ANALYSIS OF 172 EXPRESSED SEQUENCE TAGS FROM THE SHARK (*SQUALUS ACANTHIAS*) RECTAL GLAND

Stephen G. Aller[1,2], Christine M. Smith[2], David W. Towle[2,3] and John N. Forrest, Jr[1,2]
[1]Department of Medicine, Yale University School of Medicine, New Haven, CT 06510
[2]Mount Desert Island Biological Laboratory, Salisbury Cove, ME 04672
[3]Department of Biology, Lake Forrest College, Lake Forrest, IL 60045

The partial nucleotide sequence of cDNA, or Expressed Sequence Tags (ESTs), is an efficient means of identifying genes expressed in specific tissues. Because > 95% of most eukaryotic genomes is non-coding, data from genome sequencing requires computationally expensive algorithms and statistical analyses to identify putative proteins. Since introns are not present in ESTs, the sequencing of these molecules is a more efficient means of identifying both known and novel proteins expressed in a given tissue. The shark rectal gland (SRG) is a highly specialized marine epithelial organ with well defined physiology. The cloning of selective genes for membrane proteins in this model tissue has yielded important data for comparative molecular biology. Therefore, we have initiated an EST sequencing project using a SRG cDNA library.

A lambda ZAP II custom cDNA library (kindly provided by Dr. Joseph Mindell) was prepared using highly purified SRG total RNA primed with both oligo(dT) and random primers. Individual phage plaques were picked and subjected to a 30 µl polymerase chain reaction (PCR) using T3 and T7 primers to amplify the SRG gene insert. The reaction parameters were 35 cycles of (95 °C for 30 sec; 52 °C for 30 sec; 68 °C for 4 min). The resulting PCR product was purified using the QIAquick PCR purification kit (Qiagen). Two sequencing reactions (using T3 or T7 primer) were performed using 4 µl of each PCR product according to the manufacturer's protocol (ABI) on an Eppendorf Mastercycler thermocycler in the MDIBL Sequencing Center. Typical sequencing reactions yielded 800-900 bp of high quality data. After manually removing vector sequence, the remaining sequence was subjected to the Basic Local Alignment Search Tool (BLAST) algorithm against the Genbank database. In some cases, sequences were manually edited and subjected to further analysis (ORF Finder and BLASTX). All bioinformatics tools were from www.ncbi.nlm.nih.gov using the default program parameters. The Expect value (E-value) reported in the tables is a parameter that describes the number of hits "expected" to occur by chance when searching a database of a particular size. For example, an E-value of 1 assigned to a query sequence and a database match is interpreted as meaning that these two sequences relate to each other in a manner that is expected to occur one time in the database by chance.

Based on the PCR products of the first 36 picks, the average insert size of the library was 2.6 kb, with a range of 500 bp to 9 kb. Bi-directional sequencing yielded high quality data from both 5'- and 3'- ends of most EST clones. Each nucleotide sequence obtained was compared to the Genbank database and was preliminarily classified as: 1) sequence having high scoring matches to Genbank and 2) sequence having low or no scoring matches. Of 344 total sequencing reactions to date, 25% failed to produce high quality sequence, mostly due to polyadenines, 23% were identified as having significant match to either mitochondrial or ribosomal DNA, 13% had a significant match to a Genbank entry (Table 1) and 39% had no significant matches to any Genbank entry using the BLASTN comparison tool. Two possibilities exist for this last grouping: 1) The sequences are either 5'- or 3'- untranslated regions of mRNA (UTR) which are not found in Genbank, or 2) they represent open reading frames (ORF) of novel genes.

Clone #/primer	Genbank title	E-value
8 T7	Homo sapiens mRNA; cDNA DKFZp434C151	0.003
12 T7	Homo sapiens voltage dependent anion channel	1e-04
17 T3	Homo sapiens acyl-Coenzyme A dehydrogenase	2e-12
20 T7	Homo sapiens eps8 binding protein e3B1 mRNA	7e-29
21 T7	Rabbit eucaryotic release factor (eRF) mRNA	3e-28
31 T7	M.musculus mRNA for TAX responsive element binding protein 107	4e-21
54 T3	X. laevis EF-1 alpha mRNA for elongation factor 1-alpha	1e-124
68 T7	Chicken prostaglandin G/H synthase mRNA	5e-11
70 T3	**Squalus acanthias bumetanide-sensitive Na-K-Cl cotransport protein**	0
82 T3	Torpedo marmorata mRNA fragment for acetylcholinesterase C-term.	2e-13
83 T3	**Squalus acanthias cystic fibrosis transmembrane conductance regulator**	1e-154
85 T3	Homo sapiens syld709613 protein (SYLD)	1e-15
85 T7	Raja eglanteria clone 2113 Ig heavy chain (Vx, Dx1, Dx2, Jx, Cx1, and Cx2) region	2e-04
88 T3	Homo sapiens mRNA for KIAA0982 protein	3e-16
91 T7	Chicken (Na+,K+)-ATPase-beta-2 subunit mRNA	2e-05
92 T7	Homo sapiens ATQL1 pseudogene, partial sequence	4e-06
101 T3	Cynops pyrrhogaster mRNA for collagenase 3	2e-07
104 T3	Xenopus laevis occludin mRNA	0.004
118 T7	H.sapiens mRNA for transcription factor BTF 3	1e-102
122 T3	Homo sapiens splicing factor, arginine/serine-rich	5e-17
124 T3	Oryctolagus cuniculus mRNA; calmodulin-dependent protein kinase II-delta	6e-45
126 T7	Mus musculus phenylalanyl tRNA synthetase beta subunit (Frsb) mRNA	4e-06
131 T3	Raja eglanteria clone 2113 Ig heavy chain gene region	8e-26
137 T7	Human cosmid HDAC (8C10Y2) DNA	1e-18
141 T3	Human mRNA for T-cell cyclophilin	7e-14
144 T3	Raja eglanteria clone 2113 Ig heavy chain gene region	7e-04
145 T3	Raja eglanteria clone 2113 Ig heavy chain gene region	4e-09
146 T3	Chicken ubiquitin I (UbI) gene	1e-138
146 T7	B.taurus mRNA for plasmalemmal porin	3e-19
149 T3	Torpedo marmorata mRNA fragment for acetylcholinesterase C-term	1e-09
152 T3	Gallus gallus caspase-3 mRNA	2e-05
156 T3,T7	M.musculus seryl-tRNA synthetase (SERS) mRNA, 5' end	3e-25
165 T7	Rat tyrosine phosphatase (PRL-1) mRNA	3e-16
170 T3	Homo sapiens clone 23938 mRNA sequence	4e-06
172 T7	Raja eglanteria clone 2113 Ig heavy chain	3e-13
177 T3	Homo sapiens mRNA for KIAA0689 protein	1e-09
179 T3	Human DNA sequence from clone 657D16 on chromosome 1p32.2-34.2	7e-08
186 T3,T7	Bovine NADH:ubiquinone reductase 75 kD subunit of complex I	5e-52
188 T7	Human Ets transcription factors NERF-1a and NERF-1b (NERF-1a,b) mRNA	3e-25
190 T7	Mus musculus 37kDa oncofetal antigen mRNA; Rat laminin receptor mRNA, 3' end	2e-11
191 T7	Heterodontus francisci T-cell receptor J region homolog gene	9e-04

Table 1. All shark rectal gland ESTs with a significant match (E < 0.005) to a Genbank entry as determined by BLASTN search. Previously identified *Squalus acanthias* genes are in bold.

To distinguish between these two possibilities, we selected 37 of the highest quality sequences for further analysis. The original chromatogram from each reaction was manually edited and corrected for base calling errors. None of these 37 sequences had a significant match to Genbank using a nucleotide matching algorithm. All six possible reading frames were then compared to the non-redundant Genbank protein database (BLASTX) after editing. As shown in Table 2, 11 of the 37 sequences (30%) could be identified with a Genbank entry after translation into a protein.

Clone #/primer	Genbank title	E-value
18 T3	ubiquitin-like protein 8	2e-28
22 T3	tryptophanyl-tRNA synthetase (tryptophan--tRNA ligase)	3e-21
43 T7	JM1 [Homo sapiens]	8e-24
48 T3	bile salt-dependent lipase, BSDL human	1e-06
50 T7	homeobox protein [Danio rerio]	3e-69
51 T7	ooplasm specific protein [Mus musculus]	2e-05
62 T3	ORF2 [Platemys spixii] putative reverse transcriptase	1e-07
64 T3	ORF2 [Platemys spixii] putative reverse transcriptase	1e-16
67 T7	putative short chain dehydrogenase [Schizosaccharomyces pombe]	6e-16
76 T3	unknown protein - chicken	5e-34
95 T3	putative protein kinase NY-REN-64 antigen [Homo sapiens]	1e-39

Table 2. Eleven EST sequences unidentified by BLASTN were manually edited and identified as having a significant match to a Genbank protein database entry using the BLASTX search algorithm.

These 55 SRG sequences (table 1 and 2) include 53 genes identified for the first time in the SRG and two genes previously identified in this tissue (NKCC1 and CFTR).

In considering whether the 26 sequences having no match to Genbank were novel genes, we analyzed the ORF lengths of sequences obtained to date and found them to code for 140 ± 13 residues for Genbank matches and 69 ± 8 residues for no matches (values are mean ± SEM, $p<0.001$). We conclude that most of the non-matching sequences represent untranslated regions of shark genes or code for small novel peptides. Two sequences, 42 T3 and 81 T3, had large ORFs (129 and 255 amino acids, respectively) and probably represent novel proteins in the rectal gland. As more sequence is obtained, both known genes of biological interest and completely novel proteins will be discovered in this and other model marine tissues studied at MDIBL.

In summary, we report the first data from a shark rectal gland EST sequencing project. These sequences will be annotated in a database in the Bioinformatics Unit at MDIBL and clones will be made available, upon request, to MDIBL scientists and other marine biologists.

This work was supported by NIH grants DK 34208 and NIEHS P30-ES 3828 (Center for Membrane Toxicology Studies), and a Grant in Aid from the American Heart Association, Maine Affiliate.