UNIVERSIDADE FEDERAL DO AMAZONAS
PRO REITORIA DE PRESQUISA E PÓS-GRADUAÇÃO
DEPARTAMENTO DE APOIO A PESQUISA
PROGRAMA INSTITUCIONAL DE INICIAÇÃO CIENTÍFICA

ANOTAÇÃO DO TRANSCRIPTOMA DE Colossoma macropomum (TAMBAQUI).

Bolsista: Sergiane Moraes da Silva, FAPEAM

COARI-AMAZONAS
2013

UNIVERSIDADE FEDERAL DO AMAZONAS
PRO REITORIA DE PRESQUISA E PÓS-GRADUAÇÃO
DEPARTAMENTO DE APOIO A PESQUISA
PROGRAMA INSTITUCIONAL DE INICIAÇÃO CIENTÍFICA

RELATÓRIO FINAL
PIBIC-A/0072/2012
ANOTAÇÃO DO TRANSCRIPTOMA DE Colossoma macropomum
(TAMBAQUI).

Bolsista: Sergiane Moraes da Silva, FAPEAM
Orientadora: Profª. Drª. Andrea Ghelfi

COARI-AMAZONAS
2013

Esta pesquisa é financiada pela Fundação de Amparo a Pesquisa do Estado do Amazonas - FAPEAM, através do Programa Institucional de Bolsas de Iniciação Científica da Universidade Federal do Amazonas, foi desenvolvida pelo Núcleo de Estudo do Laboratório de Bioinformática e Modelagem do médio Solimões.

RESUMO

O projeto INCT/ADAPTA utilizou o sequenciador SAGE-SOLiD para analisar a expressão diferencial, sob determinados desafios ambientais, em organismos-chave para a região do Amazonas. Os dados gerados pelo sequenciador foram compostos por sequências de “tags”, contendo 21 dígitos cada, que devem ser traduzidos para nucleotídeos e posteriormente identificados com a devida anotação gênica. Para esta análise foi desenvolvido, no Laboratório de Bioinformática e Modelagem do Médio Solimões (LBMS) – UFAM-Coari, um “pipeline” denominado Tamandua, o qual conta as tags, identifica os genes, através de uma anotação automática e posteriormente faz a análise estatística da expressão diferencial dos genes. Este programa está em fase de proteção de software pelo INPA. Este PIBIC teve como objetivo realizar a anotação manual do *Colossoma macropomum* (Tambaqui), o qual pertence à família Characidae. A anotação manual das sequências encontradas neste organismo foram realizadas através do programa (BLASTn) *Basic Local Alignment Search Tool* e do Banco do Dados Gene Ontology (GO). Os resultados mostraram que a anotação automática foi eficiente, principalmente para genes que apresentaram uma identificação única. No entanto, como a tag, gerada pelo SAGE-SOLiD, apresenta apenas 21 nucleotídeos, em alguns casos mais de um gene foi identificado, de modo que nestas situações a acurácia da contagem dos genes pode ficar comprometida. As análises foram anotadas manualmente no Banco de Dados de Transcriptoma do Projeto INCT-ADAPTA, desenvolvido no LBMS-UFAM-Coari, para todos os genes expressos positivamente com um nível de expressão maior ou igual a quatro vezes (tratamento em relação ao controle), totalizando 95 genes.

Palavras Chaves: SAGE-SOLiD, *Colossoma macropomum* (Tambaqui), Transcriptoma.

ABSTRACT

The project INCT/ADAPTA used the sequencer SAGE- SOLiD to analyze the differential expression under certain environmental challenges in key organisms in the Amazonas State. The data generated were composed of sequences of tags, with 21 digits each, to be translated into nucleotides and subsequently identified with the appropriate gene name. For this analysis was developed, in the Laboratory of Bioinformatics and Modeling of the Middle Solimoes (LBMS) – UFAM-Coari, a"pipeline" called Tamandua. This program counts the generated tags, identify genes, through an automatic annotation and then makes a statistical analysis of differential expression of the genes. This program is in the process of a software protection by INPA. This PIBIC aimed to perform manual annotation Colossoma macropomum (Tambaqui), which belongs to the family Characidae. The manual annotation of the sequences found in this organism, were made through the program (BLASTn) Basic Local Alignment Search Tool and the databank of Gene Ontology (GO). The results showed that the automatic annotation was efficient, especially for genes that exhibited a unique identification. However, as the tag generated by SAGE-SOLiD, has only 21 nucleotides, in some cases more than one gene was identified. In these situations the accuracy of the count genes may be compromised. Analyses were done manually in Database Transcriptome Project INCT-ADAPTA, developed at LBMS- UFAM-Coari. It was considered all expressed genes with a positive expression level greater than, or equal to, four times (treatment compared to control), totalizing 95 genes.

Keywords: SAGE-SOLiD, *Colossoma macropomum* (Tambaqui), Transcriptoma.

LISTA DE SIGLAS

BLAST - *Basic Local Alignment and Search Tool*

INCT - Instituto Nacional de Ciência e Tecnologia

ADAPTA - Adaptações de Biotas Aquáticas

SAGE-SOLiD - *Serial Analysis of Genes Expression, Applied Biosystems*

LBMS - Laboratório de Bioinformática e Modelagem do Médio Solimões

PROTEC - Pro- Reitoria de Inovação Tecnologica

INPA - Instituto Nacional de Pesquisa da Amazônia

GO - *Gene Ontology*

KEGG - *Kyoto Encyclopedia of Genes and Genomes*

LISTA DE QUADRO

SUMÁRIO

1 INTRODUÇÃO

O projeto INCT/ADAPTA utiliza a técnica SAGE-SOLiD, o qual é um sequenciador de larga escala, para analisar as expressões diferencial em diversos organismos, sob determinados desafios ambientais. Os dados gerados pelo sequenciador são compostos por sequências de “tags”, de 21 números, que devem ser traduzidos para nucleotídeos e posteriormente identificados com a devida anotação gênica. Um programa de computador (pipeline), denominado Tamandua, foi desenvolvido no Laboratório de Bioinformática e Modelagem do Médio Solimões (LBMS), para realizar estas tarefas e posteriormente fazer a análise estatística. Este programa está em fase de proteção de software pela PROTEC (UFAM).

O organismo sequenciado pelo projeto INCT/ADAPTA no INPA foi o *Colossoma macropomum* (tambaqui), que pertence á família Characidae. Para fazer a anotação manual dessas sequências o projeto utilizou o programa *Basic Local Alignment Seararch Tool* (BLAST) on-line. Este tem por finalidade identificar os genes gerados pela análise da expressão diferencial. Existem dois tipos de anotação, a automática e a manual. A anotação automática compara os genes depositados no banco de dados através de um pipeline desenvolvido no Laboratório de Bioinformática e Modelagem do Médio Solimões (LBMS-Coari). A anotação manual requer um pesquisador, denominado curador, o qual ao consultar um ou mais bancos de dados, decide a função associada a cada gene.

Justificativa Diversos: transcriptomas amazônicos serão analisados através do Projeto INCT/ADAPTA. Com o intuito de melhor compreender os fatores que promovem as adaptações dos organismos nas mais diversas biotas aquáticas, faz-se necessário caracterizar os transcriptomas através da anotação manual dos genes expressos. Além disso, a anotação manual do transcriptoma do lábio do tambaqui iniciará uma nova linha de pesquisa no Instituto de Saúde e Biotecnologia.

Objetivo Geral Fazer a anotação manual de genes do transcriptoma do C. macropomum, incrementando as informações provenientes da anotação automática. Armazenar as informações no Banco de Dados de Transcriptoma que será criado para o projeto INCT/ADAPTA. Objetivos Específicos Verificar a acurácia da anotação automática; Identificar as vias metabólicas dos genes expressos; Fornecer os dados de anotação no Banco de Dados de Transcriptoma do Projeto INCT/ADAPTA.

2 REVISÃO BIBLIOGRÁFICA

A diversidade biológica da Amazônia é considerada uma das mais ricas do mundo, podendo ser encontrada uma variedade de espécies como (peixes, plantas terrestres e aquáticas, invertebrados, microrganismos e mamíferos aquáticos). É considerada vulnerável a alterações ambientais que podem ser naturais, como mudanças sazonais e/ou habitat, ou induzidos como contaminação por componentes petroquímicos, metais pesados, desmatamento, dentre outros, fazendo-se necessárias pesquisas para analisar a expressão diferencial desses diversos organismos, com o principal objetivo de saber se esses diferentes grupos de plantas e animais, são capazes de sobreviver ou não a esses desafios ambientais de acordo com (VAL, 2008).

O projeto INCT/ADAPTA, estuda a adaptação dos diferentes grupos de organismos existentes na Amazônia e as mudanças do meio ambiente, especialmente os ecossistemas aquáticos, por isso tem interesse em estudar o transcriptoma do *Colossoma macropomum* (tambaqui), é uma espécie que pode ser encontrada nos corpos de água branca e preta da região amazônica é uma espécie de alto valor comercial e muito apreciada pela quantidade de sua carne, é considerado útil para o monitoramento ambiental (LOPES-VASQUÉZ *et al.*, 2004; MATSUO *et al*., 2004b*;* VAL & ALMEIDA-VAL, 2004). Além disso, este organismo apresenta a vantagem de adaptar-se às condições ambientais e laboratoriais (VEINTEMILLA, 2006).

A anotação é o processo de interpretação dos dados gerados em um projeto de sequenciamento, convertendo-os em informações biologicamente relevantes. Esta pode ser feita de duas maneiras, automática e manual sendo que as duas são importantes para o resultado da pesquisa no trabalho de anotação de transcriptoma (LEWIS *et al*, 2000). A anotação automática não descreve os genes completamente, sendo necessária a visão do pesquisador para interpretação e correção das informações, requisitando assim a anotação manual (ANDRADE, 2002). Essa anotação manual vai ocorrer quando houver uma comparação com outros bancos de dados, assim o curador decidirá a função associada a cada gene (SHNEIDER & PEREIRA 2006).

3 METODOLOGIA

Foi utilizado para verificar a acurácia da anotação automática o algoritmo (BLAST) *Basic Local Alignment and Search Tool* (ALTSCHUL et al.,1990), que é uma ferramenta da bioinformática muito utilizada para alinhamentos de dados genéticos. Está disponível no *National Center for Biotechnology Information* – NCBI (www.ncbi.nlm.nih.gov/blast/), esse programa encontra similaridades entre sequências de DNA, RNA e proteínas de acordo com a linha de pesquisa determinada pelo pesquisador.

A pesquisa teve o transcriptoma do *Danio rerio* como referência, alguns parâmetros do BLAST foram estabelecidos para que o resultado da pesquisa ocorra de forma correta. O algoritmo utilizado foi o BLASTn, que compara sequências de nucleotídeos. As sequências são inseridas em formato fasta, a bases de dados foi o "RefSeq RNA", o qual é um banco de dados de sequencias de RNA anotado manualmente. Outros parâmetros como "Max Target sequences" foi utilizado o valor 10 (número de alinhamentos a serem mostrado), "Short queries" foi selecionado pois é ajuste de parâmetros para sequências pequenas, "Expect threshold" foi estabelecido 10, "Match/Mismatch Scores" scores para alinhamento correto foi igual a 1 e incorreto foi igual a menos 3, "Gap costs" penalização para abertura de gaps existente foi igual a 5 e extensão foi igual a 2.

A identificação das vias metabólicas dos genes expressos foi realizada no Banco de Dados Kyoto Encyclopedia of Genes and Genomes (KANEHISA et al., 2002). A alimentação do Banco de Dados de Transcriptoma do Projeto INCT/ADAPTA foi realizada manualmente através de uma interface gráfica.

4 RESULTADOS E DISCUSSÃO

Os resultados foram anotados corretamente pela anotação automática, no entanto, algumas tags mostraram que outros genes podem ser identificados, pois possuem o mesmo valor de "Score". Isso ocorreu porque o tamanho da sequência da tag é pequena favorecendo assim que se encontre mais de um gene com uma mesma tag. Não resultando de um erro na anotação automática, mas uma particulariedade da metodologia de sequenciamento via SAGE-SOLiD. Os resultados completos da anotação manual foram depositados no banco de dados de transcriptoma, desenvolvido pelo LBMS-Coari, os resultados parciais podem ser observados no Quadro 01.

Quadro 01 – Resultados parciais da anotação automática do Projeto INCT/ADAPTA.

NOME DO GENE	NOME DOS GENES ENCONTRADOS	CÓDIGO DE ACESSO	SYMBOL
Hypothetical protein LOC100003796	mitochondrial ribosomal protein L19, nuclear gene encoding mitochondrial protein	NM_001003544	mrpl19
	hypothetical protein LOC100002887	XM_001342521	LOC100002887
	BTB (POZ) domain containing 2	NM_001045092	BTB (POZ)
hypothetical LOC402880	Phosphatidylinositol glycan anchor biosynthesis, class	XM_003199835	Pigg
hypothetical LOC100333018	zgc:171801	NM_001110039	zgc:171801
	YTH domain family 3	NM_199870	ythdf3
LOC100534993	zgc:194207	XR_117631	zgc:194207
hypothetical LOC100535443	hypothetical protein LOC100331403	XM_003200790	LOC100331403
	provisional gene CK739236	XM_003200789	provisionalck739236
	uncharacterized LOC100536833	XR_117931	LOC100536833
	uncharacterized LOC100333278	XM_002667875	LOC100333278
	uncharacterized LOC100333326	XM_002667876	LOC100333326
	uncharacterized LOC100333249	XM_00319855	LOC100333249
hypothetical LOC100536327	si:ch73-27e22.1 microtubule-actin	XM_001335435	si:ch73-27e22.1
	crosslinking factor 1	XM_003200619	macf1
OBS.: Mais de um gene foi encontrado, no programa BLAST devido ao tamanho da tag ser de 21 pares de base.			

Após a identificação no BLASTn, os genes foram submetidos a uma nova pesquisa no Gene Ontology (GO) que caracteriza os genes em três ontologias independentes, a saber: processo biológico, função molecular e componentes celulares. Devido à preferência de anotação pelo Banco de Dados GO, em detrimento do KEGG, pelo Projeto INCT/ADAPTA, optou-se pela anotação manual do primeiro (Quadro 02).

Quadro 02 - Exemplos dos genes pesquisados no Gene Ontology.

Nome do gene	Termo de acesso	Ontologia
Endosulfine alpha a	GO:0007049 : cell cycle	biological process
	GO:0007049 : cell cycle	biological process
	GO:0051301 : cell division	biological process
	GO:0000086 : G2/M transition of mitotic cell cycle	biological process
	GO:0007067 : mitosis	biological process
	GO:0005737 : cytoplasm	cellular component
	GO:0019212 : phosphatase inhibitor activity	molecular function
	GO:0051721 : protein phosphatase 2A binding	molecular function
	GO:0004864 : protein phosphatase inhibitor activity	molecular function
	GO:0008601 : protein phosphatase type 2A regulator activity	molecular function
zgc:173615	GO:0008150 : biological_process	biological process
	GO:0005622 : intracellular	cellular component
	GO:0003676 : nucleic acid binding	molecular function
	GO:0008270 : zinc ion binding	molecular function
odorant receptor, family E, subfamily 124, member 1	GO:0007186 : G-protein coupled receptor signaling pathway	biological process
	GO:0050896 : response to stimulus	biological process
	GO:0007608 : sensory perception of smell	biological process
	GO:0007165 : signal transduction	biological process
	GO:0016021 : integral to membrane	cellular component
	GO:0016020 : membrane	cellular component
	GO:0005886 : plasma membrane	cellular component
	GO:0004930 : G-protein coupled receptor activity	molecular function
	GO:0004984 : olfactory receptor activity	molecular function
	GO:0004871 : signal transducer activity	molecular function
small nuclear ribonucleoprotein D1 polypeptide	GO:0007507 : heart development	biological process
	GO:0030529 : ribonucleoprotein complex	cellular component
	GO:0003676 : nucleic acid binding	molecular function
zinc finger protein 395	GO:0006355 : regulation of transcription, DNA-dependent	biological process
	GO:0006351 : transcription, DNA-dependent	biological process
	GO:0005737 : cytoplasm	cellular component
	GO:0005634 : nucleus	cellular component
	GO:0003677 : DNA binding	molecular function
	GO:0008270 : zinc ion binding	molecular function
immediate early response 2	GO:0060271 : cilium morphogenesis	biological process
	GO:0060026 : convergent extension	biological process

	GO:0007368 : determination of left/right symmetry	biological process
	GO:0070121 : Kupffer's vesicle development	biological process
	GO:0005634 : nucleus	cellular component
	GO:0005515 : protein binding	molecular function
protocadherin gamma-A4-like	GO:0007155 : cell adhesion	biological process
	GO:0007156 : homophilic cell adhesion	biological process
	GO:0016021 : integral to membrane	cellular component
	GO:0016020 : membrane	cellular component
	GO:0005886 : plasma membrane	cellular component

5 CONCLUSÃO

Os resultados dessa pesquisa mostraram que a anotação automática foi realizada de maneira correta, ou seja os genes e as vias metabólicas, identificados pelo “Pipeline Tamandua”, foram também encontrados no processo de anotação manual. As “tags” identificadas manualmente com mais de um gene foram devidas ao tamanho da “tag”, a qual contém apenas 21 nucleotídeos. Esta particulariedade é devida à técnica de sequenciamento SAGE-SOLiD utilizada e não a um erro na identificação do programa.

AGRADECIMENTOS

Á Deus que me deu forças para terminar este projeto, aos órgãos de apoio: INCT/ADAPTA, pelos dados fornecidos para pesquisa, ao INPA, UFAM, CNPq e principalmente a FAPEAM pela bolsa de estudo concedida e ao LBMS onde foi possível a realização deste projeto de PIBIC, bem como a orientação e supervisão da Profª. Drª. Andrea Ghelfi e a ajuda do colega Deney Araújo.

REFERÊNCIAS

ALTSCHU L, S.F.; Gish, W.; Miller, W.; Myers, E.W. & Lipman, D.J. **Basic local alignment search tool. J. Mol. Biol. 215:403-410**, 1990.

ANDRADE, R.V. **Caracterização parcial do transcriptoma do fungo dimórfico e patogênico Paracoccidioides brasiliensis**, 2002.

APPLIED BIOSYSTEMS. **Application note: SOLiD System high-throughput analysis of differential gene expression,** 2008.

APPLIED BIOSYSTEMS. **Applied Biosystems SOLiD 3 system SOLiD SADE guide**, 2009.

FORMIGHIERI, E.F. **Bioinformática e anotação de genes em Xanthomonas axonopodis pv. citri e Xilella fastidiosa: metabolismo de ferro e biossíntese de pequenas moléculas**, 2002.

JUNIOR, G.M.A. **Anotação do transcriptoma parcial de Anopheles (Nyssorhynchus) Darling Root, 1926**, 2011.

KANEHISA, M.; GOTO, S.; KAWASHIMA, S. et al. **The KEGG databases at GenomeNet. Nucleic Acids Research, v.30, p.42-46**, 2002

LEWIS, S.; ASHBURNER, M.; & REESE, M.G**. Annotating eukaryote genomes. Current Opinion in Structural Biology, 10(3), 349-354**, 2000.

LÓPES-VÁSQUEZ, K.; OLIVEIRA, S.S.; CUNHA, L.K.H.; NOZAWA, S.R.; VAL, A.L& ALMEIDA-VAL, V.M.F. **Differential gene expression on tambaqui, Colossoma macropomum Cuvier, 1818 exposed to crude oil. In: Bahavior, physiology and toxicology interactions in fish; Proceednigs of VI International Congress on the Biology of Fish, Manaus, Brasil, August 1-6, Sloman, K,; Wood, C. & MacKinlay, D (Eds), pp 13-17,** 2004.

SOUZA, A.R.B**. Análise do transcriptoma de etiquetas de sequências expressas da hipófise e parte do cérebro do tambaqui (Colossoma macropomum) e expressão do cDNA do hormônio de crescimento em Pichia pastoris,** 2009.

VAL, A.L**. National Institute of Science and Technology of the Adaptations of Aquatic Biota of the Amazon**, 2008.

VEINTEMILLA, C.A.C**. Impactos do fenantreno sobre o tambaqui (Colossoma macropomum Cuvier, 1818: CL50, crescimento e hematologia,** 2006.

ANEXO- CRONOGRAMA DE ATIVIDADES REALIZADAS

Nº	*Descrição*	*Ago 2012*	*Set*	*Out*	*Nov*	*Dez*	*Jan 2013*	*Fev*	*Mar*	*Abr*	*Mai*	*Jun*	*Jul*
01	*Revisão bibliográficas*	*R*	*R*	*R*	*R*	*R*	*R*	*R*	*R*	*R*	*R*	*R*	
02	*Acurácia da anotação automática através do BLAST*	*R*	*R*	*R*	*R*	*R*	*R*	*R*	*R*	*R*	*R*	*R*	
03	*Identificação das vias metabólicas através do KEGG*					*R*	*R*	*R*	*R*				
04	*Levantamento dos resultados Obtidos*								*R*	*R*	*R*		
05	*Alimentação do Banco de Dados de Transcriptoma do projeto INCT/ADAPTA*										*R*	*R*	
	- Elaboração do Resumo e Relatório Final (atividade obrigatória) *- Preparação da Apresentação Final para o Congresso (atividade obrigatória)*												*R*

R = Realizado