Bioinformatic tools help molecular characterization of *Perkinsus olseni* differentially expressed genes

Rita M. T. Ascenso

School of Technology and Management, Polytechnic Institute of Leiria, Morro do Lena, Alto do Vieiro, Leiria, Portugal, rita.ascenso@ipleiria.pt

Abstract

In the 80ies, in Southern Europe and in particular in Ria Formosa there was an episode of heavy mortality of the economically relevant clam *Ruditapes (R.) decussatus* associated with a debilitating disease (Perkinsosis) caused by *Perkinsus olseni*. This protozoan parasite was poorly known concerning its’ differential transcriptome in response to its host, *R. decussatus*. This laboratory available protozoan system was used to identify parasite genes related to host interaction. Beyond the application of molecular biology technologies and methodologies, only the help of Bioinformatics tools allowed to analyze the results of the study. The strategy started with SSH technique, allowing the identification of parasite up-regulated genes in response to its natural host, then a macroarray was constructed and hybridized to characterize the parasite genes expression when exposed to bivalves hemolymph from permissive host (*R. decussatus*), resistant host (*R. philippinarum*) and non permissive bivalve (*Donax trunculus*) that cohabit in the same or adjacent habitats in Southern Portugal. Genes and respective peptides full molecular characterization depended on several Bioinformatic tools application. Also a new Bioinformatic tool was developed.

1 Introduction

*Perkinsus* (*P.*) is a protozoan parasite affecting economically relevant mollusks worldwide, and was associated to clam heavy mortalities caused by Perkinsosis [1] [2] in Huelva coast [3] in South of Spain, being speculated that the higher water temperatures in the southern regions may impact on disease severity [4]. This disease had, also, negative impact in Asia, namely in Korea [5], China [6] and Japan [7]. Also, in USA, there was an oyster susceptibility to parasitism by *P. marinus* [8]. *P. olseni* [1] [2][9] was found to be the causative agent of Perkinsosis in Southern Portugal coastal region, where nearly 90 percent of the national carpet-shell clam production occurs [10].

This parasite was firstly described in Australia, infecting commercial relevant abalone species [11]. This situation led to an increase of interest to further characterize *P. olseni* response to its host, facilitated thanks to a continuous in vitro clonal cell culture of trophozoites derived from parasites infecting the carpet-shell clam, *R. decussatus* [12][13] (deposited in the American Type Cell Culture, ATCC 50984).

To characterize molecularly the parasite response to bivalves and identify differentially expressed genes from an organism from which the genome is unknown, it was used a Suppression Subtractive Hybridization (SSH). This involved the construction of a cDNA subtractive library [14], based in the Diatchenko methodology [15]. In parallel, a strategy of enrichment, Mirror Orientation Selection (MOS) was used to decrease background [16]. In order to obtain the
largest diversity of up-regulated genes, two libraries (Dfsl and efMOSl) were constructed and screened [14]. Duplicates of membranes with clones from the forward cDNA libraries were spotted and hybridized against forward and reverse libraries PCR products from the cDNA library. The autoradiography films with the impressed signals were analyzed using Bioinformatic tools. Each bacteria clone spotted represent an EST (Expressed Sequence Tag), i.e. a cDNA fragment representative of parasite gene expressed.

Among the numerous bivalves in South Portugal, R. decussatus and R. philippinarum are both susceptible bivalves to P. olseni infection but Perkinsosis does not significantly affect the wedge-shell clam Donax trunculus (Linnaeus, 1767), the blue mussel Mytilus galloprovincialis (Lamarck, 1919), the Portuguese oyster Crassostrea angulata (Lamarck, 1819) or the Japanese oyster, C. gigas (Thunberg, 1873), which also constitute important fishing resources. These bivalves have different habitats within the same coast, being D. trunculus found in the open, wave exposed sites, while the other species inhabit more sheltered areas.

To detect parasite relevant genes involved in bivalves’ response and characterize its’ expression a cDNA macroarray was constructed and hybridized [17]. Since this parasite was yet not very well described, being a protozoan, it could be relevant to compare its’ transcriptome with other parasites from Apicomplexa [18], group such as Plasmodium sp. and Toxoplasma sp., responsible for Malaria and Toxoplasmosis.

Based on relative gene expression, it were selected the most represented EST and the functionally relevant gene to further characterization. Using Bioinformatic tools was possible to interpret and treat full cDNAs and full gene sequence obtained using Molecular Biology methodologies [19]. The prediction and characterization of peptides was entirely obtained from Bioinformatic analysis. Several needs in macroarray data analysis promoted the collaboration with Department of Electronics and Informatics Engineering to develop a new Bioinformatic tool.

Although the original work was centered in Molecular Biology [20], in this paper we pretend to give an overview of the Bioinformatic tools used during the characterization of this poorly known parasite, starting from the analysis of its up-regulated transcriptome until the molecular characterization of genes and respective proteins.

2 Uncovering Parasite differential transcriptome

Transcriptome is the set of genes active in certain circumstances; P. olseni transcriptome studied in this work was a set of up-regulated genes in response to its host hemolymph, the clam R. decussatus. The differentially expressed genes were obtained using two approaches involving subtractive cDNA libraries. A cDNA subtractive library construction results in a forward and a reverse library, allowing the identification of down-regulated and up-regulated genes. In order to search for sober-expressed genes, the forward library was spotted and hybridized with forward and reverse library [15]. Two different forward libraries were constructed: (1) The Direct forward subtractive library (Dfsl), which consists on direct cloning of the forward cDNA subtractive library, followed by plate lifts screening to reveal up-regulated genes, and (2) the MOS enrichment methodology [16], used after completion of the SSH step and followed by in situ differential screening by bacteria dot blot (efMOSl) [14].
2.1 Libraries screening

Differential screening [21] was performed by double filter hybridization of each set of membranes for each approach, Dfsl and efMOSl, according to SSH protocol [15]. Numerous membranes were prepared with bacteria clones and each spot was analyzed after hybridization with forward and reverse library. After autoradiography of two replicates, the impressed signals were quantified by densitometry (Quantity One, BioRad, Richmond, USA). Quantity One is a powerful, flexible software package for imaging and analyzing 1-D electrophoresis gels, dot blots and other arrays. Available software was used to quantify gene expression, allowing the selection of genes of interest for a macroarray construction (Table 1).

For efMOSl, the scanned images were easily treated since the distribution of clones in the array and the software used a 96-well standard grid; facilitating signal detection and signal density quantification. In Dfsl, signal position was manually detected followed by signal quantification. Both data analysis involved the background quantification and subtractions from each foreground signal, performed in excel. The gene expression was obtained from the ratio between data from forward and data from reverse hybridization. Naked eye analysis was also performed to confirm information provided by the software analysis. Spots corresponding to clones showing more than two fold up-expression were further processed for sequencing. 96 and 204 clones were analyzed from Dfsl and efMOSl respectively [14].

2.2 Clones sequencing and analysis

The gene identification was possible based on sequencing at Macrogen (South Korea), which used a 3730XL DNA Sequencer and Monitoring tool, LIMS (Laboratory Information Management system). The sequences were observed one by one to check sequencing quality, using Chromas Lite, by chromatogram files observation. This permitted to ensure the sequence and the adaptors used. The adaptors sequences were removed using VecScreen\(^1\) and stored in the laboratory database, Vector NTI (Invitrogen, Carlsbad, USA), which allows to manage and analyze sequences.

2.3 Sequence identification

The major objective was to characterize the parasite response to its host, so molecular function was determined using GOblet server, which consists of a gene ontology (GO) search server based on similarity searches against known protein databases [22] (Table 1). For *P. olseni* cDNA sequences it was used a cut-off for E-value of 0.1 and 26 different GO subclasses were identified (Table 1), 25 from Dfsl and 6 from efMOSl. From BLASTx analysis the larger diversity obtained in the Dfsl library was confirmed with 82 different hits compared to 27 hits in the efMOSl [20][17]. To improve the description of the poorly characterized parasite, the sequences were submitted to BLASTx [23] taking into account the organism hit, the majority of clones corresponded to Protist kingdom. For Dfsl, 32 percent of total clones corresponded to Protist Kingdom and for efMOSl, 48 percent hits also belonged to Protozoan’s, as expected [20]. In fact *Perkinsus* genus fit in Protist Kingdom [2] although there are numerous genes, in minority, belonging to all other kingdoms.

\(^1\)http://www.ncbi.nlm.nih.gov/projects/VecScreen/
Table 1: Work-flow followed for *P. olsenii* transcriptome and molecular characterization, showing the Molecular Biology strategy aside with Bioinformatic tools used.

<table>
<thead>
<tr>
<th>MOLECULAR BIOLOGY</th>
<th>BIOINFORMATICS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parasite cells</td>
<td>Parasite cells exposed to host</td>
</tr>
<tr>
<td><strong>cDNA Subtractive Library</strong></td>
<td><strong>Two approaches</strong></td>
</tr>
<tr>
<td>Dfsl library</td>
<td>efMOSl library</td>
</tr>
<tr>
<td><strong>Differential gene expression</strong></td>
<td><strong>in silico data analysis</strong></td>
</tr>
<tr>
<td>Clones screening</td>
<td>Quantity One</td>
</tr>
<tr>
<td>Scanning image analysis</td>
<td>Quantity One</td>
</tr>
<tr>
<td>Signal quantification</td>
<td>Excel</td>
</tr>
<tr>
<td>Gene expression</td>
<td></td>
</tr>
<tr>
<td><strong>Differential Transcriptome</strong></td>
<td><strong>in silico sequence analysis</strong></td>
</tr>
<tr>
<td>EST sequence read and quality check</td>
<td>Cromas Lite</td>
</tr>
<tr>
<td>Sequence storage and treatment</td>
<td>Vector NTI</td>
</tr>
<tr>
<td>Gene Molecular function</td>
<td>GOblet</td>
</tr>
<tr>
<td>Gene identification</td>
<td>BLASTx</td>
</tr>
<tr>
<td><strong>Gene expression pattern</strong></td>
<td><strong>Macroarray analysis</strong></td>
</tr>
<tr>
<td>Macroarray construction and hybridization</td>
<td>Quantity One</td>
</tr>
<tr>
<td>Gene expression quantification</td>
<td>Developed tool: Automatic Macroarray Analysis</td>
</tr>
<tr>
<td><strong>Gene and Protein Molecular Characterization</strong></td>
<td><strong>in silico analysis</strong></td>
</tr>
<tr>
<td>Gene and Protein identification</td>
<td>BLASTn and BLASTx</td>
</tr>
<tr>
<td>Primers design</td>
<td>Primer Premier</td>
</tr>
<tr>
<td>Sequence assemble</td>
<td>Vector NTI</td>
</tr>
<tr>
<td>ORF and peptide prediction</td>
<td>ORF Finder</td>
</tr>
<tr>
<td>Intron-Exon boundaries</td>
<td>B12seq</td>
</tr>
<tr>
<td>Gene promoter characterization</td>
<td>MatInspector</td>
</tr>
<tr>
<td>Protein characterization</td>
<td>SignalP, MotifScan, TMprep</td>
</tr>
<tr>
<td>Protein domain’s motifs</td>
<td>Clustalx and WebLOGO</td>
</tr>
<tr>
<td>Protein functional family</td>
<td>Clustalx and MEGA4</td>
</tr>
</tbody>
</table>

3 Transcriptome characterization

To study genes pattern of expression, macroarrays were hybridized against transcripts from parasite exposed to permissive, resistant and non permissive bivalves [20]. A macroarray was constructed with 98 selected clones, 72 from Dfsl and 26 from efMOSl, namely those with molecular function related to cell adhesion, binding, chaperone, signal transduction, transport, cellular process, transcription, translation, structural, catalysis and basal metabolism, and clones matching hypothetical parasite genes, all over-expressed in response to host hemolymph [14].

The identification of genes and/or proteins was based on comparison with annotated databases (Table 1). BLAST was performed against the Apicomplexa Database[^2] [18]. A few genes confirmed its previously described function; others remained unknown or revealed no significant hit, below 0.07 E-value cut-off. A more generalized search was performed against non-redundant database from NCBI[^3] [23]; still, some ESTs gave no hit, which points out for an

[^2]: http://apidb.org/apidb/
atypical organism, poorly known and needing further molecular characterization. Databases updating may contribute for future comparison and improved results, for example the recently updated Apidb, now Eukaryotic Pathogen Data Base Resource or the recent ESTs described for *P. marinus* [24].

### 3.1 Parasite gene expression pattern

Two macroarrays were constructed and hybridized twice. From each there were various autoradiography degrees of exposure to get films with non saturated and saturated signal.

Gene expression was again based on image scanning, analysis and signal quantification from each experiment and replicates of macroarrays, using Quantity One (Table 1). The number of genes up-regulated one hundred fold or more in response to each bivalve hemolymph was: 58 genes up-regulated in response to *R. philippinarum* hemolymph and 84 genes up-regulated in response to *D. trunculus* hemolymph [14]. The differentially expressed genes were considered relevant in the host-parasite interaction.

### 3.2 Development of macroarray analysis tool

Having numerous autoradiography films of diverse saturation levels and with arrays of variable sizes from cDNA library screening and from macroarray; for quick macroarray data analysis, from image analysis to gene expression, it was developed a Automatic Macroarray Analysis Tool (Table 1), in cooperation with the Department of Electronics and Informatics Engineering of Faculty of Science and Technology of University of Algarve [25]. Lately this Bioinformatic tool was updated to allow more than quantification of a set of experiments with macroarrays, analyzing gene expression data. It was improved to be able for other densitometric analysis involving circular dish plates such as petri dish plates, or multiple cell plates [26]. This tool is proposed to be useful in the measurement of cell mineralization analyzing calcium quantification based on Von Kossa staining.

### 4 Characterization of most represented gene

From the 300 EST resulting from both approaches of SSH, we chose to focus on some of the highly represented. Through a cDNA subtractive library were identified those parasite genes up-regulated when challenged with host hemolymph. From these, the gene represented in more copies (15 ESTs) was obtained from the efMOSI cDNA subtracted library. First identified as *Moorella* sp. cytosine deaminase, at the time [20], following BLASTx analysis [23], was considered related to pyrimidine salvage pathway. Later a higher homology was obtained with proteins from the chlorohydrolase family. So, it was named PoClhl, Chlorohydrolase like.

#### 4.1 PoClhl gene identification and promoter characterization

The chlorohydrolase was shown to be firmly linked with a major amidohydrolase protein superfamily [27]. Homologous proteins catalyzing different reactions are being discovered at an
increasing rate with functional genomics focusing attention on the interplay between molecular sequence and function. This amidohydrolase superfamily comprehend some of the enzymes responsible for nitrogen recycle, essential metabolites for DNA and nucleic acid synthesis [27], including 30 percent of the steps in four intermediate metabolic pathways [28]. The *P. olseni* EST was identified, through *in silico* search analysis, BLAST [23], as being a chlorohydrolase protein family gene, with significant hits for TrzA, AtzA and CD.

The promoter analysis to search for putative transcription factors binding sites was used MatInspector at Genomatix Portal\(^4\). The corroboration of gene organization was obtained comparing *P. olseni* and *P. marinus* genes [29] to full characterize the PoClhl gene, but also the PoAdh and PoNHE genes. The promoter was carefully analyzed *in silico* and its characteristics compared to other *Perkinsus* spp. promoters. PoClhl and PmClhl had no canonical TATA box, but a T-rich region was identified. A comparable T-rich region was identified in other *Perkinsus* gene promoters such as the PmSOD1 gene, which contains a proximal TATA (TATTTTA) 34 bases from the transcription start point, a position typical of many Eukaryotic promoters [30], while PmNramp has a putative TATA box at -162 position [29]. The confirmation was performed by multiple alignment [31] and WebLOGO visualization [32].

### 4.2 PoClhl protein characterization

From *in silico* analysis, the PoClhl polypeptide appears as a non-secretory protein as expected, a result confirmed by SignalP analysis since no signal peptide was identified. Motif Scan predicted various putative phosphorylation sites, one cyclin recognition site and one putative tyrosine sulphating site\(^5\). The comparison of the presently described polypeptides with those from amidohydrolase family demonstrated a higher similarity with the Proteobacteria, *N. oceani*, amidohydrolase, described as a cytosine deaminase and related metal-dependent hydrolase according to predicted function. The *in silico* identification of PoClhl polypeptide as a member of the amidohydrolase family was tested by protein expression studies. The absence of an effective transfection assay for *P. olseni* limited further studies [20].

In order to search for functional domains described for metal-dependent protein family, some selected proteins were aligned with PoClhl and PmClhl. The expected motif HxH, present in the proteins suggested them to be coordinated by divalent-metals bound to specific residues, as for the other family proteins. It also contains the motif HxHxxE, relevant in substrate coordination at the active site. The metal ligands, four histidines and one aspartic acid residue, are strictly conserved in the three enzyme families, and define a subtle but sharp sequence signature of a metal-dependent hydrolases superfamily [28]. Multiple alignment using ClustalX [31] followed by WebLOGO\(^6\) [32] permitted the visual identification of conserved residues related to protein function, namely according to described examples [28]. This analysis confirmed the predicted function of PoClhl.

\(^4\)http://www.genomatix.de

\(^5\)http://myhits.isb-sib.ch/cgi-bin/motif_scan

\(^6\)http://weblogo.berkeley.edu/logo.cgi
5 Characterization of functional relevant genes

From macroarray gene expression and identification, two genes were further analyzed; adhesion related (PoAdh) and transporter, Na+/H+ antiporter (PoNHE) [20] were molecularly characterized once their function appear to be related to host response. DNA sequence, ORF, introns and exons and predicted peptide were obtained by Molecular Biology methodologies [19] and use of several Bioinformatics (Table 1). These two genes were submitted to an exhaustive molecular characterization which results won’t be presented, but only the strategy followed, as for PoClhl.

5.1 Genes molecular characterization

Full cDNA sequence was obtained based on Rapid Amplification of both 5’ and 3’ cDNA Ends (RACE) using Marathon cDNA library (Clontech, Palo Alto, USA) and primers designed based on the clones of each gene from the subtractive library. Primer Premier Software was used for all primers design (PREMIER Biosoft International, Palo Alto, USA) [20]. Numerous clones obtained with diverse sequence size, for 5’ and 3’ of the cDNA; these were assemble using the assemble tool of Vector NTI Software. The full consensus cDNA sequence was analyzed, identifying the start codon ATG and stop codon using ORF Finder at NCBI. ORF Finder tool allowed to predict the peptides coded. Having access to species P. marinus genome sequence project and ESTs [24], a local search was performed to confirm the genes presence.

Full gene sequencing was easily obtained based on P. olseni cDNA sequence aligned with P. marinus gene sequencing to predict exon region to design primers. Gene sequencing including the search of promoter region behind 5’ UTR was obtained using a Universal Genome Walker library to get upstream and downstream regions of the gene [20]. Two libraries were prepared with two different endonucleases enzymes, after several amplifications, cloning and sequencing was obtained several clones with fragments of full genes sequence. Having several sequences, these were assembled to complete each full gene consensus, after cleaning the sequence from sequencing bad quality ends and primer sequence. The assembling was performed using the Vector NTI Software. In parallel to molecular biology techniques, Bioinformatic tools allow the results analysis and presentation, namely full genes sequence and molecular organization.

5.2 Proteins characterization

Protein analysis was performed based on the predicted peptides sequences for each ORF sequence. The secondary structure prediction was performed using ExPASY available tools, namely TMprep for transmembrane regions identification [33]. More relevant than hydrophathy profile, the identification of protein domains and motifs are very useful to predict the protein function compared to other proteins of the same family. In order to comprehend post-translational modifications, SignalP [34] was used to detect secreted proteins and MotifScan [35] for protein domains prediction. To screen the protein path, MITOPROT and PATS were used to predict if one of this predicted peptides were targeted to mitochondria or apicoplast, respectively. Also the targeting to chloroplast was checked using CloroP. The knowledge about
Cation Proton Antiporter 1 super-family and NHE protein family [36] [37], allowed the comparison among 46 proteins using multiple alignments and phylogeny analysis helped to ensure protein function and cellular localization. ClustalX [31] and MEGA4 [38] analysis allowed to infer that NHE is close related to Endossomal-TNG or Plant vacuolar, part of the intracellular clade, suggesting that PoNHE may be responsible for adjustment to different environmental conditions, possibly responding to an osmotic shock related with the interface parasite-host [20]. The characterization of this parasite wouldn’t be possible without making use of Bioinformatic tools.

6 Conclusion

This paper intend to present the workflow of a molecular characterization of differential transcriptome of P. olseni parasite, in response to its host. The Molecular Biology technology and methodologies demand a multiple experiment design, with replicates and numerous sequencing data. The amount of data analysis was only possible using Bioinformatic tools. For detection and identification of parasite up-regulated genes and to clarify the full cDNA, gene and protein characterization of three genes were used diverse Bioinformatic tools. The presented strategy was used along the development of PhD work and resulted also in the development of a new Bioinformatic tool, Automatic Macroarray Analysis Tool [25]. Future work aims optimizing this instrument in order to help screening differential expressed genes based in macroarrays, for poorly characterized systems. This paper may help starting Molecular Biologists during a molecular characterization using, mainly, free web available Bioinformatic tools.

Acknowledgements

Acknowledgments to M.Sc. Ana Filipa Nogueira, PhD Adriano Benvenho and M.Sc. Mário Antunes for technical support in Latex utilization.

References


