Construction of a catalogue of amphioxus
(Branchiostoma florid) genes expressed at neurula
and gastrula stages by oligonucleotide fingerprinting.

Georgia Panopoulou

a thesis submitted for the degree of Doctor of Philosophy
in the University of London

Windeyer Institute of Medical Sciences
Department of Molecular Pathology and Clinical Biochemistry
University College London
Windeyer Building
46 Cleveland Street
London W1P 6DB
ENGLAND

Max-Planck Institute fuer Molekulare Genetik
Department Prof. Hans Lehrach
Ihnestrasse 73
D-14195 Berlin
GERMANY

June 1999
To Dimitri, Vicky and Lena
Abstract

So far 9% of the human genome has been sequenced and for more than 50% of human genes there is a representative tag sequence (61,811 Unigene clusters, NCBI release statistics 1999). The major challenge of the next phase of the human genome project will be to try to understand the function of those genes. Cross-species comparison is an effective tool for providing clues based on sequence conservation combined with the possibility of genetic manipulations such as generation of mutants or crosses.

However, the large evolutionary distance between the invertebrate and vertebrate model systems and human, from which the bulk of sequences deposited in the public databases is derived, limits the extent of such comparisons.

I have chosen Amphioxus as a key organism for comparative study of vertebrate genes since it is phylogenetically placed at the transition point from invertebrates to vertebrates within the chordate phylum. Expression studies of mainly genes for which vertebrate homologues have already been identified showed conserved expression patterns strengthening its position as a close invertebrate relative to the vertebrates. Moreover, since most of the identified amphioxus genes have been found in fewer copies than their vertebrate counterparts, it is hypothesized that amphioxus branched off from the chordate lineage before two phases of genome enlargement, the first before the emergence of lampreys while the second close to gnathostome origins. As a result, amphioxus is expected to have a quarter of the number of vertebrate genes which makes it an attractive organism for identifying novel genes quickly.

To construct an expression index of amphioxus genes, I have oligonucleotide fingerprinted a gastrula (BFLG) and a neurula (BFL26) stage cDNA libraries (110,592 clones in total). Oligonucleotide fingerprinting is based on the hybridisation of short oligonucleotides on high density grids carrying PCR products of cDNA libraries. Clones showing the same hybridisation pattern are clustered as representing the same gene. As a result, I have identified 1,783 clusters of different transcripts and 7,234 singletons in the gastrula stage while 4,674 clusters and 6,028 singletons in the neurula stage library. 5' tag sequencing of 3,870 (4,079 sequences) distinct clones from both libraries and their subsequent sequence clustering confirmed the validity of oligonucleotide fingerprinting as a pre-selection method able to achieve more than 3.5 fold normalisation and led to the identification of 1,811 from the BFLG and 1,516 from BFL26 unique transcripts. 38% of the BFL26 sequences and 20.7% of the BFLG sequences have known homologues in other organisms, among them some matching human disease genes.
Table of contents

ABSTRACT ..............................................................................................................................................3
TABLE OF CONTENTS ........................................................................................................................4
TABLE OF FIGURES ............................................................................................................................7
LIST OF TABLES ...................................................................................................................................8
ABBREVIATIONS...................................................................................................................................9
ACKNOWLEDGEMENTS ...................................................................................................................11

CHAPTER 1 ...........................................................................................................................................12

1.1. THE IMPORTANCE OF CROSS-SPECIES COMPARISONS.............................................................12

1.2. AMPHIOXUS .......................................................................................................................................19

1.3. Oligonucleotide fingerprinting ........................................................................................................29

AIM OF THIS THESIS ........................................................................................................................37

CHAPTER 2 ...........................................................................................................................................39

MATERIALS AND METHODS..............................................................................................................39

2.1. REAGENT SUPPLIERS ....................................................................................................................39

2.2. SOLUTIONS....................................................................................................................................40
### 2.3. EXPERIMENTAL PROCEDURES

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.3.1 RNA isolation</td>
<td>45</td>
</tr>
<tr>
<td>2.3.2 cDNA library construction</td>
<td>46</td>
</tr>
<tr>
<td>2.3.3 Introduction of Plasmid DNA into bacteria</td>
<td>50</td>
</tr>
<tr>
<td>2.3.4 Arraying of libraries</td>
<td>51</td>
</tr>
<tr>
<td>2.3.5 Genomic DNA extraction</td>
<td>57</td>
</tr>
<tr>
<td>2.3.6 Purification of plasmid and cosmid DNA (mini-preparation)</td>
<td>57</td>
</tr>
<tr>
<td>2.3.7 Southern Blot</td>
<td>58</td>
</tr>
<tr>
<td>2.3.8 Polymerase chain reaction</td>
<td>58</td>
</tr>
<tr>
<td>2.3.8a Purification of TAQ polymerase</td>
<td>58</td>
</tr>
<tr>
<td>2.3.8b Taq polymerase activity assay</td>
<td>59</td>
</tr>
<tr>
<td>2.3.8c PCR of a whole cDNA library (50,000 clones)</td>
<td>60</td>
</tr>
<tr>
<td>2.3.9 Probe labelling</td>
<td>61</td>
</tr>
<tr>
<td>2.3.10 Hybridisation conditions</td>
<td>64</td>
</tr>
<tr>
<td>2.3.10c Same species hybridisations using long probes</td>
<td>65</td>
</tr>
<tr>
<td>2.3.11 Automated sequencing</td>
<td>65</td>
</tr>
</tbody>
</table>

### CHAPTER 3

OLIGONUCLEOTIDE FINGERPRINTING

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTRODUCTION</td>
<td>69</td>
</tr>
<tr>
<td>3.1. OLIGONUCLEOTIDE FINGERPRINTING</td>
<td>70</td>
</tr>
<tr>
<td>3.1.1. Experimental parameters</td>
<td>70</td>
</tr>
<tr>
<td>3.1.2. Data analysis methods</td>
<td>77</td>
</tr>
<tr>
<td>3.1.3. Clustering results</td>
<td>81</td>
</tr>
<tr>
<td>3.1.4. Methods for evaluation of clustering accuracy</td>
<td>86</td>
</tr>
<tr>
<td>3.2. THEORETICAL OLIGONUCLEOTIDE FINGERPRINTING</td>
<td>102</td>
</tr>
<tr>
<td>3.3. CONCLUSIONS</td>
<td>103</td>
</tr>
</tbody>
</table>

### CHAPTER 4

EST SEQUENCING, ANALYSIS AND DATA INTERPRETATION HOW MANY AMPHIOXUS GENES HAVE BEEN ISOLATED?

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTRODUCTION</td>
<td>107</td>
</tr>
<tr>
<td>4.1. Automated processing of sequence data</td>
<td>108</td>
</tr>
<tr>
<td>4.2. Prediction of coding potential</td>
<td>109</td>
</tr>
<tr>
<td>4.3. Functional classification of EST sequences</td>
<td>111</td>
</tr>
<tr>
<td>4.4. Conclusions</td>
<td>115</td>
</tr>
</tbody>
</table>

### CHAPTER 5

AMPHIOXUS GENES EXPRESSED AT GASTRULA AND NEURULA STAGES

<table>
<thead>
<tr>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>117</td>
</tr>
</tbody>
</table>
## Table of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>FIG. 1.1</td>
<td>PHYLOGENETIC TREE OF DEUTEROSTOME PHYLA</td>
<td>20</td>
</tr>
<tr>
<td>FIG. 1.2</td>
<td>MODE OF GASTRULATION IN AMPHIOXUS</td>
<td>23</td>
</tr>
<tr>
<td>FIG. 1.3</td>
<td>AMPHIOXUS EARLY NEURULA STAGE</td>
<td>24</td>
</tr>
<tr>
<td>FIG. 1.4</td>
<td>PRINCIPLE OF OLIGONUCLEOTIDE FINGERPRINTING</td>
<td>32</td>
</tr>
<tr>
<td>FIG. 1.5</td>
<td>OUTLINE OF THE OLIGONUCLEOTIDE FINGERPRINTING PROCEDURE</td>
<td>33</td>
</tr>
<tr>
<td>FIG. 2.1</td>
<td>MAP OF THE PLASMID VECTOR P SPORT1</td>
<td>48</td>
</tr>
<tr>
<td>FIG. 2.2</td>
<td>PICKING AND THERMOCYCLER ROBOTIC SYSTEMS</td>
<td>52</td>
</tr>
<tr>
<td>FIG. 2.3</td>
<td>SPOTTING ROBOTIC SYSTEM</td>
<td>54</td>
</tr>
<tr>
<td>FIG. 3.1</td>
<td>AVERAGE INSERT SIZE OF BOTH AMPHIOXUS CDNA LIBRARIES</td>
<td>71</td>
</tr>
<tr>
<td>FIG. 3.2</td>
<td>HYBRIDISATION OF PCR FILTERS WITH PSport11mer</td>
<td>76</td>
</tr>
<tr>
<td>FIG. 3.3</td>
<td>OLIGONUCLEOTIDE HYBRIDISATION IMAGE</td>
<td>78</td>
</tr>
<tr>
<td>FIG. 3.4</td>
<td>CLUSTER SIZE DISTRIBUTION IN THE NEURULA STAGE LIBRARY ACCORDING TO bfl26.1RES CLUSTERING.</td>
<td>83</td>
</tr>
<tr>
<td>FIG. 3.5</td>
<td>CLUSTER SIZE DISTRIBUTION IN THE GASTRULA STAGE LIBRARY ACCORDING TO bfg_dr_am_1.8db AND BFLG.ALL.RES CLUSTERINGS</td>
<td>85</td>
</tr>
<tr>
<td>FIG. 3.6</td>
<td>SCHEMATIC REPRESENTATION OF THE CALMODULIN CLUSTERS IN BOTH AMPHIOXUS LIBRARIES</td>
<td>87</td>
</tr>
<tr>
<td>FIG. 3.7</td>
<td>LONG PROBE HYBRIDISATION ON COLONY FILTERS</td>
<td>90</td>
</tr>
<tr>
<td>FIG. 3.8</td>
<td>ALIGNMENT OF CALMODULIN CLONES</td>
<td>94</td>
</tr>
<tr>
<td>FIG. 3.9</td>
<td>CALMODULIN CLUSTERS FROM BOTH AMPHIOXUS LIBRARIES</td>
<td>95</td>
</tr>
<tr>
<td>FIG. 3.10</td>
<td>COMPARISON BETWEEN SEQUENCE AND FINGERPRINTING CLUSTERS</td>
<td>98</td>
</tr>
<tr>
<td>FIG. 3.11</td>
<td>ALIGNMENT OF MUSCLE ACTIN CLONES</td>
<td>100</td>
</tr>
</tbody>
</table>
List of Tables

TABLE 1.1. AMPHIOXUS DEVELOPMENTAL STAGES ................................................................. 22

TABLE 3.1. THEORETICAL CLUSTERING OF ALL AMPHIOXUS SEQUENCES DEPOSITED IN THE GENBANK .............................................................................................................. 73

TABLE 3.2. CLUSTERING RESULTS FOR BOTH AMPHIOXUS cDNA LIBRARIES ..................... 82

TABLE 3.3. EVALUATION OF CLUSTERING THROUGH HYBRIDISATION RESULTS .................. 93

TABLE 4.1. ASSESSMENT OF THE ACCURACY OF THE MZEF (M) AND GENSCAN (G) CODING PREDICTION PROGRAMS ON AMPHIOXUS SEQUENCES ...................................................... 110

TABLE 4.2. FUNCTIONAL CLASSIFICATION OF AMPHIOXUS SEQUENCES .......................... 112

TABLE 4.3. EXAMPLES OF SWISSPROT KEYWORDS USED FOR THE FUNCTIONAL CLASSIFICATION OF THE SEQUENCES ................................................................. 113

TABLE 5.1. OVERVIEW OF TRANSCRIPTS EXPRESSED IN THE AMPHIOXUS NEURULA STAGE. 118

TABLE 5.2. OVERVIEW OF TRANSCRIPTS EXPRESSED IN THE AMPHIOXUS GASTRULA STAGE 121

TABLE 5.3. DIFFERENCES IN THE EXPRESSION LEVEL OF TRANSCRIPTS BASED ON THE SIZE OF THE FINGERPRINT CLUSTER .................................................................................... 123

TABLE 5.4. IDENTIFIED GENES IN AMPHIOXUS THAT AFFECT THE FUNCTION OF THE NERVOUS SYSTEM ........................................................................................................ 126
### Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>bp</td>
<td>base pairs</td>
</tr>
<tr>
<td>BCIP</td>
<td>5-bromo-4-chloro-3-indolylphosphate p-toluidine salt</td>
</tr>
<tr>
<td>BFL26</td>
<td><em>Branchiostoma floridae</em> 26 hr</td>
</tr>
<tr>
<td>BFLG</td>
<td><em>Branchiostoma floridae</em> gastrula stage</td>
</tr>
<tr>
<td>BMP</td>
<td>bone morphogenetic protein</td>
</tr>
<tr>
<td>BSA</td>
<td>bovine serum albumin</td>
</tr>
<tr>
<td>cDNA</td>
<td>complementary deoxyribonucleic acid</td>
</tr>
<tr>
<td>Ci</td>
<td>Curie</td>
</tr>
<tr>
<td>c.v.</td>
<td>cerebral vesicle</td>
</tr>
<tr>
<td>DEPC</td>
<td>diethylpyrocarbonate</td>
</tr>
<tr>
<td>ddNTP</td>
<td>2’, 3’-dideoxyribonucleoside 5’-triphosphate</td>
</tr>
<tr>
<td>DNA</td>
<td>deoxyribonucleic acid</td>
</tr>
<tr>
<td>dNTP</td>
<td>2’-deoxyribonucleoside 5’-triphosphate</td>
</tr>
<tr>
<td>DTT</td>
<td>dithiothreitol</td>
</tr>
<tr>
<td>DIG</td>
<td>digoxigenin</td>
</tr>
<tr>
<td>EDTA</td>
<td>ethylenediamine tetra-acetic acid</td>
</tr>
<tr>
<td>EGTA</td>
<td>ethylene glycol-bis(β-amino-ethyl ether) N,N,N’,N’-tetraacetic acid</td>
</tr>
<tr>
<td>EST</td>
<td>expressed sequence tag</td>
</tr>
<tr>
<td>FISH</td>
<td>fluorescence <em>in situ</em> hybridisation</td>
</tr>
<tr>
<td>hr</td>
<td>hour</td>
</tr>
<tr>
<td>IPTG</td>
<td>isopropyl β-D-thiogalactosidase</td>
</tr>
<tr>
<td>kb</td>
<td>kilobase pair</td>
</tr>
<tr>
<td>lt</td>
<td>litre</td>
</tr>
<tr>
<td>MCS</td>
<td>multiple cloning site</td>
</tr>
<tr>
<td>MHC</td>
<td>major histocompatibility complex</td>
</tr>
<tr>
<td>min</td>
<td>minute</td>
</tr>
<tr>
<td>MOPS</td>
<td>3[N-morpholino]propanesulfonic acid</td>
</tr>
<tr>
<td>mRNA</td>
<td>messenger RNA</td>
</tr>
<tr>
<td>M.W.</td>
<td>molecular weight</td>
</tr>
<tr>
<td>NBT</td>
<td>nitroblue tetrazolium chloride</td>
</tr>
<tr>
<td>OD</td>
<td>optical density</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>PAC</td>
<td>P1 artificial chromosome</td>
</tr>
<tr>
<td>PBS</td>
<td>phosphate buffer saline</td>
</tr>
<tr>
<td>PCR</td>
<td>polymerase chain reaction</td>
</tr>
<tr>
<td>PEI</td>
<td>polyethyleneimine</td>
</tr>
<tr>
<td>PMSF</td>
<td>phenylmethylsulphenylfluoride</td>
</tr>
<tr>
<td>RH</td>
<td>radiation hybrid</td>
</tr>
<tr>
<td>RNA</td>
<td>ribonucleic acid</td>
</tr>
<tr>
<td>RNase</td>
<td>ribonuclease</td>
</tr>
<tr>
<td>rpm</td>
<td>rounds per min</td>
</tr>
<tr>
<td>RT</td>
<td>room temperature</td>
</tr>
<tr>
<td>RZPD</td>
<td>Resource Zentrum Primary Database</td>
</tr>
<tr>
<td>SBH</td>
<td>sequencing by hybridisation</td>
</tr>
<tr>
<td>SDS</td>
<td>sodium dodecylsulfate</td>
</tr>
<tr>
<td>sec</td>
<td>second</td>
</tr>
<tr>
<td>SSC</td>
<td>sodium chloride-sodium citrate</td>
</tr>
<tr>
<td>ssDNA</td>
<td>salmon sperm DNA</td>
</tr>
<tr>
<td>TAE</td>
<td>Tris-acetate-EDTA</td>
</tr>
<tr>
<td>Taq</td>
<td><em>Thermus aquaticus</em></td>
</tr>
<tr>
<td>TBE</td>
<td>Tris-borate-EDTA</td>
</tr>
<tr>
<td>TIF</td>
<td>tag image file format</td>
</tr>
<tr>
<td>Tm</td>
<td>melting temperature</td>
</tr>
<tr>
<td>tRNA</td>
<td>transfer RNA</td>
</tr>
<tr>
<td>Tris</td>
<td>Tris(hydroxymethyl)aminomethane</td>
</tr>
<tr>
<td>TIF</td>
<td>tag image file format</td>
</tr>
<tr>
<td>UTR</td>
<td>untranslated region</td>
</tr>
<tr>
<td>UV</td>
<td>ultraviolet</td>
</tr>
<tr>
<td>v/v</td>
<td>volume/volume</td>
</tr>
<tr>
<td>w/v</td>
<td>weight/volume</td>
</tr>
<tr>
<td>X-gal</td>
<td>5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside</td>
</tr>
<tr>
<td>YAC</td>
<td>yeast artificial chromosome</td>
</tr>
</tbody>
</table>
Acknowledgements

I would like to thank my supervisor Hans Lehrach for his encouragement and endless interest in this project. His belief in amphioxus as an important evolutionary organism has kept this project alive. I am grateful to Linda and Nic Holland for providing me with the RNA for the construction of the two embryonic cDNA libraries, helping me with the summer amphioxus collections in Tampa and teaching me amphioxus anatomy and embryology.

Moreover, I am grateful to Matthew Clark for making the two amphioxus libraries and for the endless enthusiastic discussions and constructive arguments throughout this project.

I would like to thank all the people who have developed the image analysis software used in this project, in particular Ruben Schattevoy, Marcus Kitzmann, Henrik Seidel and Mario Drunkowski. I am indebted to Ralf Herwig for rere- and reclustering my fingerprinting data patiently, numerous times and to Steffen Henning for his untiring assistance in analysing the sequencing data.

The robotic systems used in this project were developed, improved and are maintained by many people initially at the Imperial Cancer Research Fund in London and later at the Max-Planck Institute in Berlin. I would like to thank all these people, in particular John Curtis, Ali Ahmadi, Martin Horn and Thomas Przewieslik.

I am grateful to Dirk Schudde and Elizabeth Brundke for their excellent technical help.

Finally, my special thanks to Leo Schalkwyk who has encouraged me during the writing of this thesis.

It has been a pleasure working at the same lab with many people initially at the Genome Analysis lab at Imperial Cancer Research Fund, London and then at the Max-Planck Institute for Molecular Genetics in Berlin in particular Elmar Maier, Igor Ivanov, Carola Burgtorf, Michael Wiles, Dolores Cahill, John O’Brien, Uwe Radelof, Pia Aanstad and Alberto Musa. Especially, I want to thank Kieran and Brian, for making my life in the lab so much brighter and Albert for his patience and encouragement during the writing of this thesis.

Finally, I want to thank my parents and my sister for being always present although they were miles away.
Chapter 1

Introduction

This thesis describes the construction of a catalogue of genes from amphioxus (Branchiostoma floridae), an invertebrate chordate closely related to the vertebrate lineage by the method of oligonucleotide fingerprinting.

The first part of this chapter describes that the large amount of information that can now be quickly generated is best interpreted through its use in a combination of methods. This, coupled with the expansion in multiple organisms and cross-species comparisons is a powerful approach for identifying candidate disease genes. The choice, though, of model organisms is strongly influenced towards those that offer the possibility of genetic manipulation. However, the large evolutionary distance between those organisms along with the genetic redundancy that was introduced during the evolution of their own lineage, results in multiple paralogs which hamper such comparisons.

The next part of this chapter introduces amphioxus, as an organism that will strongly complement the above approach in the next few years. Amphioxus embryology is described as well as the nature of questions that can be answered by studying amphioxus and the reason for the need of a large scale project identifying the majority of genes in this organism.

In the third part of this chapter the oligonucleotide fingerprinting technique which is the method used for isolating amphioxus genes during this study, is described along with the progress that has been made within the five years since its first application as well as its comparison to other large scale-screening methods.

1.1. The importance of cross-species comparisons

1.1.1. Finding genes faster and understanding their function

The progress in the field of molecular biology through the development of new techniques and the advance of traditional methods has made gene isolation and function
analysis possible while the coupling of those methods with automation over the last few years has increased the throughput of information that can be generated and as a result the nature of questions that can be posed. Moreover, it has become apparent that new ways of interpreting this wealth of information should be found in order to fully exploit it.

As an example, complete sequencing of whole genomes is now an achievable target. The sequencing initially of viruses and some bacteria was followed by the 12 Mb of the yeast S. cerevisiae, the first eukaryote to be sequenced. The availability of the 100 Mb of Caenorhabditis elegans genome since the end of last year enables researchers to access all the information stored in the genome of a multicellular eukaryote [C.elegans Sequencing Consortium, 1998 #245]. Moreover, genome sequencing of a number of model organisms such as Drosophila melanogaster is in progress while 9% of the human genome has been sequenced with the aim of finishing in 2005.

The rapid generation of sequence information led to the realisation that ways of interpreting, exploring and exploiting this data have to be found but also that the knowledge of only the sequence itself is not enough. What is the information that genomic sequence can offer and to what extent it can serve as a starting point for experiments? Based on the C. elegans experience, such information will enable us to answer questions such as the importance of a gene's location and context within the genome e.g. which genes are clustered together, the variation in the gene density across specific chromosome regions or the percentage of genes that have homologs in other species thus allowing to construct phylogenetic trees. Moreover, it enables us to predict possible gene function through the prediction of structure and the presence of specific domains.

The sequence alone however, cannot offer any clue about the tissue that a specific gene is expressed, the time of expression, the pathway that it is involved or the genes that it interacts with. This information can be generated by a combination of molecular and genetic tools available in a model organism.

For example EST sequencing of clones selected from a tissue specific cDNA library indicates the range of transcripts expressed in this tissue. ESTs (expressed sequence tags, ESTs) are single-pass, partial 5' or 3' sequences of randomly selected cDNA clones (Adams et al., 1992; Adams et al., 1991; Khan et al., 1992; Okubo et al., 1992). The use of the above cDNA clones in a gene expression screen like whole-mount in situ hybridisation screen can give information about the spatial and temporal expression of this clone. Large-scale screens are currently carried out in Drosophila (Casey et al., ), Xenopus (Gawantka et al., 1998; Niehrs, 1997), zebrafish (H.Lehrach, B.Thisse, unpublished), mouse (B.Herrman, unpublished), Medaka (J. Wittbrot, unpublished), using randomly selected cDNA clones
combined with tag sequencing for the cDNA clones that show localised expression. Besides, the identification of tissue specific gene-markers, in situ screens have been useful in identifying genes that share complex expression patterns and probably function (Gawantka et al., 1998). Finally, in situ screens can be combined with large scale mutagenesis screens. Comparison of expression patterns with mutant phenotypes may lead to the identification of genes involved in the same pathway.

The above approach can be complemented by the mapping of 3’ ESTs which represent the variable 3’UTR on physical, radiation hybrid (RH) maps or by fluorescence in situ hybridisation (FISH) on chromosomes. Mapping of ESTs on radiation hybrid maps will also help identifying genes responsible for a mutant phenotype (Geisler et al., 1999).

However, the results from combined functional analysis of genes within a specific organism become more meaningful once they are integrated with data from other organisms.

1.2.2. Cross-species comparisons in gene identification and function analysis

Cross-species comparisons, even between distantly related species, have been proved to be an effective tool in identifying genes and studying their function. This strategy is based on the observation that functionally significant regions of the genome are conserved during evolution.

There are cases of mutations in genes homologous between D. melanogaster and human that cause related phenotypes in spite of the large evolutionary distance between those two species (Quiring R. et al., 1994). Based on this assumption (Banfi et al., 1996) retrieved 66 human ESTs showing high homology to D. melanogaster mutant genes. The potential involvement of those ESTs in human inherited disorders was tested by determination of their map position using both FISH and RH mapping and comparing the phenotype of the D. melanogaster mutant with the phenotype of the human disease mapped in the same region. As a result, the EST homologous for example to the D. melanogaster retinal degeneration B gene was mapped to an area of chromosome 11 where at least three types of human retinopathies were assigned.

The above is an example where gene identification can be aided through study of the mutant phenotype and therefore, it assumes the possibility of genetic manipulations. However, in the last few years, comparisons begin to extend in organisms that are attractive
for reasons other than good genetics, easy access etc. The pufferfish (Fugu rubripes) was proposed as a model organism for studying the evolution of the vertebrate genome. The Fugu genome is estimated to be around 400Mb, 7.5 times smaller than the human genome (Brenner et al., 1993). Moreover, random sequencing indicated that both the gene size and intron size are reduced which led to the speculation that its genome is compact reflecting the original vertebrate genome. Therefore, it would be easier to isolate the homologous to human genes. The above observation has been verified in many cases of genes where the Fugu homologue was isolated. For example, the human Huntington’s disease gene is 170kb while its Fugu homologue is 23kb consisting of the same number of exons. It is interesting in this case that the first coding exon which is the site of the disease-causing triplet repeat, is highly conserved (Baxendale et al., 1995). However, this is not always the rule as there are examples of genes that have retained similar size to human introns like the dopamine receptors (Macrae and Brenner, 1995).

1.1.3. Genome Duplications

Cross-species comparisons become meaningful in the case that the genes that are compared are true orthologs and not paralogs. Two proteins are orthologous if they have evolved from the same ancestor while paralogs have resulted from lineage specific gene duplications and as a result they may have adopted more divergent roles.

How many of the human disease genes have true recognisable orthologs in other model organisms? It was reported that approximately 20% of the studied human disease genes have true orthologs in yeast while 30% in the 50% of the C. elegans genome that was available at the time (Mushegian et al., 1997).

At the same time the percentage of unmatched genes or predicted proteins across multiple model organisms is impressive. 30-35% of all ORFs of the yeast genome are orphans. Are they real genes and if so, what is their nature and function? (Dujon, 1996). Moreover, 36% of the predicted 18,891 C. elegans proteins match corresponding human ones while 34% are thought to be nematode specific. Mushegian et al., 1998, constructed phylogenetic trees of 36 orthologous proteins found in Drosophila, humans, nematodes and yeast selected from genomic data. The result was that 24 proteins supported a tree where nematodes are basal to arthropods and humans (tree A), 11 proteins supported the hypothesis for a clade of arthropods and nematodes basal to humans (tree B) and one protein supported the hypothesis that arthropods are basal to nematodes and humans. In order to exclude the possibility that
tree A was a result of unequal evolutionary rates effect in the used proteins, the pairwise evolutionary distances from human to nematode were compared with those of human to arthropod for each protein. Indeed they showed that proteins that support tree A have evolved more heterogeneously while protein supporting the tree B have evolved homogeneously in all four phyla. Moreover, they showed that proteins supporting the tree A were evolving faster in the lineage from human to nematode than from human to arthropods. Therefore they concluded that *C. elegans* proteins have evolved more rapidly than the *Drosophila* ones.

Increase in the gene number is proved to be possible either by the tandem gene duplication or by duplication of specific chromosome areas (Ohno, 1970).

When genome duplications have occurred during animal evolution? Can we deduce the gene number from genome size and therefore the extent to which specific lineages have accumulated multiple gene paralogs?

Based on genome sizes within the chordates, Ohno, 1970, suggested that at least one round of tetraploidisation occurred in the lineage leading to amniotes (reptiles, birds, mammals), fish and amphibian lineages. He also suggested that an additional genome expansion (by either tetraploidy or tandem duplication) occurred earlier in the chordate lineage, in the common ancestor of cephalochordates and vertebrates and after the divergence of urochordates. The last assumption was based on the observation that the genome size of amphioxus, a cephalochordate is 17% of the human. This genome size although small, as compared to vertebrate genomes, is still three times larger than the urochordate genome, which is the third chordate subphyla.

The result of the study of eight categories of enzymes in both urochordates and cephalochordates by Schmidtke et al., 1979, where similar gene numbers were found for seven of them, has led him to suggest that the gene number between urochordates and cephalochordates must be similar. The isolation of additional amphioxus genes in the meanwhile in fewer copies than in vertebrates favors the hypothesis that vertebrates have more genes than urochordates and cephalochordates as a result of a later gene duplication or polyploidy (Holland, 1996).

While genome size can serve as an indicator of the gene number in prokaryotes, the same cannot be applied in eukaryotes (Bird, 1995). The average genome size of bacteria is 2.6 Mb. Considering *E.coli* gene size and density (0.96 genes /1000 bp) as representative, although its 4.5 Mb genome is larger than average, the average gene number in bacteria is estimated to be 2,600 (Bird, 1995). However, eukaryotic genome size is not a reliable indicator of the gene number. Nevertheless, it has been observed that eukaryotic gene numbers are similar in multiple animal phyla (Bird, 1995). *D. melanogaster* is estimated to have 16,000 genes in a 160 Mb genome, the recently sequenced *C. elegans* 19,000 genes in 97 Mb,
**Strongylocentrotus purpuratus** less than 25,000 genes in 800 Mb and the urochordate *C. intestinalis* 15,500 genes in 162 Mb (Simmen et al., 1998).

The gene numbers predicted for vertebrates are significantly higher. The pufferfish (*Fugu rubripes*), the zebrafish (*Danio rerio*), and human genomes are thought to have approximately 70,000 (Antequera and Bird, 1993) in a genome of 400 Mb (Brenner et al., 1993), 700 Mb and 3,000 Mb correspondingly. The large differences in the genome size, especially between vertebrates is due to repetitive DNA which increases with the increase in the genome size (Lewin, 1990).

What is the extent of gene duplication in the above organisms? 46% of the 4,100 *E.coli* genes (Koonin et al., 1995) and 14% of the 5,800 *S.cereviciae* genes are gene duplicates. There is no data for the recently sequenced *C. elegans* genome and neither for the *Fugu*, zebrafish, mouse and human genomes. The extent though that duplication has occurred in these lineages can already be traced through the variation in the number of members of multigene families.

Homeobox genes of the *Hox* class, is one of the best studied homeobox gene classes across multiple phyla. This is due to their conserved role in pattern specification along the anteroposterior (AP) body axis of many organisms and thus their implication in the evolution of morphology. These genes are clustered and they are expressed in a conserved order along the AP axis which coincides with their order in the genome. *C. elegans* (Bürglin and Ruvkun, 1993; Bürglin et al., 1991) and *D. melanogaster* (Kaufman et al., 1990; Lewis, 1978) have a single *Hox* cluster which is split in *Drosophila*, echinoderms have one *Hox* cluster which is devoid of genes homologous to anterior paralogous group of *Hox* genes (Dolecki et al., 1989; Martinez et al.; Wang et al., 1990), the hemichordate *Saccoglossus Kowalevsky*, has a single *Hox* cluster of nine genes (Pendleton et al., 1993). Within the chordate lineage: i) the ascidian *Ciona intestinalis* (Di Gregorio et al., 1995) has 3 *Hox* genes two of which are linked, ii) the cephalochordate amphioxus has a single *Hox* cluster containing at least 12 *Hox* genes. Each *Hox* gene is homologous to a different paralogous group of the mammalian *Hox* genes. It is striking the conservation of genomic organisation of this cluster iii) the jawless fish hagfish and lamprey, are reported to have four and three respectively ((Holland and Garcia-Fernandez, 1996) and references therein). Finally, mammals have 38 *Hox* genes organised in 4 *Hox* clusters on different chromosomes (McGinnis and Krumlauf, 1992). It has been showed that not all *Hox* clusters contain representative genes from all the 13 paralogous groups. In Fugu 4 *Hox* gene clusters (Aparicio et al., 1997) have been reported and it is expected that there is a fifth one (Holland, 1997) while zebrafish has at least six clusters (Amores 1998). Several conclusions are drawn from the above data (for an extended review (Holland and Garcia-Fernandez, 1996; Holland et al., 1994)): i) all invertebrates have
a single Hox cluster ii) there is a Hox cluster duplication after the divergence of amphioxus followed by a second one close to gnathostome origins iii) duplication of the vertebrate Hox cluster was followed by loss of some Hox genes. Study of the genomic organisation of the Fugu Hox clusters showed that nine genes have been lost while there is a new group-2 paralogue and pseudo-gene remnants of group-1 and group-3 paralogues in the Hoxc cluster (Aparicio et al., 1997).

The above pattern of double gene duplication can be also followed through examples of other multigene families. Such comparison will be more interesting once there is additional molecular data on lampreys which occupy the intermediate position between the two expected genome duplications. All the gene examples described below refer to genes identified in Drosophila, amphioxus and the vertebrate model organisms mouse, zebrafish, Xenopus.

Arthropods have one Class I paired box the poxm (poxmeso) (Bopp et al., 1989) while mouse and human have two, the Pax1 (Deutsch et al., 1988) and Pax9 (Neubüser et al., 1995). They contain one 132 aminoacid paired domain and one paired type homeobox. Amphioxus has one class I paired box gene, AmphiPax-1 (Holland et al., 1995).

Mice have two Otx genes, Otx1 and Otx2 (Simeone et al., 1993) homologous to Drosophila orthodenticle (Otd) (Finkelstein et al., 1990). Zebrafish and Xenopus, each have three Otx genes. Amphioxus has one Otx gene, AmphiOtx (Williams and Holland, 1998).

In conclusion, i) cross-species comparisons are more meaningful when orthologous genes are compared, ii) as more sequence data becomes available, the extent that multiple model organisms have lineage specific genes is revealed e.g. 34% of the C. elegans genes are thought to be nematode specific iii) additional individual gene duplications or polyploidy result in paralogous genes which become adopted to a new function. Two genome expansions, one at the origin of the vertebrate lineage followed by one at the gnathostome origin have resulted in a four-fold increase in the gene number. iv) expansion of cross species studies to organisms that are not considered as model will increase the spectrum of constructive comparisons. Especially the study of organisms that occupy the transition point between vertebrates and invertebrates will offer additional information on the mechanisms of genome expansion. Amphioxus, a cephalochordate is thought to be the closest invertebrate relative of the vertebrates. As it was discussed above, the majority of amphioxus genes isolated so far exist in one copy while in vertebrates, the same gene family consists of multiple members.
1.2. Amphioxus

1.2.1. Amphioxus phylogeny

Amphioxus or lancelet is the common name of more than 20 species of marine invertebrates living in tropical and temperate seas worldwide (Poss and Boschung, 1996), that comprise the chordate subphylum of cephalochordates. The genus name Amphioxus referring to the animal being pointed at both ends (amphis = both and oxys = sharp) was replaced by the name Branchiostoma. The later refers to the tentacles surrounding the mouth that had been falsely recognised as respiratory elements (branchiae).

18S rDNA data suggests that cephalochordates are the closest chordate subphylum to vertebrates (Wada and Satoh, 1994) (Fig.1.1). This conclusion is also supported by phylogenetic analysis based on the aldolase and triose phosphate isomerase sequences, two proteins that are thought to have evolved at constant rate over long periods (Nikoh et al., 1997). Based on the last observation, these two enzymes were used as molecular clocks for inferring divergence times. Thus, it was shown that the divergence time of cephalochordates and vertebrates was 700 million years ago just after the separation of protostomes from deuterostomes. This data agrees with the estimation based on fossil records. The oldest chordate fossils that have been found are of the early Cambrian period (520 million years ago) (Chen et al., 1995). The morphological similarities between these fossils and the living species of Amphioxus is striking.

Lancelets were initially discovered in 1770 when they were classified as mollusks (Pallas, 1774) until their rediscovery in 1830 when they were classified as primitive fish (Drach, 1948; Holland and Holland, 1989). Within the family of lancelets, from 50 names about 29 have been recognised as valid for representing different species (Poss and Boschung, 1996) based on morphological criteria such as the number of myotomes and fin chambers. Three genera have been described which differ on the arrangement of the gonads: The Branchiostoma (with bilateral series of gonads) and the Epigonichthys (with a single asymmetric series) and the Amphioxidies which may be larvae which fail to undergo metamorphosis and live for extended period in the plankton.

Amphioxus has been popular with embryologists for over a century (Conklin, 1932; Wilson, 1893) as a model for comparing and understanding the processes that lead to the formation of chordate structures because its body organisation seems intermediate in
Fig. 1.1. Phylogenetic tree of the Deuterostome phyla along with their position in respect to the postulated whole genome duplications and their hypothetical geological time (based on fossil record) of their appearance. (The branches showing the multiple vertebrates species do not indicate phylogenetic relationship). The part of the figure that shows the phylogenetic relationships between chordates (cephalochordates, urochordates and vertebrates and the other two major deuterostome phyla, hemichordates and echinoderms) is adopted from Holland and Garcia-Fernandez., 1996 while the information on the geological time of the species appearance was adopted from (Kershaw, 1983).
complexity between the tadpole larvae of tunicates and that of the jawless vertebrates, the hagfish and lamprey.

1.2.2. Amphioxus embryology

Kowalevsky (Holland and Holland, 1989; Kowalevski, 1867) claimed that the early lancelet development resembles that of an invertebrate deuterostome while the later embryology and larval development are more like the vertebrates. Wilson (Wilson, 1893) concluded that blastomere fates become determined in ascidians, lancelets and echinoids by the second, third and fourth cleavages respectively. Conklin (Conklin, 1932) in contrast to Wilson believed that all axes and poles of the future larva are irreversibly determined at or before the first cleavage and concluded that amphioxus development is mosaic like ascidians.

The classic figures of amphioxus development from the above papers have been included in every textbook describing amphioxus development. The development of techniques such as electron microscopy, dye tracing as well as the progress in molecular biology methods coupled with the possibility of induced spawning allowing access to earlier stages (Holland and Holland, 1989) have enabled more detailed studies (Hirakow and Kajita, 1991; Holland and Holland, 1992; Stokes and Holland, 1995). The result was that the general plan of development was as described. A contrasting observation worth noting was that Conklin's (Conklin, 1932) conclusion that lancelets show the same with ascidians pattern of ooplasmic segregation between fertilization and pronuclear fusion was recently challenged (Holland and Holland, 1992). In ascidians there are five distinct areas separated (larval tail muscles, mesodermal crescent, ectoderm, endoderm, notochord and neural plate) which later are to be incorporated into specific embryonic tissues.

Although the method of induced spawning by brief electroshock when the animals are ripe led to the description of the developmental stages, amphioxus research is still impaired by the inability to keep the animals in culture or inducing frequent regular spawning.

The sequence of developmental stages is described briefly below and in Table 1.1. (Holland and Holland, 1993), in order to account for the structures that have been formed and processes that have been taken place during the stages from which the libraries used in this study were prepared. The following description and figures are adopted from (Browder et al., 1991; Conklin, 1932; Whittaker, 1997) and concern mainly the B. floridæ species: i) cleavage is radial and determinate and starts after 55min (for the B. floridæ species at 25°C)
with the first cleavage occurring meridionally through the animal-vegetal axis (as defined by the second polar body). The cleavages cleavages are synchronous, at least until the 7th cleavage (128 cells) that could be followed. Difference in sizes of cells between the animal and vegetal pole are observed at the 8 cell stage and persist through blastula and gastrula. ii) blastula (begins at the seventh cleavage (128 cell-stage) where a single layer of cells form a hollow blastocoel. iii) gastrulation begins with the flattening of the vegetal pole cells forming the endodermal plate which subsequently sink into the blastocoel to form the wall of the archenteron. At the point of invagination the blastopore forms, with its dorsal lip made up of presumptive notochord cells and the ventral containing mesoderm origin cells. At late gastrula the internal notochordal plate cells flatten and the above cells of the ectoderm and neuroectoderm begin to form the neural pore. iv) Hatching occurs during neurulation which begins with the closure of the neural plate. Neural tube formation occurs when the ectodermal cells on either side of the neural plate move toward the dorsal midline until they fuse starting gradually from posterior to anterior. During this time the notochord and first muscle blocks are developing from chordamesodermal cells on the dorsal side of the archenteron and they are complete at a late-neurula-early-larva stage at about the time of hatching (9 somite pairs). v) In the one day larva (15 somites), the mouth, the club shaped gland (whose function during this stage is unknown and disappears during metamorphosis) on the opposite side of the mouth and the first gill slit below the third somites have been formed. Moreover, during this time there is a shifting in the intersomite boundaries between left and right side so that they become asymmetric. vi) Feeding starts at 2 days when the second gill slit and the anus have been formed while metamorphosis occurs at 40-50 days. Several movements of structures that have been already formed previously, occur during metamorphosis. Thus, there is an increase in the number of the gill slits but also migration from the site of the formation to their final position. Similarly, the mouth opening migrates from the left wall ventrally.

<table>
<thead>
<tr>
<th>Time</th>
<th>Stage</th>
</tr>
</thead>
<tbody>
<tr>
<td>55min</td>
<td>2-cell stage</td>
</tr>
<tr>
<td>75min</td>
<td>4-cell stage</td>
</tr>
<tr>
<td>95min</td>
<td>8-cell stage</td>
</tr>
<tr>
<td>3.5hr</td>
<td>Blastula</td>
</tr>
<tr>
<td>5.5hr</td>
<td>Gastrula</td>
</tr>
<tr>
<td>10.0hr</td>
<td>Neurula hatches</td>
</tr>
<tr>
<td>38.0hr</td>
<td>Larval mouth opens</td>
</tr>
<tr>
<td>18d</td>
<td>Larval metamorphosis begin</td>
</tr>
<tr>
<td>23d</td>
<td>Metamorphosis complete</td>
</tr>
</tbody>
</table>
At this time amphioxus has the main prominent features shared with vertebrates: i) a notochord which runs dorsally along the full length of the animal extending rostrally beyond the end of the nerve cord, ii) a dorsal hollow nerve cord with slightly enlarged cerebral vesicle at the anterior end, iii) segmented muscle blocks, iv) gill slits opening from the pharynx.

Selectively, some anatomical features found in amphioxus thought to be primitive precursors of organs found in vertebrates (Jefferies, 1986), are: i) the midgut diverticulum, which is a digestive gland that secretes enzymes into the intestine, a primitive homologue of the vertebrate liver ii) the endostyle, an organ which secretes mucus in the pharynx which functions as part of the feeding mechanism, in passing the food to the oesophagus. It is also found both in the ascidian and the lamprey larvae. iii) the circulatory system is highly developed and it is organised in a pattern that resembles that of blood vessels found in fish. There is no heart and contractility is achieved by three of the major ventral vessels being lined with myogenetically active myoepithelial cells. The blood which is colorless contains proteins but no circulatory cells, haemoglobin or other respiratory protein.

Cephalochordates though lack some major structures which prevents them from being classified as vertebrates. The most striking example is the lack of a defined cranium region which leads to the question whether the vertebrate head is the evolution of an amphioxus-like condition or whether the innovation of neural crest cells resulted in an entirely new structure attached at the anterior end of the nerve cord. The neural crest cells, thought to be unique to vertebrates, is a mass of embryonic cells that form at the border between the neural plate and the epidermis. They migrate along the embryo to form a range of structures mostly ones.
associated with the cranium and the nervous system like sensory neurons, melanocytes, smooth muscle, cartilage and bone (Baker and Bronner-Fraser, 1997).

I discuss below what is already known about amphioxus anterior nerve cord and nervous system because it is of particular interest to correlate the appearance of this structure in vertebrates with the genes that are involved in its formation and the existence of those genes in amphioxus. Such account will also point out the nature of questions that can be solved through studying amphioxus but it will also stress the genes-information that is still missing. In Chapter 5 additional genes expressed in nervous system structures, as indicated by their vertebrate counterparts, that have identified as result of this study are listed.

**Fig. 1.3.** Top: early neurula stage where the muscle blocks start forming. Middle: Early neurula stage 9.5hrs (2 somite pairs) Bottom: one day larva, 24 hrs (15 somite pairs, 1 gill slit) (After (Conklin, 1932) and (Berill, 1987))

### 1.2.3. Amphioxus as an evolutionary model system

#### 1.2.3.a. The anterior part of the amphioxus nerve cord and its equivalence to the vertebrate brain

Electron microscope reconstruction led to the observation that although elementary amphioxus anterior neural tube resembles the vertebrate brain with its three main regions being, the anterior cerebral vesicle (c.v.) including the frontal eye, the balance organ and the infundibullar cells which correlates to the floor of diencephalon, posterior c.v., including the
lamellar body thought to be homologue of epiphysis and the primary motor center (PMC) thought to be the primary locomotor center, consisting of neurons related to the vertebrate ones for reticular formation and somatic motor nuclei located at the midbrain and hindbrain (Lacalli, 1996b). Based on the above reconstructions Lacalli, 1996., proposed that the amphioxus frontal eye is equivalent of the vertebrate paired eyes. Hypothesis later supported by: i) the neural expression of amphioxus Otx gene to two clusters of cells in the anterior cerebral vesicle in the early larva (36hrs) (Williams and Holland, 1996; Williams and Holland, 1998) but also expression of the AmphiPax 6 in receptor cells and/or neurons of the frontal eye (Glardon et al., 1998) ii) Distal-less (Dll) expression (Holland et al., 1996) in the c.v. which is an indication of a part homologous to forebrain as the mouse Dlx1 and Dlx2 (Simeone et al., 1994) are expressed in the diencephalon but also extend in the ventral telencephalon, iii) amphioxus BF-1 expression in the anterior neural tube in cells underneath the eye spot which might be implicated in visual and/or olfactory perception judging from functional data in mice. Olfactory bulb is a major component of telencephalon (Toresson et al., 1998). BF-1 is a forkhead containing transcription factor ortholog to Drosophila sloppy-paired (slp-1/2) that has attracted interest due to its involvement in the development of telencephalon.

Finally, comparison of expression limits between the amphioxus Hox genes, AmphiHox-3, AmphiHox-1 and the vertebrate 1 and 3 paralogous groups that have anterior limits of expression within the hindbrain, have indicated that an extensive part of the amphioxus anterior nerve cord is homologous to the vertebrate brain (Holland et al., 1992).

The above Hox expression data has lead to the suggestion that the vertebrate hindbrain evolved from an extensive region of the neural tube rather than being a completely novel vertebrate structure (Gans, 1993) while the forebrain from an already distinctive anterior region of the neural tube (Holland, 1996).

1.2.3.b. Insights into evolution of dorsoventral axis specification through the study of amphioxus bone morphogenetic protein BMP2/4

In 1822 Geoffroy Saint-Hilaire observed in crayfish that the organisation of the main body systems like nerve cord, muscle, gut and heart resembled that of a chordate although the dorsoventral axis was inverted so that the ventral side of insects corresponds to the dorsal side of vertebrates (Geoffrey Saint-Hilaire, 1822). His idea which was rejected at the time,
was recently re-evaluated in the light of more molecular and embryological data that is now available (Arendt and Nübler-Jung, 1994; Nübler-Jung and Arendt, 1994). Thus, it is suggested that chordates and articulates (i.e. annelids and arthropods) may share a common ancestor with a ventrally located nerve cord (Arendt and Nübler-Jung, 1994). The morphology of such an ancestor as well as the mechanism of dorsoventral inversion is of major phylogenetic interest since it will indicate the time point that this inversion has occurred. Among the proposed mechanisms of dorsoventral inversion are: 1) that protostome and deuterostome lineages have been derived from an ancestor with a separate mouth and anus which are formed by the elongation of the blastopore along the ventral neurogenic side. Protostomes have retained this arrangement while deuterostomes have replaced the old mouth with a new one positioned on the opposite dorsal side. 2) the common ancestor had the mouth derived from the blastopore and a defined ventral side. Protostomes again generate the mouth by the elongation of blastopore along the ventral side. Deuterostomes create a new mouth on the dorsal side (Lacalli, 1996a).

Dorsoventral patterning is based on the antagonistic interaction between dorsal and ventral signals. Vertebrate BMP4 and Drosophila dpp from the antineural side of the embryo are antagonised, respectively, by chordin and short gastrulation expressed on the neural side. Initial dorsoventral patterning is restricted to the ectoderm of Drosophila, but occurs simultaneously in the ectoderm and mesoderm of vertebrates. Moreover, studies have indicated that BMP4 and dpp act as active inducers of epidermis at the expense of default-state neuroectoderm (Graff, 1997).

The similarity in the signalling pathways establishing the dorsoventral body axis of Drosophila and vertebrates have been further emphasized by studies showing that BMP4 and Dpp are functionally interchangeable as are chordin and short gastrulation. Thus, substitution of carboxyterminal human BMP4 with dpp can rescue Drosophila Dpp mutants (Padgett et al., 1993). Moreover, dpp injected in Xenopus promotes ventral development but injection of sog (short gastrulation) mRNA can rescue dorsal development (Holley et al., 1995).

As a conclusion, molecular data indicates that the mechanism of dorsoventral patterning is conserved although inverted between arthropods and vertebrates. This suggests that the same patterning system existed in the ancestor from which the arthropod and vertebrate lineages have diverged 600 million years ago for which the name Urbilateria (primitive bilateral animal) was proposed (De Robertis and Sasai, 1996). Since the mechanism and timing of dorsoventral inversion is of major importance in understanding the evolutionary relationships between the animal phyla, study of the dorsoventral patterning molecules like BMP-related
genes in a broad spectrum of animal phyla will help assessing whether dorsoventral inversion occurred during the deuterostome lineage as it was suggested (Arendt and Nübler-Jung, 1994).

Bone morphogenetic proteins (BMPs) originally defined as inducers of bone and cartilage in vivo in vertebrates (Wozney, 1992), are extracellular morphogens some of which are members of the transforming growth factor beta (TGFβ) superfamily. TGFβ related proteins are synthesized as large precursor proteins which are cleaved at an Arg-X-X-Arg (X being any aa) to release a mature C-terminal peptide of 110-140aa. The mature region is the most conserved region between different members of the family and contains a conserved motif of 7-9 cysteines through which it dimerizes with a second TGFβ to form an active signalling molecule (Hogan, 1996). The best known BMP of the TGFβ family is the vertebrate BMP4 which plays a key role in establishing the dorsoventral body axis at the gastrula stage and also helps pattern a wide range of organs and tissues (Hogan et al., 1994).

To date BMP-related genes have been studied only for Drosophila, C. elegans, vertebrates and ascidians (Gelbart, 1989; Hwang et al., 1997; Koster et al., 1991; Miya et al., 1996; Nikaido et al., 1997). The ascidian homologue of both vertebrate BMP2 and BMP4, functions in establishing dorsoventral polarity in the ectoderm, but seems to play no role in mesodermal patterning.

To investigate the evolution of genetic mechanisms controlling the dorsoventral patterning at the transition point between vertebrates and invertebrates, the amphioxus homologue (AmphiBMP2/4) of the vertebrate BMP4 was isolated (Panopoulou et al., 1998). AmphiBMP2/4 was isolated from the neurula stage library (BMP2/4). Amphioxus has like tunicates only a single homolog of the vertebrate BMP2 and BMP4 genes. Therefore, it is likely that a single ancestral gene has duplicated early in the vertebrate lineage to produce BMP2 and BMP4. AmphiBMP2/4 shares 55% identical aminoacids with each of the mouse BMP2 and BMP4 proteins when they are compared over their entire length. If only the C-terminal region of 116 aminoacids is compared the percentage of similarity increases to 81%. Moreover, the phylogenetic tree based on the full length proteins of Amphioxus BMP2/4, mouse BMP2 and BMP4, ascidian BMPHb and Drosophila dpp shows that AmphiBMP2/4 is closer to both mouse BMP2 and BMP4.

As it was described in 1.2.2. at the start of the amphioxus gastrula stage, the vegetal hemisphere of the blastula invaginates as the hypoblast into the animal hemisphere, which becomes the epiblast (Fig. 1.2). At the gastrula-neurula transition, the epiblast becomes subdivided into neural ectoderm dorsally and non- neural ectoderm (epidermis) laterally and ventrally. At the same time the hypoblast becomes regionalised into chordamesoderm
middorsally, paraxial mesoderm dorsolaterally, and endoderm laterally and ventrally. The paraxial mesoderm invaginates to form myocoelic somites, which constitute dorsal mesoderm (Fig. 1.3 top) while ventral mesoderm appears later.

All mesodermal regions are not formed simultaneously in most animals. Amphioxus embryos produce the dorsal mesoderm first and the ventral mesoderm later while in *Drosophila* the opposite occurs. In contrast, vertebrate embryos produce all their mesoderm simultaneously. This difference is reflected in the genetic mechanisms establishing dorsoventral polarity within the mesoderm. *In situ* hybridisation data shows that *AmphiBMP2/4* is initially detected at the midgastrula stage throughout mainly the hypoblast and to lesser extent the epiblast. By late gastrula, expression continues throughout the epiblast and hypoblast, but is more intense in dorsolateral regions of hypoblast where the presomitic grooves will form. In early neurula when the epiblast becomes subdivided into epidermis and neural ectoderm, the *AmphiBMP2/4* expression is downregulated dorsally in the neural plate.

In conclusion *AmphiBMP2/4* acts in a similar manner like the *Drosophila Dpp*, vertebrate *BMP4* and ascidian *HrBMPb* in patterning the ectoderm into the non-neural epidermis and the precursor of the central nervous system. Moreover, *AmphiBMP2/4* may be involved in patterning the hypoblast as judged by its expression in the dorsolateral region of the hypoblast in late gastrula. However, this can not be taken as a possible role in the establishment of a dorsoventral axis within the mesoderm as occurs in vertebrates. (Wilson and Hemmati-Brivanlou, 1995). Therefore, *BMP* plays a similar role in the patterning of ectoderm across invertebrates and vertebrates while the simultaneous establishment of dorsoventral polarity in the mesoderm seems to be a vertebrate added feature.

As a conclusion the study of amphioxus genes will resolve many questions as to the origin and evolution of vertebrate structures. Amphioxus however, cannot be used in isolation to infer the characteristics of the vertebrate ancestors. Common features shared between amphioxus and vertebrates show only that they were present in the last common ancestor of vertebrates and cephalochordates while differences indicate that they were derived, as the two chordate groups have evolved separately for more than 500 million years (Chen *et al.*, 1995; Nikoh *et al.*, 1997).
1.3. Oligonucleotide fingerprinting

1.3.1. Introduction

The major aim of the human genome project is to identify all genes and elucidate the regulatory mechanisms that control gene expression in a restricted spatial and temporal manner. While genomic sequencing provides information on the physical structure of the genome and the genes that are present, it does not provide any clues on the gene function. The last can be answered by studying which genes are preferentially expressed in any given cell at a specific time.

Single pass partial sequencing of randomly selected cDNA clones termed as expressed sequence tags (ESTs) has been proposed (Adams et al., 1992; Adams et al., 1991; Adams et al., 1993; Sikela and Auffray, 1993) as a method for directly assessing the diversity of genes expressed in a specific tissue. ESTs have attracted attention because of their relatively low cost in comparison to genomic sequencing, their high throughput nature along with their numerous applications in gene identification, comparative sequence analysis, gene mapping (Schuler et al., 1996), annotation of genomic sequence (Wilson et al., 1994) and candidate disease gene identification.

As a result of a collaboration between the Integrated Molecular Analysis of Genomes and their Expression (IMAGE) consortium (Lennon et al., 1996), Merck & Co., Inc., the Genome Sequencing Center (GSC) at Washington University and the National Center for Biotechnology Information (NCBI), more than 1,157,324 5' and 3' human ESTs derived from over 50 tissue specific cDNA libraries have been generated. Similarly, 350,000 ESTs have been generated from mouse tissue specific libraries (Marra et al., 1998), 85,000 from rat (http://www.ncbi.nlm.nih.gov/UniGene/Rn.stats.shtml) and 80,000 from Drosophila (http://www.fruitfly.org/EST/) (Casey et al., 1998).

However, EST sequencing is typically highly redundant. In any one tissue some genes are expressed at very high levels and others very low which is reflected in the resulting cDNA libraries e.g. in a typical somatic cell, the mRNAs of the highly and intermediate abundance class account for 50-65% of the total mRNA population (Bishop et al., 1974; Davidson and Britten, 1979). As a result, tag sequencing of random clones leads to high redundancy while
the probability of finding genes expressed at low level (which comprise about 50% of all genes) remains low.

Techniques for reducing the redundancy, such as normalisation and subtraction (Bonaldo et al., 1996) which were widely used in the IMAGE project (approximately 87% of all human ESTs were derived from these libraries), offer up to 3-fold normalisation. However, it was reported that short cDNAs which cover mainly the 3' untranslated region rather than coding, are favored over their longest counterparts which makes impossible the identification of those genes through BLAST type similarity searches. Moreover, due to the great extent of overlap in gene expression among different tissues, the use of normalised libraries is not sufficient for faster identification of novel genes. Finally, the major disadvantage of the above methods is that by eliminating redundancy, the information on the expression level of each gene is lost.

An alternative method for generating sequence information for a large number of clones in parallel which enables the preselection of a non-redundant set is the hybridisation based method of oligonucleotide fingerprinting. This method is based on the same principle of the hybridisation of short oligonucleotides on a large number of clones as the sequence by hybridisation method (SBH).

1.3.2. Sequencing by hybridisation (SBH)

Sequencing by hybridisation (SBH) is based on the idea that any sequence can be reconstructed by a series of n-mers that constitute that sequence and overlap by n-1 nucleotides. Two formats of this method were suggested where: i) the DNA to be sequenced is bound to a surface and hybridised sequentially with oligonucleotides (format 1) ii) the oligonucleotides are synthesized on a solid matrix and hybridised with sheared labeled DNA (format 2).

The first format can be used when the number of clones is larger than the number of probes while the second format is appropriate for sequencing smaller fragments of DNA.

Much effort has been invested on solving experimental problems related to either of the two approaches such as discriminative hybridisation conditions of as short as possible oligonucleotides on DNA bound on membranes (Drmanac et al., 1990) or the equality of hybrid stability irrespective of variations in sequence (Hoheisel, 1996) when oligos are attached on solid surface or the chemistry of attachment of oligonucleotides on different types of solid support.
SBH was initially applied in the reconstruction of the sequence of a 100 bp genomic fragment of human interferon gene (Strezoska et al., 1991) using 72 8-mer and 21 9-mer probes, designed to be complementary to the fragment and occurring in overlapping by one nt frames. However, it was realised that a major limitation of both formats of SBH, is that potential presence of repeats within a sequence limits the length of a fragment that can be unambiguously reconstructed to theoretically 200 bp while using 8-mer probes (Bains, 1991). Simulation experiments showed that using 1 million ordered 10mers, the size of the sequence subfragments has to be kept to less than 500, in order more than 95% of the fragments to be fully sequenced (Drmanac et al., 1992). However, this limitation disappears when part of the sequence of the analysed fragment is known. This limitation but also the efficiency of the sequencing gel-based methods has swifted the target of SBH from complete sequencing to its use as a diagnostic tool in the form of oligonucleotide chips carrying all identified point mutations in a specific gene (Maskos and Southern, 1993; Southern et al., 1992) or multiple DNA arrays of samples of affected individuals (Drmanac et al., 1998) for detecting mutations.

Except from de novo sequencing, the concept of SBH was also proposed for generating partial sequence information that could be used for:

a. creating physical maps (Poustka et al., 1986). Traditional techniques for identifying overlapping clones are based either on gel fingerprinting methods, feasible in the case of small genomic fragments (Coulson et al., 1989) where clones are ordered in the same contig if they share a percentage of common restriction patterns or hybridisation with long probes or pools of probes. Initial simulation experiments showed that using less than 100 oligomer probes with hybridisation frequency of 30% can assemble 350 cosmids distributed over an 1000 kb fragment (Hoheisel et al., 1993).

Finally, selection of shotgun genomic clones with dissimilar hybridisation oligonucleotide patterns was used as an alternative method to random shotgun sequencing, able to reduce the redundancy of the sequencing information by approximately 50% (Radelof et al., 1998). This approach is applied at the moment in a large scale sequencing project covering 1.5-2 Mb region of the 17p11.2 region of the human genome.

b. for determining sequence characteristics (Drmanac et al., 1991; Lehrach et al., 1990) that would help to predict coding potential or even function, without having to know the complete sequence (Partial SBH or oligonucleotide fingerprinting) e.g i) the GC content would indicate presence of coding sequence (vertebrate exons have 10-20% higher GC content than introns) ii) identification of characteristic patterns or motifs such as DNA binding domains iii) presence of repetitive elements e.g. ALU or LINE and in the case of an
array of cDNA clones for identifying i) novel genes ii) differences in patterns of expression between tissues or developmental stages.

The aspect of using SBH for assessing the number and relative transcript level in a given tissue or developmental stage as well as a normalisation method for reducing sequencing redundancy and thus offering faster access to novel genes, was exploited in this study.

1.3.3. Oligonucleotide Fingerprinting. Theoretical background.

Oligonucleotide fingerprinting (or partial SBH) is based on the hybridisation of short oligonucleotides on a high density array of (cDNA or genomic) clones of unknown sequence, generating a sequence based oligo-hybridisation pattern, unique for each clone. Clones with similar hybridisation pattern are clustered together as representing the same or highly overlapping sequence (Fig. 1.4.).

**Oligonucleotide hybridisation-clustering**

![Oligonucleotide hybridisation-clustering](image)

*Fig. 1.4.* Clones sharing the same oligonucleotide hybridisation pattern are clustered together. The matching oligonucleotides are spread over the entire length of the clone thus achieving the clustering of the clones with varying length.
Briefly during oligonucleotide fingerprinting, the clones to be fingerprinted, either cDNA or shotgun genomic, are PCR amplified. All PCR products are transferred (spotted) on nylon membranes according to a specified pattern, using a robotic equipment (Maier et al., 1997a). Each membrane is currently spotted with 27,648 clones, each clone in duplicate. Each membrane is then hybridised sequentially with on average 150 short oligonucleotides which have been radioactively labeled. Each filter is exposed on phosphor storage screens and the resulting hybridisation images are analysed using inhouse developed software. Clones are grouped together according to their oligonucleotide hybridisation patterns. An outline of the oligonucleotide fingerprinting procedure is given in Fig. 1.5.

**Oligonucleotide Fingerprinting**

RNA isolation from whole embryos or tissues

Plasmid cDNA library construction

Picking and arraying in 384-well MTP

PCR of cDNA inserts

Spotting of PCR products on nylon membranes

Hybridisation with radioactive oligos

Capture of hybridization data by phosphorimager

Colony Spotting on nylon membranes

Library quality control

Filter distribution as reference library

Get full length gene

RACE

in situ hybridisation of identified genes

Sequence search against nr, swissprot, dbEST...

Sequencing of one cDNA per cluster and singletons

Clustering analysis

Generation of oligohyb pattern for each cDNA

**Fig. 1.5.** Outline of the oligonucleotide fingerprinting procedure

The major advantages of oligonucleotide fingerprinting over the preselection methods (biochemical normalisation and subtraction) mentioned above are that:

- conventional oligodT cDNA libraries without any additional manipulations can be used,
• up to more than 3 fold reduction in redundancy has been achieved (Poustka et al., 1999; Radelof et al., 1998),

• the information on the relative expression levels of each transcript is retained. The size of each group of similar fingerprint clones (termed as cluster) is an indication of how abundant is the gene encoded by each of the clones that belong to the cluster.

The efficiency of oligonucleotide fingerprinting as a method for distinguishing between different sequence patterns depends largely on the sequence and the number of oligonucleotides used as well as the length of the clones to be fingerprinted.

1.3.3.a. Factors influencing the selection of oligonucleotides

The information generated in each single experiment with an oligonucleotide increases with the number of clones found positive. An ideal oligonucleotide would be one that hybridises to 50% of the clones. Meanwhile, the hybridisation frequency of an oligonucleotide increases exponentially with the decrease in its length.

The probability of an oligonucleotide of a length \( L_{\text{ol}} \) occurring \( R \) times in a random sequence of length \( L_{\text{CDNA}} \) is given by the following equation:

\[
P(R, L_{\text{CDNA}}) = \binom{L_{\text{CDNA}}}{R} \times (\frac{L_{\text{CDNA}} - R}{R})^{L_{\text{DNA}}} \times (1/4)^L_{\text{ol}} \]

According to the factor \((1/4)^L_{\text{ol}}\) the chance of an oligo of the length \( L_{\text{ol}} \) occurring \( R \) times decreases with the increase in its length (Drmanac et al., 1989). Thus the probability of an 8mer oligo occurring at least once in a 1.5 kb fragment of random sequence is 0.023.

Moreover, according to the above equation the probability of an oligo occurring in a DNA fragment increases with the increase in the length of the fragment to be fingerprinted.

However, since the DNA sequence is not random, the hybridisation frequency of an oligo will greatly vary according to its sequence. Theoretical experiments have shown that the number of oligonucleotides present in a randomly generated sequence having the same tetranucleotide frequency as a real genomic, is less than predicted (Bains, 1994). A reason for this bias in the occurrence frequency is the presence of repetitive elements. As a conclusion, not all oligonucleotides have the same hybridisation rate and thus differ in the hybridisation information that they generate.

Another factor that has to be considered when selecting the optimal oligonucleotide set is the oligonucleotide sequence composition which in combination with the oligonucleotide length affects the hybrid stability and therefore the hybridisation conditions. This problem is escalated when multiple oligonucleotides are assessed if they are present in a DNA fragment (format 2) or multiple oligonucleotides are hybridised on an array of clones (format 1, there is not published data on this). The factors considered during the hybridisation of a single oligonucleotide on an array of clones which was the format that was used during this study, are described below.
1.3.3.b. Oligonucleotide hybridisation conditions

A basic requirement of oligonucleotide fingerprinting, irrespective of the format, is the ability to reliably hybridise short oligonucleotides and distinguish between perfect matches and mismatches, especially end base mismatches which are more stable than internal. (Drmanac et al., 1990) investigated the experimental conditions under which reliable (with limited number of mismatches) hybridisation of 6mer oligonucleotides occur. They suggested that the discrimination at lower temperature depends on the ratio of the amount of the perfect match to the background and thus it is important as a first step to obtain a sufficient amount of perfect hybrid, even if simultaneously a significant amount of imperfect hybrids will form. Equilibrium during the hybridisation can be directed towards the hybrid formation by increasing the probe concentration or the amount of target DNA or decreasing the temperature. The higher amount of perfectly matched hybrid allows subsequent longer washing time where differential melting of perfect and imperfect hybrids results in difference in respective signal intensities. 8-mers and 9-mers have been found to give the better combination of duplex stability and discrimination for generating the full sequence of a clone. It has been shown that even in cases where the concentration of the non-perfectly complementary target DNA is ten times higher than that of the target containing the perfect match, the signal intensity ratio is inverted towards the favor of the perfect match by increasing the washing time at lower temperature.

After the introduction of SBH, continuous improvements have been made on the experimental side e.g PCR amplification of the target clones, increased the amount of DNA available for hybridisation but also allowed the use of oligos that would be otherwise excluded as complementary to the vector sequence (Drmanac et al., 1992; Drmanac and Drmanac, 1994; Meier-Ewert et al., 1993). Moreover, automation of many steps of the procedure has increased the number of samples that can be screened. Thus: a) the development of robotic systems that allow arraying of up to 3,000 clones/hr in microtitre plates allowed fast handling of thousands of samples (Jones et al., 1992; Maier et al., 1994b; Uber et al., 1991) b) the development of a thermocycling robot where 51,840 reactions can be performed at a time, allowing at least two runs to be performed within one day (Maier et al., 1997a; Maier et al., 1994b; Maier et al., 1997b) and c) the automated clone spotting at high density on nylon membranes where 57,600 samples (even higher densities of up to 147,456 have been achieved) can be spotted on a 22 cm x 22 cm membranes, allowed fast simultaneous handling and thus hybridisation of many membranes and thus samples per day. A detailed description of the above systems, many of which developed in the lab that this project was carried out and which were used during the course of the project, is given in the materials and methods section (chapter 2).
The first large scale project where fingerprinting was applied was in assessing the number of different transcripts present in a human fetal brain cDNA library (Drmanac et al., 1996; Milosavljevic et al., 1996). Moreover, (Meier-Ewert et al., 1998) were able to co-cluster a 9d and 12d embryo cDNA libraries and estimate the number of transcripts differentially expressed in these two stages. The possibility to recognise the identity of transcripts by comparing the experimental oligo-hybridisation patterns with the calculated theoretical ones of all the human and rodent sequences deposited in the Genbank was verified in 91% of the cases that the similarity score of the experimental fingerprint against the theoretical was $P < 10^{-7}$ (Meier-Ewert et al., 1994).
Aim of this thesis

Cross-species comparisons have been proved a major tool leading to faster identification of novel genes and the elucidation of their function. The first is achieved through the recognition of conserved domains while the second through the genetic tools available in the model organisms. This approach has been successful in the identification of functionally important domains in human disease-related genes. Moreover, in the narrowing of candidate disease regions through sequence matches of mutated gene phenotypes. The requirement though for genetic manipulations has biased the choice of model organisms. An initial overview of the recently sequenced *C. elegans* genome showed that almost 35% of its predicted proteins are nematode specific while another 35% have no known homologue in other organisms. Moreover, phylogenetic analysis of 36 orthologous proteins found in *Drosophila*, humans, nematodes and yeast and their pairwise comparison of evolutionary distances between human-nematode and human-arthropod showed that nematode genes evolve faster than the *Drosophila* ones.

As more genes are identified in model organisms through either genomic or EST sequencing, the history of the evolution of their own lineage becomes apparent by the presence of multiple paralogues or the selective absence of some.

In the case of vertebrates, functional redundancy due to the presence of multiple paralogs as a result of two cycles of presumed genome duplication in the vertebrate lineage makes difficult any interpretation or correlation of data.

In chapter 1 amphioxus was introduced as a phylogenetically important organism whose study will shed light in the pattern of genome evolution at the transition point from invertebrates to vertebrates where the first genome duplication is thought to have occurred. Nevertheless, my interest in amphioxus is not only evolutionary. The study of a genome stripped from paralogues will give fast access to an archetype chordate gene set and maybe to novel chordate genes. Study of those genes will reveal the extent of conservation within individual genes. Considering that the cephalochordate lineage separated from the vertebrate 500 million years ago, such conservation will indicate functionally important gene segments. Temporal and spatial expression studies of those genes, through *in situ* hybridisations, will point to the equivalent to vertebrate structures. Moreover, comparison of the amphioxus expression patterns with those of the vertebrate equivalent genes or gene families will show which of the roles of the genes have been adopted more recently than others. This will help to understand how the vertebrate characteristic morphological structures have appeared.

Amphioxus' embryology has been described in detail at the beginning of the century. In the last five years the interest for amphioxus has been revived. However, very selective amphioxus genes have been studied so far. These are genes that their homologues have
already been identified in other organisms. The purpose of those studies was to follow the evolutionary history of those genes or gene families and furthermore to solve long standing debates such as the origin of vertebrate brain, the evolution of the vertebrate eyes, the origin of neural crest cells etc.

The aim of this thesis is to access all amphioxus genes. Two embryonic stages, the gastrula where the morphogenetic movements leading to the formation of the three germ layers and the neurula where some of the major body structures have been formed are studied. This selection of embryonic stages rather than adult tissues will allow the identification of major patterning molecules transiently present during only the early stages of development.

Oligonucleotide fingerprinting was proposed as a method for creating sequence information (fingerprint) for a large number of clones, in parallel, enough to classify them into clusters. The experimental conditions required for the hybridisation of short oligonucleotides had been previously established (Drmanac et al., 1990). Moreover, the automation tools necessary for the screening of large number of samples as well as initial versions of data analysis software were first developed during a pilot project (Meier-Ewert, 1994). During this pilot study, oligonucleotide fingerprinting was applied on 32,000 cDNA clones from a human fetal brain cDNA library.

A further aim of this thesis is to assess the normalisation efficiency of oligonucleotide fingerprinting on a larger number of clones. Two amphioxus embryonic cDNA libraries (110,000 clones) were arrayed and screened by oligonucleotide fingerprinting. Large-scale tag sequencing of clones selected from fingerprinting clusters and further sequence clustering will indicate whether there is still redundancy in the selection of clones. The effect of parameters like the cDNA insert length or the sequence of oligonucleotides used, on the clustering will be assessed. Moreover, the extent of normalisation achieved, the ability of fingerprinting in predicting levels of expression of transcripts according to the size of the cluster they belong in and the possibility to compare clone fingerprints across multiple libraries derived from the same organism will be evaluated.
Chapter 2

Materials and methods

2.1. Reagent suppliers

Chemicals such as: Trizma-Base, Trizma hydrochloride, Trizol, Isopropyl alcohol, ethanol, chloroform, DEPC, bacto-tryptone, bacto-yeast extract, glycerol, IPTG, Hepes, mercaptoethanol, X-Gal, RNAse, PMSF, DDT, ethidium bromide, EDTA, bromophenol blue, were mainly supplied by Sigma, Merc and Pharmacia

Phenol was purchased from Roth GmbH

Agarose was purchased from Gibco BRL

Pronase was supplied by Boehringer Mannheim GmbH

Radioisotopes were supplied from Amersham

Molecular weight standards: 1 kb DNA ladder was supplied by Stratagene

Enzymes: restriction enzymes as well as T4 DNA ligase and T4 polynucleotide kinase were purchased from New England Biolabs while E.coli DNA polymerase large fragment (Klenow) was supplied by Amersham

Media: bacto-tryptone, bacto-yeast extract from Difco

Hexamers(for random priming) and dNTPs were supplied from Pharmacia

Oligonucleotides were either synthesized either by the Max-Planck sequencing group or they were purchased from TiB MolBiol (Berlin, Germany)
384-well polypropylene and polystyrene microtitre plates, 384-pin polypropylene replicating gadgets, Q-plate bilaminar sealing film were supplied by Genetix, while 96-well thermofast plates were purchased from Advanced Biotechnologies, LTD

Nylon membranes (Hybond N+) were supplied by Amersham

**RNA isolation-cDNA cloning**

Trizol used for RNA isolation and Superscript cDNA synthesis kit were purchased from Gibco BRL

mRNA purification kit (Dynabeads) from Dynal

RNase from Sigma

**Sequencing**

Dye terminator kit from Perkin Elmer, Applied Biosystems.

Thermo sequenase fluorescent labelled primer cycle sequencing kit with 7-deaza-dGTP from Amersham.

Silica columns and buffers for PCR product purification from Qiagen

**Bacterial strains**

XL1 blue: F[ΔTn10 proAB lacI9 (lacZ)ΔM15] recA1 endA1 gyrA46 thi relA1 supE44

hsdR17 lac

2.2. Solutions

2.2.1. General Buffers and solutions

1xTE: 10 mM Tris-HCl (pH8.0)

1 mM EDTA

Pronase buffer (for colony filter processing): 22.4mM Tris-HCl

27.5mM Tris-base

3% v/v Sarcosyl (30%)
Pronase stock: 50 mg/ml in sterile H2O, stored at -20°C, final concentration before use in the pronase buffer: 25mg/ml

1M Tris-buffer pH 7.4:  
18.8% v/v 1 M Tris-base (M.W. 121.1)  
81.2% v/v 1 M Tris-HCl (M.W. 157.6)

1M Tris-buffer pH 8.0:  
44.8% v/v 1 M Tris-base (M.W. 121.1)  
55.2% v/v 1 M Tris-HCl (M.W. 157.6)

Electrophoresis buffers (working concentration):

1xTAE: 40mM Tris acetate  
1 mM EDTA

1xTBE: 45mM Tris-borate  
1 mM EDTA

20xSSC: 3.0M NaCl  
0.3M sodium citrate  
Sterilise by autoclaving

X-gal: 2% in dimethylformamide. Store at -20°C, sensitive in light.

IPTG (M.W. 238.3): 0.8M dissolved in distilled H2O. Sterilise through 0.22 μm filter. Store at 20°C in 1 ml aliquots

Gel loading buffer (pH 7.5): 2.50 mM Trizma base,  
0.05 mM EDTA,  
50% v/v glycerol,  
0.008% w/v bromophenol blue.

RNase (10mg/ml): dissolve in TE, preboil for 10 min and store frozen at -20°C.

Minipreparation of plasmid DNA (alkaline lysis):

AL1: 50 mM glucose  
25 mM Tris.Cl (pH 8.0)  
10 mM EDTA
AL2: 0.2 M NaOH
1 % SDS

AL3 60.0 ml 5 M potassium acetate
11.5 ml glacial acetic acid
28.5 ml H₂O

Denaturing solution: 1.5 M NaCl, 0.5 M NaOH

Neutralising solution: 1 M Tris-HCl pH 7.5, 1.5 M NaCl

5% Linear polyacrylamide:
5 gr acrylamide (without bis-acrylamide) were dissolved in 100ml solution of TAE pH 7.8. 1/10 of the volume of 10% ammonium persulfate and 1/1000 of the volume of TEMED was added and it was let to polymerize for 30min. The polymer was precipitated by centrifuging after adding 2.5 volume of ethanol and the pellet was re-dissolved in 100ml of dH₂O.

salmon sperm DNA: 6mg/ml in 1xTE (stock solution), final concentration for spotting 600 ng/µl in 100 mM NaOH

10XPCR buffer: 100.0 ml 2.5M KCl
50.0 ml 10% Tween 20
7.5 ml 1M MgCl₂
175.0 ml 1M Tris-base
75.0 ml 1M Tris-HCl
92.5 ml H₂O
Final volume: 500 ml

End concentrations: 0.5 M KCl, 1% Tween 20, 0.015 M MgCl₂, 0.35 M Tris-base, 0.15 M Tris-HCl.

1 M sodium Phosphate(1 M Na+, 0.5 M Na₂HPO₄): 71 gr Na₂HPO₄/lt of H₂O, add approximately 4ml orthophosphoric acid (85%), pH 7.2 Sterilise by autoclaving

10xHMFM: 0.36% v/v 1 M MgSO₄
0.45% w/v Na$_3$ Citrate
0.9% w/v (NH$_4$)$_2$SO$_4$
44.0% w/v Glycerol
1.8% w/v KH$_2$PO$_4$
4.7% w/v K$_2$HPO$_4$

Phosphate salts are autoclaved separately.

**Antibiotics**

Ampicillin: 50 mg/ml stock solution dissolved either in H$_2$O (sterilise by filtration through 0.22 μm filter) or ethanol (no need to sterilise). Store at -20°C. Use at 100 μg/ml final concentration

Kanamycin: 10 mg/ml stock solution in H$_2$O. Store at -20°C. Use at 30 μg/ml final concentration.

Tetracycline: 5 mg/ml stock solution in ethanol. Store at -20°C. Use at 13 μg/ml final concentration.

DDT (M.W.154.2): 1 M stock solution dissolved in 0.01 M CH$_3$COONa pH 5.2. Sterilise by filtration through 0.2 μm filter. Store at -20°C.

**2.2.2. Hybridisation solutions**

100x Denhardts: 2% ficoll
2% polyvinylpyrrolidone
2% BSA
can be stored at -20°C

Hyb mix (stock): 80 gr Dextran sulphate
200 ml 20xSSC
50 ml 1 M Na$_2$HPO$_4$ pH7.2
2 ml 0.5 M EDTA
100 ml 100xDenhardts
5 ml salmon sperm DNA(5 mg/ml)

store at -20°C in 40 ml aliquots

For 100 ml of hybridisation solution: 40 ml of hyb mix
30-50 ml of deionised formamide
10 ml of 10% SDS
Deionised formamide: 1 hr mixing with Serdolit. Filter twice through Whatmann paper 13MM

Church buffer: 500 ml 1 M NaPi
500 ml 10% (w/v) SDS
2 ml 0.5 M EDTA

SSarc: 20% (v/v) 20xSSC
25% (v/v) N-lauryl-sarcosine (30%)
55% H$_2$O

2.2.4. Culture media

LB: 1.0% w/v bacto-tryptone
0.5% w/v bacto-yeast extract
1.0% w/v NaCl

2xYT: 1.6% w/v bacto tryptone
1.0% w/v bacto-yeast extract
0.5% w/v NaCl
Adjust pH at 7.0 with NaOH, if needed.

2xYT agar/lt: As above but before autoclaving add 9gr agar.

SOB: 2.00% w/v bacto-tryptone
0.50% w/v bacto-yeast extract
0.05% w/v NaCl
2.5 mM KCl
adjust pH at 7.0 with 5 M NaOH and autoclave
Just before use add MgCl$_2$ (from a separately autoclaved 2M MgCl$_2$) to 10 mM final concentration.

SOC: 2% w/v bacto-tryptone
0.5% w/v bacto-yeast extract
10 mM NaCl (or 0.05% w/v)
2.5 mM KCl
10 mM MgCl₂

SOC medium is identical to SOB medium, except that it contains 20 mM glucose final concentration. The glucose solution (1 M) was sterilised by filtration through a 0.22 μm filter and added after the SOC medium was autoclaved and cooled down to 60°C.

Terrific Broth: 1.2% w/v bacto-tryptone
2.4% w/v bacto-yeast extract
0.4% v/v glycerol

All media were sterilised by autoclaving at 15 lb/sq. in. on liquid cycle for 20 min.

Glycerol stocks are prepared by mixing 0.5 ml of the culture with 2YT media supplied with HMFM freezing mix and the appropriate antibiotic. Add 44 ml HMFM per 400 ml 2YT and 100 μg/ml ampicillin)

2.2.5. Genomic DNA extraction

TEN 9 (for DNA extraction): 50 mM Tris buffer pH9
100 mM EDTA
200 mM NaCl

2.3. Experimental procedures

2.3.1 RNA isolation

RNA isolation was carried out according to (Chomczynski and Sacchi, 1987) method. In order to isolate RNA from whole adult amphioxus, two amphioxus (approx. 3-4cm length, 1cm width) frozen at -80°C were mixed with liquid nitrogen and ground in a pestle and
mortar pre-cooled on dry ice. The disrupted tissues were transferred to a glass Dounce homogenizer where they were mixed with 10 ml of Trizol (1 ml of Trizol per 50-100 mg of tissue). Trizol (Life Technologies, GibcoBRL) is a mono-phasic solution of phenol and guanidine isocynate based on the acid guanidinium-phenol-chloroform (AGPC) method by (Chomczynski and Sacchi, 1987). The homogenised sample was incubated at RT for 10 min and subsequently transferred in a Corex glass tube where it was mixed with 2 ml of chloroform (0.2 ml of chloroform per 1 ml of Trizol used). The tube was shaken vigorously, incubated for another 5 min at RT and centrifuged at 12,000 g for 15 min at 4°C. Following centrifugation, the mixture was separated in a lower red organic phase (phenol-chloroform) and an upper aqueous phase. The RNA was precipitated from the aqueous phase after adding 5 ml of isopropyl alcohol (0.5 ml of isopropanol per 1 ml of Trizol used) and incubation for 15 min at RT. The RNA was visible as a gel-like colorless pellet after centrifugation at 12,000 g for 15 min at 4°C. The RNA pellet was further washed with 75% ethanol (1 ml of 75% ethanol per 1 ml of Trizol used), centrifuged at 9,000 g for 10 min at 4°C and further dissolved in DEPC-treated water.

The RNA quality was checked by a fast electrophoresis on a standard 1% agarose gel. RNA degradation was kept to minimum by soaking the gel tank and the combs in 3% H₂O₂ for 15 min. The electrophoresis buffer was also prepared in DEPC treated water.

2.3.2 cDNA library construction

2.3.2.a. Poly A+ selection

Polyadenylated (Poly A+) RNA was isolated using the Dynabeads mRNA purification Kit (Dynal). 75 µg/100 µl of total RNA was suspended in an equal volume of 2x Binding buffer (20 mM Tris-HCl, pH 7.5, 1 M LiCl, 2 mM EDTA) and heated to 65°C for 2 min to disrupt any secondary structure. The RNA was subsequently mixed with 1 mg of magnetic beads which were pre-washed and resuspended in 100 µl of 2x Binding buffer. The binding capacity of the beads is 2 µg mRNA per mg (1-5% of total RNA is mRNA). The solution was left to stand to hybridise for 3-5 min at RT. The tube was then placed in the magnetic rack where it was left to stand for 30 sec. The supernatant was removed and the sample was washed twice with 200 µl of 1xWashing Buffer (10 mM Tris.HCl, pH 7.5, 150 mM LiCl, 1 mM EDTA).
2.3.2.b. First strand synthesis

Both amphioxus cDNA libraries were constructed by Matthew Clark using the Superscript Plasmid cloning system from GibcoBRL, Life Technologies. Both libraries are oligo(dT) primed and directionally cloned into pSport1 vector (Fig. 2.1.).

First strand synthesis was carried out using the SuperScript II reverse transcriptase (GibcoBRL, Life Technologies) and the 5'pGACTAGTTCTAGATCGCGAGCGGCCGCCC(T),3' primer which binds to the 3' poly(A) tail of the mRNA.

The SuperScript II RT is a modified Moloney Murine Leukemia Virus (M-MLV) reverse transcriptase devoted from RNase H activity. RNase H activity might result in the nicking of the hybrid between the mRNA and the oligo(dT) primed first strand close to the site of polymerization which will result in the premature termination of cDNA synthesis thus yielding limited number of full-length cDNAs.

1 μg of mRNA was mixed with 1 μg Not I primer. The primer is added in excess so that each mRNA molecule binds oligo(dT),. The mixture was heated to 70°C in order to remove secondary structure of the mRNA and then quickly chilled on ice.

The eppendorf was centrifuged briefly and the 5xfirst strand buffer (250 mM Tris-HCl pH 8.3, 375 mM KCl, 15 mM MgCl₂) at a final concentration of 50 mM Tris-HCl, 75 mM KCl, 1 mM MgCl₂, together with 10 mM DTT, 500 μM each dNTP and 1 μCi α-32PdCTP were added in order to calculate later the yield of the first strand synthesis. The tube was gently vortexed, incubated at 37°C for 2 min while Superscript II RT (200U/μg of mRNA) was added. The eppendorf was further incubated at 37°C for 1 hr. The reaction was terminated by placing the tube on ice. In order to estimate the size range of products synthesized after the first and second strand synthesis the sample can be run on alkaline agarose gel. The gel is then dried and exposed on a film.

2.3.2.c. Second strand synthesis

Second strand synthesis is based on the nick translational replacement (Gubler and Hoffman, 1983) method where the RNA is partially degraded with RNaseH resulting in a series of RNA primers which are used for replacement synthesis with DNA polymerase I. Finally, the
Fig. 2.1. Map of the plasmid vector pSPORT1. The sequence of the multiple cloning site is shown in detail.
cDNA was blunt ended with T4 DNA ligase.

The second strand synthesis was carried out in 25 mM Tris-HCl pH 7.5, 100 mM KCl, 5 mM MgCl₂, 10 mM (NH₄)₂SO₄, 0.15 mM β-NAD⁺, 250 μM of each dNTP, 1.2 mM DTT with 56 U/ml of DNA ligase, 250 U/ml DNA polymerase I and 13 U/ml RNaseH. The reaction was incubated at 16°C for 2 hrs. The cDNA was end polished with T4 DNA Polymerase 5 U/ml at 16°C for 5 min. The reaction was stopped with 10 μl of 0.5 M EDTA and the mixture was extracted with an equal volume of phenol:chloroform:isoamyl alcohol (25:24:1). The cDNA was precipitated with 2.5 M CH₃COONH₄ final concentration and 2 volumes of ethanol. The pellet was washed with 70% ethanol.

2.3.2.d. Adapter addition, NotI digestion and size fractionation

The product of the first and second strand synthesis is a blunt end cDNA. Since the efficiency of ligation between blunt ended DNA is very low, restriction digestion with Not I exposes the 5' of the cDNA while the other site remains blunt ended. In order to maximize the ligation efficiency between the vector and the cDNA, Sall adapters are ligated to the cDNA. The sequence of the Sall adapter is the following:

5' -TCGACCCACGCGTCCG-3'
3' GGGTGCGCAGGCp-5'

Adapters are double stranded oligomers that are blunt ended at one end while they contain a 4-base extension at the other end. This extension corresponds to the site created after digestion with Sall and XhoI. Self ligation of adapters is prevented by phosphorylation of one strand.

In order to remove residual adapters that are present and might interfere with the ligation of the cDNA to the vector but also to select only the longer cDNAs, the produced cDNA is size fractionated using Sepharose columns. Directional cloning of the cDNA into the plasmid vector is achieved by digestion with NotI (which has a recognition site in the oligodT primer).

Briefly, to the dried cDNA pellet from the previous step, 10 μl of 5xT4 ligase buffer (250 mM Tris-HCl pH 7.6, 50 mM MgCl₂, 5 mM ATP, 5 mM DTT, 25% (w/v) PEG 8000), 10 μl Sall adapters (200 μg/μl) and 5 U of T4 DNA ligase were added in a final volume of 50 μl. The mixture was incubated at 16°C for 12-16 hrs and subsequently extracted with an equal
volume of phenol:chloroform:isoamyl alcohol (25:24:1). The cDNA was precipitated with 2.5M CH$_3$COONH$_4$ final concentration and 2 volumes of ethanol.

Digestion of the cDNA with NotI was carried out in a 50 µl reaction containing 10 U NotI, 50 mM Tris-HCl pH 8.0, 10 mM MgCl$_2$, 100 mM NaCl. The mixture was incubated at 37 °C for 2 hrs and subsequently extracted with an equal volume of phenol:chloroform:isoamyl alcohol (25:24:1). The cDNA was precipitated with 2.5M CH$_3$COONH$_4$ final concentration and 2 volumes of ethanol. The cDNA pellet was dried and dissolved in 100 µl of TEN buffer (10 mM Tris-HCl pH 7.5, 0.1 mM EDTA, 25 mM NaCl).

The cDNA was subsequently loaded on a column which had been previously equilibrated with 100 µl TEN buffer. One drop fractions of the effluent were collected separately. The first 12 fractions contain the longest fragments. The amount of cDNA in these fractions was calculated by counting the amount of radioactivity incorporated with a scintillation counter.

2.3.2.e. Ligation of the cDNA to the vector

10 ng of cDNA from the first fractions were mixed with 50 ng of Sall/NotI cut pSport1 vector, 4 µl of 5xT4 DNA ligase buffer (250mM Tris-HCl pH 7.6, 50 mM MgCl$_2$, 5 mM ATP, 5 mM DTT, 25% (w/v) PEG 8000), and 1U T4 DNA ligase in a final volume of 20 µl. The reaction was incubated at 4°C overnight. The ligated cDNA was precipitated with 0.5 volumes of CH$_3$COONH$_4$ and 2 volumes of ethanol. The pellet was washed with 70% ethanol in order to remove the salts since they will interfere with the electroporation.

2.3.3 Introduction of Plasmid DNA into bacteria

a. Preparation of electrocompetent bacteria (XL1-blue)

1lt of SOB media was inoculated with 1 ml of an overnight culture of bacteria and incubated with vigorous shaking at 37°C until A600 of 0.8. The culture was cooled on ice for 15 min and centrifuged at 4°C for 15 min at 2,000 rpm. The pellet was resuspended in 1lt of ice cold 10% glycerol and the cells pelleted again. The above step was repeated for a second time. The cells were further washed with 40 ml of 10% glycerol and pelleted after centrifuging at 8,000 rpm for 10min. The pellet was finally dissolved in 1 ml of 10%
glycerol in total. The cells were then frozen in 100 μl aliquots in 0.5 ml tubes using a dry ice-ethanol bath and stored at -70°C.

**b. Electroporation**

The electrocompetent cells were thawed on ice, mixed with 10 ng DNA to be transformed and transferred in a 0.2 cm pre-chilled cuvette. The electroporation was performed using a Gene Pulser (Biorad) at 25 μF, 200 Ω, 2.5 kV. After the pulse the cuvette was quickly removed from the electroporation chamber and the cells were transferred in 1 ml of pre-warmed SOC medium. After incubation at 37°C for 1 hr the cells were plated on selective agar plate.

**2.3.4. Arraying of libraries**

Library picking and arraying in microtitre plates, DNA spotting on nylon membranes and library clone re-arraying into a non-redundant copy were carried out using the same robotic system (Fig. 2.2 and 2.3).

The basic feature of this robot is a robotic arm which can move along all three axis, x,y,z with a speed of up to 2 m/sec and accuracy better than 5 μm (Maier et al., 1997a) above a flat surface. A different 96 or 384-metal pin gadjet is attached at the end of this arm according to the intended use.

**a. Automated systems for library arraying and spotting**

**Picking robot**

Agar trays (22.5 cm x 22.5 cm) (two trays plated with 3,000-6,000 individual colonies can be processed at a time) with grown bacterial colonies placed on specified windows (where light is illuminated from underneath) on the robot bed (Fig.2.2). A Charge Coupled Device (CCD) camera attached on the robot arm identifies the individual colonies on the agar dishes and an image analysis program calculates the xy coordinates of every colony. Up to 48 images-frames of each agar plate can be captured in approximately 3 min. Moreover, the software gives the possibility of adjusting several parameters such as colony size, roundness, colour etc. After the xy coordinates of each colony have been calculated, the robotic arm with the 96-metal pin picking gadjet is positioned over the relevant colony and an individual
Fig. 2.2. A. Bacterial colonies are picked automatically using a 96-pin gadget by a robotic system guided by the image of the plated bacteria captured by a CCD camera.

B. PCR amplification of 50,000 cDNA clones can be performed simultaneously in a thermocycler robot. The 384-plates are cycled between three waterbaths set at the appropriate temperature for denaturation, annealing and extension.
pin is extended to pick this colony. Once all the 96 pins of the gadjet have been used, they are dipped into a pre-filled 384-well microtitre plate. The gadjet is then sterilised in an ethanol bath followed by an air drying cycle. The above described system can pick 3,000 clones per hr into 384-well microtitre plates.

**Spotting robot**

For the transfer of bacterial colonies or PCR products on nylon membranes (spotting), the 96-pin gadjet is replaced with a 384-single spring pin gadjet (250 µm tip diameter). Moreover, the membranes are placed on tiles clamped on the robot bed. The microtitre plates to be spotted are fitted in a stacking system that can hold up to 56 plates. A 'grabber' attached to the robot arm removes each plate sequentially from the microtitre plate rack and places it onto a plate holder, where the lid is automatically opened and the plate barcode is read. Every 384-well plate has been sprayed with a unique barcode consisting of the library number, the library copy number and the plate number (Fig. 2.3). Finally, each plate, after being spotted is placed automatically back into the rack system. During the run a log file containing the plate numbers in the order they were spotted, as identified by the barcode reader, is created which confirms the spotting order of the plates.

All the clones were spotted in duplicate format which helps in distinguishing the true from false positive clones when scoring the filter after a hybridisation. The library clones can be spotted directly from 384-well microtitre culture plates or as PCR products (in the case of cDNA clones intended to be oligonucleotide fingerprinted).

**Spotting patterns**

The spotting density on the membranes, is determined by the nature of the clones to be spotted e.g bacterial colony filters were spotted in 4x4 duplicate pattern (blocks of 16 spots containing 8 different clones spotted in duplicate) which prevents the colonies from colliding with each other, while PCR products in 5x5 duplicate pattern (blocks of 25 spots containing 10 different clones spotted in duplicate). However, 5x5 patterns without alleyways have been used successfully for spotting colony filters. In the 4x4 pattern, 48x48 blocks, 36,864 spots in total or 18,432 different clones, are spotted. In the 5x5 spotting pattern 48x48 blocks (57,600 spots in total or 27,648 different clones) at a distance of 900µm between them are spotted on each membrane. The spot in the center of each block at the 5x5 pattern is an ink or higher complexity DNA (see 2.2.4b PCR filter) or blank dot so that it can be used as guide dot when orientating the filter while scoring it (Fig.3.4). Higher spotting densities of up to 7x7 (112,896 spots per filter) have been achieved using the same robot.
Fig. 2.3. A. Nylon membranes are overlayed on the tiles on the robot surface. Transfer of bacteria or clean DNA in the form of PCR products on the filters is carried out by a 384-pin gadjet.
B. A barcode reader records the plate number and library before it is spotted.
Re-arraying robot
Clones selected as being unique, after oligonucleotide fingerprinting the whole library, were picked from the original 384-well plates and they were re-arrayed into new 384-well microtitre plates.

The re-arraying procedure is similar to the picking procedure as described in 2.3.4.a. A 96-pin gadget picks automatically the identified clones (the clone coordinates as in the original library are specified in a file) from the original plates of the library into new 384-well plates.

a. Library arraying in microtitre plates

_E.coli_ bacteria transformed with cDNA were spread at a density of 3,000-6,000 clones on 22 cm x 22 cm bioassay trays (NUNC) filled with 200 ml of LB agar media supplied with 100 µg/ml ampicillin. The plates were incubated at 37°C for 14-16 hrs until the colonies are large enough to be picked by the picking robot, yet they do not overlap. Individual colonies were automatically picked in 384-well microtitre plates pre-filled with 2YT media and 100µg/ml ampicillin supplied with 1xHMFM freezing mix, as described above.

In order to avoid contaminating the originally picked clones of a library, four additional working copies of the whole library were made. The master or duplicating copy(Q1) was used every time when a new copy of the library was made, the picking copy for picking individual clones of interest, the spotting copy for making bacterial filters of the whole library while a copy was dedicated for preparing PCR products of the whole library for the oligonucleotide fingerprinting method.

Finally, every gridded library is given a unique number which is also sprayed on every microtitre plate of the library along with the copy number and the plate number in a barcode format. All library copies are stored at -80°C.

b. Library spotting on nylon membranes

Insitu bacterial colony filters
Hybond N+ membranes were placed on 3MM paper pre-wetted in 2YT medium. The membranes (Hybond N+) after being spotted, as described above, are overlayed on 2YT agar media supplied with 100 µg/ml ampicillin trays and incubated for 14-16 hrs at 37°C.
bacterial colonies are lysed by a modified alkaline lysis protocol (Nizetic et al., 1991) as follows:

The spotted membranes were overlayed on prewetted in denaturing solution (5M NaOH, 1.5M NaCl) 3MM Whatman paper for 5 min which was placed on a glass plate fixed above water level in a waterbath at 95°C, for 5 min.

The filter is neutralised by placing it on pre-wetted in 1M Tris-HCl pH7.5, 1.5M NaCl 3MM Whatmann, for 5min.

Proteolysis was achieved by immersing the filter in Pronase buffer containing 250 μg/ml Pronase for 20 min. The filter was dried on Whatmann and the DNA fixed by UV crosslinking.

**PCR filters**

Hybond N+ membranes (Amersham) were overlayed on pre-wetted in denaturing solution 3MM Whatmann placed on tiles in the robot. PCR products were spotted 10 times at the same spot using 400 μm in diameter pins. It was estimated that up to 100 ng of PCR is transferred on each spot. The spotting pattern was 5x5 duplicate where the center of the block was occupied by salmon sperm DNA (600 ng/μl). Since it takes approximately 8-10 hrs to spot 5 filters, it is important that the filters are kept wet throughout the spotting run. Drying of some areas will cause uneven DNA transfer and distorted pattern.

The alternative of spotting on dry membranes overlayed directly on the tiles (where due to static electricity the filters stick easily) or on the top of dry Whatman papers (using this method, it is very difficult to achieve even attachment of the filters through the complete run) was tried. It was found that filters spotted with this method, last through approximately 5 hybridisations while wet spotted filters can be used for over 15 times.

Moreover, PCR products of 1,920 partially sequenced genomic clones (average size of 500-700 bp) covering 66 kb from the human MHC class II region on chromosome 6 (Beck et al., 1992) and containing the LMP2, LMP7, TAP1 and TAP2 genes, were spotted on each membrane as controls for evaluating the quality of the oligonucleotide hybridisations.

All PCR spotted filters were hybridised initially with ssDNA in order to light up the central spots of all the blocks which help aligning the grid during automated image analysis. Moreover, an oligo (pSport1imer: 5' GCACGCGTACG-3') matching at 197-207bp of the pSport1 vector (at the 3' site of the cDNA insert) which is co-amplified along with the cDNA insert, is hybridised in order to quantify the amount of the DNA spotted.

Each PCR filter was processed by overlaying it twice for 10 min on top of two 3MM Whatman papers prewetted in denaturing solution (1.5M NaCl, 0.5M NaOH), followed by
twice for 10 min in neutralising solution (1M Tris-HCl pH 7.5, 1.5M NaCl). The filters were left to air-dry and the DNA was fixed by UV crosslinking.

2.3.5. Genomic DNA extraction

Genomic DNA was prepared from individual amphioxus since it has been observed some degree of polymorphism within the same population. Frozen adult amphioxus (approximately 3-4 cm length, 1 cm width) was mixed with liquid nitrogen and ground in a pestle and mortar pre-cooled on dry ice. The resulting powder was slowly added to 40 ml of buffer (TEN9) (see 2.2.5) while swirling constantly. Moreover, RNase up to a final concentration of 100 µg/ml was added and the solution was incubated for 10 min at RT. Moreover, ProteinaseK (1 mg/ml final concentration) and SDS (1% final conc.) were added and the solution was incubated at 50°C overnight while shaking gently. The disrupted proteins were discarded by phenol (2x) followed by chloroform extraction (1x). The DNA was precipitated by mixing with 0.8 vol. of isopropanol in the presence of CH₃COONa (100 mM final concentration) at RT. The DNA was spooled out from the solution with a glass rod, washed in 70% ethanol and dissolved in 500 µl of 1xTE (see 2.2.1).

2.3.6. Purification of plasmid and cosmid DNA (mini-preparation)

Alkaline lysis method (Birnboim and Doly, 1979).

Plasmids of high copy number e.g. pUC derivatives were grown in LB medium while cosmids which are generally found in low copy number per cell were grown in 2YT medium.

5ml of 2YT medium containing the appropriate antibiotic were inoculated with a single bacterial colony. The culture was incubated with shaking at 37°C overnight. Glycerol stocks were prepared by mixing 200 µl of the culture with 300 µl of 2YT media supplied with HMF and the appropriate antibiotic (see 2.2.4).

The remaining culture was centrifuged at 4,000 rpm at 4°C for 15 min. The supernatant was discarded and the pellet was resuspended in 100 µl of AL1 buffer(see 2.2.1), transferred to an 1.5 ml eppendorf tube and incubated at RT for 5 min. This was followed by adding 200µl of AL2 lysis buffer and gently inversion of the tube three times. After 5 min on ice, 150 µl of AL3 was added, the tube was inverted and vortexed for 10 seconds. After incubating on ice for 15 min, the tube was centrifuged for 20 min at 10,000 rpm and the
supernatant was transferred into a fresh eppendorf tube. To this 1/100 of the total volume RNase (10 mg/ml) were added and the tube was incubated at 37°C for 1 hr.

An equal volume of phenol was added to the tube which was then vortexed and centrifuged for 2 min at 10,000 rpm. The upper aqueous layer was transferred to a separate tube and residual phenol was removed by extraction with an equal volume of chloroform. The aqueous layer was transferred to a fresh eppendorf tube where it was mixed with 2 volumes of cold ethanol and 1/10 volume of 3 M sodium acetate (pH 5.2) and incubated at RT for 5 min. The DNA was visible as a gel-like pellet after centrifuging at 10,000 rpm for 20 min. The DNA pellet was washed with 70% ethanol, briefly dried in a SpeedVac vacuum desiccator and resuspended in the appropriate volume of 1xTE buffer. The above method results in DNA yields between 0.5 μg-1 μg/ml of inoculated culture.

2.3.7. Southern Blot

Agarose gels were stained with EtBr, photographed, soaked in 0.2 M HCl for 10 min (depurination facilitates the transfer of DNA fragments greater than 15 kb) and rinsed with distilled water. The gel was initially submerged in denaturing solution (1.5 M NaCl, 0.5 M NaOH) for 45 min and it was subsequently neutralised in by soaking for 45 min in 1 M Tris-HCl (pH 7.4), 1.5 M NaCl.

The DNA was transferred onto Hybond N+ membrane (Amersham) in 10xSSC overnight in a container as described in (Sambrook et al., 1989). After the transfer the membrane was rinsed in 6xSSC and dried on Whatman paper. The DNA was fixed by UV crosslinking.

2.3.8. Polymerase chain reaction

2.3.8.a. Purification of TAQ polymerase

The following protocol is from Engelke et al., 1990 after being modified by L. Schalkwyk and S. Meier-Ewert (Max-Planck Institut für Molekulare Genetik):

The *Thermus aquaticus* DNA polymerase I gene cloned in the pTP4 expression vector (as described in (Engelke et al., 1990)) was electroporated in XL1 blue cells and the transformants were transferred in 1 ml of SOC (see 2.2.4) and incubated at 37°C for 1 hr.
The cells were then streaked on petri dishes containing 100 μg/ml ampicillin and incubated at 37°C overnight. A single colony was used to inoculate 50 ml of 2xYT supplied with 100 μg/ml ampicillin and incubated with shaking at 37°C overnight. 8ml of the overnight culture were inoculated into 1lt of 2YT media containing 100 μg/ml ampicillin in a 2lt conical flask and incubated at 37°C until OD600 reached 0.2. At this point Taq expression was induced by adding 670 μl of 1M IPTG (or up to 670 μM final concentration, 750 μl of 20% stock) and it was incubated for a further 16 hrs. The cells were pelleted by centrifugation at 4,000 rpm in Beckman J6 centrifuge and resuspended in 100 ml buffer A (0.05 M Tris-HCl pH 7.9, 0.05M Dextrose, 0.001 M EDTA). The cells were pelleted again and they were resuspended in 20ml of buffer A, containing 80 mg/ml lysozyme. After 10 min at RT, PMSF to 1 mM and 20 ml buffer B (0.01 M Tris-HCl, 0.05 M KCl, 0.001 M EDTA, 0.5% Tween-20(v/v), 0.5% Nonidet P-40(v/v) was added. The sample was vortexed, incubated at 75°C for 45 min and cooled on ice for 2 min. The supernatant formed after centrifugation at 13,000 rpm was loaded on 50 ml Biorex 70 column, previously equilibrated with 6 volumes (6x50ml) of Buffer C (0.02 M Hepes pH 7.9, 0.001 M EDTA, 0.001 M PMSF, 0.5% Tween-20 (v/v), 0.5% Nonidet P-40 (v/v)) plus 0.05 M KCl.

The column was washed with 6 column volumes of buffer C plus 0.05 M KCl and the bound Taq protein was eluted with buffer C plus 0.2 M KCl. 10 fractions of 10ml each were collected. The activity of each fraction was assayed by a nick-translation assay. Most of the activity comes off in the fractions 3,4 and 5 but residual activity comes off in the next 3 fractions.

The fractions with the highest activity were pooled and dialysed against 2 lt of buffer D (0.02 M Hepes pH 7.9, 0.1 M KCl, 0.1 mM EDTA, 0.001 M DTT, 0.5% mM PMSF, 0.005% gelatine, 50% glycerol) at 4°C overnight.

2.3.8.b. Taq polymerase activity assay

100 ng of sonicated Lawr4 vector were incubated with 5 µl of isolated Taq, 0.025 M Tris-HCl pH 7.6, 0.025 M KCl, 0.002 M MgCl₂, 0.001 M 2-mercaptoethanol, 0.2 mM dATP, 0.2 mM dTTP, 0.2 mM dCTP, 0.2 mM dGTP, 30 μCi [32P-α] dATP in 50 µl volume. The reaction was incubated at 74°C for 30 min and then cooled on ice. The labelling efficiency was checked by thin layer chromatography (TLC), where 1 μl of each reaction was spotted on PEI paper (Polygram, Macherey-Nagel, Germany), air dried and vertically chromatographed in a beaker containing 0.75 M potassium di-hydrogenphosphate pH 3.5 (the running buffer
should not cover the spots). The solvent front was allowed to progress until approx. 15 cm from the start, the sheet was subsequently air dried, wrapped in cling film and exposed on phosphorimager screen for 20 min. The labeled probe should appear at the point where the sample was applied as a thick spot while the unincorporated nucleotides migrate with the solvent front.

2.3.8.c. PCR of a whole cDNA library (50,000 clones)

The PCR reactions were performed in 384-well V-shape bottom polypropylene microtitre plates (Genetix, Christchurch, Dorset, UK). The wells are arrayed in 16 rows x 24 columns, labeled A through P and 1 through 24. The spacing between the wells facilitated the automated pippeting of the PCR mix as well as the transfer of the PCR products on nylon membranes.

The PCR reaction for each cDNA clone was carried out in a 40 μl end volume which consisted of 10% (4 μl per 40 μl reaction) PCR buffer (0.5 M KCl, 1% Tween 20, 0.015 M MgCl₂, 0.35 M Tris-base, 0.15 M Tris-HCl), 100 μM (0.04 μl per 40 μl reaction of 100 mM stock solution) dNTPs, 162.5 μM (or 6.5 pmol per 40 μl reaction) of each primer, 1 U/μl of TAQ. The PCR primers used were the M13FSP (32mer: 5'GCTATTACGCCAGCTGGCGAAAGGGGGATGTG 3') from the 5' and pSport 3/86 (20mer) 5' CCAGTCCGGAATTCCGGTTGAAT 3' from the 3' (detailed map of the MCS of the pSport1 is given in Fig. 2.1).

Both amphioxus cDNA libraries were PCRed in batches of 66 plates (25,344 clones) where the PCR components (PCR buffer, dNTPs, primers, TAQ) were mixed at the concentrations described above. This PCR mix was subsequently pippetted using a robotic system, into 384-well (50 μl end volume) polypropylene plates (Genetix) labeled with numbers corresponding to the 384-well plates of the arrayed cDNA library to be PCRed. All the wells of the PCR plates were inoculated with a small amount of culture from the analogous cDNA culture plate, using 384-pin polypropylene gadjets (Genetix).

The PCR plates were then heat-sealed with a bilaminar nylon/polypropylene film (0.45mm thick), using a commercial sealing device (Genetix, Christchurch, Dorset, UK) which consists of a base with a niche where each plate is positioned and a heated plaque that is pressed on top of the plate, and the film. The sealing time for each plate was approximately 10 sec.
Afterwards, the sealed plates were inverted so that the mix from within the wells is drawn as much as possible towards the film seal where heat is transferred more efficiently in comparison to the side-walls of the wells.

PCRs were performed using an in-house constructed thermocycling robot consisting of three heated 255 L waterbaths set at the appropriate temperature for denaturation, annealing and extension. The sealed plates were loaded in a basket (Fig.2.2), attached on a robotic arm, that cycled between the baths. The waterbath temperature as well as the temperature inside the plate wells is controlled through probes which give the possibility of adjusting the time that the mix within the plate-wells is at the programmed temperature with the thickness of the plates. Finally, the plates were cycled between 96°C (30sec) and 72°C (5min) for 30 times.

Using the above robot RCR up to 51,840 (135 plates) amplification reactions could be simultaneously performed. The whole system is programmed through Visual Basic software on a PC.

Finally, the PCR plates were briefly centrifuged at 2,500 rpm using microtitre carriers on a Beckman J6 centrifuge and unsealed by heating using the heat sealing device described above. The time needed for each plate to unseal was approximately 1 min.

The success of each PCR run was tested by electrophoresis of the PCR products from the top and middle rows of a control plate. Usually a random plate of the library was PCRed twice and one of them was used as a control plate. This allows to check for any well-leakage problems across the whole plate as well as the variation in the PCR yield in relation to the well position. Moreover, aliquots of the PCR mix taken at multiple steps of the PCR setting up procedure e.g. before and after is pipetted through the robotic dispenser are run on the same gel to check for possible contaminations.

The PCR products are then transferred on nylon membranes as described in 2.2.4.I

2.3.9. Probe labelling

2.3.9.a. Random primer labeling

Long-length (more than 100 bp) cDNA clones, PCR derived fragments or cosmids were labeled by the random primer labeling method (Feinberg and Vogelstein, 1984). In this procedure, random sequence hexanucleotides are hybridised to a denatured double-stranded (or single-stranded) template at multiple sites along the whole length and serve as primers
that are extended by the 5'-3' exonuclease activity of the Klenow fragment of DNA polymerase I. Labeled (either radioactively, biotin or digoxigenin) deoxynucleotides present in the reaction mixture are incorporated in the generated double-stranded DNA.

For DNA fragments of less than 10 kb in length 50 ng were labeled while for cosmids 100 ng were used. However, even lower amount of DNA (up to 10 ng for less than 10 kb fragments and 50 ng for cosmids) have been suggested. The amount of labeled DNA influences the specific activity of the resulting probe. Large amounts of DNA used result in lower specific activities and shorter probes, while reducing the amount of DNA result in longer probes, but longer incubation times are also needed.

When PCR products were labeled, the DNA was purified away from unincorporated nucleotides and primer dimer artifacts through silica columns (QIAGEN, QIAquick columns). Alternatively, purified plasmids were enzyme digested and the fragment of interest was purified from LMP agarose.

The labeling procedure was as follows:

The DNA to be labeled was diluted in dH2O in a final volume of 30 μl and denatured by boiling for 5 min and then cooled on ice. The DNA solution volume was measured (and completed up to 20 μl) and an equivalent volume of 2xRP mix was added.

The 2xRP mix contained:

For 1140 μl mix

- 500 μl 1M Hepes-Na pH 6.6
- 500 μl TM (defined below)
- 30 μl 10 mg/ml BSA
- 60 μl TG
- 20 μl OL
- 30 μl H2O

TM: 250 mM Tris-HCl (pH 8.0)
- 25 mM MgCl2
- 50 mM β-mercaptoethanol

OL: 45 O.D. units/ml hexamers diluted 1 in 16 (i.e. A260=45)

TG: 7.6 μl each from 100 mM stock solution of dCTP, dTTP and dGTP in 1 ml of dH2O

The 2xRP mix was kept frozen at -20°C in 200 μl aliquots.
Final concentration in the labeling mix (containing equal volume of DNA solution and 2xRP mix):

- 0.219 M Hepes-Na pH 6.6
- 0.219x TM
- 0.13 mg/ml BSA
- 0.04 mM TG (nucleotides)

Finally, 1 μl (10 Ci/μl) of 32-PdATP (3,000 Ci/mmol) and 2 μl E.coli DNA polymerase I (Klenow fragment) (10 U/μl) were added to the 40 μl of the labeling mix.

The above reaction was incubated at 37°C for 3 hrs or at RT overnight. The efficiency of the labeling reaction was checked by paper chromatography as follows:

- 1 μl aliquot of the reaction was spotted onto polyethyleneimine (PEI) sheets (Polygram, Macherey-Nagel, Germany), air dried and vertically chromatographed in a beaker with 0.75 M potassium di-hydrogenphosphate pH 3.5 (the running buffer should not cover the spots).
- The solvent front was allowed to progress until approximately 15 cm from the start, the sheet was subsequently air dried, wrapped in Saran wrap and exposed on phosphorimager screen for 20 min. The labeled probe should appear at the point that the sample was applied while the unincorporated nucleotides migrate with the solvent front.

The labeling reaction was terminated by adding 0.5 M EDTA to a final concentration of 0.05 M. The probe was precipitated by adding 1/2 volume of CH₃COONH₄ 7.5 M, 1.5 volume ethanol and 1 μl 0.5% linear polyacrylamide (LP) as carrier (Gaillard and Strauss, 1990) and centrifuging at 10,000 rpm for 10 min. Higher polyacrylamide concentrations e.g. 10-20 μg of linear polyacrylamide in a DNA solution containing >0.1 M salt and 2.5 volumes of ethanol, have been also suggested (Gaillard and Strauss, 1990). Finally, the probe was dissolved in 40 μl of dH₂O. Unless used immediately, labeled DNA was stored at -20°C.

### 2.3.9.b. Oligonucleotide labeling with T4 polynucleotide kinase

Oligonucleotides were labeled at their 5' end using γ-33PdATP and T4 polynucleotide kinase in a 30 μl reaction as follows: 10 μl oligo (30 pmol), 3 μl 10x buffer (700 mM Tris-HCl pH 7.6, 100 mM MgCl₂, 50 mM DDT) 2 μl T4 polynucleotide kinase (10 U/μl), 50 μCi γ-33PdATP and 10 μl H₂O were incubated 37°C for 45 min.

The reaction was terminated with 3.5 μl of 0.5 M EDTA. The labeling efficiency was checked by pipetting 1μl of the reaction mixture onto a PEI paper (Polygram, Macherey-
Nagel, Germany) chromatography using 0.75M potassium di-hydrogenphosphate pH 3.5 as running buffer as it was described above (2.2.7.a).

2.3.10. Hybridisation conditions

2.3.10.a. Oligonucleotide hybridisation

The hybridisations were performed overnight at 4°C in glass bottles that were constantly rotating. Up to 4 filters were hybridised in the same bottle containing 12 ml 1xSSarc buffer (4xSSC, 7.2% sodium lauryl sarcosinate and 1 mM EDTA) at 4 nM oligo final concentration. The filters were subsequently washed in 1xSSarc buffer for 20 min at 4°C, dried briefly by blotting on 3MM Whatmann paper and exposed on Phosporimager screens (Molecular Dynamics) overnight.

The screens were scanned at 176 μm resolution and the resulting 16bit TIF images (4.2 Mb) were transferred to a DEC alpha Unix workstation for further analysis. Phosphor storage autoradiography is at least 10 times more sensitive than conventional film based autoradiography and allow linear measurement of the hybridisation signal over a broader range [Johnston, 1990 #296].

The amount of DNA spotted on the filters is estimated by hybridising an 11mer oligonucleotide which is complimentary to the part of the vector co-amplified with the cDNA insert and thus present in all the spotted PCR products.

After each hybridisation the membranes were cleaned from the probe by washing in 0.1xSSarc at 65°C for 20 min. Each membrane could be re-used up to 20 times without significant loss in the amount of the DNA spotted.

2.3.10.b. Cross-species hybridisations using long probes

All the hybridisations were performed in glass bottles (Hybaid). The bacterial colony filters were separated by mesh membranes (Hybaid) thus ensuring adequate diffusion of the probe between the filters. DNA probes from organisms other than amphioxus were hybridised on amphioxus cDNA libraries under the following conditions:
Up to 4 filters/bottle were pre-hybridised for 20 min at 42°C in 20 ml of hybridisation solution containing:
50% formamide
1% SDS
40% hyb-mix (20% w/v Dextran sulphate, 50% v/v 20xSSC, 12.5% 1 M Na₂HPO₄, pH7.2, 0.5% 0.5 M EDTA, 25% 100xDenhardts, 1.25% salmon sperm DNA (5 mg/ml).

The filters were hybridised overnight at 42°C in 14 ml of the same solution and they were subsequently rinsed initially twice in 3xSSC, 0.1% SDS at RT followed by double 20 min wash at 65°C.

### 2.3.10.c. Same species hybridisations using long probes

The filters were pre-hybridised in Church buffer (0.25 M Na₂HPO₄, 5% SDS, 1 M EDTA) for 20 min at 65°C and they were subsequently hybridised in the same buffer at 65°C for at least 16 hrs. Excess probe was removed by washing the filters in either 2xSSC 0.1% SDS (low stringency) or 0.04 M Na₂HPO₄, pH 7.2 (0.02 M Na₂HPO₄), 0.1% SDS (high stringency) initially briefly at RT and then twice at 65°C for 20 min.

Filters were cleaned from the radioactive probes by washing them either in 0.04 M NaPi, 0.1% SDS for 20 min at 65°C (gentle wash) until no detectable counts or in 0.2 M NaOH, 0.1% SDS for 10 mins at RT. The filters were subsequently washed in 1xTE pH 8.0 and stored either in 1xTE at 4°C or dried until they are still damp, wrapped in cling film and stored at -20°C (long term storage). The filters can be re-used for 20-25 times without significant loss in the amount of DNA present on the filter.

### 2.3.11. Automated sequencing

cDNA inserts were mostly sequenced by the dideoxytermination method (Sanger et al., 1977) where four ddNTPs each tagged with a different fluorescent dye (Prober, 1987), are incorporated during the DNA synthesis. As a result, the synthesized fragment is terminated and labeled with the dye that corresponds to that ddNTP base. The advantage of the above method is that all sequencing reactions are performed in the same tube thus reducing the
amount of sample handling. Moreover, the fluorescent sequencing fragments are resolved on a single lane on polyacrylamide gel, thus increasing the sample throughput.

All sequencing reactions were set up using a commercial kit (dye terminator ABI PRISM) which contains dNTPs and ddNTPs at the appropriate ratio as well as a thermostable polymerase (Amplitag), magnesium chloride and reaction buffer. The dye terminator molecules consist of a Rhodamine dye (R110, R6G, TAMRA or ROX) attached to a dideoxynucleotide via a linker. The above dyes have been recently replaced by new dichlororhodamine dyes (ABI) that have narrower emission spectra which is also shifted more towards the red. This reduces the spectral overlap among the four dyes in the set and, therefore, reduces background noise. The emission spectra of the four dyes is between 530nm and 620 nm. Sequencing reactions were run on 377 ABI sequencers.

A few cDNA clones were sequenced using the dye primer method while the sequencing reactions were run on a LICOR sequencer. The advantage of the LICOR system is that longer sequence reads are produced due to the different dyes and/or detection system that is used. However, the software for reading the bands is not fully automated. The dye primer method involves the use of either four differently labeled primers therefore four reactions are set up and the samples can be subsequently pooled and run in the same lane or the use of one labeled primer and the samples are run on four separate lanes. The LICOR system has two detection channels, at 700 and 800. The primers are labeled at their 5' with an infrared dye. Dyes mostly used for the LICOR sequencers are the IRD-40 or IRD-41.

2.3.11.a. Template preparation

PCR products were purified using silica membrane columns (QIAGEN). DNA is absorbed on the membrane in the presence of high salt buffer at pH 7.5 while remaining primers, primers-dimers and unincorporated nucleotides, salts, detergents such as Tween 20 are washed away by an ethanol containing buffer, and the pure DNA is eluted with buffered water (pH 7.0 and 8.5).
2.3.11.b. Sequencing reactions

**ABI**

Purified PCR products of the cDNA clones were mainly used as template for sequencing. 100 ng/kb of PCR product were mixed with 4 μl of ABI dye-terminator mix (described above), 1 μl (10 pmol/μl) of the appropriate primer and H₂O up to 20 μl final volume. The reaction was cycled between 96°C for 10 sec and 60°C for 4.05 min for 25 times. The above annealing-extension temperature applies when the pSport 3/86 primer (5'- CCG GTC CGG AAT TCC CGG GT -3', Tm: 65.4°C) was used.

The labeled fragments were precipitated with 0.1 M of CH₃COONa pH 5.2 in the presence of ethanol (to 20 μl of the sequencing reaction, 2μl of 3 M CH₃COONa pH 5.2 and 50 μl of ethanol). The mix was left for 10 min at RT and it was subsequently centrifuged for 1 hr at 4,000 rpm. The pellet was washed with 250 μl of 70% ethanol and it was dried for 5min in a SpeedVac vacuum desiccator.

All the samples were run on 66 lane gels on ABI 377 sequencers for 12 hrs and data was processed using the ABI sequence analysis software v3.0 and v.3.1 while lane tracking was carried out manually using the Perkin Elmer manual lane tracking kit.

**LICOR**

cDNA clones were sequenced from both ends in the same reaction. 100 ng/kb of template DNA were mixed with 2 pmol of each of IRD-41 or IRD700 labeled primers, 0.5 μl of DMSO (approx. 3%)and H₂O up to 13 μl final volume. Four sequencing reactions for each cDNA clone were set up consisting of 3 μl of the above mix and 1 μl of the corresponding termination mix (Tris-HCl pH 9.5, MgCl₂, Tween 20, Nonidet P-40, 2-mercaptoethanol, the corresponding ddNTP, all dNTPs at the appropriate ratio and 7-deaza-dGTP).

The following primers were used for sequencing from both ends on the LICOR sequencer: from 3' 5'- TTT TTT TTT TTT TTT TTT TTT TTT T(AGC)-3' or T28V(29mer) with a Tm of 47.8°C, labeled at 5' with IRD700 which is detectable at the 700 channel, from 5' pSport50 (17mer) 5'- TGG TAC GCC TGC AGG TA- 3' with a Tm of 55.2°C, labeled at 5' with IRD 41 detectable at 800 channel.

All the sequencing reactions were set up in 96-well thermofast plates, using the Thermo Sequenase fluorescent labeled primer cycle sequencing kit (Amersham) which except the standard dNTPs and ddNTPs, contains 7-deaza-dGTP in the nucleotide mixes. This modified nucleotide prevents the formation of 'hairpin' loops between GC residues within the newly synthesized strand since it is not able to form stable base pairs with C nucleotides. The
presence of hairpin loops changes the electrophoretic mobility of the terminated molecules causing them to run faster or slower (gel compression). Moreover, the use of the Sequenase which lacks 3'-5' exonuclease activity prevents their removal. The samples were cycled between 95°C for 2 min and 50°C for 15 sec for 25 times.
Chapter 3

Oligonucleotide Fingerprinting

Introduction

At the outset of this project the major components required for the realisation of this study were already in place. Besides the robotic equipment for example PCR thermocycler, the spotting and picking robots that had been previously developed, the experimental conditions for the hybridisation of 10mer oligonucleotides had been assayed in a pilot study on 30,000 cDNA clones of a human fetal brain cDNA library (Meier-Ewert, 1994). Moreover, initial versions of inhouse developed image analysis and clustering algorithms were available. In the meanwhile, many steps on the experimental and analytical aspect of this procedure have been further developed and refined. For example, spotting of PCR products in higher density (5x5) was introduced thus increasing the sample throughput per single experiment (Chapter 2).

During this study, oligonucleotide fingerprinting is applied on a larger number of clones thus giving the chance to identify even transcripts that are expressed at very low level. The factors that affect clustering within each library and across multiple libraries will be also described.

This chapter is divided in three parts, the first describing the experimental components that are important for the success of the oligonucleotide fingerprinting, the second outlining the data analysis procedure consisting of the image analysis and clustering and finally an account of the fingerprinting results on both amphioxus libraries.
3.1. Oligonucleotide Fingerprinting

3.1.1. Experimental parameters

The major experimental components that define how successful the oligonucleotide fingerprinting of a library will be, are the quality of the library itself, the set of the oligonucleotides that are used and finally the hybridisation conditions. Moreover, other secondary parameters, equally important for the success of the above procedure, depend mainly on the condition of the equipment used. The production of good quality PCR filters for example, depends on the spotting as well as the yield of the PCR reaction which in extent is determined by factors such as the plates used, once the PCR conditions have been established. These secondary parameters have been described in the material and methods section (Chapter 2).

3.1.1.1. Library quality assessment

Total RNA was isolated from 5-6 hr and 26 hr embryos using the (Chomczynski and Sacchi, 1987) method. Approximately 200 μg of total RNA were used for the construction of each library. The RNA was checked for degradation on 1% agarose gel. Poly A+ RNA was purified with oligo (dT)$_{30}$ coated magnetic beads (Dynabeads, Dynal, Norway). The cDNA libraries used were unidirectional, oligoT primed and cloned into pSport1 vector.

After construction, the quality of the cDNA libraries was evaluated by determining the ratio of recombinants to non-recombinants and the average insert size by PCR of approximately 100 clones.

The average insert size of the cDNA clones is on average between 1-1.5 kb for the neurula stage and 0.75-1 kb for the gastrula stage library (Fig. 3.1.). The number of clones with no insert is approximately 8-10% for both libraries.
Fig. 3.1. The average insert size of the BFLG library is 750-1000bp (A) and 1.5-2kb for the BFL26 library (B) as estimated by the size of the PCR products of randomly selected clones from both libraries. The PCR products were electrophoresed on 1% agarose gel. The standard size marker used was 1 kb DNA ladder (Stratagene).
3.1.1.2. Oligonucleotide selection

In overall 230 decamer oligonucleotides which were degenerate at both ends were used to hybridise against the PCR products of 50,688 cDNA clones from each embryonic cDNA library. The sequences of those oligos are given in Appendix I.

The above oligonucleotides were initially designed for the fingerprinting of a human fetal brain cDNA library (Meier-Ewert, 1994) while they were also subsequently used for the fingerprinting of a 9d and a 12d mouse cDNA libraries (Meier-Ewert et al., 1998). The oligonucleotides used were selected among 32,768 (4/2) possible octamer combinations as having more than 10% occurrence frequency in 2,000 human coding sequences, retrieved from the Genbank, and thus being able to efficiently distinguish between those sequences. Oligonucleotides of low complexity (e.g.poly T) as well as those matching the part of the pSport1 vector co-amplified with the insert were excluded.

The reason for using the above oligonucleotides which were designed based on human sequences for fingerprinting an amphioxus library was that:

a. less than 100 amphioxus sequences were deposited in the Genbank when this project started. This limited sequence information did not enable the design of amphioxus specific oligonucleotides.

b. the hybridisation behavior of the above oligonucleotide set was already assessed. Oligonucleotides having low hybridisation rate although they were predicted to be frequent were rejected since it was assumed that they failed to give sufficient amount of hybrid under the common experimental conditions for all the oligonucleotides.

Furthermore, the efficiency of this set of oligonucleotides to recognise and distinguish amphioxus sequences was assessed by partitioning the amphioxus sequences currently deposited in the Genbank. This was recently repeated for 162 sequences currently available (March 1999). Each of the amphioxus sequences gave either full or partial 7mer match with on average 50% of the set of oligonucleotides used.

Finally, clustering of the above sequences resulted in 134 singletons-distinct genes and 8 clusters. The above clusters of redundant sequences which were Pax6, calmodulin, muscle actin, cytoplasmic actin, IF protein A, IF protein B, troponin C, Pax-2 (Table 3.1.C.).

The above theoretical clustering is a two step process where initially the oligonucleotides present in the amphioxus sequences deposited in the Genbank are found. This result is formulated in a table where for each Genbank sequence, the oligos that are not present are
designated 0, those with 7mer match 0.7, while those with full octamer match, 1 (Table 3.1.A).

In the next step the oligonucleotide content (theoretical fingerprint) for each clone is compared to that of the rest of the clones and according to the number of their common oligos, they are grouped together in the same cluster. Table 3.1. A shows the scoring table for the amphioxus cholinesterase 2 gene. Table 3.1.B. shows how many times each oligo appears in the same sequence of amphioxus cholinesterase 2 gene. The oligos that have two partial 7mer matches are given the number 1.4 (2x0.7) or have one partial 7mer match and one full 8mer match 1.7 (0.7 +1.0). It is intended in the future to calculate the occurrence frequency of each oligonucleotide across all amphioxus sequences that will be generated as result of this project. This will enable to identify and replace the oligonucleotides that have lower number of hits and therefore are less informative. Table 3.1.C shows the clusters of sequences and some singletons of the genes currently deposited in the Genbank.

Moreover, additional oligonucleotides designed against conserved motifs of gene families were added to the above oligonucleotide set. For example 30 oligonucleotides designed against conserved domains of members of the TGF-β superfamily (Wiles, M., unpublished) were hybridised on both amphioxus libraries. The results are currently being analysed. As an example within the amphiBMP2/4 gene deposited in the Genbank the oligonucleotides that have either full or partial match are:

clone gil4039084/gblAF068750/AFO68750 total hits = 130 Amphioxus BMP2/4

partial matches 96: 0.071 0.079 0.080 0.084 0.085 0.087 0.088 0.089 0.090 0.092 0.093 0.094 0.095 0.101 0.102 0.108 0.109 0.111 0.112 0.113 0.118 0.123 0.124 0.126 0.128 0.129 0.130 0.133 0.138 0.139 0.140 0.142 0.149 0.150 0.152 0.155 0.156 0.159 0.161 0.164 0.166 0.168 0.174 0.175 0.176 0.181 0.182 0.186 0.187 0.192 0.194 0.198 0.199 0.202 0.203 0.206 0.213 0.216 0.217 0.220 0.222 0.224 0.231 0.235 0.238 0.241 0.246 0.250 0.251 0.255 0.257 0.258 0.261 0.263 0.265 0.267 0.270 0.271 0.274 0.275 0.276 0.278 0.279 0.280 0.281 0.283 0.285 0.286 0.287 0.288 0.289 0.04 0.06 0.10 0.19 0.30

full matches 34: 0.068 0.075 0.078 0.082 0.096 0.106 0.114 0.120 0.131 0.146 0.162 0.167 0.169 0.171 0.173 0.178 0.195 0.207 0.225 0.232 0.237 0.252 0.264 0.269 0.272 0.273 0.284 0.001 0.011 0.012 0.015 0.016 0.021 0.025

Within the AmphiBMP2/4 the oligonucleotides 0t01 0t11 0t12 0t15 0t16 0t21 0t25 are TGF-β oligonucleotides.

Table 3.1. A Table showing the presence of partial (0.7), full match (1.0) or absence of 207 of the oligonucleotides used in the fingerprinting of both amphioxus libraries in the amphioxus cholinesterase 2 gene sequence. B. The same as above scoring where in addition to the presence or absence of each oligonucleotide, the occurrence frequency is shown. C. Clustering of all 162 amphioxus sequences deposited in the Genbank based on their theoretical fingerprints created with the set of the oligonucleotides used in the experimental work. Only the clusters and a few of the singletons (cluster 9) are shown. The full clustering file is given in Appendix IV.
In the case of the Pax6 gene the majority of nucleotide differences between the five clones 12.1, 12.2, 4.1, 4.2, and J2 deposited in the Genbank, are mainly at the 5' within a region of 600 bp. The length of 12.2 and 4.2 is approximately 2,960 bp while the 12.1, 4.1, and J2 clones are approximately 4,170 bp. Some nucleotide differences lead to differences in the deduced amino acid sequence (Glardon et al., 1998) while others are not causing any
changes (approximately one out of 100 positions in the coding region and one out of 40 positions in the 5’ and 3’ UTR). Since the oligonucleotides matching those sequences are spread over the entire length of the clones and the clones have more than 2,000 bp of common sequence, their clustering is not affected.

3.1.1.3. Oligonucleotide hybridisations

All the oligonucleotide hybridisations were performed using filters spotted with PCR products. Among the advantages of using PCR filters instead of bacterial colony filters are that increased amount of DNA is available for hybridisation while cross hybridisation with bacterial or vector DNA is also avoided. The later allows the use of a broader selection of oligonucleotides that otherwise would be excluded as being complementary to vector or E.coli DNA.

Initially the PCR filters were hybridised with ssDNA (salmon sperm DNA) which highlights the position of the guide spots at the centre of each block (see 2.3.4. for explanation of the spotting pattern). Moreover, the position of each cDNA is lighted up by the hybridisation of the pSpot11mer oligonucleotide which is complementary at the MCS of the pSport1 vector (Fig. 2.1.), from the 3’ of the cDNA which is co-amplified with the insert (Fig. 3.2.).

At this stage, the best PCR filters were selected for the oligonucleotide hybridisations. Filters that had i) large areas with missing spots which might be due to inappropriate heat sealing of the plates before the PCR reaction ii) spots that were smeary due to the filter being dried and thus curved during the spotting run iii) uneven transfer of DNA amount during spotting due to false calibration of the spotting robot axis, were rejected.

In overall, 230 oligonucleotides were hybridised on each library. All the oligonucleotide hybridisation images were processed automatically using an inhouse developed image analysis. A brief description of the image analysis procedure is given below (3.1.2.1.).

In addition to the computational methods for assessing the image analysis procedure (see below) and oligo-hybridisation quality all the images were inspected visually. The criteria for selecting an oligonucleotide hybridisation for repetition were:

- the weakness of overall intensity signal
- whether the image analysis program has assigned properly the coordinates of the whole filter. Incorrect assignment occurs when the intensity of the signal of positives in comparison
Fig. 3.2. Hybridisation images of the ssDNA and sport11mer oligonucleotide on filters spotted with PCR products of cDNA clones from both amphioxus embryonic cDNA libraries. Each filter was spotted with 27,648 clones in duplicate, in a 5x5 spotting pattern. The above filters were spotted with the first 66 plates from each library (set 1). The BFL26 (left) filter contained areas empty of PCR product in a repeated pattern. This was due to inadequate plate sealing before the PCR reaction. The amount of PCR products from the BFLG cDNA library transferred on the filter was more even.
to that of the guide spots is stronger. In this case some of the positives that are spotted closer to the guide spot will be considered as the centre of the whole block and they will affect the automatic alignment of the whole grid. It has to be noted though, that in the majority of the cases, repetition of the oligonucleotide hybridisation did not yield better results than the first time unless if the reason for the bad hybridisation result was the quality of the filter.

3.1.2. Data analysis methods

3.1.2.1. Image analysis

The image of each oligonucleotide hybridisation (Fig. 3.3) was stored in a 16-bit TIFF grey level format which contains a header describing the image size and the pixel values for each spot. Image analysis was carried out automatically using an inhouse developed program (Ruben Schattevoy, Max-Planck Institut für Molekulare Genetik, Berlin, unpublished) as follows:

1. the four comers of each image were located manually by marking with a mouse the position of the guide spot in the most comer block. The resulting coordinates (ret file) were used to calculate the angle by which the image has to be rotated in order the pixel values (written in the TIF file) to correspond to the correct spot of the image.

2. an arguments file for every image was created. This file contained the information on the robot used for spotting, the spotting pattern, the number of the spotted plates, the plate spotting order and therefore the identity of each clone, the scanner used and the scanning resolution as well as the identity of the hybridisation probe.

3. a grid was aligned on the filter and the position of the each guide spot was found by creating a histogram of pixel values along rows and columns for the whole image. The guide spots were at the intersection of the corresponding histograms at the maximum pixel values.

4. the location of the cDNA clones within each block was found by searching for local maximum values around the expected spot position.

Attempts were made to normalise the signal intensity of each spot based on the
Fig. 3.3. The top left corner of the o158 (ggagaaga) oligonucleotide hybridisation image on 25,000 cDNA clones of the BFL26 cDNA library. The cDNA clones were spotted in a 5x5 duplicate pattern as shown in the square on the top right side of the photograph. The blockview window shows in 800% magnification a 5x5 block of 25 clones. The centre of the block is occupied by salmon sperm DNA (guide spot).
background hybridisation (i.e. the hybridisation with the sport 11mer) which gives an indication of the amount of DNA transferred on the membrane. Two methods were suggested: to normalise the mean signal value to the mean background value of both duplicate spots or to take the mean value of each individually normalised signal. However, neither of those two methods are currently used since they depend heavily on a single hybridisation the background hybridisations with pSport11mer which might not be successful.

The results of 117 oligonucleotide hybridisations for the BFL26 and 165 oligos for the BFLG were considered for clustering. The sequence of the oligonucleotides giving the best hybridisation results on both amphioxus cDNA libraries are given in Appendix II and III for each filter set. It is worth noting that oligonucleotide hybridisation results are not affected by filter effects. Oligonucleotides giving bad hybridisation results on one filter set, they give similar results on the rest of the library and even across both libraries.

A number of computational methods were developed in order to assess the probe specificity and reproducibility of an oligonucleotide hybridisation as well as the image analysis process itself.

a. \textit{m13 ratio:}

2,304 genomic clones (with an average size of 500-1000 bp) cloned into M13mp18/19 and covering 66 kb of human MHC class II region on chromosome 6 (Beck et al., 1992) were spotted on each PCR filter. The genomic sequence of this region is deposited in the EMBL database (accession number: X66401, HSMHCAPG). Moreover, the order and the 5' partial sequence of the spotted clones was known. The exact length of each clone was also determined which in combination with their already known start position within the sequenced contig allowed us to assume the entire sequence of each clone. 700 M13 clones for which the length of PCR product was visible as distinct band on the agarose gel were considered for the analysis.

An indication of the quality of an oligonucleotide hybridisation is the extent to which we can achieve discrimination between positives and negatives signals.

If the observed signal intensity ($x$) for a specific spot is a sum of the background noise signal ($\beta_1$), the signal of the partial matches ($\beta_2$) and the signal of the full matches ($\beta_3$):

$$x = \beta_1 + t_2 \beta_2 + t_3 \beta_3$$

where $t_2$ and $t_3$ represent the number of partial and full matches respectively.

Then the ratio (R) of the mean negative signal(noise) and the positive signal (full or partial match) can serve as an index for the hybridisation quality.
R = l - \beta_1 t_2 + t_3 \beta_3

Positives should have higher signal than the expected negatives. In the case of perfect data R tends to 1 while when there is no discrimination it tends to 0.

b. correlation of the intensities between two duplicate spots:

Since each clone (PCR product) is spotted in duplicate the signal intensities of each clone within a duplicate pair should be within a limited range. To calculate how well the signal intensity of each clone pair agree a correlation factor is calculated according to the following formula:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

where x and y are the signal intensities of the each duplicate clone and \( \bar{x} \) and \( \bar{y} \) are the corresponding mean values. As the difference in signal intensities between the duplicates lowers r tends to 1.

c. generation of artificial image:

The intensity value of each spot is used to recreate the oligonucleotide hybridisation image. Comparison of the signal intensities between 'experimental' and artificial hybridisation images eliminates the assignment of wrong intensity values due to miscalculation during image analysis.

3.1.2.2. Normalisation methods

The measurement of the signal intensities of clones in a single oligonucleotide experiment is more informative when it can be reproduced and it varies between different clones on the same filter. The latter depends on the specificity of the oligonucleotide probes used.

There are several factors that affect reproducibility: 1. between experiments, like differing probe concentration, washing and exposure time, hybridisation efficiency etc. 2. between clones due to differences in the amount of PCR product transferred on the filter.

These factors are eliminated by applying a normalisation procedure (double ranking) as described in (Milosavljevic et al., 1995) where in the first step, the intensity values of all clones within a hybridisation are ranked using a scale between 0 and 1. The clone having the highest signal intensity is given the value of 1 and the clone with the weakest, the value of 0.
while all other clones are given values in between. Thus the signal intensities of clones are compared within each single experiment which eliminates the artifacts introduced due to variation in exposure time, probe concentration, washing time etc.

In the next step all the signal intensity values of a single clone across multiple oligonucleotide hybridisations are re-ranked. In this normalisation step, variation in the signal intensities due to differences in the amount of PCR product, for the same clone, transferred on different filters is eliminated.

An alternative normalisation approach has been also proposed (Milosavljevic et al., 1995) where all hybridisation signal intensity values generated for each clone are divided by a 'mass' probe (an oligonucleotide probe that is complementary to the vector part amplified with every clone). This normalisation intends to limit the effect of variation in the amount of PCR products of clones spotted on the same filter. However, such normalisation depends heavily on the results and scoring of only the hybridisation of the above complementary probe.

During this study all the amphioxus filters were hybridised with an 11 mer complementary to the vector sequence co-amplified with the insert. The results of those hybridisations were used to identify the clones with the lowest intensity and therefore number of oligonucleotide hits. 25% (12,500) of the clones from each library were discarded during this step.

### 3.1.3. Clustering results

Clustering of the fingerprinted clones was attempted within each library as well as across both the libraries. Moreover, multiple clusterings were tried varying in the level of stringency in the assignment of clones to clusters (Table 3.2). Higher stringency leads to purer clusters meanwhile increasing the number of singletons and therefore the number of estimated transcripts present in a tissue. Nevertheless, this approach was favored since the average expression level of transcripts as implied from the cluster size will be unaffected. Moreover, since it is intended to tag sequence only one representative clone from each cluster, any misassigned clones would have been overlooked when clustering with lower stringency. Furthermore, the capacity of the above clustering strategy can be increased by secondary rounds of clone selection. For example selection of clones that are most dissimilar to the representative clones from the clusters.
Finally, two additional clustering approaches were tested:

- initial clustering followed by a second round of evaluation of singletons and their possible assignment into clusters
- clustering within each filter set followed by clustering of the consensus fingerprints across both sets and finally across both of the libraries. This method reduces the experimental filter effects.

The last two approaches have been recently developed (Herwig, R., et. al. 1999 in preparation) and they were used for creating the clusterings bfl26.all.res (dr), bflg.all.res (dr) and bflg26_co.res (dr).

Table 3.2. Each amphioxus cDNA library was clustered repeatedly using different clustering approaches. dr refers to the double ranking normalisation method (3.1.2.2), am refers to the arithmetic mean of duplicates, 1, 1.0, 1.2, 1.8 and 3.4 are arbitrary numbers referring to the stringency of the clustering where increasing numbers refer to lower stringency (more clones are clustered).

<table>
<thead>
<tr>
<th>Library</th>
<th>Number of clusters</th>
<th>Number of singletons</th>
</tr>
</thead>
<tbody>
<tr>
<td>BFL26</td>
<td></td>
<td></td>
</tr>
<tr>
<td>bfl26.all.res (dr)</td>
<td>3,696</td>
<td>9,806</td>
</tr>
<tr>
<td>bfl26._1.res (dr)</td>
<td>4,674</td>
<td>6,028</td>
</tr>
<tr>
<td>bfl26._1.0db.sop (dr)</td>
<td>4,133</td>
<td>9,858</td>
</tr>
<tr>
<td>bfl26._1.2db.sop (dr)</td>
<td>3,934</td>
<td>11,187</td>
</tr>
<tr>
<td>BFLG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>bflg.all.res (dr)</td>
<td>1,783</td>
<td>7,234</td>
</tr>
<tr>
<td>bfg_dr_am_1.8db.sop</td>
<td>3,177</td>
<td>11,910</td>
</tr>
<tr>
<td>bfg_dr_am_3.4db.sop</td>
<td>3,370</td>
<td>10,115</td>
</tr>
<tr>
<td>Co-Clustering</td>
<td></td>
<td></td>
</tr>
<tr>
<td>bflg26_co.res (dr)</td>
<td>4,994</td>
<td>16,097</td>
</tr>
</tbody>
</table>

A typical clustering output file has the following format:

```
# 72 122 0.101813
1 40627 0.158106 ICRFbfl26_55113
2 70378 0.157843 ICRFbfl26_128O6
3 42468 0.153105 ICRFbfl26_35N8
4 5262 0.152426 MPIbflg_29M12
5 7090 0.150179 MPIbflg_16C7
6 12043 0.149425 MPIbflg_3B19
7 28090 0.146352 MPIbflg_86E24
8 69466 0.145559 ICRFbfl26_126J9
9 42677 0.145109 ICRFbfl26_58D7
10 56886 0.142830 ICRFbfl26_79M18
11 10361 0.142219 MPIbflg_48G23
12 43749 0.140550 ICRFbfl26_16B4
13 13233 0.139400 MPIbflg_38H15
14 60469 0.138555 ICRFbfl26_76K8
```

The first line describes the cluster number (72) and its size (122) followed by a list of all clones that belong in this cluster. In the list of clones, the first column shows the number of
each clone in the cluster (1 is the closest clone to the consensus fingerprint), the second column shows the running number for this particular clone when clones are counted before being clustered, the third column shows the probability value for each clone that belongs to this cluster and the fourth column shows the library name that the clone belongs and its coordinates.

3.1.3.1. Neurula stage library (BFL26)

Out of 50,688 fingerprinted clones from the neurula stage (26 hrs) library approximately 35,482 were included in the clustering analysis while the rest were rejected as having signal intensity below a defined threshold during the hybridisation with the pSport11mer. This is due to the fact that these clones failed to give PCR product. Moreover, the above number includes clones that were discarded because of the great difference in the signal intensity of the duplicates spotted from this clone (see 3.1.2.1).

### Fig. 3.4
Cluster size distribution in the neurula stage library according to bfl26_1.res clustering. The cluster size for the majority of clusters range between 2 and 10 clones. The singletons are not included in the diagram. The column x gives the number of clones within each cluster while the column y the number of clusters for each cluster size in the bfl26_1.2db.sop (dr).
At least three different clustering approaches (Table 3.2.) were attempted using higher (bfl26_.1.0db.sop (dr)) or lower ((bfl26_.1.2db.sop(dr)) stringency. The clustering bfl26_.1.res (dr) was considered the best, based mainly on the backhybridisation data (3.1.4). According to this, 4,674 clusters (containing 29,454 clones) of different transcripts were identified, with size ranging between 815 and 2 clones as well as 6,028 singletons. An overview of the size distribution of clusters is given in Fig. 3.4. The size of the majority of the clusters is between 2 and 20 members.

1,673 clones, one per cluster, were selected from various size clusters, for 5' tag sequencing based on the bfl26_.1.2db.sop(dr) clustering. *Mitochondrial 16S rRNA, cytochrome c and b, NADH ubiquinone oxidoreductase* were among the most abundant transcripts (Table 5.1.).

### 3.1.3.2. Gastrula stage library

Similarly clustering of 35,482 clones (out of 50,688 spotted) from the gastrula stage(5-6 hrs) library, resulted in 3,177 clusters containing 23,572 clones. The clusters ranged in size between 2 and 2,896 members while 11,910 clones remained unclustered (singletons) (bfg_dr_am_1.8db.sop). *Mitochondrial 16S rRNA, muscle actin (1.78%), tubulin a-1 chain (0.88%) and NADH dehydrogenase subunit 2 (0.81%) comprised the first four larger clusters.* Table 5.2. gives the identity of some of the sequenced clones that belong to various size clusters.

The total number of different transcripts present in this tissue as estimated by the above clustering result (15,087) was considered very high. This was verified by the backhybridisation data (see 3.1.4) where it became evident that to some extent there is splitting of clones of the same transcript into multiple clusters. However, the bfg_dr_am_1.8db.sop was used for selecting clones for tag sequencing since it was the best clustering available at the time.

In the attempt to co-cluster both BFL26 and BFLG libraries, the clustering within each filter set was repeated followed by clustering of the consensus fingerprints across both sets and finally across both of the libraries. It was then realised that the bflg.all.res (dr) which resulted from this attempt is by far more accurate.

The bflg.all.res (dr) clustering predicts 1,783 clusters, containing 28,248 clones and 7,234 singletons. The number of singletons is significantly lower than the previous clustering. The
extent of accuracy of this clustering was estimated using the backhybridisation data (see 3.1.4. for explanation).

The cluster size distribution for the two clusterings of this library is given in Fig. 3.5.

### Fig. 3.5

Difference in the cluster size distribution between the bfg_dr_am_1.8db and bflg.all.res clusterings of the gastrula stage cDNA library (BFLG). The number of small clusters in the bflg.all.res clustering is decreased. Small clusters are merged with larger size clusters. Moreover, the number of singletons is significantly reduced. The column x gives the number of clones within each cluster, the column y1 the number of clusters for each cluster size in the bfg_dr_am_1.8db and the y2 column the number of clusters for each cluster size in the bfg.all. The singletons are not included in the above diagrams.

#### 3.1.3.3. Recognising the same transcripts in both libraries

Clustering of the cDNA clones from both libraries resulted in 4,994 clusters ranging in size between 4,480 and 2 members while 16,097 clones remained as singletons. The largest
cluster of 4,480 members contained *mitochondrial 16S rRNA*, the majority of the clones derived from the BFLG cDNA library.

Fig. 3.6. shows a schematic representation of the *calmodulin* fingerprint clusters from both libraries. The similarity of the hybridisation patterns between the two groups of clones is visible.

The clustering method applied for co-clustering was as described in 3.1.3.2. As a result of this approach, two new clusterings were created for both the BFL26 and BFLG libraries, the bfl26.all.res (dr) and bflg.all.res (dr). It is important to note that the estimation of the transcript number for the BFL26 cDNA library was approximately the same as in the previous clustering, almost 10,000 transcripts. However, the new BFLG clustering predicted that approximately 9,000 transcripts are present which is close to half of the number that was estimated before (see Table 3.2). The backhybridisation data verified that the bflg.all.res was more accurate. The reason for this discrepancy will be given below. This will also indicate one of the parameters that are important when fingerprinting a cDNA library.

### 3.1.4. Methods for evaluation of clustering accuracy

Three methods were used for assessing the reliability of the multiple clusterings:

1. **Hybridisation of clones selected from large and small size clusters on the whole library** and assessing the distribution of the stronger hybridisation positives within the clusters. Positive clones should belong to the same cluster as the clone used as probe.

   The neurula stage library (BFL26) was the first library to be fingerprinted and it was clustered using the double ranking normalisation method. In order to evaluate this clustering representative clones from variable size clusters were tag sequenced. These clones were subsequently used as control probes for assessing the clustering on all the libraries.

   40 different clones representing genes expected to be expressed in high levels such as *tropomyosin*, *elongation factor 1* and 2, *tubulin a* and *β chain*, *calmodulin*, *actin*, *clathrin*, *cytochrome c*, *creatine kinase*, to low level or singletons e.g. *selenium binding liver protein*, *immunoglobulin receptor*, *ferritin H-3*, *antioxidant enzyme* were hybridised. The hybridisation image of the BFL26_55P21 (*b-catenin*) clone is shown in Fig. 3.7.

   All the above hybridisations were performed on bacterial colony filters in order to avoid using PCR filters which are more difficult to produce. Each amphioxus cDNA library is spotted on three filters, in a 4x4 duplicate pattern. The hybridisation images were scored manually using an inhouse developed program (*Winclone*, Marcus Kitzman, Max-Planck...
Fig. 3.6. Schematic representation of the calmodulin clusters in both amphioxus libraries.
Institute for Molecular Genetics, unpublished) which allows the classification of the positives in three levels of intensity.

The distribution of only the strongest positives (level of intensity 1) within the clusters was assessed using a perl script (which_clusters3.prl). An example of the output of the script for assessing the distribution of the positive clones from the hybridisation the BFL26_102K19 (40S ribosomal protein S15) in the bfl26_1.res, bfg_dr_am_1.8db.sop and bfg26_co.res (dr) clustering is given below.

The first part of the script output lists the coordinates of the positives along with the number of the cluster (in the specified clustering) they belong while the second part lists the number of positives within each of the clusters as well as the cluster size (see below).

Thus 6 out of 29 40S ribosomal protein S15 positives are found in the same cluster 475 of size 10 while 13 were split in multiple clusters (bfl26_1.res clustering). It has been already mentioned above that the extra 10 (19 out of the total of 29) positives that are not found in the clusters are clones that failed to give a PCR product and thus rejected from the clustering analysis. In the bfl26.all.res clustering, the majority of the above positives (16 out of 23) are assigned in cluster 203 with size 18. This is a case where the bfl26.all.res is significantly better than bfl26_1.res

It is worth stressing that:

- the clusters include positives spotted on different filter set which implies that experimental parameters that differ between sets such as PCR product yield (since the first and the second part of the library were PCR amplified separately) do not affect clustering.
- the bfl26.all.res (dr) clustering is marginally better than the bfl26_1.res (dr). In contrast, the bfg.all.res (dr) clustering is considerably better than the bfg_dr_am_1.8db.sop since less positive clones are scattered in multiple clusters.

bfl26_1.res (dr)

```plaintext
*** /home/panopoul/oligohyb/backhyb/bfl26dgt/102k19bfl26.dgt ***

mapping clones of level 1(intensity) or stronger!


Sets 1 2 3 4

<table>
<thead>
<tr>
<th>Clones in Cluster</th>
<th>Size</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>475</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>532</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>586</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>1159</td>
<td>5</td>
<td>2</td>
</tr>
</tbody>
</table>
```

88
1 clones in cluster 1263 size 5
1 clones in cluster 2455 size 3
2 clones in cluster 3104 size 2
1 clones in cluster 4407 size 2
1 clones in cluster 4586 size 2
3 clones in cluster 4675 size 6028

Total 6 13 0 0

19/29 clones are split into 10 clusters

on bfl26.all.res

*** /project/a2z/control_dgt/102k19bf26.dgt ***

mapping clones of level 1 or stronger!

ICRFbfl26_26G18: 203; ICRFbfl26_46F20: 203; ICRFbfl26_20L5: 203; ICRFbfl26_6N11: -;
ICRFbfl26_23C11: 1862; ICRFbfl26_26F17: 3697; ICRFbfl26_44K20: 203; ICRFbfl26_16D10: 2723;
ICRFbfl26_20F13: 203; ICRFbfl26_65I20: -; ICRFbfl26_84I8: -; ICRFbfl26_80M7: -;
ICRFbfl26_60N23: 203; ICRFbfl26_94I12: 203; ICRFbfl26_76O21: 3135; ICRFbfl26_92J24: 3697;
ICRFbfl26_106I18: -;
ICRFbfl26_99L11: 203; ICRFbfl26_111C8: -; ICRFbfl26_133H6: -; ICRFbfl26_127C11: 203;
ICRFbfl26_129A2: 203; ICRFbfl26_131F1: 203; ICRFbfl26_101B14: 3697; ICRFbfl26_118N6: 203;
ICRFbfl26_103E22: 203; ICRFbfl26_134F17: -; ICRFbfl26_110K24: 1627; ICRFbfl26_133N16: -;
ICRFbfl26_103P14: -;

Sets 1 2 3 4

16 clones in cluster 203 size 18
1 clones in cluster 1627 size 3
1 clones in cluster 1862 size 2
1 clones in cluster 2723 size 2
1 clones in cluster 3135 size 2
3 clones in cluster 3697 size 9806

Total 0 0 0 0

23/34 clones are split into 6 clusters

bfg_dr_am_l.8db.sop

*** /home/panopoul/oligohyb/backhyb/bflgdgt/102k19bflg.dgt ***

mapping clones of level 1 or stronger!

95L19: 3177; 57I5: 192; 60E1: 2644; 54E13: 192; 54M21: 192; 60O10: 192; 78M3: 1147;
56I18: 192; 9D06: 3177; 80D1: 3177; 35P13: 192; 76H12: 192; 70E15: 3177; 71B8: 192;
61D23: 192; 89F3: 192; 75G5: 192; 10I21: 45; 99F8: 3177; 135N10: -; 128D4: 3177;
113P23: 192; 99B2: 192;

Sets 1 2 3 4
Fig. 3.7. Hybridisation of the amphioxus 55P21 (β-catenin) clone isolated from the neurula (BFL26) stage on bacterial colony filters from both neurula and gastrula (BFLG) cDNA libraries. The total number of positives on each library was 5. The filters after hybridisation were exposed on phosphorimager screens and scored manually using an inhouse developed software (Winclone, M. Kitzman, unpublished). The bacterial colony filters were spotted in a 4x4 spotting pattern (as shown in the square between the two filters). Each filter is carrying 18,432 distinct clones spotted in duplicate.
1 clones in cluster 45 size 63 0 1 0 0
23 clones in cluster 192 size 23 14 9 0 0
1 clones in cluster 417 size 11 0 1 0 0
1 clones in cluster 910 size 5 0 1 0 0
1 clones in cluster 1147 size 4 0 1 0 0
2 clones in cluster 2644 size 2 2 0 0 0
5 clones in cluster 3177 size 11910 0 5 0 0

Total 16 18 0 0

34/40 clones are split into 7 clusters

**on bflg.all.res**

*** /project/a22z/control_dgt/102k19bflg.dgt ***

mapping clones of level 1 or stronger!


Sets 1 2 3 4

1 clones in cluster 11 size 445 0 0 0 0
32 clones in cluster 126 size 34 0 0 0 0
2 clones in cluster 1784 size 7234 0 0 0 0

Total 0 0 0 0

35/40 clones are split into 3 clusters

**bflg26_co.res (dr)**

*** /project/a22z/control_dgt/102k19bfl26.dgt ***

mapping clones of level 1 or stronger!

ICRFbfl26_26G18: 152; ICRFbfl26_46J20: 152; ICRFbfl26_20L5: 152; ICRFbfl26_6N11: -;
ICRFbfl26_23C11: 4054; ICRFbfl26_26117: 4995; ICRFbfl26_44K20: 152; ICRFbfl26_16D10: 4711;
ICRFbfl26_20F13: 152; ICRFbfl26_65I20: -; ICRFbfl26_84I8: -; ICRFbfl26_80M7: -;
ICRFbfl26_60N23: 152; ICRFbfl26_94I2: 152; ICRFbfl26_76O21: 355; ICRFbfl26_92I24: 4995;
ICRFbfl26_129O23: 152; ICRFbfl26_99L11: 152; ICRFbfl26_111C8: -; ICRFbfl26_133H6: -;

Sets 1 2 3 4

16 clones in cluster 152 size 52 0 0 0 0
1 clones in cluster 355 size 20 0 0 0 0
1 clones in cluster 2172 size 3 0 0 0 0
1 clones in cluster 4054 size 2 0 0 0 0
1 clones in cluster 4711 size 2 0 0 0 0
3 clones in cluster 4995 size 16097 0 0 0 0

Total 0 0 0 0

23/34 clones are split into 6 clusters

*** /project/a2z/control_dgt/102k19bflg.dgt ***

mapping clones of level 1 or stronger!


Sets 1 2 3 4

1 clones in cluster 16 size 465 0 0 0 0
32 clones in cluster 152 size 52 0 0 0 0
2 clones in cluster 4995 size 16097 0 0 0 0

Total 0 0 0 0

35/40 clones are split into 3 clusters

Table 3.3. gives an overview of the distribution of all the 40 BFL26 clones selected to be hybridised on both libraries, their counted expression level according to the fingerprint cluster size (FP) and the number of positives that were accurately assigned to the correct cluster (PC) according to the bfl26_.l.res clustering.

According to Table 3.3., in the majority of hybridisations the number of positive clones is greater than the size of the cluster they belong. This is due to the fact that only 35,482 clones were considered for the clustering while the rest have been rejected due to failure to amplify or due to the inconsistent correlation of duplicates. Meanwhile, as it was noted above all the hybridisations were performed on bacterial colony filters where more clones are available for hybridisation since very few clones fail to grow.

b. sequencing of multiple members of a cluster

5' tag sequencing of 13 out of 68 members of the BFL26 calmodulin cluster proved that all clones were calmodulin with different 5' starting point (Fig. 3.8.). Moreover, the insert size of all the clones from the calmodulin clusters from both libraries was determined on agarose gel (Fig. 3.9.). Clones that were ranked within the cluster as being closer to the consensus fingerprint were shorter while the more distant are longer clones.
c. Hybridisation of a clone from a specific cluster on southern blot prepared from all the clones assigned to this cluster

The calmodulin cluster in the neurula stage library consists of 68 clones (bfl26_.l.res). Calmodulin is also expressed during the gastrula stage at lower level as estimated by the 28

Table 3.3. Identities of clones used as controls for checking the reliability of fingerprinting clustering. Transcripts expressed at low and high level (as estimated from the fingerprint cluster size they belong) were 5’tag sequenced and hybridised back on both amphioxus libraries. The distribution of the resulting positives within the clusters according to the bfl26_.l.res is given. Basic Blast Search version 2.0.6 that allows the introduction of gaps, was used for matching the sequences (Blast results were updated on 10.2.99). Abbreviations used: F.C.: fingerprint cluster, P.C.: number of positives assigned in the F.P cluster in which the clone that was used for hybridisation belonged. Only the positives of higher intensity were considered. Positives: total number of positives that were considered during clustering analysis. The 5’tag sequences of all the above clones are deposited in dbEST.

<table>
<thead>
<tr>
<th>Identity as implied from the top BLAST match</th>
<th>P-value</th>
<th>BFL26 F.C. size</th>
<th>BFL26 P.C. size</th>
<th>BFL26 Positives</th>
<th>Probe coordinates</th>
</tr>
</thead>
<tbody>
<tr>
<td>AF061443 G protein-coupled receptor LGR4</td>
<td>0.002</td>
<td>4</td>
<td>2</td>
<td>7</td>
<td>BFL26.102I13</td>
</tr>
<tr>
<td>P07181 calmodulin</td>
<td>5e-70</td>
<td>68</td>
<td>30</td>
<td>36</td>
<td>BFL26.100B15</td>
</tr>
<tr>
<td>P31674 40S ribosomal protein S15</td>
<td>3e-35</td>
<td>9</td>
<td>1</td>
<td>19</td>
<td>BFL26.102K19</td>
</tr>
<tr>
<td>AP001434 Hpast [Homo sapiens]</td>
<td>2e-19</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>BFL26.104C24</td>
</tr>
<tr>
<td>P20029 glucose regulated protein precursor</td>
<td>9e-51</td>
<td>5</td>
<td>1</td>
<td>3</td>
<td>BFL26.108B3</td>
</tr>
<tr>
<td>B57407 3alpha-hydroxysteroid dehydrogenase</td>
<td>5e-04</td>
<td>2</td>
<td>2</td>
<td>2correct</td>
<td>BFL26.14B4</td>
</tr>
<tr>
<td>A33851 alcohol dehydrogenase (NADP+) (EC 1.1.1.2)</td>
<td>1e-07</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>BFL26.15H9</td>
</tr>
<tr>
<td>O10880 5’ antioxidant enzyme AOE372</td>
<td>1e-21</td>
<td>4</td>
<td>1</td>
<td>17</td>
<td>BFL26.151I7</td>
</tr>
<tr>
<td>D86627 ferritin H-3</td>
<td>8e-07</td>
<td>1</td>
<td>1</td>
<td>28</td>
<td>BFL26.151I21</td>
</tr>
<tr>
<td>W911125 Soares mouse embryo NbME13.5 14.5</td>
<td>9.6e-22</td>
<td>7</td>
<td>1</td>
<td>3</td>
<td>BFL26.14B9</td>
</tr>
<tr>
<td>AJ005073 Alox [Mus musculus]</td>
<td>1e-17</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>BFL26.14J23</td>
</tr>
<tr>
<td>AF043415 peroxidoxin-2 [Onchocerca volvulus]</td>
<td>7e-07</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>BFL26.15B3</td>
</tr>
<tr>
<td>Q92005 translational elongation factor-1 alpha</td>
<td>8.4e-137</td>
<td>9</td>
<td>-</td>
<td>20</td>
<td>BFL26.17J6</td>
</tr>
<tr>
<td>L29259 elongation factor SHI1 p15 subunit</td>
<td>1e-56</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>BFL26.17L10</td>
</tr>
<tr>
<td>Z79639 similar to zinc metalloprotease</td>
<td>1e-21</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>BFL26.17N2</td>
</tr>
<tr>
<td>P47740 fatty aldehyde dehydrogenase</td>
<td>2e-15</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A099923 preprocathepsin P</td>
<td>1e-63</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>BFL26.17B8</td>
</tr>
<tr>
<td>P25121 60S ribosomal protein L11</td>
<td>5e-08</td>
<td>6</td>
<td>1</td>
<td>28</td>
<td>BFL26.17L16</td>
</tr>
<tr>
<td>AB013606 elongation factor 1 alpha</td>
<td>5e-88</td>
<td>10</td>
<td>1</td>
<td>131</td>
<td>BFL26.17J24</td>
</tr>
<tr>
<td>AF026402 US srRNP 100 kD protein</td>
<td>e-105</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>BFL26.20J4</td>
</tr>
<tr>
<td>U51472 superfast myosin heavy chain</td>
<td>4e-58</td>
<td>185</td>
<td>61</td>
<td>94</td>
<td>BFL26.23H3</td>
</tr>
<tr>
<td>P04866 sparc precursor (osteocencitin)</td>
<td>3e-89</td>
<td>66</td>
<td>40</td>
<td>64</td>
<td>BFL26.23O7</td>
</tr>
<tr>
<td>X06748 beta-tubulin</td>
<td>2.9e-11</td>
<td>89</td>
<td>75</td>
<td>117</td>
<td>BFL26.24A19</td>
</tr>
<tr>
<td>P19889 60S acidic ribosomal protein PO</td>
<td>1e-78</td>
<td>39</td>
<td>18</td>
<td>27</td>
<td>BFL26.36B8</td>
</tr>
<tr>
<td>U09782 myosin heavy chain</td>
<td>3e-51</td>
<td>185</td>
<td>127</td>
<td>163</td>
<td>BFL26.37E16</td>
</tr>
<tr>
<td>AB013606 elongation factor 1 alpha</td>
<td>e-112</td>
<td>314</td>
<td>196</td>
<td>282</td>
<td>BFL26.37E3</td>
</tr>
<tr>
<td>Y16474 cytochrome c oxidase subunit I</td>
<td>1e-72</td>
<td>815</td>
<td>71</td>
<td>277</td>
<td>BFL26.37G17</td>
</tr>
<tr>
<td>O18500 ACT2 SACKO ACTIN 2</td>
<td>e-113</td>
<td>125</td>
<td>55</td>
<td>204</td>
<td>BFL26.3L15</td>
</tr>
<tr>
<td>AF025330 receptor for activated protein kinase C</td>
<td>e-117</td>
<td>84</td>
<td>48</td>
<td>67</td>
<td>BFL26.48O20</td>
</tr>
<tr>
<td>AF055288 creatine kinase M1-CK</td>
<td>5e-84</td>
<td>190</td>
<td>50</td>
<td>73</td>
<td>BFL26.50E20</td>
</tr>
<tr>
<td>L22660 Rat polymeric immunoglobulin receptor</td>
<td>4.3e-12</td>
<td>121</td>
<td>25</td>
<td>53</td>
<td>BFL26.54B24</td>
</tr>
<tr>
<td>A26488 tubulin alpha-1 chain - fruit fly</td>
<td>3e-99</td>
<td>316</td>
<td>123</td>
<td>175</td>
<td>BFL26.66D19</td>
</tr>
</tbody>
</table>
Fig. 3.8. Alignment of calmodulin clones
Fig. 3.9. PCR products of the clones assigned to the calmodulin cluster from the amphioxus gastrula stage (A) and from the neurula stage (B) libraries. From the left to the right side of each row on the gel: The clones are loaded in a descending order of closeness to the consensus fingerprint of the cluster. Clones that are closer to the consensus fingerprint are shorter. All clones of the calmodulin clusters from both libraries are found positive in the hybridisation of their southern blot with the clone BFL26_49P2, which was verified to be calmodulin by 5' tag sequencing.
member cluster (bfg_dr_am_1.8db.sop). The clone BFL26_49P2 was hybridised on the southern blot prepared from the calmodulin clusters from both libraries (Fig. 3.10.) All clones of both clusters are found positive with the above clone indicating that they represent the same gene.

d. Comparison between sequence based clusters and oligonucleotide fingerprinting clusters

Clustering accuracy was also assessed by 5' tag sequencing of 3,878 clones selected from clusters from both libraries and their subsequent clustering using the PHRAP sequence assembly program (see Chapter 4).

In overall, clustering of 2,365 (of which 2,197 represent sequences of different clones) 5' ESTs from the BFLG cDNA library, indicated that 1,811 sequences did not overlap with any other sequence while 617 were grouped in 171 clusters. The majority of the clusters were 2 member clusters while the three larger size clusters contained 61, 28, 19 clones. These clusters contain mitochondrial 16S rRNA, cytochrome c oxidase, and atp synthase a chain respectively. Thus, out of 2,197 ESTs generated from clones selected from different fingerprinting clusters, 1,811 (82.4%) sequences were unique.

Similarly, out of 1,714 (of which 1,673 represent sequences of different clones) BFL26 5' ESTs, 1,516 sequences remained singletons while 264 were grouped in 92 clusters (88.4% unique sequences). The majority of the clusters (66 clusters) contained 2 members while the three larger size clusters contained 13 and 12 members. Two clusters of size 8 and 9 were muscle actin and mitochondrial 16S rRNA.

Clustering of the above 5' tag sequences from both libraries (4,079 sequences of which 3,870 represent different clones) resulted in 2,822 clones being unique while 758 clones were grouped in 380 clusters the majority of them containing between 2 and 9 members (86.8% unique sequences). The four larger size clusters of 10, 14, 21 and 55 clones were NADH dehydrogenase (consists only of BFLG clones), cytochrome c, atp synthase a chain and mitochondrial 16S rRNA respectively.

The alignment of all clusters can be viewed using the Consed program (Gordon et al., 1998) where the consensus sequence of the cluster is displayed (Fig. 3.11).

The sequence clustering results were inserted in a table (Fig. 3.10.) where the sequence clusters are listed along with their size and the clones from which the sequences were generated. Furthermore, the number and the size of the fingerprint cluster each clone belongs is shown. This enables a visual assessment of the agreement between sequence and fingerprint clusters.
The above clustering results show that there are cases where clones with overlapping 5' tag sequence are distributed in different fingerprinting clusters. Fig. 3.10.B gives an example of 10 neurula stage clones having higher score with amphioxus muscle actin that are split in 5 singletons and 5 different size fingerprinting clusters (bfl26_.l.res). The sequence alignment of the 5' tag sequences for those clones is given in Fig. 3.11. According to this alignment, clones that differ by 400 bp at their 5' site such as the first 3 clones in the alignment and the last two clones, belong in different clusters. The extra 400 bases result in a different oligonucleotide fingerprint for these two groups of clones.

Another example of clone splitting in multiple fingerprinting clusters due to their different length is described below.

Two clones having higher similarity score with the amphioxus Otx (Williams and Holland, 1998) have been identified in the neurula stage library, the MPIMBFL26_99E12 and the MPIMBFL26_27L8 as well as the MPIMBFLG_6P6 and MPIMG_58I15 in the gastrula stage library. None of the above clones is clustered together. However, it is realised by looking through the BLAST result for the 5' tag sequence of all clones that they differ considerably in their length.

The similarity region between BFLG_58I15 and the deposited in the Genbank amphioxus Otx, AF043740 (Williams and Holland, 1998) is at 780nt-1442 nt.

The similarity region for the BFLG_6P06 is at 1415-1848 nt. Both of the above clones give high similarity score only at the nucleotide level with only the amphioxus Otx which is an indication that they cover the variable 3'UTR region.

The BFL26_27L08 clone gives high similarity score at the protein level at 72aa until 135 of mouse Otx while the BFL26_99E12 with amphioxus Otx between 1015-1531nt.

The percentage of clones splitting into multiple clusters due to their varying length is higher in the case of the gastrula stage library for which it is already shown that the average insert length is approximately 0.8-1 kb.

The reason for the splitting of clones that differ in length in multiple fingerprint clusters is that the length of clone affects the number of oligonucleotide hits. Clones that are very short have limited number of hits which are not enough to recognise similarities.

To verify this, the number of oligonucleotides (from the set used in the hybridisations) having partial or full match to the amphioxus Otx sequence deposited in the Genbank was compared to the number of oligonucleotides found positive in the Otx clones identified in the BFL26 and BFLG libraries.

The Otx sequence deposited in the Genbank contains 104 partial matches and 38 full
<table>
<thead>
<tr>
<th>Cluster #</th>
<th>Library</th>
<th>Clone</th>
<th>BLAST result</th>
<th>E-value</th>
<th>Gen. no.</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>BFL26</td>
<td>10C74</td>
<td>blastx</td>
<td>4.5e-40</td>
<td>P14627</td>
<td>Homo sapiens (human), asterias armoricanus (cat), and naso scutatus (pig). 60S ribosomal protein. Length = 125</td>
</tr>
<tr>
<td>1</td>
<td>BFL26</td>
<td>10A11</td>
<td>blastx</td>
<td>3.6e-20</td>
<td>P51154</td>
<td>Gallus gallus (chicken). Vigilin. Length = 1276 (1 mark) (total: 1). Length = 649</td>
</tr>
<tr>
<td>1</td>
<td>BFL26</td>
<td>10A12</td>
<td>blastx</td>
<td>3.0e-16</td>
<td>Q76576</td>
<td>Kambalophis elegans. Id0764.5 protein. Length = 326</td>
</tr>
<tr>
<td>2</td>
<td>BFL26</td>
<td>10A24</td>
<td>blastx</td>
<td>3.6e-20</td>
<td>Q78899</td>
<td>Gallus gallus (chicken). Porphobilinogen-a-synthase. (ec 5.4.1.23). Length = 199</td>
</tr>
<tr>
<td>3</td>
<td>BFL26</td>
<td>10A35</td>
<td>blastx</td>
<td>3.5e-19</td>
<td>Q51905</td>
<td>Gallus gallus (chicken). Tropoelin. Length = 359 (5.1.15). (class: pi). Length = 269</td>
</tr>
<tr>
<td>4</td>
<td>BFL26</td>
<td>10B14</td>
<td>blastx</td>
<td>1.6e-06</td>
<td>Q02188</td>
<td>Ostreola vulgaris. odt80. 11.98 Length = 60 3 1.36 (copper pump 1). 7.98 Length = 1401.5</td>
</tr>
<tr>
<td>5</td>
<td>BFL26</td>
<td>10B21</td>
<td>blastx</td>
<td>7.5e-14</td>
<td>F23067</td>
<td>Gladiolus sp. ST265. (class: 1.232). hypothetical 53.9 kda protein (hyp). Length = 297 630 (number) (unrelated). Length = 157</td>
</tr>
<tr>
<td>6</td>
<td>BFL26</td>
<td>10D4</td>
<td>blastx</td>
<td>7.3e-14</td>
<td>Q51905</td>
<td>Gallus gallus (chicken). Tropoelin. Length = 359 (5.1.15). (class: pi). Length = 269</td>
</tr>
<tr>
<td>7</td>
<td>BFL26</td>
<td>10E1</td>
<td>blastx</td>
<td>3.5e-14</td>
<td>D138899</td>
<td>Asterias armoricanus (cat), neurokinin 3. Length = 840 Length = 85 67 415 114 (envelope protein II), major core Length = 737</td>
</tr>
</tbody>
</table>

Fig. 3.10. A. Clustering of 4,000 5' tag sequences of clones from both amphioxus libraries and their distribution in the fingerprint clusters is linked on a web page. The highest matches against the swissprot using the tblastx program are also given along with the link to the Genbank report for each match.
Fig. 3.10. B. The web page linking the clusters of the amphioxus 5’ tag sequences with the fingerprinting clusters. The numbers beside the brackets in the first and second column refer to the cluster number of the sequence or the fingerprinting cluster respectively. This was implemented in order to show the extent of consistency between the sequence and fingerprinting clustering. There are cases where the clones having overlapping sequences are assigned into different fingerprint cluster e.g. muscle actin.
Fig. 3.11. A. Alignment of the 5' tag sequences of BFL26 muscle actin that belong to the sequence cluster 264 (see fig.3.10.B). Clones have been assigned to different fingerprint clusters although they belong to the same fingerprint cluster. This is due to differences in the insert size between clones.
Fig. 3.11. B. Continue of the previous alignment of the muscle actin clones.
matches while MPIbflg_58I15 and ICRFbfl26_99E12 have 37 and 14 oligonucleotide hits respectively.

clone gi2828715gblAF043740IAF043740 total hits = 142

partial matches 104: o069 o078 o080 o081 o084 o085 o086 o087 o089 o090 o091 o092 o094 o099 o100 o103 o104 o108 o112 o113 o114 o117 o118 o119 o120 o121 o122 o125 o126 o127 o129 o131 o136 o137 o138 o140 o144 o145 o146 o148 o149 o150 o154 o160 o161 o166 o168 o177 o184 o185 o186 o188 o190 o193 o194 o199 o203 o207 o209 o212 o216 o217 o222 o223 o224 o227 o230 o233 o236 o239 o240 o247 o248 o255 o259 o263 o266 o269 o270 o271 o277 o280 o282 o284 o285 o287 o289 o291 o292 o294 o295 o296 o297 o300 o304 o305 o307 o308 o311 o312 o316 o317 o318 o321 o324 o326 o327 o329

full matches 38: o076 o088 o095 o097 o102 o105 o109 o128 o141 o143 o151 o153 o156 o164 o165 o172 o174 o182 o192 o204 o215 o220 o221 o237 o246 o250 o260 o272 o274 o275 o276 o278 o279 o281 o288 o303 o314 o318

clone MPIbflg_58I15:
o102 o103 o109 o112 o117 o120 o122 o123 o132 o135 o141 o150 o156 o174 o182 o194 o196 o202 o213 o214 o215 o217 o224 o237 o246 o247 o265 o266 o267 o268 o76 o80 o83 o84 o93 o95 o97 total hits = 37

cache ICRFbfl26_99E12:
o109 o120 o135 o139 o142 o146 o148 o163 o173 o174 o215 o232 o254 o256 total hits = 14

The above conclusion that clones representing the same transcript are split in multiple fingerprinting clusters because the extent of their overlap is not enough to cluster them together was verified by PCR amplification of 100 clones selected from the BFLG library that had the lower number of oligonucleotides hits. Thus, 32% of the 100 clones were 500bp, 49% were 750bp, 13% were between 750bp and 1000bp and 6% failed to give PCR product (data not shown).

3.2. Theoretical oligonucleotide fingerprinting

As it was mentioned in 3.2.1.2, the theoretical fingerprints for all the 168 amphioxus full length sequences which are currently deposited in the Genbank were calculated with the set of oligonucleotides used in the hybridisations. Oligonucleotides that gave a full 8mer match against a specific sequence were assigned a score of 1 while the ones with 7mer match were assigned a score of 0.7 and those with no match the value of 0. Using an algorithm based on that used in the BLAST program (Altschul et al., 1990), the consensus experimental fingerprints of all clusters in the cDNA libraries were compared to the theoretical fingerprints. It was observed that the identity of a cluster was predicted accurately when the P value was lower than 10⁻⁴.
For example by such similarity search it was predicted that the following BFL26 cluster contains transcripts of the BMP2/4 gene:

*** cluster 1808 size 4 ***

# 1808 4 0.444477
1 12583 0.539695 ICRFbfll26_1P13
2 16120 0.465279 ICRFbfll26_30B5
3 11498 0.408581 ICRFbfll26_17B23
4 13198 0.364352 ICRFbfll26_22A10

1 5.994e-07 19 gil4039084|gil|AF068750|AF068750 Branchiostoma floridae bone morphogenetic protein 2/4 (AmphiBMP2/4) mRNA, complete cds

Indeed the amphioxus BMP2/4 clone that is deposited in the database was isolated from the BFL26 library (Panopoulou et al., 1998) and it is the clone MPIMGBFL26_1P13 that belongs to this cluster.

Additional cases where the identity of the experimental cluster was accurately predicted based on the similarity score against the theoretical fingerprints are given in Appendix V. In all the listed cases it was possible to confirm that the prediction was accurate since one clone from this cluster had been already 5' tag sequenced.

The above possibility to be able to compare accurately theoretical with experimental fingerprints could be also used for adjusting the level of stringency when selecting the most reliable clustering. If a set of control clones that have been fully sequenced is spotted on all filters along with the unknown clones, comparison between the experimental and theoretical fingerprints of the control clones will point out which of the oligonucleotides should be there and which ones are false positives. This approach could be formulated using a trained neural net.

3.3. Conclusions

The generation of single pass partial sequences of randomly selected cDNA clones (ESTs) has become the main method of gene identification in the last few years. The major advantages of ESTs are their relative low cost in comparison to genomic sequencing, their high-throughput nature and their inherent feature to reflect the range of genes that are transcribed in the source cDNA library.

However, EST sequencing is highly redundant. Some genes are expressed at very high level while others very low which is reflected in the resulting cDNA libraries. This results in
some transcripts being resequenced many times. Moreover, the probability of identifying genes expressed at low level does not increase with the number of clones sequenced.

The techniques of normalisation and subtraction (Bonaldo et al., 1996) that have been proposed for reducing such redundancy, have been adopted in the construction of all libraries that are currently sequenced by the I.M.A.G.E consortium (87% of the all human ESTs are derived from these normalised libraries). Disadvantages of these normalisation and subtraction procedures include the prevalence of short 3' truncated clones or failure to normalise some transcript classes e.g. those that represent 10-20% of the total mRNA (defined as superprevalent class in (Davidson and Britten, 1979) such as mitochondrial 16S rRNA or those containing a specific sequence pattern like ALU repeat-containing cDNAs (Bonaldo et al., 1996).

Finally, the major disadvantage of the above methods is that after normalisation the information on the abundance of each transcript is lost.

An alternative approach for reducing redundancy is the use of oligonucleotide fingerprinting (Drmanac et al., 1991; Drmanac et al., 1996; Meier-Ewert et al., 1998; Meier-Ewert et al., 1993; Meier-Ewert et al., 1994 Radelof, 1998 #216; Milosavljevic et al., 1996). Oligonucleotide fingerprinting is based on the hybridisation of short oligonucleotides of known sequence onto an array of cDNA clones, thus creating a sequence based fingerprint for each clone which represents the degree of hybridisation of the different oligonucleotides to each cDNA. Clones that share the same oligonucleotide hybridisation pattern are grouped together in clusters as representing highly overlapping sequences.

During this study oligonucleotide fingerprinting was applied on approximately 110,000 cDNA clones picked from two amphioxus embryonic cDNA libraries. Multiple clusterings differing on the level of stringency in the assignment of clones in clusters were attempted (Table 3.2.). The most accurate clusterings predicted 1,783 clusters (containing 28,248 clones) and 7,234 singletons for the gastrula library and 4,674 clusters (containing 29,454 clones) and 6,028 singletons for the neurula stage library.

The number of singletons in both libraries which represent the low abundance transcripts comprise 20.38% (7,234/35,482) and 16.9% (6,028/35,482) of all the fingerprinted clones in the gastrula and neurula stage libraries respectively. Therefore, identification of those transcripts by random EST sequencing would result in the re-sequencing of the more abundant transcripts before those are identified.

Finally, co-clustering of both libraries resulted in 4,994 clusters and 16,097 singletons.

4,079 5' tag sequences from 3,870 clones selected from various size clusters from both libraries are generated so far. Based on the sequencing results mitochondrial 16S rRNA
(8.20%), muscle actin (1.77%), tubulin alpha-1 chain (0.89%), elongation factor 1-a (0.88%)
creatine kinase (0.53%) were among the most abundant transcripts in the neurula stage.
Similarly, mitochondrial 16S rRNA (8.20%), muscle actin (1.78%), tubulin a-1 chain
(0.88%), nadh dehydrogenase subunit 2 (0.81%) comprised the first four larger clusters at
the gastrula stage library.

To assess the extent of redundancy that might still occur in the selection of clones based
on the fingerprinting clustering, the 5' tag sequences were assembled using the PHRAP
program. Thus, sequence clustering of all tag sequences resulted in 2,852 (77.3% ) different
transcripts. Finally, if the sequences from each library are clustered separately, 82.4%
(1811/2197) of the sequenced BFLG clones and 90.6% (1516/1673) of the BFL26 clones
represent different transcripts.

The above percentage of non-redundant sequences when the clones are preselected by
oligonucleotide fingerprinting clustering is significantly higher than the random EST
sequencing from biochemically normalised libraries.

Briefly, the fingerprinting data generated during this study confirm the advantages of
oligonucleotid e fingerprinting. Thus:

• Clones sharing common sequence patterns can be reliably clustered. As a result, transcripts
of the higher abundance class can be recognised and clustered resulting in a massive
reduction in sequencing redundancy and therefore costs while at the same time increasing the
probability of identifying low abundance transcripts. Out of 50,688 fingerprinted clones in
each of the libraries, 8,968 and 10,702 different transcripts were recognised in the gastrula
and the neurula stage library respectively. This corresponds to an average 3.9 fold decrease
(35,482/8,968) in the overall redundancy in the gastrula stage library and 3.3 (35,482/10,702)
in the neurula stage library. Moreover, it was demonstrated that the longer clones of a cluster
can be recognised and preferably selected for sequencing. Finally it is possible to extract the
oligonucleotides that are positive for the majority of the clones in the cluster. These
oligonucleotides could be used for internal sequencing of any clone of the cluster thus
reducing the amount of time required for full length sequencing of a clone. It has been
reported that tailed octamer primers can be used in cycle sequencing (Ball et al., 1998).

• Clones representing members of gene subfamilies can be distinguished and assigned into
different clusters e.g. intermediate filament proteins A1, A2, A3 and B1, B2, B3 (Table 5.1).
• Clone fingerprints generated for one tissue library can be compared reliably across multiple tissue libraries of the same organism. This results in even more decrease in sequencing redundancy and therefore cost as only differentially expressed transcripts can be selected for sequencing.

• The expression level of a specific transcript as implied by the cluster size it belongs, is a reliable estimate of the expression level of the respective transcript in this library. As an example transcripts that are expected to be expressed at high level belong in large size clusters (Table 5.1. and 5.2.).

• The identity of clusters can be accurately predicted without sequencing by comparing the experimental consensus fingerprints with the theoretical fingerprints of sequences from the same organism. For example, it was possible to assign the identity of the BMP2/4 cluster. However, this comparison is reliable in the case that the theoretical fingerprints are created from high quality sequences. At the moment there are very few full length amphioxus sequences in the Genbank which limits the extent of such comparison.

Moreover, the parameters that affect the accuracy of the oligonucleotide fingerprinting method were pointed out with specific examples:

• the number and the sequence of the oligonucleotides used is critical in recognising and partitioning accurately an array of clones. The information generated is increased as the number of oligonucleotides used increases. However, the information imparted in an oligonucleotide hybridisation depends on the sequence of the oligo itself. Some oligonucleotides have higher occurrence frequency than others. This is shown through the example of the Branchiostoma floridae cholinesterase 2 (ChE2) gene retrieved from the Genbank (Table 3.1. B).

• the length of clones to be clustered plays a critical role in the accurate clustering of those clones while it also affects the capacity of assigning the identity of clusters reliably by comparing the experimental with theoretical fingerprints. As the length of the clones decreases, the number of the oligonucleotides necessary to recognise two clones with limited overlap, increases. The example of the differing length Otx clones from both libraries clearly illustrates the above.
Chapter 4

EST sequencing, analysis and data interpretation: How many amphioxus genes have been isolated?

Introduction

Random sequencing of cDNA clones as a method for gene identification was described more than a decade ago (Constanzo et al., 1983). This method has recently been used extensively in another form, the expressed sequence tags (ESTs). ESTs are 5’ or 3’ partial sequences of randomly selected clones from cDNA libraries constructed from a specific tissue or developmental stage (Adams et al., 1991; Adams et al., 1995; Okubo et al., 1992). EST sequencing has become popular because it enables faster access to genes in comparison to genomic sequencing while it is also a measure of the transcriptional activity of a specific tissue.

A large public effort to generate 5’ and 3’ EST sequences for the majority of human genes began in 1993 when a collaboration between the Integrated Molecular Analysis of Genomes and their Expression (IMAGE) consortium (Lennon et al., 1996), Merck & Co., Inc., the Genome Sequencing Center (GSC) at Washington University and the National Center for Biotechnology Information (NCBI) was formed, (Lennon et al., 1996; Williamson et al., 1995). To date more than 1,157,324 5’ and 3’ human ESTs derived from many different cDNA libraries are deposited in dbEST public database (Boguski et al., 1993). Their clustering resulted in 77,000 clusters (NCBI Unigene 26.5.99) representing more than 50% of human genes.

Similarly, 350,000 ESTs have been generated from mouse tissue specific libraries (Marra et al., 1998), 85,000 from rat (http://www.ncbi.nlm.nih.gov/UniGene/Rn.stats.shtml) and 80,000 from Drosophila (http://www.fruitfly.org/EST/) (Casey et al., 1998). These sequences are deposited in dbEST, as well as in Genbank.

In order to process the EST data generated from the large-scale EST projects, large genome centers have developed new software or modified the already existing software
which was developed for genomic sequencing programs. A major source of the sequence analysis programs used in this project were adopted from these large-scale EST and genomic projects either at Genome Sequencing Center at Washington University, U.S.A. or the Sanger Center at U.K.

Based on the oligonucleotide fingerprinting clustering results of the BFL26 (4,674 clusters of clones and 6,028 singletons (bfl26_1.res)) and BFLG (3,177 clusters and 11,187 singletons (bfg_dr_am_1.8db.sop)) libraries, one clone per cluster (the closest to the cluster consensus fingerprint) and all the singletons were selected for 5’ tag sequencing, 22,000 cDNA clones in total. 3,870 amphioxus cDNA clones selected from both libraries have been sequenced and processed so far.

The processing steps include the: i) cleaning of raw sequence data by removing low quality sequences, masking of vector or repeat sequences ii) database searches using BLAST programs iii) prediction of coding potential iv) classification of ESTs into functional groups.

4.1. Automated processing of sequence data

The majority of the clones were sequenced by the Max-Planck Institut internal sequencing group. The selected clones, before given for sequencing, were re-arrayed from the 384-well plates of the initial libraries into a new set of 384-well plates representing a non-redundant, normalised library. The details on the sequencing chemistry are given in Chapter 2.

The sequence traces were processed using a package of programs (cDNA PIPE-LINE) linked together by Steffen Hennig, where:

1. the PHRED program (Ewing and Green, 1998; Ewing et al., 1998) reads the sequence and performs new base calling, assigns quality values to the bases and writes the sequences in FASTA format output files. Using PHRED’s quality values low quality sequence was trimmed.

2. the above sequence was masked from vector sequences

3. using Repeatmasker (Smit and Green, ) interspersed repeats recognised in mammalian genomes (Repbase) as well as low complexity DNA sequences (e.g poly A tail) were masked.

4. finally the processed sequence was run against the locally installed Swissprot, Genembl, dbEST, human, mouse and rat Unigene sets at NCBI, the tentative consensus human Unigene from RZPD and Genecards (at the Weizmann Institute Genome Centre), databases using
gapped WU-BLAST v.2.0 (Altschul and Gish, 1996; Altschul et al., 1990) similarity searches. The parameters that are routinely used for the blastx program are: Expect value of $10^{-1}$ and BLOSUM 62 scoring matrix (Henikoff and Henikoff, 1992).

5. A results table was compiled listing the clone and sequence trace name, the sequence length after masking and the three highest scoring matches against the public databases.

Since the lane tracking of the sequences was carried out manually using the Perkin Elmer manual lane tracking kit the sequence traces were also assessed for ‘lane crossing’. Samples that had the same database match, they overlapped after sequence clustering and they had been loaded on neighbor lanes were excluded from further analysis. We found that approximately 0.36% of the BFL26 sequences and 4.8% of the BFLG sequences fall in this category. The above estimation is based on 415 sequences from the BFLG and 552 sequences from the BFLG. The same analysis will be repeated for all the amphioxus ESTs.

1,718 5' tag sequences from the BFL26 and 2,365 from the BFLG were processed using the above programs. The average sequence length after masking, which was used in the similarity searches, is between 500-600bp.

641 different sequences out 1,673 (38%) of the BFL26 sequences gave similarity score above $1.0 \times 10^{-10}$ with any of the sequences deposited in GenEmbl, Swplus, UniGene (NCBI)-Human/Mouse/Rat databases. Similarly, in the case of the BFLG library, 455 different sequences out of 2,197 (20.7%) were matched. In order to assess whether this was due to the fact that some of those sequences might be part of 3'UTR, all the sequences were run through two coding prediction programs, the GENSCAN (Burge and Karlin, 1997) and the MZEF (Zhang, 1997).

4.2. Prediction of coding potential

Since the generation of ESTs as a means of quickly accessing the genes that are expressed in a specific tissue, is a relatively new approach, the majority of the coding prediction programs that are publicly available, have been developed for the interpretation of genomic

---

1 Expect value: The statistical significance threshold for reporting matches against database sequences; the default value is 10, such that 10 matches are expected to be found merely by chance (Karlin and Altschul, 1990).

2 Substitution matrices: PAM (point accepted mutation) constructed by comparison of amino acid substitutions in closely related proteins. One PAM corresponds to an average change in 1% of all amino acid positions. BLOSUM matrices are constructed by examining multiple alignments of distantly related protein regions. The PAM scoring matrix is better for detecting the evolutionary distance of the compared proteins while the BLOSUM matrix for finding biological relationships (Altschul, 1996).
sequences. Prediction of coding potential of the first human ESTs (Adams et al., 1995) generated, was carried out using the GRAIL neural network or sequence pattern recognition program (Uberbacher and Mural, 1991). They have assessed the efficiency of this program by estimating the percentage of false negatives in a set of ESTs known to contain coding by their BLAST result. During this study two programs which were developed more recently, the GENSCAN (Burge and Karlin, 1997) and MZEF (Zhang, 1997) were used for prediction of the coding potential of the amphioxus ESTs.

Both programs were again developed for estimating the coding potential of genomic sequences which are high quality sequences. Moreover, the accuracy of prediction of both programs changes according to the organism from which the sequence is derived. For example the accuracy of GENSCAN is higher in predicting genes in human/vertebrate genomic sequences.

In order to estimate the efficiency of those programs in predicting the coding potential of amphioxus sequences, the two programs were initially run through the 162 amphioxus full length sequences which are currently deposited in the Genbank (Table 4.1). These sequences are known to be coding. Moreover, their sequence quality is over 99% and therefore sequencing errors, which are common in ESTs, will not affect the performance of those programs. As a result, 24.7% of the amphioxus full length mRNA sequences are predicted to be not coding with either of the two programs.

<table>
<thead>
<tr>
<th>Coding statistics</th>
<th>M_or_G</th>
<th>M_and_G</th>
<th>not_coding</th>
<th>nr of reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amphioxus full length in Genbank</td>
<td>122(75.3%)</td>
<td>90 (55.5%)</td>
<td>40 (24.7%)</td>
<td>162 (100 %)</td>
</tr>
<tr>
<td>Amphioxus ESTs matching the full length</td>
<td>176 (81.1%)</td>
<td>110 (50.7%)</td>
<td>41 (18.9%)</td>
<td>217 (100 %)</td>
</tr>
<tr>
<td>Amphioxus the whole EST set (BFL26)</td>
<td>1287 (74.9%)</td>
<td>706 (41.1%)</td>
<td>431 (25.1%)</td>
<td>1718 (100 %)</td>
</tr>
<tr>
<td>Amphioxus the whole EST set (BFLG)</td>
<td>1214 (51.7%)</td>
<td>344 (14.7%)</td>
<td>1133 (48.3%)</td>
<td>2347 (100 %)</td>
</tr>
</tbody>
</table>

In order to assess the efficiency of those programs on amphioxus EST sequences, 90 of the amphioxus full length sequences (the ones predicted to be coding by both programs) were
‘blasted’ against all the amphioxus ESTs generated during this project using the tblastx program. ESTs having probability score $P < e^{-25}$ were selected as being the same transcript.

217 matching ESTs were selected and subsequently run through both the MZEF and GENSCAN programs. 18.9% were falsely predicted to be non-coding. The two programs were subsequently run on all the amphioxus ESTs. As a result, 74.9% of BFL26 and 51.7% of the BFLG sequences are predicted to be coding with either of the two programs (Table 4.1). If we consider that the mean percentage of the false negative prediction of both programs is approximately 20%, then approximately 95% of the BFL26 and 70% of the BFLG sequences could be coding.

4.3. Functional classification of EST sequences.

As increasing amount of genomic sequences become available from multiple organisms the classification of sequences into functional groups allows an overview of the distribution of gene function of whole genomes across multiple phyla. So far functional classification of whole genomes was possible for small genomes such as bacteria, viruses or yeast (Tamames et al., 1996) since only these species have been completely sequenced. Moreover, the idea of functional classification, either manual or automated, of sequences has been applied in large human EST collections (Adams et al., 1995).

To assess the gene function diversity between two embryonic stages, the amphioxus 5’ EST were classified in 14 functional classes using the EUCLID annotation program (Tamames et al., 1998a; Tamames et al., 1998b) after being modified by Steffen Hennig (Max-Planck Institute for Molekular Genetics).

The classification was based on a dictionary of 823 specific keywords (Bairoch and Apweiler, 1997) found in the annotation of the swissprot sequences. ESTs were classified according to the keywords found in their highest score against the swissprot database. The 14 classes are the expansion of an initial three classes scheme consisting of the energy, information and communication functional classes (Tamames et al., 1996).

The Energy class includes the proteins related to metabolism and it is divided in the Amino acid biosynthesis, the Biosynthesis of cofactors, prosthetic groups, and carriers, the Central intermediary metabolism, the Energy metabolism, the Fatty acid and phospholipid metabolism and the Purines, pyrimidines, nucleosides, and nucleotides.

The information class includes proteins related to DNA structure replication and repair,
transcription, splicing and translation and it is divided in the Replication, Translation
Transcription subclasses.

The communication class includes the signalling molecules and it is divided in the
Regulatory functions the Transport and binding proteins, the Cellular processes and the Cell
envelope subclasses.

Examples of keywords that were used for the classification of transcripts in specific
classes are given in Table 4.3.

The class unknown contains the transcripts for which their highest score match against
swissprot contains keywords that have not been assigned into any functional class as well as
sequences that do not contain keywords in the annotation field. Sequences containing
keywords that have not been assigned to a functional class include repeat sequences and
hypothetical proteins of unknown function.

356 5' ESTs from the gastrula stage and 488 5' ESTs from the neurula stage for which the
highest probability score against swissprot sequences was lower than e^{-10} were classified
using the *EUCLID* program (Table 4.3).

In the gastrula stage 45% (162 sequences) of the identified genes are associated with basic
metabolism (Energy class), 25.5% (91 sequences) with gene/protein expression (Information
class) and 16% (57 sequences) with signalling (Communication class).

**Table 4.2.** Differences in the distribution of transcripts in functional classes between the gastrula
(BFLG) and neurula (BFL26) stages of amphioxus (*Branchiostoma floridae*). The classification of
ESTs in functional groups was carried out automatically using the *EUCLID* program (Tamames et al.,
1998a).

<table>
<thead>
<tr>
<th>Functional class</th>
<th>BFLG (356 sequences annotated)</th>
<th>BFL26 (488 sequences annotated)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amino acid biosynthesis</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>Biosynthesis of cofactors, prosthetic groups, and carriers</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>Cell envelope</td>
<td>7</td>
<td>24</td>
</tr>
<tr>
<td>Cellular processes</td>
<td>29</td>
<td>52</td>
</tr>
<tr>
<td>Central intermediary metabolism</td>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td>Energy metabolism</td>
<td>127</td>
<td>82</td>
</tr>
<tr>
<td>Fatty acid and phospholipid metabolism</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Other categories</td>
<td>5</td>
<td>9</td>
</tr>
<tr>
<td>Purines, pyrimidines, nucleosides, and nucleotides</td>
<td>14</td>
<td>51</td>
</tr>
<tr>
<td>Regulatory functions</td>
<td>14</td>
<td>22</td>
</tr>
<tr>
<td>Replication</td>
<td>21</td>
<td>13</td>
</tr>
<tr>
<td>Transcription</td>
<td>16</td>
<td>19</td>
</tr>
<tr>
<td>Translation</td>
<td>54</td>
<td>65</td>
</tr>
<tr>
<td>Transport and binding proteins</td>
<td>21</td>
<td>54</td>
</tr>
<tr>
<td>Unknown</td>
<td>27</td>
<td>67</td>
</tr>
</tbody>
</table>

In the neurula stage 33.5% (163 sequences) of the identified genes are associated with
basic metabolism (Energy class), 19.9% (97 sequences) with gene/protein expression
Table 4.3. Examples of classification of amphioxus ESTs into the EUCLID (Tamames *et al.*, 1998a) functional classes according to keywords (Bairoch and Apweiler, 1997) contained within the swissprot annotation of the highest score match.

<table>
<thead>
<tr>
<th>Functional Class</th>
<th>Extracted keywords</th>
<th>Swissprotmatch</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amino acid biosynthesis</td>
<td>ZINC LYASE</td>
<td>P00918 homo sapiens (human). carbonic anhydrase ii</td>
</tr>
<tr>
<td></td>
<td>TRANSIT PEPTIDE POLYMORPHISM CHLOROPLAST LYASE HEME</td>
<td>P48417 linum usitatissimum (flax) (linseed). allene oxide synthase precursor</td>
</tr>
<tr>
<td>Biosynthesis of cofactors, prosthetic groups, and carriers</td>
<td>ONE-CARBON METABOLISM NADP OXIDOREDUCTASE</td>
<td>P28037 rattus norvegicus (rat).10-formyltetrahyd rofolate dehydrogenase</td>
</tr>
<tr>
<td></td>
<td>LUMINESCENCE 3D-STRUCTURE</td>
<td>Q93125 aequorea victoria (jellyfish). Green fluorescent protein mutant 3.</td>
</tr>
<tr>
<td></td>
<td>ISOMERASE ENDOPLASMIC RETICULUM REPEAT REDOX-ACTIVE CENTER SIGNAL</td>
<td>P38699 mesocricetus auratus (golden hamster). probable protein disulfide isomerase p5 precursor</td>
</tr>
<tr>
<td></td>
<td>ELECTRON TRANSPORT REDOX-ACTIVE CENTER</td>
<td>P08629 gallus gallus (chicken). thioredoxin</td>
</tr>
<tr>
<td>Cell envelope</td>
<td>HEPTAD REPEAT PATTERN ANTIGEN</td>
<td>P21249 onchocerca volvulus. major antigen</td>
</tr>
<tr>
<td></td>
<td>HYDROLASE LIPID DEGRADATION PANCREAS SIGNAL</td>
<td>Q64424 myocastor coypus (coypu) (nutria). Pancreatic lipase related protein 2 precursor</td>
</tr>
<tr>
<td></td>
<td>HYDROLASE FATTY ACID BIOSYNTHESIS SIGNAL</td>
<td>Q04791 anas platyrhynchos (domestic duck).fatty acyl-coa hydrolase precursor</td>
</tr>
<tr>
<td></td>
<td>SIGNAL STORAGE PROTEIN</td>
<td>Q91062 ichthyomyzon unicuspus (silver lamprey).vitellogenin precursor</td>
</tr>
<tr>
<td>Cellular processes</td>
<td>GTP-BINDING MULTIGENE FAMILY MICROTUBULES</td>
<td>P02553 lytechinus pictus (painted sea urchin).tubulin alpha chain (fragment</td>
</tr>
<tr>
<td></td>
<td>MULTIGENE FAMILY ATP-BINDING CHAPERONE</td>
<td>P80316 mus musculus (mouse). t-complex protein 1,epsilon subunit (tcp-1-epsilon).</td>
</tr>
<tr>
<td></td>
<td>MYOSIN</td>
<td>Q17133 branchiostoma floridiae myosin, essential light chain (myosin light chain alkali).</td>
</tr>
<tr>
<td></td>
<td>CALCIUM-BINDING MYOSIN</td>
<td>P40423 drosophila melanogaster myosin regulatory light chain, nonmuscle (mrle-c).</td>
</tr>
<tr>
<td></td>
<td>PROTEASE HYDROLASE PROTEASOME</td>
<td>P12881 drosophila melanogaster (fruit fly).proteasome 29 kd subunit</td>
</tr>
<tr>
<td>Central intermediary metabolism</td>
<td>NAD OXIDOREDUCTASE MITOCHONDRION UBIQUINONE</td>
<td>P55787 gadus morhua (atlantic cod). atp synthase a chain</td>
</tr>
<tr>
<td></td>
<td>NAD OXIDOREDUCTASE</td>
<td>P49419 homo sapiens (human). antiquitin</td>
</tr>
<tr>
<td></td>
<td>OXIDOREDUCTASE</td>
<td>P37440 escherichia coli. oxidoreductase ucpa</td>
</tr>
<tr>
<td></td>
<td>MULTIGENE FAMILY LIGASE MAGNESIUM FATTY ACID METABOLISM</td>
<td>P41215 homo sapiens (human) long-chain-fatty-acid coa ligase 1</td>
</tr>
<tr>
<td>Other categories</td>
<td>RNA-BINDING NUCLEAR PROTEIN MRNA PROCESSING MRNA SPlicing</td>
<td>Q13435 homo sapiens (human). Spliceosome associated protein 145 (sap 145)</td>
</tr>
<tr>
<td>Functional class</td>
<td>Extracted keywords</td>
<td>SwissProt/Match</td>
</tr>
<tr>
<td>--------------------------</td>
<td>------------------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Purines, pyrimidines,</td>
<td>THIOL PROTEASE HYDROLASE MULTIGENE FAMILY</td>
<td>P20807 homo sapiens (human). calpain p94, large (cat alytic) subunit (ec 3.4.22.17) (calcium- activated neutral proteinase)</td>
</tr>
<tr>
<td>nucleosides, and</td>
<td>CALCIUM-BINDING</td>
<td></td>
</tr>
<tr>
<td>nucleotides</td>
<td>KINASE TRANSFERASE</td>
<td>P52623 mus musculus (mouse). uridine kinase</td>
</tr>
<tr>
<td></td>
<td>ZYMOSGEN LYPSOSOME THIOL PROTEASE HYDROLASE</td>
<td>P00787 rattus norvegicus (rat). cathepsin b precursor</td>
</tr>
<tr>
<td></td>
<td>SIGNAL GLYCOPROTEIN</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GLYCOSYLTRANSFERASE MAGNESIUM PURINE SALVAGE TRANSFERASE</td>
<td>Q64531 mus spretus hypoxanthine-guanine phosphoribosyltransferase</td>
</tr>
<tr>
<td>Regulatory functions</td>
<td>PHOSPHORYLATION NUCLEAR PROTEIN 3D-STRUCTURE</td>
<td>P23197 homo sapiens heterochromatin protein 1 homolog gamma</td>
</tr>
<tr>
<td></td>
<td>REPEAT TRANSCRIPTION REGULATION NUCLEAR PROTEIN DNA-BINDING ACTIVATOR</td>
<td>O16102 ting philia melanogaster (fruit fly).b52.</td>
</tr>
<tr>
<td></td>
<td>ATP-BINDING ANTIGEN HELICASE</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PHOSPHORYLATION INTERMEDIATE FILAMENT HEPTAD REPEAT PATTERN KERATIN COILED COIL</td>
<td>Q92676 lorida laurus (human). mesothelial keratin k filament pr</td>
</tr>
<tr>
<td></td>
<td>LIM MOTIF REPEAT METAL-BINDING ZINC-FINGER ZINC</td>
<td>O35115 rattus norvegicus (rat). dral</td>
</tr>
<tr>
<td>Replication</td>
<td>MITOSIS PHOSPHORYLATION ATP-BINDING TYROSINE-PROTEIN KINASE TRANSFERASE</td>
<td></td>
</tr>
<tr>
<td></td>
<td>NUCLEOSOME CORE DNA-BINDING NUCLEAR PROTEIN MULTIGENE FAMILY CHROMOSOMAL PROTEIN</td>
<td>Q27748 paracentrotus lividus (common sea urchin). histone h3</td>
</tr>
<tr>
<td>Transcription</td>
<td>DEVELOPMENTAL PROTEIN HOMEBOX NUCLEAR PROTEIN DNA-BINDING</td>
<td>Q64317 mus musculus (mouse). homeobox proteinidix-1</td>
</tr>
<tr>
<td></td>
<td>RNA-BINDING NUCLEAR PROTEIN ATP-BINDING HELICASE</td>
<td>P17844 homo sapiens (human). Probable rna-dependent helicase p68 (dead-box protein p68) (dead box protein 5)</td>
</tr>
<tr>
<td>Translation</td>
<td>RIBOSOMAL PROTEIN</td>
<td>P55858 sulfolobus solfataricus. 30s ribosomal protein hs6-like.</td>
</tr>
<tr>
<td></td>
<td>GTP-BINDING PROTEIN BIOSYNTHESIS ELONGATION FACTOR</td>
<td>P17508 xenopus laevis (african clawed frog). elongation factor 1-alpha, oocyte form (ef-1-alpha-01)</td>
</tr>
<tr>
<td>Transport and binding</td>
<td>REPEAT CELL ADHESION EGF-LIKE DOMAIN SIGNAL TRANSMEMBRANE GLYCOPROTEIN</td>
<td>Q28983 sus scrofa (pig). zonadhesin precursor</td>
</tr>
<tr>
<td>proteins</td>
<td>BRAIN SIGNAL CALCIUM-BINDING GLYCOPROTEIN</td>
<td>Q90399 danio aequipinnatus (giant danio) (brachydanio aequipinnatus). epyedymin</td>
</tr>
<tr>
<td>Unknown</td>
<td>NONE</td>
<td>Q14008 homo sapiens ch-tog protein (colonic and hepatic tumor over-expressed protein)</td>
</tr>
<tr>
<td></td>
<td>ACTIN-BINDING</td>
<td>P25229 xenopus laevis (african clawed frog). actin-binding protein chain a</td>
</tr>
</tbody>
</table>
(Information class) and 31% (152 sequences) with signalling (Communication class).

Therefore, while the percentage of genes controlling metabolism and gene/protein expression is similar in both stages, there is a significant increase in the cell signalling and communication molecules at the neurula stage.

In overall, for the whole embryo (844 classified ESTs), 38.5% (325 sequences) of the identified genes are associated with basic metabolism (Energy class), 22.27% with gene/protein expression (Information class) and 24.7% with signalling (Communication class).

4.4. Conclusions

3,870 clones out of 22,000 rearrayed clones from both libraries have been 5' tag sequenced. The EST sequences were initially processed using a package of programs that masks vector and repeat regions as well as low quality. Finally, the sequences were run against the Swissprot, Genembl, dbEST, human, mouse and rat Unigene sets at NCBI, and the tentative consensus human Unigene from RZPD and Genecards (at the Weizmann Institute Genome Centre), databases using gapped WU-BLAST v.2.0.

The percentage of sequences with no match was estimated in combination with the percentage of sequences that were predicted to be coding. The efficiency of two programs to predict the coding potential of full length amphioxus mRNA sequences was assessed. Both of them showed on average 20% false negative rate. Application of those programs on amphioxus ESTs matching the same full length amphioxus sequences gave the same average false negative rate. This indicates that the quality of the tested ESTs is such that it does not affect the prediction of coding regions.

Finally, the gene function diversity between the two tissues was assessed by classifying the identified transcripts into functional classes using the EUCLID program.

As a result:
• 641 different sequences out of 1,673 (38%) BFL26 sequences gave similarity score above \(1.0 \times 10^{-10}\) with any of the sequences deposited in GenEmbl, Swplus, UniGene(NCBI)-Human/Mouse/Rat databases. Similarly, in the case of the BFLG library, 455 different sequences out of 2,197 (20.7%) were matched.

• 74.9% of BFL26 and 51.7% of the BFLG sequences were predicted to be coding with either the GENSCAN or MZEF prediction programs. If we consider that the mean percentage of the false negative prediction of both programs is approximately 20%, then approximately 95% of
the BFL26 and 70% of the BFLG sequences could be coding. If the percentage of the matched sequences is deducted from the percentage of estimated coding containing sequences, then approximately 36.9% of the generated BFL26 5’ ESTs and 31% of the BFLG 5’ ESTs remain unmatched although they contain coding.

• From 356 identified ESTs from the gastrula stage, 45% (162 sequences) are associated with basic metabolism (Energy class), 25.5% (91 sequences) with gene/protein expression (Information class) and 16% (57 sequences) with signalling (Communication class). From 488 identified ESTs from the neurula stage, 33.5% (163 sequences) are associated with basic metabolism (Energy class), 19.9% (97 sequences) with gene/protein expression (Information class) and 31% (152 sequences) with signalling (Communication class). These results indicate that there is an increase in the number of signalling molecules in the neurula stage. This might be the result of this selected sequences and it will have to be re-evaluated as more clones are sequenced. Finally, 38.5% (325 sequences) of all transcripts are associated with basic metabolism (Energy class), 22.27% with gene/protein expression (Information class) and 24.7% with signalling (Communication class).
Chapter 5

Amphioxus genes expressed at gastrula and neurula stages

Introduction

In chapter 3 oligonucleotide fingerprinting was evaluated as a method which is able to cluster reliably clones according to their sequence pattern as defined by the hybridisation of short oligonucleotides. 5' tag sequencing of 3,870 cDNA clones picked from both amphioxus embryonic libraries and their sequence clustering has verified that 82.7% of those represent different transcripts. 38% of the 1,673 BFL26 and 20.7% of the 2,197 BFLG sequences gave high similarity score match while 74.9% and 51.7% of all the BFL26 and BFLG sequences are predicted to be coding with either the GENSCAN or MZEF prediction programs (Chapter 4).

This chapter focuses on the function of the transcripts as implied by their higher similarity score to genes isolated in other organisms. Which genes are expressed in higher level in both stages? Sequencing of representative clones of the largest cluster from both amphioxus libraries showed that the polyadenylated mitochondrial 16S rRNA is the transcript expressed at highest level in both stages. Out of 35,482 clones that were considered in the clustering of the neurula stage library, 815 clones were mitochondrial 16S rRNA (2.2%). The number of clones from the same transcript in the gastrula stage library is more than four times higher, 9.6%. This was a gene class that was difficult to normalise especially in the gastrula stage library. Comparison of the most abundant transcripts between the two stages shows that in the neurula stage there is an increase in transcripts expressed in muscle tissues. As an example, the number of muscle actin transcripts increases from 0.12% (46/35,482) in the gastrula stage, to 1.8% (115/35,482) in the neurula stage. Similar increase is observed for the myosin heavy chain, creatine kinase etc. Tables 5.1. and 5.2 give a sample of the transcripts that have been identified.

The presence of specific transcripts in either of the two stages insinuates which embryonic structures have been formed at these embryonic stages or even exist in amphioxus. As an example Otx and Dll, are both expressed in both stages at similar levels (Table 5.3.)
Table 5.1. Overview of transcripts expressed in the amphioxus neurula stage.

<table>
<thead>
<tr>
<th>Clone coordinates</th>
<th>Cluster size/Abundance</th>
<th>Blast Match</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>BFL26_64I23</td>
<td>815</td>
<td>Branchiostoma lanceolatum complete mitochondrial genome</td>
<td>4.6e-171</td>
</tr>
<tr>
<td>BFL26_66E14</td>
<td>631</td>
<td>Branchiostoma floridæ (florida lancelet) (amphioxus). Muscle actin</td>
<td>2.1e-142</td>
</tr>
<tr>
<td>BFL26_41G5</td>
<td>316</td>
<td>Drosophila melanogaster (fruit fly) (amphioxus). Tubulin alpha-1 chain</td>
<td>2.1e-158</td>
</tr>
<tr>
<td>BFL26_37E03</td>
<td>314</td>
<td>Branchiostoma floridæ (florida lancelet) (amphioxus). NADH dehydrogenase subunit 2</td>
<td>8.4e-137</td>
</tr>
<tr>
<td>BFL26_40G14</td>
<td>287</td>
<td>Branchiostoma floridæ (florida lancelet) (amphioxus). Muscle actin</td>
<td>2.7e-45</td>
</tr>
<tr>
<td>BFL26_50E20</td>
<td>190</td>
<td>CANIS FAMILIARIS (dog). Creatine kinase, m chain</td>
<td>1.2e-112</td>
</tr>
<tr>
<td>BFL26_23H3</td>
<td>185</td>
<td>Branchiostoma floridæ (florida lancelet) (amphioxus). Myosin chain</td>
<td>2.7e-73</td>
</tr>
<tr>
<td>BFL26_109D23</td>
<td>175</td>
<td>Drosophila melanogaster (fruit fly) (amphioxus). Cytochrome b</td>
<td>8.2e-128</td>
</tr>
<tr>
<td>BFL26_33A13</td>
<td>89</td>
<td>Homo sapiens (human). Wnt-7a protein precursor</td>
<td>3.5e-25</td>
</tr>
<tr>
<td>BFL26_120O15</td>
<td>101</td>
<td>Arabidopsis thaliana (mouse-ear cress). Actin depolymerizing factor 1</td>
<td>2.6e-30</td>
</tr>
<tr>
<td>BFL26_65O19</td>
<td>115</td>
<td>Branchiostoma floridæ (florida lancelet) (amphioxus). Essential light chain (myosin light chain alkali)</td>
<td>1.3e-94</td>
</tr>
<tr>
<td>BFL26_20A16</td>
<td>108</td>
<td>Homo sapiens (human). Complement 1</td>
<td>3.9e-22</td>
</tr>
<tr>
<td>BFL26_120O15</td>
<td>101</td>
<td>Mus musculus (mouse) ribosomal protein S3a. S97</td>
<td>2.5e-137</td>
</tr>
<tr>
<td>BFL26_20K12</td>
<td>94</td>
<td>Drosophila melanogaster (fruit fly) (amphioxus). Cytochrome oxidase subunit i. 4.5e-112</td>
<td></td>
</tr>
<tr>
<td>BFL26_76I23</td>
<td>84</td>
<td>Drosophila melanogaster (fruit fly) (amphioxus). Cytochrome oxidase subunit ii. 5.3e-130</td>
<td></td>
</tr>
<tr>
<td>BFL26_63C23</td>
<td>83</td>
<td>Volvox carteri. Sulfated surface glycoprotein 185 (ssg 185)</td>
<td>4.0e-11</td>
</tr>
<tr>
<td>BFL26_113G13</td>
<td>71</td>
<td>Rattus norvegicus (rat). 60s ribosomal protein I3 (I4).</td>
<td>2.4e-127</td>
</tr>
<tr>
<td>BFL26_94G02</td>
<td>60</td>
<td>Caenorhabditis elegans. T21h3.3 protein.</td>
<td>1.8e-50</td>
</tr>
<tr>
<td>BFL26_40D16</td>
<td>57</td>
<td>Homo sapiens (human). Mac25 protein</td>
<td>7.1e-21</td>
</tr>
<tr>
<td>BFL26_114D01</td>
<td>56</td>
<td>Branchiostoma floridæ (florida lancelet) (amphioxus). Cytochrome oxidase subunit i.</td>
<td>5.7e-148</td>
</tr>
<tr>
<td>BFL26_129D24</td>
<td>55</td>
<td>Branchiostoma floridæ (amphioxus) and branchiostoma lanceolatum (amphioxus). 1.8e-143</td>
<td></td>
</tr>
<tr>
<td>BFL26_78A16</td>
<td>54</td>
<td>Branchiostoma floridæ (florida lancelet) (amphioxus). Calcium vector protein 2.2e-108</td>
<td></td>
</tr>
<tr>
<td>BFL26_112J21</td>
<td>54</td>
<td>Saccoglossus kwalevskii. Actin 2</td>
<td>1.4e-123</td>
</tr>
<tr>
<td>BFL26_54I15</td>
<td>53</td>
<td>Homo sapiens (human). Tubulin beta-2 chain</td>
<td>4.5e-137</td>
</tr>
<tr>
<td>BFL26_78B13</td>
<td>53</td>
<td>Drosophila melanogaster (fruit fly). Qm homolog</td>
<td>1.0e-115</td>
</tr>
<tr>
<td>BFL26_92D23</td>
<td>41</td>
<td>Drosophila melanogaster (fruit fly). Myosin regulatory light chain, nonmuscle (mrlc-c)</td>
<td>4.2e-28</td>
</tr>
<tr>
<td>BFL26_66A04</td>
<td>40</td>
<td>Methanococcus jannaschii. Hypothetical protein mj1065</td>
<td>3.0e-21</td>
</tr>
<tr>
<td>Clone coordinate</td>
<td>Cluster size/ Abundance</td>
<td>Blast Match</td>
<td>P-value</td>
</tr>
<tr>
<td>------------------</td>
<td>-------------------------</td>
<td>----------------------------------------------------------------------------</td>
<td>-----------</td>
</tr>
<tr>
<td>BFL26_120C18</td>
<td>singleton</td>
<td>O21004 branchiostoma lanceolatum (common lancelet) (amphioxus). atp synthase a chain</td>
<td>1.5e-35</td>
</tr>
<tr>
<td>BFL26_11F12</td>
<td>40</td>
<td>Q99020 mus musculus (mouse). carg-binding factor-a (cbf-a).</td>
<td>6.3e-109</td>
</tr>
<tr>
<td>BFL26_129I23</td>
<td>39</td>
<td>P19889 drosophila melanogaster (fruit fly). 60s acidic ribosomal protein p0 (deoxyribonuclease (apurinic or apyrimidinic))</td>
<td>3.2e-104</td>
</tr>
<tr>
<td>BFL26_24G24</td>
<td>37</td>
<td>P61004 xenopus laevis (african clawed frog). voltage-dependent anion-selective channel protein 2</td>
<td>3.0e-15</td>
</tr>
<tr>
<td>BFL26_18C20</td>
<td>35</td>
<td>Q25148 halocynthia roretzi (sea squirt). Troponin t.</td>
<td>1.2e-99</td>
</tr>
<tr>
<td>BFL26_23J13</td>
<td>33</td>
<td>P22700 drosophila melanogaster calcium-transporting atpase sarcoplasmic/endoplasmic reticulum type (calcium pump).</td>
<td>7.3e-46</td>
</tr>
<tr>
<td>BFL26_24D19</td>
<td>33</td>
<td>P13060 drosophila melanogaster (fruit fly). elongation factor 2 (ef-2).</td>
<td>8.4e-37</td>
</tr>
<tr>
<td>BFL26_51F07</td>
<td>32</td>
<td>Q92005 brachydanio rerio (zebrafish) (zebra danio). translational elongation factor-1 alpha.</td>
<td>3.8e-137</td>
</tr>
<tr>
<td>BFL26_122B07</td>
<td>singleton</td>
<td>Q08247 homo sapiens (human). Gamma-glutamyltranspeptidase (ec 2.3.2.2) (gamma-glutamyltransferase).</td>
<td>2.1e-105</td>
</tr>
<tr>
<td>BFL26_96I14</td>
<td>30</td>
<td>P25121 rattus norvegicus (rat). 60s ribosomal protein 11</td>
<td>7.0e-82</td>
</tr>
<tr>
<td>BFL26_32C13</td>
<td>30</td>
<td>P53482 fugu rubripes (japanese pufferfish) (takifugu rubripes). actin, alpha skeletal muscle 2.</td>
<td>1.2e-95</td>
</tr>
<tr>
<td>BFL26_49F13</td>
<td>29</td>
<td>P91791 hemicentrotus pulcherrimus (sea urchin). Cyclophilin</td>
<td>6.1e-121</td>
</tr>
<tr>
<td>BFL26_9D04</td>
<td>28</td>
<td>Q93129 branchiostoma belcheri. Actin, cytoplasmic (bbca1).</td>
<td>3.5e-84</td>
</tr>
<tr>
<td>BFL26_8G08</td>
<td>28</td>
<td>Q20655 caenorhabditis elegans. 14-3-3-like protein 2.</td>
<td>2.4e-101</td>
</tr>
<tr>
<td>BFL26_89F12</td>
<td>28</td>
<td>O21728 astatheros macracanthum. Cytochrome b.</td>
<td>7.3e-102</td>
</tr>
<tr>
<td>BFL26_92K06</td>
<td>27</td>
<td>P50890 gallus gallus (chicken). 40s ribosomal protein sa (p40)</td>
<td>1.4e-113</td>
</tr>
<tr>
<td>BFL26_96B09</td>
<td>27</td>
<td>O47422 branchiostoma floridiae (florida lancelet) (amphioxus). cytochrome oxidase subunit i.</td>
<td>1.1e-107</td>
</tr>
<tr>
<td>BFL26_51L13</td>
<td>27</td>
<td>P08429 xenopus laevis (african clawed frog). 60s ribosomal protein 11a</td>
<td>8.0e-60</td>
</tr>
<tr>
<td>BFL26_37I08</td>
<td>27</td>
<td>P05548 branchiostoma lanceolatum (common lancelet) (amphioxus). cavp-target protein (cavpt).</td>
<td>2.9e-50</td>
</tr>
<tr>
<td>BFL26_59I16</td>
<td>27</td>
<td>P94041 caenorhabditis elegans. Probable 40s ribosomal protein s5</td>
<td>1.1e-127</td>
</tr>
<tr>
<td>BFL26_63C20</td>
<td>27</td>
<td>O62004 branchiostoma floridiae (florida lancelet) (amphioxus). intermediate filament protein df1.</td>
<td>1.4e-47</td>
</tr>
<tr>
<td>BFL26_79K22</td>
<td>27</td>
<td>Q24186 drosophila melanogaster (fruit fly). 40s ribosomal protein s5.</td>
<td>2.3e-65</td>
</tr>
<tr>
<td>BFL26_11D06</td>
<td>26</td>
<td>P08108 oncorhynchus mykiss (rainbow trout) (salmo gairdneri). heat shock cognate 70 kd protein (hsp70).</td>
<td>6.9e-88</td>
</tr>
<tr>
<td>BFL26_81M06</td>
<td>23</td>
<td>P41051 rattus norvegicus (rat). 60s ribosomal protein 115.</td>
<td>7.6e-42</td>
</tr>
<tr>
<td>BFL26_123P07</td>
<td>6</td>
<td>O47430 branchiostoma floridiae (florida lancelet) (amphioxus). nadh dehydrogenase subunit 5</td>
<td>7.3e-16</td>
</tr>
<tr>
<td>BFL26_63E19</td>
<td>6</td>
<td>P07452 oryctolagus cuniculus (rabbit). Carbonic anhydrase i</td>
<td>4.1e-55</td>
</tr>
<tr>
<td>BFL26_99O05</td>
<td>6</td>
<td>Q42706 unknown. glutathione s-transferase (ec 2.5.1.18) (glutathione transferase) (glutathione s-alkyltransferase)</td>
<td>3.0e-67</td>
</tr>
<tr>
<td>BFL26_11L08</td>
<td>6</td>
<td>P13942 homo sapiens (human). Calmodulin.</td>
<td>3.2e-71</td>
</tr>
<tr>
<td>BFL26_23G09</td>
<td>6</td>
<td>P19965 torpedo californica (pacific electric ray). sits-binding protein (sp105).</td>
<td>7.6e-26</td>
</tr>
<tr>
<td>BFL26_42E23</td>
<td>6</td>
<td>P12333 homo sapiens (human). ras-related protein ral-a</td>
<td>1.6e-80</td>
</tr>
<tr>
<td>BFL26_68O08</td>
<td>3</td>
<td>P35603 caenorhabditis elegans. Clathrin coat assembly protein ap50 (clathrin coat associated protein ap50).</td>
<td>3.4e-16</td>
</tr>
<tr>
<td>Clone coordinate</td>
<td>Cluster size/Abundance</td>
<td>Blast Match</td>
<td>E-value</td>
</tr>
<tr>
<td>------------------</td>
<td>------------------------</td>
<td>-------------------------------------------------------------------------------</td>
<td>---------</td>
</tr>
<tr>
<td>BFL26_54J06</td>
<td>5</td>
<td>Q05250 sus scrofa (pig), steroid membrane binding protein</td>
<td>3.6e-67</td>
</tr>
<tr>
<td>BFL26_87I14</td>
<td>3</td>
<td>Q09536 homo sapiens (human). Synaptic vesicle membrane protein vat-1 homolog.</td>
<td>1.9e-33</td>
</tr>
<tr>
<td>BFL26_56E23</td>
<td>3</td>
<td>Q15393 homo sapiens (human). Hypothetical protein kiaa0017. 7/98</td>
<td>3.4e-82</td>
</tr>
<tr>
<td>BFL26_95O17</td>
<td>3</td>
<td>P41150 oryctolagus cuniculus (rabbit). Calcyphosine (r2d5 antigen).</td>
<td>1.3e-17</td>
</tr>
<tr>
<td>BFL26_112I17</td>
<td>3</td>
<td>P23403 xenopus laevis (african clawed frog). 40s ribosomal protein s20</td>
<td>6.8e-06</td>
</tr>
<tr>
<td>BFL26_9P20</td>
<td>3</td>
<td>Q28465 macroops eugeni (tammar wallaby). atp-dependent ma-helicase (fragment).</td>
<td>6.7e-22</td>
</tr>
<tr>
<td>BFL26_8022</td>
<td>3</td>
<td>Q26630 stronglyocentrotus purpuratus (purple sea urchin). axonemal dynein light chain p33</td>
<td>2.4e-76</td>
</tr>
<tr>
<td>BFL26_7F13</td>
<td>3</td>
<td>P06603 drosophila melanogaster (fruit fly). tubulin alpha-1 chain.</td>
<td>5.9e-114</td>
</tr>
<tr>
<td>BFL26_40L03</td>
<td>3</td>
<td>O19003 bos taurus (bovine). Selonoprotein p-like protein.</td>
<td>8.9e-27</td>
</tr>
<tr>
<td>BFL26_128O01</td>
<td>3</td>
<td>P04843 homo sapiens (human). Dolichyl-diphosphooligosaccharide--protein glycosyltransferase 67 kd subunit precursor</td>
<td>4.4e-77</td>
</tr>
<tr>
<td>BFL26_68F22</td>
<td>3</td>
<td>O01305 branchiostoma floridai (florida lancelet) (amphioxus). calcium vector protein</td>
<td>1.3e-60</td>
</tr>
<tr>
<td>BFL26_58E10</td>
<td>3</td>
<td>P51975 ovis aries (sheep). Corticosteroid 11-beta-dehydrogenase, isozyme 1</td>
<td>3.5e-27</td>
</tr>
<tr>
<td>BFL26_99N11</td>
<td>3</td>
<td>Q64400 cricetulus longicauatus (long-tailed hamster) (chinese hamster). Histone h3</td>
<td>9.1e-87</td>
</tr>
<tr>
<td>BFL26_99D03</td>
<td>3</td>
<td>Q15608 homo sapiens (human). Troponin t (fragment).</td>
<td>4.1e-15</td>
</tr>
<tr>
<td>BFL26_99D21</td>
<td>3</td>
<td>Q70280 rattus norvegicus (rat). wap four-disulfide core domain protein.</td>
<td>1.4e-23</td>
</tr>
<tr>
<td>BFL26_98F13</td>
<td>3</td>
<td>P14865 rhizomucor racemosus (mucor circinelloides f. lusitanicus). elongation factor 1-alpha (ef-1-alpha).</td>
<td>3.5e-29</td>
</tr>
<tr>
<td>BFL26_25C07</td>
<td>2</td>
<td>O17497 branchiostoma belcheri. Aldolase.</td>
<td>9.5e-104</td>
</tr>
<tr>
<td>BFL26_03F09</td>
<td>17</td>
<td>branchiostoma floridai (florida lancelet) (amphioxus). intermediate filament protein b1 (fragment).</td>
<td>3.0e-26</td>
</tr>
<tr>
<td>BFL26_08P05</td>
<td>5</td>
<td>SP_IN:O62004 O62004 branchiostoma floridai (florida lancelet) (amphioxus).intermediate filament protein d1</td>
<td>7.2e-30</td>
</tr>
<tr>
<td>BFL26_08D20</td>
<td>3</td>
<td>SP_IN:O62004 O62004 branchiostoma floridai (florida lancelet) (amphioxus).intermediate filament protein d1</td>
<td>4.5e-45</td>
</tr>
<tr>
<td>BFL26_40O23</td>
<td>10</td>
<td>AJ22381 Brachiaostoma floridai mRNA for intermediate filament protein D1</td>
<td>2.3e-65</td>
</tr>
<tr>
<td>BFL26_20O10</td>
<td>2</td>
<td>SW;WNT2 DROME P28465 drosophila melanogaster (fruit fly). protein dwnt-2</td>
<td>8.0e-05</td>
</tr>
<tr>
<td>BFL26_38C2</td>
<td>3</td>
<td>SW;WNT3 XENLA P31284 xenopus laevis (african clawed frog). wnt-3 protein (xwnt-3)</td>
<td>5.4e-96</td>
</tr>
<tr>
<td>BFL26_31A1</td>
<td>3</td>
<td>SW;WNT6 XENLA P31287 xenopus laevis (african clawed frog). xwnt-6 protein</td>
<td>4.2e-69</td>
</tr>
<tr>
<td>BFL26_41F14</td>
<td>3</td>
<td>SW;WNT8 BRARE P51029 brachydanio rerio (zebrahiko) (zebra danio). wnt-8 protein precursor</td>
<td>3.1e-44</td>
</tr>
<tr>
<td>BFL26_2K11</td>
<td>2</td>
<td>Q90744 gallus gallus alpha-n-acetylgalactosaminidase (fragment). associated with I Schindler disease Kanzaki disease</td>
<td>1.4e-92</td>
</tr>
<tr>
<td>BFL26_50J12</td>
<td>4</td>
<td>AF014010 Mus musculus polycystic kidney disease 2 (Pkd2) mRNA, complete cds</td>
<td>2.1e-39</td>
</tr>
<tr>
<td>BFL26_053L24</td>
<td>3</td>
<td>U50928 Human autosomal dominantpolyzystic kidney disease type II (PKD2) mRNA, complete cds</td>
<td>4.5e-40</td>
</tr>
</tbody>
</table>
Table 5.2. Overview of transcripts expressed at the amphioxus gastrula stage

<table>
<thead>
<tr>
<th>Clone name</th>
<th>Cluster size/ Abundance</th>
<th>Blast Match</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>BFLG_42J22</td>
<td>2938</td>
<td>U97377 Trichechus manatus tRNA-Val and 16S ribosomal RNA genes, mitochondrial genes encoding mitochondrial RNAs</td>
<td>1.3e-46</td>
</tr>
<tr>
<td>BFLG_21G24</td>
<td>563</td>
<td>Q61755 daphnia pulex, cytochrome c oxidase polypeptide ii</td>
<td>4.2e-88</td>
</tr>
<tr>
<td>BFLG_16G18</td>
<td>312</td>
<td>B. flordiae mRNA for AmphiBrf52.</td>
<td>1.7e-12</td>
</tr>
<tr>
<td>BFLG_4E23</td>
<td>272</td>
<td>P11233 homo sapiens (human), ras-related protein ra-1.</td>
<td>1.6e-80</td>
</tr>
<tr>
<td>BFLG_22L06</td>
<td>253</td>
<td>Q07822 homo sapiens( human),mac25 protein</td>
<td>2.2e-35</td>
</tr>
<tr>
<td>BFLG_132N12</td>
<td>188</td>
<td>P2806 gryllus pennsylvanicanus, cytochrome oxidase i</td>
<td>1.2e-29</td>
</tr>
<tr>
<td>BFLG_127L13</td>
<td>117</td>
<td>O47434 branchiostoma floridai(florida lancelet) (amphioxus), nadh dehydrogenase subunit 2.</td>
<td>9.0e-107</td>
</tr>
<tr>
<td>BFLG_20L15</td>
<td>90</td>
<td>O21004 branchiostoma lanceolatum (common lancelet) (amphioxus), atp synthase a chain</td>
<td>1.5e-123</td>
</tr>
<tr>
<td>BFLG_18J18</td>
<td>80</td>
<td>P49154 urechis caupo (inkeeper worm) (spoonworm), 40s ribosomal protein 2</td>
<td>5.7e-140</td>
</tr>
<tr>
<td>BFLG_61E19</td>
<td>80</td>
<td>P38384 canis familiaris (dog), and mus musculus (mouse), protein transport protein sc61 gamma subunit</td>
<td>5.7e-39</td>
</tr>
<tr>
<td>BFLG_131J12</td>
<td>75</td>
<td>P29472 branchiostoma lanceolatum (common lancelet) (amphioxus), cytochrome b</td>
<td>7.3e-62</td>
</tr>
<tr>
<td>BFLG_39C16</td>
<td>70</td>
<td>P46405 sus scrofa (pig), 40s ribosomal protein 12</td>
<td>9.8e-78</td>
</tr>
<tr>
<td>BFLG_10B23</td>
<td>70</td>
<td>Q17133 branchiostoma floridai (amphioxus), essential light chain (myosin light chain alkali)</td>
<td>1.0e-62</td>
</tr>
<tr>
<td>BFLG_18E08</td>
<td>64</td>
<td>Q36148 trachemys scripta, cytochrome c oxidase subunit 1</td>
<td>3.5e-83</td>
</tr>
<tr>
<td>BFLG_69G15</td>
<td>64</td>
<td>O21004 branchiostoma lanceolatum (common lancelet) (amphioxus), atp synthase a chain</td>
<td>6.0e-115</td>
</tr>
<tr>
<td>BFLG_38C15</td>
<td>62</td>
<td>P52821 caenorhabditis elegans, probable 40s ribosomal protein 25</td>
<td>5.8e-56</td>
</tr>
<tr>
<td>BFLG_21E18</td>
<td>62</td>
<td>P32046 gallus gallus (chicken), 60s ribosomal protein 137a</td>
<td>1.9e-46</td>
</tr>
<tr>
<td>BFLG_99K17</td>
<td>61</td>
<td>P39032 rattus norvegicus (rat), 60s ribosomal protein 136</td>
<td>4.0e-49</td>
</tr>
<tr>
<td>BFLG_13K03</td>
<td>60</td>
<td>P33947 homo sapiens (human), er lumen protein retaining receptor 2 (kdel receptor 2)</td>
<td>4.6e-66</td>
</tr>
<tr>
<td>BFLG_116A22</td>
<td>59</td>
<td>Q09510 caenorhabditis elegans, probable myosin regulatory light chain</td>
<td>1.7e-47</td>
</tr>
<tr>
<td>BFLG_61A18</td>
<td>57</td>
<td>Q96175 daphnia pulex, cytochrome c oxidase polypeptide ii</td>
<td>4.4e-93</td>
</tr>
<tr>
<td>BFLG_116I07</td>
<td>57</td>
<td>P03912 xenopus laevis (african clawed frog), nadh-ubiquinone oxidoreductase chain 4</td>
<td>1.0e-70</td>
</tr>
<tr>
<td>BFLG_13OB07</td>
<td>56</td>
<td>P29316 homo sapiens (human), and rattus norvegicus (rat), 60s ribosomal protein I23a</td>
<td>3.5e-47</td>
</tr>
<tr>
<td>BFLG_62N09</td>
<td>51</td>
<td>P38663 homo sapiens (human), and rattus norvegicus (rat), 60s ribosomal protein I24</td>
<td>2.5e-71</td>
</tr>
<tr>
<td>BFLG_62K01</td>
<td>50</td>
<td>P00160 xenopus laevis (african clawed frog), cytochrome b</td>
<td>1.0e-111</td>
</tr>
<tr>
<td>BFLG_15K08</td>
<td>9</td>
<td>Q15392 homo sapiens (human), apoptosis inhibitor survivin.</td>
<td>1.9e-34</td>
</tr>
<tr>
<td>BFLG_67D10</td>
<td>9</td>
<td>Q93131 branchiostoma floridai (florida lancelet) (amphioxus), cytoplasmic actin.</td>
<td>4.3e-119</td>
</tr>
<tr>
<td>BFLG_72K18</td>
<td>46</td>
<td>Q48643 homo sapiens (human), t-complex protein 1, epsilon subunit (tcp-1-epsilon) (cct-epsilon)</td>
<td>2.8e-128</td>
</tr>
<tr>
<td>BFLG_38J21</td>
<td>46</td>
<td>Q93132 branchiostoma floridai (florida lancelet) (amphioxus), muscle actin.</td>
<td>2.4e-105</td>
</tr>
<tr>
<td>Clone name</td>
<td>Cluster size</td>
<td>Abundance</td>
<td>P-value</td>
</tr>
<tr>
<td>------------</td>
<td>-------------</td>
<td>-----------</td>
<td>------------</td>
</tr>
<tr>
<td>BFLG_63G23</td>
<td>9</td>
<td>Q93130</td>
<td>2.8e-128</td>
</tr>
<tr>
<td>BFLG_10020</td>
<td>9</td>
<td>Q25130</td>
<td>6.3e-61</td>
</tr>
<tr>
<td>BFLG_36C10</td>
<td>9</td>
<td>Q04499</td>
<td>2.1e-24</td>
</tr>
<tr>
<td>BFLG_54I15</td>
<td>8</td>
<td>P05217</td>
<td>4.5e-137</td>
</tr>
<tr>
<td>BFLG_37B13</td>
<td>8</td>
<td>P51571</td>
<td>1.6e-53</td>
</tr>
<tr>
<td>BFLG_54I06</td>
<td>7</td>
<td>P15094</td>
<td>1.6e-17</td>
</tr>
<tr>
<td>BFLG_84A11</td>
<td>7</td>
<td>P42577</td>
<td>7.8e-18</td>
</tr>
<tr>
<td>BFLG_4J20</td>
<td>7</td>
<td>P70667</td>
<td>6.0e-96</td>
</tr>
<tr>
<td>BFLG_33I01</td>
<td>singleton</td>
<td>O35047</td>
<td>1.3e-44</td>
</tr>
<tr>
<td>BFLG_31N23</td>
<td>5</td>
<td>O15145</td>
<td>6.9e-88</td>
</tr>
<tr>
<td>BFLG_61L03</td>
<td>10</td>
<td>Q13765</td>
<td>1.0e-83</td>
</tr>
<tr>
<td>BFLG_57D16</td>
<td>253</td>
<td>P11831</td>
<td>2.8e-66</td>
</tr>
<tr>
<td>BFLG_62O11</td>
<td>2</td>
<td>P54843</td>
<td>1.0e-16</td>
</tr>
<tr>
<td>BFLG_71N01</td>
<td>2</td>
<td>P20290</td>
<td>7.0e-66</td>
</tr>
<tr>
<td>BFLG_10E03</td>
<td>3</td>
<td>P50901</td>
<td>3.0e-15</td>
</tr>
<tr>
<td>BFLG_63L12</td>
<td>singleton</td>
<td>U85970</td>
<td>1.1e-14</td>
</tr>
<tr>
<td>BFLG_66O23</td>
<td>singleton</td>
<td>Q29426</td>
<td>7.9e-17</td>
</tr>
<tr>
<td>BFLG_10I19</td>
<td>singleton</td>
<td>A223581</td>
<td>5.9e-200</td>
</tr>
<tr>
<td>BFLG_49K20</td>
<td>singleton</td>
<td>P08729</td>
<td>4.0e-47</td>
</tr>
<tr>
<td>BFLG_115P13</td>
<td>10</td>
<td>O62007</td>
<td>4.4e-16</td>
</tr>
<tr>
<td>BFLG_103C01</td>
<td>3</td>
<td>A223581</td>
<td>5.9e-202</td>
</tr>
<tr>
<td>BFLG_123P22</td>
<td>2</td>
<td>A223581</td>
<td>3.5e-110</td>
</tr>
<tr>
<td>113L16</td>
<td>2</td>
<td>O60496</td>
<td>8.4e-17</td>
</tr>
</tbody>
</table>
Table 5.3. Differences in the expression level of transcripts based on the size of the fingerprint cluster.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Clone coordinates</th>
<th>Fp cluster size</th>
</tr>
</thead>
<tbody>
<tr>
<td>mitochondrial 12S and 16S ribosomal RNA genes</td>
<td>BFL26</td>
<td>BFLG</td>
</tr>
<tr>
<td>Q93132 (amphioxus). Muscle actin.</td>
<td>64I23</td>
<td>815</td>
</tr>
<tr>
<td>P06603 drosophila melanogaster tubulin alpha-1 chain</td>
<td>41G05</td>
<td>316</td>
</tr>
<tr>
<td>Q25052 heliothisodes diminutivus. elongation factor 1 alpha</td>
<td>37E03</td>
<td>314</td>
</tr>
<tr>
<td>O47434 (amphioxus), nadh dehydrogenase subunit 2</td>
<td>40G14</td>
<td>287</td>
</tr>
<tr>
<td>P05123 canis familiaris (dog). creatine kinase, m chain</td>
<td>50E20</td>
<td>190</td>
</tr>
<tr>
<td>O47431 branchiostoma floridai (amphioxus). Cytochrome b</td>
<td>109D23</td>
<td>175</td>
</tr>
<tr>
<td>Q17133 (amphioxus). Myosin, essential light chain</td>
<td>65O19</td>
<td>115</td>
</tr>
<tr>
<td>P97351mus musculus (mouse). ribosomal protein s3a.</td>
<td>120015</td>
<td>101</td>
</tr>
<tr>
<td>P25444 mus musculus (mouse). 40s ribosomal protein s2</td>
<td>20K12</td>
<td>94</td>
</tr>
<tr>
<td>O42248 brachydanio rario guanine nucleotide-binding protein beta</td>
<td>76J23</td>
<td>88</td>
</tr>
<tr>
<td>subunit-like protein (receptor of activated protein kinase c) (rack).</td>
<td></td>
<td></td>
</tr>
<tr>
<td>002367 ciona intestinalis. Calmodulin. 6/98</td>
<td>94G02</td>
<td>68</td>
</tr>
<tr>
<td>P12829 human. myosin light chain 1, embryonic muscle/atrial isoform.</td>
<td>130M06</td>
<td>56</td>
</tr>
<tr>
<td>(amphioxus). Homeobox protein dll homolog.</td>
<td>16N03</td>
<td>13D02</td>
</tr>
<tr>
<td>U82487 Branchiostoma floridai engrailed protein (AmphiEn) mRNA</td>
<td>3P11</td>
<td>-</td>
</tr>
<tr>
<td>P22810 drosophila melanogaster homeotic protein orthodenticle</td>
<td>27L08</td>
<td>58I15</td>
</tr>
<tr>
<td>P20290 human transcription factor btf3</td>
<td>101J13</td>
<td>1L11</td>
</tr>
<tr>
<td>P54864 serinus canaria transcription factor ap-1 (proto-oncogene c-jun).</td>
<td>93P22</td>
<td>66M10</td>
</tr>
<tr>
<td>Q63836 mus musculus liver protein, selenium binding</td>
<td>129M21</td>
<td>66E20</td>
</tr>
<tr>
<td>Q61123 mus musculus maternal embryonic message 3 (mem3).</td>
<td>84C03</td>
<td>-</td>
</tr>
<tr>
<td>X68045 B. floridai gene for amphihox3</td>
<td>8E4</td>
<td>3</td>
</tr>
<tr>
<td>X15243 Xenopus maternal G10 mRNA. 9/93 Length = 777</td>
<td></td>
<td></td>
</tr>
<tr>
<td>O15394 homo sapiens neural cell adhesion protein.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
As the components of whole pathways are identified it is tempting to put the pieces together. Below, I try to assemble the components of the nervous system, mostly the genes encoding molecules that transmit signal between neurons for which little is known. The list is by far not complete. However, it gives a first glimpse and an opportunity for a first comparison with the homologous genes from C.elegans (Bargmann, 1998) which will indicate which members of the families have been conserved.

5.1. Unwiring the amphioxus nervous system

Little is known about the molecular nature of the nervous system of amphioxus. Electron microscope studies have shown that the anterior part of the nerve cord in Branchiostoma floridae is partially homologous to the vertebrate brain in the sense that it is organised in a brain-like manner containing three main regions: the anterior and posterior part of the cerebral vesicle and the primary motor center. Moreover, the isolation and expression studies of genes which are homologues of vertebrate genes expressed in organs of the anterior nerve cord has indicated the similarities and differences between amphioxus and higher vertebrates.

Thus AmphiOtx expression underlines the frontal eye region in Amphioxus and corresponds to the expression of vertebrate Otx genes in the eyes (Williams and Holland, 1996). Pax-6 (Glardon et al., 1998) is detected in two of the four types of amphioxus photoreceptors, the lamellar organ and the frontal eye. AmphiDll (Distal-less) is expressed in the anterior cerebral vesicle in a region suggested to be homologous to vertebrate forebrain (Holland et al., 1996).

In addition, during this study homologues of the mouse basic domain leucine-zipper (bZIP)-type transcription factors maf2 (c-maf) and ets-1 have been isolated from the gastrula stage library. The gene disrupted in the recessive mouse mutation kreasler which affects hindbrain segmentation encodes for a bZIP transcription factor of the Maf family (Cordes and Barsh, 1994). Mouse Ets-1 is expressed initially (day 8.0 of gestation) in the developing nervous system including the presumptive hindbrain, neural tube, neural crest and the first and second branchial arches. Later, ets-1 expression is described in developing vascular structures, including the heart, arteries, capillaries and meninges (Maroulakou et al., 1994).
Transmission of nerve impulse

*Neurotransmitters* are the molecules which by diffusing at the junction between two neurons, the synapse, transmit an impulse from cell to cell. The most widespread excitatory neurotransmitters outside the central nervous system are acetylcholine and the catecholamines, adrenaline and noradrenaline. In the central nervous system the aminoacids glutamate, aspartate, glycine and taur inne and the amines, dopamine, 5-hydroxytryptamine, 4-aminobutyric acid and histamine function as neurotransmitters.

When the membrane of a cell becomes depolarised due to the arrival of an impulse the neurotransmitter is discharged between the two cells(synaptic cleft) This causes the depolirisation of the postsynaptic membrane(by changing its permeability to sodium ions) thus initiating an impulse in the second cell (excitatory neurotransmitters). Other neurotransmitters decrease the ease by which the post-synaptic membrane can be depolirized and they are known as inhibitory neurotransmitters. *Neurotransmitter gated ion channels* are transmembrane receptors that transiently form an ionic channel after the binding of a specific neurotransmitter. These channels have the advantage that they open quickly and they desensitise rapidly which allows fast signal transmission as opposed to G-coupled receptors that generate slower but longer lasting response. G-coupled receptors have seven transmembrane domains which are linked with a cytoslasmic heterotrimeric G protein. Finally, it is important that the neurotransmitter is removed as soon as the signal is transmitted. This can be achieved by uptake of the neurotransmitter back into the presynaptic or postsynaptic neurone by the neurotransmitter transporters or by conversion of the neurotransmitter into an inactive metabolite through its enzymatic catabolism.

During this study members from the majority of gene families implicated in neural transmission are isolated (Table 5.4.). All the families consist of multiple components e.g. there are three major classes of potassium channels, the inward rectifier channels with two transmembrane domains (TM), the two pore/TWIK channels with 4TM and the voltage-regulated potassium channels with 6TM.

Among the identified genes are:

One amphioxus inward potassium channel showing higher similarity score to human, expressed at the neurula stage.

One type of inhibitory neurotransmitter transporter expressed at the neurula stage showing higher similarity to drosophila transporter *ine* (*inebriated*)/rosA gene is identified. Mutations in *Drosophila ine* confer increased excitability of the motor neuron.

No member of the mechanosensory channel gene family is identified so far e.g from the degenerin family. There are two *C.elegans* degenerin genes, the mec-4 and mec-10 required
for touch sensation which when mutated, they give dominant hyperactive channels that cause the touch cells to swell and degenerate (Driscoll et. al. 1991). However, a transcript giving higher similarity score to the erythrocyte band 7.2b integral membrane protein stomatin is identified in amphioxus. The *C.elegans mec-2* genes which is also implicated in touch sensation is related to the above protein (Huang et al., 1995).

Finally, it is worth drawing the attention to the *glutamate receptor* identified in the neurula stage library. L-glutamate acts as an excitatory neurotransmitter. Its action is mediated by a range of receptors named according to their agonists. The amphioxus identified receptor as implied by its highest score to rat glutamate receptor, ionotropic kainate 2, binds kainate and it may be involved in the transmission of light from the retina to the hypothalamus.

Table 5.4. Identified genes in amphioxus that affect the function of the nervous system as implied by their similarity to the vertebrate or invertebrate corresponding genes.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Embryonic stage</th>
<th>Cluster size/ Abundance</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Neurotransmitter transporters</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dros inebriated (ine/roA)-homologue of human BGT-1 (GABA/betaine transporter)</td>
<td>050M17 BFL26</td>
<td>singleton</td>
</tr>
<tr>
<td>vesicular acetylcholine transporter</td>
<td>095J21 BFL26</td>
<td>2</td>
</tr>
<tr>
<td>sodium- and chloride-dependent glycine transporter 1</td>
<td>013K01 BFL26</td>
<td>2</td>
</tr>
<tr>
<td><strong>Neurotransmitter-gated ion channels</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>acetylcholine receptor like protein, alpha-type precursor (non-alpha-3 chain)</td>
<td>064P22 BFL26</td>
<td>3</td>
</tr>
<tr>
<td>neuronal acetylcholine receptor protein, beta-4 chain precursor (non-alpha-3 chain)</td>
<td>042M14 BFLG</td>
<td>3</td>
</tr>
<tr>
<td>neuronal acetylcholine receptor protein, alpha-3 chain precursor (non-alpha-3 chain)</td>
<td>089A21 BFL26</td>
<td>5</td>
</tr>
<tr>
<td>neuronal acetylcholine receptor protein, alpha-9 chain precursor (non-alpha-3 chain)</td>
<td>075J03 BFL26</td>
<td>4</td>
</tr>
<tr>
<td>neuronal acetylcholine receptor protein, alpha-6 chain precursor (non-alpha-3 chain)</td>
<td>089A21 BFL26</td>
<td>5</td>
</tr>
<tr>
<td>serotonin-gated ion channel receptor</td>
<td>075J03 BFL26</td>
<td>4</td>
</tr>
<tr>
<td>serotonin transporter</td>
<td>050M17 BFL26</td>
<td>singleton</td>
</tr>
<tr>
<td>glutamate receptor 6 precursor (glutamate receptor, ionotropic kainate 2)</td>
<td>053H18 BFL26</td>
<td>singleton</td>
</tr>
<tr>
<td>cyclic nucleotide-gated channel beta subunit</td>
<td>070D22 BFL26</td>
<td>2</td>
</tr>
<tr>
<td><strong>G-coupled receptors</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G protein-coupled receptor</td>
<td>035K22 BFLG</td>
<td>2</td>
</tr>
<tr>
<td>dopamine receptor</td>
<td>062G02 BFLG</td>
<td>3</td>
</tr>
<tr>
<td><strong>Ion channels</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dros. putative potassium channel subunit (elk)</td>
<td>60E18 BFL26</td>
<td>7</td>
</tr>
<tr>
<td>chloride intracellular channel protein 2</td>
<td>128E18 BFL26</td>
<td>3</td>
</tr>
<tr>
<td>calcium entry channel 1</td>
<td>020I07 BFLG</td>
<td>4</td>
</tr>
<tr>
<td>inward rectifier potassium channel 4</td>
<td>013K03 BFL26</td>
<td>7</td>
</tr>
<tr>
<td><strong>Enzymes catalysing neurotransmitter synthesis</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dopamine beta- hydroxylase</td>
<td>028D03 BFL26</td>
<td>3</td>
</tr>
<tr>
<td>acetylcholinesterase precursor</td>
<td>010E04 BFL26</td>
<td>2</td>
</tr>
<tr>
<td>aspartate aminotransferase</td>
<td>051J24 BFL26</td>
<td>7</td>
</tr>
</tbody>
</table>

No member of the gap junction gene family has been identified so far. It is interesting that
the gap junction genes found in invertebrates have not been identified in vertebrates where there is the connexin gap junction gene family. This implies that the gap junctions in vertebrates and invertebrates have arisen independently. In order to investigate this, specific hybridisations with 4 innexins, the *Shaking-B neural*, *Shaking-B lethal*, *Innexin I* and *Innexin II* and *Connexin 26, 32 and 43* were carried out on both libraries. The results of the hybridisations are currently analysed.

It is striking that with very few exceptions the majority of the amphioxus identified genes implicated in the regulation of nervous system show higher similarity score to related vertebrate genes. Moreover, the majority of those genes are expressed at the neurula stage. The above conclusion though has to be re-evaluated taking into consideration the co-clustering results of both libraries. It might be that short transcripts from the gastrula stage library have not been recognised due to their short length which covers mostly the 3' UTR which varies even between members of the same gene family.

The fact that more components of the neural signal transmission pathway might be present at earlier stage is also pointed out by presence of *acetylcholine* and *dopamine* receptors, *calcium channels* and *G-coupled receptors* at the gastrula stage.

5.2. Conclusions

The 5' tag sequencing of only 3,870 clones from both amphioxus cDNA libraries gives already an indication of the spectrum of genes that are transcribed in the gastrula and neurula stages and in extension the structures that have formed. Components of whole pathways have been identified. As an example I have described transcripts involved in the transmission of signal between neurons for which little is known.
Chapter 6

Discussion

6.1. Conclusions

Of the estimated 70,000 -100,000 vertebrate genes, the majority is represented by a human or mouse EST sequence. However, a great proportion of those sequences remain unidentified.

On the other hand, as more sequence and functional data become available from multiple organisms, comparisons across species is proved to be a valuable tool for identifying novel genes and elucidating their function. The major bulk though of information is derived from the traditional model organisms, D.melanogaster, C.elegans, Xenopus, zebrafish and mouse. These organisms offer the possibility of functional analysis through mutagenesis screens, whole mount in situ expression screens, targeted gene knockouts etc. However, in the case of vertebrate model organisms, the presence of multiple paralogs that substitute for each other makes difficult the interpretation of data. Moreover, emerging data suggests that some of the vertebrate lineages have undergone additional lineage specific gene duplications which makes difficult the correlation of data. Finally, the great evolutionary distance between the well-studied invertebrate organisms, C. elegans and D. melanogaster and vertebrates, results in limited sequence similarity. More than 35% of the predicted proteins of the recently sequenced C. elegans are thought to be nematode specific.

Therefore, expansion of studies to organisms that are not considered as model will increase the spectrum of constructive comparisons. For example the study of organisms that occupy the transition point between vertebrates and invertebrates will offer additional information on the mechanisms of genome expansion.

Cephalochochordates are considered as the last invertebrate chordate sub-phylum separated from vertebrates before the double genome expansion for which molecular evidence suggests that occurred along the vertebrate lineage. Therefore, amphioxus’ 600 Mb genome is estimated to contain a quarter of the number of vertebrate genes. Expression studies of all amphioxus genes isolated so far indicate that they are expressed in structures homologous, although simplified, to the vertebrate ones, implying that their function is
conserved.

The aim of this study was to give access to all amphioxus genes even those that are expressed at very low level and in a very short time window. Oligonucleotide fingerprinting, a normalisation method based on the clustering of clones according to their sequence based hybridisation pattern with short oligonucleotides, was used. Two amphioxus embryonic cDNA libraries (110,000 clones) were arrayed and oligonucleotide fingerprinted. The result of this study is not only a catalogue of amphioxus genes expressed at gastrula and neurula stages but also an evaluation of the oligonucleotide fingerprinting method.

As a result of this study (conclusions are listed more extensively at the end of each chapter):

• a non-redundant catalogue of amphioxus genes expressed in each of the two embryonic stages, gastrula and neurula stages was constructed:

⇒ 8,968 different transcripts were recognised in the gastrula stage and 10,702 in the neurula stage. This is the first ever published estimation on the number of genes present at these embryonic stages.

⇒ the extra information of the estimation of the expression level is attached to each of those transcripts

⇒ 4,079 5’ tag sequences were generated from 3,870 cDNA clones picked from both embryonic stages. This has led to the identification of 2,467 unique transcripts, of which 28% matched a sequence deposited in any of the GenEmbl, Swplus, UniGene(NCBI)-Human/Mouse/Rat databases. Analytically, 38% BFL26 sequences and 20.7% of the BFLG sequences gave similarity score higher than 1.0e-10. These sequences already show which genes are expressed at the respective stages.

⇒ 74.9% of BFL26 and 51.7% of the BFLG sequences are predicted to be coding with either the GENSCAN or MZEF prediction programs. Considering that both programs showed a false negative prediction rate of approximately 20%, it is possible that higher percentage of the above ESTs contain coding region.

⇒ finally, the sequences that gave a significant similarity score with the swissprot database were classified in functional classes based on their highest score match. Thus, 38.5% (325 sequences) of all transcripts are associated with basic metabolism (Energy class), 22.27% with gene/protein expression (Information class) and 24.7% with signalling (Communication class).
• It was demonstrated with specific examples the already mentioned advantages of oligonucleotide fingerprinting. In addition, the parameters that affect reliable clustering were determined. Finally, for the first time the extent of redundancy that still exists in the selection of clones based on the fingerprinting clustering was investigated. Thus it was demonstrated that:

  ⇒ up to 4 fold normalisation is achieved.
  ⇒ the information on the abundance of a specific transcript within the screened clones which in extent reflects its abundance in the tissue from which the library, is a reliable estimate.
  ⇒ fingerprints can be compared reliably across multiple libraries. Therefore, the already sequenced transcripts can be recognised within every new fingerprinted library and excluded from further sequencing. This results in additional reduction in the sequencing costs.
  ⇒ the specific set of oligonucleotides used, can partition effectively the amphioxus sequences.
  ⇒ clones with insert size smaller than 500 bp are difficult to be clustered using the same number of oligonucleotides as in this study.
  ⇒ experimental fingerprints can be compared with the theoretical fingerprints of a database of sequences. Thus the identity of cluster can be assigned without sequencing.

• and through the selective tag sequencing

  ⇒ multiple members of the same gene family. For example Wnt2, Wnt3, Wnt6, Wnt7a, Wnt7b, Wnt8.
  ⇒ disease-related genes have been identified

The results of this study constitute a tremendous resource not only for the amphioxus research field. They give the opportunity for the first time to a broader spectrum of people to have an overview of a large number of sequences of amphioxus genes, an organism that is considered to be the closest invertebrate relative of vertebrates.

The results of this study enable answering questions such as how many and which genes are present in the gastrula and neurula stages of amphioxus which implies which structures are present at this time of development or even at all in amphioxus, what is the estimated expression level of those genes, are they vertebrate or invertebrate like or which genes have undergone duplication before the divergence of amphioxus and vertebrates.
Moreover, since the libraries are arrayed, the clones can be distributed thus giving the chance to more people to have access to such valuable material. There are very few people currently working exclusively on amphioxus, mainly due to the difficulty in obtaining embryos or tissues as amphioxus cannot be kept in culture and breeds only during a period of two months every year.

All results mentioned will be available through RZPD (Resource Zentrum Primary Database) and an amphioxus project web page. As it was shown in Fig. 3.10 this page will include:

- for each tag sequence, the sequence cluster it belongs and its size as well as the sequence alignment for all the sequences in this cluster,
- the fingerprint cluster of every clone that has been tag sequenced. Thus, it will be possible to compare whether sequences that belong in the same sequence cluster are split in multiple fingerprint clusters. Moreover, it will be possible to find which oligonucleotides are positive for the majority of the clones in the cluster. These oligonucleotides could be used for internal sequencing of any clone of the cluster thus reducing the amount of time required for full length sequencing of a clone. It has been reported that tailed octamer primers can be used in cycle sequencing (Ball et al., 1998). Moreover, the longer clone of a cluster could be selected for sequencing.
- the three highest score matches against each of the Swissprot, Genembl and dbEST databases along with the respective alignments.

6.2. Future Work

The results of the work described in this thesis have already assisted a number of studies. The isolation of additional transcripts of the intermediate filament (IF) protein gene family will show whether the organisation of this gene family in amphioxus is similar to vertebrates (in collaboration with Prof. K. Weber, Max-Planck Institute for Biophysical Chemistry, Goettingen) (Erber et al., 1998; Luke and Holland, 1999). Expression studies of the identified during this study, homologues of the mouse basic domain leucine-zipper (bZIP)-type transcription factors maf2 (c-maf) and ets-1 will shed light into the evolution of the vertebrate hindbrain (in collaboration with Bill Jackman) and neural crest. The Wnt (wingless and int (Nusse et al., 1991) ) gene family consist of multiple members, some of which have
duplicated in specific lineages, with very diverse roles (Wodarz and Nusse, 1998). During this study, transcripts of \textit{Wnt2}, \textit{Wnt3}, \textit{Wnt6}, \textit{Wnt7a}, \textit{Wnt7b}, \textit{Wnt8}, have been isolated. This data provides new evidence that the duplication of \textit{Wnt7} occurred before the divergence of amphioxus from the vertebrates (Sidow, 1992). (expression studies will be carried out in collaboration with Linda Holland, Scripps Institution of Oceanography, La Jolla).

Besides, the study of the amphioxus transcripts for which there is already some indication for their function through their similarity score with a gene previously identified in another organism, the expression study of the transcripts that are unmatched is of great interest. An amphioxus \textit{in situ} hybridisation screen will be organised in order to study the expression of those transcripts.

As it was mentioned in the introduction (see 1.1.3.), increase in the gene number is thought to have occurred either by tandem gene or whole chromosome area duplications (Ohno, 1970). The study of the genomic organisation of genes that have undergone multiple duplications provides evidence on the mechanism responsible for this duplication. The homeotic \textit{Hox} genes are the best example of genes for which there is extensive information on their genomic organisation across multiple species. It has therefore been observed that besides the conservation in the order of the \textit{Hox} genes within a cluster, there is a conservation in the order of genes surrounding the clusters (Bentley \textit{et al.}, 1993; Lundin, 1993; Rabin \textit{et al.}, 1986). (Ruddle \textit{et al.}, 1994) have tabulated all the mouse genes that are members of gene families and of which at least one member maps to the chromosomes where the \textit{Hox} clusters are located, the chromosomes 2, 6, 11 and 15. Seventy four families, 323 genes in total were identified. Amphioxus transcripts showing high similarity score to many of the genes mentioned in (Ruddle \textit{et al.}, 1994), have been identified during this study. For example \textit{acetylcholine receptor}, \textit{crystallin}, \textit{glutamate receptor}, \textit{Wnt2}, \textit{Wnt3} etc. Phylogenetic analysis of the identified transcripts will indicate whether they are homologs to the mapped mouse genes. It will be interesting to investigate whether those genes are linked with the single amphioxus \textit{Hox} cluster (Garcia-Fernandez and Holland, 1994).

However, there are additional to \textit{Hox} genes cases where there is conserved synteny. These regions might be even more interesting to test since in the case of \textit{Hox} clusters it might be argued that the extensive conservation within and around the cluster is due to the important role of the \textit{Hox} genes in body organisation.

Three known members of the retinoic acid receptor map to human chromosomes 1,9 and 6 (Almasan \textit{et al.}, 1994; Fitzgibbon \textit{et al.}, 1993). The same organisation is present in mouse
as well. Thus the above genes map to chromosomes 2, 17 and 1. (Katsanis et al., 1996) observed that these are within larger regions of synteny between human and mouse which implies that this region triplicated within an ancestral genome prior to human-mouse divergence. Thus, they were able to predict the presence of gene family members within these paralogous regions. e.g. Notch.

In the meanwhile, the large-scale mapping of human and mouse ESTs has generated additional information (Boguski and Schuler, 1995). It will be interesting to blast all amphioxus generated ESTs against 'The gene map of the human genome' (http://www.ncbi.nlm.nih.gov/genemap99/) or 'The human mouse homology maps' at NCBI (http://www.ncbi.nlm.nih.gov/Homology/) databases which contain all the mapping information. Amphioxus transcripts that match human or mouse ESTs that have been mapped very close on the same chromosome will be selected for hybridisation on the amphioxus genomic library (Burgtorf et al., 1998).

The results of this thesis, besides being the incitement for future work on amphioxus, they can serve as the basis for future amphioxus oligonucleotide fingerprinting projects.

As it was shown in the example of the amphioxus cholinesterase gene, the occurrence frequency of oligonucleotides differs. Based on hybridisation results, only the oligonucleotides that have higher occurrence frequency and therefore are more informative (Appendix I and II) could be selected for future hybridisations. However, the lower occurrence frequency of some oligonucleotides as implied by the hybridisation data can be also attributed to the different hybridisation behavior of the oligonucleotides due to their different sequence characteristics. Finally, it has been observed during this study that false positives do occur (noise) under the hybridisation conditions used. The hybridisation conditions that were used in this study as well as in others were identical and standardised for all oligonucleotides, which is important in high throughput project. An alternative approach where the oligonucleotides are divided in groups according to their sequence characteristics and hybridised under different hybridisation conditions like temperature or washing time, could be tested. It is possible to reduce even more the hybridisation temperature from 4° to 0°C without causing precipitation of the hybridisation buffer, or increasing the washing time for example for the GC rich oligonucleotides, or including tetra-alkyl ammonium salts to reduce the difference in the effect in the duplex stability of A-T or G-C regions. Moreover, it is possible to increase the hybrid stability or specificity by using universal bases. Inosine unlike guanosine, forms two hydrogen bonds instead of three with cytosine, thus lowering the Tm of the hybrid (Varshney et al., 1988). Moreover, the universal
base containing 5-nitroindole is shown to base pair with more equal affinity than inosine, with all four natural bases (Loakes et al., 1997).

Finally, the non-radioactive labeling of oligonucleotides with digoxigenin (DIG) will increase the throughput of oligonucleotide fingerprinting (Maier et al., 1994a). A time-consuming step in the oligonucleotide hybridisation process is the handling of the radioactive filters. The combination of DIG labeling with the plastic backed (laminated) membranes, will facilitate the handling of the filters and it will increase the number of hybridisations that can be performed per day. So far random primed or PCR DIG-dUTP labeled probes are routinely hybridised on PAC libraries (Bancroft et al., 1997).

Furthermore, additional oligonucleotides could be designed specifically after amphioxus sequences based on the amphioxus 5' ESTs. This has to be tested though, since the ESTs are single pass partial sequences and they have some percentage of errors (substitutions, insertions, deletions).

Finally, hybridisation of oligonucleotides that have been designed after conserved domains of gene families is an approach that has been already tested. Approximately 30 short oligonucleotides designed after the TGF-β family (Wiles, M., unpublished) have been hybridised on both amphioxus cDNA libraries. The results of this study are being currently analysed. The example of the oligonucleotides found positive for the amphioxus BMP2/4 (see 3.1.1.2) include some of the TGF-β oligonucleotides.

Finally, a main area of this project, for future developments is the analysis and interpretation of sequences.

• The frequency of errors (substitutions, insertions, deletions) in the generated ESTs will be estimated by comparing the 219 amphioxus ESTs (see 4.2.) that gave high similarity score against the corresponding amphioxus full length genes.

• The codon preference will be estimated based on the large number of sequences that were generated. Current estimation (http://www.dna.affrc.go.jp/~nakamura/CUTG.html) is based only on the 160 sequences currently deposited in the Genbank database.

• Alternative sequence clustering approaches will be tested. For example clustering the sequences based on BLAST similarity search against each other (Casey et al., 1998).

• Finally, all function prediction programs were based on the assumption that since sequence is the major factor that determines the gene function, sequence similarity implies function similarity. Although this is true to a great extent, there have been many groups of genes reported with similar sequence that have diverse (although related) function e.g. gene duplications are accompanied by functional divergence although the resulting paralogs share
some sequence features which are characteristic of the general gene family. Thus it is necessary to evaluate the level of similarity between a gene with the unknown function and genes of similar sequence in the database with different function before function prediction.

One method proposed (Eisen, 1998) is to focus on how the genes became similar in sequence through evolution instead of using sequence similarity only (Phylogenomics). The initial step in the above approach is the generation of a phylogenetic tree that represents the evolutionary history of the gene of interest and its homologs and then biologically determined functions of the homologs are overlaid on the tree. Finally, the function of the uncharacterised gene is predicted based on the history of functional changes of the homologs as inferred by the tree.
Appendices

APPENDIX I

List of the sequences of the oligonucleotides hybridised against both the BFL26 and BFLG cDNA libraries, 110,000 clones in total.
APPENDIX II

List of the oligonucleotides that were included (oligos (I)) and those rejected (oligos (II)) from the clustering analysis of the neurula stage library (BFL26). Set 1 filters were spotted with the PCR products of the clones in plates 1-66, while set 2 filters with those in plates 67-132.
<table>
<thead>
<tr>
<th>BFL26</th>
<th>Set 1 filters oligos(I)</th>
<th>Set 2 filters oligos(I)</th>
<th>Set 1 filters oligos (II)</th>
<th>Set 2 filters oligos (II)</th>
</tr>
</thead>
<tbody>
<tr>
<td>o100</td>
<td>o194</td>
<td>o100</td>
<td>o124</td>
<td>o118</td>
</tr>
<tr>
<td>o101</td>
<td>o199</td>
<td>o101</td>
<td>o129</td>
<td>o124</td>
</tr>
<tr>
<td>o102</td>
<td>o201</td>
<td>o204</td>
<td>o130</td>
<td>o129</td>
</tr>
<tr>
<td>o103</td>
<td>o202</td>
<td>o205</td>
<td>o131</td>
<td>o120</td>
</tr>
<tr>
<td>o104</td>
<td>o203</td>
<td>o207</td>
<td>o132</td>
<td>o131</td>
</tr>
<tr>
<td>o105</td>
<td>o204</td>
<td>o208</td>
<td>o133</td>
<td>o132</td>
</tr>
<tr>
<td>o106</td>
<td>o205</td>
<td>o212</td>
<td>o138</td>
<td>o133</td>
</tr>
<tr>
<td>o107</td>
<td>o206</td>
<td>o213</td>
<td>o141</td>
<td>o138</td>
</tr>
<tr>
<td>o108</td>
<td>o207</td>
<td>o215</td>
<td>o144</td>
<td>o141</td>
</tr>
<tr>
<td>o109</td>
<td>o208</td>
<td>o217</td>
<td>o145</td>
<td>o143</td>
</tr>
<tr>
<td>o110</td>
<td>o212</td>
<td>o219</td>
<td>o147</td>
<td>o144</td>
</tr>
<tr>
<td>o111</td>
<td>o213</td>
<td>o220</td>
<td>o156</td>
<td>o156</td>
</tr>
<tr>
<td>o112</td>
<td>o215</td>
<td>o223</td>
<td>o165</td>
<td>o165</td>
</tr>
<tr>
<td>o113</td>
<td>o217</td>
<td>o231</td>
<td>o166</td>
<td>o167</td>
</tr>
<tr>
<td>o114</td>
<td>o219</td>
<td>o232</td>
<td>o167</td>
<td>o169</td>
</tr>
<tr>
<td>o115</td>
<td>o220</td>
<td>o235</td>
<td>o169</td>
<td>o176</td>
</tr>
<tr>
<td>o116</td>
<td>o223</td>
<td>o236</td>
<td>o175</td>
<td>o179</td>
</tr>
<tr>
<td>o117</td>
<td>o225</td>
<td>o237</td>
<td>o176</td>
<td>o182</td>
</tr>
<tr>
<td>o119</td>
<td>o231</td>
<td>o238</td>
<td>o179</td>
<td>o186</td>
</tr>
<tr>
<td>o120</td>
<td>o232</td>
<td>o240</td>
<td>o182</td>
<td>o191</td>
</tr>
<tr>
<td>o121</td>
<td>o235</td>
<td>o241</td>
<td>o184</td>
<td>o193</td>
</tr>
<tr>
<td>o122</td>
<td>o236</td>
<td>o242</td>
<td>o185</td>
<td>o196</td>
</tr>
<tr>
<td>o123</td>
<td>o237</td>
<td>o243</td>
<td>o186</td>
<td>o197</td>
</tr>
<tr>
<td>o125</td>
<td>o238</td>
<td>o250</td>
<td>o187</td>
<td>o198</td>
</tr>
<tr>
<td>o126</td>
<td>o241</td>
<td>o251</td>
<td>o188</td>
<td>o199</td>
</tr>
<tr>
<td>o127</td>
<td>o243</td>
<td>o253</td>
<td>o189</td>
<td>o200</td>
</tr>
<tr>
<td>o128</td>
<td>o250</td>
<td>o254</td>
<td>o190</td>
<td>o202</td>
</tr>
<tr>
<td>o134</td>
<td>o253</td>
<td>o255</td>
<td>o191</td>
<td>o203</td>
</tr>
<tr>
<td>o135</td>
<td>o254</td>
<td>o256</td>
<td>o192</td>
<td>o206</td>
</tr>
<tr>
<td>o136</td>
<td>o255</td>
<td>o257</td>
<td>o193</td>
<td>o209</td>
</tr>
<tr>
<td>o137</td>
<td>o256</td>
<td>o258</td>
<td>o195</td>
<td>o210</td>
</tr>
<tr>
<td>o139</td>
<td>o257</td>
<td>o259</td>
<td>o196</td>
<td>o211</td>
</tr>
<tr>
<td>o140</td>
<td>o258</td>
<td>o260</td>
<td>o197</td>
<td>o214</td>
</tr>
<tr>
<td>o142</td>
<td>o259</td>
<td>o262</td>
<td>o198</td>
<td>o216</td>
</tr>
<tr>
<td>o143</td>
<td>o260</td>
<td>o263</td>
<td>o209</td>
<td>o218</td>
</tr>
<tr>
<td>o146</td>
<td>o261</td>
<td>o266</td>
<td>o210</td>
<td>o222</td>
</tr>
<tr>
<td>o148</td>
<td>o262</td>
<td>o267</td>
<td>o211</td>
<td>o224</td>
</tr>
<tr>
<td>o149</td>
<td>o266</td>
<td>o268</td>
<td>o214</td>
<td>o226</td>
</tr>
<tr>
<td>o150</td>
<td>o267</td>
<td>o271</td>
<td>o216</td>
<td>o227</td>
</tr>
<tr>
<td>o151</td>
<td>o268</td>
<td>o275</td>
<td>o218</td>
<td>o228</td>
</tr>
<tr>
<td>o152</td>
<td>o276</td>
<td>o253</td>
<td>o221</td>
<td>o229</td>
</tr>
<tr>
<td>o153</td>
<td>o278</td>
<td>o254</td>
<td>o222</td>
<td>o230</td>
</tr>
<tr>
<td>o154</td>
<td>o279</td>
<td>o255</td>
<td>o224</td>
<td>o233</td>
</tr>
<tr>
<td>o155</td>
<td>o281</td>
<td>o257</td>
<td>o226</td>
<td>o234</td>
</tr>
<tr>
<td>o157</td>
<td>o298</td>
<td>obz</td>
<td>o228</td>
<td>o239</td>
</tr>
<tr>
<td>o158</td>
<td>o59</td>
<td>ocz</td>
<td>o229</td>
<td>o245</td>
</tr>
<tr>
<td>o159</td>
<td>o97</td>
<td>o60</td>
<td>o230</td>
<td>o246</td>
</tr>
<tr>
<td>o160</td>
<td>o98</td>
<td>o61</td>
<td>o233</td>
<td>o247</td>
</tr>
<tr>
<td>o161</td>
<td>o99</td>
<td>o62</td>
<td>o234</td>
<td>o248</td>
</tr>
<tr>
<td>o162</td>
<td>o63</td>
<td>o64</td>
<td>o239</td>
<td>o249</td>
</tr>
<tr>
<td>o163</td>
<td>ocz</td>
<td>o65</td>
<td>o240</td>
<td>o252</td>
</tr>
<tr>
<td>o164</td>
<td>o66</td>
<td>ot10</td>
<td>o242</td>
<td>o261</td>
</tr>
<tr>
<td>o168</td>
<td>o68</td>
<td>ot11</td>
<td>o244</td>
<td>o264</td>
</tr>
<tr>
<td>o170</td>
<td>o70</td>
<td>ot12</td>
<td>o245</td>
<td>o265</td>
</tr>
<tr>
<td>o171</td>
<td>o71</td>
<td>ot13</td>
<td>o246</td>
<td>o269</td>
</tr>
<tr>
<td>o172</td>
<td>orap</td>
<td>ot14</td>
<td>o247</td>
<td>o270</td>
</tr>
<tr>
<td>o173</td>
<td>ot10</td>
<td>o248</td>
<td></td>
<td></td>
</tr>
<tr>
<td>o174</td>
<td>ot11</td>
<td>o249</td>
<td></td>
<td></td>
</tr>
<tr>
<td>o177</td>
<td>ot12</td>
<td>ot17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>o178</td>
<td>ot13</td>
<td>o263</td>
<td></td>
<td></td>
</tr>
<tr>
<td>o180</td>
<td>ot15</td>
<td>o265</td>
<td></td>
<td></td>
</tr>
<tr>
<td>o181</td>
<td>ot16</td>
<td>o269</td>
<td></td>
<td></td>
</tr>
<tr>
<td>o183</td>
<td>ot17</td>
<td>o270</td>
<td>o277</td>
<td></td>
</tr>
</tbody>
</table>
APPENDIX III

List of the oligonucleotides that were included in the clustering analysis of the gastrula stage library (BFLG). Set 1 filters were spotted with the PCR products of the clones in plates 1-66, while Set 2 filters with those in plates 67-132.
<table>
<thead>
<tr>
<th>BFLG</th>
<th>set 1 filters</th>
<th>set 2 filters</th>
</tr>
</thead>
<tbody>
<tr>
<td>o100</td>
<td>o171</td>
<td>o243</td>
</tr>
<tr>
<td>o102</td>
<td>o172</td>
<td>o244</td>
</tr>
<tr>
<td>o103</td>
<td>o173</td>
<td>o245</td>
</tr>
<tr>
<td>o104</td>
<td>o174</td>
<td>o246</td>
</tr>
<tr>
<td>o105</td>
<td>o176</td>
<td>o247</td>
</tr>
<tr>
<td>o106</td>
<td>o177</td>
<td>o249</td>
</tr>
<tr>
<td>o107</td>
<td>o178</td>
<td>o250</td>
</tr>
<tr>
<td>o108</td>
<td>o179</td>
<td>o251</td>
</tr>
<tr>
<td>o109</td>
<td>o180</td>
<td>o252</td>
</tr>
<tr>
<td>o110</td>
<td>o181</td>
<td>o253</td>
</tr>
<tr>
<td>o111</td>
<td>o182</td>
<td>o254</td>
</tr>
<tr>
<td>o112</td>
<td>o183</td>
<td>o255</td>
</tr>
<tr>
<td>o113</td>
<td>o184</td>
<td>o256</td>
</tr>
<tr>
<td>o114</td>
<td>o185</td>
<td>o257</td>
</tr>
<tr>
<td>o115</td>
<td>o186</td>
<td>o259</td>
</tr>
<tr>
<td>o116</td>
<td>o187</td>
<td>o260</td>
</tr>
<tr>
<td>o117</td>
<td>o188</td>
<td>o261</td>
</tr>
<tr>
<td>o118</td>
<td>o189</td>
<td>o262</td>
</tr>
<tr>
<td>o119</td>
<td>o191</td>
<td>o262</td>
</tr>
<tr>
<td>o120</td>
<td>o193</td>
<td>o263</td>
</tr>
<tr>
<td>o121</td>
<td>o194</td>
<td>o265</td>
</tr>
<tr>
<td>o122</td>
<td>o195</td>
<td>o266</td>
</tr>
<tr>
<td>o123</td>
<td>o196</td>
<td>o267</td>
</tr>
<tr>
<td>o124</td>
<td>o197</td>
<td>o268</td>
</tr>
<tr>
<td>o125</td>
<td>o198</td>
<td>o275</td>
</tr>
<tr>
<td>o126</td>
<td>o199</td>
<td>o276</td>
</tr>
<tr>
<td>o127</td>
<td>o201</td>
<td>o277</td>
</tr>
<tr>
<td>o128</td>
<td>o202</td>
<td>o278</td>
</tr>
<tr>
<td>o129</td>
<td>o203</td>
<td>o279</td>
</tr>
<tr>
<td>o130</td>
<td>o204</td>
<td>o280</td>
</tr>
<tr>
<td>o131</td>
<td>o205</td>
<td>o281</td>
</tr>
<tr>
<td>o132</td>
<td>o206</td>
<td>o282</td>
</tr>
<tr>
<td>o133</td>
<td>o207</td>
<td>o283</td>
</tr>
<tr>
<td>o134</td>
<td>o208</td>
<td>o284</td>
</tr>
<tr>
<td>o135</td>
<td>o209</td>
<td>o285</td>
</tr>
<tr>
<td>o136</td>
<td>o210</td>
<td>o286</td>
</tr>
<tr>
<td>o137</td>
<td>o211</td>
<td>o287</td>
</tr>
<tr>
<td>o138</td>
<td>o212</td>
<td>o288</td>
</tr>
<tr>
<td>o139</td>
<td>o213</td>
<td>o289</td>
</tr>
<tr>
<td>o140</td>
<td>o214</td>
<td>o290</td>
</tr>
<tr>
<td>o141</td>
<td>o215</td>
<td>o291</td>
</tr>
<tr>
<td>o142</td>
<td>o216</td>
<td>o292</td>
</tr>
<tr>
<td>o143</td>
<td>o217</td>
<td>o293</td>
</tr>
<tr>
<td>o144</td>
<td>o218</td>
<td>o294</td>
</tr>
<tr>
<td>o145</td>
<td>o219</td>
<td>o295</td>
</tr>
<tr>
<td>o146</td>
<td>o220</td>
<td>o296</td>
</tr>
<tr>
<td>o147</td>
<td>o221</td>
<td>o297</td>
</tr>
<tr>
<td>o148</td>
<td>o222</td>
<td>o298</td>
</tr>
<tr>
<td>o149</td>
<td>o223</td>
<td>o299</td>
</tr>
<tr>
<td>o150</td>
<td>o224</td>
<td>o300</td>
</tr>
<tr>
<td>o151</td>
<td>o225</td>
<td>o301</td>
</tr>
<tr>
<td>o152</td>
<td>o226</td>
<td>o302</td>
</tr>
<tr>
<td>o153</td>
<td>o227</td>
<td>o303</td>
</tr>
<tr>
<td>o154</td>
<td>o228</td>
<td>o304</td>
</tr>
<tr>
<td>o155</td>
<td>o229</td>
<td>o305</td>
</tr>
<tr>
<td>o156</td>
<td>o230</td>
<td>o306</td>
</tr>
<tr>
<td>o157</td>
<td>o231</td>
<td>o307</td>
</tr>
<tr>
<td>o158</td>
<td>o232</td>
<td>o308</td>
</tr>
<tr>
<td>o159</td>
<td>o233</td>
<td>o309</td>
</tr>
<tr>
<td>o160</td>
<td>o234</td>
<td>o310</td>
</tr>
<tr>
<td>o161</td>
<td>o235</td>
<td>o311</td>
</tr>
<tr>
<td>o162</td>
<td>o236</td>
<td>o312</td>
</tr>
<tr>
<td>o163</td>
<td>o237</td>
<td>o313</td>
</tr>
<tr>
<td>o164</td>
<td>o238</td>
<td>o314</td>
</tr>
<tr>
<td>o165</td>
<td>o239</td>
<td>o315</td>
</tr>
<tr>
<td>o166</td>
<td>o240</td>
<td>o316</td>
</tr>
<tr>
<td>o167</td>
<td>o241</td>
<td>o317</td>
</tr>
<tr>
<td>o168</td>
<td>o242</td>
<td>o318</td>
</tr>
<tr>
<td>o170</td>
<td>o243</td>
<td>o319</td>
</tr>
</tbody>
</table>
APPENDIX IV

Clustering of all 158 amphioxus sequences deposited in the Genbank (March 1999) after construction of their sequence based theoretical fingerprints with the same set of oligonucleotides used in the experimental oligonucleotide fingerprinting of the amphioxus BFL26 and BFLG libraries.
## Results

### Clustering Summary

- **Rows**: 158 (number of sequences used)
- **Columns**: 207 (number of oligonucleotides used)
- **Total number of clusters**: 8
- **Total number of singletons**: 134 gathered in cluster 9

### Cluster Information

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Size</th>
<th>Highest E-Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>0.814736</td>
<td>gil3204117embJ223444BF223444 Branchiostoma floridae Pax6 gene (AmphiPax6), clone 12.2</td>
</tr>
<tr>
<td>2</td>
<td>32</td>
<td>0.946206</td>
<td>gil3204113embJ223444BF223444 Branchiostoma floridae Pax6 gene (AmphiPax6), clone 1.4</td>
</tr>
<tr>
<td>3</td>
<td>34</td>
<td>0.832577</td>
<td>gil3204111embJ223444BF223444 Branchiostoma floridae Pax6 gene (AmphiPax6), clone 7.3</td>
</tr>
<tr>
<td>4</td>
<td>33</td>
<td>0.722088</td>
<td>gil3204116embJ223444BF223444 Branchiostoma floridae Pax6 gene (AmphiPax6), clone 4.2</td>
</tr>
</tbody>
</table>

### Cluster Details

<table>
<thead>
<tr>
<th>#</th>
<th>Size</th>
<th>E-Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.818621</td>
<td>109 0.44479 gil2055247embJ223444BF223444 Branchiostoma floridae Branchiostoma lanceolatum mRNA for calmodulin, complete cds</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.385014</td>
<td>108 0.234707 gil2055249embJ223444BF223444 Branchiostoma floridae Branchiostoma lanceolatum mRNA for calmodulin, complete cds</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0.87237</td>
<td>107 0.87237 gil2055249embJ223444BF223444 Branchiostoma floridae Branchiostoma lanceolatum mRNA for calmodulin, complete cds</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1.092240</td>
<td>106 1.092240 gil4006935embJ223444BF223444 Branchiostoma floridae Branchiostoma lanceolatum mRNA for troponin C, complete cds</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1.178723</td>
<td>105 1.178723 gil3025751embJ223444BF223444 Branchiostoma floridae Branchiostoma lanceolatum mRNA for troponin C, complete cds</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1.062699</td>
<td>104 1.062699 gil3201494embJ223444BF223444 Branchiostoma floridae Branchiostoma lanceolatum mRNA for troponin C, complete cds</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>0.744538</td>
<td>103 0.744538 gil3201492embJ223444BF223444 Branchiostoma floridae Branchiostoma lanceolatum mRNA for troponin C, complete cds</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>0.640733</td>
<td>102 0.640733 gil3025753embJ223444BF223444 Branchiostoma floridae Branchiostoma lanceolatum mRNA for troponin C, complete cds</td>
</tr>
</tbody>
</table>

### Additional Information

- **E-Values are reported for sequences with high similarity to known genes.**
APPENDIX V

Prediction of the identity of some of the experimental fingerprinting clusters from both amphioxus cDNA libraries based on their comparison with the theoretical fingerprints of the amphioxus sequences currently deposited in the Genbank using a BLAST type algorithm. T (theoretical) denotes the prediction of the identity of the clones that belong in this cluster based on the blast search of the experimental against the theoretical fingerprints. E (experimental) denotes the identity of a clone of the same cluster based on the blast search of its 5' tag sequence against the swissprot database.
BFL26 library

**T *** cluster 2 size 631 ***

1 4.732e-09 33 gil1526484dbjID87407ID87407 Branchiostoma floridiae mRNA for muscle actin, complete cds,
2 6.305e-06 36 gil1552219dbjID87739ID87739 Amphioxus mRNA for muscle actin BbMA1, complete cds,
3 2.608e-04 11 gil2113815iemblZ83271IBFZ83271 B.floridiae mRNA for actin,

E 066E14 BFL26/set1/s7a13b24 MPIMG531 673 s7a13b24.blastx_swplus 2.1e-142 Q93132
###>branchiostoma floridiae (florida lancelet) (amphioxus). muscle actin. 2/97 Length = 378 <###

**T *** cluster 10 size 175 ***

1 2.387e-07 10 gil2655387gbAF035173IAF035173 Branchiostoma floridiae cytochrome b (Cytb) gene, mitochondrial gene encoding mitochondrial protein, complete cds,

109D23 BFL26/RRset24/k8c14b10 MPIMG531 604 k8c14b10.blastx_swplus 8.2e-128 O47431
###>branchiostoma floridiae (florida lancelet) (amphioxus). cytochrome b. 6/98 Length = 380 <###

**T *** cluster 19 size 115 ***

1 6.505e-10 17 gil726505gbU22529IBFU22529 Branchiostoma floridiae alkali myosin light chain (AmphiMLC-alk) mRNA, complete cds,

065O19 BFL26/RRset22/t8c10a04 MPIMG531 582 t8c10a04.blastx_swplus 3.6e-96 Q17133
###>branchiostoma floridiae (florida lancelet) (amphioxus). myosin, essential light chain (myosin light chain alkali). 11/97 Length = 149 <###

**T *** cluster 37 size 68 ***

1 2.087e-06 18 gil2055249dbjIA0003082IAB0003082 Branchiostoma floridiae mRNA for calmodulin, complete cds,
2 2.087e-06 18 gil2440038iemblY09863IBFRNAC B.floridiae mRNA for calmodulin,
3 2.910e-04 22 gil2055247dbjIAB0003081IAB0003081 Branchiostoma lanceolatum mRNA for calmodulin, complete cds,

094G02 BFL26/RRset21/w8c09a12 MPIMG531 401 w8c09a12.blastx_swplus 1.8e-50 O16305
###>caenorhabditis elegans. t21h3.3 protein. 6/98 Length = 149 <###
094G02 BFL26/RRset21/w8c09a12 MPIMG531 401 w8c09a12.blastx_swplus 1.8e-50 O02367
###>ciona intestinalis. calmodulin. 6/98 Length = 149 <###

**T *** cluster 49 size 54 ***
1 3.680e-08 11 gill188350dbjlAB001688lAB001688 Branchiostoma lanceolatum mRNA for calcium vector protein, complete cds,

2 3.680e-08 11 gill188352dbjlAB001689lAB001689 Branchiostoma floridae mRNA for calcium vector protein, complete cds,

E 078A16 BFL26/RRset26/s8c17a53 MPIMG531 553 s8c17a53.blastx_swplus 2.2e-108 O01305

###>branchiostoma floridae (florida lancelet) (amphioxus). calcium vector protein.

T *** cluster 74 size 37 ***

1 1.688e-06 19 gil2113813leblZ83269IBFZ83269 B.floridae mRNA for AmphiBrf52,

E 093J01 BFL26/RRset26/p8c17a55 MPIMG531 318 p8c17a55.blastn_genembl 1.4e-77 Z83269

###>B.floridae mRNA for AmphiBrf52. 5/98 Length = 346 <###

093JO 1 BFL26/RRset26/p8c17a55 MPIMG531 318 p8c17a55.blastx_swplus 9.9e-07 Q39682

###>daucus carota (carrot). glycine-rich protein (fragment). 8/98 Length = 111 <###

T *** cluster 106 size 28 ***

1 6.014e-08 36 gil1526482dbjd87406ID87406 Branchiostoma floridai mRNA for cytoplasmic actin, complete cds,

2 2.877e-06 37 gil1552217dbjd87738ID87738 Amphioxus mRNA for cytoplasmic actin BbCA1, complete cds,

3 8.654e-05 36 gil2653409leblY13663IBLY13663 Branchiostoma lanceolatum mRNA for actin, 1832 bp,

E 009D04 BFL26/RRset3/8a07b63 MPIMG531 609 t8a07b63.blastx_swplus 6.1e-121
Q93129###>branchiostoma belcheri. actin, cytoplasmic (bbca1). 7/98 Length = 375 <###

T *** cluster 118 size 27 ***

1 1.756e-18 49 gil3201496leblAJ223581IBRF223581 Branchiostoma floridai mRNA for intermediate filament protein D1,

2 8.458e-07 64 gil3201490leblAJ223577IBRF223577 Branchiostoma floridai mRNA for intermediate filament protein C1, partial,

3 2.007e-06 39 gill3201512leblAJ223579IBRL223579 Branchiostoma lanceolatum mRNA for intermediate filament protein D1,

E 063C20 BFL26/RRset18/m8b17b26 MPIMG531 601 m8b17b26.blastx_swplus 1.1e-127 O62004

###>branchiostoma floridai (florida lancelet) (amphioxus). intermediate filament protein d1.
T *** cluster 149 size 22 ***


E 082B03  BFL26/RRset21/t8c04a26  MPIMG531  449  t8c04a26.blastn_genembl  3.4e-162
Z83263  ###>B.floridae mRNA for AmphiP2. 598 Length = 515 <###
082B03  BFL26/RRset21/t8c04a26  MPIMG531  449  t8c04a26.blastx_swplus  2.6e-47  001725
###>branchiostoma floridae (florida lancelet) (amphioxus). 60s acidic ribosomal protein p2. 798 Length = 116 <###

T *** cluster 228 size 17 ***

1  1.044e-09  42 gil3201494emblAJ223580IBRF223580 Branchiostoma floridae mRNA for
intermediate filament protein B1,

2  3.064e-09  42 gil5724lemblX64522IBLCIFIL B.lanceolatum mRNA for cytoplasmic intermediate
filament protein,

3  9.895e-04  15 gil2995654lgblAF052464IAF052464 Branchiostoma floridae homeobox protein
AmphiXlox gene, partial cds,

E 038109  BFL26/RRset16/v8b09b71  MPIMG531  753  v8b09b71.blastx_swplus  1.7e-72  062003
###>branchiostoma floridae (florida lancelet) (amphioxus). intermediate filament protein b1 (fragment).

T *** cluster 320 size 13 ***

1  4.962e-06  36 gill526482ldbjlD87406ID87406 Branchiostoma floridae mRNA for cytoplasmic
actin, complete cds,

2  4.606e-05  37 gil1552217dbjlD87738ID87738 Amphioxus mRNA for cytoplasmic actin BbCA1,
complete cds,

3  9.909e-04  36 gil2653409lemblY13663IBLY13663 Branchiostoma lanceolatum mRNA for actin,
1832 bp,

E 090E05  BFL26/RRset21/p8c03a15  MPIMG531  537  p8c03a15.blastx_swplus  2.5e-95  P30163
###>onchocerca volvulus. actin 2. 793 Length = 376 <###
090E05  BFL26/RRset21/p8c03a15  MPIMG531  537  p8c03a15.blastx_swplus  2.5e-95  P53486
###>fugu rubripes (japanese pufferfish) (takifugu rubripes). actin, cytoplasmic 3 (beta-actin 3).

T *** cluster 473 size 10 ***

1  5.637e-08  33 gil15526484dbjlD87407ID87407 Branchiostoma floridae mRNA for muscle actin,
complete cds,

2  2.746e-06  36 gil1552219dbjlD87739ID87739 Amphioxus mRNA for muscle actin BbMA1,
complete cds,
3.8330e-06 34 gi|2653407|emb|Y13662|BLY13662| Branchiostoma lanceolatum mRNA for actin, 1826 bp,

E092P05 BFL26/RRset26/p8c17a29|MPIMG531|544| p8c17a29.blastx_swplus 5.8e-118 Q93132
###>branchiostoma floridae (florida lancelet) (amphioxus). actin, muscle.

T *** cluster 517 size 9 ***
1 8.022e-06 10 gi|2113799|emb|Z83260|BFZ83260| B.floridae mRNA for alpha tubulin,

E044N08 BFL26/RRset16/k8b09b40|MPIMG531|420| k8b09b40.blastx_swplus 1.0e-52 P02553
###>lytechinus pictus (painted sea urchin). tubulin alpha chain (fragment).

BFLG library
T *** cluster 40 size 70 ***
1 3.742e-10 10 gi|726505|gb|U22529|BFU22529| Branchiostoma floridae alkali myosin light chain (AmphiMLC-alk) mRNA, complete cds,

E010B23 BFLG/RRset17/t8513b21|MPIMG498|673| t8513b21.blastx_swplus 1.0e-62 Q17133
###>branchiostoma floridae (florida lancelet) (amphioxus). myosin, essential light chain (myosin light chain alkali).

T *** cluster 71 size 46 ***
1 3.617e-08 22 gi|1526484|dbj|D87407|D87407| Branchiostoma floridae mRNA for muscle actin, complete cds,
2 4.654e-05 23 gi|1552219|dbj|D87739|D87739| Amphioxus mRNA for muscle actin BbMA1, complete cds,
3 4.262e-04 21 gi|2653407|emb|Y13662|BLY13662| Branchiostoma lanceolatum mRNA for actin, 1826 bp,

E038J21 BFLG/RRset23/m8519a24|MPIMG498|617| m8519a24.blastx_swplus 2.4e-105 Q93132
###>branchiostoma floridae (florida lancelet) (amphioxus). muscle actin.
References


Lehrach, H., Drmanac, R., Hoheisel, J., Larin, Z., Lennon, G., Monaco, A., Nizetic, D.,
and Sequencing, in: Genome Analysis Genetic and Physical mapping, Cold Spring Harbor
Laboratory Press.

Consortium: an integrated molecular analysis of genomes and their expression. Genomics


276:565.

DNA oligonucleotides containing non-specific base analogues. Journal of Molecular
Biology 270(3):426-35.


Lundin, L. (1993). Evolution of the vertebrate genome as reflected in paralogous

Compact genome of the puffer fish Fugu Ribripes. Genomics 25:436-446.

high density DNA and in situ colony filters based on fluorescence detection. Nucleic
Acids Research 22(16):3423-4.

of robotic technology to automated sequence fingerprint analysis by oligonucleotide


Nikoh, N., Iwabe, N., Kuma, K., Ohno, M., Sugiyama, T., Watanabe, Y., Yasui, K., Shi-cui,


Smit, A., and Green, P. RepeatMasker.


incorporation in GC rich RNA probes increases hybridisation sequence specificity. *Nucleic Acids Research* 16:4162.


Relevant Publications
AmphiBMP2/4, an Amphioxus Bone Morphogenetic Protein Closely Related to Drosophila decapentaplegic and Vertebrate BMP2 and BMP4: Insights Into Evolution of Dorsoventral Axis Specification

GEORGIA D. PANOPOULOU,* MATTHEW D. CLARK,1 LINDA Z. HOLLAND,2 HANS LEHRACH,1 AND NICHOLAS D. HOLLAND2
1Max-Planck-Institut für Molekulare Genetik, Berlin (Dahlem), Germany
2Scripps Institution of Oceanography, University of California at San Diego, La Jolla, California

ABSTRACT Amphioxus AmphiBMP2/4 appears to be a single gene closely related to vertebrate BMP2 and BMP4. In amphioxus embryos, the expression patterns of AmphiBMP2/4 suggest patterning roles in the ectodermal dorsoventral axis (comparable to dorsoventral axis establishment in the ectoderm by Drosophila decapentaplegic and vertebrate BMP4). In addition AmphiBMP2/4 may be involved in somite evagination, hindgut regionalization, differentiation of olfactory epithelium, patterning of the anterior central nervous system, and establishment of the heart primordium. One difference between the developmental role of amphioxus AmphiBMP2/4 and vertebrate BMP4 is that the former does not appear to be involved in the initial establishment of the dorsoventral polarity of the mesoderm. Dev. Dyn. 1998;213:130–139. © 1998 Wiley-Liss, Inc.

Key words: cephalochordate; TGFβ; dorsoventral body axis; heart

INTRODUCTION

Bone morphogenetic proteins (BMPs) were originally defined as inducers of ectopic bone in the vertebrate integument (Wozney et al., 1988). Although BMPs are molecularly diverse, most of them are extracellular morphogens belonging to the transforming growth factor beta (TGFβ) superfamily. All TGFβs are synthesized and secreted as relatively large precursor proteins which are proteolytically cleaved extracellularly to release a mature C-terminal portion that dimerizes with a second mature TGFβ to form an active signalling molecule (Hogan, 1996). The best known BMP of the TGFβ superfamily is vertebrate BMP4, which plays a key role in establishing the dorsoventral body axis at the gastrula stage and also helps pattern a wide range of organs and tissues during later development (Hogan, 1996). BMP4 is closely related both to Drosophila decapentaplegic (DPP), which also patterns the dorsoventral axis of the embryo (St. Johnston and Gelbart, 1987), and to vertebrate BMP2, which plays a role in dorsoventral patterning that is essential in fish (Kishimoto et al., 1997) but subsidiary in Xenopus (Suzuki et al., 1997).

During establishment of the dorsoventral axis, vertebrate BMP4 and Drosophila DPP from the anti-neural side of the embryo are antagonized, respectively, by chordin and short gastrulation protein (homologs of one another) produced on the neural side (Nellen et al., 1996; Piccolo et al., 1996). In both of these morphogenetic systems, the signals from the neural side of the body bind and, thereby, inactivate signals from the anti-neural side. Initial dorsoventral patterning is limited to the ectoderm of Drosophila, but occurs nearly simultaneously in the ectoderm and mesoderm of vertebrates (Wilson and Hemmati-Brivanlou, 1995; Holley et al., 1996; Neave et al., 1997). Work on the dorsoventral patterning of the ectoderm has led to the concept that BMP4 and DPP act as active inducers of epidermis at the expense of default-state neuroectoderm (Graff, 1997).

The similarities in the signalling pathways establishing dorsoventral body axis of Drosophila and vertebrates have been emphasized, especially by studies showing that BMP4 and DPP are functionally interchangeable, as are chordin and short gastrulation (Padgett et al., 1993; Holley et al., 1985, 1996). Such remarkable conservation strongly suggests that the molecular basis of dorsoventral patterning is phylogenetically ancient and may have arisen early in an ancient bilateral animal (the urbilaterian) that preceded the separation of the arthropod and vertebrate lineages (De Robertis and Sasai, 1996). The polarities of the morphogenetic systems establishing dorsoventral axes correspond between Drosophila and vertebrates. In contrast, the topographies of the Drosophila and vertebrate body plans are reversed dorsoventrally with respect to one another, i.e., thus, the neural side of the former is ventral, whereas the neural side of the latter

Grant sponsor: Max-Planck-Gesellschaft; Grant sponsor: NSF; Grant number: IBN 96-309938.
*Correspondence to: G.D. Panopoulou, Max-Planck-Institut für Molekulare Genetik, Berlin (Dahlem), Germany.
Received 9 April 1997; Accepted 2 June 1998
is dorsal. This pattern has favored the idea that the animal body became inverted dorsally with respect to the substratum during deuterostome evolution (Arendt and Nübler-Jung, 1994; Holley et al., 1995; De Robertis and Sasai, 1996).

The dorsoventral inversion hypothesis is of central importance for understanding the evolutionary relationships of the animal phyla. One way to test this hypothesis further is to examine the structure and developmental expression of BMP-related genes in a wide spectrum of animals. To date, however, such genes have been studied only for Drosophila, Caenorhabditis, vertebrates, and ascidian tunicates. The tunicate gene, a homolog of both vertebrate BMP2 and BMP4, functions in establishing dorsoventral polarity in the ectoderm, but appears to play no role in mesodermal patterning (Miya et al., 1997).

To shed more light on the evolution of genetic mechanisms establishing the dorsoventral body axis of animals generally, we studied an amphioxus gene (AmphiBMP2/4), which is closely related to Drosophila DPP and to vertebrate BMP2 and BMP4. Amphioxus (subphylum Cephalochordata; phylum Chordata) is widely believed to be the closest living invertebrate relative of the vertebrates (Wada and Satoh, 1994) and has long occupied a key position in discussions of the origin of the vertebrates from the invertebrates (Presley et al., 1996).

Amphioxus morphogenesis, which consists of foldings of unstratified cell layers, contrasts with vertebrate morphogenesis, which includes conspicuous delaminations of stratified epithelia and migrations of individual cells. As a result, there are some differences in the formation of the germ layers between amphioxus and vertebrates. At the start of the amphioxus gastrula stage, the vegetal hemisphere of the blastula invaginates as the hypoblast into the animal hemisphere, which becomes the epiblast. At the gastrula-neurula transition, the epiblast becomes subdivided into neural ectoderm dorsally and non-neural ectoderm (epidermis) laterally and ventrally. The paraxial mesoderm soon evaginates to form myocoelic somites, which constitute dorsal mesoderm (Fig. 1A,B). Ventral mesoderm is lacking and does not appear until later, when it is produced (as perivisceral mesothelia) by ventrad growth from the somites (Fig. 1C).

In most animals, all the mesodermal regions are not produced virtually simultaneously at gastrulation. For example, amphioxus embryos produce the dorsal mesoderm first and the ventral mesoderm later; and Drosophila produces the ventral mesoderm first and the dorsal mesoderm later. In contrast, vertebrate embryos produce all their mesoderm virtually simultaneously by delamination and cell migrations as they gastrulate (Fig. 1D). Thus, the advent of the ventral mesoderm, which is delayed in amphioxus, is precocious in vertebrates. The present study indicates that this difference is reflected in the genetic mechanisms establishing dorsoventral polarity within the mesoderm. During amphioxus development, AmphiBMP2/4 may be involved in patterning the hypoblast, but not in the early establishment of a dorsoventral axis within the mesoderm (although this would not necessarily preclude AmphiBMP2/4 from patterning the mesoderm in other ways). In sum, although both vertebrates and amphioxus apparently use BMP homologs similarly to pattern the dorsoventral axis in the ectoderm, vertebrates may be the only animals in which BMPs play concurrent roles in both the ectoderm and mesoderm for establishing dorsoventral polarity.

**RESULTS**

**Sequence Analysis**

Figure 2 shows the structure of full-length cDNA of AmphiBMP2/4 from Branchiostoma floridae. The nucleotide sequence is 1,981 bp long and has a noncanonical polyadenylation signal (AATAAA) near the 3' end (Sheets...
et al., 1990). The longest open reading frame encodes a protein of 361 amino acids, when assumed that translation starts at the first ATG downstream from the in-frame stop codons. The C-terminal 102 amino acids include seven invariant cysteines characteristic of the mature region of BMPs. The most probable site where the mature region is cleaved proteolytically from the protein of 361 amino acids, if one assumes that translation starts at the first invariant cysteine (between RQKR and the following alanine).

**Phylogenetic Analysis of BMP-Related Proteins**

A phylogenetic tree of BMP-related proteins (Fig. 4) includes a distinct clade (the DPP subclass of TGF(Bs) with the branching order: *Drosophila* DPP, tunicate *AmphiBMP2/4*, and vertebrate BMP2 and BMP4. Therefore, *AmphiBMP2/4* lies at the root of the divergence between vertebrate BMP2 and vertebrate BMP4.

**Southern Blot Analysis**

Figure 5 shows a Southern blot hybridized at low stringency with an 817-bp fragment of *AmphiBMP2/4* DNA, including the conserved 3' half of the coding region and part of the 3' UTR. Under such low stringency, we would have expected multiple bands in all lanes if amphibioxus possessed a second *BMP2/4* gene. Mouse *BMP2* and *BMP4* have a 66% nucleotide identity over the 545 most 3' bases of the coding region (83% identity over the most conserved 336 bases), and we have previously shown that probing a muscle actin gene at higher stringency with a shorter probe gives the two-headed arrow marks the most probable proteolytic maturation site. The arrowheads delimit the EcoRI fragment used to probe the Southern blot.
has largely roofed over the neural plate, grooves are giving rise to the paraxial mesoderm. Fianking regions of ectoderm that are overgrowing it are composed of endoderm, except dorsolaterally where expression is undetectable in the neural plate or dorsal epidermis and is weak in the ventral epidermis (Fig. 6E-G). By this stage, the hypoblast can be considered as composed of endoderm, except dorsolaterally where the presomitic grooves constitute paraxial mesoderm and middorsally where the notochord is forming. Expression continues in the endoderm and presomitic grooves, but not detectable in the presumptive notochord (Fig. 6F-G).

Developmental Expression of AmphiBMP2/4

Transcripts of AmphiBMP2/4 are first detectable by in situ hybridization at the midgastrula stage (Fig. 6A), strongly throughout the invaginated hypoblast and less conspicuously in the epiblast. By the late gastrula stage, expression continues throughout the epiblast and hypoblast, but is most conspicuous in dorsolateral regions of the hypoblast (Fig. 6B, arrowheads) where transcripts of AmphiBMP2/4 are presumable due to polymorphism.

Fig. 3. Amino acids encoded by AmphiBMP2/4 compared with amino acids encoded by mouse BMP2 (Feng et al., 1994) and BMP4 (Takaoka et al., 1993). Clustal V was used to introduce gaps (indicated by dashes). Identical amino acids and conserved amino acid substitutions are indicated, respectively, by colons and single dots. In the mature region, the seven invariant cysteines are indicated by arrowheads.

Therefore, it is likely that amphioxus has only one BMP2/4 gene and that the two bands resulting from digestion with EcoRI1 and BstXI are presumably due to polymorphism.
mBMPS
mBMP7
mBMP6
mBMP8
HrBMPa
d60A
dSCREW
mBMP2
mBMP4
AmphiBMP2/4
HrBMPb
dDPP

Fig. 4. Phylogenetic tree constructed by neighbor joining (Saitou and Nei, 1987) and based on the full-length proteins of amphioxus AmphiBMP2/4 and 12 related TGFα proteins from mouse (m), Drosophila (d), and the ascidian tunicate Halocynthia roretzi (Hr). The scale bar indicates the number of substitutions per amino acid position.

Fig. 5. Genomic Southern blot analysis of DNA pooled from 20 adults of Branchiostoma floridae. Numbers at top of lanes refer to digestion in the following restriction enzymes: 1, Sall; 2, PstI; 3, EcoRI; 4, EcoO109I; 5, BstXI; 6, BglII; 7, Apal. Blot probed at low stringency with 3' half of the coding region plus part of 3' UTR of AmphiBMP2/4.

ceral coelom has grown ventrad from the myocoels. At this stage, some of the splanchnic mesothelial cells of the perivisceral coelom contain conspicuous transcripts of AmphiBMP2/4. These expressing mesothelial cells are invariably situated to the right of the midventral line of the body, markedly to the right in the region of the pharynx (Fig. 6N) but only slightly to the right in the region of the midgut and hindgut (Fig. 6O).

Figure 6P–U shows an early larva (with primary pigment spot, mouth, and first gill slit). Expression is still detectable in the cells of Hatschek’s right diverticulum, but not the left one (now part of the preoral ciliated pit). Transcripts are also detectable in the club-shaped gland, posteriorly in the mesoderal growth zone posterior to the last somite, at the posterior extremity of the hindgut, and in splanchnic mesothelial cells (skewed, as before, markedly to the right in the pharyngeal region and slightly to the right along the midgut and hindgut). In addition, some cells at the anterior end of the neural tube contain transcripts of AmphiBMP2/4, but there is no longer any detectable expression in the anterodorsal patch of epidermal cells. During the first week of larval life, the expression signals gradually fade until they can no longer be detected by in situ hybridization.

DISCUSSION

BMP2 and BMP4 Probably Arose by Gene Duplication in Early Vertebrate Evolution

Mouse BMP2 and BMP4 encode proteins sharing overall amino acid identities of 61% (about 85% in the mature region). Despite this structural similarity, the two proteins have divergent functions, although the possibility that they form heterodimers remains open (Jones et al., 1991; Furuta et al., 1997). In contrast, amphioxus (present results) and tunicates (Miya et al., 1997) probably have only a single homolog of these two vertebrate genes. It is thus likely that a single ancestral gene comparable to AmphiBMP2/4 duplicated early in the vertebrate evolutionary line to produce BMP2 and BMP4. The functions of the ancestral gene were evidently parcelled out between the two vertebrate paralogs, although with some overlap and redundancy. Such a gene duplication appears to be an example of the wide-spread multiplication of genes that is correlated with the marked increase in body plan complexity at the origin of the vertebrates (Holland et al., 1994).

Dorsoventral Patterning in Bilaterian Animals

Ectodermal transcription of amphioxus AmphiBMP2/4, which is ubiquitous at the late gastrula stage, is down-regulated in the dorsal epiblast at the early neurula stage. This transcription pattern suggests that AmphiBMP2/4—like Drosophila DPP (Frasch, 1995; Brehs et al., 1996), vertebrate BMP4 (Graff, 1997; Hemmati-Brivanlou and Melton, 1997), and tunicate HrBM PB (Miya et al., 1997)—may participate in a signalling system that subdivides the embryonic ectoderm into the non-neural epidermis and the precursor of the central nervous system.

AmphiBMP2/4 is also expressed in the hypoblast of the late gastrula, most strongly on either side dorsolaterally, where the presomitic grooves of the paraxial mesoderm will evaginate; expression is weaker ventrally and laterally (presumptive endoderm) and middorsally (presumptive notochord). This expression pattern suggests that AmphiBMP2/4 expression might play a role in patterning the hypoblast (and perhaps in early somite evagination); however, this is not the same as establishing a dorsoventral axis within the meso-
In the early neurula of amphioxus, the presomitic grooves become divided up along the rostrocaudal axis to form myocoelic somites, which constitute the dorsal, paraxial mesoderm; at this stage, ventral mesoderm is lacking and only appears later when the perivisceral coelom buds off ventrally from the myocoels (Fig. 1C). In contrast, an important early function of vertebrate BMP2 is the establishment of dorsoventral polarity in the nascent mesoderm (De Robertis and Sasai, 1996; Holley et al., 1996; Graff, 1997; Tonegawa et al., 1997). The vertebrate mesoderm is extensive from the outset (Fig. 1D), and BMP4 helps determine its dorsoventral axis by means of a signalling system parallel to the one acting in the ectoderm.

The present results suggest that Drosophila, tunicates, amphioxus, and vertebrates all establish the dorsoventral axis within the ectoderm around the time of gastrulation, but no such common theme unites dorsoventral patterning within the mesoderm of these animal groups. In Drosophila, mesodermal patterning is subsequent to and evidently driven by the dorsoventral patterning in the ectoderm (St. Johnston and Gelbart, 1987; Ferguson and Anderson, 1992; De Robertis and Sasai, 1996); in tunicates, the coelomic mesoderm is degenerate and largely replaced by a diffuse mesenchyme that apparently never becomes dorsoventrally patterned; in amphioxus, an AmphiBMP2/4-expressing dorsal mesoderm develops before there is any ventral mesoderm. Therefore, only the vertebrates appear to produce a mesoderm that is extensive at the outset and is patterned by a genetic mechanism paral­leling the BMP-based one acting simultaneously in the ectoderm. The most parsimonious evolutionary scenario suggested by these data is that the urbilaterian used BMP-like morphogens to help establish the dorsoventral axis of the body within the ectoderm. To this original patterning system in the ectoderm, the vertebrates evidently added a parallel genetic system to ensure the early establishment of dorsoventral polarity within their mesoderm.

AmphiBMP2/4 in Tailbud, Gut, and Neural Structures

AmphiBMP2/4 expression rapidly declines in the somites and soon remains detectable only in the most posterior paraxial mesoderm, which is part of a tail-bud like growth zone. Therefore, in contrast to vertebrate BMP4 (Pourquié et al., 1996), AmphiBMP2/4 is apparently not involved in regionalization within individual somites. However, sustained expression of AmphiBMP2/4 in the mesodermal growth zone posterior to the last somite is very similar to BMP2 and BMP4 expression in the tail bud mesenchyme of vertebrates (Jones et al., 1991; Fainsod et al., 1994).

The expression pattern of AmphiBMP2/4 is identical in the pharynx and hindgut of amphioxus embryos and larvae suggests that the gene may be involved in anteroposterior regionalization of the endoderm. In Drosophila, DPP is also expressed in the developing foregut and hindgut (St. Johnston and Gelbart, 1987), and BMP4 is expressed in some of the visceral pouches of the developing pharynx of vertebrates (Hemmati-Brivanlou and Thomsen, 1995). In amphioxus, the AmphiBMP2/4-expressing region of pharyngeal wall develops into the club-shaped gland, a highly modified gill slit (Goodrich, 1930) for which a number of functions have been proposed (reviewed by Olsson, 1983).

The developmental expression pattern of AmphiBMP2/4 also suggests that this gene is involved in differentiation of anterior neural structures after neurulation has formed the central nervous system of amphioxus. At the late neurula stage, expression in a patch of anterodorsal epidermis might be related to the differentiation of primary sensory neurons, which occur there in high concentrations (Baatrup, 1981; Fritzsch, 1996). This epidermal expression of AmphiBMP2/4 might be comparable to the expression of BMP2 in the nasal placodes of vertebrates (Hemmati-Brivanlou and Thomsen, 1995).

In early larvae of amphioxus, AmphiBMP2/4 is expressed at the anterior of the neural tube, which is a diencephalon-like region thought to include homologs of the vertebrate pineal and lateral eyes (Lacalli et al., 1994). This neural expression has parallels in tunicate HrBMP6, which is expressed far anteriorly in the larval nervous system (Miya et al., 1997) and in vertebrate BMP4, which is expressed in the developing diencephalon, pineal, and lateral eyes (Jones et al., 1991; Fainsod et al., 1994; Hemmati-Brivanlou and Thomsen, 1995; Winnier et al., 1995). However, whereas vertebrate BMP4 plays a role in establishing dorsoventral patterning within the neural tube (Liem et al., 1995; Furuta et al., 1997; Mehler et al., 1997), amphioxus AmphiBMP2/4 evidently does not.

AmphiBMP2/4 and Heart Development in Amphioxus

At the late neurula stage of amphioxus, some of the splanchnic mesothelial cells of the newly produced visceral coelom begin to express AmphiBMP2/4. The right-left axis of amphioxus embryos and larvae is distorted by excessive growth in breadth of the left side (probably due to their highly specialized method of filter feeding, as discussed by Presley et al., 1996). Thus, the morphologic ventral midline is displaced to the right of the topographical ventral midline, most markedly at the level of the pharynx (van Wijhe, 1919).

The row of AmphiBMP2/4-expressing mesothelial cells runs right along the morphologic ventral midline (i.e., is canted conspicuously to the right along the pharynx and slightly to the right along the midgut and hindgut); this line of cells exactly marks the course of the developing subenteric blood vessel and its anterior continuation, the truncus arteriosus. These vessels are, at least in part, precursors of the subendostylar artery, the probable homolog of the heart of tunicates and vertebrates (Hirakow, 1985), despite statements in
many textbooks that amphioxus has no heart. Thus, AmphiBMP2/4 evidently plays a part in amphioxus heart formation, just as BMP2 and BMP4 function in vertebrate heart formation (Jones et al., 1991; Fainsod et al., 1994; Hemmati-Brivanlou and Thomsen, 1995; Martinez-Barberá et al., 1997). The developmental gene cascades involved in early patterning of the heart could well be very similar in amphioxus and vertebrates. However, much of later cardiogenesis is no doubt unique to vertebrates (Fishman and Chien, 1997); for example, vertebrate BMP4 is involved in making a symmetrical heart rudiment asymmetrical (Chen et al., 1997), but, conversely, amphioxus AmphiBMP2/4 appears to help establish an originally asymmetric heart.
rudiment that later becomes symmetric in the topological ventral midline at larval metamorphosis.

EXPERIMENTAL PROCEDURES

Larval cDNA Library Construction; Screening; Sequencing

Ripe adults of the Florida amphioxus, *B. floridæae*, were collected by shovel and sieve in 1 m of water in Tampa Bay, Florida, during the summer breeding season. Eggs from electrically stimulated animals were fertilized, and the embryos and larvae were cultured in the laboratory (Holland and Holland, 1993).

Total RNA was isolated from 26-hr embryos (Chomczynski and Sacchi, 1987), and poly(A)* RNA was purified with Oligo (dT)30 Dynabeads (Dynal, Oslo, Norway). The cDNA was prepared with the SuperScript Plasmid kit (Gibco BRL, Gaithersburg, MD) and directionally cloned into plasmid vector pSP65 (Gibco BRL). A total of 50,000 cDNA clones were randomly picked into 384-well microtiter plates and robotically spotted onto Hybond N* nylon membranes (Amersham, UK) (Lehrach et al., 1990, 1997; Zehetmeier and Lehrach, 1994). This library was screened with a 1.1-kb polymerase chain reaction fragment, including the 3' end of mouse BMP4 labeled with [32P]dATP and [32P]dCTP at 42°C overnight in 30% deionized formamide, 4X standard saline citrate (SSC), 1 mM ethylenediaminetetra-acetic acid, 50 mM Na2HPO4 (pH 7.2), 1% SDS, 8% dextran sulfate, and 10% Denhardt's. Hybridized filters were washed in 2 X sodium dodecyl sulfate (SSC) with 0.1% SDS twice for 20 min at room temperature and twice more for 20 min at 65°C. Two full-length cDNA clones were obtained, and one (MPIMGEFL26-1P13 = AmphioEMPF2/4) was sequenced on both strands (ABI PRISM, Perkin Elmer, Langen, Germany).

Sequence Analysis and Phylogenetic Tree Construction

Full-length amino acid sequences were compared among AmphibMP2/4 mouse BMP2 (Feng et al., 1994) and mouse BMP4 (Takaoka et al., 1993). Clustal V was used to introduce gaps, and amino acid substitutions with Dayhoff scores of nine or more were considered conserved (Doolittle, 1987). A phylogenetic tree was constructed based on the full length proteins of amphioxus AmphioEMPF2/4 and twelve related TGFβ proteins from mouse (m), *Drosophila* (d), and the tunicate *Halocynthia roretzi* (Hr). The tree was constructed with a Lasergene program (DNASTAR, Madison, WI) based on the neighbor-joining method of Saitou and Nei (1987). References for the sequences were: mBMP5 (Lyons et al., 1989), mBMP6 (Gitelman et al., 1994), mBMP7 (Ozkaynak et al., 1991), mBMP8 (Ozkaynak et al., 1992), HrBMPa (Miya et al., 1996), d60A (Wharton et al., 1991), dSCREW (Arora et al., 1994), mBMP2 (Feng et al., 1994), mBMP4 (Takaoka et al., 1993), HrBMPb (Miya et al., 1997), dDPP (Padgett et al., 1987), and dNODAL (Zhou et al., 1993).

Obtaining Adult Genomic DNA and Southern Blot Analysis

Genomic DNA was extracted from a sample of 20 adults of *B. floridæae*, and seven 10-μg aliquots were digested, respectively, with *SalI*, *PstI*, *EcoRI*, *EcoO19I*, *BstXI*, *BglII*, and *ApoI*. Southern blots were prepared according to Holland et al. (1996) and hybridized with a 817-bp *EcoRI* fragment (between the arrowheads in Fig. 2) that included the highly conserved 3' half of the coding region of *AmphiBMP2/4*. Hybridization and focusing on the mesothelial cells in the region of the pharynx. L: Cross-section through level x in 6J with detectable expression in anterior dorso-sal epidermis and in Hatschek's right (arrowhead) and left (arrow) diverticula. M: Cross-section through level y in 6J showing strong expression in ventral and right wall of pharynx (the rudiment of the club-shaped gland). N: Cross-section through level x in 6K showing expression in splanchnic mesothelial cells of the perivisceral coelom in the region of the pharynx. O: Cross-section through level y in 6K showing expression in splanchnic mesothelial cells of the perivisceral coelom in the region of the midgut. P: Early larva (arrow indicates the primary pigment spot, which does not represent *AmphiBMP2/4* expression) with expression in right Hatschek's diverticulum, anterior end of nerve cord, club-shaped gland, posterior mesoderm, posterior extremity of hindgut, and splanchnic mesothelial cells of the perivisceral coelom. Q: Midgut region of preceding larva showing expression in mesothelial cells just to the right of the ventral midline. R: Anterior portion of 6P showing expression at anterior end of nerve cord (arrowhead), in wall of Hatschek's right diverticulum (single arrow), and in club-shaped gland (twin arrows). S: Anterior part of 6P with right side in focus. T: Cross-section through level x in 6S showing expression in splanchnic mesothelial cells in pharyngeal region. U: Section through level y in 6S showing expression in splanchnic mesothelial cells in the midgut region (the arrow indicates the primary pigment spot in the nerve cord). Scale bars = 40 μm in A–E, H, J, K, P–S, 10μm in F, G, I, L–O, T, U.
washing were at low stringency according to Holland et al. (1997).

**In Situ Hybridization and Histology**

Gene expression was studied by in situ hybridization of developmental stages of *B. floridiae* fixed at frequent intervals. Fertilization envelopes of unchopped embryos were removed with insect pins to facilitate penetration of reagents. Methods of fixation and in situ hybridization were according to Holland et al. (1996). The riboprobe, synthesized starting at the 3′ end of the full-length clone of *AmphioBMP2/4*, was approximately 1,600 bp long. After the in situ preparations had been photographed as whole mounts, they were counterstained and prepared as 3.5-μm histologic sections according to Holland et al. (1996).

**ACKNOWLEDGMENTS**

We are deeply indebted to Dr. John M. Lawrence of the University of South Florida for providing laboratory facilities in Tampa and to Dr. Michael V. Wiles for the mBMP4 probe and for valuable discussions and comments. This work was supported in part by the Max-Planck-Gesellschaft (G.D.P. and M.D.C.) and in part by NSF research grant IBN 96-309938 to N.D.H. and L.Z.H.

**REFERENCES**


Fritzsch B. Similarities and differences in lanceolate and craniate nervous systems. Lar J. Zool. 1996;42(Suppl.):147–160.

**Furuta Y, Piston DW, Hogan BML. Bone morphogenetic proteins (BMPs) as regulators of dorsal forebrain development. Development 1997;124:2203–2212.**


**Goodrich ES. “Proboscis pores” in craniate vertebrates, a suggestion concerning the premandibular somites and hypophysis. Q. J. Microsc. Sci. 1917;82:539–563.**


**Graff JM. (1997). Embryonic patterning: to BMP or not to BMP, that is the question. Cell 1997;89:171–174.**

**Hemmati-Brivanlou A, Melton D. Vertebrate embryonic cells will become nerve cells unless told otherwise. Cell 1997;86:13–17.**

**Hemmati-Brivanlou A, Thomsen GH. Ventral mesodermal patterning in *Xenopus* embryos: Expression patterns and activities of BMP-2 and BMP-4. Dev. Genet. 1995;17:76–89.**


**Hogan BML. Bone morphogenetic proteins: Multifunctional regulators of vertebrate development. Genes Dev. 1996;10:1580–1594.**


**Holland LZ, Kene M, Williams NA, Holland ND. Sequence and embryonic expression of the amphioxus engrailed gene (*AmphiEn*): The metameric pattern of transcription resembles that of its segment-polarity homolog in *Drosophila*. Development 1997;124:1723–1732.**


**Holland PWH, Garcia-Pernádez J, Williams NA, Sidow A. Gene duplications and the origins of vertebrate development. Development 1994(Suppl.):125–133.**


**Jones CM, Lyons KM, Hogan BML. Involvement of Bone Morphogenetic Protein-4 (BMP-4) and Vgr-1 in morphogenesis and neurogenesis in the mouse. Development 1991;111:531–542.**


