Nestor Cubas Wendt

ANÁLISE DO TRANSCRIPTOMA DA VIEIRA PATA-DE-LEÃO *Nodipecten nodosus* (Linnaeus, 1758): MONTAGEM, ANOTAÇÃO DOS TRANSCRITOS E CARACTERIZAÇÃO ESTRUTURAL DO CYP30E1

Dissertação submetida ao Programa de Pós-Graduação em Bioquímica da Universidade Federal de Santa Catarina para a obtenção do Grau de Mestre em Bioquímica.

Orientador: Prof. Dr. Afonso Celso Dias Bainy
Coorientador: Dr. Guilherme de Toledo e Silva

Florianópolis
2017

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Wendt, Nestor Cubas
ANÁLISE DO TRANSCRIPTOMA DA VIEIRA PATA-DE-LEÃO Nodipecten nodosus (Linnaeus, 1758): MONTAGEM, ANOTAÇÃO DO TRANSCRITOS E CARACTERIZAÇÃO ESTRUTURAL DO CYP30E1 / Nestor Cubas Wendt ; orientador, Afonso Celso Dias Bainy ; coorientador, Guilherme de Toledo-Silva. - Florianópolis, SC, 2017.
114 p.

Dissertação (mestrado) - Universidade Federal de Santa Catarina, Centro de Ciências Biológicas. Programa de Pós Graduação em Bioquímica.

Inclui referências

1. Bioquímica. 2. Nodipecten nodosus. 3. RNA-Seq. 4. Citocromo P450. I. Bainy, Afonso Celso Dias. II. Toledo Silva, Guilherme de. III. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Bioquímica. IV. Título.

"Análise do transcriptoma da vieira pata-de-leão Nodipecten nodosus (Linnaeus, 1758): montagem, anotação dos transcritos e caracterização estrutural do CYP30E1"

Por

Nestor Cubas Wendt

Dissertação julgada e aprovada em sua forma final pelos membros titulares da Banca Examinadora (006/PPGBQA/2017) do Programa de Pós-Graduação em Bioquímica - UFSC.

Prof(a). Dr(a). Ariane Zamoner Pacheco de Souza
Coordenador(a) do Programa de Pós-Graduação em Bioquímica

Banca examinadora:

Dr(a) Afonso Celso Dias Bainy (Universidade Federal de Santa Catarina)
Orientador(a)

Dr(a) Andrea Rita Marrero (Universidade Federal de Santa Catarina)

Dr(a) Patrícia Hermes Stoco (Universidade Federal de Santa Catarina)

Dr(a) Karim Hahn Lüchmann (Universidade Federal de Santa Catarina)

Florianópolis, 24 de fevereiro de 2017.

AGRADECIMENTOS

Aos meus pais, pelo suporte que possibilitou a realização deste curso de mestrado, especialmente no primeiro ano. Ao professor Afonso Bainy, por ter me aceito como aluno e me orientado durante este período. Agradecimentos especiais ao Guilherme Toledo, por ter me trazido para esta área e me ensinado muitas das coisas que sei hoje. Ao Guilherme Razzera, pela ajuda e discussões sobre biologia estrutural. Á Universidade Federal de Santa Catarina, por toda a estrutura que fornece aos alunos, especialmente a Biblioteca e o Restaurante Universitário. À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior pelo apoio financeiro concedido através da bolsa de mestrado.

RESUMO

Os ecossistemas costeiros fornecem bens e serviços valiosos. Entretanto, a maioria destes ecossistemas já apresentam impactos antropogênicos. Historicamente os ecossistemas costeiros são alvo do descarte de resíduos industriais, agrícolas e de efluentes urbanos. Esta contaminação, além de gerar riscos à saúde humana, modifica a ecologia dos ecossistemas marinhos. Esta dissertação está dividida em dois capítulos. O primeiro trata do sequenciamento e caracterização do transcriptoma de glândula digestiva de vieiras da espécie *Nodipecten nodosus*. As leituras passaram por verificação de qualidade e foram montadas através de três diferentes métodos. A montagem de melhor qualidade (Velvet k = 45) apresentou um tamanho de contig N50 de 2.301 pares de base (pb), compreendendo 76.861 transcritos. Destes, 33,72% foram anotados em bancos de dados públicos. Diversos transcritos de genes envolvidos na biotransformação de xenobióticos e com atividade antioxidante foram identificados. No segundo capítulo foram apresentados os resultados da classificação dos 33 transcritos citocromos P450 (CYP) buscados no transcriptoma de glândula digestiva da vieira *N. nodosus*. Seis novas famílias CYP foram identificadas. Além disso, a proteína CYP30E1 foi caracterizada estruturalmente, através de métodos computacionais. As sequências analisadas apresentaram, em geral, os motivos característicos desta superfamília conservados. A caracterização do CYP30E1 revela uma alta semelhança com o CYP3A4 humano, indicando funções semelhantes no metabolismo de alguns xenobióticos. Esta dissertação apresenta importantes contribuições na caracterização do transcriptoma da vieira *N. nodosus* e de seus transcritos CYP.

Palavras-chave: 1. *Nodipecten nodosus* 2. RNA-Seq 3. Citocromo P450

ABSTRACT

Coastal ecosystems provide valuable goods and services. However, most of these ecosystems already show anthropogenic impacts. Historically, coastal ecosystems have been subject to the disposal of industrial, agricultural and urban wastes. This contamination, in addition to producing risks to human health, modifies the marine ecosystem's ecology. This work is divided into two chapters. The first deals with the sequencing and characterization of the transcriptome of the digestive gland of scallops from the species *N. nodosus*. High-throughput sequencing reads went through quality checking and were assembled through three different methods. The assembly with the highest quality (Velvet k = 45) had a N50 contig size of 2,301 base pairs, comprising 76,861 transcripts. Of these, 33.72% were annotated in public databases. Several transcripts from genes involved in xenobiotic transformation or that had antioxidant activity were identified. In the second chapter, the results of the classification of 33 cytochrome P450 transcritos, searched in the digestive gland transcriptome of the scallop *N. nodosus*, were presented. Six new CYP families were identified. In addition, the protein CYP30E1 was characterized structurally, by computational methods. The analyzed sequences presented, in general, well conserved motifs from this superfamily. The characterization of the CYP30E1 revealed a high similarity with the human CYP3A4, indicating similar function in the metabolism of xenobiotics. This dissertation produced important contributions in the transcriptome characterization of the scallop *N. nodosus* and its CYP transcripts.

Keywords: 1. *Nodipecten nodosus* 2. RNA-Seq 3. Cytochrome P450

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

aa	Aminoácido
BIC	Critério Bayesiano de Informações
BWA	Alinhador *Burrows-Wheeler*
cDNA	DNA complementar
CYP	Citocromo P450
DNA	Ácido desoxirribonucleico
FPKM	Fragmentos por Quilobase de transcritos por Milhão de leituras mapeadas
GO	*Gene Ontology*
HPA	Hidrocarboneto policíclico aromático
IED-BIG	Instituto de Ecodesenvolvimento da Baía da Ilha Grande
INDEL	Inserção ou deleção de bases no DNA
NCBI	Centro Nacional de Informação Biotecnológica
NCBInr	Banco de sequências não redundantes de proteínas do NCBI
NIH	Instituto Nacional de Saúde
pb	Pares de bases
PDB	Banco de Dados de Proteínas
qPCR	PCR quantitativo
RNA	Ácido ribonucleico
RNAm	Ácido ribonucleico mensageiro
RMSD	Raiz média quadrática
SSH	Hibridização subtrativa supressiva
SNP	Polimorfismos de nucleotídeo único
TPM	Transcritos por Milhão
XML	*eXtensible Markup Language*

SUMÁRIO

INTRODUÇÃO GERAL

A poluição aquática é definida como a introdução, de forma direta ou indireta, de energia ou compostos em um ecossistema aquático, resultando em dano (GESAMP, 1991). Por definição, poluentes são substâncias que causam efeitos biológicos adversos em organismos. Dentro destes, os xenobióticos são considerados poluentes que não fazem parte da bioquímica de um organismo (WALKER et al., 2001).

No Brasil, em 2006, um quarto da população do país vivia nos municípios da zona costeira (SERAFIM; HAZIN, 2006), o que produz diversas alterações aos ambientes estuarinos e marinhos, além de riscos à saúde humana. Estes ecossistemas costeiros fornecem bens e serviços valiosos, sendo utilizados no lazer e em diversas atividades industriais. De fato, a maioria dos ecossistemas costeiros no mundo já apresentam sinais de impacto antropogênico (HALPERN et al., 2008) (Figura 1), devido à sobrepesca, descarte de poluentes e degradação da qualidade da água (JACKSON et al., 2001; LOTZE et al., 2006).

Figura 1 – Mapa mundial do impacto antropogênico em 20 ecossistemas costeiros e marinhos. Regiões fortemente impactadas em destaque: Caribe, Mar do Norte, sul do Japão. Destaque para a Austrália (quadrado à direita), região pouco impactada. Imagem modificada de (HALPERN et al., 2008).

Infelizmente, os impactos humanos nestes ecossistemas têm aumentado nas últimas décadas (HALPERN et al., 2008). Diversas vias introduzem compostos de origem antrópica no meio ambiente: esgoto, agricultura, aquicultura, deposição atmosférica, atividades portuárias, extração de petróleo ou descartes acidentais de produtos químicos (GOLDENBURG; ELLIOTT; NAYLOR, 2001; PRÓSPERI; NASCIMENTOS, 2006). Assim, organismos que vivem em ambientes aquáticos são continuamente expostos a contaminantes oriundos de diversas atividades humanas (SARKAR et al., 2006). Além disso, muitos poluentes são persistentes e podem ser acumulados em diversas espécies marinhas (SARKAR et al., 2006). Níveis tróficos mais altos também sofrem alta exposição a estas moléculas, devido a biomagnificação (HOLSBEEK et al., 1999; SOLÉ; PORTE; ALBAIGÉS, 2001; KUCKLICK et al., 2002). Neste contexto, é imprescindível a avaliação dos impactos que estes poluentes geram no ambiente aquático.

A exposição de organismos aos xenobióticos pode gerar diversas alterações moleculares, histológicas e comportamentais, eventualmente resultando em um efeito negativo (SARKAR et al., 2006). Estas modificações são chamadas de biomarcadores. Nas últimas décadas, diversas agências de fiscalização ambiental passaram a favorecer a utilização biomarcadores no monitoramento de contaminação ambiental, pois a quantificação direta de poluentes não necessariamente indica efeitos deletérios em organismos aquáticos (CAJARAVILLE et al., 2000).

Como ferramentas no monitoramento, biomarcadores moleculares são importantes pois atuam como um alerta precoce, permitindo a realização de medidas corretivas em tempo adequado. Além disso, biomarcadores apresentam especificidade, sensibilidade e podem ser utilizados em diversas espécies (SARKAR et al., 2006). Geralmente, a modificação da expressão gênica é o primeiro tipo de resposta em organismos expostos a poluentes (BRULLE et al., 2008). Assim, tecnologias de biologia molecular como PCR quantitativo (qPCR), hibridização subtrativa supressiva (SSH), microarranjos e RNA-Seq têm sido utilizadas na seleção de novos biomarcadores (BULTELLE et al., 2002; MEDEIROS et al., 2008; LIANG et al., 2009; ZHANG et al., 2012; JIN et al., 2015; PIAZZA et al., 2016).

Dentre os organismos utilizados em programas de biomonitoramento, os moluscos bivalves se destacam por apresentarem ampla distribuição geográfica, serem filtradores e em maioria sésseis, e utilizados em diversas pesquisas (BAINY et al., 2000; CAJARAVILLE et al., 2000; RADLOWSKA; PEMPKOWIAK, 2002; BOCCHETTI et al., 2008; PAN et al., 2011; LIU et al., 2012; ZHANG et al., 2012 ZHENG et al., 2015). Além disso, algumas espécies são capazes de acumular contaminantes em seus tecidos (PÉREZ-CADAHÍA et al., 2004; SOLÉ; BUET; ORTIZ, 2007).

Especificamente, as espécies da família Pectinidae, popularmente conhecidas como vieiras, têm sido utilizados em muitos estudos de respostas biológicas a poluentes, especialmente petróleo bruto, hidrocarbonetos aromáticos policíclicos (HPA) e metais, devido a sua capacidade de acumular HPAs em seus tecidos, grande distribuição geográfica e dificuldade em escapar de áreas contaminadas (REN; LIU, 2006; HANNAM et al., 2009; LIU et al., 2012; PAN; PIAZZA et al., 2016). No Brasil, a família Pectinidae possui 6 gêneros e 16 espécies (RIOS, 1994). Dentre estas, a vieira *N. nodosus* gera interesse econômico, por possuir crescimento rápido, podendo chegar até 17,8 cm, e ser bem aceita no mercado brasileiro (RUPP; PARSONS, 2006) (Figura 2). Devido ao estímulo à malacocultura no Brasil, criou-se uma demanda por águas não poluídas, estimulando consciência ambiental e ações públicas com o objetivo de melhorar a qualidade das águas em comunidades costeiras (RUPP; PARSONS, 2006). Santa Catarina é, atualmente o segundo maior produtor da vieira *N. nodosus* no Brasil. O maior problema no cultivo desta espécie, hoje, é a produção de sementes, já que larvas são sensíveis às condições do cultivo.

Figura 2 – A esquerda, foto da concha de vieira *N. nodosus*, com aproximadamente 14 cm. A direita, vista da gônada e músculo adutor. Fotos de vieiras do Instituto de Ecodesenvolvimento da Baía da Ilha Grande (IED-BIG), obtidas na página da Prefeitura de Angra dos Reis (https://goo.gl/1ZfsmR).

O primeiro registro desta espécie no Brasil foi feito por HAAS (1953), na Ilha Grande, Rio de Janeiro. Esta vieira é distribuída principalmente na costa Atlântica da América Central e do Sul (RUPP; PARSONS, 2006), sendo geralmente encontrada dentro de pequenas cavernas ou entre rochas (DÍAZ; PUYANA, 1994). Vieiras do gênero *Nodipecten* possuem, caracteristicamente, nós bulbosos ocos em suas conchas. Não existem trabalhos estudando a longevidade desta espécie. Com relação a reprodução, estas vieiras são hermafroditas simultâneas que liberam gametas durante todo o ano de forma assíncrona, com picos na primavera e no verão (SCHLEDER et al., 2008). A filogenia da espécie *N. nodosus* é descrita a seguir:

Reino: Animalia
Filo: Mollusca
Classe: Bivalvia
Ordem: Ostreoida
Família: Pectinidae
Gênero: *Nodipecten*
Espécie: *Nodipecten nodosus*

Embora *N. nodosus* demonstre grande importância econômica, poucas sequências nucleotídicas e proteicas para esta espécie estão disponíveis nos bancos de dados públicos, com raros esforços para reverter este cenário (AMERICO et al., 2015). Neste contexto, avanços

em conhecimentos genômicos e transcriptômicos para esta vieira podem auxiliar na produção extensiva e no entendimento de mecanismos moleculares de resposta à poluição ambiental.

O objetivo desta dissertação foi caracterizar o transcriptoma da glândula digestiva da vieira *N. nodosus*, produzindo um repositório de informações nucleotídicas nesta espécie, que poderão ser utilizadas em futuros trabalhos. No primeiro capítulo, as leituras obtidas por sequenciamento Illumina foram montadas e anotadas, utilizando informações de bancos de dados públicos. Diversos transcritos de genes envolvidos em processos de biotransformação ou genes tradicionalmente utilizados como biomarcadores foram identificados. No segundo capítulo, foram selecionadas e classificadas os transcritos CYPs presentes no transcriptoma da glândula digestiva da vieira *N. nodosus*. Seis novas famílias foram identificadas, e este foi o primeiro esforço na classificação de CYPs nesta vieira. Além disso foi realizada uma caracterização funcional do CYP30E1.

OBJETIVOS

OBJETIVO GERAL

- Caracterizar o transcriptoma da glândula digestiva da vieira *N.* nodosus, especialmente transcritos de genes importantes em estudos ecotoxicológicos.

OBJETIVOS ESPECÍFICOS

- Produzir o primeiro transcriptoma, via RNA-Seq, da glândula digestiva da vieira *N. nodosus*.

- Investigar a diversidade de transcritos da superfamília dos CYPs, enzimas importantes no metabolismo de xenobióticos, expressas na glândula digestiva da vieira *N. nodosus.*

CAPÍTULO I:

MONTAGEM E ANOTAÇÃO DO TRANSCRIPTOMA DA GLÂNDULA DIGESTIVA DA VIEIRA *Nodipecten nodosus* (Linnaeus, 1758)

1. INTRODUÇÃO

As diferenças fenotípicas de células geneticamente idênticas sempre foi um intrigante campo de pesquisa na biologia molecular. O estudo do conjunto e nível de genes transcritos em uma célula, o transcriptoma, é fundamental para o nosso entendimento de diversos fenômenos biológicos. Desde o primeiro ácido ribonucleico (RNA) sequenciado nos anos 1960 (HOLLEY et al., 1965), diversos métodos para o estudo desta molécula foram desenvolvidos. Notavelmente, o método de sequenciamento desenvolvido por SANGER et al. (1977) foi o mais utilizado por décadas. Entretanto, ao estudar uma resposta molecular, além do conhecimento da sequência de um transcrito, é interessante conhecer seu nível de transcrição. Para isso, diversos métodos foram desenvolvidos, distintamente o qPCR (SAKATSUME et al., 1989) e o método de microarranjos (SCHENA et al., 1995). No final da década passada, entretanto, os chamados métodos de sequenciamentos de segunda geração foram introduzidos no mercado (MARGULIES et al., 2005). Estes métodos reduziram o custo por base e a complexidade dos experimentos, além de aprimorar a cobertura dos sequenciamentos (MOROZOVA; HIRST; MARRA, 2009), permitindo que cada vez mais amostras sejam sequenciadas (Figura 3).

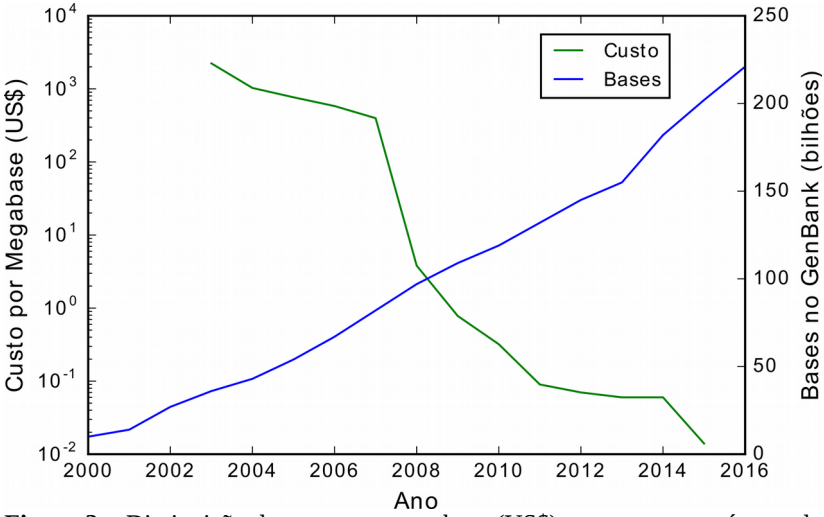


Figura 3 – Diminuição do custo por megabase (US$) e aumento no número de bases depositadas no *GenBank*, a partir do ano 2000. Os dados foram obtidos do Instituto Nacional de Saúde (NIH) americano (https://goo.gl/jh0Gxz).

O RNA-Seq é um destes métodos de segunda geração, destacando-se por ter um bom custo-benefício e produzir grandes quantidades de dados, que podem ser utilizados para estimar os níveis de transcrição gênica (WANG; GERSTEIN; SNYDER, 2009). Nele, primeiramente o RNA é isolado de uma amostra e convertido em DNA complementar (cDNA). Este então é ligado a adaptadores, que são complementares aos iniciadores utilizados nas reações do sequenciamento. Além disso, por permitir o sequenciamento de todos os transcritos de uma amostra, esta metodologia pode ser utilizada na identificação de todos os tipos de RNAs, codificantes ou não, em estudos de marcadores moleculares como polimorfismos de nucleotídeo único (SNP) ou microssatélites e variações genéticas entre diferentes espécies ou populações (MOROZOVA; HIRST; MARRA, 2009). Tradicionalmente, a maior área de aplicação dos sequenciamentos de segunda geração é na caracterização de transcriptomas de espécies não-modelo (EKBLOM; GALINDO, 2010). Especificamente em bivalves, o RNA-Seq já foi utilizado em diversos estudos (HOU et al., 2011; CHEN

et al., 2013; PAIRETT; SERB, 2013; LÜCHMANN et al., 2015; PAN et al., 2015).

Este tipo de estudo geralmente é bastante descritivo, mas estes são um ponto de partida necessário para aplicações posteriores, como desenvolvimento de marcadores moleculares, estudo de *splicing* alternativo, avaliação da expressão de genes, seja através de qPCR ou microarranjos (EKBLOM; GALINDO, 2010). Métodos de sequenciamentos de alto desempenho tem grande potencial em estudos ecotoxicológicos, pois eles fornecem informações acerca das adaptações moleculares, incluindo regulação gênica dos organismos expostos aos contaminantes (SCHIRMER et al., 2010). Infelizmente, a maioria das tecnologias de sequenciamento, hoje, só permitem o sequenciamento de fragmentos pequenos de cDNA. Assim, as leituras dos transcritos fragmentados precisam ser montadas novamente em suas sequências originais. Este processo, conhecido como montagem, consiste na sobreposição de regiões similares nestas leituras para produção de sequências consenso ou contigs. Assim, em um transcriptoma, a montagem ideal é aquela em que o número e tamanho dos contigs é igual ao número e tamanho dos transcritos da espécie. Após montagem, geralmente é realizado o processo de anotação dos contigs montados. O objetivo deste processo é agregar, ao máximo, informações sobre as sequências. Isto inclui identificar sequências homólogas em bancos de dados, marcadores genéticos putativos, regiões conservadas, entre outros.

Marcadores genéticos podem ser utilizados na identificação de variabilidade genética (LABORDA, 2011). Os microssatélites são pequenas sequências compostas de 1 a 6 nucleotídeos repetidos em *tandem* (LITT e LUTY, 1989). Este tipo de marcador genético pode ser conservado dependendo da região em que está presente e, consequentemente, pode ser utilizado em diversos estudos, como genética de populações (CAIXETA et al., 2009). Já os SNPs são mais abundantes, e sua identificação foi facilitada pelo advento dos sequenciamentos de segunda geração. Considerando que diferentes populações podem responder, ou metabolizar, de forma diferente xenobióticos (BICKHAM et al., 2000; BRAMMELL et al., 2004; LUKKARI et al., 2004), a incorporação deste tipo de informação em estudos ecotoxicológicos é importante.

O objetivo do presente trabalho é caracterizar o transcriptoma de glândula digestiva da vieira *N. nodosus*, através de RNA-Seq gerado por tecnologia Illumina. As leituras foram montadas através de três métodos diferentes, e as sequências resultantes da melhor montagem foram anotadas em bancos de dados públicos. Além disso, foram reconhecidos marcadores genéticos putativos. Este é o primeiro esforço para o estudo do transcriptoma da vieira *N. nodosus* através de sequenciamentos de segunda geração, assim sendo uma importante fonte de informações para pesquisas futuras nesta espécie.

2. OBJETIVOS

2.1 OBJETIVO GERAL

- Fornecer um transcriptoma de qualidade da glândula digestiva da vieira *N. nodosus*, produzindo um repositório de informações para esta espécie.

2.2 OBJETIVOS ESPECÍFICOS

- Montar as amostras sequenciadas em um transcriptoma referência de qualidade da glândula digestiva da vieira *N. nodosus*.

- Identificar transcritos de genes envolvidos em processos de biotransformação de xenobióticos e transcritos utilizados tradicionalmente como biomarcadores, através de anotação a partir de diversos bancos de dados públicos.

- Mapear marcadores genéticos putativos nos transcritos, fornecendo um ponto de partida para estudos de genética populacional nesta espécie.

3. METODOLOGIA

3.1 EXTRAÇÃO DE RNA, PREPARO DAS BIBLIOTECAS E SEQUENCIAMENTO

Seis vieiras da espécie *N. nodosus* foram obtidas do IED-BIG, localizado em Angra dos Reis, Rio de Janeiro. Três destas foram acondicionadas, durante 22 dias, em gaiola próxima a uma plataforma de extração de petróleo que descartava água produzida no mar. As outras três foram acondicionadas em área, ao mar, que não recebe diretamente descarte de água produzida, por 24 dias. Todos os animais estavam no mesmo estágio de desenvolvimento e nenhum morreu. O RNA mensageiro (RNAm) foi extraído a partir de glândulas digestivas de todos os animais, armazenadas em *RNAlater* (Sigma) e mantidas a 4 ºC para a preparação das bibliotecas de RNA-Seq. As bibliotecas foram produzidas na empresa Helixxa (Campinas, São Paulo), de acordo com o *TruSeq RNA Sample Preparation kit*. A avaliação da qualidade foi realizada utilizando um 2100 Bioanalyzer e a quantificação através de qPCR, seguindo as instruções do *Library qPCR Quantification kit* (Illumina Inc., San Diego, EUA). As bibliotecas foram clusterizadas utilizando cBOT, através do protocolo descrito no *cBOT User Guide*. O sequenciamento foi realizado em um HiSeq 2000. Através deste protocolo foram obtidas leituras pareadas de 100 pb.

3.2 VERIFICAÇÃO DE QUALIDADE E MONTAGEM *DE NOVO*

A lista completa de programas e bancos de dados utilizados neste capítulo é apresentada no Apêndice A. A qualidade das leituras foi verificada pelo FastQC (ANDREWS, 2016). Adaptadores, bases de baixa qualidade (Q < 30 nota phred) e leituras pequenas (tamanho < 25 pares de base) foram removidas através do programa Trimmomatic (BOLGER; LOHSE; USADEL, 2014). As leituras resultantes passaram pelo método de normalização *in silico* disponibilizado no pacote Trinity (GRABHERR et al., 2011) (cobertura máxima 20). Então, as leituras normalizados das seis amostras foram combinadas e montadas *de novo* por três métodos: através do programa Trinity com o tamanho de *kmer* 25 e dos programas Velvet (ZERBINO; BIRNEY, 2008) e Oases (SCHULZ et al., 2012) com tamanhos de *kmer* 25 e 45. Valores de

FPKM (Fragmentos por Quilobase de transcritos por Milhão de leituras mapeadas) das três montagens foram calculados pelo RSEM (LI; DEWEY, 2011), depois do alinhamento das leituras não normalizados através do Bowtie (LANGMEAD; SALZBERG, 2012). Transcritos com baixo FPKM (FPKM < 0,5) foram removidos e então os programas Transrate (SMITH-UNNA et al., 2016) e BUSCO (SIMÃO et al., 2015) foram utilizados para avaliar a qualidade das montagens. Depois de mapear as leituras nos transcritos montados, o Transrate calcula uma pontuação para a montagem baseado na qualidade das bases, cobertura das leituras, segmentação das sequências e exatidão (SMITH-UNNA et al., 2016). O BUSCO procura, na montagem, ortólogos universais de cópia única. Para a análise do BUSCO, foram utilizados os ortólogos universais de cópia única de metazoários. A partir da montagem escolhida (Velvet, k = 45), as fases abertas de leitura (ORF) dos transcritos foram preditas pelo TransDecoder (HAAS; PAPANICOLAOU, 2016), utilizando parâmetros padrões.

3.3 ANOTAÇÃO

Os transcritos foram alinhados através do programa BLAST+ (RAMSAY et al., 2000) em três bancos de dados públicos: banco de sequências não redundantes de proteínas do NCBI (NCBInr), TrEMBL e SwissProt (APWEILER et al., 2004). O valor máximo de *e-value* aceito para estes alinhamentos foi de 10^{-5}. O banco Pfam-A (SONNHAMMER; EDDY; DURBIN, 1997) também foi consultado através do programa HMMER (FINN; CLEMENTS; EDDY, 2011) com *e-value* máximo de 10^{-10}. A partir dos resultados do TrEMBL e SwissProt, termos Gene Ontology (GO) (ASHBURNER et al., 2000) foram obtidos a partir do arquivo XML (eXtensible Markup Language) disponível na página do UniProt. WEGO (YE et al., 2006) foi utilizado para representar graficamente as anotações GO. Transcritos alinhados à proteínas virais e bacterianas foram removidos das análises subsequentes.

3.4 IDENTIFICAÇÃO DE MARCADORES PUTATIVOS

Microssatélites foram identificados utilizando o programa MISA (THIEL et al., 2003). Somente motivos dinucleotídeos de seis ou mais repetições foram considerados. Para motivos trinucleotídeos até

pentanucleotídeos, a repetição mínima aceita foi de quatro vezes, e para motivos hexanucleotídeos três repetições mínimas. Para a detecção de SNPs, inicialmente as leituras foram alinhadas aos *contigs* através do Alinhador *Burrows-Wheeler* (BWA) (LI; DURBIN, 2009), com parâmetros padrões. Então, o programa SAMtools (LI et al., 2009) e o BCFtools (NARASIMHAN et al., 2016) foram utilizados para prever todos os variantes. Depois da remoção das inserções e deleções (INDEL), VCFtools (DANECEK et al., 2011) foi utilizado na identificação dos SNPs putativos.

4. RESULTADOS

4.1 SEQUENCIAMENTO E MONTAGEM

Foram obtidas 130.320.965 leituras pareadas através do sequenciamento Illumina, cada qual com 100 pb. Durante o controle de qualidade 2.903.457 leituras com adaptadores, bases de baixa qualidade ou tamanho foram removidos (Tabela 1). As métricas separadas por amostra estão disponíveis na Tabela 2, e não apresentaram diferenças significativas.

Tabela 1 – Métricas dos dados brutos e filtrados do sequenciamento, agrupando todas as amostras.

Métricas	Dados brutos	Dados filtrados
Leituras (milhões)	130,32	127,45
Pares de base (bilhões)	26,36	25,47
Bases Q20	96,65%	98,86%

Tabela 2 – Métricas dos dados brutos produzidos por sequenciamento Illumina separadas por amostra.

Amostras	1	2	3	1	5	6
Leituras (milhões)	28,84	37,18	43,34	41,21	55,69	48,58
Bases Q20	98,84%	98,81%	98,89%	98,87%	98,88%	98,87%

A normalização resultou em um conjunto de 5.122.812 leituras pareadas. As métricas de qualidade das três montagens *de novo* testadas neste trabalho estão disponíveis na Tabela 3 e Apêndice B. Os transcritos montados através dos programas Trinity (k = 25) e Velvet (k = 25 e 45) apresentaram tamanho de contig N50 de 2.051, 2.408 e 2.301, respectivamente. Velvet (k = 25) teve 53,9% de transcritos com, pelo menos, uma base sem cobertura, enquanto que a montagem do Trinity teve 3,9% de transcritos com cobertura média menor que 1. Com relação aos genes ortólogos universais, a montagem que aprestou o maior número foi a Velvet (k = 25), com 656 dos 978 genes presentes. A montagem escolhida para as análises posteriores foi a Velvet (k = 45).

Tabela 3 – Métricas de qualidade das três montagens produzidas neste trabalho: Trinity (k = 25), Velvet (k = 25) e Velvet (k = 45)..

Métricas	Trinity (k =25)	Velvet (k = 25)	Velvet (k = 45)
Transcritos	94.074	60.062	76.861
N50	2.051	2.408	2.301
Fragmentos mapeados	92,85%	90,02%	89,94%
Transcritos com, pelo menos, uma base sem cobertura	37,23%	53,90%	39,01%
Transcritos com cobertura média menor que 1	3,90%	2,04%	1,54%
Score Transrate	0,259	0,282	0,302
Ortólogos universais	634 (64,83%)	656 (67,76%)	651 (66,56%)

Os resultados da predição de ORFs estão na Tabela 4. No total, 25.812 ORFs com tamanho médio de 1086 pb foram preditas dos transcritos montados. 47,27% das ORFs preditas eram completas.

Tabela 4 – Métricas das ORFs completas e parciais preditas pelo programa TransDecoder, identificadas na montagem Velvet (k = 45) do transcriptoma de glândula digestiva de *N. nodosus*.

Métricas	Completas	Parciais	Internas
ORFs	12.202	10.856	2.754
Tamanho médio (pb)	1.110	1.161	681
N50 (pb)	1.395	1.473	777
Maior ORF (pb)	12.828	14.001	21.786

4.2 ANOTAÇÃO

Os resultados da anotação dos 76.861 transcritos montados (Velvet k = 45) para o transcriptoma de *N. nodosus* estão na Tabela 5 e Figuras 4 e 5. Para o banco de dados NCBInr e TrEMBL, respectivamente, 25.605 (33,31%) e 25.366 (33,0%) dos transcritos tiveram *hits* significantes. Com relação ao SwissProt, 18.796 (24,45%) foram anotados. A maioria dos transcritos com mais de 1.500 pb foram

anotados (Figura 4). Aproximadamente 1.000 ORFs preditas não foram anotadas. Um gráfico de setores com as 4 espécies de maiores *hits* na anotação NCBInr é apresentado na Figura 5. Neste banco, a maioria das sequências tiveram *hits* com espécies relativamente próximas, como *Crassostrea gigas* e *Lottia gigantea*. Para a anotação do banco SwissProt, a maioria dos *hits* foi em sequências de espécies modelos como *Homo sapiens*, *Mus musculus* e *Rattus norvegicus*.

Tabela 5 – Transcritos da montagem Velvet (k = 45) anotados nos bancos de dados públicos utilizados neste trabalho.

Banco de dados	*Hits*
NCBInr	25.605 (33,31%)
TrEMBL	25.366 (33,00%)
SwissProt	18.796 (24,45%)
Pfam-A	14.330 (55,52%)
Gene Ontology	17.856 (23,23%)
Anotados em todos os bancos	6.466 (8,41%)
Anotados em, pelo menos, um banco	25.920 (33,72%)
Não anotados	50.941 (66,28%)

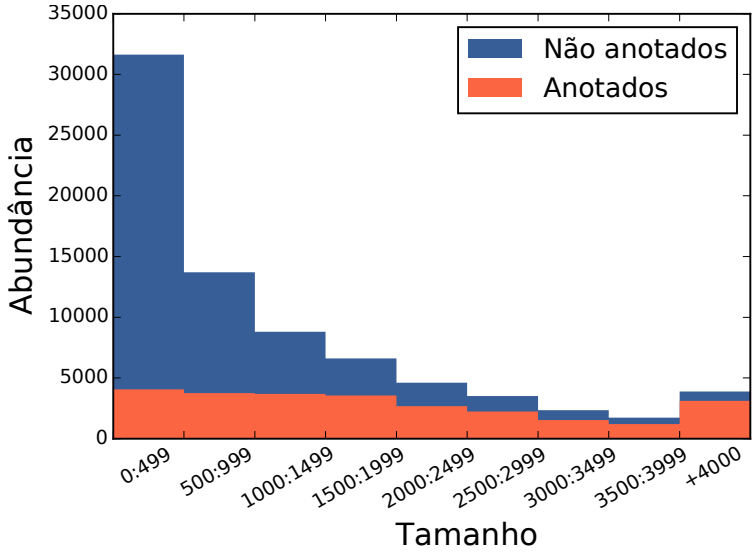


Figura 4 – Histograma da distribuição de tamanhos dos transcritos pela presença de anotação.

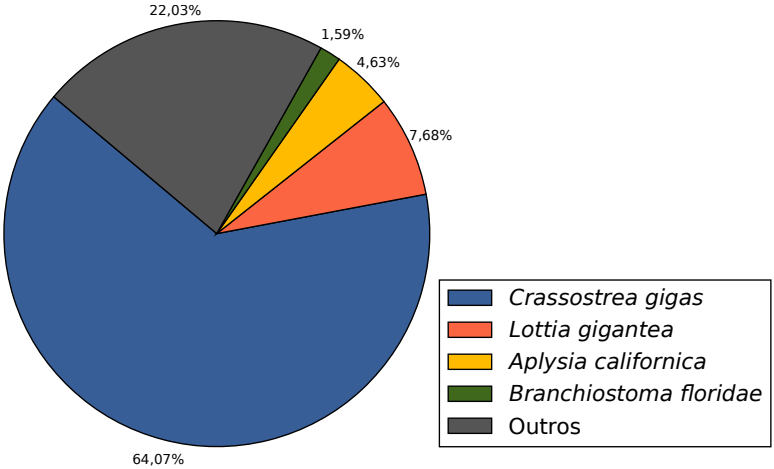


Figura 5 – Gráfico de setores das espécies com o maior número de hits na anotação NCBInr, para os transcritos montados.

A representação gráfica dos termos GO produzida pelo WEGO apresenta a diversidade de transcritos anotados neste banco (Figura 6). Os contigs foram distribuídos em 34 categorias do GO, pertencentes as três grandes classes: 5 em Componente Celular, 10 em Função Molecular e 19 em Processo Biológico. Dentre as categorias mais representativas estão “célula” (GO:0005623), “organela” (GO:0043226), “ligação” (GO:0005488), “catálise” (GO:0003824), “regulação biológica” (GO:0065007) e “processo metabólico” (GO:0008152).

Diversos transcritos de genes envolvidos em processos de biotransformação e com atividade antioxidante foram identificados (Tabela 6). Muitos destes genes são utilizados como biomarcadores de contaminação aquática ou estresse oxidativo, e sua identificação possibilita futuros trabalhos neste sentido na vieira *N. nodosus*.

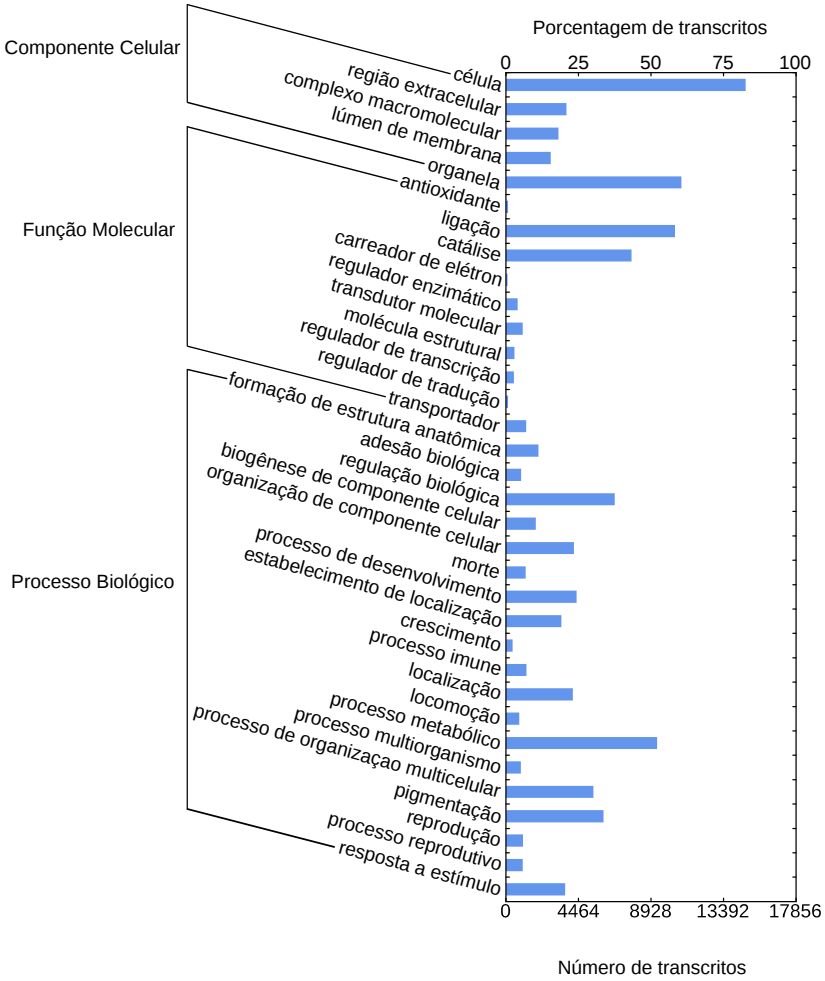


Figura 6 – Representação gráfica dos termos *Gene Ontology* associados aos transcritos presentes no transcriptoma. Os termos foram traduzidos.

Tabela 6 – Transcritos anotados de genes envolvidos em processos de biotransformação ou com atividade antioxidante.

Grupo	Transcritos
CYPs	229

GSTs	48
SULTs	67
Catalase	5
SODs	10
Glutationa peroxidase	2
Glutationa redutase	2
Glutarredoxina	9
Peroxirredoxina	3
Tiorredoxina	16
Tiorredoxina redutase	2
Total	393

4.3 MARCADORES MOLECULARES PUTATIVOS

Foram identificados 10.959 microssatélites nos transcritos montados (Tabela 7). Os tipos mais comums de motivos foram dinucleotídeos (43,1%), seguidos de trinucleotídeos (39,16%) e tetranucleotídeos (10,97%). Motivos do tipo pentanucleotídeos (1,47%) e hexanucleotídeos (3,23%) foram os menos comuns. 131.643 SNPs também foram identificados no transcriptoma de glândula digestiva (Tabela 8). Transições foram 1,45 vezes mais comuns (59,22%) que transversões (40,78%).

Tabela 7 – Sumário dos diferentes tipos de microssatélites identificados na montagem *de novo* (Velvet = 45) do transcriptoma de glândula digestiva da vieira *N. nodosus*.

Tipo de motivo microssatélite	Abundância
Dinucleotídeos	4.732 (43,18%)
Trinucleotídeos	4.291 (39,16%)
Tetranucleotídeos	1.202 (10,97%)
Pentanucleotídeos	161 (1,47%)
Hexanucleotídeos	573 (5,23%)

Tabela 8 – Sumário dos SNPs identificados na montagem *de novo* (Velvet = 45) do transcriptoma de glândula digestiva da vieira *N. nodosus*.

Tipo de SNP	Abundância
Transições	77.964 (59,22%)
A-G/G-A	39.071 (29,69%)
C-T/T-C	38.893 (29,54%)
Transversões	53.679 (40,78%)
A-C/C-A	13.411 (10,19%)
A-T/T-A	19.280 (14,65%)
T-G/G-T	13.148 (09,99%)
G-C/G-C	7.840 (05,56%)
Transições/Transversões	1,45

5. DISCUSSÃO

5.1 SEQUENCIAMENTO E MONTAGEM

Com relação à qualidade das leituras, a Tabela 1 revela um sequenciamento de alta qualidade, visto que somente 3,23% das leituras foram removidas após a filtragem e nenhuma diferença significativa foi observada entre as amostras (Tabela 2). Testes e avaliações extensivas de diversas estratégias de montagem frequentemente levam a melhores montagens e resultados em análises subsequentes como anotação e expressão diferencial (ZHAO et al., 2011). Embora a montagem Velvet (k = 45) tenha tido menos leituras mapeadas em comparação as outras montagens *de novo*, ela teve menos transcritos com cobertura média menor que 1 e somente 39,01% de transcritos com, pelo menos, uma base sem cobertura (Tabela 3). A cobertura é um fator importante se o objetivo do sequenciamento é a identificação de transcritos diferencialmente expressos. O número de ortólogos universais recuperados na análise do programa BUSCO foi similar em ambas as montagens *de novo* do Velvet. Além disso, Velvet (k = 45) teve a maior pontuação do Transrate (0,302). Baseado nesta pontuação e nas métricas de mapeamento, a montagem *de novo* Velvet (k = 45) foi escolhida como a mais adequada para as aplicações subsequentes. Espera-se que a melhor montagem seja uma representação mais fiel dos transcritos originais, com menos problemas de quimeras e segmentações.

Com respeito a outros transcriptomas de bivalves, o tamanho médio de transcrito obtido neste trabalho foi maior que o da *Corbicula fluminea* (791 pb) (CHEN et al., 2013), *Argopecten orradians* (529 pb) (PAIRETT; SERB, 2013), *Cyclina sinensis* (980 pb) (PAN et al., 2015), *Crassostrea brasiliana* (575 pb) (LÜCHMANN et al., 2015), *Patinopecten yessoensis* (618 pb) (HOU et al., 2011) e menor que o da *Crassostrea gigas* (2.328 pb) (versão de fevereiro, 2015, disponível no RefSeq). O maior tamanho médio de transcrito obtido no transcriptoma da *Crassostrea gigas* se deve ao grande esforço de sequenciamento já realizado para descrever este organismo, que possui o genoma sequenciado (ZHANG et al., 2012) e a possível diferença natural entre as espécies. Todavia, as métricas obtidas neste trabalho demonstram que a estratégia de montagem *de novo* utilizada foi satisfatória e serve como repositório de transcritos para a vieira *N. nodosus*, auxiliando aplicações

posteriores como análise de transcrição diferencial, estudos filogenéticos e de biologia estrutural.

5.2 ANOTAÇÃO

Neste presente trabalho, foram preditas 25.812 ORFs para o transcriptoma de glândula digestiva de *N. nodosus* (Tabela 4). Em comparação, no transcriptoma de todos os tecidos da *Corbicula fluminea*, 105.737 ORFs foram identificadas (CHEN et al., 2013). Para *Cyclina sinensis* 20.877 ORFs foram preditas no transcriptoma da hemolinfa (PAN et al., 2015) e para a *Crassostrea gigas* 45.406, a partir de diversos transcriptomas e do genoma (ZHANG et al., 2012). O número de ORFs preditas depende de diversos fatores, desde da qualidade e redundância da montagem, programa utilizado, tecido sequenciado e esforço de sequenciamento. Neste sentido, o valor de 45.406 ORFs, calculados com a ajuda do genoma, identificadas na *Crassostrea gigas* é um bom guia. O grande número de ORFs observado na *Corbicula fluminea* provavelmente é resultado do uso da ferramenta *getorf* (WESTERLUND; BJØRNHOLM, 2009), que é menos restritiva que o TransDecoder, utilizado neste trabalho. O número de ORFs preditas aqui é parecido com a da espécie *Cyclina sinensis*, chegando próximo a metade do número identificado na *Crassostrea gigas*. Isto provavelmente é resultado do sequenciamento de um único tecido.

Com relação à anotação, possivelmente muitos dos transcritos de menor tamanho foram erros de montagem, e por isso não foram anotados. Além disso, a maioria das ORFs preditas foram anotadas. Entretanto, ainda existe a possibilidade destes serem transcritos ainda desconhecidos. Finalmente, as espécies mais comuns na anotação dos transcritos no banco de dados NCBInr foram próximas à vieira *N. nodosus*, considerando as sequências depositadas nestes bancos de dados, como a *Crassostrea gigas*, *Lottia gigantea* e a *Aplysia californica*. Ainda assim, o SwissProt, sendo um banco de dados manualmente anotado e revisado, esperava-se uma anotação majoritariamente de espécies modelo.

A Figura 6, através da classificação GO, revela diversos processos biológicos e funções moleculares das proteínas codificadas pelos transcritos sequenciados neste transcriptoma. Espera-se que categorias mais gerais sejam mais frequentes, devido a formatação hierárquica dos

termos e foi o caso aqui. Termos de hierarquia maior, e portanto mais generalistas, como “célula” (GO:0005623) ou “organela” (GO:0043226) para classe Componente Celular ou “ligação” (GO:0005488) e “processo metabólico” (GO:0065007) para as classes Função Molecular e Processo Biológico, respectivamente, foram os mais comuns. Considerando a importância da vieira pata-de-leão na malacocultura e na ecotoxicologia, diversos termos GO despertam interesse. Transcritos com anotação GO “reprodução” (GO:0000003), “processo reprodutivo” (GO:0022414), “crescimento” (GO:0040007) são possíveis alvos para futuros estudos que tratam do cultivo desta espécie para alimentação humana. Da mesma maneira, os termos “resposta a estímulo” (GO:0050896), “morte” (GO:0008219), “processo imune” (GO:0002376), “regulação biológica” (GO:0065007), “regulador de transcrição” (GO:0006355), “regulador de tradução” (GO:0045182), “regulador enzimático” (GO:0030234) são de possível interesse em estudos ecotoxicológicos. Esta análise realizada pelo WEGO é uma maneira simplificada de apresentar os contigs anotados. Além disso, é importante reconhecer as limitações do GO, especialmente em transcriptomas de espécies não-modelo. O principal problema, neste caso, é a falta de informações sobre proteínas não estudadas, levando a um viés, nesta classificação, a proteínas conservadas.

5.3 MARCADORES MOLECULARES PUTATIVOS

Dentre os microssatélites, 2.151 foram identificados no transcriptoma de todos os tecidos da *Corbicula fluminea* (CHEN et al., 2013), 1.400 destes trinucleotídeos (Tabela 7). Para o transcriptoma da vieira *Patinopecten yessoensis*, gerado a partir de amostras do músculo aditor, glândula digestiva e gônadas, 2.768 microssatélites foram identificados, sendo o tipo mais comum os trinucleotídeos (1.091) (HOU et al., 2011). Repetições de trinucleotídeos são o tipo mais comum de microssatélites em transcritos porque sua replicação não gera uma mudança na fase de leitura (SUTHERLAND; RICHARDS, 1995). Embora isto seja verdade para o transcriptoma da *Corbicula fluminea* e *Patinopecten yessoensis*, a vieira *N. nodosus* apresentou mais repetições de dinucleotídeos, por uma margem estreita. Apesar deste resultado poder ser simplesmente consequência de erros de montagem, uma maior

frequência de microssatélites que não são trinucleotídeos cria um maior risco de mudança na fase de leitura de genes.

Somente no transcriptoma de *Patinopecten yessoensis* foram identificados SNPs. Nesta espécie foram identificados 34.841 SNPs, sendo 20.958 destes transições e 12.804 transversões (razão de 1,64). Vieiras, em geral, possuem alta variabilidade genética (BEAUMONT, 2006). Os dados deste trabalho sugerem que a vieira *N. nodosus* segue o mesmo padrão. Microssatélites e SNPs podem ser úteis como marcadores moleculares em futuros estudos de genética de populações.

6. CONCLUSÕES

O presente estudo apresenta o transcriptoma, gerado a partir do sequenciamento de glândula digestiva, da vieira *N. nodosus*. Os resultados demonstram uma montagem de alta qualidade. Este é um dos primeiros esforços realizados para sequenciar e anotar as sequências nucleotídicas nesta espécie. Dentre os resultados aqui apresentados, destacam-se a alta qualidade da montagem *de novo* do transcriptoma da vieira *N. nodosus*. Além disso, genes de diversos processos biológicos foram sequenciados, o que torna este transcriptoma um importante repositório para futuros estudos nesta espécie.

CAPÍTULO II:

CLASSIFICAÇÃO E CARACTERIZAÇÃO DOS TRANSCRITOS CYPS PRESENTES NO TRANSCRIPTOMA DA GLÂNDULA DIGESTIVA DA VIEIRA *Nodipecten nodosus* (Linnaeus, 1758)

1. INTRODUÇÃO

A superfamília dos citocromos P450 agrupa diversas hemoenzimas que catalisam reações metabólicas de substratos endógenos e xenobióticos. Estes substratos incluem moléculas endógenas como esteróides, eicosanóides e sais biliares, mas também drogas e outros poluentes (GONZALEZ; KIMURA, 2003). As primeiras evidências acerca deste tipo de proteína foram geradas em 1958, após a observação de um espectro com pico de absorbância em 450 nm em amostras hepáticas de ratos (KLINGENBERG, 1958). Após décadas de estudos, a importância destas enzimas nas mais diversas vias metabólicas é evidente. Nelson (2013) mostrou a importância dos CYPs durante a evolução dos organismos e em aplicações na indústria farmacêutica, agricultura e biotecnologia, uma vez que estas enzimas metabolizam diversos fármacos, hormônios e outras moléculas importantíssimas nas mais variadas vias metabólicas.

Considerando a presença dos CYPs nos três reinos biológicos, é seguro afirmar que o gene CYP ancestral surgiu muito cedo na história evolutiva das espécies, antes de separação de eucariotos, bactérias e arqueobactérias (SEZUTSU; LE GOFF; FEYEREISEN, 2013). Em animais, há evidências que todos os CYPs foram originados a partir de duplicações em *tandem* de um único gene (NELSON; GOLDSTONE; STEGEMAN, 2013). Devido ao grande esforço empregado no sequenciamento de genomas e transcriptomas, são conhecidas mais de 18.500 exemplares desta superfamília, compreendendo centenas de espécies dos todos os reinos (NELSON, 2009). Diversos artigos já foram publicados sobre este grupo de proteínas, em diferentes áreas de pesquisa.

No nível de sequência, os CYPs apresentam imensa diversidade. Apesar de existirem motivos conservados nesta superfamília, hoje não existe um resíduo universalmente conservado (SEZUTSU; LE GOFF; FEYEREISEN, 2013). Contudo, o aumento do número de estruturas experimentalmente determinadas nesta superfamília revela um enovelamento conservado (DE MONTELLANO, 2005). A despeito de exceções, os CYPs possuem regiões características (WERCK-REICHHART; FEYEREISEN, 2000) (Figura 7). Na região N-terminal possuem uma região membranar e hidrofóbica. Esta é seguida de dois *clusters*: o primeiro de resíduos

básicos e o segundo de prolinas, geralmente PPGP. Na hélice C está presente o motivo WxxxR, responsável por formar a ponte salina com o propianato do grupo heme. Na região central da hélice I está presente o motivo A(A,G)x(E,D)T, que está envolvido na protonação do oxigênio distal do complexo ferro-hidroperoxo. Na hélice K, o motivo ExxR é considerado importante para o enovelamento correto e incorporação do grupo heme. Então, está presente o motivo PERF ou suas variações. Por fim, o motivo FxxGxxxCxG, presente na hélice L, é importante na ligação do o grupo heme-tiolato, através de sua cisteína.

Figura 7 – Esquema evidenciando as regiões e os motivos conservados na maioria das enzimas CYPs. Os valores de “x” indicam qualquer aminoácido. Imagem produzida no programa Inkscape.

Estudos filogenéticos classificam os genes CYP em diferentes clãs, famílias e subfamílias (NELSON et al., 1993). A família é definida por um número arábico, enquanto que a subfamília por uma letra. Os clãs são grupos de famílias. Sugere-se que novos genes CYPs sejam submetidos ao Cômite de Nomenclatura de Citocromos P450 para classificação (NELSON, 2006). Como regra geral, criada após classificação dos CYPs de mamíferos, genes de uma mesma família precisam de ao menos 40% de identidade, enquanto que em subfamílias este valor sobe para 55% (NELSON, 2006). É necessário ter em mente que, apesar desta regra funcionar bem para sequências de mamíferos, ela pode e já foi quebrada em diversas situações (NELSON, 2006).

Especificamente em mamíferos, o clado mais estudado neste aspecto, as famílias CYP2, CYP3 e CYP4 possuem o maior número de genes, quando comparadas com as famílias restantes (NELSON et al., 2004). Estas famílias, juntamente com a CYP1, são capazes de responder a estímulos ambientais, como exposição aos xenobióticos (NEBERT; WIKVALL; MILLER, 2013). As enzimas da família CYP3 metabolizam aproximadamente 30% das drogas utilizadas clinicamente,

hoje (ZANGER; SCHWAB, 2013). Em humanos, essa família possui somente uma subfamília, CYP3A, que consiste de 4 genes. Devido a alta similaridade entre os genes CYP3A, estes metabolizam substratos similares (ANDREW WILLIAMS et al., 2002). Dentre os substratos estão principalmente moléculas lipofílicas grandes como ciclosporina A, eritromicina e paclitaxel (ZANGER; SCHWAB, 2013). Entretanto, moléculas pequenas também são metabolizadas: tamoxifeno, diversos antidepressivos e opioides, entre outros (ZANGER; SCHWAB, 2013).

O esgoto recebe, todo dia, diversos contaminantes oriundos de atividades industriais, agricultura ou de áreas residenciais. Tradicionalmente, os estudos ecotoxicológicos sempre priorizaram poluentes clássicos, como pesticidas e moléculas que demonstram persistência no meio ambiente. Recentemente, entretanto, a presença de fármacos em efluentes tem gerado preocupação, devido aos seus possíveis efeitos em organismos marinhos (BRAUSCH et al., 2012). Os primeiros artigos tratando da ocorrência, destino e efeito de fármacos no ambiente aquático sugiram no final da década de 1990 (HALLING-SORENSEN et al., 1998; DAUGHTON e TERNES, 1999). Após administração, os fármacos são excretados em sua forma original ou metabolizada. Como os métodos de tratamento de esgoto não foram desenhados para a eliminação deste tipo de molécula, estas não são necessariamente removidas e podem, assim, ser lançadas aos ambientes aquáticos (DEBLONDE; COSSU-LEGUILLE; HARTEMANN, 2011). De fato, vários estudos já identificaram, em diferentes concentrações, inúmeros fármacos em ambientes aquáticos, inclusive no lençol freático (TERNES, 1998; BENDZ et al., 2005; THOMAS, 2006; KIM et al., 2007; ROBERTS). Além disso, estudos já identificaram bioacumulação de fármacos em organismos aquáticos (BRAUSCH et al., 2012). Apesar do número de publicações estudando os efeitos ecotoxicológicos de fármacos ter aumentado, menos de 10% dos fármacos prescritos hoje possuem informações publicadas na literatura (BRAUSCH et al., 2012).

Um dos métodos computacionais utilizados, atualmente, no estudo de ligação entre uma proteína e seu ligante é o atracamento molecular. Assim é possível estudar a interação entre uma proteína de interesse e um fármaco, por exemplo. Basicamente, este método é dividido em duas etapas, a primeira consistindo na busca das possíveis posições do ligante no sítio ativo da proteína em questão, seguida da estimativa da força desta ligação (BROOIJMANS; KUNTZ, 2003). É

fundamental notar que o aumento do número de proteínas com modelos estruturais conhecidos, através de métodos como cristalografia de raios-X, é crucial para este tipo de estudo (BERMAN et al., 2000). O atracamento molecular foi idealizado em 1982 (KUNTZ et al., 1982) e é utilizado em diversas áreas de pesquisa biológica. A indústria farmacêutica, por exemplo, utiliza esta metodologia há muitos anos para o desenvolvimento de novos fármacos. Mais do que isso, esta metodologia é um campo de pesquisa por si só (KITCHEN et al., 2004). De fato, simular, em um computador, o fenômeno complexo de ligação entre duas moléculas é desafiador. Para isto, diversos métodos de atracamento molecular já foram desenvolvidos, com o intuito de simular adequadamente a flexibilidade molecular e quantificar a energia de ligação entre duas moléculas (KITCHEN et al., 2004).

Diversas teorias que tratam do mecanismo de ligação entre uma enzima e seu substrato já foram criadas. A primeira hipótese para explicar esta ligação foi o mecanismo chave-fechadura (FISCHER, 1894). Aqui, a complementariedade estrutural das moléculas é que permitiria esta ligação, não haveria flexibilidade. Eventualmente a ideia de encaixe induzido surgiu, introduzindo o conceito de que a interação entre a enzima e substrato induziriam mudanças conformacionais na proteína (KOSHLAND, 1958). Por fim, o conceito de seleção conformacional foi proposto em 1964, o qual sugere que proteínas possuem diversos estados conformacionais em solução, cabendo ao ligante selecionar as conformações mais compatíveis para a ligação. Somente na década de 1990, entretanto, este conceito ganhou fôlego a partir da publicação de um artigo referência de FRAUENFELDER et al. (1991). De fato, proteínas são sistemas flexíveis. Esta proposta de que existe uma seleção conformacional já acumula diversas evidências, principalmente através de estudos de ressonância magnética nuclear (CSERMELY; PALOTAI; NUSSINOV, 2010).

Com relação ao atracamento molecular, já foi demonstrado extensivamente que a utilização de diversos estados conformacionais é benéfico na predição de complexos (SANDAK; WOLFSON; NUSSINOV, 1998; ERICKSON et al., 2004; FERRARI et al., 2004; ALBERTS; TODOROV; DEAN, 2005; KOSKA et al., 2008;). Especificamente na ecotoxicologia, o atracamento molecular já foi utilizado em diversos estudos, evidenciando o potencial desta metodologia na área (WALKER; MCELDOWNEY, 2013). Yang e

colaboradores (2010) demonstraram que este método é capaz de identificar potenciais estrógenos; WU et al. (2010) utilizaram modelagem e atracamento molecular na predição de interações de contaminantes e diversos receptores em diferentes níveis tróficos; WU et al. (2009) verificaram que o atracamento molecular de receptores nucleares gerou resultados que condizem com os dados experimentais. Finalmente, o atracamento molecular já foi sugerido como uma ferramenta para a avaliação do impacto ambiental de fármacos (WALKER; MCELDOWNEY, 2013).

Neste contexto, este capítulo buscou classificar as sequências de CYPs encontrados no transcriptoma de glândula digestiva da vieira *N. nodosus* e utilizou modelagem e atracamento molecular para caracterizar uma proteína CYP30E1, frente a possíveis substratos. Os resultados aqui produzidos são interessantes considerando o baixo número de CYPs já classificados em invertebrados não-insetos. Além disso, a caracterização de um CYP30E1, família pouco estudada, é importante para o entendimento das possíveis funções que esta pode ter nesta espécie.

2. OBJETIVOS

2.1 OBJETIVO GERAL

- Identificar a caracterizar os transcritos CYPs presentes no transcriptoma da glândula digestiva da vieira *N. nodosus*.

2.2 OBJETIVOS ESPECÍFICOS

- Relatar a diversidade e padrões de expressão dos transcritos CYPs presentes no transcriptoma da glândula digestiva da vieira *N. nodosus*.

- Estudar a capacidade de metabolização das drogas eritromicina, bromoergocriptina e midazolam do CYP30E1, o transcrito CYP mais semelhante ao CYP3A4 humano identificado no transcriptoma da glândula digestiva da vieira *N. nodosus*.

3. METODOLOGIA

3.1 IDENTIFICAÇÃO E CLASSIFICAÇÃO DAS SEQUÊNCIAS CYPs

A lista completa de programas utilizados neste capítulo é apresentada no Apêndice C. As sequências CYPs da vieira *N. nodosus*, obtidas a partir do sequenciamento RNA-Seq de glândula digestiva, foram selecionadas através de busca da família p450 (PF00067) no banco Pfam-A (SONNHAMMER; EDDY; DURBIN, 1997), pelo programa HMMER (FINN; CLEMENTS; EDDY, 2011). Para remoção de pseudogenes, erros de montagem e transcritos incompletos, somente foram considerados aqueles transcritos com tamanho entre 350 e 600 aminoácidos (aa), valores em que estão contidos todos os CYPs funcionais das principais espécies modelo.

Para cada grupo de transcritos, agrupados na montagem pelo Velvet, o transcrito de maior TPM (Transcritos por Milhão) foi escolhido. As sequências selecionadas foram então alinhadas através do programa MUSCLE (EDGAR, 2004) e a conservação dos motivos da superfamília CYP foi checada manualmente. Logos de conservação dos principais motivos dos CYPs foram criados através da ferramenta WebLogo (CROOKS et al., 2004). Todas as sequências CYPs obtidas foram enviadas ao Dr. David Nelson para classificação oficial.

3.2 MODELAGEM ESTRUTURAL

A modelagem da estrutura do transcrito CYP30E1 foi realizada através pacote I-TASSER (YANG et al., 2015). Este CYP foi escolhido por ser o CYP mais semelhante ao CYP3A4 humano identificado no transcriptoma da glândula digestiva da vieira *N. nodosus*. Três modelos foram produzidos utilizando as estruturas cristalográficas 2J0D, 3UA1 e 5TE8, obtidas do Banco de Dados de Proteínas (PDB) (BERMAN et al., 2000). Estas são estruturas do CYP3A4 humano complexado com os substratos eritromicina, bromoergocriptina e midazolam, respectivamente. O CYP3A4 foi escolhido devido a semelhança entre as famílias CYP30 e CYP3, e por ser o CYP3 mais estudado. Após modelagem, o grupo heme foi adicionado através de minimização realizada pelo UCSF Chimera (PETTERSEN et al., 2004), com

parâmetros padrão. A carga do átomo de ferro foi definida em +2. A avaliação da qualidade dos modelos foi realizada com o Cscore, TM-Score e gráfico de Ramachandran produzido pelo RAMPAGE (LOVELL et al., 2003).

3.3 ATRACAMENTO MOLECULAR

O atracamento molecular foi realizado através do AutoDock Vina (TROTT; OLSON, 2010). Os três modelos produzidos foram avaliados quanto a capacidade de se ligar às moléculas das estruturas cristalográficas (Figura 8).

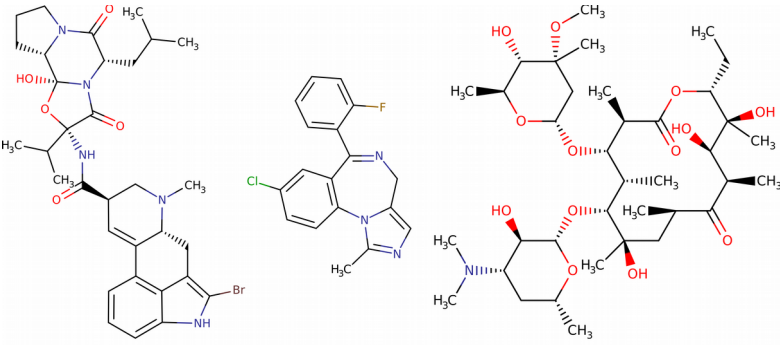


Figura 8 – Estruturas das moléculas utilizadas no atracamento molecular dos modelos estruturais do CYP30E1 da viera.

A caixa para o atracamento molecular foi centralizada no sítio ativo da enzima, usando como referência as estruturas cristalográficas. Todos os parâmetros da caixa e os aminoácidos considerados flexíveis na análise estão disponíveis no Apêndice D. Os três estados conformacionais modelados e os ligantes foram preparados para o atracamento molecular através das ferramentas acessórias *prepare_receptor4.py*, *prepare_flexreceptor4.py* e *prepare_ligand4.py* incluídos no pacote MGLtools. Em todos os cálculos de atracamento, a carga do átomo de ferro do grupo heme foi definida como +2. As estruturas dos fármacos foram obtidas dos modelos experimentais.

4. RESULTADOS

4.1 IDENTIFICAÇÃO E CLASSIFICAÇÃO DAS SEQUÊNCIAS CYPs

No total foram identificados 229 transcritos, em 100 *clusters*, da classe p450 no transcriptoma de glândula digestiva da vieira *N. nodosus*, através do banco de dados Pfam-A. Destes, foram selecionados 33 transcritos, sendo aqueles de ORFs completas e de tamanho entre 350 e 600 aa. Após o alinhamento destes, o logo dos seis principais motivos característicos dos CYPs foi gerado (Figura 9).

Figura 9 – Logo dos motivos conservados nos 33 transcritos CYPs identificados no transcriptoma da vieira *N. nodosus*. Em (A), o *cluster* de prolinas PPGP; em (B), o motivo WxxxR; em (C), o motivo A(A,G)x(E,D)T; em (D), o motivo ExxR, importante no enovelamento e incorporação do grupo heme; em (E), o motivo PERF; em (F), o motivo característico dos CYP FxxGxxxCxxG, cuja cisteína se liga ao grupo heme-tiolato.

O alinhamento e a Figura 9 revelam uma alta conservação destes motivos de CYPs. O motivo PPGP apresenta algumas alterações, principalmente no primeiro resíduo. Já no WxxxR, somente 3 sequências apresentam alterações no resíduo triptofano, e 4 no resíduo arginina. O motivo A(A,G)x(E,D)T é menos conservado. Foram observadas 10 alterações na primeira alanina. O importante resíduo treonina, envolvido no protonamento do oxigênio distal do complexo

ferro-hidroperóxido, foi substituído por serina em 4 sequências. Os aminoácidos ácido glutâmico e arginina do motivo ExxR estão todos conservados nas sequências selecionadas. Com relação ao PERF, somente o resíduo arginina foi completamente conservado. Finalmente, o motivo FxxGxxxCxG apresentou os resíduos fenilalanina e cisteína conservados em todas as sequências, enquanto que as glicinas variaram em 1 e 3 sequências, respectivamente. Os resultados completos da classificação realizada pelo Dr. David Nelson estão disponível na Tabela 9.

Tabela 9 – Transcritos CYPs selecionados identificados no transcriptoma da vieira *N. nodosus*. A classificação foi realizada pelo Dr. David Nelson. Transcritos com asteriscos pertencem às novas famílias identificadas.

Transcrito	Clã	TPM médio
CYP4JU1	4	2,315
CYP4JU2	4	2,071
CYP4JU3	4	2,265
CYP4JU4	4	0,454
CYP4JV1	4	2,4027
CYP4JX1	4	6,7772
CYP4JX2	4	40,9721
CYP4JW1	4	1,837
CYP20A1	20	3,306
CYP30E1	3	11,745
CYP30G1	3	8,5412
CYP30G2	3	8,0826
CYP30F1	3	5,264
CYP30F2	3	11,3438
CYP30F3	3	13,988
CYP30F4	3	9,2596
CYP44G1	mitocondrial	0,4521
CYP356B1	2	5,146

CYP356B2	2	8,247
CYP356B3	2	6,148
CYP3315A1*	2	23,009
CYP3315A2*	2	2,922
CYP3315B1*	2	29,380
CYP3315C1*	2	5,380
CYP3315D1*	2	2,014
CYP3316A1*	2	1,792
CYP3317A1*	2	2,317
CYP3318A1*	3	0,942
CYP3319A1*	46	2,061
CYP3320A1*	mitocondrial	0,214
CYP3067B4	7	4,614
CYP3072B1	mitocondrial	0,634
CYP3072C1	mitocondrial	1,203

Dentre os 33 transcritos avaliados, foram classificados CYPs em 7 clãs: CYP2, CYP3, CYP4, CYPmito, CYP7, CYP20 e CYP46. Transcritos dos clãs CYP19, CYP26, CYP51 e CYP74, cuja presença é esperada na espécie *N. nodosus* (NELSON; GOLDSTONE; STEGEMAN, 2013) não foram identificados neste transcriptoma. Seis famílias inéditas foram identificadas, CYP3315-CYP3320, compondo 10 transcritos. Dentre, os transcritos da família CYP30, o CYP30E1 desperta interesse devido a expressão alta e semelhança desta família com a família CYP3 de vertebrados, importante no metabolismo de fármacos e poluentes. Com relação aos motivos CYP no transcritos CYP30E1, todas as posições estão conservadas, com exceção de uma prolina no PPGP, substituída por uma isoleucina.

4.2 MODELAGEM ESTRUTURAL

A Tabela 10 contém as medidas utilizadas na avaliação da qualidade dos modelos gerados. Os valores de Cscore para todos os modelos são próximos a 1,00. Já para o TM-Score, os modelos

obtiveram em média um valor de 0,85. Os quatro modelos obtiveram mais de 96% de resíduos com ângulos permitidos. Todos os gráficos de Ramachandran estão disponíveis no Apêndice E.

Tabela 10 – Métricas de qualidade utilizadas na avaliação dos modelos produzidos para o transcrito CYP30E1.

Modelos	Cscore	TM-Score	Ramachandran permitidos
m2J0D	0,91	0,84±0,08	96,4%
m3UA1	1,05	0,86±0,07	96,4%
m5TE8	1,03	0,85±0,08	96,9%

No CYP3A4 humano, diversos resíduos importantes para a ligação e orientação dos substratos no sítio ativo desta enzima já foram identificados: Phe108, Ser119, Ile120, Leu211, Asp214, Ile301, Phe304, Ala305, Thr309, Ala370 e Leu373 (HALPERT, 1998; FOWLER et al., 2000, 2002; KHAN et al., 2002). A sequência CYP30E1 possui 6 destes 11 aminoácidos conservados (Figura 10). Os resíduos respectivos, na sequência da vieira, são: Phe109, Asn121, Asn122, Leu214, Lys217, Val324, Phe327, Ala328, Thr332, Ala392 e Thr395. Como demonstra a Figura 10, dois sítios, na vieira, possuem substituições de pequeno impacto em propriedades físico-químicas, Asn121 e Val324. Entretanto, três substituições apresentam mudanças mais significativas: a troca de dois resíduos hidrofóbicos por resíduos hidrofílicos em Asn122 e Thr395, e a troca de um resíduo de carga negativa por um de carga positiva em Lys217.

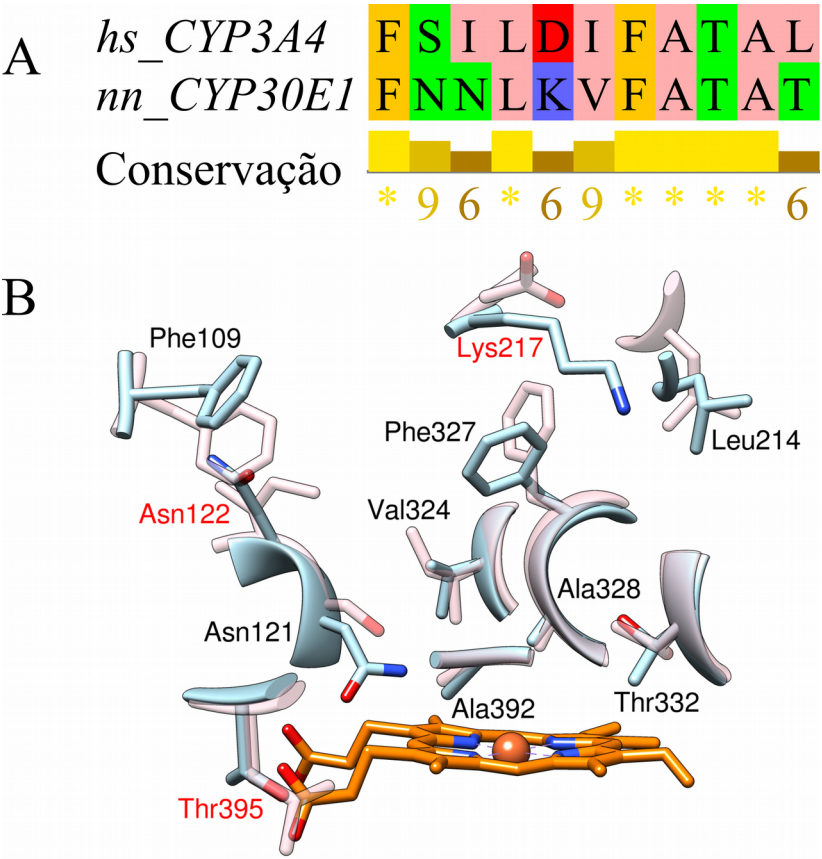


Figura 10 – Aminoácidos importantes para a função do CYP3A4 humano, no CYP30E1 da vieira *N. nodosus*. Em (A), alinhamento e conservação destes resíduos. As posições do CYP3A4 são: Phe108, Ser119, Ile120, Leu211, Asp214, Ile301, Phe304, Ala305, Thr309, Ala370 e Leu373. Aminoácidos conservados são marcados com um asterisco. O alinhamento foi colorido de acordo com o esquema Zappo, que colore de acordo com propriedades físico-químicas dos aminoácidos. Imagem gerada pelo Jalview. Em (B), posição destes resíduos no sítio ativos do CYP3A4 humano (rosa) e do modelo m5TE8 da vieira. Em detalhe, grupo heme em laranja. Imagem produzida pelo Chimera.

A sequência do transcrito CYP30E1 também apresenta, quando comparada com o CYP3A4 humano, duas inserções de 9 e 12

aminoácidos, logo em seguida uma da outra (Figura 11). Estas inserções estão em uma região externa da enzima, longe do sítio ativo e da região membranar.

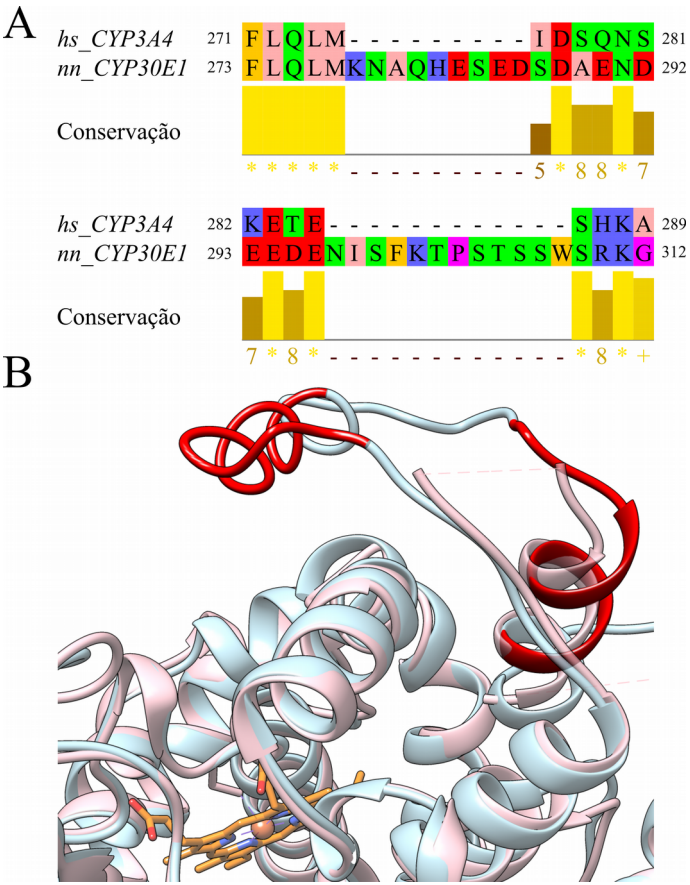


Figura 11 – Inserções presentes no transcrito CYP30E1 da vieira *N. nodosus*, não observadas no CYP3A4 humano. Em (A), alinhamento da região, colorida de acordo com esquema Zappo. Aminoácidos marcados com o sinal + possuem as mesmas propriedades físico-químicas. Figura produzida pelo Jalview. Em (B), as estruturas do CYP3A4 humano (rosa) e do modelo m2J0D gerado para a sequência da vieira (azul). A região em vermelho indica as duas inserções observadas. Em laranja, o grupo heme. Imagem gerada pelo Chimera.

4.3 ATRACAMENTO MOLECULAR

As três moléculas testadas apresentaram, após atracamento molecular, poses coerentes com a possível metabolização destas moléculas pelo CYP30E1 da vieira *N. nodosus*. No caso do antibacteriano eritromicina, a afinidade de ligação estimada foi de -10,1 kcal/mol. A pose predita foi bastante diferente da estrutura experimental, com o grupo desoamina da eritromicina mais próximo do grupo heme (Figura 12). Nesta pose, a molécula interage de forma hidrofóbica com os resíduos Asn121 e Leu218. Além disso, há formação de ponte de hidrogênio entre a Lys217 e a eritromicina.

Já com relação à bromoergocriptina, a pose predita foi bastante parecida com a estrutura experimental do CYP3A4 humano (Figura 13). A afinidade de ligação foi estimada em -11,9 kcal/mol. O resíduo Ala392 atua como aceptor em uma ponte de hidrogênio. Interações hidrofóbicas são observadas entre a molécula e os resíduos Leu218 e Phe109. Além disso, uma interação π ocorre entre a Arg394 e os anéis mais distantes do grupo heme (Figura 13).

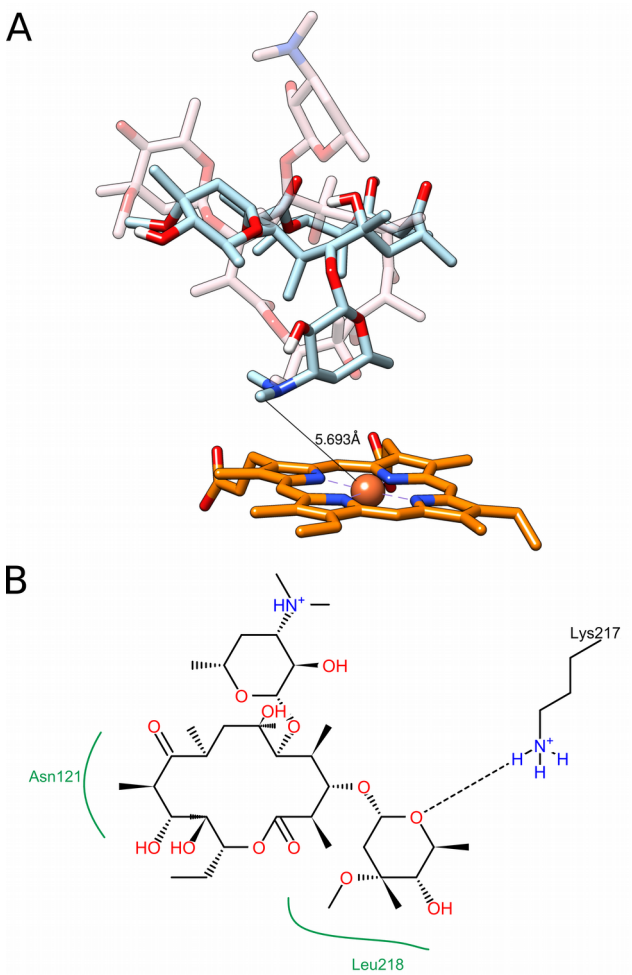


Figura 12 – Pose predita para a eritromicina no sítio ativo do CYP30E1 da vieira. A afinidade de ligação estimada foi de -10,1 kcal/mol. Em (A), comparação com a pose não-produtiva da estrutura experimental (rosa) e da pose predita pelo atracamento molecular (azul). Heme em laranja. A distância, do grupo metil com o átomo de ferro está marcada. Este grupo é removido durante a metabolização desta molécula. Imagem gerada pelo Chimera. Em (B), interações da eritromicina com resíduos do sítio ativo. Imagem gerada pelo PoseView (STIERAND; RAREY, 2010).

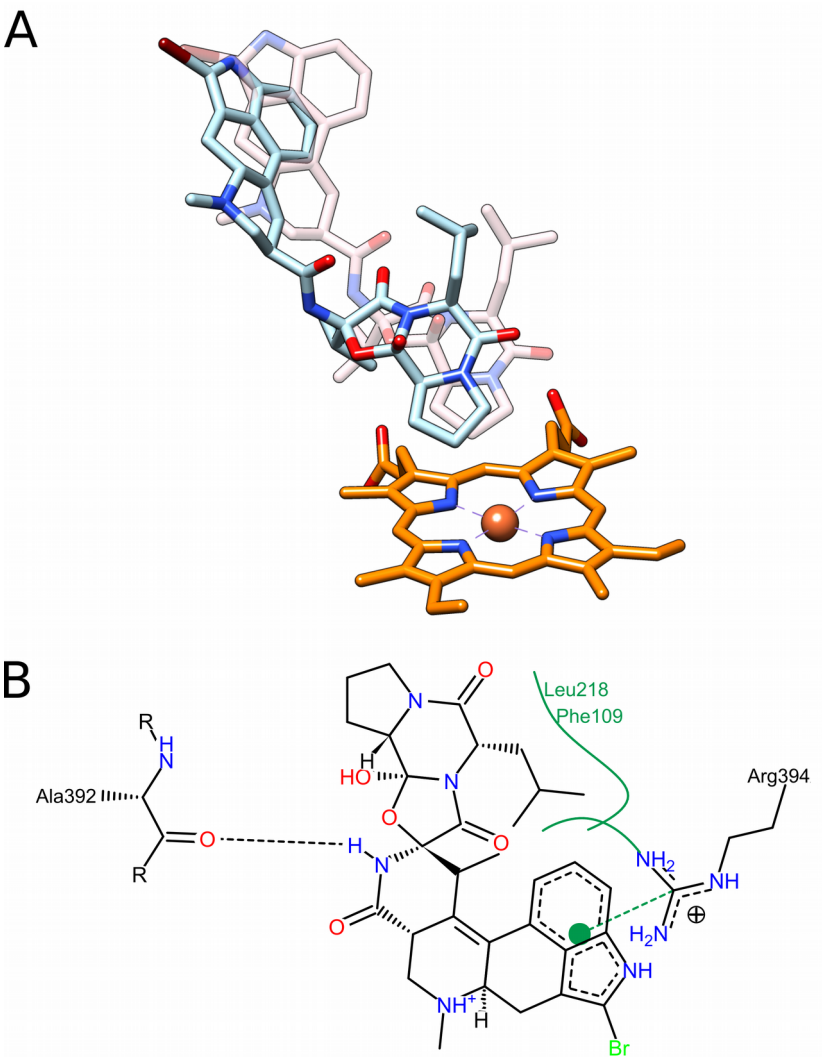


Figura 13 – Pose predita para a bromoergocriptina no sítio ativo do CYP30E1 da vieira. A afinidade de ligação estimada foi de -11,9 kcal/mol. Em (A), poses semelhantes para a estrutura experimental (rosa) e a pose predita (azul). Heme em laranja. Imagem gerada pelo Chimera. Em (B), interações da bromoergocriptina com resíduos do CYP30E1. Imagem gerada pelo PoseView.

O benzodiazepínico diazolam não apresentou uma pose semelhante a já observada no CYP3A4 humano (Figura 14). No CYP30E1 da vieira, o fármaco apresenta outra conformação, com afinidade estimada de -9,6 kcal/mol.

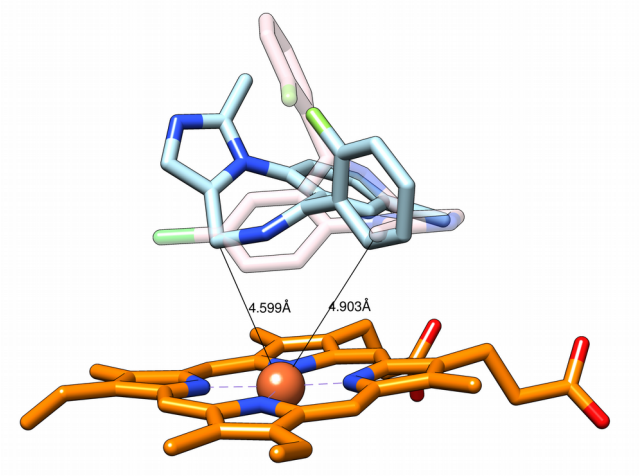


Figura 14 – Comparação entre a pose da estrutura experimental (rosa) e a pose predita (azul) para o fármaco midazolam. A afinidade de ligação estimada foi de -9,6 kcal/mol. A distâncias dos grupos que podem ser hidroxilados durante a metabolização desta molécula foram representadas, para as duas moléculas. Imagem gerada pelo Chimera.

5. DISCUSSÃO

5.1 IDENTIFICAÇÃO E CLASSIFICAÇÃO DAS SEQUÊNCIAS CYPs

Historicamente, os alinhamentos produzidos para sequências de CYPs foram ancorados pelos motivos WxxxR, a treonina do motivo ExxR e a cisteína do FxxGxxxCxG (SEZUTSU; LE GOFF; FEYEREISEN, 2013). Assim, a identificação de membros desta superfamília sempre utilizaram estas regiões invariáveis. Com o aumento no número de sequências CYPs a serem classificadas, especialmente de procariotos e plantas, diversas exceções foram sendo identificadas. Hoje, não existe nenhum resíduo invariável na superfamília dos CYPs (SEZUTSU; LE GOFF; FEYEREISEN, 2013).

A partir da publicação dos genomas dos primeiros invertebrados, como a *Drosophila melanogaster* e *Caenorhabditis elegans*, foi observada uma imensa diversidade de CYPs em invertebrados. Em *Drosophila melanogaster*, foram classificados 90 genes e em *Caenorhabditis Elegans* 84 genes (NELSON, 2009). Até 2013, 3.452 sequências CYPs haviam sido classificadas em insetos, 1.056 em mamíferos e 883 em invertebrados não-insetos (NELSON, 2009). Embora exista um esforço para a classificação de CYPs em invertebrados, ela está focada em insetos, grupo com grande diversificação nesta superfamília. Ênfase deve ser dada para a caracterização do grande número de CYPs em invertebrados marinhos, já que estas enzimas podem ser importante para a adaptação de animais frente a exposição aos xenobióticos (REWITZ et al., 2006).

Neste trabalho os CYPs identificados apresentam conservação nos motivos característicos deste grupo de proteínas (Figura 9). Com relação ao número de genes, estima-se que invertebrados não-insetos possuam, em média, de 76 a 89 CYPs em seus genomas (NELSON, 2011). O número baixo, 33, de CYPs identificados aqui é resultado do fato do transcriptoma ser exclusivo da glândula digestiva, além da baixa cobertura do sequenciamento. É importante notar que, dependendo da função biológica, CYPs podem ser expressos em diversos tecidos. Em moluscos, a glândula digestiva é o tecido que apresenta os maiores níveis de atividade na metabolização de xenobióticos (SOLÉ; LIVINGSTONE, 2005). Assim, os CYPs identificados aqui são aqueles

expressos na glândula digestiva da vieira, provavelmente possuindo funções relacionadas a este tecido. Em função desta limitação não é possível propor hipóteses sobre a diversificação dos CYPs nesta espécie.

Neste trabalho, 33 transcritos CYP foram classificados pelo Dr. David Nelson em diversas famílias, algumas inclusive inéditas. Esta foi a primeira vez que um CYP do clã 46 foi descrito em moluscos. Estas são descobertas importantes, pois possibilitam a identificação de membros destas famílias em outros taxa próximos. É importante notar que a diversidade observada na superfamília dos CYPs surgiu como resultado de sucessivas duplicações gênicas e subsequente divergência (NELSON, 1998).

Muitos dos CYPs que metabolizam xenobióticos possuem baixa especificidade de substratos, e seu nocaute geralmente não é letal (GONZALEZ; KIMURA, 2003). Dentre os CYPs capazes de metabolizar xenobióticos, a famíla CYP3 é bastante importante pois, em humanos, metaboliza aproximadamente 30% dos fármacos utilizados clinicamente (ZANGER; SCHWAB, 2013). O CYP3A4 humano, enzima mais importante neste aspecto, é bastante promíscua com relação a seus substratos, e pode acomodar estruturas de diversos tamanhos devido ao seu grande sítio ativo (SEVRIOUKOVA; POULOS, 2013). Curiosamente, a família CYP3 parece não estar presente em moluscos (NELSON; GOLDSTONE; STEGEMAN, 2013). Entretanto, a família CYP30, exclusiva de moluscos, apresenta alta semelhança com a família CYP3. Assim, cria-se a hipótese de que membros da família CYP30 possam agir no metabolismo de xenobióticos da mesma forma como os membros da família CYP3 o fazem, em vertebrados. O transcrito CYP30E1 possui alta expressão digital, considerando todos os transcritos avaliados, o que torna sua caracterização estrutural interessante. Além disso, ele apresentou todos os motivos característicos dos CYPs bem conservados, indicando que a proteína codificada possivelmente é funcional, já que diversos destes motivos são importantes na funcionalidade do grupo heme e enovelamento correto da enzima.

5.2 MODELAGEM ESTRUTURAL E ATRACAMENTO MOLECULAR

Os valores de Cscore e TM-Score, além dos resíduos em posições permitidas segundo o gráfico de Ramachandran demostram que os modelos gerados, para o CYP30E1, foram de alta qualidade (Tabela 10, Apêndice E). Estes resultados indicam que esta proteína apresenta, teoricamente, as características necessárias par assumir conformações que acomodam os seus respectivos ligantes.

Dentre os resíduos já identificados como importantes no metabolismo de substratos do CYP3A4, os resíduos Ile120, Asp214 e Leu373 atuam no metabolismo de midazolam (KHAN et al., 2002). Na vieira, estes foram substituídos por resíduos hidrofílicos (Asn122 e Thr395) e de carga contrária (Lys217) (Figura 10). Estas substituições, que modificam os parâmetros físico-químicos dos seus respectivos sítios, devem modificar a especificidade de substratos que esta enzima pode metabolizar. Segundo estudos computacionais, o resíduo Ser119, no CYP3A4 humano, é importante na estabilização de substratos e inibidores através de pontes de hidrogênio (EKROOS; SJOGREN, 2006). No CYP30E1, este sítio foi substituído por uma asparagina, que também é capaz de realizar este tipo de interação (Figura 10). Visto que 8 dos 11 resíduos não apresentam modificações significativas, o CYP30E1 da vieira *N. nodosus* deve apresentar ao menos uma sobreposição de substratos com o CYP3A4 humano. Entretanto, deve-se ter cautela em assumir funcionalidade em sequências similares pois uma única substituição, em regiões funcionalmente importantes, pode modificar a competência da enzima (CHEN; BERENBAUM; SCHULER, 2002).

Com relação aos resultados do atracamento molecular, a eritromicina apresentou uma pose diferente da estrutura experimental (2J0D) (Figura 12). A pose do cristal, não-produtiva, não apresenta interações polares. Entretanto, a proximidade de quatro fenilalaninas neste modelo sugere que o complexo é estabilizado parcialmente por interações hidrofóbicas (EKROOS; SJOGREN, 2006). A metabolização da eritromicina é realizada através da desmetilação do grupo desoamina. Na pose predita no complexo CYP30E1/eritromicina, este metil está bastante próximo do ferro do grupo heme, a somente 5,693 angstroms (Å) de distância (Figura 12). Isto indica uma pose produtiva, o que cria a hipótese de que a enzima da vieira, assim como a humana, é capaz de metabolizar este fármaco antibacteriano. Os resultados que dizem respeito à bromoergocriptina foram bastante semelhantes ao complexo

humano (Figura 13). SEVRIOUKOVA e POULOS (2012) apresentaram a estrutura cristalográfica do complexo CYP3A4/bromoergocriptina. Neste trabalho, os dados cristalográficos sugerem que os aminoácidos Arg212 e Thr224 são importantes na ligação desta molécula. Além disso, foram observados contatos com os resíduos Ile301, Phe304, Ala305, Arg105, Arg212, Ala370 e Arg372 (SEVRIOUKOVA; POULOS, 2012). Apesar do CYP30E1 apresentar os resíduos Arg212 e Thr224 substituídos por isoleucinas, hidrofóbicas, a pose predita foi bastante semelhante, indicando a capacidade de metabolização desta molécula. Entretanto, devido às modificações físico-químicas nestes dois sítios, é importante considerar se estas não impedem ou dificultam a entrada deste fármaco no sítio ativo da enzima. Já o midazolam foi acomodado no sítio ativo em uma posição contrária ao complexo apresentado na estrutura experimental do CYP3A4 (Figura 14). Sabe-se que os resíduos Ile120, Asp214 e Leu373 são considerados importantes no metabolismo de midazolam (KHAN et al., 2002). Portanto, devido às modificações físico-químicas presentadas nestes sítios já descritas na Figura 10, espera-se que o CYP30E1 não seja capaz de metabolizar, da mesma forma, este fármaco. A metabolização do midazolam pode ocorrer através de hidroxilação de dois grupos, produzindo 1-hidroximidazolam ou 4-hidroximidazolam. O CYP3A4 produz primariamente 1-hidroximidazolam, apresentando o grupo hidroxilado próximo ao ferro do grupo heme (Figura 14). No caso do CYP30E1, da vieira, a pose gerada pelo atracamento molecular apresenta o grupo que, quando hidroxilado, forma a molécula 4-hidroximidazolam (Figura 14). Assim, os resultados do atracamento molecular, para esta molécula, sugerem que este CYP metaboliza primariamente midazolam em 4-hidroximidazolam, ao contrário do CYP3A4 humano.

6. CONCLUSÕES

Neste capítulo foram identificados e classificados os CYPs identificados no transcriptoma de glândula digestiva de *N. nodosus* e uma destas isoformas, a proteína CYP30E1, foi caracterizada estruturalmente, através de métodos computacionais. Os resultados aqui apresentados representam o primeiro esforço na classificação de CYPs nesta espécie. Seis novas famílias foram descobertas. Esta foi a primeira vez que um CYP do clã 46 foi identificado em moluscos. Além disso, são uma contribuição na caracterização da diversidade que os invertebrados não-insetos apresentam nesta superfamília. Considerando a falta de estudos na caracterização funcional de membros da família CYP30, os resultados aqui apresentados revelam indícios em relação ao mecanismo de ação destas proteínas, na vieira.

CONSIDERAÇÕES FINAIS

- Este foi o primeiro transcriptoma, gerado a partir do sequenciamento RNA-Seq de glândula digestiva, da vieira *N. nodosus*. Grande esforço foi colocado na montagem adequada das leituras obtidas, e na anotação dos contigs. Estes resultados geram um grande avanço nas informações nucleotídicas já produzidas nesta espécie, o que possibilita novos estudos em diversas disciplinas.

- Esta dissertação também apresentou a primeira classificação dos transcritos CYPs da glândula digestiva da vieira *N. nodosus*. Os resultados revelam CYPs de diversas famílias, algumas até então desconhecidas. Considerando o baixo número de publicações sobre a família CYP30, estudos deste tipo são extremamente importantes.

- A caracterização do transcrito CYP30E1 revelou motivos e um enovelamento conservado evolutivamente. Em relação ao metabolismo de alguns substratos, a comparação realizada com a enzima CYP3A4 humana revelou algumas semelhanças entre as duas enzimas.

REFERÊNCIAS

ALBERTS, I. L.; TODOROV, N. P.; DEAN, P. M. Receptor flexibility in de novo ligand design and docking. **Journal of Medicinal Chemistry**, v. 48, n. 21, p. 6585–6596, 2005.

AMERICO, J. A. et al. Gene discovery in the tropical scallop Nodipecten nodosus: Construction and sequencing of a normalized cDNA library. **Marine Environmental Research**, v. 91, p. 34–40, 2015.

ANDREW WILLIAMS, J. et al. Comparative metabolic capabilities of CYP3A4, CYP3A5, and CYP3A7. **Drug Metabolism and Disposition**, v. 30, n. 8, p. 883–891, 2002.

APWEILER, R. et al. UniProt: the Universal Protein knowledgebase. **Nucleic acids research**, v. 32, n. Database issue, p. D115-9, 2004.

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

BAINY, A. C. D. et al. Biochemical responses in farmed mussel Perna perna transplanted to contaminated sites on Santa Catarina Island, SC, Brazil. **Marine Environmental Research**, v. 50, n. 1–5, p. 411–416, 2000.

BENDZ, D. et al. Occurrence and fate of pharmaceutically active compounds in the environment, a case study: Höje River in Sweden. **Journal of Hazardous Materials**, v. 122, n. 3, p. 195–204, 2005.

BERMAN, H. M. et al. The Protein Data Bank. **Nucl. Acids Res.**, v. 28, n. 1, p. 235–242, 2000.

BICKHAM, J. W. et al. Effects of chemical contaminants on genetic diversity in natural populations: implications for biomonitoring and

ecotoxicology. **Mutation research/Reviews in Mutation research**, v. 463, n. 1, p. 33-51, 2000.

BOLGER, A. M.; LOHSE, M.; USADEL, B. Trimmomatic: A flexible trimmer for Illumina sequence data. **Bioinformatics**, v. 30, n. 15, p. 2114–2120, 2014.

BRAMMELL, B. F. et al. Apparent lack of CYP1A response to high PCB body burdens in fish from a chronically contaminated PCB site. **Marine environmental research**, v. 58, n. 2, p. 251-255, 2004.

BRAUSCH, J. M. et al. **Reviews of Environmental Contamination and Toxicology. Volume 218**. [s.l: s.n.]. v. 218

BROOIJMANS, N.; KUNTZ, I. D. Molecular recognition and docking algorithms. **Annual review of biophysics and biomolecular structure**, v. 32, p. 335–373, 2003.

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

BULTELLE, F. et al. Identification of differentially expressed genes in Dreissena polymorpha exposed to contaminants. **Marine Environmental Research**, v. 54, n. 3–5, p. 385–389, 2002.

CAIXAETA, E. et al. Tipos de marcadores moleculares. In: BORÉM, A.; CAIXETA, E. (Eds.). **Marcadores moleculares**. Viçosa – MG: Editora Folha de Viçosa, 2009. p. 11-94.

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

CHEN, H. et al. Sequencing and de Novo assembly of the Asian clam (Corbicula fluminea) transcriptome using the illumina GAIIx method. **PLoS ONE**, v. 8, n. 11, p. 1–12, 2013.

CHEN, J. S.; BERENBAUM, M. R.; SCHULER, M. A. Amino acids in SRS1 and SRS6 are critical for furanocoumarin metabolism by CYP6B1v1, a cytochrome P450 monooxygenase. **Insect Molecular Biology**, v. 11, n. 2, p. 175–186, 2002.

CROOKS, G. et al. WebLogo: a sequence logo generator. **Genome Res**, v. 14, p. 1188–1190, 2004.

CSERMELY, P.; PALOTAI, R.; NUSSINOV, R. Induced fit, conformational selection and independent dynamic segments: An extended view of binding events. **Trends in Biochemical Sciences**, v. 35, n. 10, p. 539–546, 2010.

DANECEK, P. et al. The variant call format and VCFtools. **Bioinformatics**, v. 27, n. 15, p. 2156–2158, 2011.

DE MONTELLANO, P. R. O. Cytochrome P450: Structure, mechanism, and biochemistry: Third edition. **Cytochrome P450: Structure, Mechanism, and Biochemistry: Third edition**, p. 1–689, 2005.

DEBLONDE, T.; COSSU-LEGUILLE, C.; HARTEMANN, P. Emerging pollutants in wastewater: A review of the literature. **International Journal of Hygiene and Environmental Health**, v. 214, n. 6, p. 442–448, 2011.

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

EKBLOM, R.; GALINDO, J. Applications of next generation sequencing in molecular ecology of non-model organisms. **Heredity**, v. 107, n. 1, p. 1–15, 2010.

EKROOS, M.; SJOGREN, T. Structural basis for ligand promiscuity in cytochrome P450 3A4. **Proceedings of the National Academy of Sciences**, v. 103, n. 37, p. 13682–13687, 2006.

ERICKSON, J. A. et al. Lessons in Molecular Recognition: The Effects of Ligand and Protein Flexibility on Molecular Docking Accuracy. **Journal of Medicinal Chemistry**, v. 47, n. 1, p. 45–55, 2004.

FERRARI, A. M. et al. Soft docking and multiple receptor conformations in virtual screening. **Journal of Medicinal Chemistry**, v. 47, n. 21, p. 5076–5084, 2004.

FINN, R. D.; CLEMENTS, J.; EDDY, S. R. HMMER web server: Interactive sequence similarity searching. **Nucleic Acids Research**, v. 39, n. SUPPL. 2, p. 1–9, 2011.

FISCHER, E. Einfluss der Configuration auf die Wirkung der Enzyme. **Ber. Dtsch. Chem. Ges.**, v. 27, p. 2985–2993, 1894.

FOWLER, S. M. et al. Amino acid 305 determines catalytic center accessibility in CYP3A4. **Biochemistry**, v. 39, n. 15, p. 4406–4414, 2000.

FOWLER, S. M. et al. CYP3A4 active site volume modification by mutagenesis of leucine 211. **Drug Metabolism and Disposition**, v. 30, n. 4, p. 452–456, 2002.

FRAUENFELDER, H.; SLIGAR, S. G.; WOLYNES, P. G. The energy landscapes and motions of proteins. **Science (New York, N.Y.)**, v. 254, n. 5038, p. 1598–603, 1991.

GONZALEZ, F. J.; KIMURA, S. Study of P450 function using gene knockout and transgenic mice. **Archives of Biochemistry and Biophysics**, v. 409, n. 1, p. 153–158, 2003.

GRABHERR, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. **Nature biotechnology**, v. 29, n. 7, p. 644–52, 2011.

HALLING-SORENSEN, B. et al. Occurence, fate and effects of pharmaceuticals substance in the environment - A review. **Chemosphere**, v. 36, n. 2, p. 357–393, 1998.

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

HALPERT, J. R. Anthony Y. H. Lu Commemorative Issue Preface. **Drug metabolism and disposition: the biological fate of chemicals**, v. 26, n. 12, p. 1167, 1998.

HANNAM, M. L. et al. Immune function in the Arctic Scallop, Chlamys islandica, following dispersed oil exposure. **Aquatic Toxicology**, v. 92, n. 3, p. 187–194, 2009.

HOLLEY, R. W. et al. Structure of a ribonucleic acid. **Science (New York, N.Y.)**, v. 147, n. 3664, p. 1462–5, 1965.

HOLSBEEK, L. et al. Heavy metals, organochlorines and polycyclic aromatic hydrocarbons in sperm whales stranded in the southern North Sea during the 1994/1995 winter. **Marine Pollution Bulletin**, v. 38, n. 4, p. 304–313, 1999.

HOU, R. et al. Transcriptome sequencing and De Novo analysis for Yesso Scallop (Patinopecten yessoensis) using 454 GS FLX. **PLoS ONE**, v. 6, n. 6, 2011.

JACKSON, J. B. C. et al. Historical Overfishing and the Recent Collapse of Coastal Ecosystems. **Science**, v. 293, n. 5530, p. 629–638, 2001.

JIN, Q. et al. RNA-seq based on transcriptome reveals differ genetic expressing in Chlamys farreri exposed to carcinogen PAHs. **Environmental Toxicology and Pharmacology**, v. 39, n. 1, p. 313–320, 2015.

KHAN, K. K. et al. Midazolam oxidation by cytochrome P450 3A4 and active-site mutants: an evaluation of multiple binding sites and of the metabolic pathway that leads to enzyme inactivation. **Mol Pharmacol**, v. 61, n. 3, p. 495–506, 2002.

KIM, S. D. et al. Occurrence and removal of pharmaceuticals and endocrine disruptors in South Korean surface, drinking, and waste waters. **Water Research**, v. 41, n. 5, p. 1013–1021, 2007.

KITCHEN, D. B. et al. Docking and Scoring in Virtual Screening for Drug Discovery: MEthods and Applications. **Nat. Rev. Drug Disc.**, v. 3, n. 11, p. 935–949, 2004.

KLINGENBERG, M. Pigments of rat liver microsomes. **Archives of Biochemistry and Biophysics**, v. 75, n. 2, p. 376–386, 1958.

KOSHLAND, D. E. Application of a Theory of Enzyme Specificity to Protein Synthesis. **Proceedings of the National Academy of Sciences of the United States of America (PNAS)**, v. 44, n. 2, p. 98–104, 1958.

KOSKA, J. et al. Fully automated molecular mechanics based induced fit protein - Ligand docking method. **Journal of Chemical Information and Modeling**, v. 48, n. 10, p. 1965–1973, 2008.

KUCKLICK, J. R. et al. Persistent organochlorine pollutants in ringed seals and polar bears collected from northern Alaska. **Science of the Total Environment**, v. 287, n. 1–2, p. 45–59, 2002.

KUNTZ, I. D. et al. A geometric approach to macromolecule-ligand interactions. **Journal of Molecular Biology**, v. 161, n. 2, p. 269–288, 1982.

LABORDA, P. **Marcadores moleculares microssatélites na investigação do genoma de *Drosophila mediopunctata*: desenvolvimento e construção de mapa genético de ligação**. Universidade Estadual de Campinas – [S.1.] 2011.

LANGMEAD, B.; SALZBERG, S. L. Fast gapped-read alignment with Bowtie 2. **Nat Methods**, v. 9, n. 4, p. 357–359, 2012.

LI, B.; DEWEY, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. **BMC bioinformatics**, v. 12, n. 1, p. 323, 2011.

LI, H. et al. The Sequence Alignment/Map format and SAMtools. **Bioinformatics**, v. 25, n. 16, p. 2078–2079, 2009.

LI, H.; DURBIN, R. Fast and accurate short read alignment with Burrows-Wheeler transform. **Bioinformatics**, v. 25, n. 14, p. 1754–1760, 2009.

LITT, M.; LUTY, J. A. A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. **American journal of human genetics**, v. 44, n. 3, p.397, 1989.

LIU, N. et al. Application of the biomarker responses in scallop (Chlamys farreri) to assess metals and PAHs pollution in Jiaozhou

Bay, China. **Marine Environmental Research**, v. 80, p. 38–45, 2012.

LOTZE, H. K. et al. Depletion, Degredation, and Recovery Potential of Estuaries and Coastal Seas. **Science**, v. 312, n. 5781, p. 1806–1809, 2006.

LOVELL, S. C. et al. Structure validation by Cα geometry: φ,ψ and Cβ deviation. **Proteins: Structure, Function and Genetics**, v. 50, n. 3, p. 437–450, 2003.

LÜCHMANN, K. H. et al. Key metabolic pathways involved in xenobiotic biotransformation and stress responses revealed by transcriptomics of the mangrove oyster Crassostrea brasiliana. **Aquatic Toxicology**, v. 166, p. 10–20, 2015.

LUKKARI, T. et al. Biomarker responses of the earthworm Aporrectodea tuberculata to copper and zinc exposure: differences between populations with and without earlier metal exposure. **Environmental Pollution**, v. 129, n. 3, p. 377-386, 2004.

MARGULIES, M. et al. Genome sequencing in microfabricated high-density picolitre reactors. **Nature**, v. 437, n. 7057, p. 376–80, 2005.

MOROZOVA, O.; HIRST, M.; MARRA, M. A. Applications of new sequencing technologies for transcriptome analysis. **Annu Rev Genomics Hum Genet**, v. 10, p. 135–151, 2009.

NARASIMHAN, V. et al. BCFtools/RoH: A hidden Markov model approach for detecting autozygosity from next-generation sequencing data. **Bioinformatics**, v. 32, n. 11, p. 1749–1751, 2016.

NEBERT, D. W.; WIKVALL, K.; MILLER, W. L. Human cytochromes P450 in health and disease. **Philosophical transactions**

of the Royal Society of London. Series B, Biological sciences, v. 368, n. 1612, p. 20120431, 2013.

NELSON, D. et al. The P450 Superfamily: Update on New Sequences, Gene Mappings, Accession Numbers, Early Trivial Names of Enzymes, and Nomenclature. **DNA and Cell Biology**, v. 12, n. 1, p. 1–51, 1993.

NELSON, D. R. Metazoan cytochrome P450 evolution. **Comparative Biochemistry and Physiology - C Pharmacology Toxicology and Endocrinology**, v. 121, n. 1–3, p. 15–22, 1998.

NELSON, D. R. et al. Comparison of cytochrome P450 (CYP) genes from the mouse and human genomes, including nomenclature recommendations for genes, pseudogenes and alternative-splice variants. **Pharmacogenetics**, v. 14, n. 1, p. 1–18, 2004.

NELSON, D. R. **Cytochrome P450 Nomenclature, 2004Cytochrome P450 Protocols**, 2006. Disponível em: <http://link.springer.com/10.1385/1-59259-998-2:1>

NELSON, D. R. The cytochrome p450 homepage. **Human genomics**, v. 4, n. 1, p. 59–65, 2009.

NELSON, D. R. Progress in tracing the evolutionary paths of cytochrome P450. **Biochimica et Biophysica Acta - Proteins and Proteomics**, v. 1814, n. 1, p. 14–18, 2011.

NELSON, D. R. A world of cytochrome P450s. **Philosophical transactions of the Royal Society of London. Series B, Biological sciences**, v. 368, n. 1612, p. 20120430, 2013.

NELSON, D. R.; GOLDSTONE, J. V; STEGEMAN, J. J. The cytochrome P450 genesis locus: the origin and evolution of animal cytochrome P450s. **Philosophical transactions of the Royal Society**

of London. Series B, Biological sciences, v. 368, n. 1612, p. 20120474, 2013.

OF, E.; GESTAGENS, S.; FISH, O. N. Pharmaceuticals and Personal Care Products in the Environment. **Environmental Toxicology**, v. 28, n. 12, p. 2663–2670, 2009.

PAIRETT, A. N.; SERB, J. M. De Novo Assembly and Characterization of Two Transcriptomes Reveal Multiple Light-Mediated Functions in the Scallop Eye (Bivalvia: Pectinidae). **PLoS ONE**, v. 8, n. 7, 2013.

PAN, B. et al. De novo RNA-seq analysis of the venus clam, cyclina sinensis, and the identification of immune-related genes. **PLoS ONE**, v. 10, n. 4, p. e0123296, 2015.

PAN, L. et al. Identification of a novel P450 gene belonging to the CYP4 family in the clam Ruditapes philippinarum, and analysis of basal- and benzo(a)pyrene-induced mRNA expression levels in selected tissues. **Environmental Toxicology and Pharmacology**, v. 32, n. 3, p. 390–398, 2011.

PAN, L. Q.; REN, J.; LIU, J. Responses of antioxidant systems and LPO level to benzo(a)pyrene and benzo(k)fluoranthene in the haemolymph of the scallop Chlamys ferrari. **Environmental Pollution**, v. 141, n. 3, p. 443–451, 2006.

PÉREZ-CADAHÍA, B. et al. Evaluation of PAH bioaccumulation and DNA damage in mussels (Mytilus galloprovincialis) exposed to spilled Prestige crude oil. **Comparative Biochemistry and Physiology - C Toxicology and Pharmacology**, v. 138, n. 4, p. 453–460, 2004.

PETTERSEN, E. F. et al. UCSF Chimera - A visualization system for exploratory research and analysis. **Journal of Computational Chemistry**, v. 25, n. 13, p. 1605–1612, 2004.

PIAZZA, R. S. et al. Exposure to phenanthrene and depuration: Changes on gene transcription, enzymatic activity and lipid peroxidation in gill of scallops Nodipecten nodosus. **Aquatic Toxicology**, v. 177, p. 146–155, 2016.

RAMSAY, L. et al. A simple sequence repeat-based linkage map of Barley. **Genetics**, v. 156, n. 4, p. 1997–2005, 2000.

REWITZ, K. F. et al. Marine invertebrate cytochrome P450: Emerging insights from vertebrate and insect analogies. **Comparative Biochemistry and Physiology - C Toxicology and Pharmacology**, v. 143, n. 4, p. 363–381, 2006.

ROBERTS, P. H.; THOMAS, K. V. The occurrence of selected pharmaceuticals in wastewater effluent and surface waters of the lower Tyne catchment. **Science of the Total Environment**, v. 356, n. 1–3, p. 143–153, 2006.

SAKATSUME, O. et al. Absolute mRNA quantification using the polymerase chain reaction (PCR). A novel approach by a PCR aided transcript titration assay (PATTY). **Nucleic Acids Research**, v. 17, n. 1, p. 3689–3697, 1989.

SANDAK, B.; WOLFSON, H. J.; NUSSINOV, R. Flexible docking allowing induced-fit in proteins: Insights from an open to closed isomers. **Proteins: Struct. Funct. Genet.**, v. 32, n. January, p. 159–174, 1998.

SANGER, F.; NICKLEN, S.; COULSON, A R. DNA sequencing with chain-terminating inhibitors. **Proceedings of the National**

Academy of Sciences of the United States of America, v. 74, n. 12, p. 5463–7, 1977.

SARKAR, A. et al. Molecular Biomarkers: Their significance and application in marine pollution monitoring. **Ecotoxicology**, v. 15, n. 4, p. 333–340, 2006.

SCHENA, M. et al. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. **Science (New York, N.Y.)**, v. 270, n. 5235, p. 467–70, 1995.

SCHIRMER, K. et al. Transcriptomics in ecotoxicology. **Analytical and Bioanalytical Chemistry**, v. 397, n. 3, p. 917–923, 2010.

SCHLEDER, D. D. et al. Evaluation of hemato-immunological parameters during the reproductive cycle of the scallop Nodipecten nodosus in association with a carotenoid-enriched diet. **Aquaculture**, v. 280, n. 1–4, p. 256–263, 2008.

SCHULZ, M. H. et al. Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels. **Bioinformatics**, v. 28, n. 8, p. 1086–1092, 2012.

SEVRIOUKOVA, I. F.; POULOS, T. L. Structural and mechanistic insights into the interaction of cytochrome P4503A4 with bromoergocryptine, a type I ligand. **Journal of Biological Chemistry**, v. 287, n. 5, p. 3510–3517, 2012.

SEVRIOUKOVA, I. F.; POULOS, T. L. Understanding the mechanism of cytochrome P450 3A4: recent advances and remaining problems. **Dalton Transactions**, v. 40, n. 12, p. 3116–3126, 2013.

SEZUTSU, H.; LE GOFF, G.; FEYEREISEN, R. Origins of P450 diversity. **Philosophical transactions of the Royal Society of**

London. Series B, Biological sciences, v. 368, n. 1612, p. 20120428, 2013.

SIMÃO, F. A. et al. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. **Bioinformatics**, v. 31, n. 19, p. 3210–3212, 2015.

SMITH-UNNA, R. et al. TransRate: Reference-free quality assessment of de novo transcriptome assemblies. **Genome Research**, v. 26, n. 8, p. 1134–1144, 2016.

SOLÉ, M.; BUET, A.; ORTIZ, L. Bioaccumulation and biochemical responses in mussels exposed to the water-accommodated fraction of the Prestige fuel oil. **Scientia Marina**, v. 71, n. June, p. 373–382, 2007.

SOLÉ, M.; LIVINGSTONE, D. R. Components of the cytochrome P450-dependent monooxygenase system and “NADPH-independent benzo[a]pyrene hydroxylase” activity in a wide range of marine invertebrate species. **Comparative Biochemistry and Physiology - C Toxicology and Pharmacology**, v. 141, n. 1, p. 20–31, 2005.

SOLÉ, M.; PORTE, C.; ALBAIGÉS, J. Hydrocarbons, PCBs and DDT in the NW Mediterranean deep-sea fish Mora moro. **Deep-Sea Research Part I: Oceanographic Research Papers**, v. 48, n. 2, p. 495–513, 2001.

SONNHAMMER, E. L. L.; EDDY, S. R.; DURBIN, R. Pfam: A comprehensive database of protein domain families based on seed alignments. **Proteins: Structure, Function and Genetics**, v. 28, n. 3, p. 405–420, 1997.

STIERAND, K.; RAREY, M. PoseView -- molecular interaction patterns at a glance. **Journal of Cheminformatics**, v. 2, n. Suppl 1, p. P50, 2010.

SUTHERLAND, G. R.; RICHARDS, R. I. Simple tandem DNA repeats and human genetic disease. **Proceedings of the National Academy of Sciences of the United States of America**, v. 92, n. 9, p. 3636–3641, 1995.

TERNES, T. A. Occurrence of drugs in German sewage treatment plants and rivers. **Water Research**, v. 32, n. 11, p. 3245–3260, 1998.

THIEL, T. et al. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.). **Theoretical and Applied Genetics**, v. 106, n. 3, p. 411–422, 2003.

TROTT, O.; OLSON, A. J. Software news and update AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. **Journal of Computational Chemistry**, v. 31, n. 2, p. 455–461, 2010.

WALKER, S. D.; MCELDOWNEY, S. Molecular docking: A potential tool to aid ecotoxicity testing in environmental risk assessment of pharmaceuticals. **Chemosphere**, v. 93, n. 10, p. 2568–2577, 2013.

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

WERCK-REICHHART, D.; FEYEREISEN, R. Cytochromes P450: a success story. **Genome biology**, v. 1, n. 6, p. REVIEWS3003, 2000.

WESTERLUND, F.; BJØRNHOLM, T. Directed assembly of gold nanoparticles. **Current Opinion in Colloid & Interface Science**, v. 14, n. 2, p. 126–134, 2009.

WU, B. et al. In silico predication of nuclear hormone receptors for organic pollutants by homology modeling and molecular docking. **Toxicology Letters**, v. 191, n. 1, p. 69–73, 2009.

WU, B. et al. Computational studies of interactions between endocrine disrupting chemicals and androgen receptor of different vertebrate species. **Chemosphere**, v. 80, n. 5, p. 535–41, 2010.

YANG, J. et al. The I-TASSER Suite: Protein structure and function prediction. **Nature Methods**, v. 12, n. 1, p. 7–8, 2015.

YANG, W. et al. Molecular docking and comparative molecular similarity indices analysis of estrogenicity of polybrominated diphenyl ethers and their analogues. **Environmental Toxicology and Chemistry**, v. 29, n. 3, p. 660–668, 2010.

ZANGER, U. M.; SCHWAB, M. Cytochrome P450 enzymes in drug metabolism: Regulation of gene expression, enzyme activities, and impact of genetic variation. **Pharmacology and Therapeutics**, v. 138, n. 1, p. 103–141, 2013.

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.

ZHANG, G. et al. The oyster genome reveals stress adaptation and complexity of shell formation. **Nature**, v. 490, n. 7418, p. 49–54, 2012.

ZHAO, Q.-Y. et al. Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study. **BMC Bioinformatics**, v. 12, n. Suppl 14, p. S2, 2011.

ZHENG, J. et al. Identification of two isoforms of CYP4 in Marsupenaeus japonicus and their mRNA expression profile response to benzo[a]pyrene. **Marine Environmental Research**, v. 112, p. 96–103, 2015.

APÊNDICE A – Programas e bancos de dados públicos, junto com suas respectivas versões, utilizados nas análises apresentadas no capítulo I da presente dissertação.

Programas e versões utilizados.

Programa	Versão
FastQC	0.10.1
Trimmomatic	0.33
Trinity	2.0.6
Velvet	1.2.10
Oases	0.2.8
RSEM	1.2.20
Bowtie	2.1.0
Transrate	1.0.0-beta3
BUSCO	1.1b1
Transdecoder	2.0.1
BLAST+	2.2.30
HMMER	3.1b2
WEGO	Versão não definida
MISA	1.0
BWA	0.7.12-r1039
SAMtools	0.1.19-96b5f2294a
BCFtools	0.1.19-96b5f2294a
VCFtools	0.1.11

Bancos de dados públicos e versões utilizados.

Banco de dados	Versão
NCBInr	Julho, 2015
UniProtKB-TrEMBL	Abril, 2015
UniProtKB-SwissProt	Abril, 2015

Pfam-A	27.0
Gene Ontology	Abril, 2015

APÊNDICE B – Todas as métricas geradas para a avaliação de qualidade das três estratégias de montagem *de novo* testadas.

Montagem	Trinity (k = 25)	Velvet (k = 25)	Velvet (k = 45)
n_seqs	94.074	69.062	76.861
smallest	200	200	200
largest	19.173	18.412	21.786
n_bases	92.284.530	94.660.450	96.511.712
mean_len	980,978	1.370,659	1.255,666
n_under_200	0	0	0
n_over_1k	28.866	31.487	31.504
n_over_10k	38	69	49
n_with_orf	17.654	18.964	20.220
mean_orf_percent	39,088	36,469	37,774
n90	308	621	509
n70	1.192	1.559	1.438
n50	2.051	2.408	2.310
n30	3.086	3.502	3.364
n10	5.147	5.681	5.330
gc	0,369	0,367	0,370
gc_skew	0,012	0,003	0,003
at_skew	0,005	0,004	0,004
cpg_ratio	1,494	1,509	1,494
bases_n	0	63.371	5.401
proportion_n	0	6,690E-004	5,600E-005
linguistic_complexity	0,155	0,213	0,197
fragments	127.417.512	127.417.512	127.417.512
fragments_mapped	118.301.136	114.707.458	114.602.820
p_fragments_mapped	0,928	0,900	0,899
good_mappings	108.250.832	104.474.768	103.909.975
p_good_mapping	0,850	0,820	0,816

bad_mappings	10.050.304	10.232.690	10.692.845
potential_bridges	43.317	25.865	28.456
bases_uncovered	2.395.886	3.420.125	3.432.191
p_bases_uncovered	0,026	0,036	0,036
contigs_uncovbase	35.020	37.225	29.984
p_contigs_uncovbase	0,372	0,539	0,390
contigs_uncovered	3.673	1.410	1.185
p_contigs_uncovered	0,039	0,020	0,015
contigs_lowcovered	46.856	31.839	37.394
p_contigs_lowcovered	0,498	0,461	0,487
contigs_segmented	20.782	18.599	18.837
p_contigs_segmented	0,221	0,269	0,245
score	0,259	0,282	0,302
optimal_score	0,338	0,326	0,328
cutoff	0,031	0,026	0,032

APÊNDICE C – Programas e bancos de dados públicos, junto com suas respectivas versões, utilizados nas análises apresentadas no capítulo II da presente dissertação.

Programas e versões utilizados.

Programa	Versão
HMMER	3.1b2
MUSCLE	3.8.31
Jalview	2.7
I-TASSER	5.0
UCSF Chimera	1.11.2
RAMPAGE	Versão não definida
MGLTools	1.5.6
AutoDock Vina	1.1.2
PoseView	Versão não definida

Bancos de dados públicos e versões utilizados.

Banco de dados	Versão
Pfam-A	27.0

APÊNDICE D – Parâmetros da caixa para o atracamento molecular e resíduos flexíveis utilizados nos atracamentos dos substratos do CYP3A4 humano, através do programa AutoDock Vina.

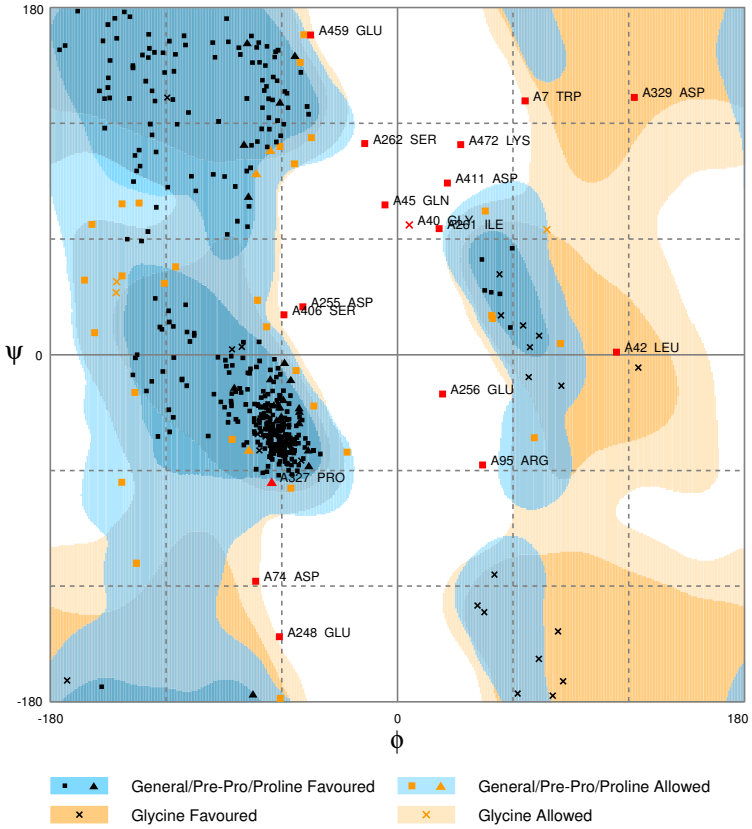
m2J0D:
center_x = 63.598
center_y = 73.034
center_z = 58.345
size_x = 24.209
size_y = 23.047
size_z = 27.033
ILE107_VAL108_PHE109_LYS110_ASN121_ASN122_ILE215_LEU216_LYS217_LEU218_PHE243_LEU244_VAL324_PHE327_ALA328_THR332_ALA392_THR393_ASP396

m3UA1:
center_x = 62.442
center_y = 74.927
center_z = 58.521
size_x = 28.628
size_y = 27.663
size_z = 28.299
PHE58_ARG106_ILE107_VAL108_PHE109_ASN121_ASN122_LEU216_LYS217_LEU218_ILE223_ILE224_LEU226_ILE227_VAL324_PHE327_ALA328_PRO391_ALA392_ASP396

m5TE8:
center_x = 62.642
center_y = 72.733
center_z = 60.049
size_x = 22.992
size_y = 18.068
size_z = 29.321
ILE107_VAL108_PHE109_ASN121_ASN122_LEU218_ALA219_SER220_PHE327_ALA328_GLU331_THR332_PRO391_ALA392_THR393_ARG394_LEU502

APÊNDICE E – Gráficos de Ramachandran para os diferentes estados conformacional produzidos para o CYP30E1 da vieira *N. nodosus*.

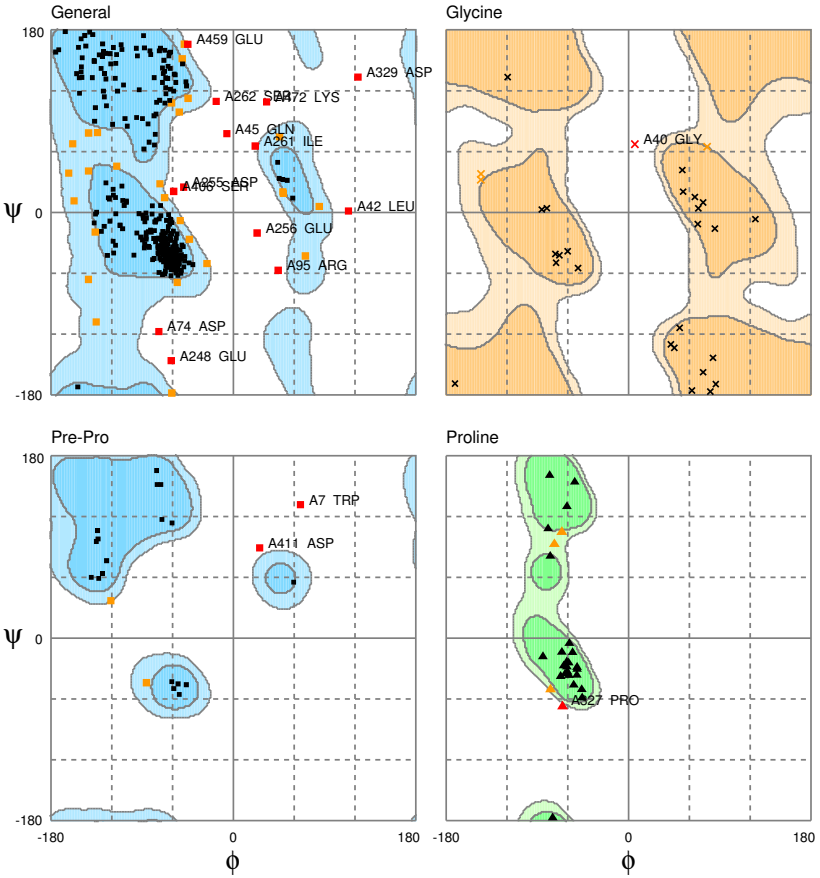
m2J0D:



Number of residues in favoured region (~98.0% expected)	:	425 (89.1%)
Number of residues in allowed region (~2.0% expected)	:	35 (7.3%)
Number of residues in outlier region	:	17 (3.6%)

RAMPAGE by Paul de Bakker and Simon Lovell available at http://www-cryst.bioc.cam.ac.uk/rampage/

Please cite: S.C. Lovell, I.W. Davis, W.B. Arendall III, P.I.W. de Bakker, J.M. Word, M.G. Prisant, J.S. Richardson & D.C. Richardson (2002) Structure validation by Cα geometry: ϕ/ψand Cβ deviation. *Proteins: Structure, Function & Genetics.* **50**: 437-450

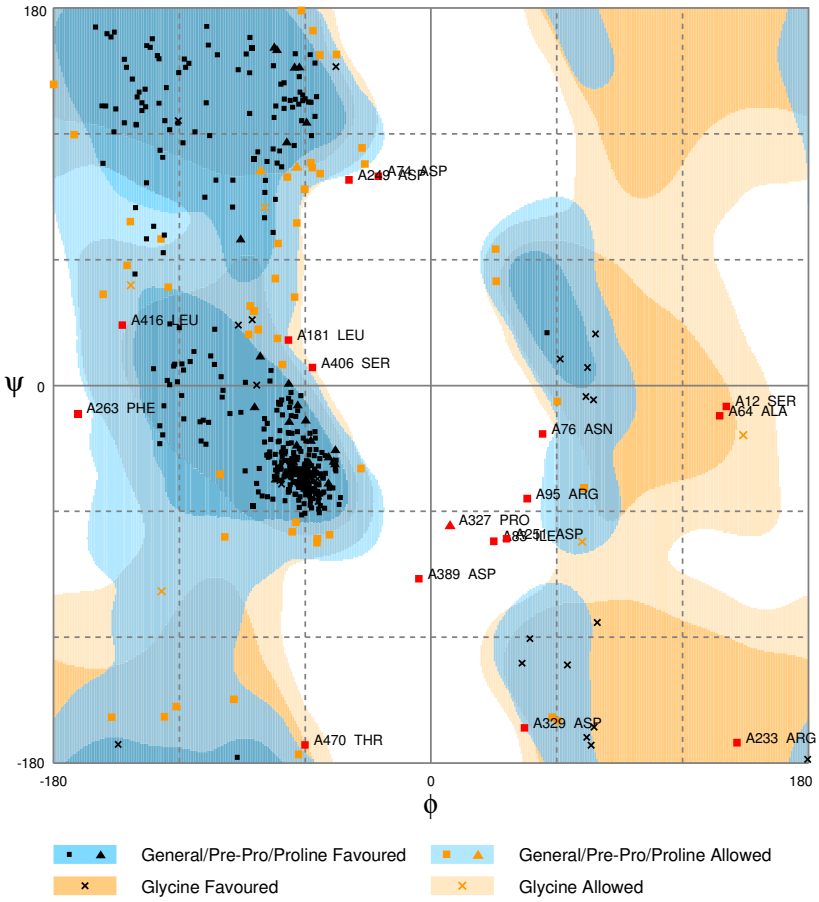


Number of residues in favoured region (~98.0% expected)	: 425 (89.1%)
Number of residues in allowed region (~2.0% expected)	: 35 (7.3%)
Number of residues in outlier region	: 17 (3.6%)

RAMPAGE by Paul de Bakker and Simon Lovell available at http://www-cryst.bioc.cam.ac.uk/rampage/

Please cite: S.C. Lovell, I.W. Davis, W.B. Arendall III, P.I.W. de Bakker, J.M. Word, M.G. Prisant, J.S. Richardson & D.C. Richardson (2002) Structure validation by Cα geometry: ϕ/ψ and Cβ deviation. *Proteins: Structure, Function & Genetics*. **50**: 437-450

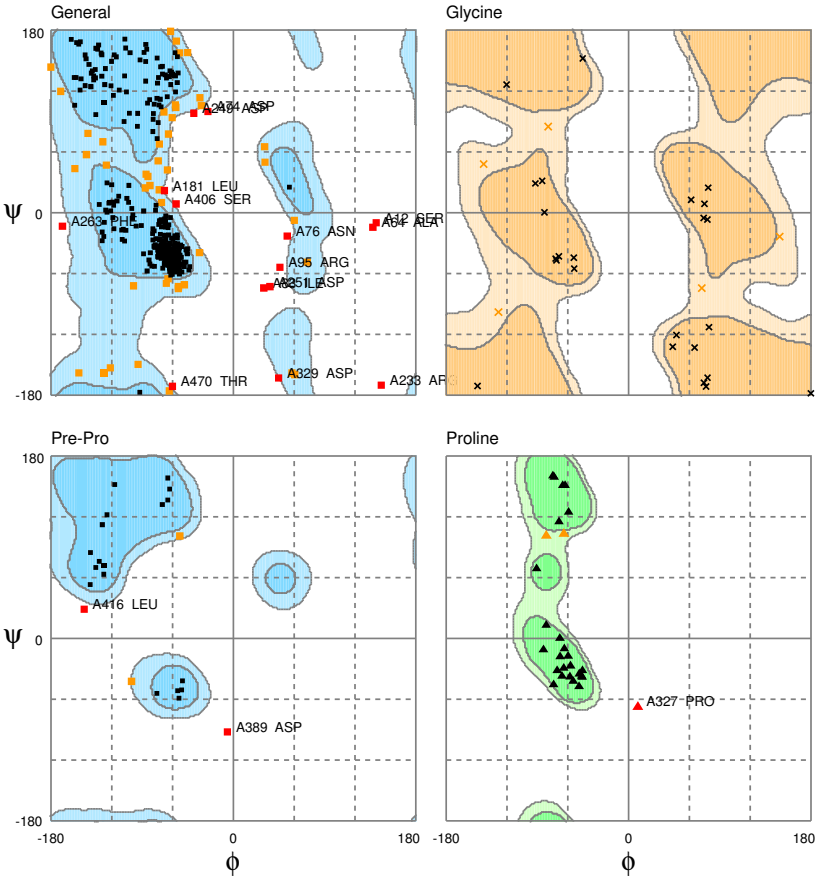
m3UA1:



Number of residues in favoured region (~98.0% expected)	:	406 (85.1%)
Number of residues in allowed region (~2.0% expected)	:	54 (11.3%)
Number of residues in outlier region	:	17 (3.6%)

RAMPAGE by Paul de Bakker and Simon Lovell available at http://www-cryst.bioc.cam.ac.uk/rampage/

Please cite: S.C. Lovell, I.W. Davis, W.B. Arendall III, P.I.W. de Bakker, J.M. Word, M.G. Prisant, J.S. Richardson & D.C. Richardson (2002) Structure validation by Cα geometry: ϕ/ψand Cβ deviation. *Proteins: Structure, Function & Genetics.* **50:** 437-450

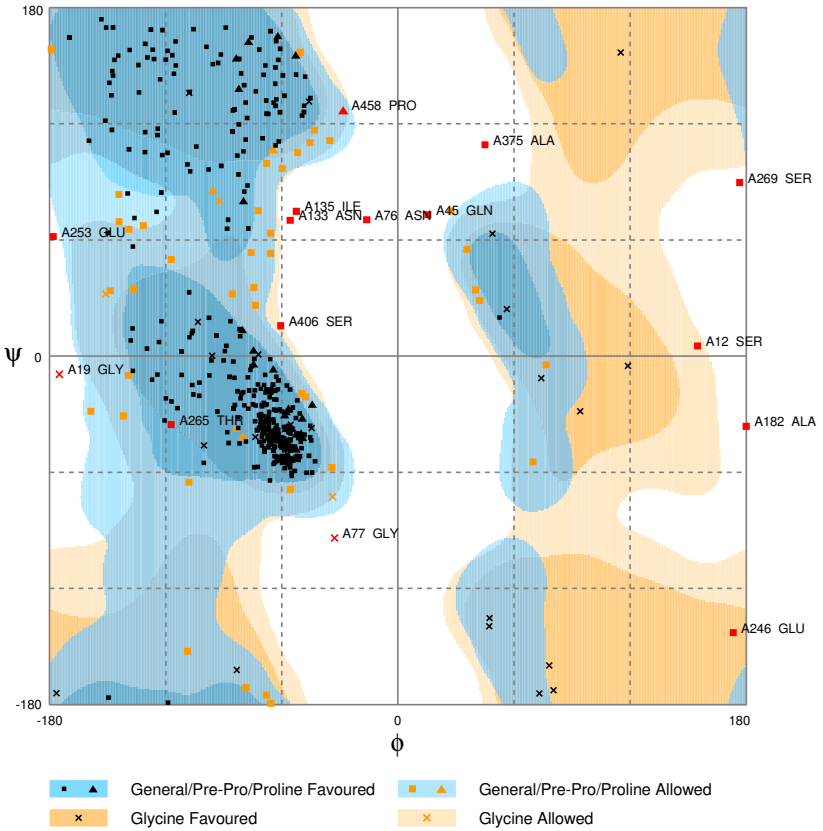


Number of residues in favoured region (~98.0% expected) : 406 (85.1%)
Number of residues in allowed region (~2.0% expected) : 54 (11.3%)
Number of residues in outlier region : 17 (3.6%)

RAMPAGE by Paul de Bakker and Simon Lovell available at http://www-cryst.bioc.cam.ac.uk/rampage/

Please cite: S.C. Lovell, I.W. Davis, W.B. Arendall III, P.I.W. de Bakker, J.M. Word, M.G. Prisant, J.S. Richardson & D.C. Richardson (2002) Structure validation by Cα geometry: ϕ/ψ and Cβ deviation. *Proteins: Structure, Function & Genetics*. **50**: 437-450

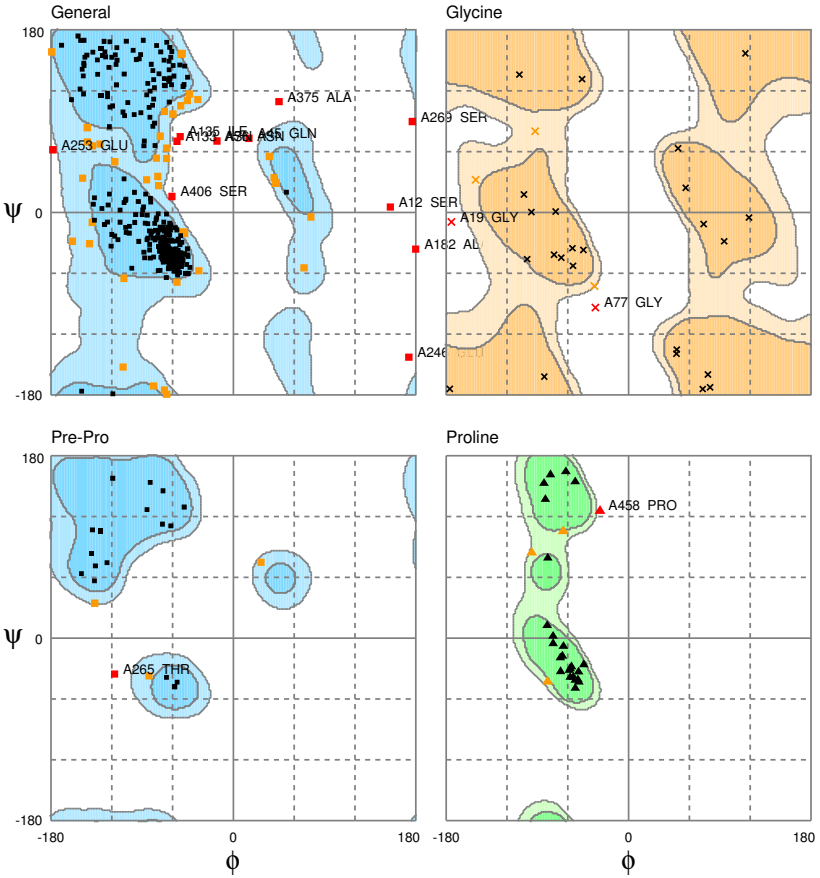
m5TE8:



Number of residues in favoured region (~98.0% expected)	: 414 (86.8%)
Number of residues in allowed region (~2.0% expected)	: 48 (10.1%)
Number of residues in outlier region	: 15 (3.1%)

RAMPAGE by Paul de Bakker and Simon Lovell available at http://www-cryst.bioc.cam.ac.uk/rampage/

Please cite: S.C. Lovell, I.W. Davis, W.B. Arendall III, P.I.W. de Bakker, J.M. Word, M.G. Prisant, J.S. Richardson & D.C. Richardson (2002) Structure validation by Cα geometry: ϕ/ψ and Cβ deviation. *Proteins: Structure, Function & Genetics.* **50**: 437-450



Number of residues in favoured region (~98.0% expected)	: 414 (86.8%)
Number of residues in allowed region (~2.0% expected)	: 48 (10.1%)
Number of residues in outlier region	: 15 (3.1%)

RAMPAGE by Paul de Bakker and Simon Lovell available at http://www-cryst.bioc.cam.ac.uk/rampage/

Please cite: S.C. Lovell, I.W. Davis, W.B. Arendall III, P.I.W. de Bakker, J.M. Word, M.G. Prisant, J.S. Richardson & D.C. Richardson (2002) Structure validation by Cα geometry: ϕ/ψ and Cβ deviation. *Proteins: Structure, Function & Genetics*. **50**: 437-450