco, onde o usuário só precisa inserir alguma palavra-chave, como o nome de um gene, ou um termo GO de interesse.

Vários sistemas de análise para bioinformática já foram desenvolvidos (OVERBEEK et al., 2000; MARKOWITZ et al., 2008; DE OLIVEIRA et al., 2005; WAGNER et al., 2014; D'ANTONIO et al., 2013). O objetivo destes sistemas é facilitar o armazenamento, análise e apresentação dos resultados das análises que propõem. Aqui, todo o processo de análise foi automatizado em um único *script* chamado sanger. Além disso, pode-se gerar um relatório em formato HTML contendo todos os resultados das análises. Todos os *scripts* foram desenvolvidos de forma simples e são de uso fácil por pessoal com pouca experiência na linha de comando.

4.3 Banco de dados MySQL

A estrutura para o banco de dados MySQL desenvolvida contém seis tabelas, sendo capaz de armazenar várias informações sobre as diversas análises contidas no sistema (Figura 16). A tabela

completa, contendo os nomes de todas as colunas, pode ser encontrada no Apêndice A.

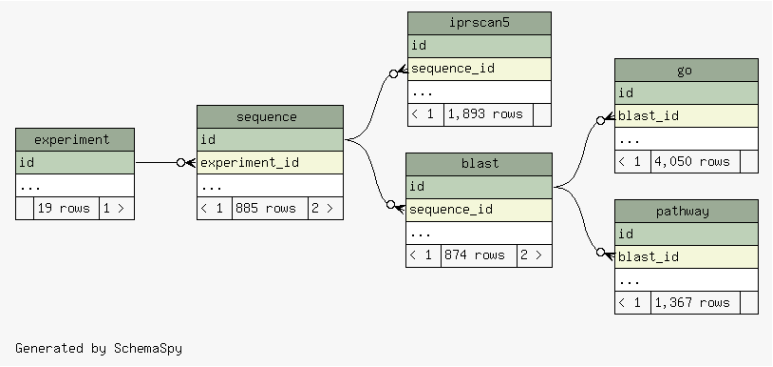


Figura 16 – Versão simplificada do esquema MySQL para o banco de dados. A versão completa contendo todas as informações pode ser encontrada no Apêndice A.

O esquema de tabelas foi desenvolvido com o intuito de separar os resultados de forma intuitiva para pessoas não familiarizadas com a ciência da computação. As seis tabelas contém dados sobre as análises em ordem cronológica, começando com as informações dos experimentos, passando às sequências e então às anotações. A nomenclatura das colunas foi feita obedecendo os nomes já

definidos pelos programas, facilitando a busca. Os tipos de dados para as mais variadas informações foram escolhidas buscando boa precisão e uso mínimo de espaço de armazenamento.

Uma das etapas mais importantes no desenvolvimento de bancos de dados relacionais é a normalização (NELSON; REISINGER; HENRY, 2003). No presente esquema, esforços foram feitos para diminuir ao máximo a redundância de dados.

4.4 Comparação das respostas biológicas dos peixes *Prochilodus lineatus* e *Poecilia vivipara* expostos a atrazina, cobre e fenantreno

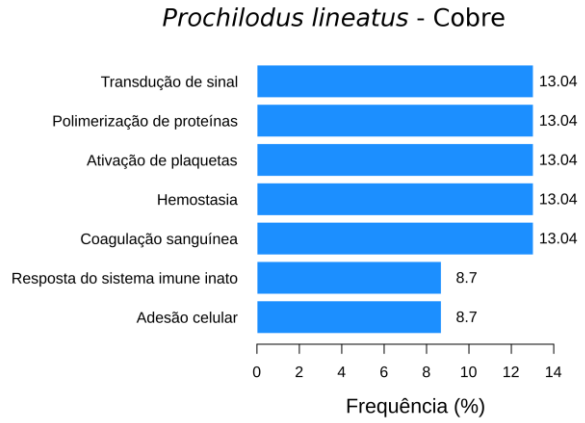
Através dos sistema de análise, foram montadas sequências de seis experimentos SSH, dos peixes *Prochilodus lineatus* e *Poecilia vivipara* expostos a atrazina, cobre e fenantreno (Tabela 6).

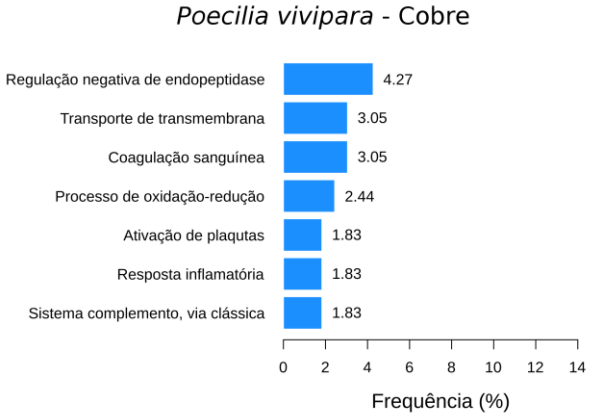
Tabela 6 – Estatísticas dos resultados das análises dos experimentos SSHs para *Prochilodus lineatus* e *Poecilia vivipara* expostos a atrazina, cobre e fenantreno.

Experimento	Contigs	N50 (pb)	*Hits* nr	Termos GO	Termos Reactome
P. lineatus – Atrazina	66	427	35	249	47
P. lineatus – Cobre	30	370	15	65	37
P. lineatus – Fenantreno	67	406	28	283	98
P. vivipara – Atrazina	77	453	62	533	180
P. vivipara – Cobre	49	471	40	483	220
P. vivipara – Fenantreno	50	466	34	285	47

Todas as sequências anotadas no banco de proteínas nr foram classificadas segundo seus respectivos processos biológicos, através de informações encontradas na literatura científica. Os Apêndice B-G apresentam estas sequências anotadas separadas pelo processo biológico a qual pertencem, nome da proteína, código de acesso do *GenkBank*, *e-value* e a porcentagem da proteína alvo que foi alinhada. Os resultados das análises dos processos biológicos GO e termos Reactome

estão apresentados na Figura 17 e Tabela 7, respectivamente.


Prochilodus lineatus - Cobre
Transdução de sinal
Polimerização de proteínas
Ativação de plaquetas
Hemostasia
Coagulação sanguínea
Resposta do sistema imune inato
Adesão celular
13.04
13.04
13.04
13.04
13.04
8.7
8.7
0 2 4 6 8 10 12 14
Frequência (%)


Poecilia vivipara - Cobre
Regulação negativa de endopeptidase
Transporte de transmembrana
Coagulação sanguínea
Processo de oxidação-redução
Ativação de plaqutas
Resposta inflamatória
Sistema complemento, via clássica
4.27
3.05
3.05
2.44
1.83
1.83
1.83
0 2 4 6 8 10 12 14
Frequência (%)

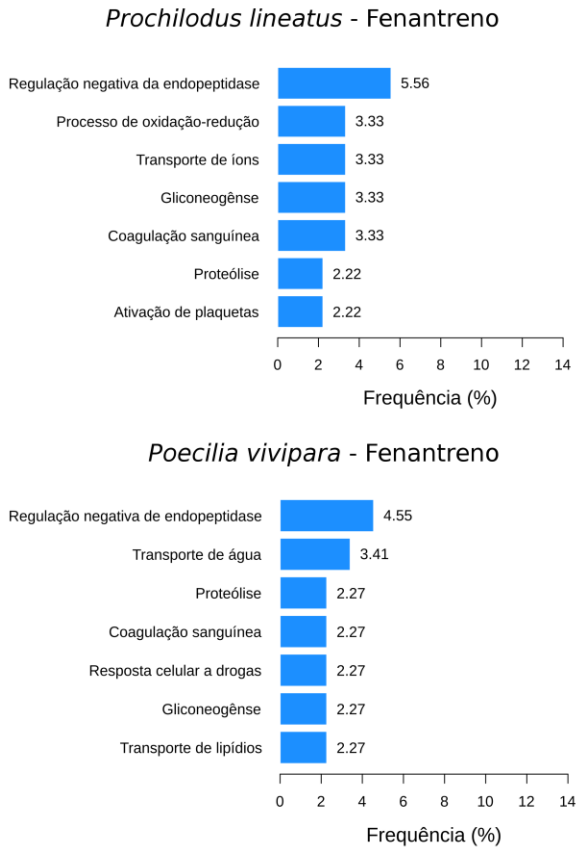


Figura 17 – Termos GO mais frequentes para a categoria processo biológico das SSH de *Prochilodus linetaus* e *Poecilia vivipara* expostos a atrazina, cobre e fenantreno.

Tabela 7 – Termos Reactome mais frequentes das SSH de Prochilodus linetaus *e* Poecilia vivipara *expostos a atrazina, cobre e fenantreno.*

Nome	Termo	Frequência (%)	Nome	Termo	Frequência (%)
Prochilodus lineatus – Atrazina			*Poecilia vivipara* - Atrazina		
Via intrínseca	REACT_289856	11,76	Via intrínseca	REACT_326	6,12
Degranulação de plaquetas	REACT_318	11,76	Via comum	REACT_1439	6,12
Via comum	REACT_327369	11,76	Gliconeogênese	REACT_286441	4,08
Transportadores de íons de metal SLC	REACT_20547	5,88	Ativação de C3 e C5	REACT_7972	4,08
Carboxilação de precursores proteicos	REACT_351465	5,88	Ativação inicial do complemento	REACT_8024	4,08
Prochilodus lineatus – Cobre			*Poecilia vivipara* - Cobre		
Degranulação de plaquetas	REACT_307071	18,75	Degranulação de plaquetas	REACT_307071	7,94
Interações da integrina	REACT_319261	12,50	Via comum	REACT_286713	4,76
Sinalização alfa IIb beta 3 da integrina	REACT_330010	12,50	Interações da integrina	REACT_319261	4,76
Via comum	REACT_286713	12,50	Sinalização MAPK para integrinas	REACT_332492	4.76
Sinalização MAPK para integrinas	REACT_332492	12,50	Sinalização alfaIIb beta3 de integrinas	REACT_330010	4,76
Prochilodus lineatus – Fenantreno			*Poecilia vivipara* - Fenantreno		
Via comum	REACT_327369	7,69	Contração de músculo	REACT_324616	16,67

Nome	Termo	Frequência (%)	Nome	Termo	Frequência (%)
Prochilodus lineatus – Atrazina			*Poecilia vivipara* - Atrazina		
			estriado		
Interações da integrina	REACT_13552	7,69	Biogênese de vesículas lisossomais	REACT_308445	5,56
Gliconeogênese	REACT_115534	3,85	Gliconeogênese	REACT_305537	5,56
Associação de TriC/CCT com proteínas	REACT_326227	3,85	Transporte por aquaporinas	REACT_23826	5,56
Sinalização alfaIIb beta3 de integrinas	REACT_15523	3,85	Contração de músculo liso	REACT_20558	5,56

Peixes são o grupo mais diversos de vertebrados, com quase 25.000 diferentes espécies ocupando a maioria dos nichos aquáticos. Assim, o monitoramento da saúde de peixes oferece uma boa compreensão da condição do ambiente aquático (ZELIKOFF et al., 2000). Esta comparação tem como objetivo encontrar algum padrão de resposta entre os experimentos.

Primeiramente, é clara a semelhança das respostas entre as duas espécies de peixes. Isso talvez se deve ao fato de ambos serem

relativamente próximos evolutivamente, já que ambos pertencem à classe Actinopterygii. Nos experimentos de exposição a atrazina e cobre, tanto na classificação putativa (Apêndices B-G), nos termos GO mais comuns para processos biológicos (Figura 17) e termos Reactome (Tabela 7) é clara a presença de genes envolvidos em processos de imunidade inata, coagulação sanguínea e inflamação, como os fibrinogênios e proteínas do sistema complemento. Proteínas com capacidade de inibir endopeptidases também são frequentes. Já para os experimentos envolvendo fenantreno, apesar da coagulação estar presente, a imunidade inata não é tão observada (Figura 17).

Vários estudos em seres humanos demostram a ativação do sistema complemento por toxinas e poluentes (PATON; ROWAN-KELLY; FERRANTE, 1984; ORTH et al., 2009; HEIDEMAN, 1979; HULANDER et al., 2009). Muitos estudos também revelam que mesmo exposição a doses baixas de xenobióticos podem inibir a imunidade

adaptativa ao longo do tempo e assim aumentar o risco de infecção (ILBÄCK e FRIMAN, 2007).

O sistema complemento é um dos principais representantes da imunidade inata, com mais de 30 componentes, em maioria proteínas plasmáticas, agindo como enzimas ou proteínas de ligação (MÜLLER-EBERHARD, 1988). Estas proteínas interagem de maneira altamente regulada, participando de diferentes funções como eliminação de antígenos, lise de células estranhas, fagocitose, quimiotaxia e mediação de respostas inflamatória (MÜLLER-EBERHARD, 1988).

Uma interação entre os sistemas de coagulação e complemento já havia sido proposta, pois ambas descendem do mesmo sistema ancestral, e em 2008 a via molecular desta interação foi descrita (AMARA et al., 2008). Uma inflamação sistêmica gera uma ativação maciça da cascata de coagulação, devido a geração de trombina, repressão de mecanismos anticoagulantes e inibição da fibrinólise (LEVI et al., 2003; AMARA et al., 2008), como também uma

hiperativação do complemento, liberando anafilatoxinas, promovendo inflamação aguda, ativando quimiotaxia e ativação de mastócitos. Proteínas envolvidas na fase aguda da inflamação como o fibrinogênio estão bastante presentes nas respostas observadas. Gabay e colaboradores (1999) descobriram que estas proteínas aumentam em até 25% em resposta a dano, infecção ou inflamação. XIE e colaboradores (2009) também afirmam que o aumento na expressão de fibrinogênio é uma evidência de inflamação sistêmica.

As proteínas do sistema complemento também tem expressão induzida com estímulos inflamatórios (KATAGIRI; HINDRA; AOKI, 1999). Em 2008, Mattos e colaboradores observaram indução da expressão da proteína C3 em *Poecilia vivipara* exposta à fração acomodada de óleo diesel.

Componentes da imunidade inata aparentam ser bastante conservados entre espécies (HOFFMANN et al., 1999; ULEVITCH, 2000). Isto

significa que a sensibilidade da imunidade inata a determinado contaminante é similar entre diversas espécies, o que pode tornar monitoramentos de impactos ambientais mais eficientes (BOLS et al., 2001).

4.4.1 Atrazina

Rohr e McCoy (2010) realizam uma revisão da literatura sobre os efeitos da exposição de atrazina em peixes e anfíbios. Eles relataram que 16 dos 18 estudos observados demonstravam uma redução da imunidade adaptativa em peixes e anfíbios. Por outro lado, a exposição a atrazina também aumentava o risco de infecção pode nematódeos, vírus ou bactérias em 12 de 14 estudos.

Outra revisão crítica da literatura envolvendo efeitos da exposição a atrazina foi realizada por Solomon e colaboradores em 2008. Muitos trabalhos que observaram os efeitos da atrazina em mamíferos revelam a supressão da imunidade

adaptativa (FOURNIER et al., 1992; WHALE et al., 2003; FILIPOV et al., 2005; KARROW et al., 2005), diminuição do número de linfócitos (WALSH e RIBELIN, 1975) e aumento na atividade fagocítica de macrófagos (FOURNIER et al., 1992), indicando uma repressão da imunidade adaptativa e aumento da imunidade inata pela fagocitose.

4.4.2 Cobre

Com relação a imunidade, Dethloff e Bailey (1998) relatam redução dos linfócitos circulantes em 15%, células B em 56% e elevação de neutrófilos e monócitos em *Oncorhynchus mykiss*. Dick e Dixon (2000) também observaram redução de linfócitos, mas aumento de trombócitos e neutrófilos. Vários outros estudos também observaram aumento de neutrófilos circulantes em organismos expostos a cobre (WEYTS et al., 1998; ZEEMAN e BRINDLEY, 1981). Assim como na exposição a atrazina, aqui temos uma diminuição

da imunidade adaptativa e aumento da inata pela fagocitose, através dos neutrófilos.

4.4.3 Fenantreno

Peixes coletados de águas poluídas por HPAs possuem lesões em brânquias, pele e nadadeiras causadas por infecções (SEELEY e WEEK-PERKINS, 1991). Uma revisão da literatura sobre os efeitos de HPAs em peixes revela que estes suprimem tanto a imunidade inata quanto a adaptativa em peixes (REYNAUD e DESCHAUX, 2006). Carlson e colaboradores (2002) demonstraram aumento na suscetibilidade de infecção por bactérias pela exposição de HPAs em peixes.

Apesar de termos relacionados a imunidade inata não estarem entre os mais comuns nos experimentos envolvendo fenantreno (Figura 18), vários genes deste sistema foram encontrados (Apêndices D e G). Este resultado é contrário aos resultados anteriores observados na literatura.

5. CONCLUSÕES E PERSPECTIVAS

- Os programas CAP3, iAssembler e Phrap geraram montagens melhores a partir de sequências obtidas através de metodologia de sequenciamento Sanger, com maior número de sequências e tamanho do contig N50. Ainda, as sequências geradas pelo CAP3 têm o diferencial de possuírem maior qualidade.

- O sistema de análise aqui desenvolvido facilita o armazenamento, análise e apresentação de análises de sequenciamentos do tipo Sanger. No futuro, pretende-se também dar suporte aos sequenciamentos de nova geração.

- O esquema de tabelas relacionais para o banco de dados MySQL passou por normalização, evitando redundância nos dados. A nomenclatura das tabelas e colunas é de fácil compreensão a pessoas inexperientes em SQL.

- A exposição dos peixes *Poecilia vivipara* e *Prochilodus lineatus* a atrazina, cobre e fenantreno gerou respostas semelhantes nas duas espécies. Muitos genes envolvidos em processos de imunidade inata, como o sistema complemento e o sistema de coagulação sanguínea foram ativados na exposição a atrazina, cobre e fenantreno, indicando danos celulares causados pelos compostos. A indução da expressão desses genes provavelmente é realizada através do aumento na susceptibilidade a infecções causada pela exposição a xenobióticos. Esse aumento gera estímulos inflamatórios que podem ativar a expressão de proteínas do sistema complemento e de coagulação.

REFERÊNCIAS

AMARA, U. et al. Interaction between the coagulation and complement system. In: **Current topics in complement II**. Springer US, 2008. p. 68-76.

ASHBURNER, M. et al. Gene Ontology: tool for the unification of biology. **Nature genetics**, v. 25, n. 1, p. 25-29, 2000.

ATTWOOD, T. K. et al. PRINTS--a database of protein motif fingerprints. **Nucleic acids research**, v. 22, n. 17, p. 3590, 1994.

AUSTEN, M. C.; WARWICK, R. M.; CARMEN ROSADO, M. Meiobenthic and macrobenthic community structure along a putative pollution gradient in southern Portugal. **Marine Pollution Bulletin**, v. 20, n. 8, p. 398-405, 1989.

BERNSTEIN, F. C. et al. The protein data bank. **European Journal of Biochemistry**, v. 80, n. 2, p. 319-324, 1977.

BOLS, N. C. et al. Ecotoxicology and innate immunity in fish. **Developmental & Comparative Immunology**, v. 25, n. 8, p. 853-873, 2001.

BRULLE, F. et al. Identification and expression profile of gene transcripts differentially expressed during metallic exposure in Eisenia fetida coelomocytes. **Developmental & Comparative Immunology**, v. 32, n. 12, p. 1441-1453, 2008.

CAJARAVILLE, M. P. et al. The use of biomarkers to assess the impact of pollution in coastal environments of the Iberian Peninsula: a practical approach. **Science of the Total Environment**, v. 247, n. 2, p. 295-311, 2000.

ALTSCHUL, S F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. **Nucleic acids research**, v. 25, n. 17, p. 3389-3402, 1997.

CARLSON, E. A.; LI, Y.; ZELIKOFF, J. T. Exposure of Japanese medaka (Oryzias latipes) to benzo [a] pyrene suppresses immune function and host resistance against bacterial challenge**. Aquatic Toxicology**, v. 56, n. 4, p. 289-301, 2002.

CHEVREUX, B.; WETTER, T.; SUHAI, S. Genome sequence assembly using trace signals and additional sequence information. In: **German Conference on Bioinformatics**. 1999. p. 45-56.

CLEARWATER, S. J.; FARAG, A. M.; MEYER, J. S. Bioavailability and toxicity of dietborne copper and zinc to fish. **Comparative Biochemistry and Physiology Part C: Toxicology & Pharmacology**, v. 132, n. 3, p. 269-313, 2002.

COLLINS, F. S. et al. A vision for the future of genomics research. **Nature**, v. 422, n. 6934, p. 835-847, 2003.

COOKE, M.; DENNIS, A. J. PAH-X: Polynuclear aromatic hydrocarbons: A decade of progress. 1988.

COOPER, A. JL; PLUM, F. R. E. D. Biochemistry and physiology of brain ammonia. **Physiol Rev**, v. 67, n. 2, p. 440-519, 1987.

COORAY, M. P. N. S. Molecular biological databases: evolutionary history, data modeling, implementation and ethical background. **Sri Lanka Journal of Bio-Medical Informatics**, v. 3, n. 1, p. 2-11, 2012.

CRICK, F. et al. Central dogma of molecular biology. **Nature**, v. 227, n. 5258, p. 561-563, 1970.

DE GROOT, R. et al. (2012). Global estimates of the value of ecosystems and their services in monetary units. *Ecosystem Services*, *1*(1), 50-61. Global estimates of the value of ecosystems and their services in monetary units. **Ecosystem Services**, v. 1, n. 1, p. 50-61, 2012.

DE OLIVEIRA, T. et al. An automated genotyping system for analysis of HIV-1 and other microbial sequences. **Bioinformatics**, v. 21, n. 19, p. 3797-3800, 2005.

DETHLOFF, G. M.; BAILEY, H. C. Effects of copper on immune system parameters of rainbow trout (Oncorhynchus mykiss). **Environmental toxicology and chemistry**, v. 17, n. 9, p. 1807-1814, 1998.

DICK, P. T.; DIXON, D. G. Changes in circulating blood cell levels of rainbow trout, Salmo gairdneri Richardson, following acute and chronic exposure to copper. **Journal of Fish Biology**, v. 26, n. 4, p. 475-481, 1985.

D'ANTONIO, Mattia et al. WEP: a high-performance analysis pipeline for whole-exome data. **BMC bioinformatics**, v. 14, n. Suppl 7, p. S11, 2013.

EDDY, S. HMMER3: a new generation of sequence homology search software. **URL: http://hmmer. janelia. Org**, 2010.

EDGAR, R. C.; BATZOGLOU, S. Multiple sequence alignment. **Current opinion in structural biology**, v. 16, n. 3, p. 368-373, 2006.

EWING, B.; GREEN, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. **Genome research**, v. 8, n. 3, p. 186-194, 1998.

EWING, B. et al. Base-calling of automated Sequencer traces usingPhred. I. Accuracy assessment. **Genome research**, v. 8, n. 3, p. 175-185, 1998.

FALQUET, L. et al. The PROSITE database, its status in 2002.**Nucleic acids research**, v. 30, n. 1, p. 235-238, 2002.

FINN, R. D. et al. Pfam: the protein families database. **Nucleic acids research**, v. 42, n. D1, p. D222-D230, 2014.

FLEEGER, J. W.; CARMAN, K. R.; NISBET, R. M. Indirect effects of contaminants in aquatic ecosystems. **Science of the Total Environment**, v. 317, n. 1, p. 207-233, 2003.

FOURNIER, M. et al. Limited immunotoxic potential of technical formulation of the herbicide atrazine (AAtrex) in mice. **Toxicology letters**, v. 60, n. 3, p. 263-274, 1992.

FOWLER, B. A. Molecular biomarkers: Challenges and prospects for the future. **Toxicology and Applied Pharmacology**, v. 206, n. 2, p. 97, 2005.

GABAY, C.; KUSHNER, I. Acute-phase proteins and other systemic responses to inflammation. **New England journal of medicine**, v. 340, n. 6, p. 448-454, 1999.

GARANTZIOTIS, S. et al. Inter-α-trypsin inhibitor attenuates complement activation and complement-

induced lung injury. **The Journal of Immunology**, v. 179, n. 6, p. 4187-4192, 2007.

GILBERTSON, M. et al. Immunosuppression in the northern leopard frog (Rana pipiens) induced by pesticide exposure. **Environmental Toxicology and Chemistry**, v. 22, n. 1, p. 101-110, 2003.

HAFT, D. H.; SELENGUT, J. D.; WHITE, O. The TIGRFAMs database of protein families. **Nucleic acids research**, v. 31, n. 1, p. 371-373, 2003.

HALPERN, B. S. et al. A global map of human impact on marine ecosystems. **Science**, v. 319, n. 5865, p. 948-952, 2008.

HANDY, R. D.; DEPLEDGE, M. H. Physiological responses: their measurement and use as environmental biomarkers in ecotoxicology. **Ecotoxicology**, v. 8, n. 5, p. 329- 349, 1999.

HARDIN, J. A.; HINOSHITA, F.; SHERR, D. H. Mechanisms by which benzo a pyrene, an environmental carcinogen, suppresses B cell lymphopoiesis. **Toxicology and applied pharmacology**, v. 117, n. 2, p. 155-164, 1992.

HEIDEMAN, M. Complement activation in vitro induced by endotoxin and injured tissue. **Journal of Surgical Research**, v. 26, n. 6, p. 670-673, 1979.

HOFFMANN, J. A. et al. Phylogenetic perspectives in innate immunity. **Science**, v. 284, n. 5418, p. 1313-1318, 1999.

HUANG, X.; MADAN, A. CAP3: A DNA sequence assembly program. **Genome research**, v. 9, n. 9, p. 868-877, 1999.

HULANDER, M. et al. Blood interactions with noble metals: coagulation and immune complement activation. **ACS applied materials & interfaces**, v. 1, n. 5, p. 1053-1062, 2009.

IANNELLI, R. et al. Assessment of pollution impact on biological activity and structure of seabed bacterial communities in the Port of Livorno (Italy).**Science of the Total Environment**, v. 426, p. 56-64, 2012.

ILBÄCK, N.; FRIMAN, G. Interactions among infections, nutrients and xenobiotics. **Critical**

reviews in food science and nutrition, v. 47, n. 5, p. 499-519, 2007.

ISLAM, S.; TANAKA, M. Impacts of pollution on coastal and marine ecosystems including coastal and marine fisheries and approach for management: a review and synthesis. **Marine pollution bulletin**, v. 48, n. 7, p. 624-649, 2004.

JOSHI-TOPE, G. et al. Reactome: a knowledgebase of biological pathways. **Nucleic acids research**, v. 33, n. suppl 1, p. D428-D432, 2005.

KATAGIRI, T.; HIRONO, I.; AOKI, T. Molecular analysis of complement component C8β and C9 cDNAs of Japanese flounder, Paralichthys olivaceus. **Immunogenetics**, v. 50, n. 1-2, p. 43-48, 1999.

KIELY, T.; DONALDSON, D.; GRUBE, A. Pesticides industry sales and usage. **Washington, DC: Office of Prevention, Pesticides and Toxic Substances, United States Environment Protection Agency**, p. 16, 2004.

KHATRI, P.; SIROTA, M.; BUTTE, A. J. Ten years of pathway analysis: current approaches and

outstanding challenges. **PLoS computational biology**, v. 8, n. 2, p. e1002375, 2012.

LEVI, M. et al. Infection and inflammation and the coagulation system. **Cardiovascular research**, v. 60, n. 1, p. 26-39, 2003.

LI, Z. et al. Comparison of the two major classes of assembly algorithms: overlap–layout–consensus and de-bruijn-graph. **Briefings in functional genomics**, v. 11, n. 1, p. 25-37, 2012.

LIANG, F. et al. An optimized protocol for analysis of EST sequences. **Nucleic acids research**, v. 28, n. 18, p. 3657-3665, 2000.

LETUNIC, I. et al. SMART 4.0: towards genomic data integration. **Nucleic Acids Research**, v. 32, n. suppl 1, p. D142-D144, 2004.

LÜCHMANN, K. H. **Expressão diferencial de genes em ostras *Crassostrea brasiliana* expostas a fração solúvel de óleo diesel**. 2012. Tese (Doutorado em Bioquímica) – Centro de Ciências Biológicas, UFSC, Florianópolis, 2012.

MADERA, M. et al. The SUPERFAMILY database in 2004: additions and improvements. **Nucleic acids research**, v. 32, n. suppl 1, p. D235-D239, 2004.

MARKOWITZ, V. M. et al. IMG/M: a data management and analysis system for metagenomes. **Nucleic acids research**, v. 36, n. suppl 1, p. D534-D538, 2008.

MARX, V. Biology: The big challenges of big data. **Nature**, v. 498, n. 7453, p. 255-260, 2013.

MATTOS, J. J. **Respostas bioquímicas e moleculares no peixe *Poecilia vivipara* exposto à fração de óleo diesel acomodada em água**. 2010. Dissertação (Mestrado em Bioquímica) – Centro de Ciências Biológicas, UFSC, Florianópolis, 2010.

MOORE, M. et al. An integrated biomarker-based strategy for ecotoxicological evaluation of risk in environmental management. **Mutation Research**, v. 552, n. 1-2, p. 247-268, 2004.

MOSER, J. R. **Biomarcadores moleculares no camarão branco, *Litopenaeus vannamei* (CRUSTACEA: DECAPODA), submetido a**

estresse ambiental e infectado pelo vírus da síndrome da mancha branca (*white spot syndrome virus*, WSSV). 2011. Tese (Douturado em Biotecnologia e Biociências) – Centro de Ciências Biológicas, UFSC, Florianópolis, 2011.

MUDZINSKI, S. P. Effects of Benzo a pyrene on Concanavalin A-Stimulated Human Peripheral Blood Mononuclear cells in Vitro: Inhibition of Proliferation but No Effect on Parameters Related to the G 1 Phase of the Cell Cycle. **Toxicology and applied pharmacology**, v. 119, n. 2, p. 166-174, 1993.

MULLER-EBERHARD, H. J. Molecular organization and function of the complement system. **Annual review of biochemistry**, v. 57, n. 1, p. 321-347, 1988.

NELSON, M. R.; REISINGER, S. J.; HENRY, S. G. Designing databases to store biological information. **Biosilico**, v. 1, n. 4, p. 134-142, 2003.

ORTH, D. et al. Shiga toxin activates complement and binds factor H: evidence for an active role of complement in hemolytic uremic syndrome.**The Journal of Immunology**, v. 182, n. 10, p. 6394-6400, 2009.

OVERBEEK, R. et al. WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. **Nucleic acids research**, v. 28, n. 1, p. 123-125, 2000.

PATON, J. C.; ROWAN-KELLY, B.; FERRANTE, A. Activation of human complement by the pneumococcal toxin pneumolysin. **Infection and immunity**, v. 43, n. 3, p. 1085-1087, 1984.

PIAZZA, C. E. **Identificação e análise da transcrição gênica diferencial em peixes *Poecilia vivipara* (Bloch & Scheider, 1801), expostos ao esgoto sanitário**. 2012. Dissertação (Mestrado em Biotecnologia) – Centro de Ciências Biológicas, UFSC, Florianópolis, 2012.

REYNAUD, S.; DESCHAUX, P. The effects of polycyclic aromatic hydrocarbons on the immune system of fish: a review. **Aquatic Toxicology**, v. 77, n. 2, p. 229-238, 2006.

ROBERTSON, G. et al. De novo assembly and analysis of RNA-seq data. **Nature methods**, v. 7, n. 11, p. 909-912, 2010.

ROHR, J. R.; MCCOY, K. A. A qualitative meta-analysis reveals consistent effects of atrazine on freshwater fish and amphibians. **Environmental Health Perspectives**, p. 20-32, 2010.

SANGER, F.; NICKLEN, S.; COULSON, A. R. DNA sequencing with chain-terminating inhibitors. **Proceedings of the National Academy of Sciences**, v. 74, n. 12, p. 5463-5467, 1977.

SCHIRMER, K. et al. Transcriptomics in ecotoxicology. **Analytical and bioanalytical chemistry**, v. 397, n. 3, p. 917-923, 2010.

SEELEY, K. R.; WEEKS-PERKINS, B. A. Suppression of natural cytotoxic cell and macrophage phagocytic function in oyster toadfish exposed to 7, 12-dimethylbenz anthracene. **Fish & Shellfish Immunology**, v. 7, n. 2, p. 115-121, 1997.

SERVANT, F. et al. ProDom: automated clustering of homologous domains. **Briefings in Bioinformatics**, v. 3, n. 3, p. 246-251, 2002.

SHENDURE, J.; JI, H. Next-generation DNA sequencing. **Nature biotechnology**, v. 26, n. 10, p. 1135-1145, 2008.

SMITH, V. H.; TILMAN, G. D.; NEKOLA, J. C. Eutrophication: impacts of excess nutrient inputs on freshwater, marine, and terrestrial ecosystems. **Environmental pollution**, v. 100, n. 1, p. 179-196, 1999.

SOLOMON, K. R. et al. Effects of atrazine on fish, amphibians, and aquatic reptiles: a critical review. **Critical reviews in toxicology**, v. 38, n. 9, p. 721-772, 2008.

TAYLOR, E. W. et al. Lethal and sub-lethal effects of copper upon fish: a role for ammonia toxicity?. In: **SEMINAR SERIES-SOCIETY FOR EXPERIMENTAL BIOLOGY**. Cambridge University Press, 1996. p. 85-114.

TOLEDO-SILVA, G. **Análise da expressão gênica diferencial em ostras-dopacífico *Crassostrea gigas* expostas a esgoto doméstico *in situ*.** 2009. Dissertação (Mestrado em Biotecnologia) – Centro de Ciências Biológicas, UFSC, Florianópolis, 2009.

ULEVITCH, R. J. **Molecular mechanisms of innate immunity. Immunologic research**, v. 21, n. 2-3, p. 49-54, 2000.

UNIPROT CONSORTIUM. The universal protein resource (UniProt). **Nucleic acids research**, v. 36, n. suppl 1, p. D190-D195, 2008.

UNESCO – IOC/UNESCO, IMO, FAO, UNDP. **A Blueprint for Ocean and Coastal Sustainability**. Paris: IOC/UNESCO. 43 p. 2011. Disponível no endereço http://www.unesco.org/new/fileadmin/MULTIMEDIA/HQ/SC/pdf/interagency_blue_paper_ocean_rioPlus20.pdf.

U.S. EPA. Interim Reregistration Eligibility Decision for Atrazine. **United States Environmental Protection Agency**, 2003.

U.S. EPA. Preliminary Interpretation of the Ecological Significance of Atrazine Stream-water Concentrations Using a Statistically Designed Monitoring Program. **United States Environmental Protection Agency**, 2007.

VERLI, H. (2014). **Bioinformática da Biologia à flexibilidade molecular**. Porto Alegre, 2014. 282 p.

XIE, F. J. et al. Identification of immune responsible fibrinogen beta chain in the liver of large yellow croaker using a modified annealing control primer system. **Fish & shellfish immunology**, v. 27, n. 2, p. 202-209, 2009.

WAGNER, G. et al. STINGRAY: system for integrated genomic resources and analysis. **BMC research notes**, v. 7, n. 1, p. 132, 2014.

WALKER, C. H. et al. **Principles of Ecotoxicology**. Taylor & Francis, 2a ed, Londres, p.309, 2001.

WANG, Z.; GERSTEIN, M.; SNYDER, M. RNA-Seq: a revolutionary tool for transcriptomics. **Nature Reviews Genetics**, v. 10, n. 1, p. 57-63, 2009.

WELCH, L. et al. Bioinformatics curriculum guidelines: Toward a definition of core competencies. **PLoS computational biology**, v. 10, n. 3, p. e1003496, 2014.

WEYTS, F. A. A. et al. Cortisol induces apoptosis in activated B cells, not in other lymphoid cells of the common carp, Cyprinus carpio L. **Developmental & Comparative Immunology**, v. 22, n. 5, p. 551-562, 1998.

WHALEN, M. M. et al. Immunomodulation of human natural killer cell cytotoxic function by triazine and carbamate pesticides. **Chemico-biological interactions**, v. 145, n. 3, p. 311-319, 2003.

WILLIAMS, C. Combatting marine pollution from land-based activities: Australian initiatives. **Ocean & coastal management**, v. 33, n. 1, p. 87-112, 1996.

WU, R. S. S. Eutrophication, water borne pathogens and xenobiotic compounds: environmental risks and challenges. **Marine Pollution Bulletin**, v. 39, n. 1-12, p. 11-22, 1999.

ZEEMAN, M. G.; BRINDLEY, W. A. Effects of toxic agents upon fish immune systems: a review. **Immunologic considerations in toxicology**, v. 2, p. 1-60, 1981.

ZELIKOFF, J. T. et al. Biomarkers of immunotoxicity in fish: from the lab to the ocean. **Toxicology letters**, v. 112, p. 325-331, 2000.

ZERBINO, D. R.; BIRNEY, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. **Genome research**, v. 18, n. 5, p. 821-829, 2008.

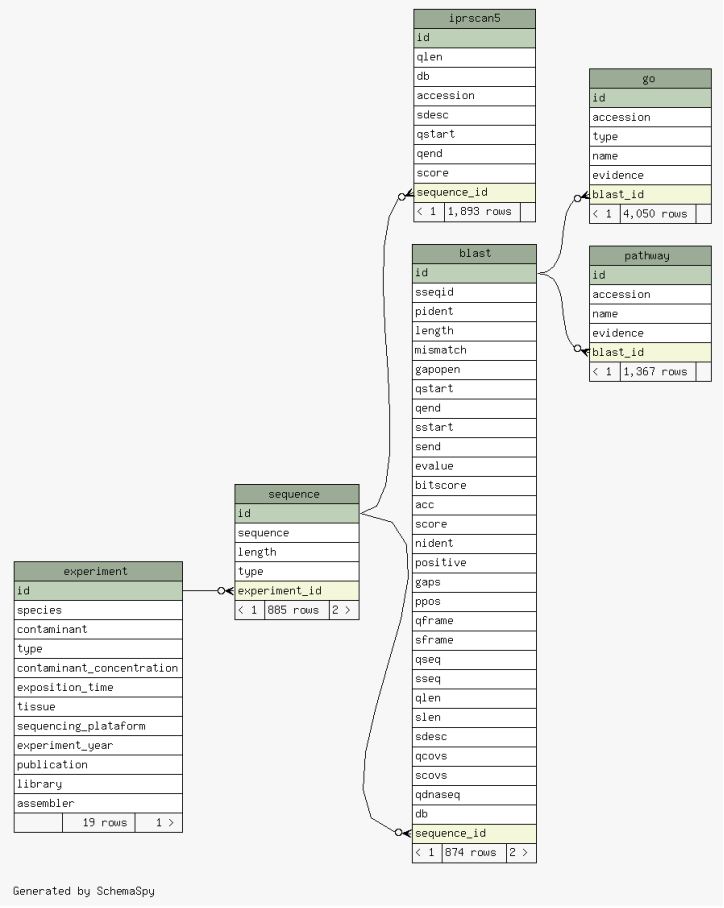
ZDOBNOV, E. M.; APWEILER, R. InterProScan–an integration platform for the signature-recognition methods in InterPro. **Bioinformatics**, v. 17, n. 9, p. 847-848, 2001.

ZHENG, Y. et al. iAssembler: a package for de novo assembly of Roche-454/Sanger transcriptome sequences. **BMC bioinformatics**, v. 12, n. 1, p. 453, 2011.

ZOU, S. et al. Occurrence and distribution of antibiotics in coastal water of the Bohai Bay, China: impacts of river discharge and aquaculture activities. **Environmental Pollution**, v. 159, n. 10, p. 2913-2920, 2011.

APÊNDICES

Apêndice A – Esquema para o banco de dados MySQL desenvolvido.



Apêndice B – Lista de genes induzidos pela exposição a atrazina em fígado de peixes *Prochilodus lineatus*. Os genes homólogos foram encontrados através de alinhamentos (BLAST) no banco de dados nr.

Processo biológico/descrição	Código de acesso GenBank	E-value	Proteína alvo alinhada (%)
Coagulação sanguínea			
Antitrombina 3	XP_007240173.1	1e-31	17,23
Imunidade			
Colectina 12	XP_007255963.1	4e-26	25,00
Complemento C3	XP_007238158.1	1e-63	35,37
Fibrinogênio alfa	AGO58845.1	2e-28	49,55
Fibrinogênio beta	NP_997939.1	3e-20	9,28
Imunidade e coagulação sanguínea			
Inibidor de protease do plasma C1	XP_007237022.1	3e-39	17,51
Interação vírus-hospedeiro			
Proteína AMBP	NP_957412.2	1e-	47,40

Processo biológico/descrição	Código de acesso GenBank	E-value	Proteína alvo alinhada (%)
		67	
Metabolismo de retinol			
Proteína de ligação ao retinol 4	XP_007232922.1	3e-23	25,26
Resposta a compostos tóxicos			
Proteína de ligação saxitoxinas e tetrodotoxinas 2	XP_007238277.1	3e-20	33,16
Transporte de lipídios			
Apolipoproteína C-I	XP_007231203.1	2e-14	65,88
Transporte de oxigênio			
Hemoglobina beta	XP_007238254.1	3e-17	26,35

Apêndice C - Lista de genes induzidos pela exposição a cobre em fígado de peixes *Prochilodus lineatus*. Os genes homólogos foram encontrados através de alinhamentos (BLAST) no banco de dados nr.

Processo biológico/descrição	Código de acesso GenBank	E-value	Proteína alvo alinhada (%)
Adesão celular			
Caderina	XP_007239384.1	2e-40	41,45
Glicoproteína associada a microfibrilas	XP_007238404.1	9-25	25,20
Apoptose			
Proteína tumoral traducionalmente controlada	XP_007246736.1	4e-32	34,50
Comunicação celular			
Proteína de junção celular GAP Cx32.2	XP_007255451.1	4e-38	32,67
Imunidade			
Fibrinogênio alfa	AAH75895.1	9e-48	14,33
Fibrinogênio beta	XP_007242808.1	2e-	34,92

Processo biológico/descrição	Código de acesso GenBank	E-value	Proteína alvo alinhada (%)
		110	
Imunidade e coagulação sanguínea			
Inibidor de protease do plasma C1	XP_007237022.1	3e-34	15,62

Apêndice D – Lista de genes induzidos pela exposição a fenantreno em fígado de peixes *Prochilodus lineatus*. Os genes homólogos foram encontrados através de alinhamentos (BLAST) no banco de dados nr.

Processo biológico/descrição	Código de acesso GenBank	E-value	Proteína alvo alinhada (%)
Enovelamento de proteínas			
Proteína complexo-T 1 eta	XP_009949993.1	6e-61	35,53
Imunidade			
Fibrinogênio beta	XP_007242808.1	1e-92	62,34
Lectina tipo-C 4-D	XP_006642436.1	1e-12	14,54
Glicólise e imunidade			
Glicose 6 fosfato isomerase	XP_007228249.1	2e-130	35,44
Gliconeogênese			
Fosfoenolpiruvato carboxilase	XP_007244225.1	2e-84	23,40

Processo biológico/descrição	Código de acesso GenBank	E-value	Proteína alvo alinhada (%)
Imunidade e coagulação sanguínea			
Antitripsina alfa 1	XP_007253008.1	3e-53	26,40
Glicoproteína rica em histidina	XP_007252644.1	3e-63	38,47
Prototrombina	XP_007244969.1	3e-16	10,41
Inibição de protease			
Inter alfa inibidor de tripisina cadeia pesada H2	XP_007235043.1	3e-105	20,09
Resposta a estresse			
Aldeído desidrogrenase 7-A1	XP_003390782.1	3e-7	37,31
Transporte de íons			
Ceruloplasmina	XP_006007918.1	1e-7f	3,04
Serotransferina 2	XP_007257884.1	1e-12	15,54

Apêndice E – Lista de genes induzidos pela exposição a atrazina em fígado de peixes *Poecilia vivipara*. Os genes homólogos foram encontrados através de alinhamentos (BLAST) no banco de dados nr.

Processo biológico/descrição	Código de acesso GenBank	E-value	Proteína alvo alinhada (%)
Apoptose			
Inibidor bax 1	XP_007566952.1	2e-29	22,73
Proteína guanina de ligação a nucleotídeos beta 2	ACO14497.1	6e-12	11,04
Biomineralização			
Ectonucleotideo pirofosfatase/fosfodiesterase 2	XP_007569097.1	7e-40	7,95
Biosíntese de ácidos biliares			
Esterol alfa 12 hidroxilase	XP_008432528.1	1e-43	15,46

Processo biológico/descrição	Código de acesso GenBank	E-value	Proteína alvo alinhada (%)
Coagulação sanguínea, quimiotaxia			
Cofator de heparina 2	XP_007563577.1	4e-148	40,79
Glicólise			
Frutose bifosfato aldolase B	XP_007548330.1	2e-121	46,70
Homeostase do colesterol			
Proteína transmembrana 97	XP_008423439.1	4e-80	80,23
Imunidade			
Complemento C3	XP_008426008.1	1e-120	11,67
Complemento C4	XP_007559679.1	9e-94	11,31
Imunoglobulina de alta afinidade gama Fc receptor IB	XP_007568184.1	7e-35	18,79
Imunidade e coagulação sanguínea			
Cininogênio 1	XP_008404505.1	3e-59	26,95

Processo biológico/descrição	Código de acesso GenBank	E-value	Proteína alvo alinhada (%)
Fibrinogênio gama	XP_008402544.1	8e-20	11,29
Prototrombina	XP_007541445.1	7e-06	15,72
Inibidor de protease do plasma C1	XP_005807821.1	3e-16	6,53
Inibição de proteases			
Inter alfa inibidor de tripisina cadeia pesada H2	XP_007552244.1	4e-97	5,83
Macroglobulina alfa 2	XP_005807548.1	5e-78	9,28
Inter alfa inibidor de tripisina cadeia pesada H2	XP_007552244.1	2e-95	19,30
Proteólise			
Elastase 1	XP_007577589.1	3e-100	64,55
Tripsina 1	XP_007553823.1	1e-97	59,50
Interação vírus-hospedeiro			
Proteína AMBP	XP_008422211.1	2e-47	20,92

Processo biológico/descrição	Código de acesso GenBank	E-value	Proteína alvo alinhada (%)
Metabolismo de aminoácidos			
Homogentisato dioxigenase	XP_008436825.1	1e-20	11,29
Metabolismo de cetonas			
Succinil CoA coenzima trasferase	XP_008416554.1	2e-61	22,39
Metabolismo de colesterol			
Apolipoproteína B-100	XP_005475014.1	1e-22	2,03
Metabolismo de lipídios			
Lipoproteína lipase	XP_005806437.1	7e-19	7,54
Sinalização celular			
Proteína de superfície celular A33	XP_003969896.1	3e-40	40,21
Transporte de carboidratos			
Glicose-6-fosfato	XP_007575988.1	3e-	30,77

Processo biológico/descrição	Código de acesso GenBank	E-value	Proteína alvo alinhada (%)
translocase		80	
Transporte de lipídios			
Apolipoproteína A-I	XP_008424299.1	5e-68	44,87
Transportador de ânion orgânicos em soluto 2A1	XP_007571452.1	6e-10	7,65

Apêndice F – Lista de genes induzidos pela exposição a cobre em fígado de peixes *Poecilia vivipara*. Os genes homólogos foram encontrados através de alinhamentos (BLAST) no banco de dados nr.

Processo biológico/descrição	Código de acesso GenBank	E-value	Proteína alvo alinhada (%)
Adesão celular			
Complexo protéico ligador a fator de crescimento subunidade ácida lábil	XP_007560522.1	5e-07	15,25
Proteína de ligação ao hialuronato 2	XP_007551516.1	9e-49	13,46
Apoptose			
Proteína guanina de ligação a nucleotídeos beta 2	XP_005795376.1	2e-68	35,33
Coagulação sanguínea			
Plasminogênio	XP_007573313.1	1e-82	16,92

Processo biológico/descrição	Código de acesso GenBank	E-value	Proteína alvo alinhada (%)
Glicólise			
Beta enolase	XP_005816431.1	1e-86	29,49
Glucuronidação de xenobióticos			
UDP glucuronosiltransferase 1-1	XP_007570405.1	2e-95	28,27
UTP glicose 1 fosfato uridililtransferase	XP_008416114.1	7e-111	33,94
Imunidade			
Complemento C3	AEJ08067.1	4e-34	3,81
Fribrinogênio alfa	XP_008401331.1	4e-30	7,76
Fibronectina	XP_007560176.1	2e-39	2,83
Imunidade e coagulação sanguínea			

Processo biológico/descrição	Código de acesso GenBank	E-value	Proteína alvo alinhada (%)
Fibrinogênio gama	XP_008402544.1	1e-179	61,81
Inibição de proteases			
Inter alfa inibidor de tripisina cadeia pesada H3	XP_007563409.1	1e-88	16,04
Inter alfa inibidor de tripisina cadeia pesada H2	XP_007574610.1	2e-46	10,83
Macroglobulina alfa 2	XP_005807548.1	5e-47	5,74
Metabolismo de colesterol			
Apolipoproteína B-100	XP_005475014.1	4e-20	3,69
Metabolismo de lipídios			
Fosfolipase A2	XP_008415777.1	5e-56	61,11
Oxidação-redução			
Citocromo P450 2F2	XP_007553247.1	8e-117	35,99

Processo biológico/descrição	Código de acesso GenBank	E-value	Proteína alvo alinhada (%)
Citocromo P450 3A30	XP_007544553.1	2e-65	67,36
Citocromo P450 3A56	XP_007576306.1	1e-11	6,46
Glutationa S-transferase 3 microssomal	XP_007574665.1	6e-40	52,14
Regulação imunidade			
Fucolectina 7	XP_007569946.1	4e-17	43,17
Transporte de lipídios			
Apolipoproteína Eb	XP_008429391.1	5e-67	50,19
Apolipoproteína C-I	XP_008429390.1	3e-31	79,52
Biosíntese de piridina			
Quinureninase	XP_007568732.1	4e-94	29,59

Apêndice G – Lista de genes induzidos pela exposição a fenantreno em fígado de peixes *Poecilia vivipara*. Os genes homólogos foram encontrados através de alinhamentos (BLAST) no banco de dados nr.

Processo biológico/descrição	Código de acesso GenBank	E-value	Proteína alvo alinhada (%)
Adesão celular			
Complexo protéico ligador a fator de crescimento subunidade ácida lábil	XP_007560522.1	5e-07	15,25
Biomineralização			
Ectonucleotideo pirofosfatase/fosfodiesterase 2	XP_007569097.1	4e-85	15,56
Coagulação sanguínea, quimiotaxia			
Cofator de heparina 2	XP_007563577.1	4e-120	35,05
Glicólise			

Processo biológico/descrição	Código de acesso GenBank	E-value	Proteína alvo alinhada (%)
Frutose bifosfato aldolase B	XP_007548330.1	5e-76	37,36
Gliconeogênese			
Frutose 1,6 bifosfato 1	XP_007548901.1	4e-34	18,10
Glucuronidação de xenobióticos			
UTP glicose 1 fosfato uridililtransferase	XP_006806870.1	4e-62	79,37
Imunidade			
Complemento C3	XP_006806870.1	2e-41	4,45
Complememto C8 gama	XP_005803373.1	5e-39	30
Imunidade, inflamação			
Lisozima C	XP_007575081.1	3e-77	81,51

Processo biológico/descrição	Código de acesso GenBank	E-value	Proteína alvo alinhada (%)
Inibição de proteases			
Macroglobulina alfa 2	XP_007544742.1	6e-26	7,13
Inter alfa inibidor de tripisina cadeia pesada H2	XP_007574610.1	6e-47	10,83
Metabolismo de aminoácidos			
Triptofano 2,3 dioxigenase	XP_008419272.1	6e-21	11,14
Proteólise			
Carboxipeptidase A1	XP_007570371.1	9e-37	16,59
Transporte			
Aquaporina 7	XP_008407704.1	7e-102	61,38
Transporte de íons			

Processo biológico/descrição	Código de acesso GenBank	E-value	Proteína alvo alinhada (%)
Serotransferina 2	XP_007559512.1	1e-09	4,49
Transporte de lipídios			
Apolipoproteína A-I	XP_008429393.1	3e-16	44,87
Apolipoproteína C-I	XP_008429390.1	3e-31	79,52
Proteína de ligação a ácidos graxos	XP_008429302.1	2e-75	92,86
Transporte de proteínas			
Fator de ADP-ribosilação	XP_009439877.1	2e-50	36,40