

Genome Analysis
Based on EST Collections:
A Clustering Pipeline and a Database on
Xenopus laevis

Dissertation zur Erlangung des akademischen Grades eines
Doktors der Naturwissenschaften (Dr. rer. nat.)
der Technischen Fakultät der Universität Bielefeld

vorgelegt von
Alexander Sczyrba
Dezember 2006

Dipl.-Inform. Alexander Sczyrba
Universität Bielefeld
Technische Fakultät
AG Praktische Informatik
D-33594 Bielefeld
email: asczyrba@techfak.uni-bielefeld.de

Genehmigte Dissertation zur Erlangung des akademischen Grades
Doktor der Naturwissenschaften (Dr. rer. nat.).
Der Technischen Fakultät der Universität Bielefeld
am 04.12.2006 vorgelegt von Alexander Sczyrba,
am 30.03.2007 verteidigt und genehmigt.

Gutachter:

Prof. Dr. Robert Giegerich, Universität Bielefeld
Prof. Dr. Curtis R. Altmann, Florida State University, USA

Prüfungsausschuss:

Prof. Dr. Thomas Noll, Universität Bielefeld
Prof. Dr. Robert Giegerich, Universität Bielefeld
Prof. Dr. Curtis R. Altmann, Florida State University, USA
Dr. Dirk J. Evers, Universität Bielefeld

Gedruckt auf alterungsbeständigem Papier nach ISO 9706.

Acknowledgments

I would like to thank my supervisor Robert Giegerich for his guidance, ideas and support throughout my time as graduate student. Thanks for giving me the opportunity to do this work at the Practical Computer Science group.

I was lucky to meet Curtis Altmann during my time at Terry Gaasterland's lab at Rockefeller University. His interest in developmental biology inspired me to pay regard to *Xenopus*, a species not much appreciated by the bioinformatics community. Curtis constantly came up with new ideas that seemed so simple to solve by writing some Perl scripts...

Special thanks go to Stefan Kurtz not only for providing his *Vmatch* tool, but also for the dinner sessions we had at his place when he was still in Bielefeld.

Thanks to colleagues and friends, especially Jomuna Choudhuri, Dirk Evers, Thomas Fiedler, Andreas Freier, Arne Hauenschild, Matthias and Thomas Höchsmann, Carsten Meyer, Jens Reeder, Marc Rehmsmeier, Peter Steffen, and Thomas Töller, and the BiBiServ team with Jörn Clausen, Sven Hartmeier, Susanne Konermann, and Jan Krüger for many helpful discussions and a nice time. Special thanks to Michael Beckstette, who not only kept me caffeinated, but also has been great help during the data analysis for XenDB providing his Genlight system.

Finally, I would like to thank my parents and family for their support and patience, and Petra for putting up with me over the past years, filling my life with joy.

Contents

Acknowledgments	iii
1. Introduction	1
1.1. Motivation	1
1.2. Structure of the Thesis	3
I. Suffix-Array Based EST Clustering	5
2. EST Clustering	7
2.1. Expressed Sequence Tags (ESTs)	8
2.1.1. cDNA Cloning	8
2.1.2. EST Sequencing	11
2.1.3. EST Quality	13
2.1.4. EST Databases	15
2.1.5. EST Uses	17
2.2. EST Clustering	20
2.2.1. Clustering Goals	20
2.2.2. Clustering Procedure	21
2.3. Clustering Tools	24

2.3.1.	<i>CAP3</i>	24
2.3.2.	<i>d2_cluster</i>	25
2.3.3.	<i>PaCE</i>	25
2.3.4.	<i>BLASTclust</i> (megaBLAST)	26
2.3.5.	<i>TGICL</i>	26
2.4.	Gene Indices	26
2.4.1.	NCBI UniGene	27
2.4.2.	TIGR Gene Index	28
2.4.3.	STACK	29
3.	Suffix Array Based EST Mapping and Clustering	31
3.1.	Motivation	31
3.2.	Enhanced Suffix Arrays	32
3.2.1.	Suffix Trees and Suffix Arrays	32
3.2.2.	Enhanced Suffix Arrays	35
3.2.3.	<i>Vmatch</i>	36
3.3.	e2g - EST Mapping	37
3.3.1.	Design Rationale	38
3.3.2.	Implementation	39
3.3.3.	Web interface	41
3.3.4.	Performance Evaluation	44
3.3.5.	Utility	44
3.4.	EST Clustering using <i>Vmatch</i>	44
3.4.1.	Clustering Parameters	45
3.4.2.	<i>X. laevis</i> EST Data Set	46
3.5.	Validation of Clustering Results	56
3.5.1.	Hubert and Arabie Adjusted Rand Index	57
3.5.2.	EST Clustering Benchmark Data Set	59
3.5.3.	Quality Evaluation	61
3.5.4.	Performance Evaluation	78
3.6.	Summary	80
4.	EST Clustering Pipeline	81
4.1.	Design Rationale	81

4.1.1. Data Import	82
4.1.2. Pre-Processing	84
4.1.3. Repeat Masking	84
4.1.4. Clustering	85
4.1.5. Assembly	86
4.1.6. Annotation of Contig Sequences	86
4.1.7. Web interface	90
4.2. Database Schema	91
4.3. Implementation	96
4.3.1. Clustering Pipeline	96
4.3.2. User Interface	97
4.4. Summary	106

II. Applications to *Xenopus laevis* 107

5. XenDB: A *Xenopus laevis* Gene Index 109

5.1. Motivation	109
5.2. Generation of a <i>Xenopus laevis</i> Gene Index	110
5.2.1. Sequence retrieval and Cleanup	110
5.2.2. Repeat Masking	111
5.2.3. Clustering	112
5.2.4. Assembly	113
5.3. Sequence Analysis of <i>Xenopus laevis</i> Gene Index	114
5.3.1. Identification of Chimeric Sequences	115
5.3.2. Gene Ontology prediction and Functional Classification	115
5.4. Clone Selection	117
5.4.1. Identification of full length contigs	117
5.4.2. Identification of full length clones	119
5.5. Utility	121
5.5.1. User Interface	121
5.5.2. Homeobox Gene Identification	122
5.5.3. Homologue Identification from the Cancer Genome Anatomy Project	123
5.5.4. Homologues of <i>Drosophila</i> Eye Development Genes	125

5.5.5. Application of the IsoSVM classifier to <i>X. laevis</i> EST data	127
5.6. Summary	127
6. Computational Identification of miRNAs in <i>X. laevis</i> EST clusters	129
6.1. microRNAs: Biogenesis and Prediction	130
6.2. Computational Identification of miRNAs	133
6.3. Results: <i>X. laevis</i> miRNAs	134
6.4. Summary	136
7. Conclusion and Outlook	139
A. Structures of Predicted miRNA Precursors	141
Bibliography	145

List of Tables

2.1. Number of GenBank and dbEST entries	18
3.1. <i>Vmatch</i> clustering parameters for <i>X. laevis</i> data set.	46
3.2. Notation for contingency table representing cluster overlap	58
3.3. Formulae for calculating the number of object pairs for the four different types of pairs.	59
3.4. Number of clusters and singlets for the 16 <i>A. thaliana</i> benchmark data sets .	62
3.5. <i>Vmatch</i> clustering parameters for <i>A. thaliana</i> data set.	62
3.6. Maximum values of the Adjusted Rand indices for <i>A. thaliana</i> benchmark data sets clustered with <i>Vmatch</i> option <code>-identity</code> and <code>-leastscore</code>	68
3.7. Mean Adjusted Rand Index for <i>A. thaliana</i> benchmark data sets, <i>Vmatch</i> option <code>-identity</code>	68
3.8. Mean Adjusted Rand Index for <i>A. thaliana</i> benchmark data sets, <i>Vmatch</i> option <code>-leastscore</code>	73
3.9. Mean Adjusted Rand Index for <i>A. thaliana</i> benchmark data sets, <i>Vmatch</i> option <code>-leastscore</code> and different X-Drop values	74
3.10. <i>Vmatch</i> 'default' parameter settings for EST clustering.	75
3.11. Friedman analysis of variance by ranks applied to the Rand Index values of the different clustering tools	77

5.1. Tissue types and developmental stages in <i>X. laevis</i> ESTs	111
5.2. Summary of <i>X. laevis</i> EST cleanup and clustering.	113
5.3. Number of full length <i>X. laevis</i> contigs as derived by BLASTX and FASTY . .	119
5.4. Average length of <i>X. laevis</i> contigs for different BLASTX and FASTY full length contig categories	120
5.5. Homeobox genes in <i>X. laevis</i>	124
5.6. <i>Xenopus</i> matches to Pax6/ey Regulated Genes	126
6.1. miRNAs identified in EST clusters	135

List of Figures

2.1. Essential steps in the cDNA cloning procedure.	9
2.2. Directional cloning	10
2.3. Sequence and trace file of EST CF549456	13
2.4. NCBI's Entrez view of an EST entry, accession BG410207.	16
2.5. Alternative splicing of genes	20
3.1. Suffix tree	33
3.2. Enhanced suffix array	34
3.3. Data flow in the EST mapping tool e2g	40
3.4. Screenshot of the e2g web interface	43
3.5. Number of clusters and singlets for various settings of <i>Vmatch</i> parameters with 94% <i>identity</i>	47
3.6. Number of clusters and singlets for various settings of <i>Vmatch</i> parameters with 96% <i>identity</i>	48
3.7. Number of clusters and singlets for various settings of <i>Vmatch</i> parameters with 98% <i>identity</i>	49
3.8. Number of clusters and singlets for various settings of <i>Vmatch</i> parameters with <i>leastscore</i> representing 94% identity	53
3.9. Number of clusters and singlets for various settings of <i>Vmatch</i> parameters with <i>leastscore</i> representing 96% identity	54

3.10. Number of clusters and singlets for various settings of <i>Vmatch</i> parameters with <i>leastscore</i> representing 98% identity	55
3.11. Adjusted Rand Index for <i>Vmatch</i> clustering results of data set 10k (option -identity)	64
3.12. Adjusted Rand Index for <i>Vmatch</i> clustering results of data set 20k (option -identity)	65
3.13. Adjusted Rand Index for <i>Vmatch</i> clustering results of data set 40k (option -identity)	66
3.14. Adjusted Rand Index for <i>Vmatch</i> clustering results of data set 80k (option -identity)	67
3.15. Adjusted Rand Index for <i>Vmatch</i> clustering results of data set 10k (option -leastscore)	69
3.16. Adjusted Rand Index for <i>Vmatch</i> clustering results of data set 20k (option -leastscore)	70
3.17. Adjusted Rand Index for <i>Vmatch</i> clustering results of data set 40k (option -leastscore)	71
3.18. Adjusted Rand Index for <i>Vmatch</i> clustering results of data set 80k (option -leastscore)	72
3.19. Adjusted Rand Index for clustering tools applied on different data sets . . .	76
3.20. Multiple comparison analysis for quality evaluation results	78
3.21. Running times for tools <i>CAP3</i> , <i>d2.cluster</i> , <i>PaCE</i> , <i>TGICL</i> , <i>BLASTclust</i> and <i>Vmatch</i> for different data sets	79
4.1. Design of EST clustering pipeline	83
4.2. Full length clone selection and consensus sequence categories	89
4.3. EST clustering database schema (part 1 of 3).	92
4.4. EST clustering database schema (part 2 of 3).	93
4.5. EST clustering database schema (part 3 of 3).	94
4.6. Query interface for the clustering and analysis results	98
4.7. XenDB search result for cluster number 2341	100
4.8. XenDB contig view	101
4.9. XenDB: graphical alignment visualization	102
4.10. XenDB: Result for a search for GO term eye	103

4.11. XenDB species mapping: Identification of potential <i>Xenopus</i> homologues to <i>Drosophila</i> genes.	104
4.12. SQL code for mapping accessions of FASTY hits to cluster contigs.	105
5.1. Contig identified as potential chimera	116
5.2. Comparison of a BLASTX alignment with corresponding full length FASTY alignment	118
5.3. Two examples of contigs derived from clones predicted to have a full length insert (P5P)	121
6.1. miRNA biogenesis	132
6.2. miR-17 cluster in contig 16044	137

Introduction

1.1. Motivation

Since its discovery by Friedrich Miescher in 1869 DNA has been the central object of research for molecular biologists. It took another 84 years until in 1953 Watson and Crick solved the molecular structure of DNA [162] and again 24 more years until the establishment of a sequencing technique by Sanger *et al.* [131], which allowed to determine the nucleotide order of a given DNA fragment. Since then, the sequencing of complete genomes has undergone an amazing development. The genome sequence of *Haemophilus influenzae* was the first complete genome to be deciphered in 1995 and the first eukaryote followed soon after in 1997 with yeast. In 2001 the completion of the human genome was announced and recently the first megabyte of the neanderthal genome could successfully be sequenced [55].

According to the *Genomes OnLine Database* (GOLD) [99], 460 complete genomes have been sequenced and published to date, including 29 archaea, 385 bacteria, and 43 eukaryotes. 1345 genomes projects are currently ongoing. The human genome alone consists of 3 billion nucleotides, a decent part of the more than 67 billion bases reported in the public GenBank database. With next generation sequencing techniques like pyrosequencing and

454 sequencing, and the cost of existing technologies continuing to decline, the amount of data will even grow faster in the future, surpassing Moore's law for the increase of microprocessor computational power. Therefore, it is obvious that the need for efficient algorithms and well organized databases to store and cross-link the information will only increase.

Additionally, a big effort is made to determine the sequences of fragments of genes that have been copied from DNA to RNA (*Expressed Sequence Tags, ESTs*). 227 ongoing EST projects produce huge amounts of data and submit tens of thousands of sequences to public databases each day. ESTs provide the most extensive available survey of the transcriptome of an organism and with it evidence for the existence of genes. They are indispensable for gene discovery, gene structure prediction, and genomic mapping. The price of the low-cost high-throughput data is that ESTs contain high error rates and are not very well annotated. The low quality sequence data can be improved by several processing steps and by clustering into gene-oriented clusters, which again can be assembled to contig sequences for further analyses.

In the first part of this thesis, we will describe an EST clustering pipeline that makes use of enhanced suffix arrays, a data structure that has been shown to be as powerful as suffix trees, with the advantage of a reduced space requirement and reduced processing time. Further on, enhanced suffix arrays have been shown to be superior to other matching tools for a variety of applications. We will validate the clustering results based on a "gold-standard" EST data set of *A. thaliana*. The implemented clustering pipeline takes advantage of the underlying database and enables unique batch functionality of mapping results from other organisms to the species of interest.

For some species, EST projects provide the only information about their gene content. One of these species is the African clawed frog *Xenopus laevis*. Research using this model system has provided critical insights into the mechanisms of early vertebrate development and cell biology. In the former, *X. laevis* has led the way in establishing the mechanisms of early fate decisions, patterning of the basic body plan, and organogenesis. Contributions in cell biology include work on chromosome replication, cell cycle components, and signaling pathways. Despite of the interest in this model organism, no genome project is planned, and EST and cDNA sequences are the only resource available. The Trans-NIH *Xenopus* Initiative therefore agreed on recommendations for future goals to further improve *Xenopus* as a non-mammalian model system. One of the goals of highest priority is the generation of ESTs and full length cDNA collections, as they facilitate functional

assays, one of the particular strengths of *Xenopus*.

We have applied the EST clustering pipeline described in the first part to *X. laevis*, both to identify full length protein encoding sequences and full length cDNA clones. The unique database system supports comparative approaches between *X. laevis* and other model systems, and enables the retrieval of their potential full length clones.

1.2. Structure of the Thesis

Chapter 2 starts with an introduction into the techniques of cDNA cloning and sequencing of expressed sequence tags (ESTs), including common problems. Existing EST databases and typical applications arising from the availability of ESTs are introduced, including EST clustering. The most widely used EST clustering tools are described together with their resulting gene indices.

Chapter 3 introduces the concept of enhanced suffix arrays and the tool *Vmatch*. The application of *Vmatch* to EST mapping and clustering is demonstrated and clustering results qualitatively validated and compared to other clustering tools, using a “gold-standard” data set of *A. thaliana* ESTs.

Chapter 4 presents the design and implementation of a clustering pipeline including a corresponding database schema to store clustering results in a persistent way. The pipeline implements *Vmatch* as the central clustering tool.

Chapter 5 demonstrates the application of the clustering pipeline to a *X. laevis* EST data set. Resulting cluster contigs undergo a comprehensive sequence analysis and the web-based interface XenDB is presented together with a number of examples emphasizing the utility of the clustering and sequence analysis results.

Chapter 6 describes the computational identification of miRNA genes in the clustered EST data set.

Chapter 7 summarizes the conclusion we draw and ends in an outlook of future work arising from this thesis.

1. Introduction

Part I.

Suffix-Array Based EST Clustering

EST Clustering

When Putney *et al.* [125] recognized in 1983 that random cloning and sequencing could provide rapid access to the mRNAs in the cell, they probably did not realize that this would become one of the most widely used methods later called *EST (expressed sequence tag) sequencing*. In 1990, Sydney Brenner proposed that an obvious way of finding at least a large part of the important fraction of the human genome were to sequence the messenger RNAs of expressed genes to provide rapid access to the genes [23]. Critics of this idea countered that cDNA sequencing would miss the regulatory elements that could only be found in the genomic DNA sequence.

The term 'EST' was introduced by Adams *et al.* in 1991 [4] in a publication describing the identification of 337 genes expressed in human brain. Soon after that The Institute of Genome Research (TIGR) generated EST data on a massive scale [3, 5]. Although access to the data was initially restricted, TIGR released more than 100,000 ESTs shortly after to the dbEST database [18] maintained by the NCBI. Among the first large projects in the 1990ies contributing their data were The Genexpress Index [64] with 25,000 brain- and muscle-derived sequences, Merck and Company with more than 528,000 human ESTs [62, 42, 165] and a project funded by the Howard Hughes Medical Institute which produced 216,000 mouse ESTs.

2.1. Expressed Sequence Tags (ESTs)

2.1.1. cDNA Cloning

One of the fundamental techniques of molecular biology is the enzymatic conversion of mRNA to double-stranded DNA and the insertion of this DNA into vectors. After the discovery of reverse transcriptase in 1970 [12, 150], the first clones of complementary DNA (cDNA) were obtained in the mid 1970s. Since then, improvements have been made to the efficiency of synthesis of double-stranded cDNA and the vector systems. The method of cDNA cloning at the time was most widely used to synthesize and clone full-length double stranded cDNAs. Later, Putney *et al.* [125] realized that random cloning and sequencing could provide rapid access to almost all the mRNAs in the cell. Today, cDNA cloning is within the range of any competent laboratory (Figure 2.1 shows the essential steps in the cloning procedure). cDNA libraries can be routinely prepared and methods to identify clones of extremely rare species of mRNAs are available. Improvements of the protocols led to increased sizes of cloned cDNAs, so that it is now possible to isolate full-length cDNAs from all but the longest mRNAs [130].

After isolation of the RNA from the cell, the first strand of cDNA is synthesized by an RNA-dependent DNA polymerase reverse transcriptase using poly(A)⁺ mRNA as a template. Primers used for this reaction are typically: (1) oligo(dT) 12-18 nucleotides in length that bind to the poly(A) tract at the 3' terminus of the eukaryotic mRNA molecules, (2) primer-adaptors containing a homopolymeric oligo(dT) tract at the 3' terminus and a restriction site at the 5' terminus, (3) oligo(dT) primers covalently linked to a plasmid and (4) random primers.

The product of first-strand synthesis (the mRNA-cDNA hybrid) is treated with RNase H which produces gaps in the mRNA strand of the hybrid. The resulting series of RNA primers are used by DNA polymerase I during the synthesis of the second cDNA strand. Next, linker molecules are ligated to the cDNA molecule termini. These molecules are cleaved at a restriction site in the linker and ligated to a vector carrying the cohesive termini compatible with those of the linker (see Figure 2.2).

Primer-adaptors allow for *directional cloning* of the cDNA as shown in Figure 2.2. In this cloning strategy the termini in the ligation reaction are not all equivalent, but produced by digestion with two restriction enzymes with different recognition sequences. The termini of the cDNA fragments will not be complementary and unable to ligate to each other.

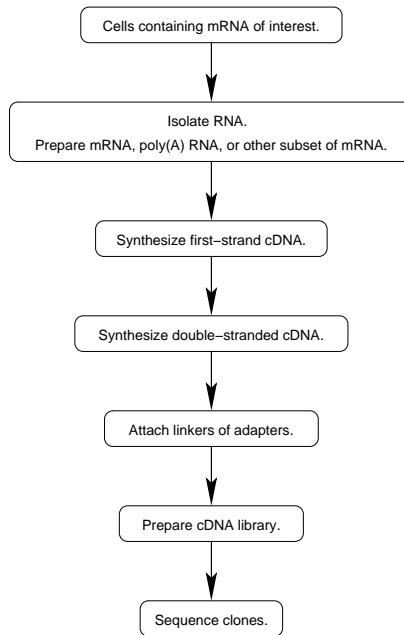


Figure 2.1.: Essential steps in the cDNA cloning procedure.

However, they will ligate to a vector that has been prepared by cleavage with the same two enzymes, generating recombinants containing an insert in a predefined orientation.

If the first-strand cDNA synthesis was incomplete (e.g. in the case of very long mRNAs), the clones will lack sequences from the 5' end of the mRNA. If synthesis of the second-strand cDNA was blocked, the 3' end of the mRNA will be underrepresented. Therefore, sometimes random oligonucleotides are used as primers in the construction of the cDNA library.

Methods of Enrichment

Mammalian cells typically contain between 10,000 and 30,000 different transcribed sequences. Alternative splicing can produce even more different species of mRNA per cell. Not all of these sequences are represented equally: genes that are actively transcribed make a greater contribution to the pool of mRNAs than genes that are transcribed more rarely. Bishop *et al.* define three frequency classes of distributions of mRNAs in a typical somatic cell based on reassociation kinetics analysis: (1) superprevalent (10-15 mRNAs representing 10-20% of the total mRNA mass); (2) intermediate (1,000-2,000 mRNAs; 40-45%); and (3) complex (15,000-20,000 mRNAs; 40-45%) [16, 21].

2. EST Clustering

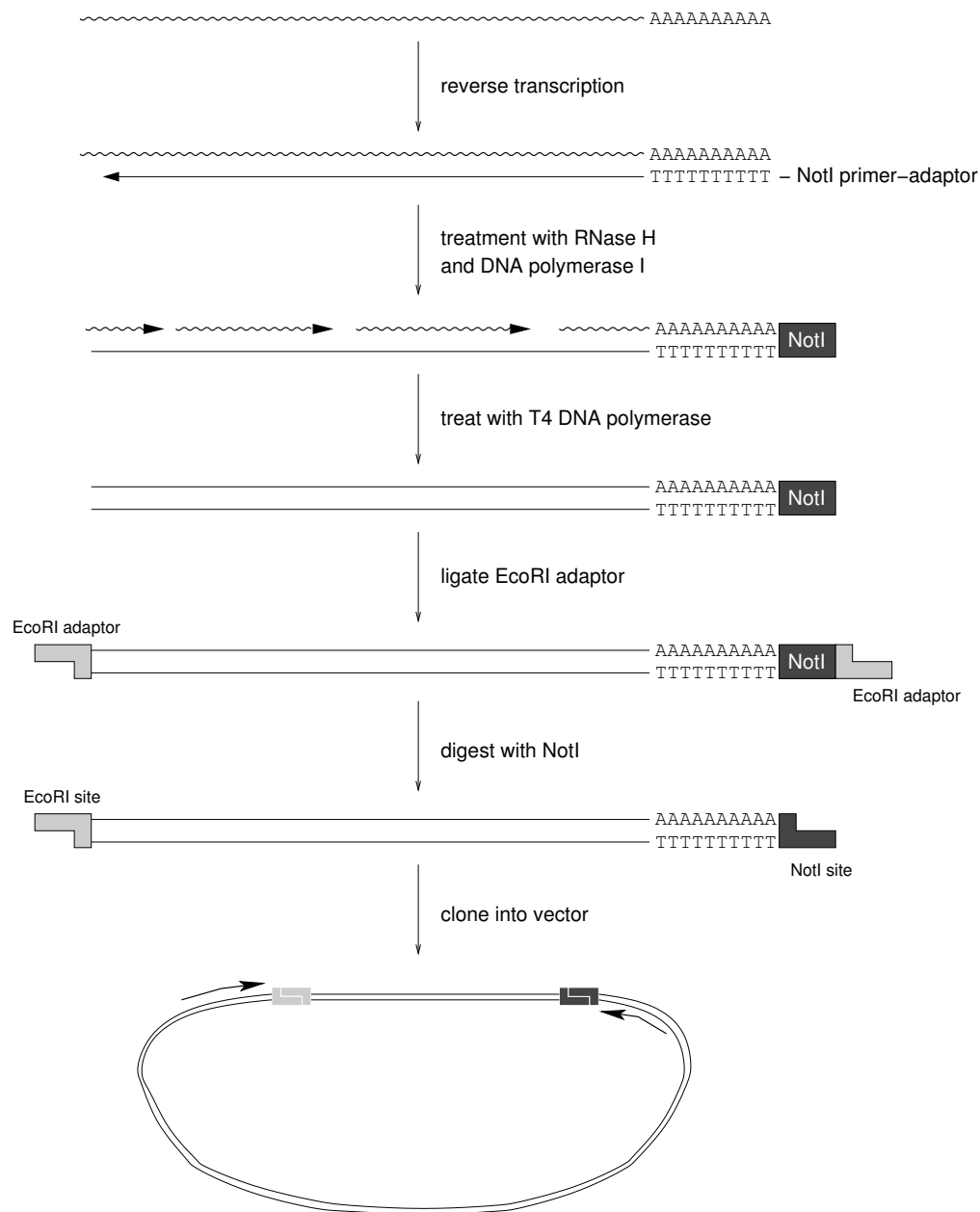


Figure 2.2.: Directional cloning: The first strand of the cDNA is primed by oligo(dT) sequences linked to a primer-adaptor encoding a restriction endonuclease recognition site (in this case, *NotI*). The mRNA sequence is represented by the wiggly strand, the oligo(dT)-adaptor is shown in black. Treatment of the RNA-DNA hybrid with RNase H and DNA polymerase I nicks the RNA moiety so that T4 DNA polymerase can complete the second-strand synthesis. Ligation of *EcoRI* adaptors to the cDNA product and cleavage with *NotI* allow the product to be inserted in a directed manner into the appropriate vector. (Adapted from [130]).

The number of clones required to achieve a given probability that a low-abundance mRNA will be present in a cDNA library is [35]:

$$N = \frac{\ln(1 - P)}{\ln(1 - [1/n])}$$

where N is the number of clones required, P is the probability desired and $1/n$ is the fraction of the total mRNA that is requested by a single type of rare mRNA. For example, Williams [164] analyzed the mRNA population of human fibroblast cells that contain 12,000 different transcribed sequences. mRNAs with <14 copies/cell constitute 30% of the total mRNA, and 11,000 different mRNAs belong to this class. To achieve a 99% probability of obtaining a cDNA clone of an mRNA from this cell line at that low frequency would require a library of 170,000 clones.

Unfortunately, many mRNAs are present at even lower levels (1 molecule/cell is not unusual [130]). Therefore, methods have been developed to enrich the starting population of mRNA molecules for sequences of interest. This allows the size of the library to be reduced.

Subtractive cloning removes sequences from the library that are of no interest. This is achieved by hybridizing single-stranded cDNA prepared from mRNA extracted from the tissue of interest prepared from another source that do not express the genes of interest.

Normalized libraries bring the frequency of occurrence of clones of individual mRNAs into a narrow range of one order of magnitude [139, 21]. The method is based on reassociation kinetics. Rarer species will anneal less rapidly and the remaining single-stranded fraction of cDNA will become progressively normalized during the procedure, reducing significantly the high variation in abundance among the clones of the cDNA library.

2.1.2. EST Sequencing

Expressed Sequence Tags (ESTs)

After the cDNA library has been constructed, individual clones are picked from the library and the cDNA insert is sequenced from each end using primers that hybridize to the vector sequence. The length of these sequences average about 500 bases, which led to the term *expressed sequence tag* (EST) [4]. As the length of an mRNA is usually longer than 500 bases (for human the mean lengths are: coding sequence 1340 bases, 5' UTR 300 bases, and 3' UTR 770 bases [68]), the ESTs represent only fragments of genes and not the

2. EST Clustering

complete coding sequence.

The ESTs can be classified as derived from the 5' or 3' end of the clone if the cDNA has been directionally cloned into the vector. Initially, EST sequencing projects favored the 5' end of the cDNAs because the 5' sequences are likely to contain more protein coding sequence than the 3' ends. 3' ends are mostly poly(dT) primed and therefore contain significant untranslated regions (UTRs). Many genes have very long 5' and 3' UTRs, so that single read sequencing of the 3' end will not reveal any information about the coding potential of the gene. On the other hand, the 3' end of the cDNA clone is sometimes preferred because the 3' UTR offers more unique sequence as it is the most diverse region of the transcript [83, 129], which can be used to distinguish between individual genes and paralogous gene family members that may be closely related in their coding sequences. These unique sequences are better suited for the design of cDNA array-based experiments where the cDNA and its EST are often used if the complete genome sequence is not available. Today, more and more projects sequence both ends of the cDNA.

EST sequencing is an established high-throughput method at many sequencing centers, e.g. the Genome Sequencing Center at Washington University generates about 20,000 ESTs per week.

I.M.A.G.E. Consortium

The *I.M.A.G.E. Consortium* [96] was initiated in 1993 by four academic groups on a collaborative basis after informal discussions led to a common vision of how to achieve an important goal in the study of the human genome: the Integrated Molecular Analysis of Genomes and their Expression. Specifically, the consortium shares high-quality, arrayed cDNA libraries used for EST sequencing and place sequence, map, and expression data on the clones in these arrays into the public domain.

Today, more than half of the ESTs in GenBank are from IMAGE clones. The human and mouse genomes were the first to be studied, and the collection now contains clones from rat, zebrafish, Fugu, *Xenopus* and rhesus macaque. A majority of the clones are publicly available and free of any royalties.

```

>gnl|ti|286854524 name:15596501 CF549456
GCTGGACCGGTCCGGAATTCCCGGATCGAGAAGAGAAGGAGAGATATAGGCACAGCAGAACTGGTTTGTGGTATGTATAATG
GCTAAGCCTTGTGGTAAGTTTACTGCAATCCTCGGGCTAATGGCTTACACTAATTGTATTAGCAATATTCATGTATTATAGCAAT
TATCCCATACAAAAATGTTATTTGATCCAAATANCCGGANGAACATTTTCATCGTTTGTATGTGCACATTACTTTACATTGTAATAT
TAATAATACAAAGTCTGCACCCAAAACAAGTTTATTGGANNAANCTATGTGCCACGACTTATTGTATTATAGATTCCCTGTGCAT
TATGGGTTCCCTAAGGAGCCACAAGAGCTGCTACTATAATAGAACATTTTCccttggatccaattgcaaaatgnngtccttgnanacc
ggntcacnaaaggcacacnggtataaaacngtgggtctgtngggncntnttttaaataacgatatttcaatgcagctcatttattt
taaaaccccttactagncggttaaccagaattaagtacaatgctggcggttttcattggcgacagncatggtttaaaacctgtccaa
tggggcaaccacataaaagncatanggacatggctccttgggcntttttgggggggtggcctggncggntttccccggggcct
ttaaataattcctccangggccctnntaaaaaaaagnagccaaggaannangttncccttttgaaaaagggggaaaaaaaaccccc
tnccttttgggggaaaaaaatnanaatttttnccttttccccnggggncnncncttnaaaaccccccaaaaaacccgnaaaa
aaaaaaannngggggaaaaaaccccccnagnanttttttgggggggggncccccccccggtntntcnggggganaantt
ttaaggggggtgggggnnccggnccaac

```

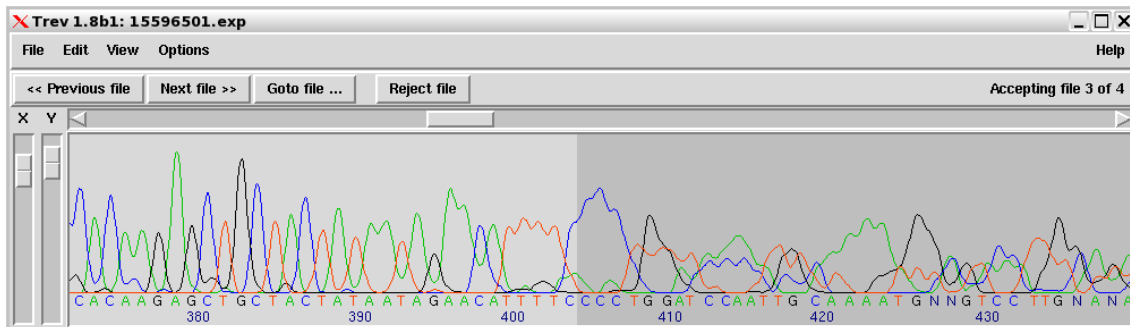


Figure 2.3.: Sequence of EST CF549456 and the corresponding trace file obtained from NCBI's trace archive. The sequence shows a tremendous quality drop around position 405, indicated by the dark gray background color in the trace viewer and lowercase letters in the sequence.

2.1.3. EST Quality

Sequence Quality

ESTs are typically automatically processed, unedited, single-read sequences derived by rapidly sequencing cDNA clones. Consequently, they have a higher error rate than sequences that are verified by multiple sequencing reactions. Compared to the finished portion of the draft human genome sequence which has an error rate of less than 1 in 10,000 bases [68], ESTs have errors on the order of 3% [18]. They contain insertions, deletions and substitutions compared to the mRNA sequence. These errors are usually highest at the beginning and end of the reads and the highest quality portion of the EST sequence is between 100 and 300 bases [62]. Figure 2.3 shows an example of a low quality EST sequence. Database entries sometimes contain annotations about the high quality segments of a sequence read which, when available, can be used to clip the sequences.

Contamination

In addition to the incorrect nucleotides within the EST sequence, parts or even the complete sequence can be incorrect. ESTs can contain vector, bacterial, or mitochondrial (or other structural or regulatory RNA) sequence contamination. Hillier *et al.* [62] found that normalized libraries show comparatively low levels of contamination compared to non-normalized, which had up to 20% bacterial contamination. All of the libraries contained mitochondrial sequences, ranging from 1% to 16% of the ESTs.

Internal Priming

Most cDNA clones are poly(dT) primed and therefore the 3' ESTs should represent the 3' untranslated region of the mRNA. However, a potential problem with this approach is the occurrence of false 3' ends because of internal priming. This can occur as a result of priming to an A-rich region upstream of the poly(A) tail during the reverse transcription step. Aaronson *et al.* [1] and Hillier *et al.* [62] found that 1.5-3% of oligo(dT)-primed 3' ESTs do not align with the known 3' end of the mRNA.

Inverted Clones

Another source of error are inverted clones due to failures either in the directional cloning procedure or in the association tracking between primers and sequence, that lead to mislabeled 5' and 3' ESTs. Up to 6% of the ESTs match a known mRNA in an inverted orientation [62].

Chimeric Clones

Chimeric clones are also a main concern in EST sequencing projects. They could arise during the cloning procedure as a result of artificial fusion of cDNAs derived from different genes. Lane tracking errors, which introduce incorrect associations between sequence and clone, can also be perceived as chimeric clones, when 5' and 3' ESTs are incorrectly assigned to the same clone. Aaronson *et al.* [1] and Hillier *et al.* [62] estimated the frequency of chimeric clones at 1%, another study at 11% [166, 167].

Full-length cDNAs

The only alternative to EST sequencing to circumvent the problems mentioned above and the incomplete sequence coverage of ESTs are full-length cDNA sequencing projects [30, 149, 67]. Full-length cDNA sequences are obtained by shotgun sequencing of the cDNA clones that have been selected for 5' and 3' ends. The underlying redundancy in the shotgun sequences increases the coverage of each individual nucleotide and allows for correcting sequencing errors. Because many reads are necessary, the costs and time for such projects are much higher compared to EST projects.

2.1.4. EST Databases

dbEST [18, 20] is a division of GenBank that contains sequence data and other information on Expressed Sequence Tags from a number of organisms. A brief account of the history of human ESTs in GenBank is available in [17].

Like all sequences in GenBank, ESTs can be accessed in the databases by their unique accession or GI number. Additional functionality can limit the search to sequences from e.g. particular clones they have been derived from, the tissue or cell types, developmental stages, libraries, etc. As it is not generally known in advance if the ESTs come from coding or non-coding parts of the mRNA, sequence characterization and annotation are minimal.

Figure 2.4 shows the Entrez version of an NCBI flat file entry of an EST sequence. In the top block, different IDENTIFIERS, including the accession number and GenBank GI, are shown. The CLONE INFO part of the entry shows the 5' or 3' orientation of the EST, if known (here, 5') and the plate the clone is located in. The PRIMERS section specifies the primer used for the sequencing reaction and if a poly(A) tail could be identified. Next, the SEQUENCE of the EST is shown, in some cases along with a note supplied by the submitter about where the high-quality sequence starts and ends (not shown here). A PUTATIVE ID or annotation can be assigned by the submitter. The LIBRARY block tells from which library the clone was derived (organism, tissue or cell type, developmental stage, etc.), including which vector was used for cloning. If the cDNA was cloned directionally, the different restriction sites are indicated. The entry ends with information about the SUBMITTER and corresponding CITATIONS.

Since the first EST projects (see page 2) the number of ESTs in the public databases increased dramatically. In 1995 the ESTs surpassed the number of non-EST entries and as of November 2006 there are 39 million EST records with 21.4 billion bases in GenBank,

2. EST Clustering

IDENTIFIERS	
dbEST Id:	8131787
EST name:	S10-8-H10
GenBank Acc:	BG410207
GenBank gi:	13506213
CLONE INFO	
Clone Id:	(5')
Plate:	S10-8 Row: H Column: 1
DNA type:	cdNA
PRIMERS	
Sequencing:	SP6-22 5' ctt gat tta ggt gac act ata g 3'
PolyA Tail:	Unknown
SEQUENCE	
	TTACTTCACTTCCACGACCATACCCTCATAGCCGTTTTTCTTATTAGTACGCTAGTTCTT TACATTATTACTATTATAATAACTACTAACTAACTAATAACAACCTCCATAGACGCCCAA GAGATCGAAATAGTGTGAACTATTATACCAGCTATTATCCTTATCATAAATGCCCCCTCCA TCCCTTCGTATTCTATATTTAATAGATGAAGTTAATGATCCACACTTAACAATTAAGCA ATCGGCCACCAATGATACTGAAGCTACGAATATACTAACTATGAGGATCTCTCATTGAC TCTTATATAAATCCAATAATGACCTTACCCCTGGACAATTCGGCTGCTAGAAGTTGAT AATCGAATAGTAGTCCCAATAGAATCTCCAACCCGACTTTTAGTTACAGCCGAAGACGTC CTCCACTCGTGAGCTGTACCCTCCTTAGGNGNCAAAACAGATGCAATCCCAGGACGACTT CATCAAACATCATTTATTGNTACTCGTCCGGGAGTATTTTACGGACAATGTTTCAGAAATT TGGCGGAGNCCACCACAGCTTTATACCAATTGGAGGTGAAGCAGACCGCTAACCGACTTT GAAACTGATCTTTATCAATACTAGAN
Entry Created:	Apr 1 2001
Last Updated:	Apr 1 2001
PUTATIVE ID	
	Assigned by submitter cytochrome c oxidase subunit II (nFL) U33552
LIBRARY	
dbEST lib id:	8754
Lib Name:	Stage 10+ Gastrula Library
Organism:	Xenopus laevis
Develop. stage:	10 - 10.5
Lab host:	DH5alpha
Vector:	pDH105/CS2++
R. Site 1:	Sal I
R. Site 2:	Not I
Description:	Weinstein,D.C., Honore,E., and Hemmati-Brivanlou,A. (1997). Epidermal induction and inhibition of neural fate by translation initiation factor 4AIII. Development 124, 4235-4242.
SUBMITTER	
Name:	Brivanlou, AH
Lab:	Laboratory of Molecular Vertebrate Embryology
Institution:	The Rockefeller University
Address:	1230 York Avenue, New York, NY 10021, USA
Tel:	212 327 8684
Fax:	212 327 8685
CITATIONS	
PubMed ID:	11456444
Title:	Microarray-based analysis of early development in Xenopus laevis
Authors:	Altmann,C.R., Bell,E., Sczyrba,A., Pun,J., Bekiranov,S., Gaasterland,T., Brivanlou,A.H.
Citation:	Dev. Biol. 236 (1): 64-75 2001

Figure 2.4.: NCBI's Entrez view of an EST entry, accession BG410207.

comprising 62% of all sequences. EST projects are available for a diverse collection of organisms, currently there are ESTs from 904 different organisms in dbEST. Table 2.1 shows the number of GenBank nucleotide and dbEST entries as of November 2006 for the top 30 organisms. ESTs provide more than 90% of all GenBank entries for some of these organisms.

2.1.5. EST Uses

The wide range of usage of EST sequences can be seen as complement, but also as alternative to sequencing whole genomes:

Gene Identification

Gene identification remains the most popular use of ESTs. Although sequencing an organism's complete genomic DNA is the only way to access all genes, it is still time consuming and expensive. Early examples show how useful ESTs are for cloning important genes by 'hopping' across taxonomic boundaries from model organisms like *S. pombe* or *D. melanogaster* to human [154, 105, 117]. In these cases database searches provide a more sensitive gene identification than traditional hybridization- or PCR-based gene cloning strategies permit for evolutionary diverged organisms.

In the absence of the complete genomic sequence, cDNAs (and ESTs) remain the only link back to the genome. Boguski *et al.* state that an immediate practical value of interest to a broad range of biomedical researchers was the accelerated cloning of human genes for which homologues in other organisms have already been functionally characterized [18].

Physical Map Construction

PCR or hybridization essays developed from ESTs were used to identify YACs, BACs or other large-insert clones from which genome physical maps are constructed [134]. The mapping of the ESTs onto the physical map immediately identify the genomic regions that contain the corresponding genes. If the genomic regions are linked to disease genes, ESTs can help to identify mutations in the candidate genes.

2. EST Clustering

Organism	GenBank	dbEST	%
<i>Homo sapiens</i> (human)	11,529,395	7,895,572	68.48%
<i>Mus musculus</i> + <i>domesticus</i> (mouse)	8,274,347	4,722,069	57.06%
<i>Oryza sativa</i> (rice)	1,657,455	1,211,064	73.06%
<i>Zea mays</i> (maize)	3,290,677	1,160,485	35.26%
<i>Danio rerio</i> (zebrafish)	1,487,983	1,152,269	77.43%
<i>Bos taurus</i> (cattle)	1,940,528	1,141,099	58.80%
<i>Xenopus tropicalis</i>	1,187,340	1,039,143	87.51%
<i>Rattus norvegicus</i> + sp. (rat)	1,960,756	871,144	44.43%
<i>Triticum aestivum</i> (wheat)	888,045	855,067	96.28%
<i>Arabidopsis thaliana</i> (thale cress)	1,427,866	734,275	51.42%
<i>Ciona intestinalis</i>	699,355	686,396	98.14%
<i>Sus scrofa</i> (pig)	1,266,196	640,034	50.54%
<i>Gallus gallus</i> (chicken)	967,324	599,171	61.94%
<i>Xenopus laevis</i> (African clawed frog)	559,409	542,288	96.93%
<i>Drosophila melanogaster</i> (fruit fly)	703,540	514,613	73.14%
<i>Hordeum vulgare</i> + subsp. <i>vulgare</i> (barley)	470,267	437,321	92.99%
<i>Salmo salar</i> (Atlantic salmon)	433,309	428,803	98.96%
<i>Canis familiaris</i> (dog)	2,590,947	365,909	14.12%
<i>Glycine max</i> (soybean)	645,342	359,402	55.69%
<i>Caenorhabditis elegans</i> (nematode)	384,223	346,064	90.06%
<i>Pinus taeda</i> (loblolly pine)	333,746	329,469	98.71%
<i>Vitis vinifera</i> (wine grape)	427,234	316,756	74.14%
<i>Oryzias latipes</i> (Japanese medaka)	676,968	309,868	45.77%
<i>Aedes aegypti</i> (yellow fever mosquito)	463,072	298,060	64.36%
<i>Branchiostoma floridae</i> (Florida lancelet)	344,718	277,538	80.51%
<i>Gasterosteus aculeatus</i> (three spined stickleback)	306,548	276,992	90.35%
<i>Oncorhynchus mykiss</i> (rainbow trout)	264,387	260,886	98.67%
<i>Malus x domestica</i> (apple tree)	255,763	254,422	99.47%
<i>Pimephales promelas</i>	250,033	249,941	99.96%
<i>Solanum lycopersicum</i> (tomato)	577,798	249,392	43.16%

Table 2.1.: Number of GenBank nucleotide and dbEST entries (November 2006) for the top 30 organisms. Percentages show the fraction of EST sequences compared to all nucleotide entries for each organism.

Annotation of Genomic Sequence

Although ESTs are imperfect, they help predicting and confirming the intron-exon organization of genes, complementing gene prediction programs. ESTs provide valuable experimental evidence of transcription.

Differential Expression

A central question in genomics is the identification of genes associated with tissue differentiation and ontogeny by developing profiles of sequences that are differentially expressed in particular cell types or at different developmental stages. Hybridization methods can be applied to determine differential expression of genes by using cDNA clones as markers, which correspond to EST sequences.

Identification of Gene Homologues

The identification of a gene homologue of a protein derived from a (different) species can elucidate the function of the gene in a model organism. In case of human, identifying an homologue in an animal species allows the evaluation of the potential utility as a model organism for disease studies.

Alternative Splicing

RNA splicing is a post-transcriptional process in eukaryotes prior to mRNA translation. During the transcription, a pre-messenger RNA (pre-mRNA) is produced as a copy of the genomic DNA. It contains the intronic regions, which are removed during the following processing (RNA splicing), as well as exon sequences. During RNA splicing, exons are usually retained in the mature mRNA, but sometimes targeted for removal (see Figure 2.5).

The different combinations can create a variety of mRNAs from a single pre-mRNA, a process referred to as alternative RNA splicing. The alternative splicing events affect the protein coding region of the mRNA and thereby different proteins will be produced during translation. Alternative splicing in non-coding regions of the RNA can result in changes in regulatory elements [100]. EST analyses have shown, that among seven different eukaryotes the amount of alternative splicing is comparable, with no large differences between humans and other animals [24].

2. EST Clustering

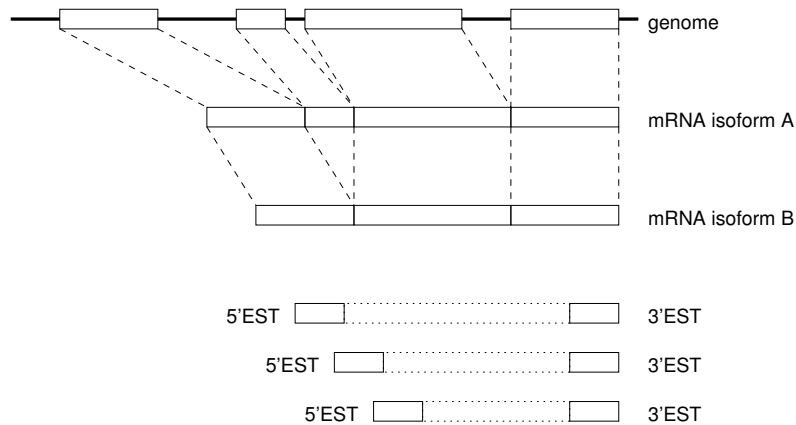


Figure 2.5.: Alternative splicing of genes. The exons of the gene are spliced into two different isoforms, one including the second exon (isoform A), the other missing it (isoform B). The different isoforms can result in two different protein sequences, if the start codon is located in or before the second exon. The different isoforms can be detected by EST analysis, given that different ESTs span the boundaries of exons 1-2, 2-3, and 1-3, which is the case for the 5' ESTs in the figure.

Single Nucleotide Polymorphisms (SNP)

ESTs can also be used as a starting point for detection of single nucleotide polymorphisms (SNPs). A SNP occurs at a frequency of about one in 1,2 Kb between any two individual human genomes [68, 151, 160]. SNPs are an important research tool for genetic association studies and may be used in the development of diagnostic or therapeutic approaches to diseases [156, 108, 98].

2.2. EST Clustering

2.2.1. Clustering Goals

Looking at the large datasets of sequences produced by EST sequencing projects it is obvious that the ESTs cannot each represent a different gene. Even with methods of library normalization abundant transcripts are overrepresented in the EST databases compared to rare ones. The sequences are generally of poor quality with many errors and artifacts, but for many organisms EST collections provide the only information about the coding potential of the genome. Unfortunately, the partial nature of ESTs makes full length cDNA discovery difficult.

Soon after the establishment of the first large EST sequencing projects it became clear

that instead of examining all of the raw data, *gene indices* had to be generated to answer questions like how many transcript groups exist in the total pool of EST sequences or in what tissue are which genes transcribed.

Gene indices try to cluster ESTs into groups that represent the same gene or gene isoform. Reads which are of poor quality or show potential contamination are eliminated. Clusters may be assembled to contig (consensus) sequences, depending on the goal of the gene index. The resulting clusters with all useful information are stored in a query-able database, allowing access to the underlying EST sequences and their corresponding annotations. The reduction of the huge dataset to unique transcripts generates a new source of information with much higher quality as the original input set. A gene index can provide an organized view of the transcriptional state of the tissue and organism from which the ESTs are derived and, hence, makes information available that cannot be derived from sequencing a whole genome alone, but rather supports the analysis of genomic data.

2.2.2. Clustering Procedure

In general, the clustering procedure can be split into four steps: (1) quality control steps are performed to reduce artefactual sequences, (2) pairwise comparisons of all ESTs are conducted to group sequences based on identity, (3) clusters are assembled to contig sequences, and (4) clusters can optionally be joined by further information such as clone annotation.

Pre-processing

The first step in the clustering procedure is the assessment of the sequence quality. Low quality segments of the sequences are often not clipped (see Figure 2.3) thoroughly. Sequencing errors do not necessarily occur as N's but some defined nucleotide in the published sequence. If the trace file is not available, the distinction from SNPs is extremely difficult, unless several reads are available for the same clone.

As described earlier (Section 2.1.3, p. 14) ESTs can contain contamination. Frequently, vector and linker sequences have not been removed properly before submission to the public databases, especially in high-throughput projects. Also, bacterial and mitochondrial sequences might have not been identified and removed.

Another naturally occurring sequence artifact are repeats. These repetitive sequences include LINES (long interspersed elements), SINES (short interspersed elements), ALUs

2. EST Clustering

and satellite repeats. By virtue of the fact that multigene families exist, genes themselves may be repetitive in nature [44]. Fortunately, many of the well-studied members of gene families (HOX and hemoglobin genes) appear to be sufficiently divergent [119] to be distinguished from repeat sequences.

The described sequence artifacts can cause problems during the clustering phase as they would influence the pairwise comparison. Sequences with nearly identical parts but from different genes would be grouped together, creating spurious clusters. Therefore, these sequences must be masked or eliminated during the first pre-processing step.

The removal of contaminants requires a collection of sequences to screen against. VectorDB¹ contains annotations and sequence information for many vectors commonly used in molecular biology. Information for more than 2600 vectors is available. A common problem, though, is that repeats are only well known for model organisms. For those, RepBase [74] is a comprehensive database of repetitive elements from diverse eukaryotic organisms. Currently, it contains over 3600 annotated sequences representing different families and subfamilies of repeats. It can be used with RepeatMasker² to mask repeats in the EST collection.

After identification of an artefactual sequence, a decision has to be made if the ambiguous regions are removed and remaining sequences clustered or if the whole EST sequence is discarded from the data set. Usually, a minimal length of 100 nucleotides remaining after masking is used as cutoff for most gene indices.

Clustering

During the clustering step, the EST sequences are partitioned into subsets (or clusters) based on sequence similarity. Depending on the particular goal of the clustering procedure, two approaches can be distinguished: while stringent clustering parameters using high sequence similarity produce clusters where different isoforms of the gene transcripts are separated, looser clustering joins these isoforms into one cluster. Each approach has its advantages and disadvantages: loose clustering gives a better estimation of the number of genes by grouping different isoforms of the same gene together, but has the risk of clustering paralogous genes.

Genome-based clustering approaches map the ESTs to the genomic sequence and use

¹<http://seq.yeastgenome.org/vectordb/>

²unpublished, A. F. A. Smit, R. Hubley, and P. Green: RepeatMasker at <http://repeatmasker.org>

this information to cluster the sequences. Whenever the genomic sequence is available, this method has the advantage that chimeric sequences and paralogous genes might be identified more easily.

Transcript-based methods do not use genomic information, mostly because it is not available. This approach is the most widely used. Based on a similarity function and a suitable cutoff, sequences are compared pairwise and placed into groups by applying single linkage clustering. The similarity functions are local alignment algorithms like BLAST [7] or Smith-Watermann [138], others use exact substrings [27, 60, 75, 76] or co-linear sets of these [101], or word-based distances [61, 26, 124].

Assembly

Following the clustering step, clusters can be assembled into contig sequences to reconstruct the originating mRNA sequence. This step is not necessarily implemented in the different public gene index databases. An advantage of an assembly is the correction of possible sequencing errors, if enough EST sequences are in a particular cluster and cover the mRNA in a suitable way. Also, full length cDNA sequences can be reconstructed that again can be subject to subsequent sequence analysis or clone selection.

Assembly of EST clusters is commonly performed by genome assembly tools like CAP3 [65] or PHRAP [54]. These tools are designed for assembly of genomic contigs and have some problems with EST clusters, especially when loose clustering is performed. Different isoforms of the same gene result in multiple contigs of the same cluster. More recently, specialized tools for EST assembly became available [32, 102].

Cluster joining

A last step in the pipeline is the optional cluster joining. Here, the information about the shared clone id from 5' and 3' ESTs can be used to further join clusters which do not show sequence similarity, usually because of non-overlap between the 5' and 3' end reads of the clone. Linking by clone information, though, is an error-prone procedure in the EST clustering pipeline as it relies on the accuracy of the annotation and the uniqueness of the clone ids which are not standardized between different sources. Another problem arises from chimeric clones: 5' and 3' reads have their origins in different genes and would join clusters that should not be merged.

2.3. Clustering Tools

Different tools have been used in the past for clustering EST data sets. Some of the tools were originally developed for genome assembly. Since the two problems are related, both EST clustering and assembly tools first try to detect similarities in pairs of sequences. While the assembly tools generate consensus sequences as the last step, EST clustering tools usually produce only clusters of sequences. Hence, in many cases assembly tools are used to generate consensus sequences for the clusters as additional last step after running the clustering tool (see Section 2.4). However, it is not always desirable to form a single consensus for a given cluster of ESTs, as the cluster might represent different isoforms of the same gene due to alternative splicing.

In the following sections the most widely used tools for EST clustering will be briefly introduced. Here, we focus on tools we used in the comparison during the quality assessment in Section 3.5.3. Section 2.4 describes gene indices, that use these tools for EST clustering.

2.3.1. *CAP3*

CAP3 [65] was originally designed as a genome assembly program. The algorithm consists of three major phases: (1) after identification and removal of poor regions of reads, overlaps between reads are computed; (2) reads are joined to form contigs in decreasing order of overlap scores; (3) multiple sequence alignments of the reads of a contig are used to construct consensus sequences.

To speed up the overlap detection of sequence pairs, an overlapping alignment between two reads is simplified as an ordered chain of ungapped parts of the alignment of sufficient length. For the alignment, common words between the reads are identified and (gap-free) extended as far as possible in both directions. Matches are chained and only pairs of reads with chains scoring above a threshold are considered for the next step, where a global alignment is generated for the pair using a banded dynamic programming approach. In the last step, a multiple alignment of overlapping reads is used to generate a consensus as contig sequence.

2.3.2. *d2_cluster*

d2_cluster [61, 26] is the central clustering tool in the STACK pipeline (see Section 2.4.3). It uses the d^2 distance function [152] to detect overlaps between sequences. d^2 is a dissimilarity measure and is derived from the degree to which subsequences are shared between two sequences. The basic distance function for two sequences v and w is defined as

$$d_n^2(v, w) = \sum_{|s|=n} (m_v(s) - m_w(s))^2 \quad (2.1)$$

where n is the length of the subsequences s and $m_i(s)$ the multiplicity of s in sequence i . While the d^2 score combines the values for $d_n^2(v, w)$ for different values of n as follows

$$d^2(v, w) = \sum_{n=l}^u d_n^2(v, w) \quad (2.2)$$

in practice, n is fixed to a suitable value. *d2_cluster* performs pairwise comparisons between all sequences based on this measure and clusters sequences with distances smaller than a given threshold.

2.3.3. *PaCE*

PaCE [75, 76] is a parallel EST clustering algorithm based on the generalized suffix tree [57] data structure. It tries to avoid as many pairwise alignments as possible by first detecting “promising pairs” based on maximal matches. Maximal matches (in contrast to fixed-length word-based approaches) are exact matches of variable length, that cannot be extended at either end. The idea for this approach is that pairs of ESTs with longer exact matches are more likely to produce a suitable alignment in the next phase.

The promising pairs of ESTs are generated in parallel on demand in decreasing order of the maximal match lengths from a distributed representation of the generalized suffix tree. It is not mandatory to perform pairwise alignments of each generated pair, because the sequences might have been clustered already through the agglomerative clustering procedure. This reduces the number of pairs considered for the alignment. When aligning two sequences, the already computed maximal exact match is extended at either end (this time allowing for gaps and mismatches) using banded dynamic programming. The size of the band is determined by the number of errors tolerated, controlling the quality of the alignment.

2.3.4. *BLASTclust* (megaBLAST)

BLASTclust is contained within the standalone BLAST [7, 8] package and can be used to cluster nucleotide sequences using the megaBLAST algorithm. The program determines pairwise matches and places a sequence in a cluster if the sequence matches at least one sequence already in the cluster.

megaBLAST is a special version of BLAST that uses the same greedy algorithm as *Vmatch* (see Section 3.2.3) for the extension of matches during the sequence alignment [170]. It is optimized for aligning sequences that differ only slightly as a result of sequencing or other similar "errors".

BLASTclust takes as input a FASTA file of sequences, produces a temporary BLAST database, runs megaBLAST to identify matching sequences and performs the clustering. *BLASTclust* has several parameters that can be used to control the stringency of clustering: thresholds for score density, percent identity, and alignment length.

2.3.5. *TGICL*

The TIGR Gene Indices clustering tools (*TGICL*) [122] cluster and create assemblies (contigs) from a set of DNA sequences given in a FASTA file. The clustering phase performs pairwise alignments (using a slightly modified version of megaBLAST), which are then filtered and used to build subsets of sequences by a transitive closure approach. In the assembly phase each cluster is sent to the assembly program (*CAP3*) which generates a multiple alignment of the sequences in the cluster and creates one or more contig sequences. Both clustering and assembly phases can be executed in parallel on multiple CPU machines or in a PVM (Parallel Virtual Machine) environment.

2.4. Gene Indices

After the clustering procedure, which reduces the mass of the raw low quality data into manageable quantities, *gene indices* help to organize and analyze the EST datasets by storing the results in query-able databases. While many gene indices were established in the last years (e.g. [42, 29]), only the most popular three will be discussed further: NCBI's UniGene, TIGR's Gene Indices and STACK. Most of them perform the series of steps described earlier (see Section 2.2), although each index makes different assumptions on what forms a cluster.

2.4.1. NCBI UniGene

UniGene is an “experimental system for automatically partitioning GenBank sequences into a non-redundant set of gene-oriented clusters” [19, 133, 123]. Each UniGene cluster contains sequences that represent a unique gene, and is linked to related information, such as the tissue types in which the gene is expressed, model organism protein similarities, the LocusLink report for the gene and its map location. Currently, UniGene clusters ESTs from a total of 63 organisms. UniGene databases are updated weekly with new EST sequences, and bimonthly with newly characterized sequences.

As first step during the build procedure, sequences of foreign origin (such as *E. coli*) are eliminated, and regions derived from cloning vectors or linkers and rRNAs or mitochondrial sequences identified. High-quality segments of each sequence are identified through the assigned base-level error probabilities if available at NCBI’s Trace Archive³. Simple repeats and low complexity regions are masked using DUST [58], transposable repetitive elements by comparison with libraries of organism-specific repeats. After screening, a sequence must contain at least 100 high quality unmasked bases to be included in UniGene. Builds are either transcript based or genome based, here we focus on the transcript based procedure.

The input dataset of UniGene not only contains ESTs but also mRNA sequences. After screening, mRNAs are clustered into *gene links*. Pairs that are “sufficiently similar” are linked to form initial clusters, however, the amount of similarity has not been defined exactly. Next, ESTs are compared to these initial clusters using megaBLAST [170], and “sufficiently similar” sequences added to these clusters. Links that would join initial mRNA-based clusters are discarded, to prevent non-biological chimeric sequences from creating artefactual clusters. Also, EST-to-EST links are generated that either extend the initial clusters or generate clusters containing ESTs only.

The third step adds clone-based edges to allow non-overlapping 5’ and 3’ ESTs to be assigned to the same cluster. Because of imperfect clone labeling, double linkage clustering is used in this step, i. e. at least two 5’ ESTs from one cluster have to be linked to at least two 3’ ESTs from another cluster by their clone ID.

The next step discards clusters that are not *anchored* at the 3’ end of a transcription unit. Therefore, a cluster has to contain a sequence with a polyadenylation signal, a poly(A) tail, or at least two ESTs labeled as having been derived using the 3’ primer.

³<http://www.ncbi.nlm.nih.gov/Traces/trace.cgi?>

2. EST Clustering

Finally, ESTs that do not belong to a cluster are rechecked at a lower stringency and added to the cluster that contains the sequence it best matches to. These non-stringent parameters produce clusters that represent a single gene and may contain more than one alternative-splice form. No attempt is made to produce contigs or consensus sequences, one reason being that clusters can contain 5' and 3' sequences from the same clone, but these sequences do not always overlap. In contrast, each cluster gets a *representative sequence* assigned, which is the longest sequence within the cluster. If the cluster does not contain a full length mRNA, the representative sequence will only be a portion of it, losing the 5' or 3' information.

2.4.2. TIGR Gene Index

The TIGR Gene Index Project “creates organism specific databases aiming to provide an analysis of publicly available EST and gene sequence data to identify transcripts” [126]. Currently, TIGR clusters ESTs from a total of 89 organisms.

For each gene index, the EST sequences are extracted from NCBI's dbEST database. The new sequences are screened for vector and *E. coli* sequences, poly(A) trimmed, and sequences with at least 100bp and less than 3% Ns remain. *Expressed Transcript* (ET) sequences are extracted from appropriate divisions of GenBank and participate in the clustering and assembly process along with the cleaned ESTs. ESTs and ETs are clustered if they share a minimum of 95% identity over at least 40 base pairs, identified by megaBLAST. Additionally, a maximum unmatched overhang of 20 base pairs is allowed.

A significant difference to UniGene is that TIGR clusters are assembled into *Tentative Consensus* (TC) sequences using CAP3. These assemblies tend to represent one transcript, i. e. alternative isoforms end up in different clusters. For that reason, several TIGR TCs may be contained within one UniGene cluster.

TCs are searched against non-redundant protein and nucleotide databases to provide provisional functional assignments, Gene Ontology (GO) terms are assigned based on the best hits.

The TIGR Gene Indices are updated three times a year subject to the condition that more than 10% or more than 25,000 new sequences became available since the last release (ESTs or gene sequences), or the index is more than one year old.

2.4.3. STACK

The STACK project at the South African National Bioinformatics Institute (SANBI) “aims to generate a comprehensive representation of the sequence of each of the expressed genes in the human genome by extensive processing of gene fragments to make accurate alignments, highlight diversity and provide a carefully joined set of consensus sequences for each gene. The STACK project is comprised of the STACKdb human gene index, a database of virtual human transcripts, as well as stackPACK, the tools used to create the database” [107, 33]. Currently, clusters are available for human ESTs only.

Unlike UniGene or TIGR Gene Indices, STACK separates ESTs by tissue type in the first step, allowing to explore transcript expression in specific tissues. Next, sequences are masked against human repeat sequences available at RepBase [73], vector sequences, mitochondrial and ribosomal sequences removed once identified using `cross_match` [54]. Sequences with less than 50 base pairs are discarded.

Clusters are formed by pairwise comparison of all sequences using the word-based, greedy clustering algorithm `d2_cluster` [61, 26, 39]. Two sequences fall into the same cluster if they share a 150 base pair segment with at least 96% identity. Clusters are assembled using PHRAP [54].

2. EST Clustering

Suffix Array Based EST Mapping and Clustering

3.1. Motivation

Apart from clone pair information, sequence similarity is the main indicator for clustering ESTs. Basically, two different approaches can be used for EST clustering: (1) *genome-based* and (2) *transcript-based* clustering methods. Genome-based methods can be applied if (part of) the genomic sequence is available. ESTs are then mapped to genomic loci by comparing the transcripts to the genomic sequence. This method also helps in annotating exon-intron structures of genes. If genomic sequence is not available, transcript-based methods perform an all-against-all pairwise comparison of the ESTs, where significant similarity in an overlap is interpreted as indication that ESTs are derived from the same gene.

Sequence similarity can be determined by pairwise alignment algorithms using standard dynamic programming methods allowing for insertions, deletions and mismatches. The running time of these algorithms is quadratic in the lengths of the sequences. Applied to all possible pairs of ESTs in an EST index this approach is too expensive for most data

sets, especially for organism with hundreds of thousands of ESTs available. Therefore, alternative methods of fast overlap detection are needed.

Genome assembly tools like CAP3 and PHRAP for instance use exact string matching to identify potential pairs which will be further aligned using standard dynamic programming. As a matter of fact they are frequently used to cluster ESTs and produce satisfactory results if the sequence sets are not too large. These tools work well when the fragments represent a random sampling of the DNA, but for ESTs the coverage is not uniform, as it depends on the level of expression of the gene. The transitive closure clustering can avoid an all-against-all comparison, but in the worst case, the number of overlaps is still quadratic in the number of ESTs [76].

We use a different approach for overlap detection: *enhanced suffix arrays* [2]. Enhanced suffix arrays are related to suffix trees, which were first described by Weiner [163]. Gusfield [57] devotes 120 pages of his book to the description and applications of suffix arrays, impressively demonstrating that suffix trees are one of the most important data structures in string processing. In this chapter we will describe the basic idea of enhanced suffix arrays and present how they can be successfully applied to the problem of EST mapping and clustering.

3.2. Enhanced Suffix Arrays

3.2.1. Suffix Trees and Suffix Arrays

Basic Definitions

Let S be a string of length $|S| = n$ over a finite alphabet Σ . The special symbol $\$$ is an element of Σ but does not occur in S . Following [2] we suppose that the size of the alphabet is constant, and that $n < 2^{32}$, which implies that an integer in the range $[0, n]$ can be stored in 4 bytes. If $S = uvw$ for some strings u, v, w , we call u a *prefix* of S , v a *substring* of S , and w a *suffix* of S .

Suffix Trees

A *suffix tree* for the string S is a rooted directed tree with $n + 1$ leaves numbered 0 to n . It represents a substring index for the string S . Each internal node, other than the root, has at least two children and each edge is labeled with a nonempty substring of $S\$$. No

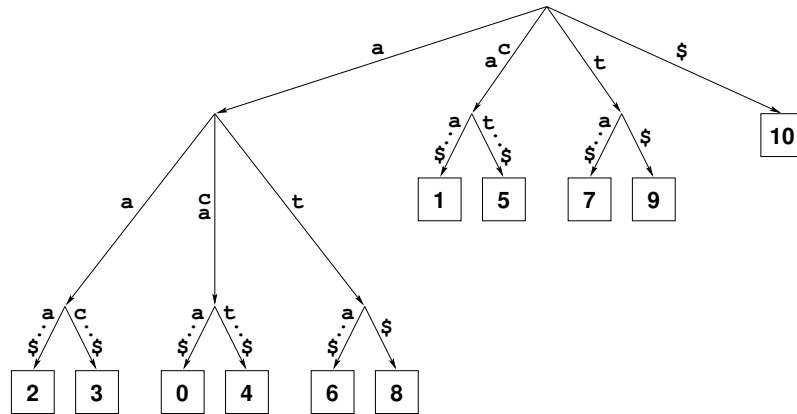


Figure 3.1.: The suffix tree for string $S = acaaacatat$ (adopted from [2]).

two edges out of a node can have edge-labels beginning with the same character. The suffix tree represents all suffixes of $S\$$, which can be retrieved by concatenating the edge-labels on the path from the root to a leaf. Figure 3.1 shows the suffix tree for the string $S = acaaacatat$.

The suffix tree can be built in $O(n)$ time and space [163, 155] if the alphabet is assumed to be constant. The suffix tree can then be used to solve a countless number of matching tasks, e.g.:

- the occurrence of a string P of length m in the string S can be checked in $O(m)$ time, i.e. the running time is independent of the length of the string S
- all z occurrences of P in S can be found in $O(m + z)$
- the longest common substring of two strings S_1 and S_2 can be found in $O(S_1 + S_2)$ time

This time bound is not achievable by most other algorithms like BLAST or *d2.cluster*, which are at most linear in the length of the text. In typical bioinformatics applications n is usually huge compared to m , making even those algorithms impractical for some applications. Especially if the text is a fixed set of strings (like genomic or EST sequences), suffix trees are the favorite data structure for matching tasks.

There are, however, some practical drawbacks: although the theoretical asymptotic space efficiency is linear, the space consumption of a suffix tree is quite large in practice. Even improved implementations still require 20 bytes per input character in the worst

3. Suffix Array Based EST Mapping and Clustering

i	suftab	lcptab	$S_{\text{suftab}[i]}$
0	2		aaacatat\$
1	3	2	aacatat\$
2	0	1	acaaacatat\$
3	4	3	acatat\$
4	6	1	atat\$
5	8	2	at\$
6	1	0	caaacatat\$
7	5	2	catat\$
8	7	0	tat\$
9	9	1	t\$
10	10	0	\$

Figure 3.2.: The enhanced suffix array of the string $S = \text{acaaacatat}$ (adopted from [2]).

case [88]. (In case of the human genome for instance, Kurtz [88] estimates the memory requirements to build a suffix tree for the complete genomic sequence to be approximately 45 gigabytes.) Another drawback is the poor locality behavior of memory reference, causing a significant loss in efficiency on cached processor architectures and makes it difficult to store in secondary memory [2].

Suffix Arrays

A more space efficient data structure which is highly related to suffix trees are suffix arrays [53, 103]. A *suffix array* `suftab` of a string S is an array of integers in the range 0 to n , specifying the lexicographic ordering of the $n + 1$ suffixes of S ; namely $S_{\text{suftab}[0]}, S_{\text{suftab}[1]}, \dots, S_{\text{suftab}[n]}$ where `suftab`[k] denotes the start position of the k th smallest suffix in the set of suffixes of S , and $S_i = S[i..n - 1]$ is the i th suffix of S . $S_{\text{suftab}[0]} < S_{\text{suftab}[1]} < \dots < S_{\text{suftab}[n]}$, where “ $<$ ” denotes the lexicographical order. See Figure 3.2 for an example of a suffix array for the string $S = \text{acaaacatat}$.

A suffix array for a string of length n can be built in $O(n)$ time [77, 80, 84] and requires only 4 bytes per input character. All occurrences of a pattern P of length m in a string S can be found in $O(m \log n)$ time. By adding an extra table holding the information about the longest common prefixes (*lcps*), this running time can be improved to $O(m + \log n)$ [103].

3.2.2. Enhanced Suffix Arrays

Abouelhoda *et al.* [2] took the idea of adding additional tables to a suffix array a step further and developed the idea of enhanced suffix arrays, comparable to the approach of Manber *et al.* [103] of “enhancing” a suffix array by information about the *lcps* of adjacent elements, which reduces the time complexity for a trivial search from $O(m \log n)$ to $O(m + \log n)$.

An *enhanced suffix array* is a data structure consisting of a suffix array and additional tables:

- *lcp-table*

The lcp-table is an array of integers that defines the length of the longest common prefix of $S_{\text{sufstab}[i-1]}$ and $S_{\text{sufstab}[i]}$. It can be computed as a by-product during the construction of the suffix array and requires $4n$ bytes.

- *lcp-interval trees*

The lcp-interval tree is a conceptual (or virtual) tree that allows to simulate all kinds of bottom-up traversals of a suffix tree.

- *child-table*

The child-table allows to simulate any top-down traversal of the suffix tree by means of the enhanced suffix array. It can be computed in linear time and requires n bytes in practice

- *suffix link-table*

The concept of suffix links is incorporated by the suffix link table which requires $2n$ bytes in practice.

Abouelhoda *et al.* show that every algorithm that is based on a suffix tree data structure can be systematically replaced with an algorithm based on an enhanced suffix array that solves the same problem in the same time complexity [2]. E.g. the basic suffix array enhanced with the lcp-table and the child-table allows to find all z occurrences of a pattern P in S in optimal $O(m + z)$ time.

3.2.3. *Vmatch*

The concept of enhanced suffix arrays has been implemented in the versatile software tool *Vmatch*¹, which allows for efficiently solving large scale sequence matching tasks. The most important features of *Vmatch* are [89]:

- Persistent index
Often, large portions of the sequence set under consideration are static and do not change much over time. Therefore it makes sense to preprocess and extract information that is then stored in a data structure that allow efficient access. *Vmatch* preprocesses the set of sequences into an index structure which is stored as a collection of several files constituting the *persistent index*. The index efficiently represents all substrings of the preprocessed sequences and, unlike many other sequence comparison tools, allows matching tasks to be solved in time, *independent* of the size of the index.
- Alphabet independency
Unlike other software tools, *Vmatch* can process sequences over any user defined alphabet not larger than 250 symbols. *Vmatch* implements the concept of *symbol mappings*, denoting alphabet transformations. These allow the user to specify that different characters in the input sequences should be considered identical in the matching process. This feature, for example, will be used to keep the sequence information of masked repeat sequences in the input, but prevent these regions from being matched.
- Versatility
Vmatch allows a multitude of different matching tasks to be solved using the persistent index. Every matching task is basically characterized by (1) the *kind of sequences* to be matched, (2) the *kind of matches* sought, (3) additional *constraints* on the matches, and (4) the *kind of postprocessing* to be done with the matches.
- Match selection and customized output
Matches can be selected according to their length, E-value, identity, or match score. Postprocessing allows e.g. *masking* of substrings covered by a match, *inverse output* (i.e. out put of substrings *not* covered by a match), or *clustering* of sequences

¹see <http://www.vmatch.de/>

according to the matches found. Several options allow for output customization, e.g. XML output.

X-Drop Extension Strategy

The suffix array allows to rapidly identify exact matches between a query sequence and the index. These initial matches can be extended further to produce gapped local alignments. Standard dynamic programming algorithms for pairwise alignments perform a fixed amount of computations to fill the dynamic programming matrix. To further increase the speed of the search, *Vmatch* can extend exact matches by either of two strategies: (1) the *maximum error* extension strategy, as described in [90] for repeat detection, and (2) the *greedy* extension strategy described by Zhang *et al.* [170].

Here, we will briefly introduce the second strategy, which can be much faster than traditional dynamic programming approaches when aligning DNA sequences that differ only by few errors, such as sequencing errors. It is related to a banded dynamic programming algorithm like the one described in [31] which restricts the region of the matrix to be explored. The main idea of [170] is to consider only matrix positions for which the optimal local alignment score does not fall more than X below the best score seen so far, hence the name *X-Drop*. Starting from an exact match (the *seed*), the alignment is calculated forward and backward. The advantage of the X-Drop approach is the adaption of the region being explored while the alignment is being constructed. The choice of X influences the size of the search space. When the aligned sequences are similar, the region is small, mismatches cause the region to be expanded, depending on the value of X . The greedy algorithm does not guarantee that the highest scoring alignment is found, but in practice the alignments score very close to the optimal.

3.3. e2g - EST Mapping

High throughput cDNA and EST sequencing projects have generated a vast amount of data representing the transcribed portion of the organisms in study. As soon as (parts of) the sequence of the associated genome becomes available, the cDNA and ESTs are mapped to the genomic sequence to e.g. detect new genes, verify the exon-intron structure of predicted genes, and determine splice variants.

Mapping ESTs or cDNAs to a genomic sequence is a standard task in molecular biology,

and there are several tools available for this task (see e.g. [113, 45, 158, 51, 78, 40, 93]). Most of these tools are developed for small scale tasks, where the sensitivity was the main design goal. The essential step is usually a costly dynamic programming method with a running time quadratic in the size of the input. Therefore, some tools apply filtering methods first. These scan the ESTs in linear time to find regions containing highly conserved matches to the genomic sequence. Unfortunately, none of the existing tools can efficiently handle complete EST collections of vertebrates with millions of ESTs. This is because the fast filtering methods (if any) still have to scan the entire EST collection. Moreover, there are only few tools available which provide a comprehensive graphical representation of the sometimes contradicting mappings of the ESTs or cDNAs to the genomic sequence. A good example of such a visualization tool is SpliceNest [36], which however only allows to visualize static datasets.

3.3.1. Design Rationale

To make use of *Vmatch* for efficiently mapping EST sequences to their genomic locations, we developed the web-based tool *e2g*, which provides both, efficient mapping of user provided genomic sequence and convenient visualization. *e2g* is conceptually different from other mapping tools in that it provides an *EST collection as target database*. All other mapping tools provide the genomic sequence as target and only a limited number of ESTs can be uploaded for a single matching task. Especially in cases where the focus of the application is on identifying new genes or splice variants of known genes in a genomic region of interest, *e2g* allows rapid identification of ESTs for further analyses.

Figure 3.3 shows the seven phases of an *e2g* analysis:

1. *Data import*
Import ESTs and cDNAs and build index structure.
2. *User upload*
Users upload genomic sequence and optionally ESTs and annotation.
3. *Pre-processing*
Masking of simple and organism-specific repeats in genomic sequence.
4. *Gene prediction*
Gene prediction in genomic sequence.

5. *EST Mapping*

Mapping of ESTs and cDNAs to genomic sequence.

6. *Filter*

A filter is applied to remove spurious hits and reduce output size.

7. *Visualization*

Matches are graphically visualized with further analysis options.

The first step of downloading the EST collections and building the enhanced suffix arrays is only done once. These indices are then used each time a user performs steps 2 to 7 via the Web interface. EST and cDNA collections are downloaded from NCBI. The indices allow to find highly conserved matches between the genomic sequence and the EST collection much faster than a scanning based method. Indices are precomputed using *mkvtree* and stored on disk. The user uploads a genomic sequence and either chooses an EST collection of an existing index, or optionally uploads a custom EST/cDNA data set. After pre-processing of the genomic sequence (repeat masking and gene prediction), it is then matched against the index using *Vmatch*. Results are stored in a relational DBMS. The web server generates overviews of the mapping, for the region chosen by the user. The mapping of the ESTs or cDNAs is visualized as colored blocks (representing the length and direction of the matches) relative to the genomic sequence. The user can interactively explore the set of matches by zooming into regions of interest. Alignments of selected ESTs are computed on demand by *Vmatch* or *sim4*, respectively.

3.3.2. Implementation

e2g can be used in two different basic modes:

1. the user uploads a genomic DNA sequence and chooses one of the EST collections available on the server. Currently, EST collections for *H. sapiens* ($\approx 6 \cdot 10^6$ ESTs), *M. musculus* ($\approx 4.3 \cdot 10^6$ ESTs), *C. elegans*, *C. briggsae* ($\approx 3 \cdot 10^5$ ESTs), *X. laevis* contigs from EST clustering (see Chapter 5), and corresponding index structures are available.
2. the user uploads a genomic DNA sequence as well as the cDNA/EST collection to be mapped. In this case, the index structure for the collection is first computed by the server.

3. Suffix Array Based EST Mapping and Clustering

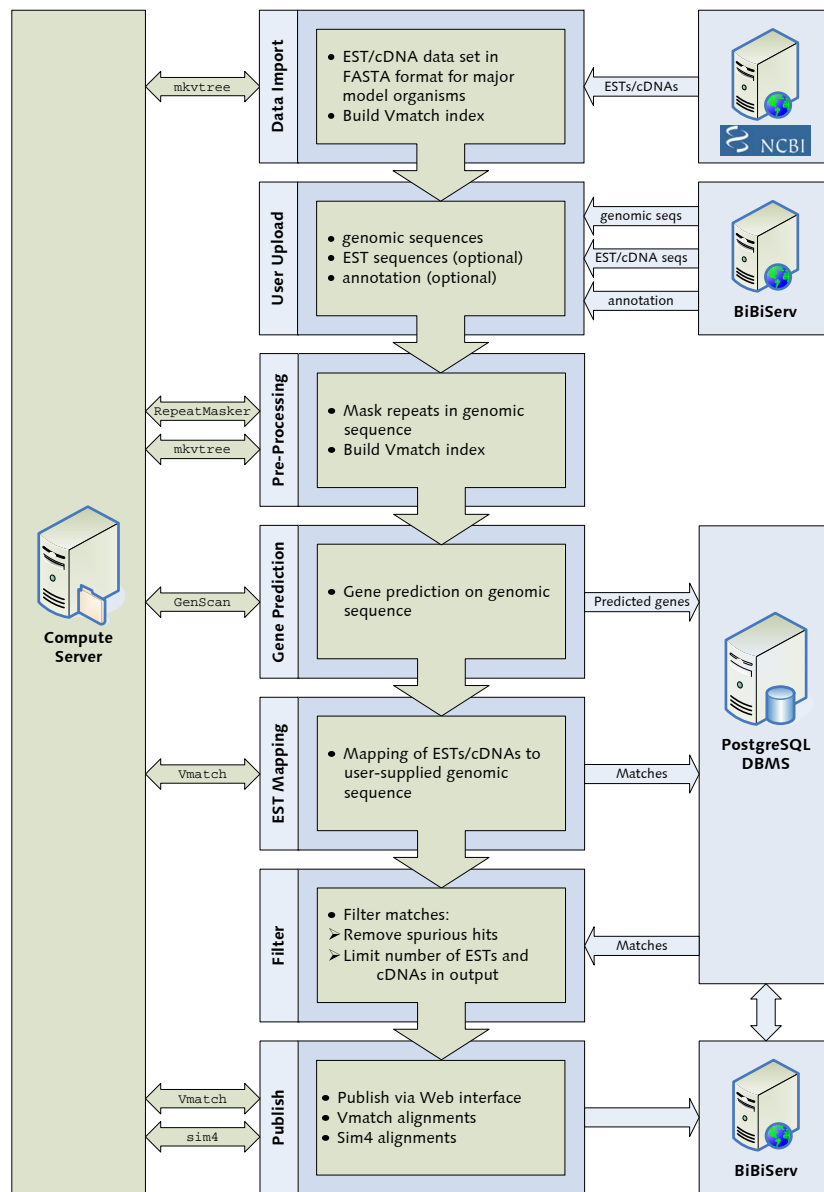


Figure 3.3.: Design rationale of e2g. EST and cDNA collections are downloaded from NCBI, indices precomputed using *mkvtree* and stored on disk. The user uploads a genomic sequence and either chooses an EST collection, or uploads a custom EST/cDNA set. After pre-processing of the genomic sequence (repeat masking and gene prediction), it is then matched against the index using *Vmatch*. Results are stored in a relational DBMS. The web server generates overviews of the mapping, for the region chosen by the user. An alignment of the matching sequences, and a spliced alignment of a selected EST is computed on demand by *Vmatch* and *sim4*, respectively.

In both modes, *RepeatMasker* is optionally applied to the uploaded genomic sequence to mask repeats. Also, *GenScan* [25] is run to obtain an initial ab-initio gene prediction. Furthermore, the user can upload a file containing a gene annotation of the corresponding genomic sequence. In the following we will assume the first mode, because we expect it to be the standard mode.

The index structure is an enhanced suffix array stored on several files. It provides rapid access to all substrings of the ESTs of the given collection. The enhanced suffix array is precomputed only once, and can be used many times for different genomic sequences.

Given the enhanced suffix array, *Vmatch* matches the genomic DNA against the enhanced suffix array to obtain exact matches. These are extended using the *X-drop* extension strategy. This gives highly conserved matches between ESTs and the genomic sequences. Sometimes there is a large number of spurious hits, typically caused either by DNA contaminations in the EST library, or by repeats missed by *RepeatMasker*. Therefore, a match is discarded if there is no other match in the same EST. For convenient and fast access, the positions of the remaining matches are stored in a relational database. The matches represent a mapping of a subset of the ESTs to specific positions on the genomic sequence. The user can select an EST and a spliced alignment for the EST is computed on the fly using *sim4* [45]. This allows for the detection of splice site signals. Running *sim4* on a single EST and a small region of the genomic sequence does not add much to the running time of *e2g*.

All steps in the *e2g* data flow (Figure 3.3) are implemented using web services technology. *e2g* is available at the Bielefeld University Bioinformatics Server² [87]. The *Vmatch* jobs run on a Sun Solaris compute server with eight 800 MHz UltraSparc III CPUs and 64 GB of RAM. The web interface and its underlying CGI framework are implemented as messaging services. This allows to easily integrate more servers if necessary and develop standalone clients which are independent of the web interface and can be used in an automated way.

3.3.3. Web interface

Figure 3.4 shows a screenshot of a graphical overview produced by *e2g* when uploading a 16.5Kbp genomic sequence from *M. musculus* (Genbank GI: 28515921, bases 60,000-76,500) to compare it to 4.1 million ESTs from the same species. The overview is split

²<http://bibiserv.techfak.uni-bielefeld.de/e2g/>

3. Suffix Array Based EST Mapping and Clustering

into five panels, arranged from top to bottom:

1. *General Information Panel*: The top of the window provides general information about the current task. The user can zoom into a region of interest within the submitted genomic sequence. The positions of the highlighted matches in the EST and the genomic sequence are displayed. This part of the overview also provides links to download the sequences or GI numbers of matching ESTs.
2. *Annotation Panel*: The second section of the window shows gene predictions for the genomic sequence, as uploaded by the user (orange colored) and delivered by *GenScan* (blue colored). If the prediction refers to the forward strand, then the exons are shown above the line representing the genome, otherwise below.
3. *cDNA Mapping Panel*: cDNA matches on the genomic sequence are shown as colored blocks. Forward matches are shown in green, reverse complemented matches are shown in red.
4. *EST Mapping Panel*: EST matches on the genomic sequence are shown in the same way as cDNA matches. The two kinds of matches are separated since cDNAs are usually of higher quality and thus matches to the genomic sequence are more reliable.
5. *Mapping Summary Panel*: The bottom panel provides a summary of all matches, shown as colored boxes. The color code represents the coverage of a region, i.e. the relative number of matches in the region. For example, in Figure 3.4, regions with high coverage are represented by red boxes and regions with low coverage by blue boxes.

The *GenScan* and uploaded annotation from the annotation panel can be superimposed to the cDNA and EST matches, by dragging a transparent image over the lower part of the window. The transparent image conveniently allows the user to compare the gene prediction to the matches found.

By clicking on a match, an alignment (computed by *Vmatch*) between this individual region of the EST and the genomic sequence is shown in a popup window. The alignment is supplemented by additional information such as positions in the genomic sequence and in the EST, scores, identity values, and E-values (see Figure 3.4, bottom). Additionally, *sim4* can be invoked to produce a spliced alignment over the whole EST sequence.

3.3. e2g - EST Mapping

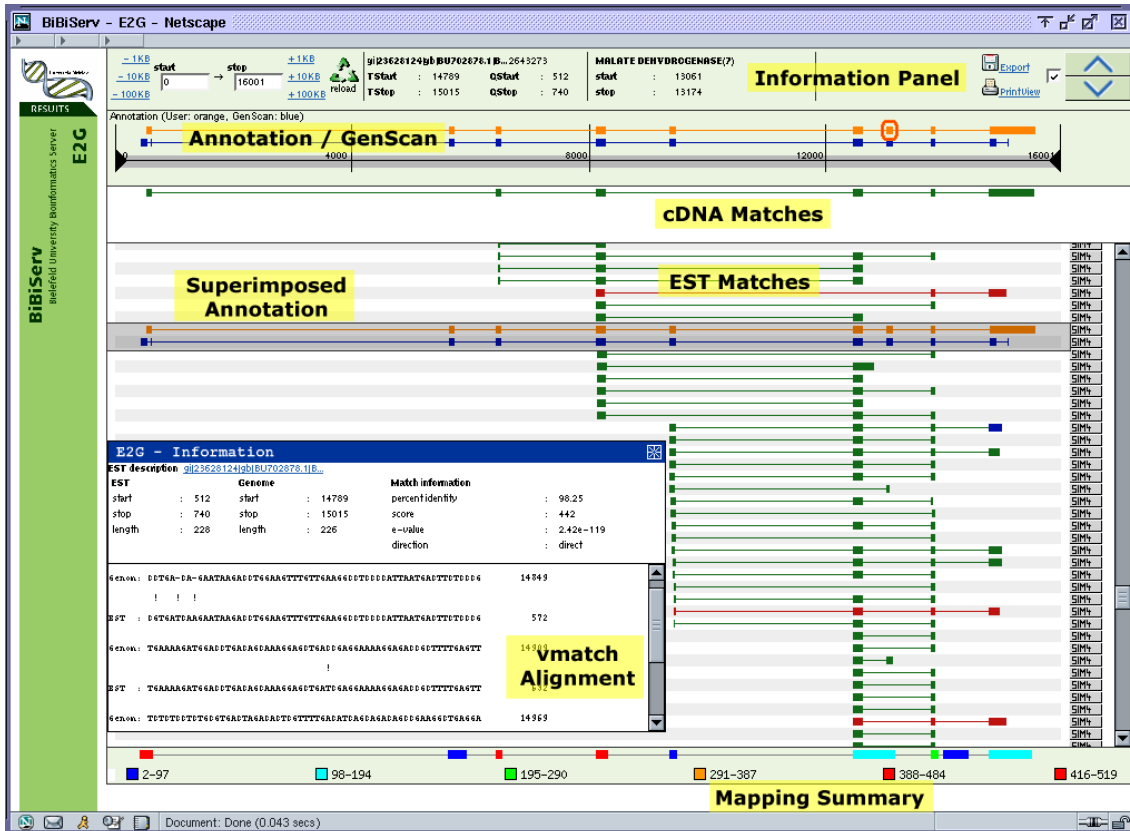


Figure 3.4.: Screenshot of the e2g web interface showing the mapping of mouse ESTs/cDNAs to mouse genomic sequence (Genbank GI: 28515921, bases 60,000-76,500). The information is split into five panels: general information, gene annotation, cDNA matches, EST matches, and mapping summary (from top to bottom). Forward matches are shown in green, reverse complemented in red. The transparent gray shaded box in the middle of the image contains the annotation uploaded by the user. It can be moved over matching ESTs to further inspect its exon/intron structure. The popup window shows the *Vmatch* alignment of the blue highlighted exon in the EST match panel.

3.3.4. Performance Evaluation

For our performance experiments, we mapped all ESTs from *M. musculus* (total length 1.87 Gbp) to a 16.5Kbp genomic sequence (same as above) from the same species. Analyses were run on a SUN-UltraSparc III CPU (800 MHz). To show the limits of a scanning based approach we first applied *sim4* to this data set. Since *sim4* cannot handle files larger than 2GB, we split the 2.4 GB file containing the mouse ESTs into two files. The total running time of *sim4* was 3.5 hours, which is unacceptable for a web service application.

e2g delivers a mapping (without any spliced alignment) for the same data size in much less time, using a pre-computed index structure. With the default parameter setting, all 3883 matches of length ≥ 30 , containing exact seeds of length ≥ 20 , and with identity $\geq 98\%$ are computed in 30 seconds. This is about the same time as required for storing the match positions in the database and generating the graphical overview.

3.3.5. Utility

As a brief example of the utility of *e2g*, we mention the work by Drepper *et al.* [41] who successfully used *e2g* to identify unknown genes in the wobbler (*wr*) mouse, an animal model of amyotrophic lateral sclerosis [132]. The *wr* mutation was mapped to mouse chromosome 11 between markers *BAC147N22* and *Murr1*. In this region, two new genes could be identified using *e2g*: *NM.172792* and *Tmem17*. The candidate interval for *wr* was further narrowed to 0.9 Mb between *D11Hjk30* and *D11Hjk29*. Finally, in exon 23 of *Vps54*, another gene in this region, an A-T transversion in the *wr/wr* genomic DNA could be identified that results in the amino acid substitution L967Q and causes motor neuron disease and defective spermiogenesis in the wobbler mouse [132].

3.4. EST Clustering using Vmatch

Vmatch provides an option to cluster the database sequences according to the matches found in a self comparison of the index. Based on the computed matches, single linkage clustering is applied: starting with each EST in a single cluster, clusters are merged if two ESTs from two clusters have significant similarity. The process is continued until no further clusters can be merged. This procedure is also referred to as *transitive closure*, which means that any two sequences above a similarity threshold will end up in the same

cluster. Two sequences A and B are in the same cluster even if they do not overlap at all, but there exists a third sequence C with significant similarity to both A and B.

To identify matching sequences, *Vmatch* first computes all maximal exact matches of a given minimal length (*seedlength*) between all sequences. These seeds are extended in both directions allowing for matches, mismatches, insertions, and deletions using the X-Drop alignment strategy as described above (see Section 3.2.3). To speed up the identification of the initial seeds, the seed length was optimized for the type of match according to the following formula:

$$seedlength = \left\lfloor \frac{l}{l - (l \cdot p/100) + 1} \right\rfloor \quad (3.1)$$

where l is the minimal match length and p the minimal percent identity of the match. In case of $l = 100$ and $p = 98$ the seed can be as long as 33, whereas for $l = 40$ and $p = 94$ the minimal length of the seed is 11, which results in a significantly increased running time, because many more initial exact matches are found and have to be extended by the X-Drop alignment.

3.4.1. Clustering Parameters

A general problem in clustering a data set is the choice of the correct clustering parameters. Often, parameters are a result of trial and error by manually inspecting the clustering results. UniGene, one of the most widely used databases of clustered ESTs describes the clustering procedure in a very vague statement: "Sequence pairs which are sufficiently similar are linked together to form initial clusters." Neither an exact definition of "sufficiently similar" nor the exact procedure used for constructing the UniGene clusters is satisfyingly documented to allow complete reproduction. The same holds true for most other EST indices like TIGR Gene Indices or STACK.

In an attempt to objectively define appropriate clustering criteria, we will take advantage of the speed of the *Vmatch* clustering approach to systematically vary the relevant parameters: minimal match length, percent identity and X-drop value. (Remember that the underlying index has to be constructed once and then many matching tasks can rapidly be performed on the same index.)

It is hypothesized that the 'correct' parameterization is inherent in the data set and will be revealed as an abrupt change in the curve on the resulting graph: the structure of the data becomes apparent when ESTs of the same gene cluster together, which again form

3. Suffix Array Based EST Mapping and Clustering

Name	Parameter	Values
Percent identity	-identity	94, 96, 98
Overlap length	-l	30, 50, 60, 80, 100, 120, 140, 160, 180, 200
X-Drop	-exdrop	1, 2, 4, 8

Table 3.1.: *Vmatch* clustering parameters for *X. laevis* data set.

clusters of gene (i.e. protein) families and then again superfamilies.

3.4.2. *X. laevis* EST Data Set

The first data set used for assessing appropriate clustering parameters is a set of 243,981 ESTs and mRNAs (138,405,765 nt) of the African clawed frog *X. laevis*. Sequences were downloaded from GenBank and pre-processed as described in Sections 2.2.2 and 4 to mask contaminants and repeat sequences. An enhanced suffix array is constructed for all processed sequences and *Vmatch* is used for clustering the sequences using the parameters defined in Table 3.1.

An overall of 120 different clusterings were computed, using the pipeline described in Section 4. As will be shown in Section 3.5.4, such analyses can hardly be done for other clustering algorithms because of the long running times. *Vmatch* instead allows an exhaustive exploration of the parameter space. The results of these analyses showing the effect of varying the minimal match length (-l), percent identity (-identity) and X-Drop value (-exdrop) are presented in Figures 3.5 (94% identity), 3.6 (96% identity), and 3.7 (98% identity), respectively. The top part of each figure shows the number of clusters in blue and the number of singlets in red. Each data point is annotated by the running time (in minutes) for the corresponding parameter setting. The bottom part of each figure again depicts the running times for each setting (logarithmic scale).

Clustering Results

The goal in this gene-oriented clustering of the EST data set was to group all sequences from one gene into one cluster. Ideally, the sequences would have no sequencing errors or other artifacts and a clear separation could be made by pure overlap detection of a certain

3.4. EST Clustering using Vmatch

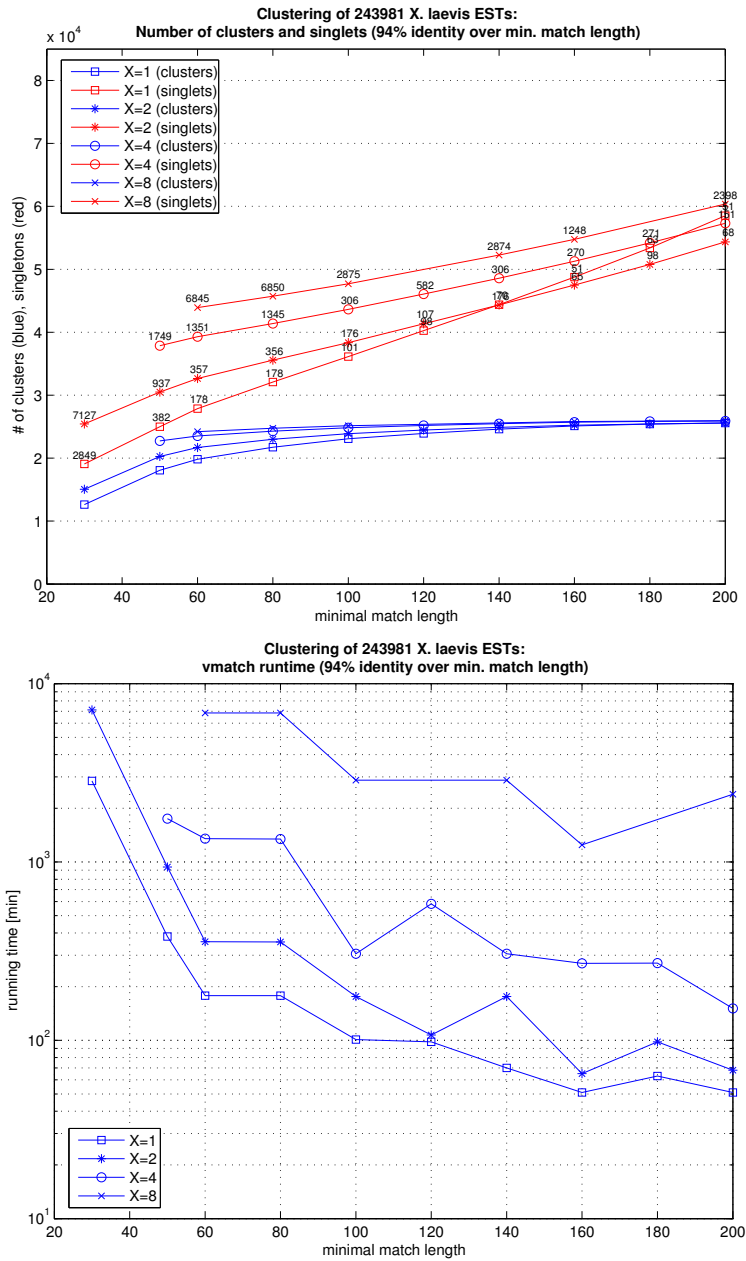


Figure 3.5.: Top: Number of clusters (blue) and singlets (red) for various settings of *Vmatch* parameters *overlap length* (-1) and *X-Drop* (-exdrop) with given *identity* (-identity) of 94%. Bottom: Running times for corresponding parameter settings (time for index construction excluded, logarithmic scale).

3. Suffix Array Based EST Mapping and Clustering

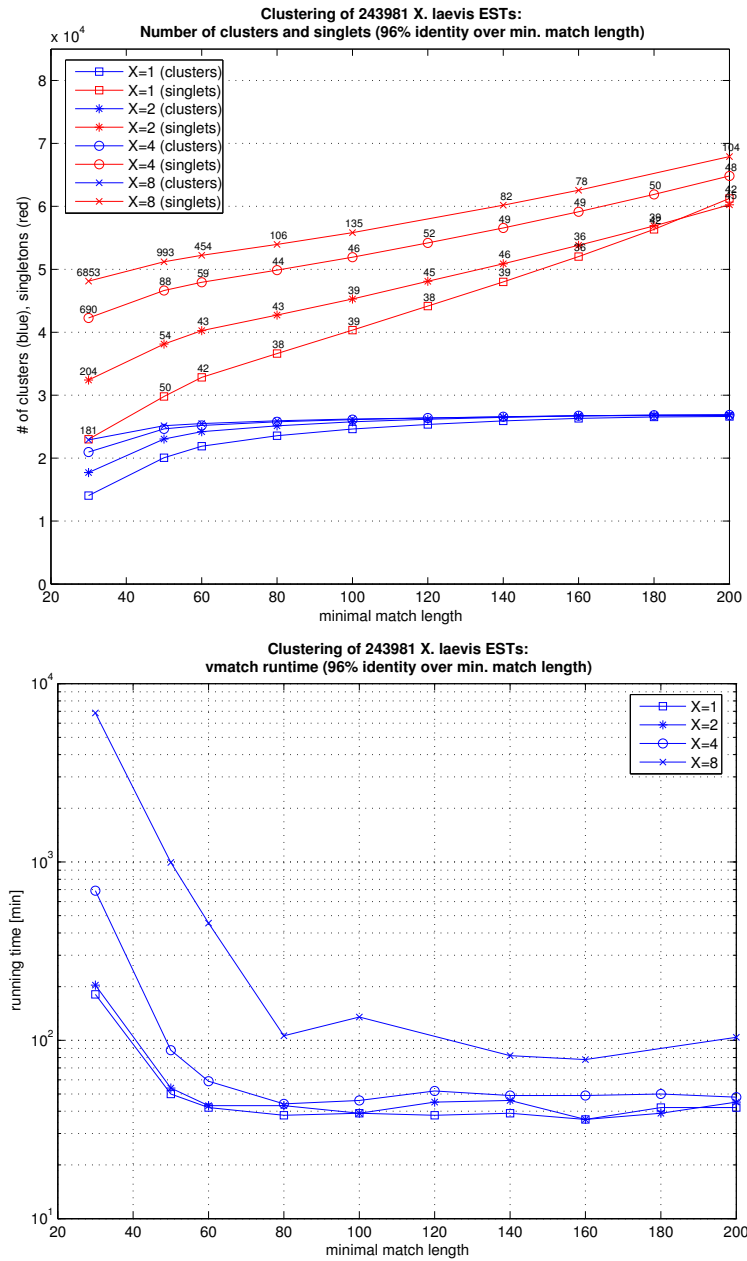


Figure 3.6.: Top: Number of clusters (blue) and singlets (red) for various settings of *Vmatch* parameters *overlap length* (-l) and *X-Drop* (-exdrop) with given *identity* (-identity) of 96%. Bottom: Running times for corresponding parameter settings (time for index construction excluded, logarithmic scale).

3.4. EST Clustering using Vmatch

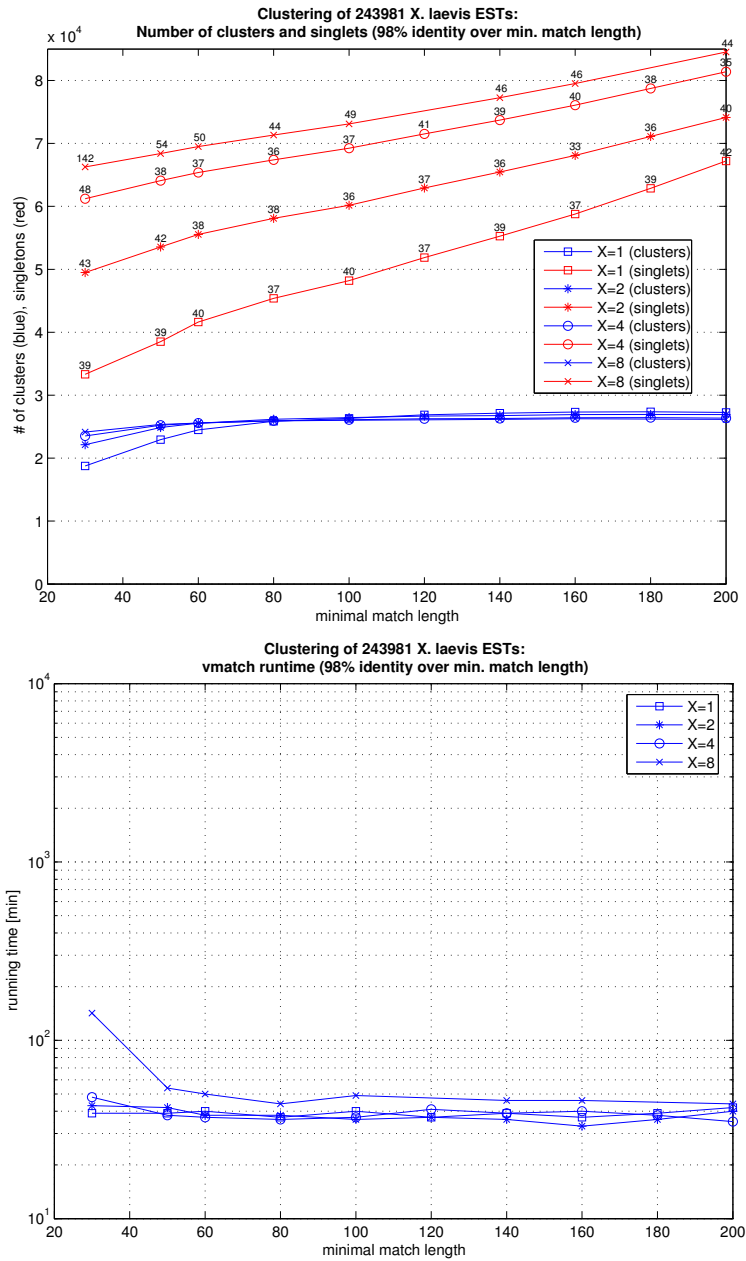


Figure 3.7.: Top: Number of clusters (blue) and singlets (red) for various settings of *Vmatch* parameters *overlap length* (-1) and *X-Drop* (-exdrop) with given *identity* (-identity) of 98%. Bottom: Running times for corresponding parameter settings (time for index construction excluded, logarithmic scale).

length. In this case, the minimal overlap length has to be chosen such that matches are unique within a gene and therefore allow a perfect separation. Unfortunately, repeats, linker sequences and other sequencing artifacts may cause problems here.

A number of conclusions become apparent: at this level of resolution (120 independent clusterings), a distinct point indicating the 'correct' parameters does not become readily apparent. The total number of clusters stays relatively constant (approx. 25,000) starting at a minimal match length of $l = 60$. Decreasing the minimal match length to $l = 30$ ($exdrop = 1$) results in 12,621 clusters ($identity = 94\%$), 14,043 clusters ($identity = 96\%$) and 18,750 clusters ($identity = 98\%$), respectively. For $exdrop = 8$ and $l = 30$, the number of clusters is 22,909 ($identity = 96\%$) and 24,148 ($identity = 98\%$), respectively. As expected, the number of clusters increases with a higher percent identity cutoff.

The number of singlets increases almost linearly with increasing minimal match length from 19,072 ($identity = 94\%$, $exdrop = 1$, $l = 30$) to 58,529 ($identity = 94\%$, $exdrop = 1$, $l = 200$). More stringent percent identity cutoffs produce more singlets: 22,993 ($identity = 96\%$, $exdrop = 1$, $l = 30$) up to 61,259 ($identity = 96\%$, $exdrop = 1$, $l = 200$) and 33,315 ($identity = 98\%$, $exdrop = 1$, $l = 30$) up to 67,214 ($identity = 98\%$, $exdrop = 1$, $l = 200$), respectively.

A remarkable effect can be observed in the number of singlets while modifying the X-Drop value: for all percent identity cutoffs, the number of clusters and singlets increases with higher X-Drop values. Intuitively, a higher X-Drop value should allow for more errors and therefore longer matches between the sequences. This again should result in more sequences clustered together and thus the number of clusters and singlets should decrease with higher X-Drop values. The observed finding could be the effect of the approach *Vmatch* uses to filter the matches found: seeds are extended to both directions, until the alignment score drops more than X from the best score seen so far. The percent identity filter is then applied on these matches, which results in a rejection of matches with good seeds but relatively poor extensions, although a shorter match (though longer than the minimal match length) could have passed the filter.

Running Time

The *Vmatch* runs with 120 different parameter settings were conducted on a Sun Ultra-Sparc III (900 MHz) CPU and 64GB RAM. The total run-time as a function of minimal match length is shown in Figures 3.5 (94% identity), 3.6 (96% identity), and 3.7 (98%

identity), respectively (bottom plots). The run-times do not include the construction of the index, which has to be done just once and took 795 seconds. For 98% identity (Fig. 3.7) the running time stays fairly constant for minimal match lengths between 60 and 200 (approx. 40 seconds). Same holds true for an identity cutoff of 96% and X-Drop values between 1 to 4. With smaller minimal match length ($l \leq 60$) the running time increases exponentially (50 sec for $X = 1$, $l = 50$; and 181 sec for $l = 30$). This effect starts earlier ($l \leq 80$) for $X = 8$ (993 sec for $X = 8$, $l = 50$; and 6852 sec for $l = 30$). For an identity cutoff of 94%, running times grow exponentially starting at 68 sec ($X = 1$) and 2398 sec ($X = 8$) for $l = 200$, increasing to 2849 sec ($X = 1$) and 7127 sec ($X = 2$) for $l = 30$, respectively. Raising the X-Drop value from 1 to 8 results in an increase of the running time of two orders of magnitudes.

The effect of exponentially increasing running time with smaller match length can be explained by the number of seeds found in the first matching phase before the extension of each match using the X-Drop approach: the number of matching seeds grows exponentially with decreasing seed length. The seed length is chosen optimally for the match task using Equation 3.1. In case of a 94% identity match of length 30, the maximal seedlength is 10: in a random sequence one could expect 1 match per MB sequence data. As ESTs are highly redundant, many more initial seeds can be expected that are extended by the costly X-Drop alignment. The reverse effect explains the shorter running time for increasing percent identity: higher identity allows for longer seeds, fewer initial matches and therefore reduced run-time.

Percent identity vs. leastscore

The effect of having more singlets with higher X-Drop values led to the question if a different *Vmatch* parameter might be more suitable for clustering ESTs than the percent identity. In a second experiment, the *leastscore* (option `-leastscore`) was used instead of the `-identity` option. The *leastscore* was chosen in the following way:

$$\textit{leastscore} = \left\lfloor \frac{2l \cdot \textit{identity}}{100} \right\rfloor \quad (3.2)$$

where l is the minimal length and *identity* the minimal percent identity of the match. When using the X-Drop strategy, alignments are scored such that each mismatch has score -1, an indel (i.e. insertion or a deletion) has score -2, and each match has score 2.

For 98% identity and length 100 the *leastscore* is then defined as $2 \cdot 100 \cdot 0.98 = 196$.

3. Suffix Array Based EST Mapping and Clustering

The shortest possible match has a seed of 33 nucleotides (see Eq. 3.1), 65 matches and 2 mismatches. In the worst case, a match having this score would have length 293: a seed of length 33, 130 matches and 130 mismatches, where the matches and mismatches would alternate perfectly. This case however is very unlikely to occur in evolutionary related sequences and not to be expected.

Figures 3.8, 3.9, and 3.10 show the results of different parameter settings with *leastscore*s adjusted for 94%, 96% and 98% identity over the minimal match length l using Equation 3.2. The overall shape of the curves look very similar to those from Figures 3.5, 3.6, and 3.7. Comparable to the results of clustering using percent identity, the number of clusters do not exceed 25,000 clusters. For $exdrop = 1$ and $l = 100$, the number of clusters is 21,629 (*leastscore* = 188), 21,911 (*leastscore* = 192), and 22,318 (*leastscore* = 196). For increasing $exdrop = 4$ and $l = 100$, the number of clusters decreases to 12,386 (*leastscore* = 188), 16,443 (*leastscore* = 192), and 19,727 (*leastscore* = 196), respectively.

The number of singlets is compared to Figures 3.5, 3.6, and 3.7 overall lower, while the increase with higher minimal match lengths scales linear again: the number of singlets increases from 12,751 (*leastscore* = 192, $exdrop = 1$, $l = 30$) to 57,247 ($l = 200$, Figure 3.9). Higher X-Drop values result in less singlets: 5,466 (*leastscore* = 192, $exdrop = 4$, $l = 50$) up to 43,167 ($l = 200$).

An X-Drop value of $exdrop = 8$ is clearly too high, as too many sequences get clustered, especially for low percent identities (94% and 96%, Figures 3.8 and 3.9). For $l = 100$ and $exdrop = 8$, the 243,981 sequences get group into only 290 clusters and 6,139 singlets. Although the *leastscore*s of 188 and 192, respectively, are the same as for lower $exdrop$ values, the X-Drop approach allows for too many errors while extending the seeds.

The bottom plots of Figures 3.8, 3.9, and 3.10 depict the running times for the *Vmatch* clustering when using the `-leastscore` parameter. The running times are almost the same compared to the `-identity` based clustering.

These results suggest to use the `-leastscore` instead of the `-identity` parameter for clustering ESTs. The behavior of producing less clusters and singlets with higher X-Drop values is much more intuitive than the results obtained by using the percent identity filter instead. Unfortunately, again no inherent “correct parameter setting” becomes apparent for the parameter space shown in Figures 3.8, 3.9, and 3.10. The increase in number of singlets and almost constant number of clusters makes a choice difficult. The next section deals with the problem of finding a feasible parameter setting for *Vmatch*-based

3.4. EST Clustering using Vmatch

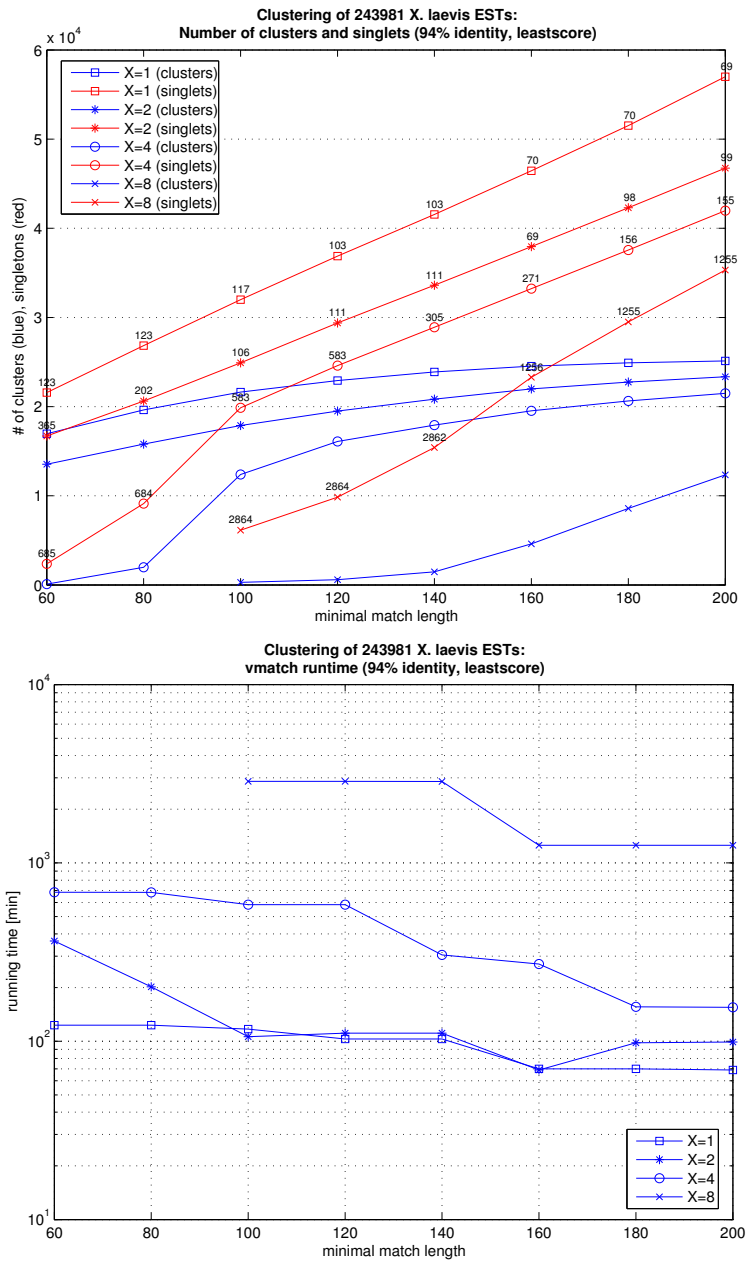


Figure 3.8.: Top: Number of clusters (blue) and singlets (red) for various settings of *Vmatch* parameters *overlap length* (-l) and *X-Drop* (-exdrop) with given *leastscore* (-leastscore) adjusted for 94% identity over minimal match length. Bottom: Running times for corresponding parameter settings (time for index construction excluded, logarithmic scale).

3. Suffix Array Based EST Mapping and Clustering

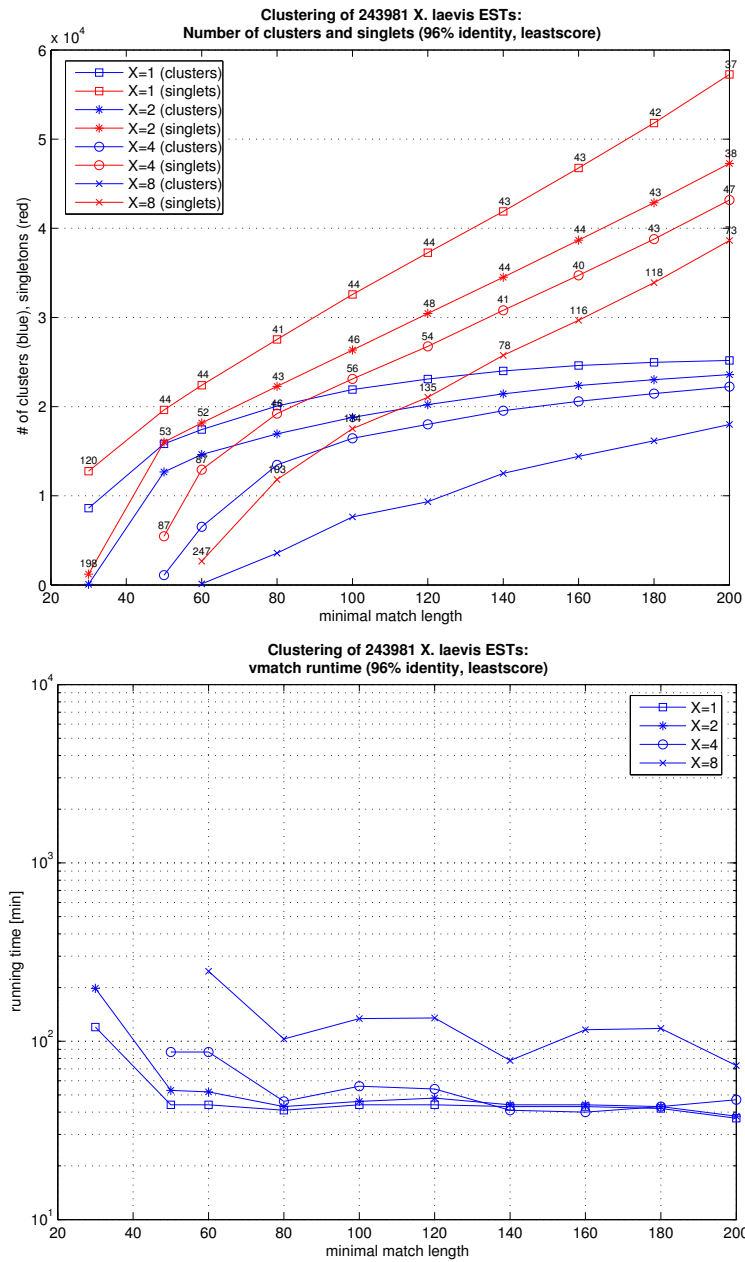


Figure 3.9.: Top: Number of clusters (blue) and singlets (red) for various settings of *Vmatch* parameters *overlap length* (-l) and *X-Drop* (-exdrop) with given *leastscore* (-leastscore) adjusted for 96% identity over minimal match length. Bottom: Running times for corresponding parameter settings (time for index construction excluded, logarithmic scale).

3.4. EST Clustering using Vmatch

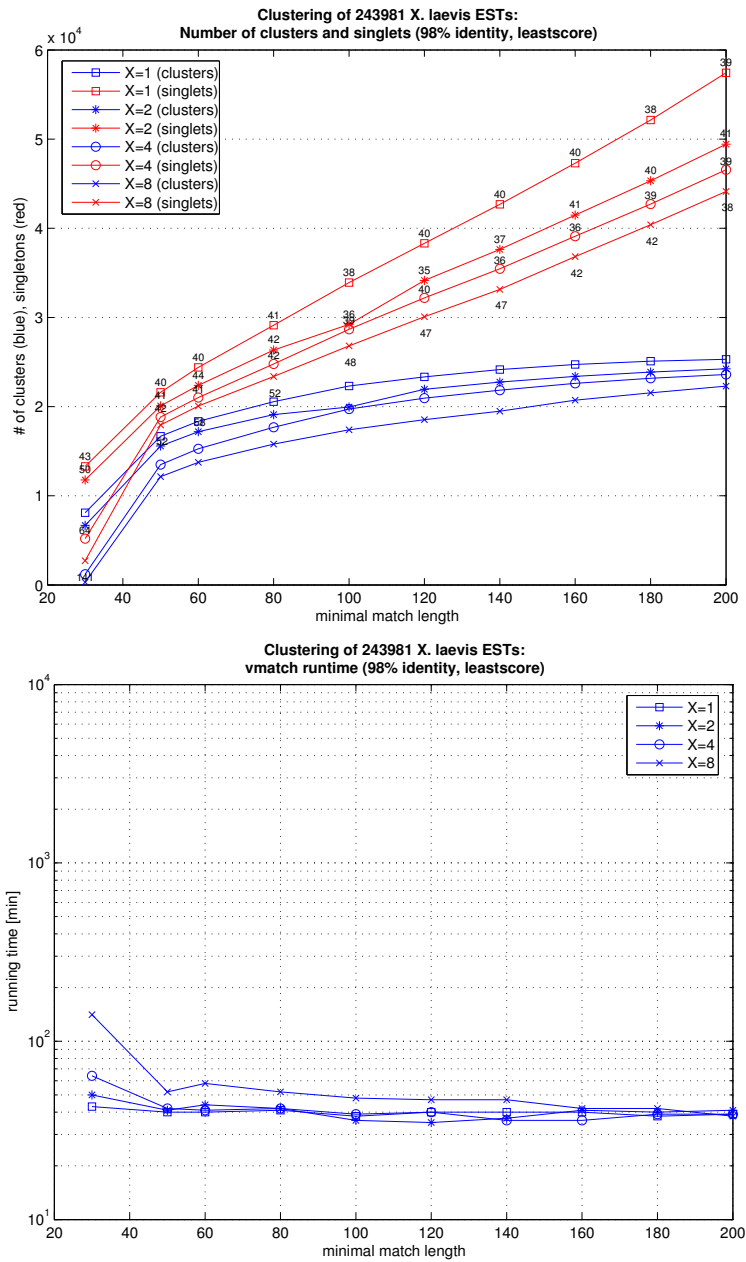


Figure 3.10.: Top: Number of clusters (blue) and singlets (red) for various settings of *Vmatch* parameters *overlap length* (-l) and *X-Drop* (-exdrop) with given *leastscore* (-leastscore) adjusted for 98% identity over minimal match length. Bottom: Running times for corresponding parameter settings (time for index construction excluded, logarithmic scale).

EST clustering and compares the results to those of different other algorithms.

3.5. Validation of Clustering Results

In the previous section we have shown that different parameters produce different partitions of the data set. The degree in which the partitions differ is not apparently known, and postulating the “best” parameter setting not feasible from the clustering alone. In general, it is a non-trivial task to compare different partitions of a given set. Especially if the space of all possible clusterings grows with the size of the data set, the clustering soon becomes too complex to draw reasonable conclusions just by human intuition alone.

Once we obtained sets of clusters by different algorithms or parameters, a measure of similarity allowing to compare and evaluate the performance is needed. An *index for cluster validity* measures the adequacy of a structure through cluster analysis in terms that can be interpreted objectively. The adequacy refers to how well the clustering structure provides true information about the data.

Indices for attacking this question in a quantitative manner can be categorized as follows [72]:

- An *external index* assesses the degree to which two partitions of n objects agree. One partition comes from a clustering solution. The second is assigned a priori, independent of the data and the first partition.
- *Internal indices* measure the degree to which a clustering obtained by a clustering algorithm is justified in light only of the pattern or proximity matrix. They measure the fit between the partition imposed by a clustering algorithm and the data themselves.

An external index is the index of choice, whenever the correct partition of the data is known. In case of EST clustering, this partition can be constructed in fairly good quality by mapping the ESTs against the genomic sequence. In Section 3.5.2 we describe a benchmark data set for EST clustering established by Zhu *et al.* [171]. We will use this data set as an a priori classification of the ESTs. Therefore, we will use an external index to validate the partitions obtained from different parameters and algorithms.

Several external criterion statistics have emerged in the literature:

- Jaccard [71]

- Rand [127]
- Fowlkes and Mallows [46]
- Morey and Agresti adjusted Rand [112]
- Hubert and Arabie adjusted Rand [66]

Milligan *et al.* [109] give a comprehensive overview of these five different external indices. Similar to other studies [46, 110] they find that the mean Rand index value increased as the number of clusters in the hierarchical solution increased. The Fowlkes and Mallows index decreased as the number increased. The Jaccard index has the same performance pattern as the Fowlkes and Mallows measure and thus, does not appear to be a suitable measure for comparing across hierarchy levels. The Hubert and Arabie adjusted Rand index was particularly effective for conditions of random noise data sets and produced mean values quite close to 0 for these cases in contrast to the other measures.

Based on empirical comparisons, Milligan *et al.* conclude that, for partitions having different numbers of clusters, the Hubert and Arabie adjusted Rand index is the index of choice. In case of EST clustering different parameters and algorithms almost always result in different numbers of clusters, therefore we will use this index in the following sections to validate the clusterings produced by *Vmatch* and other methods in contrast to other authors, who mostly use the Jaccard index (e.g. [76, 101]) or count the number of exact matching clusters in different partitions [26]. First, we give a formal definition of the Hubert and Arabie Adjusted Rand Index.

3.5.1. Hubert and Arabie Adjusted Rand Index

Suppose $U = \{u_1, \dots, u_R\}$ and $V = \{v_1, \dots, v_C\}$ are two different partitions of a set $S = \{O_1, \dots, O_n\}$ of n objects, i.e. the entries of U and V are subsets of S such that $\bigcup_{i=1}^R u_i = S = \bigcup_{j=1}^C v_j$ and $u_i \cap u_{i'} = \emptyset = v_j \cap v_{j'}$ for $1 \leq i \neq i' \leq R$ and $1 \leq j \neq j' \leq C$. n_{ij} denotes the number of objects that are common to clusters u_i and v_j . Let $n_{i\cdot}$ and $n_{\cdot j}$ denote the number of objects in clusters u_i and v_j , respectively. The information of cluster overlap between the two partitions U and V can be written in form of a contingency table as in Table 3.2.

Measures of correspondence between U and V are frequently based on how object pairs are classified in the $R \times C$ contingency matrix. Among all $\binom{n}{2}$ distinct pairs there are four different types:

3. Suffix Array Based EST Mapping and Clustering

		Partition V				
Class		v_1	v_2	\cdots	v_C	Sums
Partition U	u_1	n_{11}	n_{12}	\cdots	n_{1C}	$n_{1\cdot}$
	u_2	n_{21}	n_{22}	\cdots	n_{2C}	$n_{2\cdot}$
	\cdot	\cdot	\cdot		\cdot	\cdot
	\cdot	\cdot	\cdot		\cdot	\cdot
	\cdot	\cdot	\cdot		\cdot	\cdot
	u_R	n_{R1}	n_{R2}	\cdots	n_{RC}	$n_{R\cdot}$
Sums		$n_{\cdot 1}$	$n_{\cdot 2}$	\cdots	$n_{\cdot C}$	$n_{\cdot\cdot} = n$

Table 3.2.: Notation for the contingency table representing the cluster overlap of partitions U and V .

- (a) pairs of objects placed in the same cluster in U and in the same cluster in V
- (b) pairs of objects placed in different clusters in U and in the same cluster in V
- (c) pairs of objects placed in the same cluster in U and in different clusters in V
- (d) pairs of objects placed in different clusters in U and in different clusters in V

Types (a) and (d) can be interpreted as agreements, (b) and (c) as disagreements. Based on the contingency table, explicit formulae can be given for calculating the number of object pairs for each type (see Table 3.3).

Commonly used measures are the Jaccard index $\frac{a}{a+b+c}$ [71] and the Rand index $\frac{a+d}{\binom{n}{2}}$ [127]. The Jaccard and the Rand indices both lie between 0 and 1. When two partitions are identical, the index is 1. A problem is that the expected value for these indices for two random partitions do not take a constant value, e.g. zero. (In fact, it depends on the number of clusters.) Consequently, the relative sizes for each of these raw measures are difficult to compare. The Hubert and Arabie adjusted Rand index [66] corrects for this by using the general form of an index corrected for chance:

$$\frac{\text{Index} - \text{Expected Index}}{\text{Maximum Index} - \text{Expected Index}}, \quad (3.3)$$

which is bounded above by 1 and takes on value 0 when the index equals its expected value.

Type	Formula
(a)	$\frac{1}{2} \sum_{i=1}^R \sum_{j=1}^C n_{ij}(n_{ij} - 1)$
(b)	$\frac{1}{2} \left(\sum_{j=1}^C n_{.j}^2 - \sum_{i=1}^R \sum_{j=1}^C n_{ij}^2 \right)$
(c)	$\frac{1}{2} \left(\sum_{i=1}^R n_{i.}^2 - \sum_{i=1}^R \sum_{j=1}^C n_{ij}^2 \right)$
(d)	$\frac{1}{2} \left(n^2 + \sum_{i=1}^R \sum_{j=1}^C n_{ij}^2 - \left(\sum_{i=1}^R n_{i.}^2 + \sum_{j=1}^C n_{.j}^2 \right) \right)$

Table 3.3.: Formulae for calculating the number of object pairs for the four different types of pairs.

Hubert and Arabie Adjusted Rand Index

The Hubert and Arabie adjusted Rand index has the form:

$$\frac{\sum_{i,j} \binom{n_{ij}}{2} - \sum_i \binom{n_{i.}}{2} \sum_j \binom{n_{.j}}{2} / \binom{n}{2}}{\frac{1}{2} \left[\sum_i \binom{n_{i.}}{2} + \sum_j \binom{n_{.j}}{2} \right] - \sum_i \binom{n_{i.}}{2} \sum_j \binom{n_{.j}}{2} / \binom{n}{2}} \quad (3.4)$$

Referring to the results of Milligan *et al.* [109], we adopt the Hubert and Arabie adjusted Rand index as our measure of agreement between the external classes and clustering results.

3.5.2. EST Clustering Benchmark Data Set

Different algorithms and parameters used for clustering explain the differences between EST clusters provided by different gene indices. For species with genome sequence available, mapping of ESTs to the genomic sequence can be used as “gold standard” to compare the algorithms and calibrate parameters.

3. Suffix Array Based EST Mapping and Clustering

To assess the accuracy of the clustering results of *Vmatch* and other methods, the benchmark data set of Zhu *et al.* [171]³ will be used as an external measure of cluster validity. The data set consists of 168,200 *A. thaliana* ESTs and was created by spliced alignment of ESTs to their cognate locations in the Arabidopsis genome, with subsequent clustering based on genome location.

A. *thaliana* EST Mapping

Zhu *et al.* confidently mapped 169,888 ESTs onto the Arabidopsis genome by spliced alignments using GeneSeqer [158, 157], a spliced alignment program incorporating sequence similarity and splice site scoring. The mapping provides verified sets of EST clusters for evaluation of EST clustering programs. Results were divided into three different groups of partitions, including clusters based solely on sequence alignment, and clusters based on sequence alignment and clone pair information.

- I. Clusters (including singlets) based solely on sequence alignment (all putative cognate EST locations; 172,137 pcSPAs⁴ from 169,888 hqESTs⁵)
- Ia. Clusters (including singlets) based solely on sequence alignment (uniquely mapped ESTs only; 146,527 pcSPAs from 146,527 hqESTs)
- Iib. Clusters (including singlets) based on sequence alignment and clone pair information (uniquely mapped ESTs only; 146,527 pcSPAs from 146,527 hqESTs)
- IIIa. Clusters (including singlets) based solely on sequence alignment (unique putative cognate EST locations only; 168,200 pcSPAs from 168,200 hqESTs)
- IIIb. Clusters (including singlets) based on sequence alignment and clone pair information (unique putative cognate EST locations only; 168,200 pcSPAs from 168,200 hqESTs)

Clustering was based on genomic locations: Let *est1* map to region $[a, b]$ and *est2* to region $[c, d]$, where $a \leq c$, on the same chromosome; then *est1* and *est2* are clustered if $c \leq b + G + 1$, where G is the clustering parameter. G could be negative (overlap required) or positive (specifying the maximal allowed gap). For ESTs giving multiple exon spliced

³available at: <http://www.plantgdb.org/AtGDB/prj/ZSB03PP/ESTclustering.php>

⁴pcSPAs: putative cognate spliced alignments

⁵hqESTs: high quality ESTs

alignments, the overlap rule is superseded by the requirement for consistency of strand orientation as indicated by GeneSeqer. Thus, ESTs from overlapping genes in opposite transcriptional directions can be separated into different clusters. Additionally, ESTs from the same plasmid (clone pairs) were used to join clusters independent of their local map coordinates.

3.5.3. Quality Evaluation

As we have seen in Section 3.4, options and parameter settings can hardly be optimized in an appropriate manner by manually exploring the parameter space. In the following we describe the evaluation of quality and performance of EST clustering based on *Vmatch*. Different *Vmatch* options and parameters are used to find the “best” setting, which is then compared to other widely used EST clustering tools.

Benchmark Data Set

The accuracy of the results are assessed using the *A. thaliana* data set described in [158, 157] (see Section 3.5.2). This data set was chosen following Kalyanaraman *et al.* [76], who used the same data set to evaluate their EST clustering tool *PaCE*. The data set consists of 168,200 ESTs (group III, see Section 3.5.2).

The benchmark data set represents clusters based on spliced alignments of the ESTs to their cognate locations in the *A. thaliana* genome. ESTs were clustered if their genomic locations overlapped by at least 40bp. 146,527 ESTs mapped to unique locations, 21,673 ESTs mapped to multiple locations on the genome. ESTs aligning to multiple locations were mapped to the cluster corresponding to the location with maximum alignment score (see [76] for details). The genome based clustering results in 18,727 clusters and 10,803 singlets.

To evaluate quality and run-time performance as functions of data size, different subsets of the 168,200 ESTs were extracted as follows: from the set of clusters represented in the benchmark set, smaller sets of 10,000, 20,000, 40,000 and 80,000 ESTs were derived. To achieve this, whole clusters instead of single sequences were chosen randomly from the benchmark set until the desired number of total sequences was reached. This prevents breaking clusters apart by possibly removing sequences that join subclusters, such as ESTs from 5' and 3' ends of the same mRNA. For each of the four data sizes, 4 cluster sets were

3. Suffix Array Based EST Mapping and Clustering

Set	Clusters	Singlets	Set	Clusters	Singlets
10ka	937	544	40ka	4877	2872
10kb	1194	661	40kb	4376	2670
10kc	1199	709	40kc	4654	2751
10kd	1011	588	40kd	4425	2557
20ka	2245	1232	80ka	8756	5137
20kb	2198	1193	80kb	8712	5067
20kc	2502	1386	80kc	8903	5170
20kd	1961	1089	80kd	9017	5099

Table 3.4.: Number of clusters and singlets for the 16 benchmark data sets derived from 168,200 ESTs mapped to their positions on the *A. thaliana* genome.

Name	Parameter	Values
Percent identity	-identity	96, 98
Minimal overlap	-l	40, 60, 80, 100, 120, 140, 160, 180, 200
X-Drop	-exdrop	1, 2, 4, 8

Table 3.5.: *Vmatch* clustering parameters for *A. thaliana* data set.

derived, allowing for averaging of the results. Table 3.4 gives an overview of the data sets and their numbers of clusters and singlets.

These data sets are used as a priori partitions in external index assessments to evaluate the accuracy of the clustering results. The mapping to genomic positions should prevent sequences from overclustering by spurious hits between the ESTs.

Experiments

The 16 different data sets were first clustered using *Vmatch* similar to the setting in Section 3.4. Using identity cutoffs (option `-identity`) of 98% and 96%, the minimal match length was varied from 40 to 200 in steps of 20. For each combination of percent identity and minimal match length, X-Drop values of 1, 2, 4 and 8 were used. Table 3.5 gives an overview of the parameter space. The optimal seedlength for each combination of minimal match length and X-Drop value was calculated according to Equation 3.1, optimizing the run-time of the matching step.

Following the approach in Section 3.4, the percent identity was also simulated by the `-leastscore` parameter instead of the `-identity` cutoff to avoid the effect seen in the *X. laevis* data set, which generated more singlets with higher X-Drop values. As described before, such an effect is not expected, because allowing for more errors in a match should result in less singlets. An objective measure of the quality of both approaches can now be given by the following cluster validation.

Overall, 2,304 different parameter settings have been explored for *Vmatch*, each setting providing a different partition of the data set. The resulting partitions are compared to the “gold standard” benchmark data set by calculating the adjusted Rand index as defined in Section 3.5.1, Equation 3.4.

The results of *Vmatch* based EST clustering were compared to other widely used clustering programs, namely *CAP3*, *PaCE*, *d2.cluster*, *TGICL* and *BLASTclust*. For this purpose, the 16 different data sets were clustered with each tool using the default parameters. For each resulting partition, the adjusted Rand index was determined by comparing the result to the benchmark data set.

Results for `-identity` option

Figures 3.11, 3.12, 3.13, and 3.14 show the adjusted Rand indices for data sets *10k*, *20k*, *40k*, and *80k*, respectively. For each data set, eight different combinations of the options `-identity` (98% or 96%) and X-Drop ($X = 1, 2, 4$ or 8) are plotted. The minimal match length was varied from 40 to 200 in steps of 20. Table 3.6 shows the maximal adjusted Rand indices achieved for all parameter settings within each data set.

It is obvious that for the `-identity` option of *Vmatch*, in the vast majority of cases the best results are produced when using an identity cutoff of 96%. Only for very short minimal match lengths of $l = 40$ or in some cases of data set *80k*, an cutoff of 98% results in better Rand indices. Especially, all maximum adjusted Rand indices (see Table 3.6) were achieved for an identity cutoff of 96%.

For about 88% of these best cases, the best X-Drop value turned out to be $X = 1$, in the remaining 12% $X = 2$. Minimal match lengths of $l = 40$ or $l = 60$ produced best results in more than 62%. With increasing size of the data sets, a tendency of better Rand indices for slightly larger l (80 to 120 instead of 40 to 60) can be observed.

Overall, the results for data sets *10k*, *20k*, and *40k* are very similar. The adjusted Rand index decreases within each data set for minimal match lengths of $l \geq 60$. The minimal

3. Suffix Array Based EST Mapping and Clustering

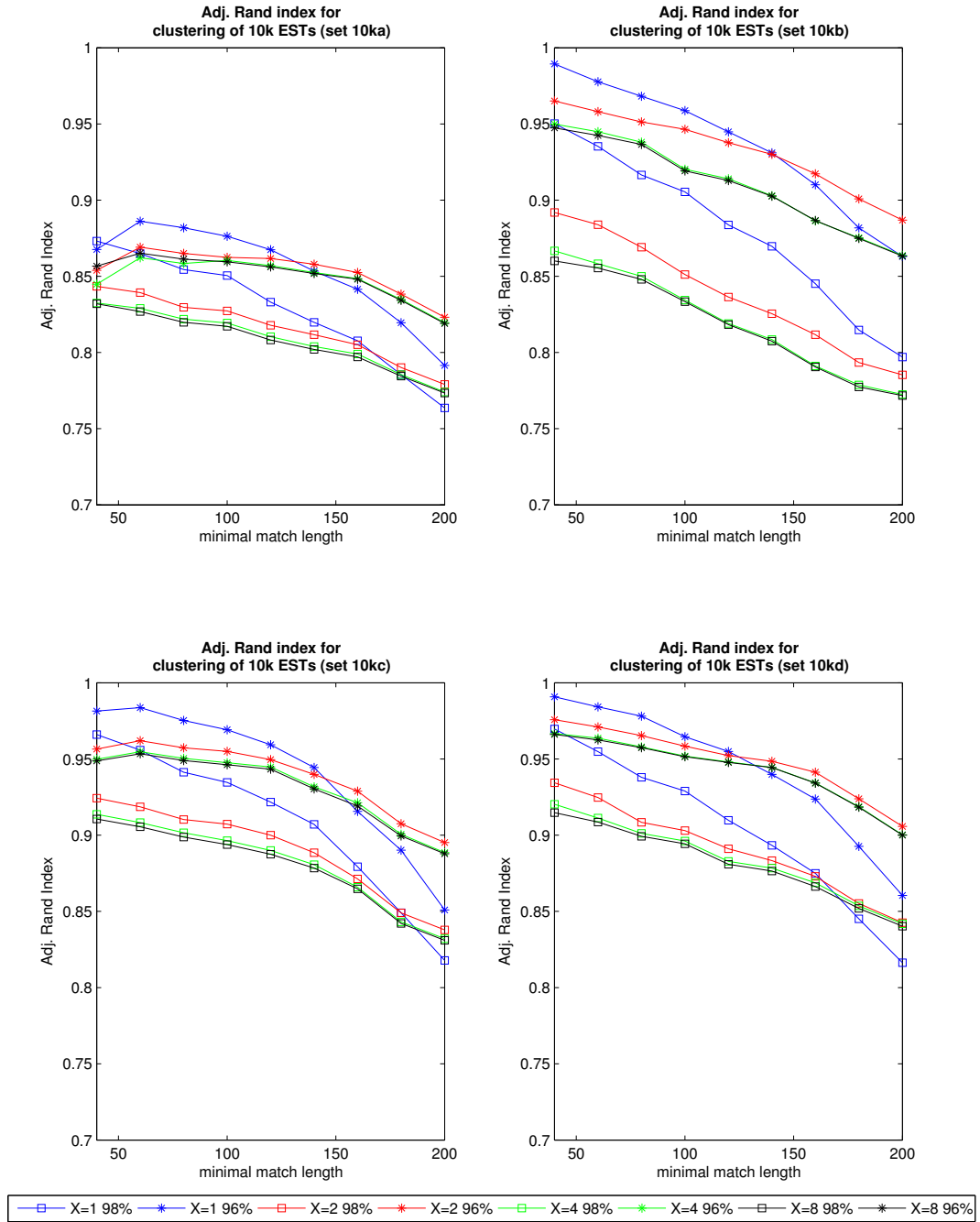


Figure 3.11.: Adjusted Rand Index for *Vmatch* clustering results of data set 10k (option -identity)

3.5. Validation of Clustering Results

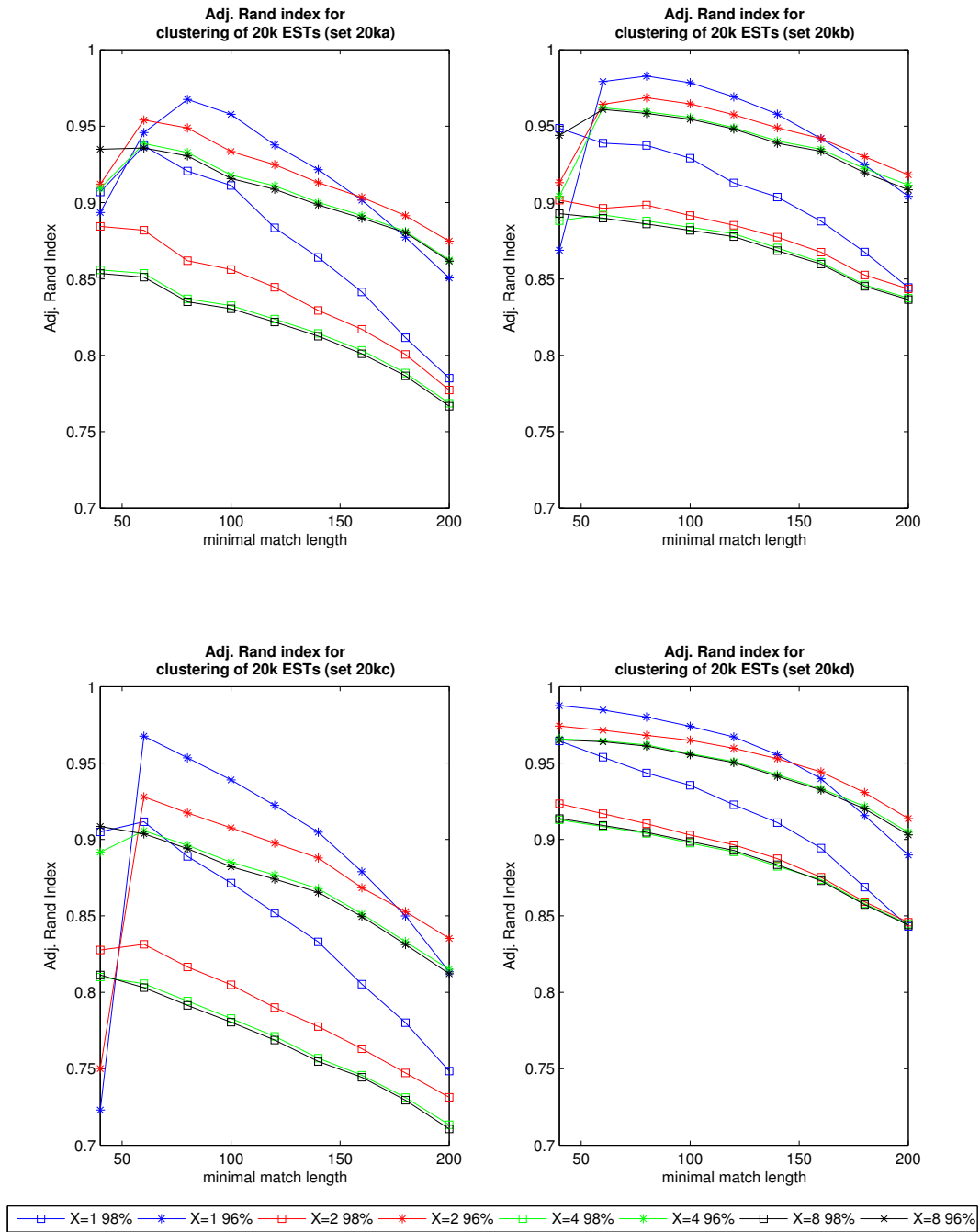


Figure 3.12.: Adjusted Rand Index for *Vmatch* clustering results of data set 20k (option -identity)

3. Suffix Array Based EST Mapping and Clustering

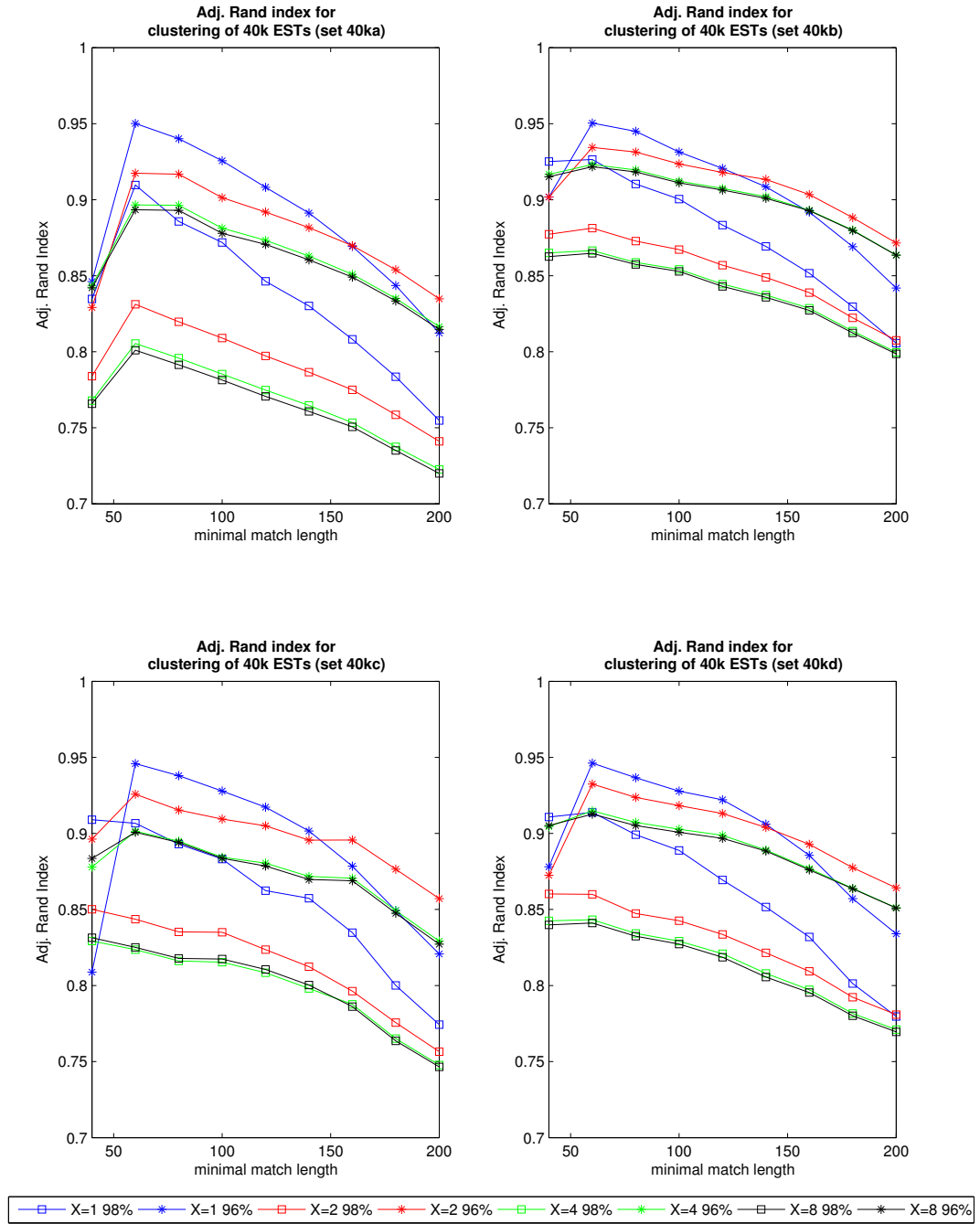


Figure 3.13.: Adjusted Rand Index for *Vmatch* clustering results of data set 40k (option -identity)

3.5. Validation of Clustering Results

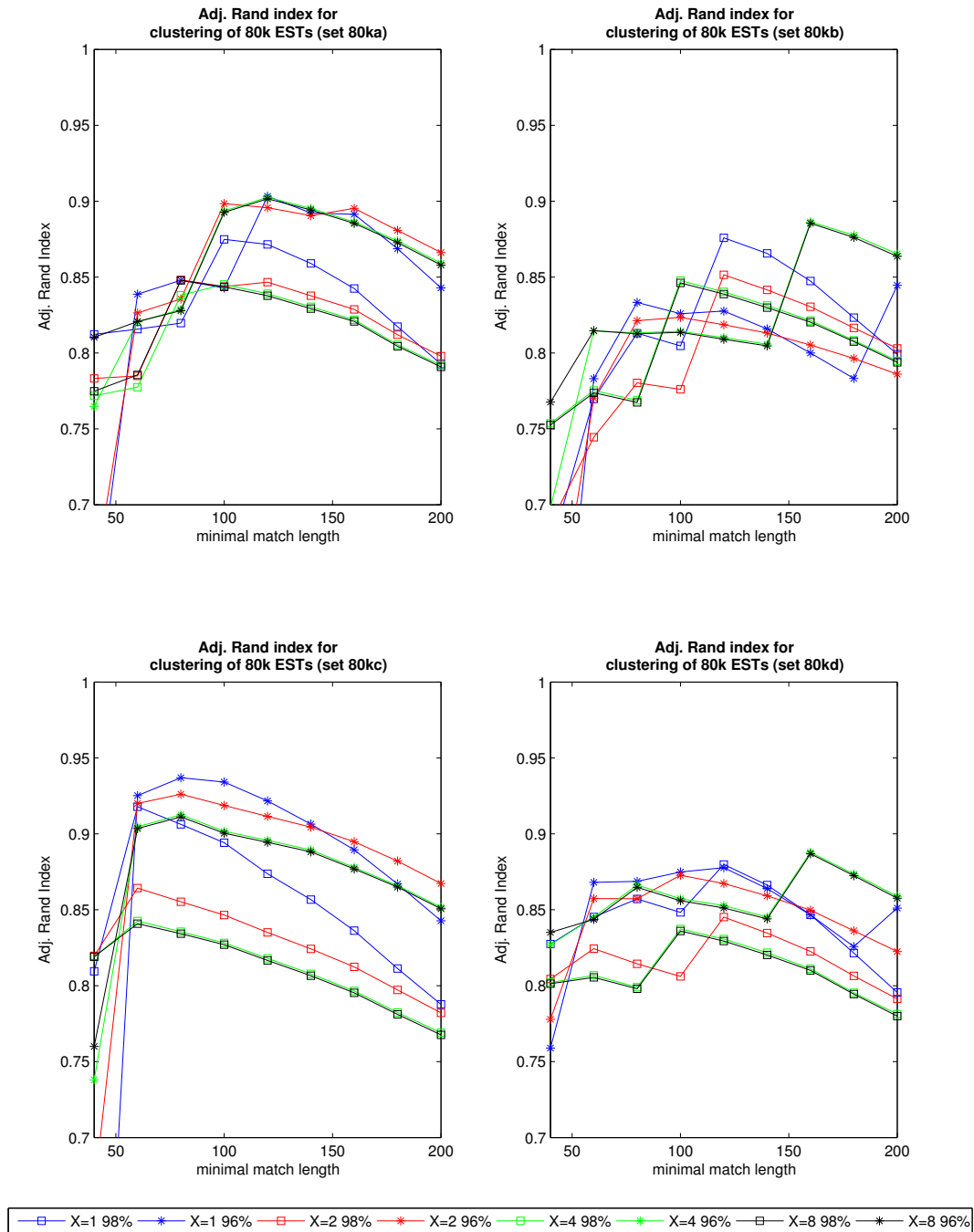


Figure 3.14.: Adjusted Rand Index for *Vmatch* clustering results of data set 80k (option -identity)

3. Suffix Array Based EST Mapping and Clustering

Set	-identity				-leastscore			
	a	b	c	d	a	b	c	d
10k	0.8861	0.9894	0.9837	0.9908	0.8933	0.9923	0.9825	0.9959
20k	0.9675	0.9828	0.9676	0.9875	0.9643	0.9900	0.9780	0.9918
40k	0.9501	0.9504	0.9459	0.9462	0.9537	0.9525	0.9378	0.9236
80k	0.9032	0.8864	0.9370	0.8878	0.8485	0.8225	0.9189	0.8624

Table 3.6.: Maximum values of the Adjusted Rand indices for *A. thaliana* benchmark data sets clustered with *Vmatch* option `-identity` (left part) and option `-leastscore` (right part).

Set	a	b	c	d	Mean
10k	0.83 (± 0.02)	0.88 (± 0.03)	0.91 (± 0.03)	0.92 (± 0.03)	0.89 (± 0.03)
20k	0.88 (± 0.03)	0.91 (± 0.02)	0.83 (± 0.04)	0.92 (± 0.03)	0.89 (± 0.03)
40k	0.83 (± 0.03)	0.88 (± 0.03)	0.85 (± 0.03)	0.86 (± 0.03)	0.86 (± 0.03)
80k	0.84 (± 0.04)	0.80 (± 0.05)	0.85 (± 0.06)	0.84 (± 0.02)	0.83 (± 0.04)

Table 3.7.: Mean Adjusted Rand Index for *A. thaliana* benchmark data sets, option `-identity` used for matching. Other Parameters as specified in Table 3.5.

Rand index measured was 0.7106 for data set *20k* with parameters $l = 200$, $X = 8$ and an identity cutoff of 98%. The results for the *80k* data set are different. The Rand index increases to a maximum at $l \approx 120$. For $l < 120$ or $l > 150$, the Rand index decreases again.

Table 3.7 depicts the mean adjusted Rand index for all parameter settings within each data set. The mean values vary from 0.80 (*80kb*) to 0.92 (*10kd*, *20kd*). The standard deviation within each data set size stays fairly constant at ≈ 0.03 . Overall, the mean of the means decreases from 0.89 to 0.83 with increasing data size. The standard deviation increases also from 0.03 to 0.04.

Results for `-leastscore` option

In the next experiment, the *Vmatch* option `-identity` was replaced by the `-leastscore` option. The leastscore was calculated as described in Equation 3.2 (see page 51) to simulate percent identities of 96% or 98%, respectively, over the minimal match length. For each data set, eight different combinations of the options `-leastscore` and `X-Drop` ($X = 1, 2, 4$ or 8) were used. Figures 3.15, 3.16, 3.17, and 3.18 show the adjusted Rand indices for data sets *10k*, *20k*, *40k*, and *80k*, respectively. The minimal match length was

3.5. Validation of Clustering Results

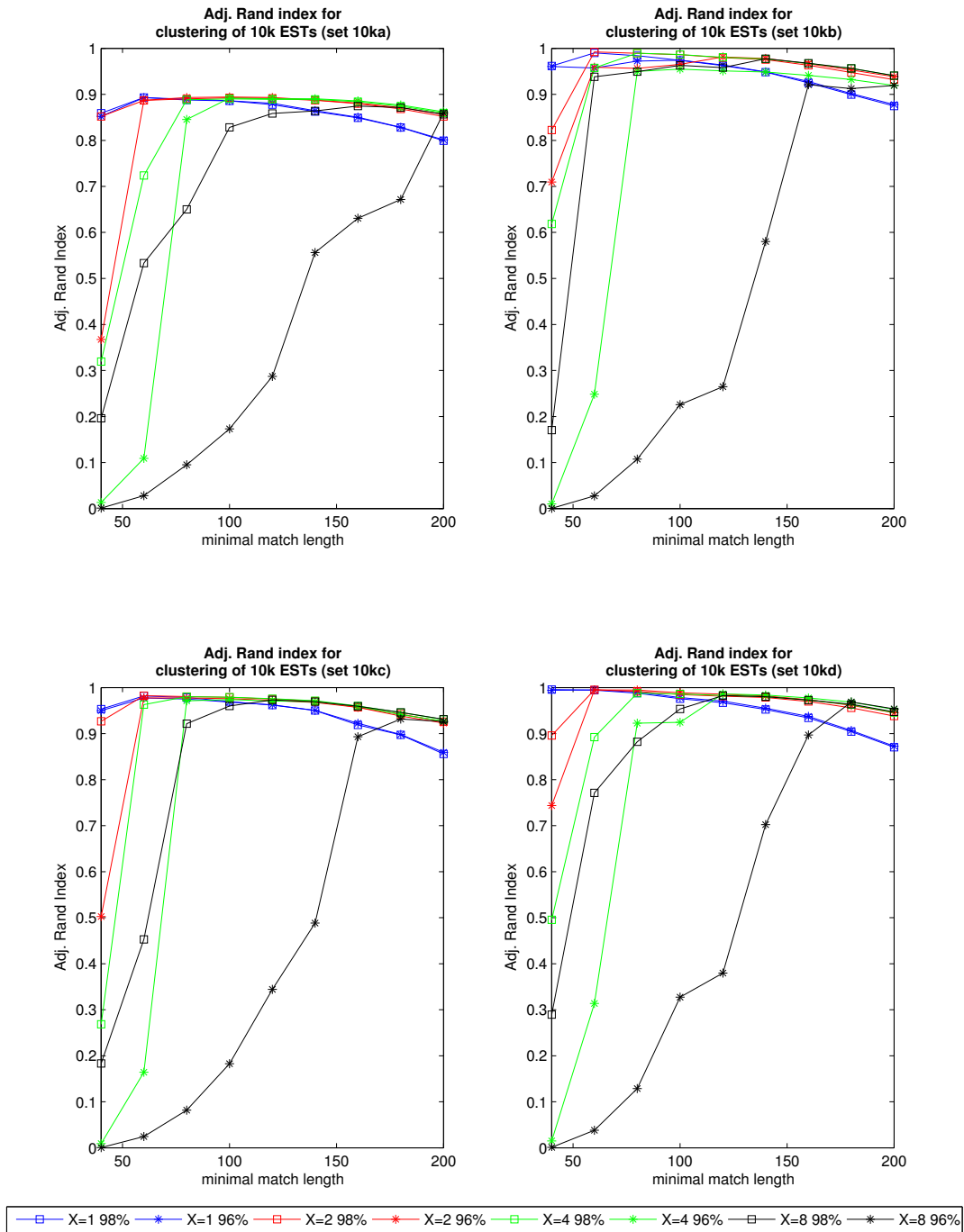


Figure 3.15.: Adjusted Rand Index for *Vmatch* clustering results of data set 10k (option `-leastscore`)

3. Suffix Array Based EST Mapping and Clustering

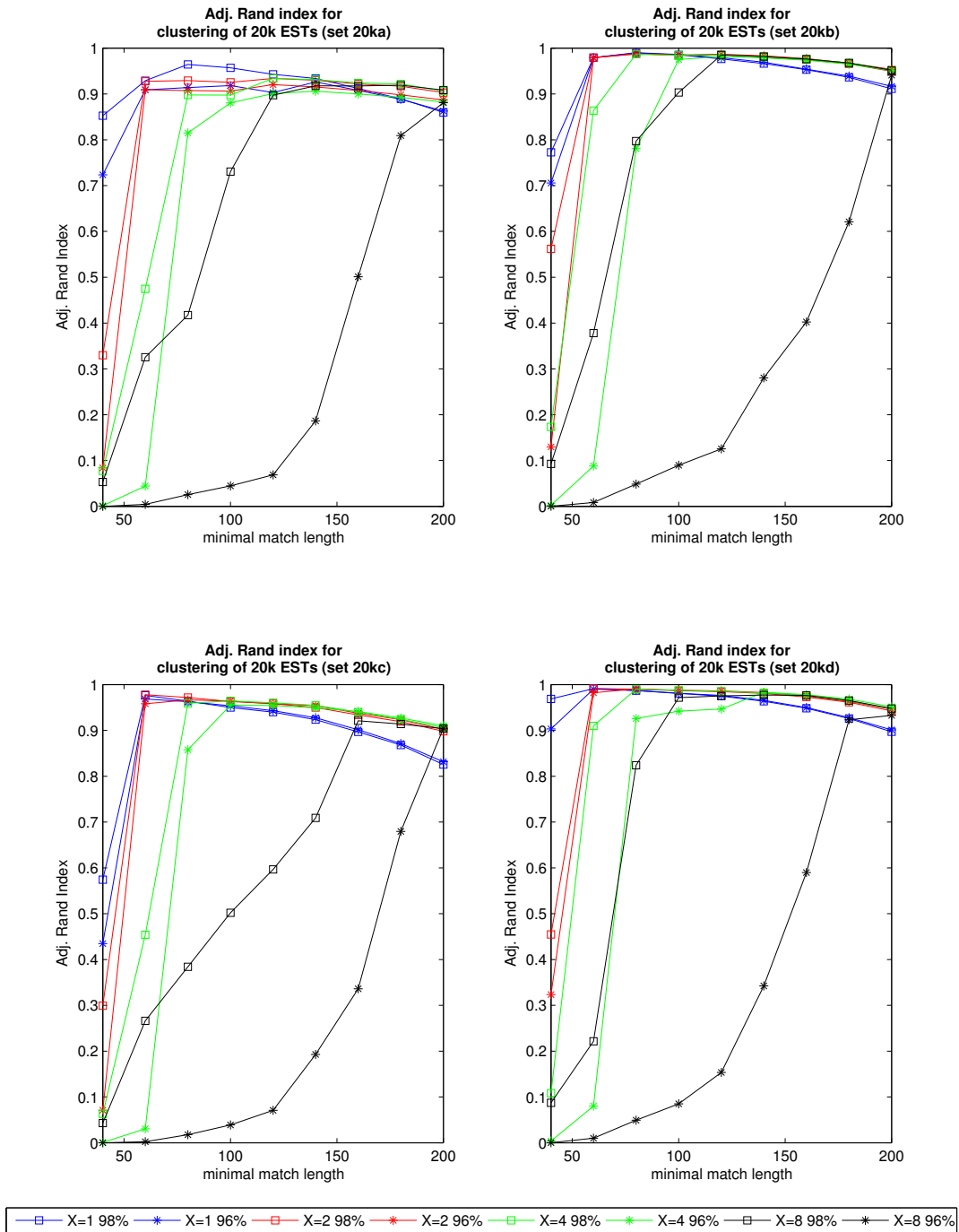


Figure 3.16.: Adjusted Rand Index for *Vmatch* clustering results of data set 20k (option `-leastscore`)

3.5. Validation of Clustering Results

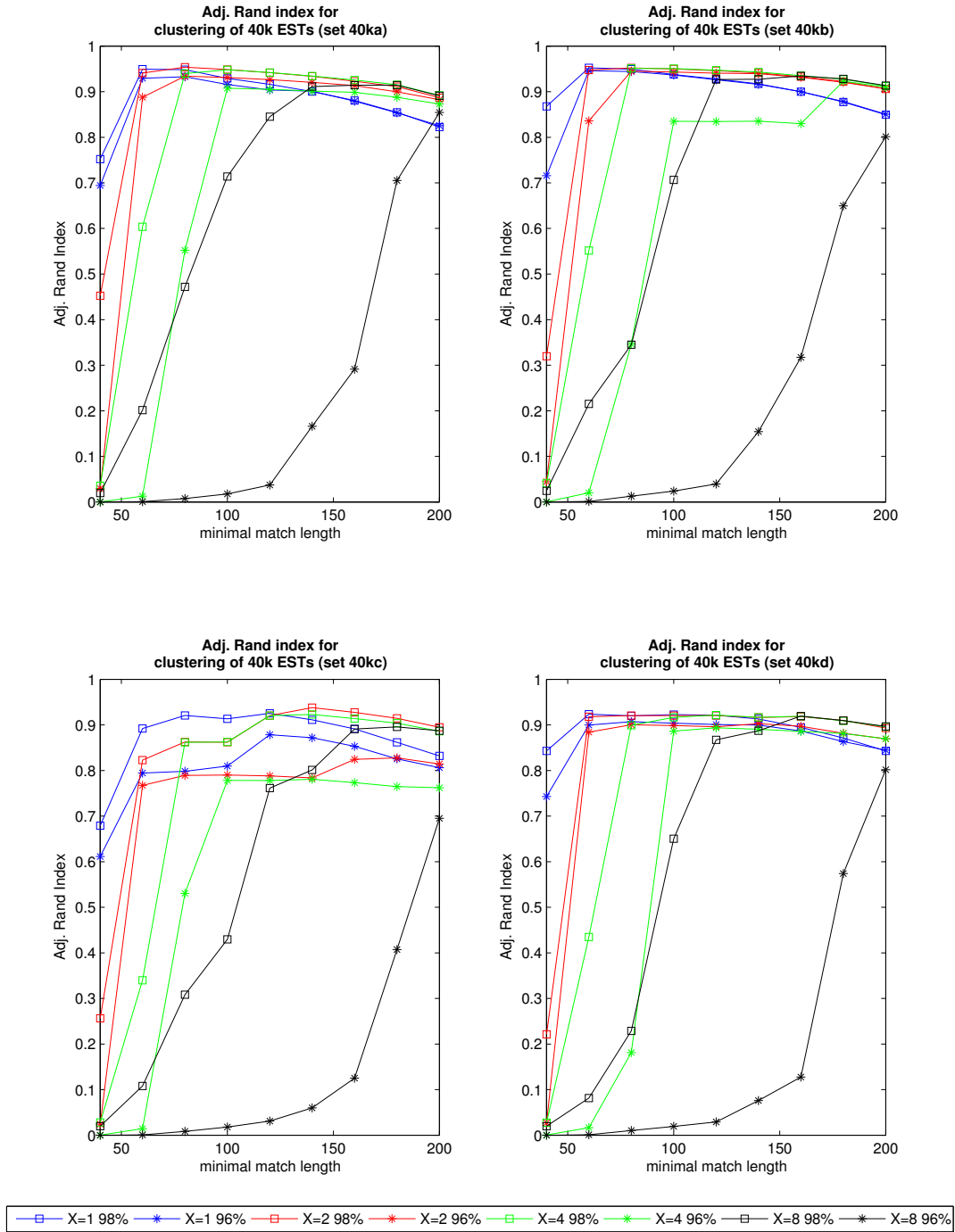


Figure 3.17.: Adjusted Rand Index for *Vmatch* clustering results of data set 40k (option `-leastscore`)

3. Suffix Array Based EST Mapping and Clustering

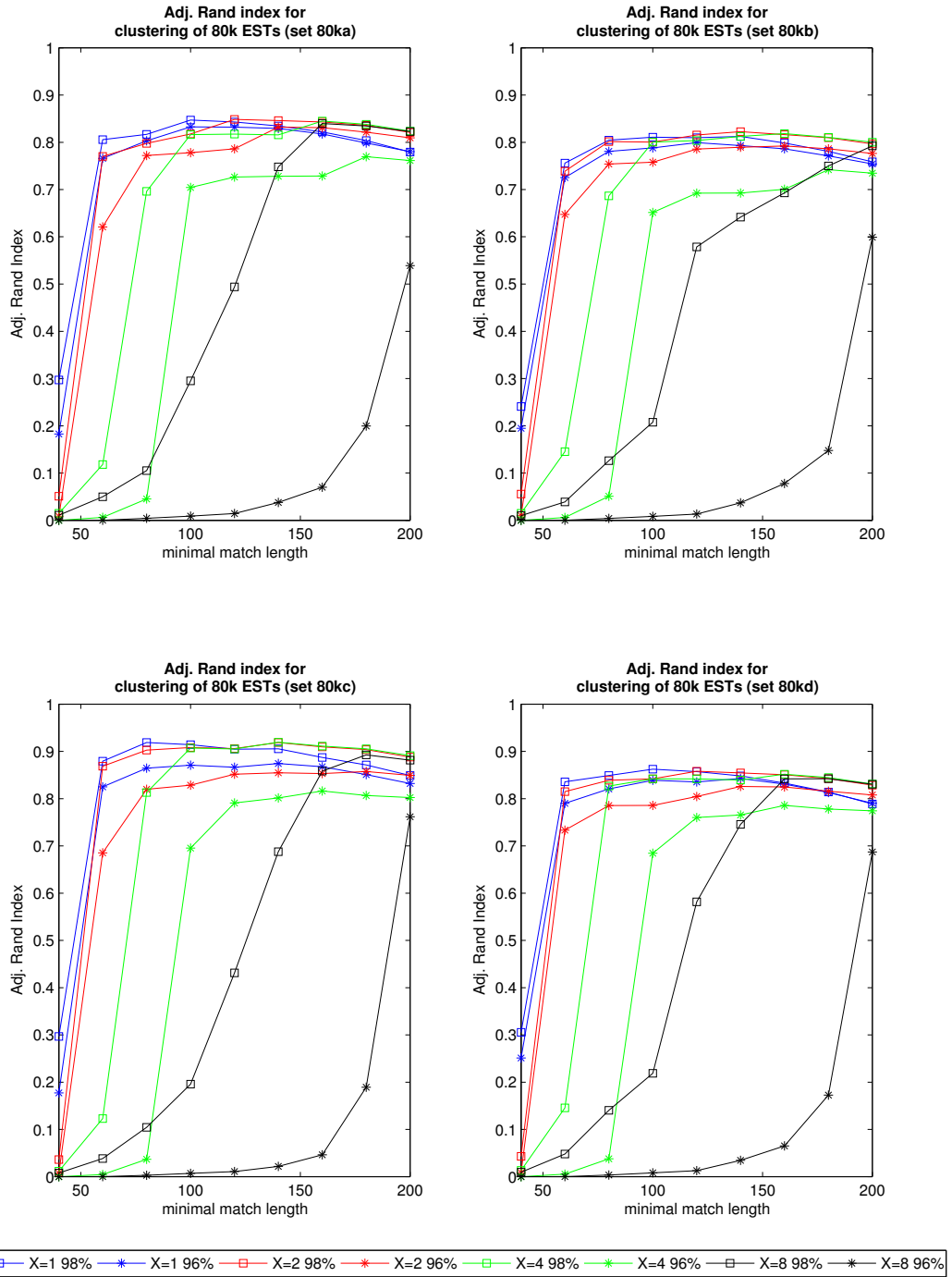


Figure 3.18.: Adjusted Rand Index for *Vmatch* clustering results of data set 80k (option `-leastscore`)

Set	a	b	c	d	Mean
10k	0.75 (± 0.16)	0.85 (± 0.16)	0.83 (± 0.18)	0.86 (± 0.16)	0.82 (± 0.17)
20k	0.74 (± 0.23)	0.80 (± 0.22)	0.73 (± 0.26)	0.81 (± 0.22)	0.77 (± 0.23)
40k	0.72 (± 0.23)	0.72 (± 0.23)	0.66 (± 0.22)	0.70 (± 0.24)	0.70 (± 0.23)
80k	0.58 (± 0.25)	0.56 (± 0.24)	0.62 (± 0.28)	0.60 (± 0.26)	0.59 (± 0.26)

Table 3.8.: Mean Adjusted Rand Index for *A. thaliana* benchmark data sets, option `-leastscore` used for matching. Other Parameters as specified in Table 3.5.

varied again from 40 to 200 in steps of 20. Table 3.6 shows the maximal adjusted Rand indices achieved for all parameter settings within each data set.

For the `-leastscore` option of *Vmatch*, a `leastscore` simulating an identity of 98% produces better adjusted Rand indices in all cases compared to the corresponding 96% parameter setting. The maximum adjusted Rand indices are shown in Table 3.6.

For 50% of these best cases, the best X-Drop is $X = 1$, for the other half $X = 2$. Minimal match lengths of $l = 60$ or $l = 80$ produced best results in more than 68% of these cases. Again, with increasing size of the data sets, a tendency of better Rand indices for slightly larger l (100 to 150 instead of 60 to 80) can be observed.

In contrast to the `-identity` option, the overall shape of the adjusted Rand index plot is highly similar in all data sets, including the largest set *80k*. X-Drop values of $X = 4$ and $X = 8$ turn out to be undesirable (black and green curves in Figures 3.15 to 3.18). Also, a minimal match length of $l = 40$ combined with X-Drop values of $X = 1$ or $X = 2$ is not suitable. Table 3.8 depicts the mean adjusted Rand indices for all parameter settings within each data set, ranging from 0.56 (*80kb*) to just 0.86 (*10kd*). Compared to the best results in Table 3.6, this explains the high standard deviations of up to 0.28 (4.7 times larger compared to a maximum of 0.06 in Table 3.7). The mean of the means for each data set size decreases from 0.82 (*10k*) to 0.59 (*80k*), the standard deviation increases for larger data sets from 0.17 to 0.26.

It is obvious that some choice of X-Drop values lead to an extreme worsening of the Rand indices and are undesirable for that kind of EST clustering. Using the same X-Drop values with the `-identity` option can still lead to acceptable results. Although the basic method of extending the exact seeds is the same in both cases, the identity filter removes hits with too many errors afterwards, which leads to sufficient clustering results in the latter case. If such a filter is not applied, too many spurious hits result in over-clustering of the ESTs.

3. Suffix Array Based EST Mapping and Clustering

Set	a	b	c	d	Mean
10k	0.87 (± 0.02)	0.96 (± 0.03)	0.95 (± 0.03)	0.96 (± 0.03)	0.94 (± 0.03)
20k	0.91 (± 0.02)	0.97 (± 0.02)	0.93 (± 0.04)	0.97 (± 0.02)	0.95 (± 0.02)
40k	0.91 (± 0.03)	0.92 (± 0.03)	0.85 (± 0.03)	0.90 (± 0.02)	0.90 (± 0.03)
80k	0.81 (± 0.03)	0.78 (± 0.03)	0.87 (± 0.03)	0.82 (± 0.02)	0.82 (± 0.03)

Table 3.9.: Mean Adjusted Rand Index for *A. thaliana* benchmark data sets, option `-leastscore` used for matching. For the calculation of the means, clusterings with X-Drop values of $X = 4$ and $X = 8$, and minimal match length $l = 40$ were excluded.

Therefore, X-drop values of $X = 4$ and $X = 8$, as well as a minimal match length of $l = 40$ were excluded from the analysis of the `-leastscore` option, and mean Rand indices calculated for the remaining parameter settings again. Table 3.9 shows the results: as can be deduced from Figures 3.15 to 3.18, the Rand indices are much more consistent for the remaining settings. The indices vary from 0.78 (*80kb*) to 0.96 (*10kb*). The mean of the means within each data size decreases now from 0.95 to 0.82. The standard deviations show more consistency and increase slightly from 0.02 to 0.03. Compared to the `-identity` option, the mean Rand indices increase by over 5% when using the `-leastscore` option.

Choice of Parameters

The performance of *Vmatch* allowed a thorough analysis of the impact of the choice of parameters used for EST clustering. Nevertheless, still no “best” parameter setting can be defined, as the clustering results clearly depend on, amongst others, the size, the composition, and the quality of the data set. In most applications the latter is unknown a priori, so that some kind of “default” setting has to be chosen.

The benchmark results suggest to prefer the `-leastscore` over the `-identity` option, if some constraints are considered. X-Drop values of $X \geq 4$ are not desirable and should be avoided. This leads to a smaller variation in the clustering results, which are then not as dependent on the remaining parameters. A minimal match length of $l = 100$, as often defined by different other clustering tools as the default match length, turned out to be a good choice. Therefore, for the upcoming analyses, the “default” *Vmatch* parameters for EST clustering were defined as shown in Table 3.10, although for some applications these default values might have to be adjusted.

<i>Vmatch</i> Option	Value
-1	100
-seedlength	33
-leastscore	196
-exdrop	2

Table 3.10.: *Vmatch* 'default' parameter settings for EST clustering.

Comparison to Different Tools

To assess the quality of *Vmatch* based EST clustering compared to other tools widely used for this purpose, we applied five different tools (see Section 2.3) to the same 16 data sets described before (see page 61): *CAP3*, *PaCE*, *d2_cluster*, *TGICL*, and *BLASTclust*. All tools were used with their default parameters settings, assuming that these settings have already been analyzed and appropriately set by the authors. As has been described above, an optimal parameter setting can vary between different data sets, and this can obviously hold true for the tools we compare to, here. Nevertheless, an impression of the quality performance of the different tools should still be possible. To achieve a fair comparison, *Vmatch* was used with the parameters described in Table 3.10, referred to as 'default' here. Additionally, the Rand index for an 'optimal' *Vmatch* parameter setting for the corresponding data set is shown.

Figure 3.19 shows the Rand index values for the various algorithms on all 16 data sets. In general, there is a clear tendency of lower quality for larger data sets. For the *10k* data sets, the 6 different tools perform almost equally well. *CAP3* is in terms of quality the best tool, it performs best in almost all cases, even for larger data sets. When using the optimal parameter setting, *Vmatch* outperforms *CAP3* for the *10k* and *20k* data sets. The 'default' parameter setting of *Vmatch* is comparable to tools like *PaCE*, *d2_cluster* and *TGICL* in terms of quality. The adjusted Rand index for *BLASTclust* drops significantly for larger data sets, showing the lowest quality in all data sets.

Except for *BLASTclust*, the Rand indices for the different tools are quite similar. Therefore, we performed a Friedman test [47] to assess whether there is a significant difference between the scores of the different tools. Friedman's test is (similar to classical balanced two-way ANOVA) a nonparametric test for data having a two-way layout (data grouped by two categorical factors). Unlike two-way analysis of variance, Friedman's test does not treat the two factors symmetrically and it does not test for an interaction between them.

3. Suffix Array Based EST Mapping and Clustering

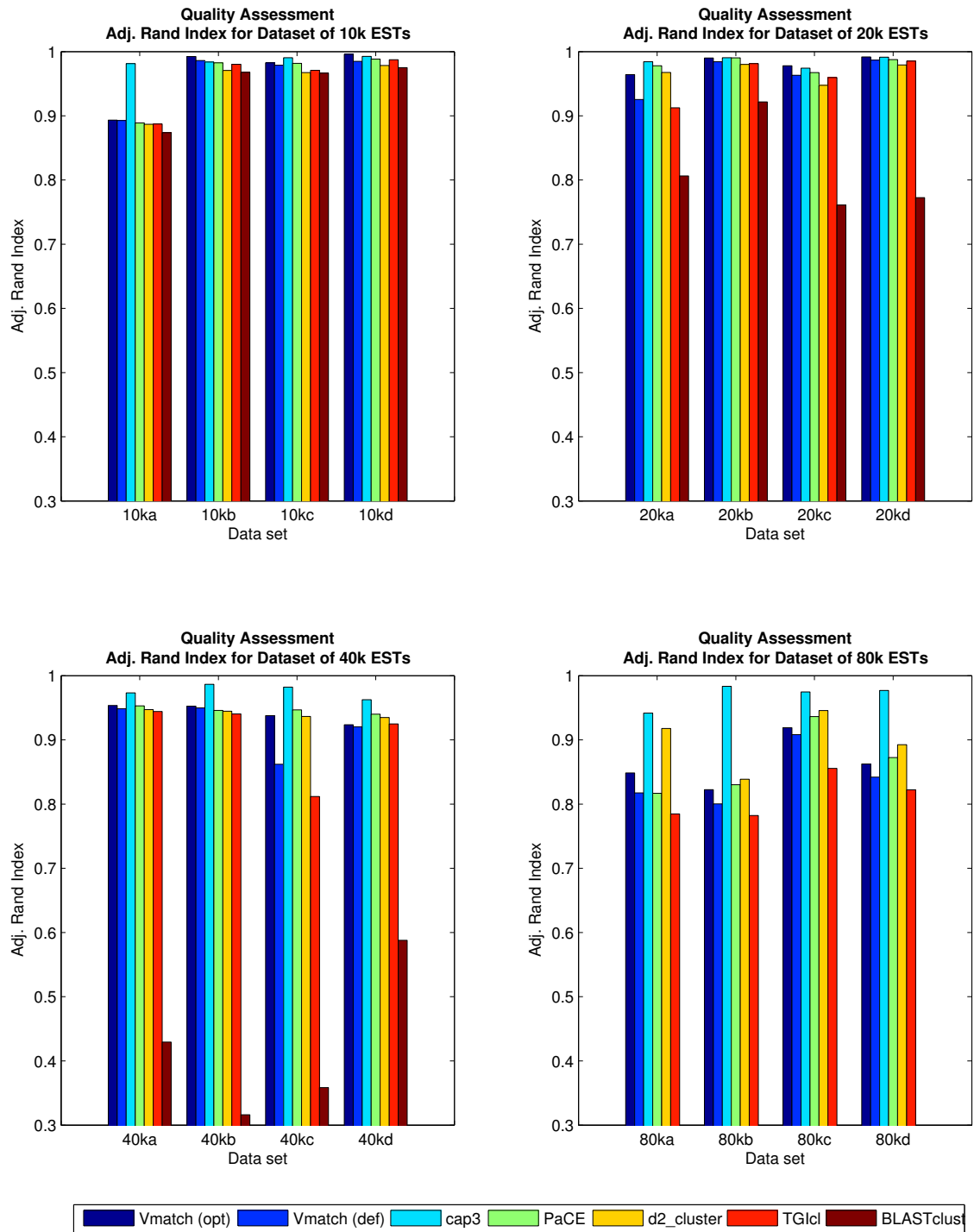


Figure 3.19.: Adjusted Rand Index for clustering tools *Vmatch*, *CAP3*, *PaCE*, *d2_cluster*, *TGICL*, and *BLASTclus* applied on *A. thaliana* data sets of different sizes.

Source	SS	df	MS	χ^2	$P > \chi^2$
Columns	337.3750	6	56.2292	72.2946	1.3822e-13
Error	110.6250	90	1.2292		
Total	448	111			

Table 3.11.: Friedman analysis of variance by ranks applied to the Rand Index values of the different clustering tools (see Figure 3.19). The p-value of 1.3822e-13 suggests that there is a significant ($\alpha = 0.05$) difference between the Rand Index values obtained for the 6 clustering tools.

Instead, it is a test for whether the columns are different after adjusting for possible row differences. The test is based on an analysis of variance using the ranks of the data across categories of the row factor, which does not depend on any distribution models.

Using Friedman’s analysis, we can test the null hypothesis that there is no significant difference between the Rand values of the different tools. We used the *Matlab* function *friedman* to perform the test. It returns a p-value for the null hypothesis that $\mu_i = \mu_j$ for all $i \neq j$. If the p-value is near zero, this casts doubt on the null hypothesis. A sufficiently small p-value suggests that at least one column-sample median is significantly different than the others. A 95% confidence interval was used ($\alpha = 0.05$).

Table 3.11 shows the results of the Friedman test. The p-value of 1.3822e-13 suggests that there is a significant difference between the Rand Index values obtained for the 6 clustering tools and therefore a significant difference in the quality of the resulting clusterings.

The rejection of H_0 does not imply that the means of the Rand values of all 6 tools are significantly different. To identify which pairs of tools are different, a multiple comparison test was conducted using *Matlab*’s *multcompare* function. It returns a matrix of pairwise comparison results and also displays a graph with each group mean represented by a symbol and an interval around the symbol. Two means are significantly different if their intervals are disjoint, and are not significantly different if their intervals overlap. The resulting graph is shown in Figure 3.20.

The figure shows as example all tools in red (*Vmatch* ’def’, *d2_cluster*, *TGICL*, and *BLASTclust*), whose Rand index values are significantly different from the Rand index values of *CAP3* (blue). In contrast, the Rand values of *Vmatch* ’opt’ and *PaCE* are not significantly different, suggesting that the quality of the tools is comparable to that of *CAP3*.

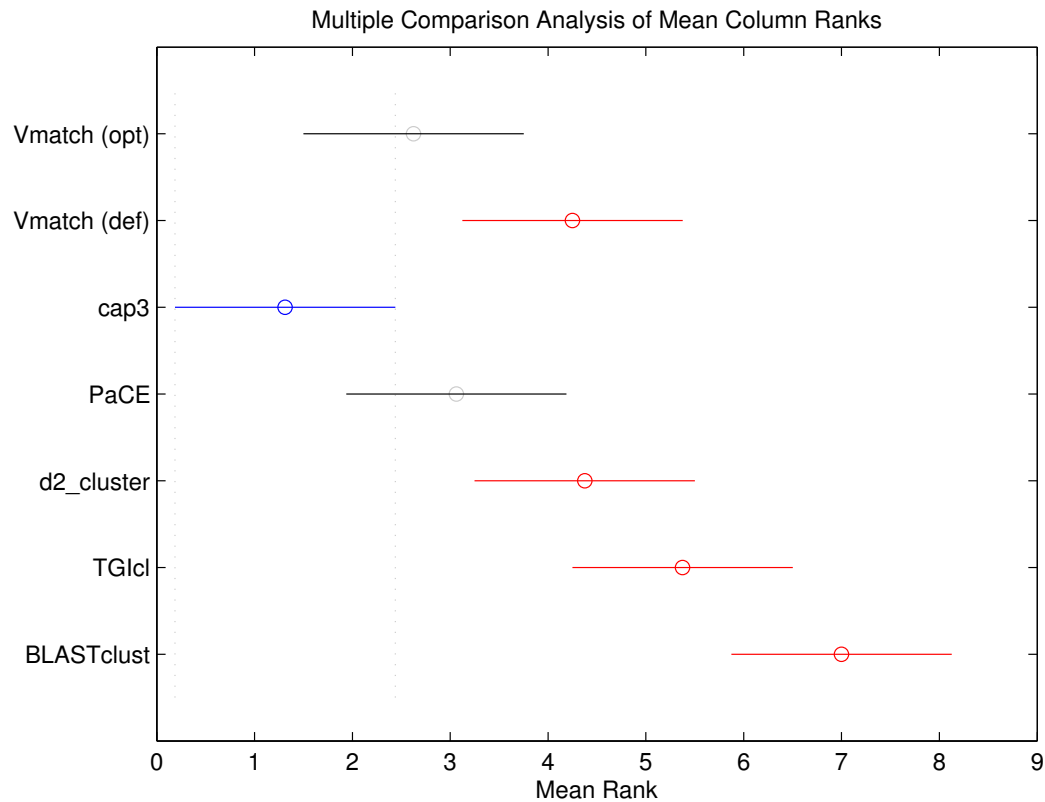


Figure 3.20.: Multiple comparison analysis of mean column ranks using the statistics of the Friedman test (Table 3.11). Each group mean is represented by a symbol and an interval around the symbol. Two means are significantly different if their intervals are disjoint.

In summary, *CAP3* is clearly the best tool to cluster ESTs, unfortunately its running time makes it unusable for larger data sets (see Section 3.5.4). A parameter-adjusted *Vmatch* and *PaCE* do not perform significantly worse. Using *Vmatch* with not-optimized parameters is in terms of quality comparable to *PaCE* and *d2_cluster*.

3.5.4. Performance Evaluation

High quality results are a desired outcome of an EST clustering tool. We have seen so far, that *CAP3* is the best tool in terms of quality, but as we will show in this section, *CAP3* is not capable of clustering tens of thousands of ESTs in acceptable time. To assess the run-time performance of each tool, we measured the running time for each tool for each of

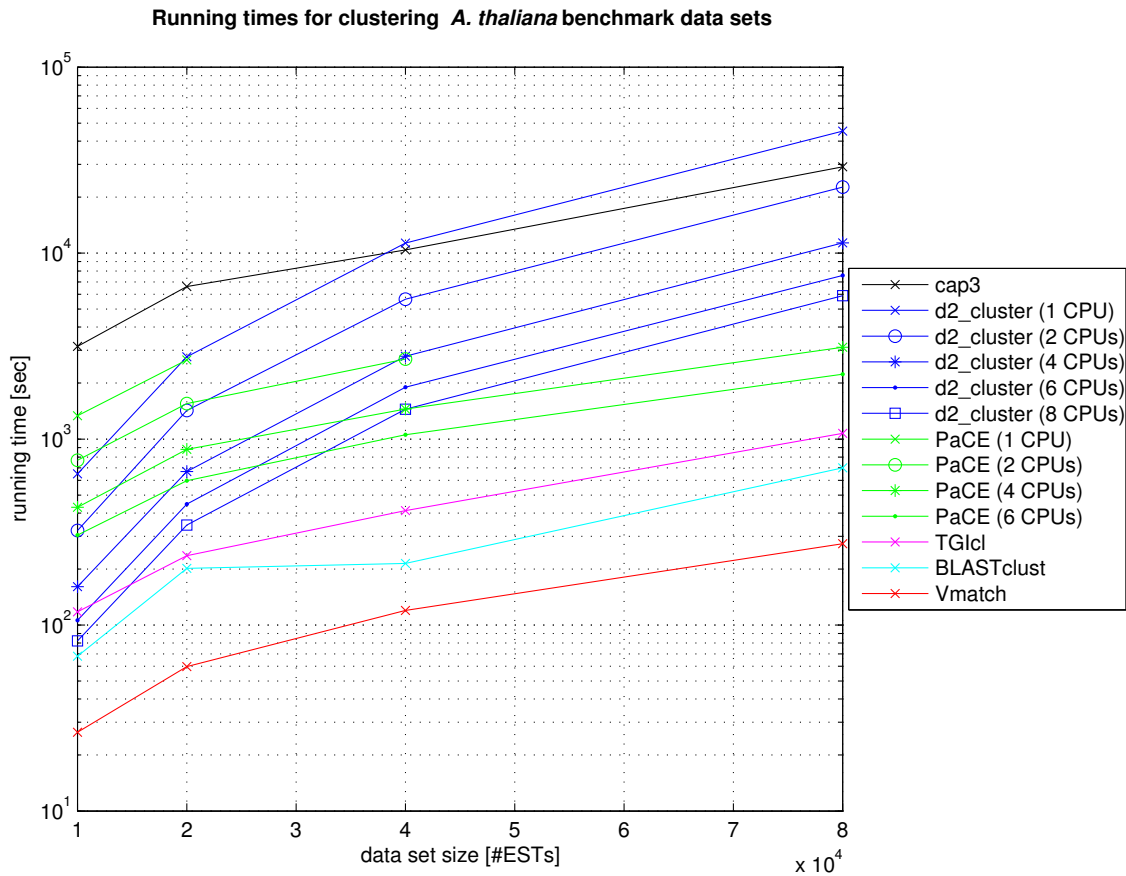


Figure 3.21.: Running times for tools *CAP3*, *d2_cluster*, *PaCE*, *TGICL*, *BLASTclust* and *Vmatch* for different data sets. Times are averaged for each of the groups 10k, 20k, 40k and 80k.

the 16 data sets on a Sun V880 with 8 SPARC-CPU (900 MHz) and 64GB RAM. Running times were averaged for each of the groups 10k, 20k, 40k and 80k. For the tools *PaCE* and *d2_cluster*, we ran the analyses with different numbers of CPUs.

Figure 3.21 shows the mean running times for the 4 different data set sizes and the 6 different tools. *Vmatch* is clearly the best tool in terms of performance. It outperforms all other tools (even if ran on multiple CPUs) by up to two orders of magnitude. With 26 seconds for 10,000 ESTs, *Vmatch* is twice as fast as *BLASTclust* (68 sec) and almost 120 times faster than *CAP3* (3147 sec). The parallel version of *PaCE* needs 1333 seconds using 1 CPU and still 304 seconds on 6 CPUs. *d2_cluster* is also capable of running in parallel mode and takes 650 seconds for 10k ESTs on 1 CPU; using 6 CPUs the running time decreases to 106 seconds.

For 80,000 ESTs, *Vmatch* is the only tool that allows for a parameter space evaluation in reasonable time. It clusters the sequences in 4.5 minutes. The parallel versions (6 CPUs) of *PaCE* and *d2_cluster* need 37 minutes and 2.1 hours, respectively. Even when using 6 CPUs, these tools are more than one order of magnitude slower than *Vmatch*. Running *d2_cluster* on a single CPU is even slower than *CAP3*: while *d2_cluster* needs 12.6 hours to finish, *CAP3* clusters the sequences in 8.1 hours.

3.6. Summary

One of the most popular applications of ESTs is the mapping to genomic sequence. It allows gene discovery, gene structure prediction and identification of alternative splicing. As the mapping procedure is a computationally expensive process, most genome browsers like NCBI's Map Viewer or the UCSC Genome Browser provide only static views of the data. The web-based tool *e2g* enables users to not only map a single EST to a genomic sequence, but instead find *all* matching ESTs for a region of interest. Using the high performance of enhanced suffix arrays it allows fast identification of ESTs and provides options to further analyze single ESTs in more detail.

For the application of EST clustering we have shown that *Vmatch* is a suitable and highly efficient tool to identify matching sequences in the EST data set. The choice of optimal parameters remains to be difficult, as results can vary significantly between different data sets and depend on its size, composition, and quality. Compared to 5 different EST clustering tools, in terms of clustering quality *Vmatch* performs slightly worse than *CAP3*, which turned out to be the best tool for EST clustering, albeit the worst in terms of running time. Comparing the running time, *Vmatch* outperforms all other tools by up to two orders of magnitude, which makes *Vmatch* the tool of choice for growing EST data sets. In the next chapter, we will describe an EST clustering pipeline using *Vmatch* both for pre-processing and clustering ESTs.

EST Clustering Pipeline

In the previous chapters we have described the general goal of EST clustering and a very efficient suffix array based approach of determining the clusters using *Vmatch*. As described in Section 2.2.2, the EST clustering procedure comprises not only the clustering step itself, but also a variety of pre- and post-processing steps. For an automated analysis, a pipeline calling different tools for each step is needed, preferable using a database to store information and data generated. In this chapter, we will describe a system which integrates *Vmatch* into such a pipeline.

4.1. Design Rationale

The clustering system comprises a computational pipeline and a database which stores the raw data, analysis results, and monitors the progress of the pipeline. A relational database management system (RDBMS) is a central part of the system as it holds a persistent view of the state of all tasks. Job parameters, in- and outputs are kept in the database, ensuring data integrity and concurrent user access. New categories of data can be added to the database without disruption to the existing system. The SQL query language allows an efficient retrieval of the results and powerful combination of the features either imported

4. EST Clustering Pipeline

or derived from the various analyses.

Figure 4.1 depicts an overview of the design of the pipeline. All data is stored in the central RDBMS (right column the figure). In each step, the data is either retrieved from or stored to the database. For tools working on flat or index files, data is exported to the local file system (left column in the figure). Computationally expensive tasks can be distributed on a compute cluster. Users can access the clustering and analysis results via a Web frontend, which again makes extensive use of the RDBMS' querying language.

The computational pipeline consists of the following steps, each of these will be described in further detail in the following sections:

1. *Data import*
Read and validate input sequences and store data into the database.
2. *Pre-Processing*
Clip high quality sequence and build *Vmatch* index.
3. *Repeat Masking*
Mask repeats and remove vector and mitochondrial sequence.
4. *Clustering*
Build index structures, cluster sequences and store cluster information into database.
5. *Assembly*
Assemble clusters to contig sequences, store assembly information into database.
6. *Annotation*
Annotate and functionally classify contigs, derive Gene Ontologies.
7. *Web Interface*
Make results available via Web interface, provide extensive search capabilities.

4.1.1. Data Import

To enhance the usability and search capabilities of the system, complete GenBank flat files are preferred as data import. In case where GenBank files are not available or sequences have not been published yet, the pipeline also accepts simple FASTA files. If GenBank files are imported, annotations including but not limited to library source, tissue type, cell type

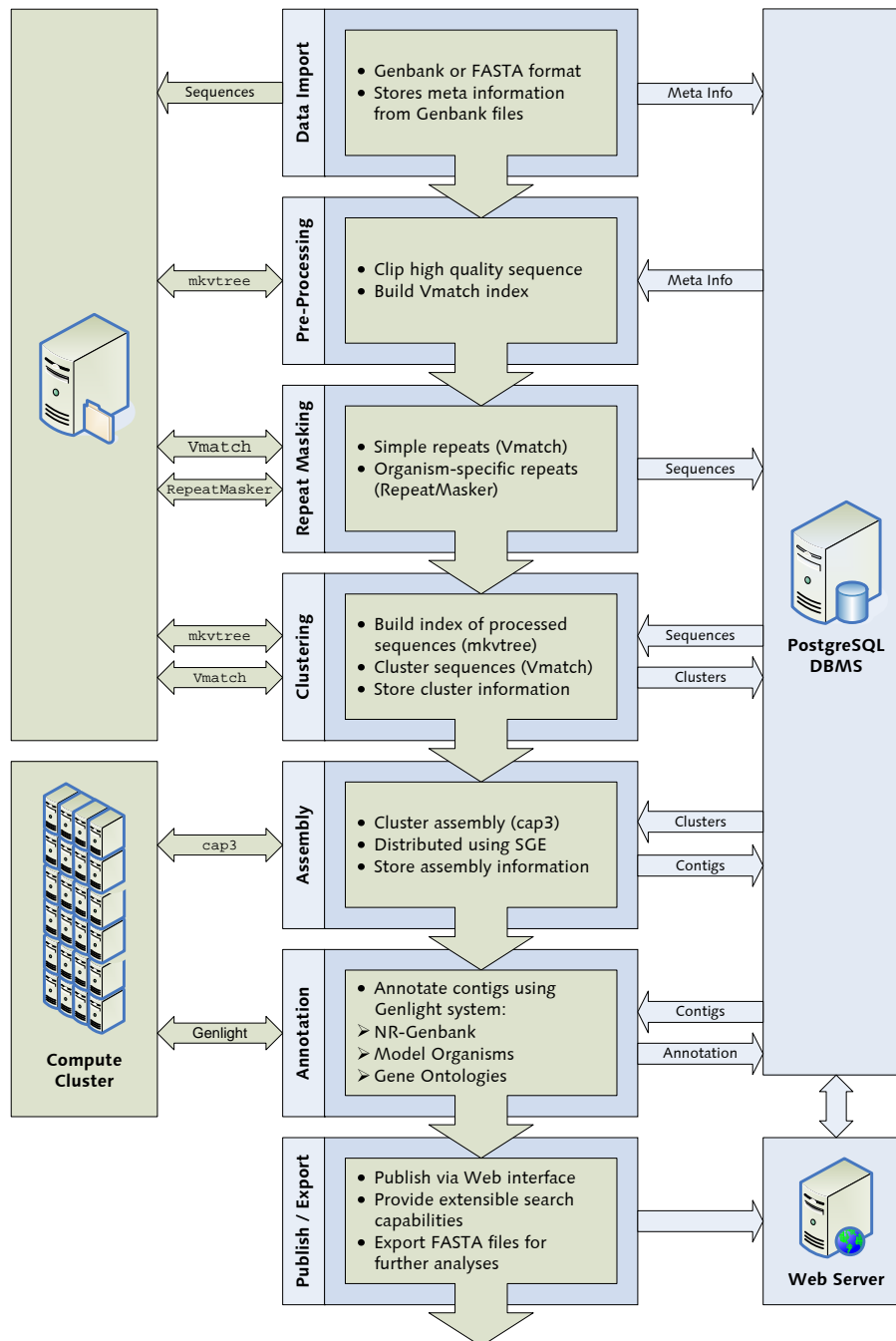


Figure 4.1.: A pipeline for EST clustering. The logical data flow is shown in the middle. Data and meta information is stored in a PostgreSQL DBMS. Sequence data is exported to the file system directly if efficient access is desired, e.g. in case of index files. Annotation for contig sequences is generated externally by the Genlight system. All analysis results are accessible through a Web interface.

4. EST Clustering Pipeline

and developmental stage are extracted directly from the flat files (feature: `source`, qualifiers: `clone_lib`, `tissue_type`, `cell_type` and `dev_stage`) and imported to the database. Where suitable, this information is used in following steps of the analysis.

While the meta information is stored in the database, FASTA files are generated and stored in the local file system to allow several tools in the following steps easy access to the sequence data (see Figure 4.1, left column).

4.1.2. Pre-Processing

Pre-processing is performed to remove low quality sequences. The high-throughput EST sequencing projects produce vast amounts of sequence data but do not always trim the sequences appropriately, instead provide information about the regions of best quality for each sequence. If available, information about high quality start and end of sequencing reads is used to trim sequences according to high quality regions to insure best sequence quality.

To identify each sequence, unique keys have to be generated for each sequence. We use SHA1 checksums [118] as identifiers, as we cannot rely on the IDs users provide in their input data. This way, each sequence is assigned a unique SHA1 key, generated from the description and sequence itself.

As last operation in this step, an enhanced suffix array is built for the sequence set using `mkvtree` for further processing with `Vmatch` in the next step.

4.1.3. Repeat Masking

After removal of low quality sequence regions, vector and other contaminant sequences have to be removed. Vector sequences are available from GenBank and VectorDB¹. VectorDB contains the sequences of almost all available vectors and many of these sequences are highly similar except for the cloning sites. This high redundancy leads to very high running times when using programs like RepeatMasker or `cross_match`, because many initial hits are found and extended by a costly Smith-Waterman alignment [138]. As this again is a perfect application for a suffix array, we use `Vmatch` for vector and contaminant masking. `Vmatch` identifies the initial seeds in all redundant sequences very rapidly and extends the matches using the X-Drop approach described earlier. As a result the running time of this step can be reduced significantly. If vector sequence is found at either the 5' or

¹<http://seq.yeastgenome.org/vectordb/>

3' end of the ESTs, it is trimmed in addition to additional sequence preceding or following the vector sequence. ESTs that have vector remaining in the middle of the sequences are discarded completely.

Repetitive elements as obtained from Repbase [73, 74] and GenBank are masked using RepeatMasker. Here, RepeatMasker performs considerably better than *Vmatch* in terms of repeat detection. Especially more distantly related repeats from related organisms are identified better if the Smith-Waterman algorithm is used instead of the X-Drop approach. The inclusion of related organisms' repeats is necessary in cases where no or very few repeats are known for the organism under consideration.

If hits against ribosomal RNA and mitochondrial sequences are found in the imported sequence set, the corresponding sequences are removed completely. Sequences that have less than 100 consecutive bases left after cleanup are discarded completely. Identified repeat sequences are masked as lower case characters to preserve the sequence information for the assembly step. During the clustering, a special character mapping prevents taking the lower case characters into account.

Information (including type of repeats found, rejected sequences, etc.) about the masking step as well as the processed sequences are stored in the database. Again, SHA1 keys are generated for the processed sequences to guarantee unique access to the sequences.

4.1.4. Clustering

After pre-processing and repeat masking, the ESTs can be clustered. First, an index has to be built for the processed sequences which can be used by *Vmatch*. Therefore, the sequences that passed the masking step are exported from the database and stored in the local file system. An index is built using `mkvtree`, this time using a special character mapping (`mkvtree` option `-smap`) that prevents lower case characters (i. e. masked regions) within the sequences to be matched.

Next, *Vmatch* is used to cluster the sequences as described in Section 3.4. Although the clustering is run on a single CPU, the pipeline allows for distributing the clustering step on a compute grid. This is especially useful in case of attempts to define appropriate clustering criteria. The evaluation described in Section 3.5 has been performed that way. A batch of clustering jobs (e. g. with different clustering parameters or different sequence sets) can be distributed on a computing grid. The process is database driven, i. e. all parameters for each clustering job are stored in the database. Each job stores the results

4. EST Clustering Pipeline

of the clustering back in the database, including information about the number of clusters and singlets, cluster membership for each sequence, running time of the clustering job, etc. This way, a comprehensive analysis can be done on the clustering results as demonstrated in Section 3.5.

4.1.5. Assembly

After the clustering step, clusters are assembled into contig sequences. An advantage of the assembly is the correction of sequencing errors in the resulting consensus sequence, given enough sequences in a cluster and sufficient coverage of the reconstructed mRNA sequence.

We use *CAP3* as assembly tool. Benchmarks have shown that *CAP3* is the best tool in terms of EST cluster assembly [97]. In agreement, our own quality assessment has demonstrated *CAP3* to be the best tool in terms of quality (see Section 3.5.3). The poor run-time performance of *CAP3* has been overcome by clustering the sequences with the much more efficient *Vmatch*. Nevertheless, assembly of hundreds of clusters with large amounts of sequences still takes a considerable amount of time. Therefore, the assembly process can also be distributed on a computing grid. Again, Perl scripts generate the appropriate *CAP3* calls. The jobs are distributed on the compute cluster and each job retrieves the cluster and sequence information directly from the database server.

The assembly results are stored back in the database. This includes information about the number of contigs and singlets for each cluster, and the contig and consensus sequences, including the position of each EST in the consensus. The position of each EST in the contig sequence will be important later on for identification of full length clones (see Section 4.1.6).

4.1.6. Annotation of Contig Sequences

To enhance the suitability of the system, a variety of sequence comparisons are performed at the protein level. The cluster consensus sequences and all singletons are subject to extensive BLASTX [8] and FASTY [120] homology searches vs. the non-redundant protein database (NR) from NCBI and the proteomes of various major model organisms using the high throughput analysis pipeline of the *Genlight* system [14, 15]. Proteome sets for model organisms are obtained from the *International Protein Index* (IPI) [79]. The IPI provides a top-level guide to the main databases: Swiss-Prot, TrEMBL, RefSeq and

Ensembl. It curates minimally redundant yet maximally complete sets of the indexed organisms.

Performing separate comparisons to different organisms allows a search for matching sequences based on the identity of any gene known from each species as well as query for genes which have matches in some but not all databases. This aids in the discovery and analysis of conserved and unique genes.

Functional Classification

In addition to these databases, we have included BLASTX searches in the COG and KOG databases which are used to functionally classify the contig sequences. The collection of *Clusters of Orthologous Groups* (COGs) of proteins is an approach to the identification of orthologous protein sets based on clustering of consistent genome-specific best hits [148]. COG currently contains 66 sequenced prokaryotic genomes. The KOG (eukaryotic orthologous groups) is an extension of the COG to complex, multicellular eukaryotes. It contains 7 sequenced genomes of animals, fungi, microsporidia, and plants. Each sequence in the COG and KOG database is assigned to at least one out of 25 functional categories, which consist of the main classes *Information Storage and Processing* (5 categories), *Cellular Processes and Signaling* (10 categories), *Metabolism* (8 categories), and *Poorly Characterized* (2 categories).

All sequences resulting from the clustering and assembly processes are compared to these protein sets using BLASTX and FASTY. The reason for incorporating FASTY analyses is that ESTs are often of low sequence quality, and sequencing errors can still exist in the assembled consensus sequences. FASTY is a version of FASTA that compares a DNA sequence to a protein sequence database, translates the DNA sequence in three forward (or reverse) frames and allows (in contrast to BLASTX) for frame shifts, maximizing the length of the resulting alignments.

Gene Ontologies

The Gene Ontology (GO) project [11, 59] is an ongoing international collaborative effort to generate consistent descriptions of gene products in different databases using a set of three controlled vocabularies or ontologies: *Biological Processes*, *Cellular Components*, and *Molecular Functions*. The GO vocabulary allows consistent searching of databases

4. EST Clustering Pipeline

using uniform queries. The availability of such vocabularies can be critical to the interpretation of high throughput approaches.

GO terms are imported into the system from the *Gene Annotation Database* (GOA) [28]. The GOA project aims to provide high-quality GO annotations to proteins in the UniProt/IPI database. Based on FASTY similarities with both mouse and human IPI sequences, GO annotations can be mapped to the contig sequences. This enriches the system by search capabilities for GO terms.

Identification of Full Length ORF Containing Contigs

A special interest lies in full length hits of the consensus sequences vs. known proteins. For this purpose, BLASTX and FASTY hits are categorized into four classes, representing the quality of the full length matches (see Figure 4.2): *Class 1*: Matches cover 100% of the sequence of a known protein. Additionally, the matched protein sequence begins with the conserved methionine and ends at a conserved STOP codon. *Class 2*: Matches covering 100% of the sequence of a known protein. Additionally, the matched protein sequence includes the initial methionine. *Class 3*: Matches capable of covering 100% of the matched protein sequence with no additional constraints. *Class 4*: Matches that cover the protein over almost its full length, allowing the match to start or end maximal 10 amino acids after/before the start or end of the protein. (See Figure 4.9 for an example of a Class 1 match.)

Identification of putative Full Length Insert Containing Clones

Often, biologists are interested in identifying a full length clone for further study and this desire has been met by the establishment of a number of the Gene Collections (the Mammalian Gene Collection, the *Xenopus* Gene Collection and the Zebrafish Gene Collection). The analysis described above has been extended to select potential full length insert containing *clones* that are available through the IMAGE consortium [96] and provide a simple yet powerful search tool to rapidly match homologous genes of interest to their counterparts in the organism under examination. The Gene Collections are an NIH initiative that supports the production of cDNA libraries, clones and 5'/3' sequences to provide a set of full-length (ORF) sequences and cDNA clones of expressed genes for a variety of model systems. Since the average length of the characterized full length vertebrate protein is 1,400 bases and the average sequence length of the consensus sequences is considerably

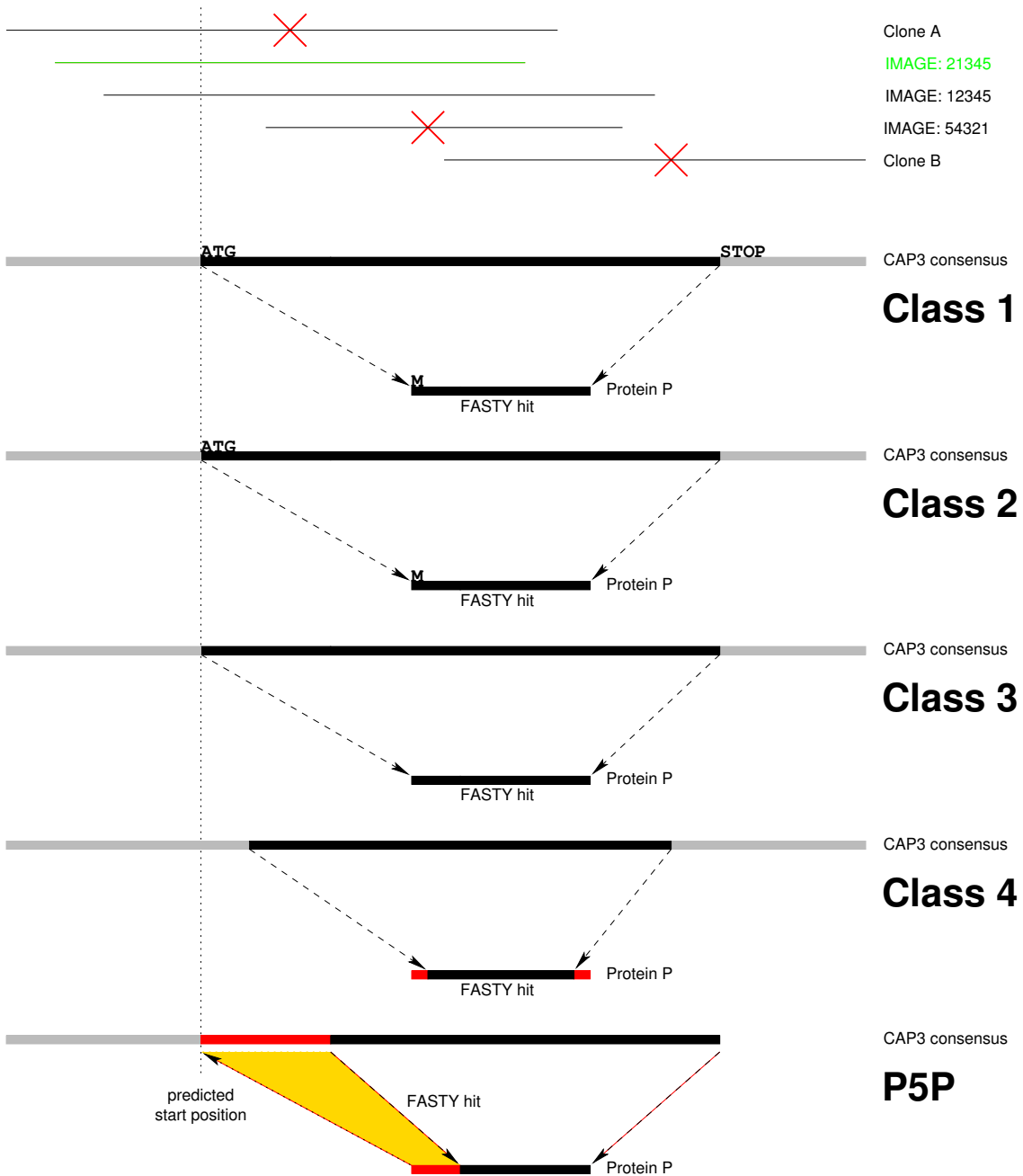


Figure 4.2.: Full length clone selection (top) and consensus sequence categories (bottom). The CAP3 contig sequence is compared to protein databases using BLASTX and FASTY, and hits categorized in 4 categories (see text for details). Predicted 5' contigs (P5P) have to have enough sequence to fill up the missing 5' end of the protein sequence. Clone selection: Clone A and B get discarded because of missing IMAGE id. Clone 54321 does not span 5' end of protein match. Clone 21345 is selected as most 5' clone fulfilling the requirements.

4. EST Clustering Pipeline

smaller in most cases, many sequences which are full length will not be detected by the previous approach. To identify additional clones that potentially carry a full length insert, we searched the database for sequence matches which were sufficiently long to include the start methionine but which did not have sufficient homology to be detected by the previous methods. Thus, a sequence with a query start position ($Start_q$) which is greater than the subject start site ($Start_s$) is potentially a full length open reading frame (hereafter referred to as *P5P*, *predicted 5 prime*). Clearly, the value of such a prediction decreases as the values of $Start_q$ increases and the predictive value increases with lower values of $Start_s$. Full length clones predicted by this method are subject to 3' truncations due to mispriming in poly(A) rich regions rather than at the polyA tail. Such regions would be characterized by the presence of the amino acid lysine (codons AAA, AAG) or asparagine (codons AAU, AAC).

Best FASTY hits are extracted for consensus sequences from all four full length categories as well as the *P5P* categories as described above. For sequences matching these categories, the most 5' EST contributing to the *CAP3* contig sequence is selected. In addition, the selected clone has to span the amino-terminal end of the FASTY protein match. Finally, to ensure the ready availability of the clones and therefore the utility of the analysis, the selected clone has to be available through the IMAGE consortium. See Figure 4.2 for an illustration of 5' clone selection.

4.1.7. Web interface

The results of the analyses described above are incorporated into the SQL database amenable to complex queries. The database can be accessed through a user friendly web based interface which allows individual and batch queries using accession, GI, and UniGene and TIGR cluster IDs. In addition, the user can query the sequence hits using any protein accession/GI number both singly and in batch mode. This allows a rapid identification of consensus sequences and their corresponding clones with hits to given protein sequences. The output of various queries displays the matching cluster(s) and links to a web page as presented in Section 4.3.2. For each cluster, links to the best hit for a number of model organisms are provided as well as links to the assembly result, consensus sequence generated by *CAP3*, and visual alignments of all FASTY results. GenBank accession numbers for each EST in the cluster and whether the corresponding clone has been identified as full length are provided. Additionally, for each cluster consensus the COG and KOG classification, as

well as the GO terms are available.

The analysis and database system provides a very powerful tool which allows users to take advantage of a number of technical and experimental advances. We have selected a couple of examples to illustrate possible types of queries in Section 4.3.2.

4.2. Database Schema

Figures 4.3 to 4.5 show the schema of the underlying database. Logically, it consists of 7 different sections, corresponding to the different steps in the clustering pipeline: *EST Sequence Sets*, *Clustering*, *Assembly*, *Sequence Analysis Data Sets*, *BLAST and FASTA Analyses*, *Full Length Clone Prediction*, *COG and KOG Classification*, and *Gene Ontology*. In the following sections we will give a description of the information stored in each section's entities.

EST Sequence Sets

The central tables in the *EST Sequence Sets* section of the schema (yellow tables in Figure 4.3) are the entities `seq_sets_catalog` and `sequences`, which keep information about the imported sequences like name, description, number of sequences, taxon, etc. and the actual sequences with clipping and masking positions. If sequences were imported from GenBank files, additional features are stored in the `est_source`, `est_xref`, and `est_seqinfo` tables. Here, among others, clones and libraries, tissue and cell types, developmental stages, cross references to other databases and information about high quality regions of the sequences are kept.

Clustering

The next step in the pipeline is the clustering of the EST sequences (red tables in Figure 4.3), which is based on an index structure. The parameters for building the index are stored in the `indexes` table including the name and the path to the index in the local file system. Based on these indexes, `clustersets` can be defined with a number of different parameters, the most important being *length*, *exdrop*, *leastscore*, *identity* and *seedlength* as described in Section 3.4. The clustering results are stored in the tables `clusters` and `sequences_clusters`. For each *set.id/seq.id* pair exactly one *clusterset.id/cluster.id* relation defines the resulting clusters.

4. EST Clustering Pipeline

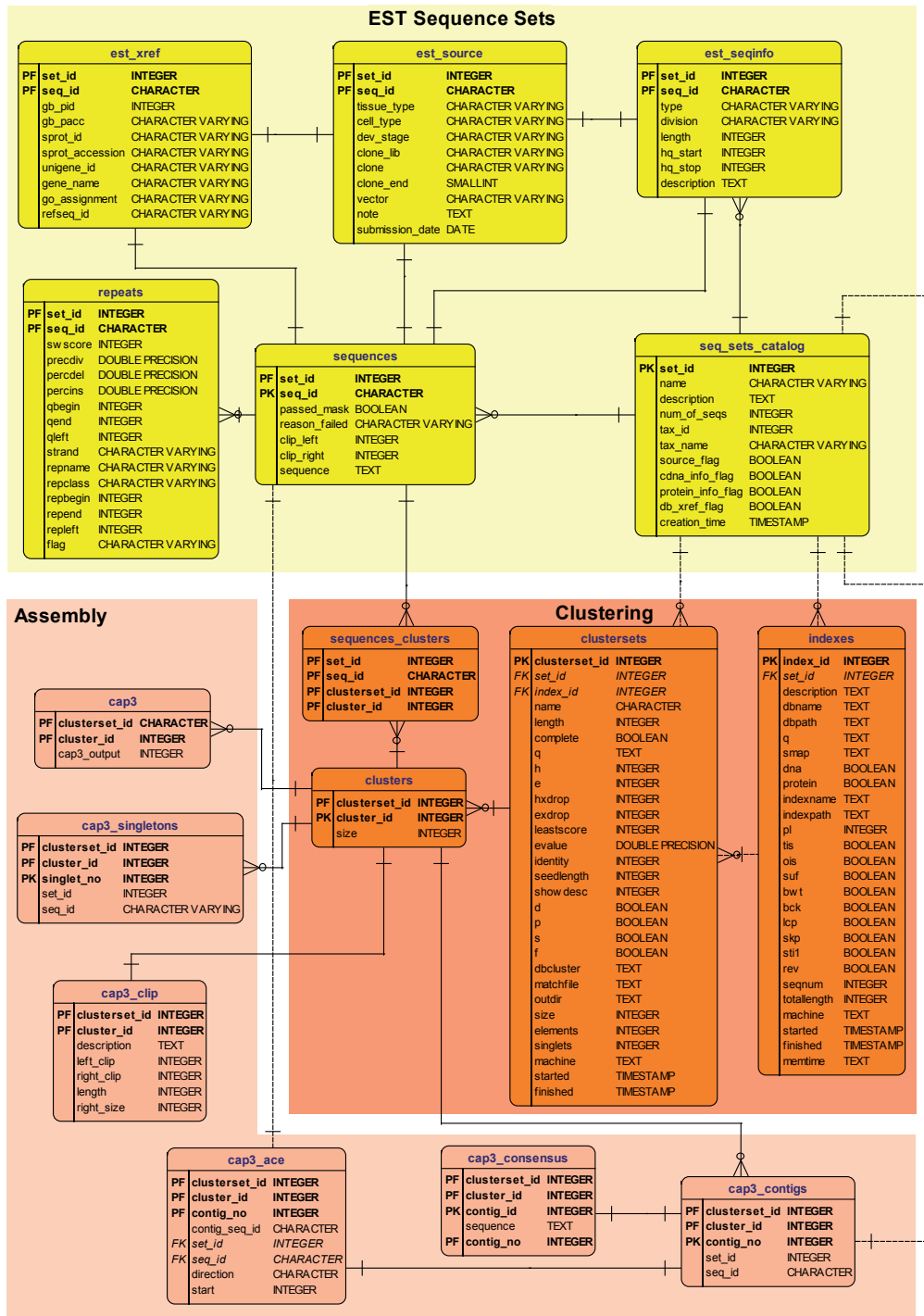


Figure 4.3.: EST clustering database schema (part 1 of 3).

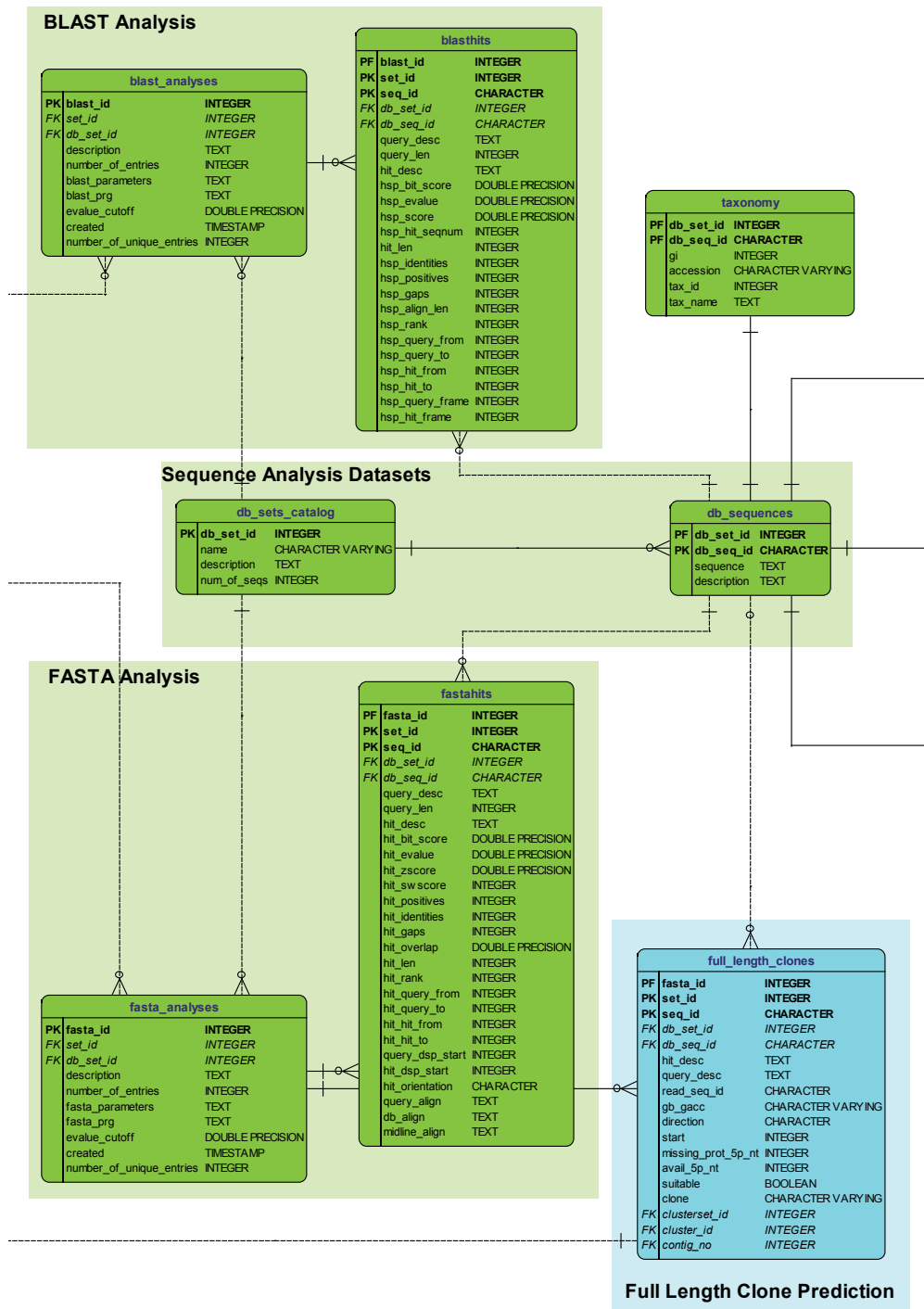


Figure 4.4.: EST clustering database schema (part 2 of 3).

4. EST Clustering Pipeline

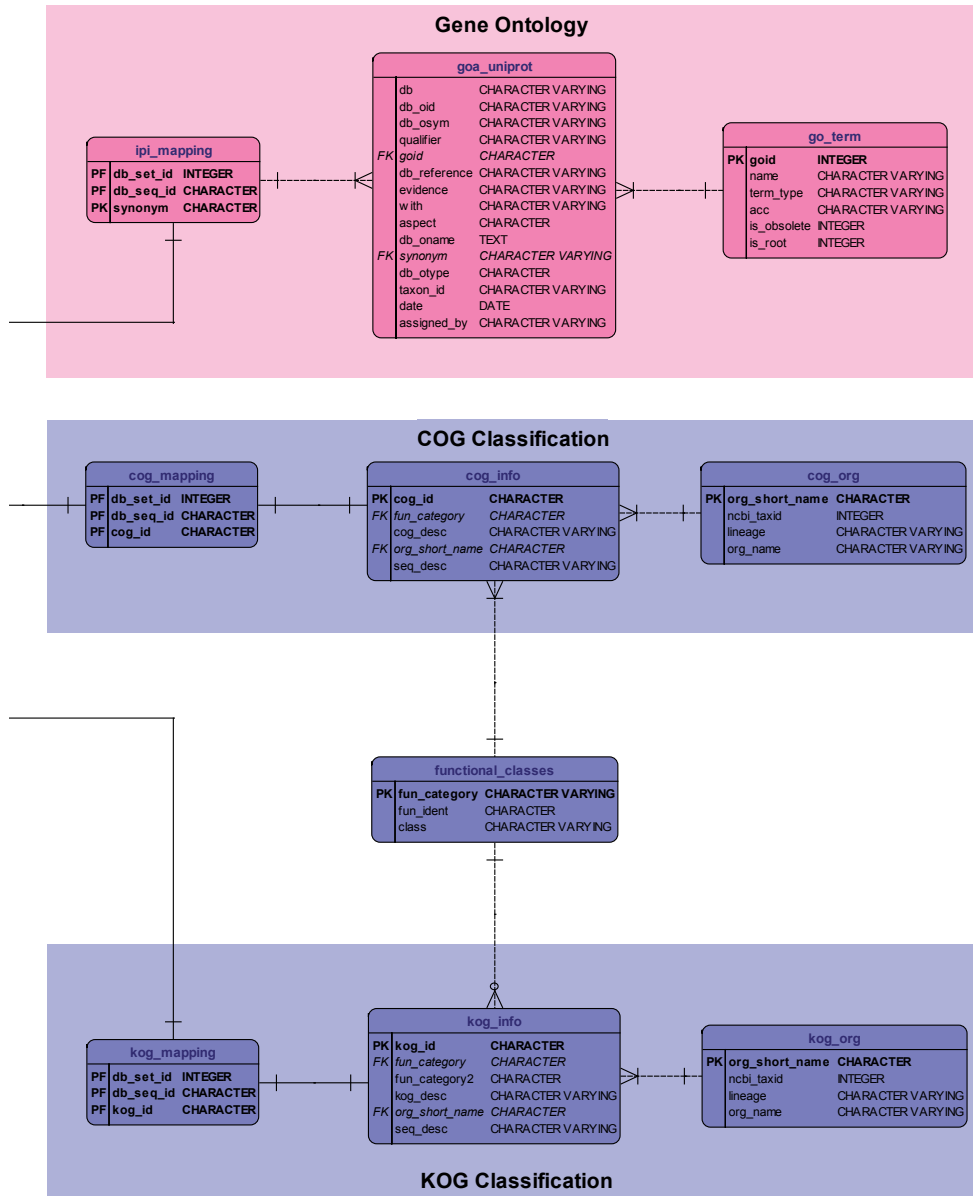


Figure 4.5.: EST clustering database schema (part 3 of 3).

Assembly

Following the clustering, clusters are assembled using *CAP3* (pink tables in Figure 4.3). The complete *CAP3* output is parsed and stored in different tables. While *cap3* holds the textual output, *cap3_contigs* stores the relation between the input sequences (*set_id* / *seq_id*) and the resulting contigs and their consensus sequences (*cap3_consensus*) for each cluster. The clipping information (*cap3_clip*) allows for selecting full length clones in the subsequence analysis (see below).

Sequence Analysis Data Sets

The cluster consensus sequences and singletons are merged to a new sequence set and subject to extensive sequence analyses (see below). Comparable to the EST data sets, target datasets are imported to the database (*db_sequences*), which are cataloged in the *db_sets_catalog* table. *taxonomy* information is kept for the data sets, if available, to allow for organism specific queries about the results (see green tables in Figure 4.4, middle).

BLAST and FASTA Analysis

The cluster consensus sequences and singletons are compared to the target datasets using BLAST and FASTA jobs, whose parameter settings are stored in the *blast_analyses* (green tables in Figure 4.4, top) and *fasta_analyses* tables (Figure 4.4, bottom). Resulting hits are then kept in the *blasthits* and *fastahits* tables, which allow efficient queries about the BLAST and FASTA results.

Full Length Clone Prediction

The full length clone prediction allows the identification of clones which very likely contain the full open reading frame of the protein. The prediction is based on the *cap3_acc* and *cap3_contigs* tables, which store information about the exact location of the clones in the assembled contig sequence. Combined with the FASTA analysis, full length clones can be identified. Results are stored in the *full_length_clones* table.

COG and KOG Classification

COG and KOG classifications are based on BLASTX hits to proteins from the COG or KOG databases, respectively (blue tables in Figure 4.5). The `cog_mapping` and `kog_mapping` tables allow the mapping from the protein sequences to the functional categories stored in `cog_info` and `kog_info`.

Gene Ontology

Gene Ontology terms are based on FASTY hits to IPI sequence sets (pink tables in Figure 4.5). `ipi_mapping` allows to map hits to the imported IPI sequences to Gene Ontology terms by the GO synonym. The GO terms and associated IPI sequences are imported from the GOA database.

4.3. Implementation

4.3.1. Clustering Pipeline

The open-source object-relational DBMS PostgreSQL² is used as the central DBMS in the clustering pipeline. PostgreSQL was chosen because it supports a large part of the SQL standard and offers many modern features like complex queries, foreign keys, triggers, views and transactional integrity.

The pipeline is implemented in Perl, as its support for regular expressions makes the handling of sequences and parsing of textual output fairly simple. At various places we make use of BioPerl [141]. BioPerl is a collection of Perl modules that supports the development of Perl scripts for bioinformatics applications. It provides reusable Perl modules that facilitate writing Perl scripts for sequence manipulation, accessing of databases using a range of data formats and execution and parsing of the results of various molecular biology programs.

Perl's DBI module is used for access to the database. DBI is the standard database interface module for Perl. It defines a set of methods, variables and conventions that provide a consistent and transparent database interface, independent of the actual database being used. Therefore, the scripts of the clustering pipeline will work on different database types (e.g. MySQL, MSSQL, Oracle, Informix, Sybase, etc.) by using the API defined by DBI.

²<http://www.postgresql.org>

The Perl scripts handle the data import and processing. External programs are called from within the scripts and results parsed and stored in the database. The scripts allow to generate a batch of clustering jobs with different clustering parameters. These clustering jobs, as well as the assembly process, can be distributed on a computing grid.

For job distribution, the Sun Grid Engine (SGE) [145] is used as resource management software. The Grid Engine project is an open source community effort to facilitate the adoption of distributed computing solutions. It provides a single point of access to the computing grid by accepting jobs submitted by users and scheduling them for execution on appropriate systems in the grid based upon resource management policies.

4.3.2. User Interface

While the clustering pipeline is driven by command-line Perl scripts, all analysis results can be accessed by a user-friendly web interface. The CGI-based scripts are hosted on the *Bielefeld University Bioinformatics Server* (BiBiServ). The server is connected to the underlying DBMS, which hosts the results of the clustering and all analyses. The user can query the database in various ways and download data for further processing. In the next sections we will give a overview of the capability of the system.

Query interface

Figure 4.6 shows the query interface of the *XenDB* database, which is an example of the described pipeline applied to *X. laevis* ESTs. The query interface allows basic queries for accession and GI numbers of the imported sequences, cluster IDs of the generated clusters. UniGene or TIGR clusters can be searched if the information was available in the imported files. This allows a comparison of clustering results to other EST databases. GO terms can be searched for as well as description lines. These queries are available in single or batch mode through file uploads.

Figure 4.7 shows the result of a search for cluster number 2341. As can be seen on the result page, the cluster contains 9 contigs as well as 9 singletons. The high number of contigs is a result of the *CAP3* assembly of the cluster sequences. A reason for such a high number of contigs can be misassembly, but much more often *CAP3* splits clusters apart that contain different transcript isoforms of the same gene.

For each contig, the best FASTY hit with the corresponding description and E-values is shown in the overview of the results. For singletons, also links to the original GenBank

4. EST Clustering Pipeline

XenDB - Search XenDB - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://bibiserv.techfak.uni-bielefeld.de/cgi-bin/xendb-search?cl_ Go

Universität Bielefeld **XenDB**

RESULT

BiBiServ
Bielefeld University Bioinformatics Server
XenDB

Search for *X. laevis* GenBank Accessions, GIs or Clusters:

Single	Multiple (File Upload)
Accession: <input type="text"/>	<input type="text"/> Browse...
GI number: <input type="text"/>	<input type="text"/> Browse...
XenDB cluster: <input type="text"/>	<input type="text"/> Browse...
UniGene cluster: <input type="text"/>	<input type="text"/> Browse...
TIGR cluster: <input type="text"/>	<input type="text"/> Browse...
Singleton: <input type="text"/>	<input type="text"/> Browse...
Gene Ontologies: <input type="text"/>	<input type="text"/> Browse...
Search in Description: <input type="text"/>	

Search for GenBank Accessions or GIs in FASTY hits:

Single	Multiple (File Upload)
Accession: <input type="text"/>	<input type="text"/> Browse...
GI number: <input type="text"/>	<input type="text"/> Browse...

Search for keywords in *X. laevis* annotation data:

Tissue-Type

Reset Search Database

Figure 4.6.: Query interface for the clustering and analysis results. Single queries as well as batch uploads can be used to search the database. *XenDB* is the name of a particular database viewed here.

entries are provided, as well as the UniGene and TIGR clusters, if available.

Contig View

The contig view (Figure 4.8) shows detailed information about a contig of a certain cluster. The original *CAP3* output can be accessed, the contig sequence downloaded. Best FASTY hits for the non-redundant database and for each of 9 model organisms are shown. GO terms and COG/KOG classifications for the contig sequence are provided. A hyperlink brings up a graphical overview of the hits. If a full length clone is available (see Section 4.1.6), the clone is highlighted in green in the *Clone* column of the sequence information.

Alignment Visualization and Full length clone hit

A more detailed visualization of the FASTY alignment can be generated by following the hyperlink from the contig view. The visualization shows the start and end of the protein match within the contig sequence. In the example (Figure 4.9), the contig has a full length hit (class 1) against a protein sequence. The contig is also long enough, so that the corresponding clone on the 5' end has good chances to include the full length insert (see Section 4.1.6).

Gene Ontology Results

Figures 4.10 (top) shows the results of the query for GO term *eye*. If the query is too general, as is the case in the example, a list of all matching GO terms is shown with the numbers of contigs matching that term. From the list, a more specific term can be chosen and all contigs matching the term are retrieved from the database.

Species Mapping

The web interface allows also for searching for accession or GI numbers in the FASTY hits. The comparative query allows the identification of the set of contig sequences most related to a set from another organism. Thus, the database is designed to address many researchers facing a critical issue: the comparison of genomic studies in one organism and their application to studies in another model organism. This task is faced by many laboratories attempting to extract the information gained in human, mouse, fly and worm microarray and library sequencing studies which often consist of large tables of genes.

4. EST Clustering Pipeline

XenDB - Search Results for XenDB cluster 2341

#	Cluster	Contig	Hit-Description	E-value
1	2341	7	gi 28839578 gb AAH47840.1 Unknown (protein for IMAGE:5413486) [Danio rerio]	0
2	2341	3	gi 27370889 gb AAH41262.1 MGC52825 protein [Xenopus laevis]	6.4e-105
3	2341	1	gi 27881711 gb AAH44715.1 MGC52578 protein [Xenopus laevis]	1.9e-88
4	2341	6	gi 9931991 emb CAC04528.1 DYSKERIN [Mus musculus]	6.4e-71
5	2341	9	gi 9931991 emb CAC04528.1 DYSKERIN [Mus musculus]	1.8e-52
6	2341	4	gi 539691 dbj BAA06440.1 HMG-X protein [Xenopus laevis]	3e-39
7	2341	2	gi 27370889 gb AAH41262.1 MGC52825 protein [Xenopus laevis]	8.2e-28
8	2341	5		
9	2341	8		

#	Accession	GI	Cluster	UG	TGI	Singleton	Hit-Description	E-value
1	BE188811	9730515	2341	21410		9	gi 28839578 gb AAH47840.1 Unknown (protein for IMAGE:5413486) [Danio rerio]	8.9e-73
2	CA983095	27515749	2341	42618	TC259849	2	gi 27370889 gb AAH41262.1 MGC52825 protein [Xenopus laevis]	5.5e-57
3	BJ092916	17592256	2341	42618	TC259864	7	gi 27370889 gb AAH41262.1 MGC52825 protein [Xenopus laevis]	3.4e-51
4	AJ009279	3821755	2341	47703	TC285474	3	gi 28839578 gb AAH47840.1 Unknown (protein for IMAGE:5413486) [Danio rerio]	4.8e-44
5	AW768029	7700043	2341	4150		4	gi 27881711 gb AAH44715.1 MGC52578 protein [Xenopus laevis]	8.7e-32
6	BI446923	15271630	2341	47703		8	gi 27881711 gb AAH44715.1 MGC52578 protein [Xenopus laevis]	6.2e-27
7	BE507071	9726846	2341	4150	TC259867	5	gi 27881711 gb AAH44715.1 MGC52578 protein [Xenopus laevis]	8e-25
8	BF025588	10756007	2341	4150	TC259867	1		
9	BG019103	12474972	2341	na	TC285477	6		

Figure 4.7.: Search result for cluster number 2341. The overview shows all contigs and singletons of the cluster. Links to GenBank entries, UniGene and TIGR clusters are provided. Information about best FASTY hits are shown.

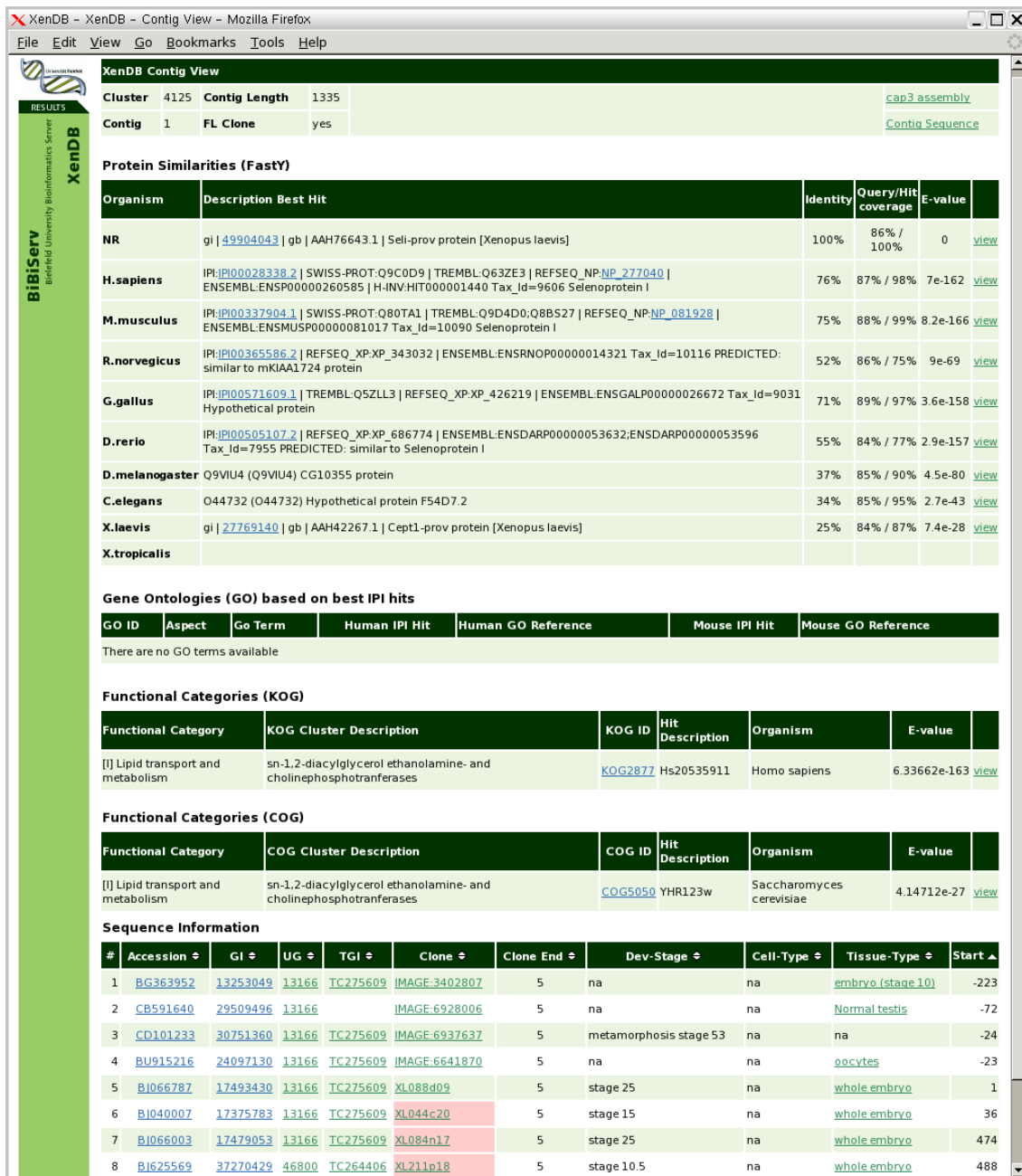


Figure 4.8.: Contig view: Summary of clustering information including hyperlinks to CAP3 output and FASTA sequences. FASTY similarities to NR protein database and model organisms. Shown are also GO terms and COG/KOG classifications. Detailed information about the sequences in the cluster are available, including potential full length clones.

4. EST Clustering Pipeline

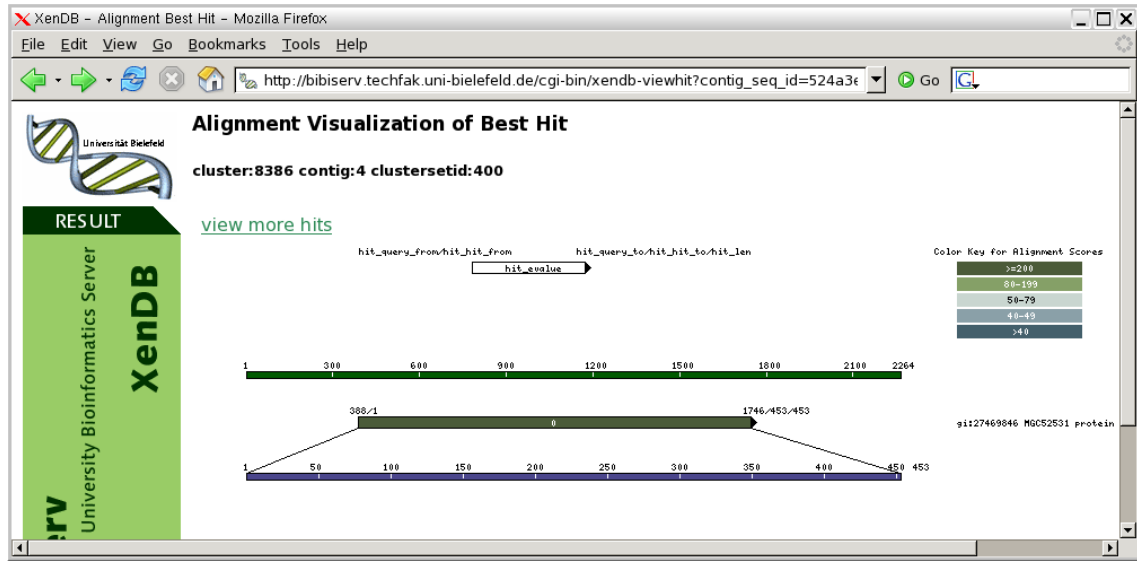


Figure 4.9.: Graphical visualization of a FASTY alignment. The example shows a full length protein hit (class 1), where the contig sequence contains a clone which very likely has a full length insert.

While other databases such as UniGene or TIGR Gene Indices also provide collections of clustered ESTs, the unique batch functionality of mapping results from other organisms to EST sequences of the organism in study and retrieving their potential full length clones was not available before. Moreover, our implementation is specifically designed and focused on relating the analyzed sequence data to the major model organisms.

This way the user can perform a comparative analysis and mapping of experimental results or other external sources from popular model organisms to the organism in study. For example, in many instances, the outcome of e. g. a microarray type of experiment is a variety of tables listing regulated genes and the associated expression changes. In cases where few published array studies are available, the results of existing extensive databases of expression studies for a variety of model organisms can possibly be mapped to the organism of interest, if the corresponding genes could be identified.

The NCBI maintains a common database, the Gene Expression Omnibus [13], which contains data from over 98,000 samples including over 2,800 human, 1,800 mouse and 200 Drosophila data sets. To identify likely homologues of these genes in the organism of interest, GenBank accession numbers can be obtained from the NCBI Gene Expression

XenDB - Search Results for Gene Ontology eye

#	GO ID	Name	Sequences
1	GO:0042706	eye photoreceptor cell fate commitment	5
2	GO:0001747	eye morphogenesis (sensu Mammalia)	32
3	GO:0001654	eye morphogenesis	5
4	GO:0005212	structural constituent of eye lens	14
5	GO:0042462	eye photoreceptor cell development	1
6	GO:0048075	positive regulation of eye pigmentation	2

XenDB - Search Results for Gene Ontology GO:0001654

GO ID	Aspect	GO Term
GO:0001654	biological_process	eye morphogenesis

#	Cluster	Contig	Singleton	Human IPI Hit	Human Hit-Description	Mouse IPI Hit	Mouse Hit-Description
1	14830	1		IPI00010862	Bone morphogenetic protein receptor type IB precursor	IPI00128197	Ensembl_locations(Chr-bp):3-142484451 Bone morphogenetic protein receptor type IB precursor
2	2735	1		IPI00010862	Bone morphogenetic protein receptor type IB precursor	IPI00128197	Ensembl_locations(Chr-bp):3-142484451 Bone morphogenetic protein receptor type IB precursor
3	731	1				IPI00121519	Ensembl_locations(Chr-bp):2-80635255 Neurogenic differentiation factor 1
4			123968	IPI00010862	Bone morphogenetic protein receptor type IB precursor	IPI00128197	Ensembl_locations(Chr-bp):3-142484451 Bone morphogenetic protein receptor type IB precursor
5			215288			IPI00114230	Ensembl_locations(Chr-bp):10-130473460 Neurogenic differentiation factor 4

Figure 4.10.: Result for a search for GO term eye. A list of matching GO terms is shown if the query was too general (top). From the list, a more specific term can be chosen (bottom).

4. EST Clustering Pipeline

#	Input	XenDB Cluster	XenDB Contig	Xenopus Accession	FL clone	Hit-Description	E-value	Organism Rank	Overall Rank
1	O18381	8386	[E]		✓	gj 12643549 sp O18381 PAX6_DROME Paired box protein Pax-6 (Eyeless protein)	9.9e-69	9	100
2	Q01071	7661		BG810687	✗	gj 7301418 gb AAF56544.1 CG8328-PA [Drosophila melanogaster] gj 24650220 ref NP_524503.2 CG8328-PA [Drosophila melanogaster] gj 55715433 gb AAV59225.1 enhancer of split complex mdelta protein [Drosophila melanogaster] gj 55715420 gb AAV59213.1 enhance	4.1e-11	44	96
3	Q01071	3876	1		✓	gj 7301418 gb AAF56544.1 CG8328-PA [Drosophila melanogaster] gj 24650220 ref NP_524503.2 CG8328-PA [Drosophila melanogaster] gj 55715433 gb AAV59225.1 enhancer of split complex mdelta protein [Drosophila melanogaster] gj 55715420 gb AAV59213.1 enhance	1.7e-10	85	89
4	Q01070	7661	1		✓	gj 7301419 gb AAF56545.1 CG8333-PA [Drosophila melanogaster] gj 24650222 ref NP_524504.2 CG8333-PA [Drosophila melanogaster] gj 55715434 gb AAV59226.1 enhancer of split complex mgamma protein [Drosophila melanogaster] gj 55715421 gb AAV59214.1 enhance	2.2e-18	45	78
5	AAD52845	18882	1		✗	gj 23093959 gb AAF50508.2 CG8114-PA, isoform A [Drosophila melanogaster] gj 24660486 ref NP_729306.1 CG8114-PA, isoform A [Drosophila melanogaster] gj 5817604 gb AAD52845.1 Pebble [Drosophila melanogaster]	5.6e-61	7	28
6	AAB61239	10868	1		✓	gj 2196776 gb AAB61239.1 bunched gene product [Drosophila melanogaster]	2.1e-21	9	65
7	AAF48990	10585	1		✓	gj 45447061 gb AAF48990.3 CG12238-PA [Drosophila melanogaster] gj 45550083 ref NP_608334.3 CG12238-PA [Drosophila melanogaster] gj 62901062 sp Q9VWF2 SAYP_DROME Supporter of activation of yellow protein (Enhancer of yellow protein 3)	6.7e-22	4	29
8	P48554	9517	1		✓	gj 607070 emb CA84710.1 RacB [Drosophila melanogaster] gj 21430054 gb AAM50705.1 GM13874p [Drosophila melanogaster] gj 7295237 gb AAF50559.1 CG8556-PA [Drosophila melanogaster] gj 21356563 ref NP_648121.1 CG8556-PA [Drosophila melanogaster] gj 134695	1.5e-109	6	33
9	CAA76941	18485	1		✗	gj 4377454 emb CAA76941.1 UNC-13 protein [Drosophila melanogaster]	2.5e-161	1	33
10	P36872	1777	1		✓	gj 23170875 gb AAN13455.1 CG6235-PF, isoform F [Drosophila melanogaster] gj 7299303 gb AAF54498.1 CG6235-PA, isoform A [Drosophila melanogaster] gj 16768962 gb AAL28700.1 LD12394p [Drosophila melanogaster] gj 24645612 ref NP_731451.1 CG6235-PF, isoform	0	22	65

Figure 4.11.: Species Mapping: Identification of potential *Xenopus* homologues to *Drosophila* genes. The page shows *Xenopus* contigs (column: *XenDB contig*) producing the top ranked hit to the potential *Drosophila* homologue (column *Input*). Ranks are shown for the organism in question (*Organism Rank*) and the for all hits found by the FASTY analysis (*Overall Rank*). If the contig is identified as full insert containing clone it will be marked as such in the output (*FL clone*).

```

SELECT DISTINCT
  tt4.gi,
  COALESCE(t4.cluster_id,t5.cluster_id) AS cluster_id,
  t4.contig_no,
  t6.gb_gacc,
  COALESCE(t7.suitable,t8.suitable) AS fl_clone,
  tt4.hit_desc,
  tt4.hit_evalue,
  tt4.count AS rank_org,
  tt4.hit_rank AS rank_all
FROM
  (SELECT
    tt1.query_shal_id,
    tt1.hit_shal_id,
    tt1.gi,
    tt1.hit_desc,
    tt1.hit_evalue,
    tt1.hit_rank,
    COUNT(tt2.query_shal_id)
  FROM
    (SELECT
      t1.query_shal_id,
      t2.hit_shal_id,
      gi,
      hit_desc,
      hit_evalue,
      hit_rank,
      taxid
    FROM
      fasta_obj_2032_q339_db328 t1,
      (SELECT
        t1.gi,
        t1.hit_shal_id,
        MAX(hit_bit_score),
        taxid
      FROM
        nr328_gis t1,
        fasta_obj_2032_q339_db328 t2,
        nr328_gi_tax t1a
      WHERE
        t1.gi IN ('607070') AND
        t1.hit_shal_id = t2.hit_shal_id AND
        t1.gi = t1a.gi
      GROUP BY
        t1.hit_shal_id,
        t1.gi,
        t1a.taxid
      ) t2
    WHERE
      t1.hit_shal_id=t2.hit_shal_id AND
      t1.hit_bit_score=t2.max
    ) tt1,
    fasta_obj_2032_q339_db328 tt2,
    nr328_tax tt3
  WHERE
    tt2.query_shal_id=tt1.query_shal_id AND
    tt2.hit_shal_id=tt3.hit_shal_id AND
    tt3.taxid=tt1.taxid AND
    tt2.hit_rank<=tt1.hit_rank
  GROUP BY
    tt1.query_shal_id,
    tt1.hit_shal_id,
    tt1.gi,
    tt1.hit_desc,
    tt1.hit_evalue,
    tt1.hit_rank,
    tt1.taxid
  ) tt4
LEFT OUTER JOIN
  cap3_ace_1 t4 ON tt4.query_shal_id=t4.contig_seq_id
LEFT OUTER JOIN
  cap3_singletons_1 t5 ON t5.seq_id=tt4.query_shal_id
LEFT OUTER JOIN
  db_xref_2 t6 ON t6.seq_id=tt4.query_shal_id
LEFT OUTER JOIN
  cl400_5p_clones t7 ON t7.query_shal_id=tt4.query_shal_id AND
  t7.suitable=true
LEFT OUTER JOIN
  cl400_5p_clones_reverse t8 ON t8.query_shal_id=tt4.query_shal_id AND
  t8.suitable=true
ORDER BY tt4.hit_evalue ASC

```

Figure 4.12.: SQL code for mapping accessions of FASTY hits to cluster contigs.

Omnibus and used to query the database to identify potential homologues of the regulated genes and predicted full length clones. As these sequences are available from commercial sources, they can be readily obtained and tested using the various experimental approaches.

Figure 4.11 shows as example the result table of a mapping from *Drosophila* genes to *Xenopus* contigs (see Section 5.5.4 for details). It includes links to the matching cluster and contig, the E-value and rank in the FASTY result list and whether a full length clone has been identified. The contig web link leads to additional information including the consensus analysis and the top FASTY hits.

The execution time of the query is extremely fast, as pre-computed FASTY hits are analyzed and no sequence based matching has to be performed. We take full advantage of the PostgreSQL DBMS and its capability of sub-queries and outer joins for performing the query, which combines a large number of tables of the database to obtain the necessary information. Figure 4.12 shows exemplarily such a mapping from a *Drosophila* gene to a *Xenopus* contig. Given the right indexes on the tables, this query is executed in about 52 ms by the DBMS, outperforming any sequence similarity search by far.

4.4. Summary

We have designed and implemented a clustering pipeline with a relational database management system as a central part of the system. The data flow is controlled by status information which is kept in the database. Several steps of the processing pipeline can be distributed on a compute cluster. *Vmatch* is used as the EST clustering tool and *CAP3* to generate contig sequences for each cluster. Contig sequences undergo a comprehensive sequence analysis, which is facilitated by the Genlight system. Functional classifications and Gene Ontologies are derived for each contig and stored in the database. Additionally, full length ORF containing contig sequences and full insert containing clones are identified. The system allows to query the database via a web interface and presents the results in a user-friendly way, including the assembly results, contig sequences, best hits to major model organisms, COG and KOG classifications, and GO terms. The system provides a unique functionality of comparative queries to rapidly identify potential homologues to other model organisms in the clustered EST data set.

Part II.

Applications to *Xenopus laevis*

XenDB: A *Xenopus laevis* Gene Index

5.1. Motivation

The African clawed frog *Xenopus laevis* is a major model organism which strongly contributed in two areas of vertebrate biology: early embryonic development and cell biology. *X. laevis* has led the way in establishing the mechanisms of early fate decisions, patterning of the basic body plan, and organogenesis. Contributions in cell biology and biochemistry include work on chromosome replication, chromatin and nuclear assembly, cell cycle components, and signaling pathways.

Xenopus is one of the primary resources for understanding early vertebrate development due to some unique advantages. A single female can produce hundreds of embryos each day, whose development can easily be studied from the time of fertilization because of external development and the large size, which again allows microsurgery and injections. All cells have an autonomous supply of nutrients which makes the embryos ideal for experimental approaches.

Extensive research has been carried out on signaling pathways in *Xenopus*. Especially gain or loss of function experiments have helped in resolution of these pathways. New components of pathways have been identified using expression cloning approaches. Gain

of function screens try to identify the functions of genes. In *Xenopus*, they are often based on injecting RNA made from cDNA libraries, obtained from tissues or developmental stages of particular interest. Optimally, these libraries should contain full-length cDNA clones to guarantee the initiation at the 5' end, which is not necessarily the case in poly-dT primed libraries. Therefore, there is a need for full-length cDNA libraries in expression-ready vectors.

In summer 2006, the Trans-NIH *Xenopus* Initiative agreed on recommendations¹ for future resources and goals which are necessary to improve *Xenopus* as a non-mammalian model system. One of the goals of highest priority is the generation of ESTs and full length cDNA collections:

“The ready availability of the sequenced clones through the IMAGE consortium [...] has provided new molecular markers, and full length clones for functional analyses. Full length cDNAs are particularly important to the Community because they facilitate the generation of Unigene sets which can be powerful for functional assays, one of the particular strengths of Xenopus, and as a collection, made for highly efficient expression cloning assays. [...] The priority for future EST and cDNA resources is to identify, in an expression-ready vector, a full length clone set for as large a fraction of the genes in the X. tropicalis genome as is feasible.”

X. laevis and *X. tropicalis* EST sequencing projects have produced 1.5 million ESTs, with *X. tropicalis* currently being seventh in the number of entries per organism in November 2006 (see Table 2.1). Many tissues have been sampled including embryonic stages and adult tissues. We have exploited the *X. laevis* sequence data to address the need for full length clones by EST clustering, extensive sequence analyses and full length clone prediction as described in the following sections.

5.2. Generation of a *Xenopus laevis* Gene Index

5.2.1. Sequence retrieval and Cleanup

350,468 Sequences were downloaded from GenBank release 138 and stored in the OR-DBMS PostgreSQL. The following divisions were included: Vertebrate Sequences (VRT,

¹<http://www.nih.gov/science/models/xenopus/>

5.2. Generation of a *Xenopus laevis* Gene Index

Tissue Type	Sequences	Developmental Stage	Sequences
N/A	120653	N/A	127219
whole embryo	105295	Adult	43436
Egg, oocyte libraries	59533	Gastrula	42394
Gastrula libraries	22490	Neurula	30094
Neurula Libraries	8164	Embryo, Stage 19-25	40346
Embryos, stages 19-26	6658	Embryo, Stage 31-32	20745
egg, subtr. by stage 13-17 animal cap	3806	Metamorphosis Stage 62	13894
embryo, animal cap	2907	Metamorphosis 50-53	10648
head, stage 30	2761	Tadpole	1950
pooled embryos (stage 10-14)	2672	Embryo	60

Table 5.1.: Ten most abundant tissue types (left) and developmental stages (right) in the *X. laevis* EST data set.

5,506 sequences), EST (344,747 sequences) and High Throughput cDNA (HTC, 215 sequences). 228,496 sequences were annotated as 5' ESTs and 116,122 as 3' ESTs. 245,415 different cDNA clones were represented in the data set, out of which 92,463 had both 5' and 3' sequences. Entries annotated as being genomic sequences were excluded from the analysis. To enhance the usability and search capabilities of the database, complete GenBank entries were incorporated. Annotations including library source, tissue type, cell type and developmental stage were extracted directly from GenBank entries. Table 5.1 shows a clear bias to early developmental stages from which the libraries were collected. Unfortunately, the sequences are not very well annotated in GenBank. 34% of the sequences do not have a tissue type assigned and 36% have no developmental stage information.

5.2.2. Repeat Masking

197,888 ESTs (57.4% of the EST sequences) had information about high quality start or end of sequencing reads. This information was used to trim sequences according to high quality regions to insure best sequence quality. Vector sequence was downloaded from GenBank and VectorDB and the sequence masked using *Vmatch*. ESTs were trimmed to eliminate vector sequence located at either the 5' or 3' end (6678 ESTs, 1.9% of total sequence set). In some cases, additional non vector sequence preceded or followed known vector sequence. If such non-vector sequence was less than 20 bases long, it was trimmed from the EST together with the vector sequence. ESTs that had vector sequences left after trimming were discarded completely. Repetitive elements were obtained from

Rebase and GenBank and masked using RepeatMasker. In addition, if hits against ribosomal RNA and mitochondrial sequences were found in the downloaded sequence set, the corresponding sequences were removed. The availability of complete mitochondrial genomic and ribosomal sequences makes the inclusion of these sequences unnecessary while masking was performed to minimize possible clustering errors arising from these common sequences. Sequences that had less than 100 consecutive bases left after cleanup were discarded completely (21,039 sequences, 6.0%). The resulting sequence set consisted of 317,242 sequences (90.5%) with an average length of 536 bases (see Table 5.2).

5.2.3. Clustering

The cleaned *X. laevis* EST sequence set was grouped into gene specific clusters using *Vmatch* as described in Section 3.4. Due to the efficiency of *Vmatch*, we were able to perform the clustering for a wide variety of parameters on the complete sequence set. This allowed us to study the effect of the parameter choice on the clustering (similar to data shown in Section 3.4.2).

For the current data set, we tried to select parameters which mimic the parameters that were probably used for generating the UniGene clusters. Unfortunately, the algorithm used for constructing the UniGene clusters is not sufficiently documented to allow complete reproduction. We selected parameters designed to produce a stringent clustering of the available sequences. For the described data set, sequences were clustered when a pairwise match of at least 150 nucleotides and 98% identity was found (*seedlength* = 33, *exdrop* = 3). The construction of the enhanced suffix array took 33 minutes on a SUN UltraSparc III (900 MHz) CPU. Clustering took another 17 minutes. This resulted in 25,971 clusters containing 276,365 sequences (87.11% of the input set) and 40,877 singletons (12.89%). The average cluster size was 10.6 (std. dev. 51.8) sequences. The distribution of cluster sizes is shown in Table 5.2. 22,834 clusters were composed of ESTs only, 61 clusters of mRNA sequences (VRT and HTC divisions) only and 3,076 clusters of both mRNAs and ESTs. Among the singletons are 4,262 sequences which contain less than 150 nt (after sequence cleanup described above) and would therefore be incapable of being joined in a cluster.

Total number of ESTs and cDNAs		350,468	Cluster sizes		#ESTs
Number of distinct clones	245,415		4,097 - 8,192		1
Number of good sequences	317,242		2,049 - 4,096		1
Average trimmed EST length (bp)	536		1,025 - 2,048		2
Number of 3' EST sequences	116,122		513 - 1,024		15
Number of 5' EST sequences	228,496		257 - 512		35
Clones with 5' and 3' sequences	92,463		129 - 256		116
Number of clusters	25,971		65 - 128		414
Number of singletons	40,877		33 - 64		973
Number of CAP3 contigs	31,353		17 - 32		1,755
Number of CAP3 singletons	4,801		9 - 16		2,974
Average CAP3 contig length (bp)	1,045		5 - 8		4,571
Max. cluster size (no. of ESTs)	6,332		3 - 4		6,444
Average cluster size (no. of ESTs)	10.6		2		8,670

Table 5.2.: Summary of *X. laevis* EST cleanup and clustering.

5.2.4. Assembly

In the next step, a consensus sequence was generated for each cluster using *CAP3*. The aim of this approach was to both refine the number of clusters and to improve the overall sequence quality. The 25,971 clusters produced 31,353 contig sequences (avg. length: 1,045bp, std. dev.: 729 bp) and 4,801 singlets (avg. length: 664 bp, std. dev.: 424 bp). The longest contig was 13,130 bp (DNA-dependent protein kinase catalytic subunit, accession: [GenBank:AB016434]), while the smallest contig was 154 bases long. Here, it became obvious that *CAP3* is a genome assembly program not designed to assemble EST clusters containing potential splice variants: *CAP3* assembly subsequently split a fraction of the clusters into separate contigs and singletons. On average, a cluster was split into 1.2 (std. dev 3.0) contigs and 1.8 (std. dev 11.3) singlets by *CAP3*. As illustrated in Table 5.2, the average length of the sequences increased from 536 bp (average for input ESTs) to 1,045 bp (average for *CAP3* contig sequences) which was lower than the average length for previously characterized *Xenopus* full length sequences (sequences selected as full length by NIH's *Xenopus Gene Collection* (XGC) initiative had an average length of 2,115 bp). There are many genes whose transcript is significant longer than two times the current state of the art sequencing run of ≈ 1000 bp. This means that 5' and 3' sequences derived from a >2 kb transcript are unable to be joined without sequence from incomplete cDNA clones which provide a source of nested deletions. Sequences from both ends can

be linked by annotation, and this has been done by a variety of clustering approaches including NCBI UniGene which uses a double linkage rule. Non-overlapping 5' and 3' ESTs are assigned to the same cluster if clone IDs are found that link at least two 5' ends from one cluster with at least two 3' ends from another cluster and the two clusters are merged. We have examined the effect of double linkage joining using the clone annotation. In this analysis, 17,588 clusters were stable and the total number of clusters was reduced from 25,971 to 21,249. Most of the joined clusters (3,122) were created from two clusters while three clusters were combined 456 times. While the number of clusters is decreased by this joining, our overall analysis is not affected. Potential full length clones selected as part of the P5P group (see Section 4.1.6) are also unaffected by annotation linkage.

5.3. Sequence Analysis of *Xenopus laevis* Gene Index

We have performed a variety of sequence comparisons at the protein level including translation analysis. The sequences of cluster contigs and all singletons were subject to extensive BLASTX and FASTY homology searches vs. the non-redundant protein database (NR) from NCBI and the proteomes of five major model organisms using the high throughput analysis pipeline of the *Genlight* system [14]. Proteome sets for *H. sapiens*, *M. musculus* and *R. norvegicus* were obtained from the International Protein Index (IPI). *C. elegans* and *D. melanogaster* protein sequences were retrieved from the UniProt database. Additionally, all available protein sequences for *X. laevis* and *X. tropicalis* were extracted from GenBank. In addition to these databases, we have included BLASTX searches in the COG and KOG database and have used the results to functionally classify the *Xenopus* sequences. All sequences resulting from the clustering and assembly processes were compared to these protein sets using BLASTX with an E-value cutoff of $1.0e^{-6}$. ESTs are often of low sequence quality, and sequencing errors can still exist in the assembled contig sequences. Therefore, all analyses against the protein databases were also done using FASTY (E-value cutoff: $1.0e^{-6}$) a version of FASTA that compares a DNA sequence to a protein sequence database, translates the DNA sequence in three forward (or reverse) frames and allows (in contrast to BLASTX) for frame shifts, maximizing the length of the resulting alignments.

5.3.1. Identification of Chimeric Sequences

A significant issue in EST clustering methods is the presence of chimeric sequence which inappropriately joins unrelated genes into a single cluster. While the number of chimeric sequences is estimated at less than 1% [1, 62], their presence has disproportionate effects on the clustering outcome. To identify potential chimeric sequences, we analyzed the FASTY hits in the protein NR database and applied the following simple procedure: Matches of at least 100 bp in length were mapped back to the contig sequences to identify the regions that are covered by a match. If two matches overlap, the region will be extended accordingly. If after the mapping two clearly separated regions remain, the contig is flagged as potential chimera. Figure 5.1 shows an example of a contig identified as chimera.

Examination of the identified chimeric sequences reveals three major classes. In the first, two distinct FASTY hits can be identified which do not overlap and are in opposite orientation. In the second class, the second identified FASTY hit matches retroviral or transposable element related sequences. This suggests the possibility that these may reflect real transcripts in which a mobile element has been inserted into the genome. A close evaluation of such sequences may provide some insights into the evolutionary history of various populations of *Xenopus*. The final class of potential chimeric sequences identified contains short predicted or hypothetical proteins. This class may in fact not be chimeric at all but may reflect errors in protein coding prediction methods.

The described procedure identified 113 potential chimeric contigs (0.3% of the 33,034 sequences with matches against the protein NR database), which are flagged in the database as such. We do not eliminate these potential chimeras, as they do not significantly affect the results of the sequence analyses done later on, which are mainly based on the best hit only. In fact, the analysis underestimates the number of full length sequences, as some chimeras cover two full length protein matches. A complete identification of chimeric sequences is practically impossible without a comparison to the underlying genome sequence. And even then, polycistronic transcripts which may exist cannot be separated from chimeras perfectly [85].

5.3.2. Gene Ontology prediction and Functional Classification

The Gene Ontology (GO) project is an ongoing international collaborative effort to generate consistent descriptions of gene products using a set of three controlled vocabularies or

5. XenDB: A *Xenopus laevis* Gene Index

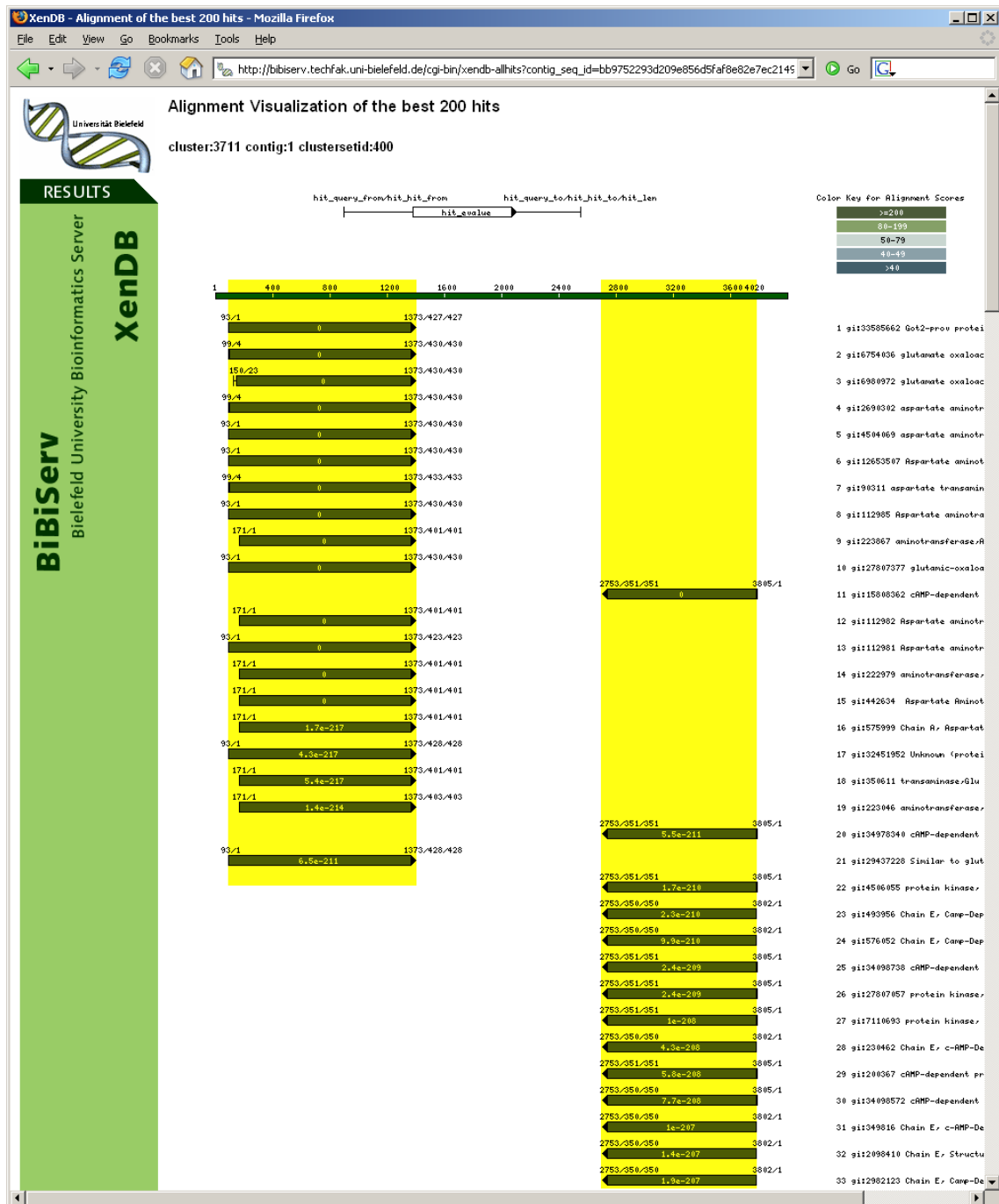


Figure 5.1.: Example of a contig identified as potential chimera. The matches to two distinct proteins do not overlap.

ontologies: biological processes, cellular components, and molecular functions. The GO vocabulary allows consistent searching of databases using uniform queries. The availability of such vocabularies can be critical to the interpretation of high throughput approaches such as microarrays. Based on FASTY homologies with both mouse and human sequence, we have mapped GO annotations to the *Xenopus* sequences. Of the 30,683 contigs with matches to mouse (29,971) or human IPI sequences (29,963), 19,721 contigs have been assigned putative GO annotations. Among the 10,500 potential full length ORF containing IMAGE clones (see Section 5.4.2), 6,886 have been assigned GO annotations. The non-redundant *X. laevis* data set was then classified based on their homology to known proteins from the KOG database (BLASTX $1.0e^{-5}$ E-value cutoff, best hit selection). 17,624 sequences (67.3 %) had a hit against the KOG database and could be assigned a functional category.

5.4. Clone Selection

Acknowledging the special interest in full length clones, we focused during the analysis of the contig sequences to full ORF containing contigs and full length clones [136]. The identification of complete ORFs allows further analyses on not only the corresponding protein sequences, but also to UTR sequences flanking the coding regions. The value of having direct access to full length clones has been described in Section 5.1.

5.4.1. Identification of full length contigs

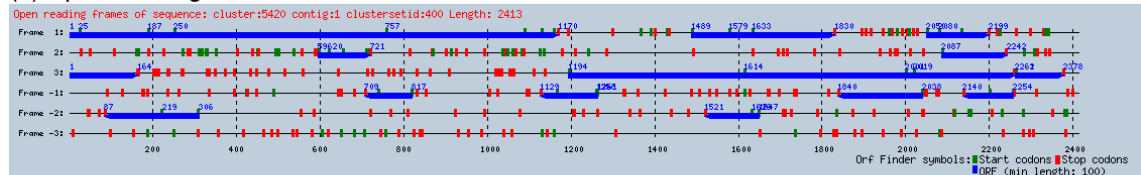
Table 5.3 (top) shows the number of identified *full length ORF containing contigs* using BLASTX (see Section 4.1.6 and Figure 4.2). 3,942 contigs were *Class 1* hits in the non-redundant protein database. As the stringency of the full length definition was relaxed, the number of contigs characterized as full length increases to 5,050 (*Class 2*), 7,792 (*Class 3*) and 12,389 (*Class 4*) contigs, respectively. As EST sequences have many sequencing errors, and even the assembly of clusters cannot correct all of these, FASTY comparisons were done for the same data set (Table 5.3, bottom). This way, the length of the resulting alignments could be maximized. A comparison of BLASTX and FASTY results shows the effect of frame shift corrections obtained by FASTY. The number of contigs having *Class 1* hits could be increased to 5,139 while the less stringent categories increased similarly by an average of 20%. The effect of frameshift correction can clearly be seen in Figure

5. XenDB: A *Xenopus laevis* Gene Index

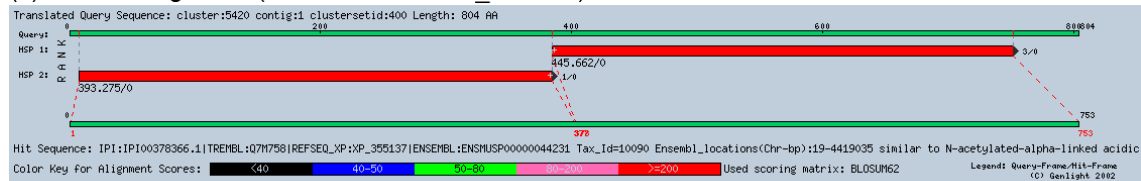
5.2. Table 5.4 shows the average lengths of contigs for each of the four categories. Here, the average length of the contig is 2,210 bp for *Class 1* contigs having FASTY matches against *X. laevis*, corresponding very well to already known *Xenopus* proteins. Overall, the average length of sequence decreases with lower quality categories as expected, especially for *Class 4*, where the alignment can miss 20 amino acids on both ends of the matching protein. The only exceptions are *D. melanogaster* and *C. elegans*, where the average length increases for *Class 4* sequences.

X. laevis cluster:5420 contig:1

(a) Open Reading Frames



(b) BLASTX alignment (hit: *M. musculus* XP_355137)



(c) FASTY alignment (hit: *M. musculus* XP_355137)



Figure 5.2.: Comparison of a BLASTX alignment with corresponding full length FASTY alignment, as generated by the Genlight system. Blue boxes in (a) indicate open reading frames, green boxes start and red boxes stop codons, respectively. The assembled contig sequence has a frameshift at position 1150 from frame 1 to 3, generating two distinct HSPs in the BLASTX alignment (b). FASTY clearly corrects this frameshift and generates a full length alignment (c).

Comparing the numbers of full length sequences in Table 5.3, the matches in human, mouse, rat and *X. laevis* are in general agreement (2,619 full length sequences for *Class 1* on average). What is striking is the deviation of both the number of full length contigs as well as the average length of contigs (Table 5.4) having matches against *D. melanogaster*

BLASTX								
Class	Protein NR	Human	Mouse	Rat	Fruitfly	<i>C. elegans</i>	<i>X. laevis</i>	<i>X. tropicalis</i>
1	3942	1760	1765	1455	219	140	2918	495
2	5050	2067	2076	1736	311	233	3104	541
3	7792	2647	2919	2592	392	283	3898	590
4	12389	5587	5841	3078	2071	1856	5024	1033
P5P	15870	13942	14179	13113	8425	8117	9227	4334

FASTY								
Class	Protein NR	Human	Mouse	Rat	Fruitfly	<i>C. elegans</i>	<i>X. laevis</i>	<i>X. tropicalis</i>
1	5139	2347	2337	1930	268	190	3862	660
2	6243	2692	2671	2248	383	296	4119	721
3	9576	3528	3774	3374	473	357	4967	796
4	14094	6467	6701	6341	2249	1918	5701	1241
P5P	15651	13578	13954	13085	8108	7746	9055	4159

Table 5.3.: Number of *X. laevis* contigs with full length BLASTX (top) and FASTY (bottom) hits in the non-redundant protein database (NCBI), five model organisms, and available *X. laevis* and *X. tropicalis* proteins. Lower quality categories include sequences from higher, more stringent categories.

and *C. elegans*: only 268 and 190 full length sequences with average lengths of 1,659 and 1,575 bp for *Drosophila* and *C. elegans* in *Class 1*, respectively. Only within the *Class 4* category there are 2,249 and 1,918 contigs with average lengths of 1,611 bp and 1,563 bp, respectively. A possible explanation for this difference is the divergence of the vertebrate species from these invertebrate model systems.

5.4.2. Identification of full length clones

Best FASTY hits were extracted for contigs from all four full length categories as well as the P5P categories as described above. For contigs matching these categories, the most 5' EST contributing to the *CAP3* contig sequence was selected. In addition, the selected clone had to span the amino-terminal end of the FASTY protein match. Finally, to ensure the ready availability of the clones and therefore the utility of the analysis, the selected clone had to be available through the IMAGE consortium. (See Figure 4.2 for an illustration of 5' clone selection.) The P5P criteria selected 15,651 potential full length insert containing clones out of which 10,500 are distinct IMAGE clones, which represents an additional 1,557 sequences compared to *Class 4*. Two examples of such predicted protein coding sequences are presented in Figure 5.3. We have mapped these clones to 7,782

5. XenDB: A *Xenopus laevis* Gene Index

BLASTX								
Class	Protein NR	Human	Mouse	Rat	Fruitfly	<i>C. elegans</i>	<i>X. laevis</i>	<i>X. tropicalis</i>
1	1984	1835	1805	1788	1620	1541	2171	1743
2	1831	1806	1776	1775	1541	1391	2120	1697
3	1630	1813	1775	1834	1560	1429	1981	1693
4	1393	1680	1675	496	1638	1640	1879	1660

FASTY								
Class	Protein NR	Human	Mouse	Rat	Fruitfly	<i>C. elegans</i>	<i>X. laevis</i>	<i>X. tropicalis</i>
1	2007	1888	1859	1843	1659	1575	2210	1807
2	1837	1856	1821	1819	1563	1440	2152	1774
3	1553	1790	1772	1804	1569	1441	2019	1768
4	1329	1683	1673	1664	1611	1563	1910	1703

Table 5.4.: Average length of *X. laevis* contigs for different BLASTX (top) and FASTY (bottom) full length contig categories.

distinct clusters. To assess the quality of the full length (FL) prediction method, we compared our set to the IMAGE clone set selected by the Xenopus Gene Collection (XGC²) for full length sequencing. At that time, the XGC had selected 10,482 IMAGE clones for sequencing. Our analysis selected 3,152 IMAGE clones that were identical to clones selected by the XGC. Of the remaining 7,348 clones from our set, 4,866 selected IMAGE clones were found in an identical cluster as 4,465 XGC selected clones (note that some of these clones are in the same cluster). In addition, 1,154 XGC clones did not have sequence available to be included in our analysis. The remaining 1,711 IMAGE clones selected for sequencing by XGC are not found in our predicted set while 2,482 clones were unique to our set. In an effort to examine why the 1,711 sequences selected for sequencing were not identified as full length, we compared the $start_q$ and $start_s$ values as described above. Using the P5P prediction criteria described above, we identify 107 XGC selected IMAGE clones that we predict are not full length but have an alternative clone which we predict is full length. Though final confirmation of the results requires additional sequencing, our method appears to be successful at identifying full length sequences and distinguishing non-full length sequences identified by an independent method. The FL clones are labeled in the XenDB web interface, allowing a rapid identification of potential FL clones for a gene of interest.

Due to the large number of sequences, we are unable to examine each sequence individually. Since the analysis depends on the overall degree of conservation among the

²<http://xgc.nci.nih.gov/>

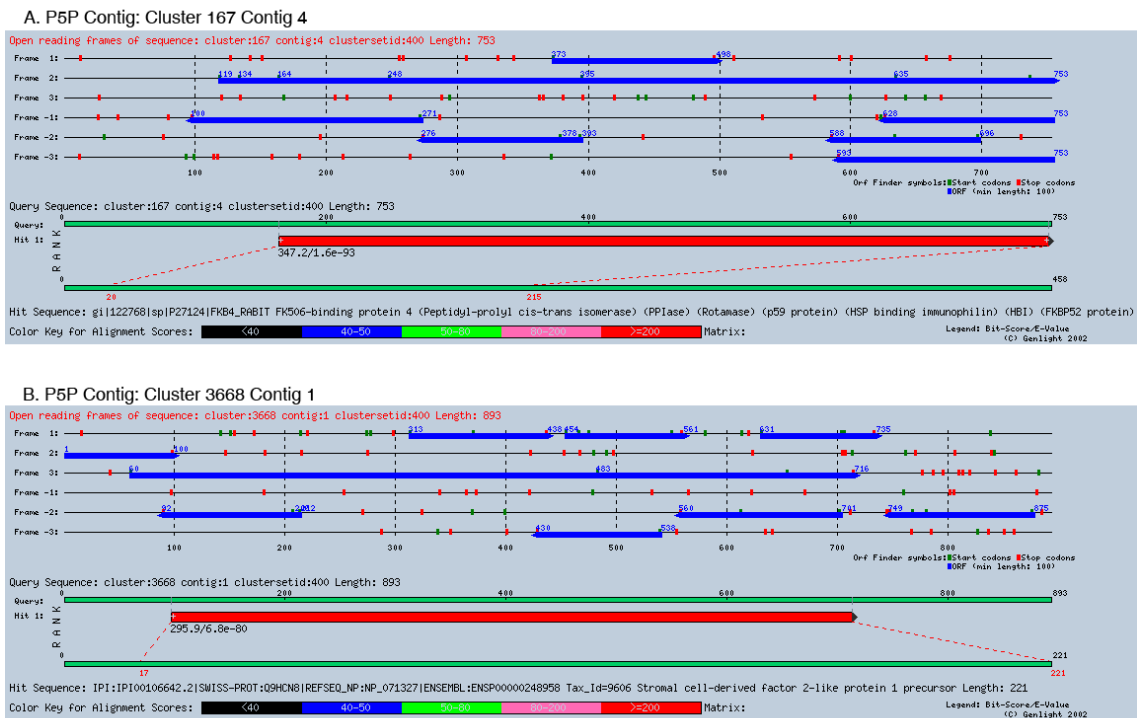


Figure 5.3.: Two examples of contigs derived from clones predicted to have a full length insert (P5P). The start positions in the hit suggest that the unmatched amino-terminal protein sequence is not well conserved between *X. laevis* and the matched organisms, here rabbit (top) and human (bottom), but the open reading frames (blue boxes) indicate that the clones the sequences were derived from do actually contain a full length insert. (Screenshots of the results were generated by the Genlight system.)

sequences, such an approach will not be as successful on weakly conserved genes. In general, it seems likely that decreasing e-values correspond to higher quality predictions. On a global basis, the results need to be carefully considered, as an independent assessment of the distribution of conservation among the ensemble of sequences is not available.

5.5. Utility

5.5.1. User Interface

The results of the analyses described above have been incorporated into an SQL database amenable to complex queries. The database can be accessed through the user friendly

web based interface of XenDB³ [135]. XenDB allows individual and batch queries using *Xenopus* accession, GI, and XenDB, UniGene and TIGR cluster IDs. In addition, the user can query the *Xenopus* sequence hits using any protein accession/GI number both singly and in batch mode. This allows a rapid identification of *Xenopus* contigs and their corresponding clones with hits to given protein sequences. The output of various queries displays the matching *Xenopus* cluster(s) and links to a web page as presented in Section 4.3.2. For each cluster, links to the best hit for a number of model organisms are provided as well as links to the assembly result, consensus sequence generated by *CAP3*, and visual alignments of all FASTY results. GenBank accession numbers for each EST in the cluster and whether the corresponding clone has been identified as full length are provided. Additionally, for each contig the COG and KOG classification, as well as the GO terms are available.

The analysis and database system provides a very powerful tool which will enable the *Xenopus* community to take advantage of a number of technical and experimental advances. We have selected a couple of examples to illustrate possible types of queries. In considering the results, it is important to bear in mind that these examples can be combined to further refine the sequence set. In the first example, we sought to identify all the genes of a known type or class. In the second example, we wished to identify the set of *Xenopus* sequences which best matched a set of genes from another species identified using the CGAP database administered by the National Cancer Institute (NCI) [92, 143]. Another example demonstrates the ability of the system to translate results identified by microarray technologies, or other related high throughput technologies, to identify likely *Xenopus* homologues.

5.5.2. Homeobox Gene Identification

Homeobox containing proteins are a very important group of transcriptional regulators that play key roles in developmental processes. They can be divided into a 'complex' and a 'dispersed' super class representing the homeotic genes and the large number of homeodomain containing proteins dispersed (and diverged) within the genome [49]. The homeotic (Hox) genes play key roles in the anterior-posterior patterning of both vertebrate and invertebrate embryos and in *Xenopus* are often used as markers of anterior-posterior

³<http://bibiserv.techfak.uni-bielefeld.de/xendb>

development [37, 168, 70]. The vertebrate homeotic genes are organized into four clusters arranged in the same order in which they are expressed in the anterior-posterior axis [49]. Of the 39 vertebrate Hox genes, we have identified 28 homologues in *X. laevis*, while 19 are present in the protein database (Table 5.5). For those sequences not identified, we sought to determine whether they had been identified in the genome of *X. tropicalis*. To do so, we used TBLASTX, provided as a tool on the *X. tropicalis* website⁴ to search for the missing sequences. Strong matches were identified for all of the remaining Hox genes except HoxD12. Using the BLASTN tool on the genome site, we confirmed that the gene order was conserved within each scaffold. Interestingly, we were unable to identify HoxD12 within the predicted region though both HoxD11 and HoxD13 were recognized.

5.5.3. Homologue Identification from the Cancer Genome Anatomy Project

A second example takes advantage of the CGAP database⁵ administered by the National Cancer Institute (NCI). This database and resource incorporates a large number of interconnected modules aimed at gene expression in cancer. Among the modules are a Serial Analysis of Gene Expression (SAGE) database [22]. The SAGE approach counts polyadenylated transcripts by sequencing a short 14 bp tag at the genes 3' end and is a quantitative method to examine gene expression [159]. Another module is the Digital Gene Expression Displayer (DGED) which distinguishes statistical differences in gene expression between two pools of libraries [91]. Each method generates tables of genes based on a wide variety of selection criteria. As would be expected, the source for the vast majority of the available data comes from either human or mouse thus demanding a tool to cross match the results in *Xenopus*.

For this particular example, we selected a tissue based query (DGED) derived from SAGE data in which we sought a set of genes that might include potential markers for glial or astrocyte fates. For this query, we selected all brain, cortex, cerebellum and spinal cord libraries excluding any libraries derived from cell lines. This yielded 58 potential libraries. From this we selected any library labeled as a glioblastoma for pool A and libraries labeled astrocytoma for pool B while excluding the remaining libraries (which included medulloblastomas, ependymomas, etc.). We did not distinguish between cancer grades. This limited the total number of libraries to six glioblastoma and nine astrocytoma libraries

⁴<http://genome.jgi-psf.org/Xentr4/>

⁵<http://cgap.nci.nih.gov/>

5. XenDB: A *Xenopus laevis* Gene Index

IPI Accession	Gene	Xenopus cluster/contig	FASTY e-value	BLASTX e-value	FL Clone	Protein Accession
IPI00027694	HOX-A1	cluster:4123 contig:1	$4e^{-85}$	$1.99e^{-99}$	5536792	AAH44984
IPI00012049	HOX-A2	cluster:7495 contig:1	$4.1e^{-130}$	$7.64e^{-145}$	3556495	AAG30508
IPI00012050	HOX-A3	cluster:10945 contig:1	$6.5e^{-91}$	$2.89e^{-111}$	4683538	AAH41731
IPI00020926	HOX-A4	fgenesh.C_1023000005				
IPI00302291	HOX-A5	cluster:25739 contig:1	$6.9e^{-44}$	$1.27e^{-38}$		
IPI00010742	HOX-A6	fgenesh.C_1023000003				
IPI00010743	HOX-A7	cluster:3210 contig:1	$5.8e^{-40}$	$1.17e^{-64}$	XL071e19	AAA49753
IPI00010744	HOX-A9	vm_singlet:264323	$1.2e^{-33}$	$3.48e^{-29}$		
IPI00010731	HOX-A10	fgenesh.C_1487000003				
IPI00010754	HOX-A11	cluster:6499 contig:1	$7.2e^{-42}$	Was C11	XL088b06	
IPI00305850	HOX-A13	vm_singlet:174355	$3.8e^{-57}$	$1.22e^{-97}$		
IPI00294724	HOX-B1	fgenesh.C_2225000001				
IPI00027261	HOX-B2	fgenesh.C_2225000002				
IPI00027259	HOX-B3	fgenesh.C_2225000003				
IPI00014540	HOX-B4	cluster:22503 contig:1	$1.2e^{-27}$			
IPI00012514	HOX-B5	vm_singlet:57425	$8.5e^{-35}$	$3.92e^{-59}$		
IPI00015075	HOX-B6	cluster:2339 contig:1	$6.2e^{-42}$	$2.52e^{-72}$	XL098k02	
IPI00172584	HOX-B7	cluster:1985 singlet:1	$2.6e^{-65}$	$8.16e^{-77}$	4201615	P04476
IPI00014536	HOX-B8	cluster:16406 contig:1	$2.8e^{-28}$	$9.9e^{-43}$		
IPI00014539	HOX-B9	cluster:8543 contig:1	$4e^{-30}$	$1.05e^{-50}$	XL069k06	P31272
IPI00030703	HOX-B10	cluster:24736 contig:1	$5.6e^{-48}$	$6.95e^{-74}$		
IPI00295561	HOX-C4	fgenesh.C_202000010				
IPI00022893	HOX-C5	vm_singlet:33065	$1.5e^{-41}$	$6.14e^{-32}$		
IPI00015921	HOX-C6	cluster:9871 singlet:1	$4.2e^{-93}$	$3.16e^{-109}$	4202432	P02832
IPI00010756	HOX-C8	cluster:11257 contig:1	$5.2e^{-95}$	$9.74e^{-118}$	XL045l21	AAB71818
IPI00010757	HOX-C9	fgenesh.C_202000007				
IPI00020947	HOX-C10	cluster:3243 contig:1	$1.3e^{-51}$	$1.63e^{-127}$	4970594	AAO25534
IPI00011610	HOX-C11	fgenesh.C_202000005				
IPI00010758	HOX-C12	vm_singlet:240042	$2.4e^{-46}$	$2.75e^{-22}$		
IPI00010759	HOX-C13	cluster:21388 contig:1	$2e^{-80}$	$5.86e^{-89}$	XL064e01	
IPI00001551	HOX-D1	cluster:9419 contig:1	$2.5e^{-50}$	$1.68e^{-65}$	3475513	AAA49745
IPI00215882	HOX-D3	cluster:4099 contig:1	$2.6e^{-114}$	$3.48e^{-121}$	4684054	
IPI00012390	HOX-D4	cluster:21685 contig:1	$7.1e^{-67}$	$7.99e^{-83}$	5571854	AAQ95789
IPI00008481	HOX-D8	cluster:11793 contig:1	$5.8e^{-62}$	$2.08e^{-74}$	5543040	AAH60408
IPI00292734	HOX-D9	cluster:13847 contig:1	$6.5e^{-38}$	$5.28e^{-55}$	XL045k22	CAC44973
IPI00292735	HOX-D10	cluster:6503 contig:1	$3.8e^{-135}$	$3.97e^{-143}$	4032032	CAC44974
IPI00305856	HOX-D11	fgenesh.C_1333000003				
IPI00018803	HOX-D12	missing				
IPI00018806	HOX-D13	cluster:13386 contig:1	$1.7e^{-93}$	$2.17e^{-112}$	3399571	AAO25535

Table 5.5.: Homeobox genes in *X. laevis*: for each HOX gene the corresponding cluster and contig is shown, as well as the most 5' clone in the assembly and the protein accession number, if available. When *X. laevis* genes were not identified, an identifier corresponding *X. tropicalis* sequence is provided.

containing 487,197 and 863,610 SAGE tags each, respectively. Submission of the query resulted in the identification of 395 tags with a 2x expression factor and a 0.05 significance factor (default CGAP query values). These 395 tags represented 308 different sequences (180 were >2 fold higher in glioblastoma and 128 were >2 fold higher in astrocytoma) which corresponded to 278 proteins in the public database (115 glioblastoma, 163 astrocytoma) and were matched using the batch GenBank accession module available online in XenDB to 100 and 142 *Xenopus* sequences, respectively. Among the genes identified are *vimentin* (15x, $P = 0.01$) and *sox10* (7.6x, $P = 0.03$), genes previously established as markers of glial and oligodendrocyte fate respectively as well as genes downstream of the Notch signaling pathway, known to be important for glia formation. Thus the system developed and presented here allows 'in silico' based tools established for the study and analysis of other organisms, particularly human and mouse, to be easily and rapidly applied to the *Xenopus* model system.

5.5.4. Homologues of *Drosophila* Eye Development Genes

In the final example, we take advantage of the database to perform a comparative analysis of microarray expression data. In many instances, the outcome of an array type experiment is a variety of tables listing regulated genes and the associated expression changes. Currently, there are few published *Xenopus* array studies available [6, 38, 116, 153, 82, 34, 10, 121, 137] while there exist extensive databases of expression for a variety of model organisms. The NCBI maintains a common database, the Gene Expression Omnibus [43] which contains data from over 15,000 samples including 337 Human, 92 mouse and 12 *Drosophila* experiments (average 25 samples/experiment).

We selected a recent paper which examined gene expression changes induced by ectopic expression of the eyeless gene (*ey/Pax-6*) in *Drosophila* imaginal disks [106]. The development of the eye is evolutionarily conserved among both vertebrates and invertebrates [52, 50]. Many important insights into eye development have come from studies in *Drosophila* which has defined a genetic cascade of evolutionarily conserved regulatory factors [48]. One such factor is Pax-6/eyeless which is capable of inducing ectopic eyes on both flies and vertebrates. In the study ([106]), 371 eye-induced genes are detected using two different oligonucleotide based array platforms (Affymetrix and Hoffmann-LaRoche) and 73 are discussed in detail within the text ([106], Tables 1 and 2).

To identify likely homologues of these genes in *Xenopus*, GenBank accession numbers

5. XenDB: A *Xenopus laevis* Gene Index

#	Cluster	Ctg	FL clone	Protein Accession	Description	E-value	Fly Rank	All Rank
1	21344	1	YES	AAA19592	Lola protein short isoform	$3.9e^{-10}$	42	508
2	21344	1	YES	AAA19593	Lola protein long isoform	$7.0e^{-10}$	67	553
3	22774		NO	AAA21879	atonal protein	$1.2e^{-17}$	3	21
4	5646	1	YES	AAA28528	fasciclin II	$4.3e^{-28}$	5	61
5	3838	1	NO	AAA28723	eyes absent	$1.8e^{-118}$	1	49
6	10868	1	YES	AAB61239	bunched gene product	$2.0e^{-21}$	6	39
7	BJ063320		NO	AAC46506	Dachshund	$1.6e^{-16}$	8	44
8	10334	1	YES	AAC47196	Lozenge	$2.9e^{-56}$	4	77
9	7019	1	YES	AAD38602	scratch	$4.4e^{-35}$	15	83
10	4763	2	YES	AAD38642	BcDNA.GH11415	$2.9e^{-146}$	3	15
11	16925	1	YES	AAD38646	BcDNA.GH11973	$8.2e^{-14}$	1	14
12	18882	1	NO	AAD52845	Pebble	$7.4e^{-62}$	2	14
13	3666	1	YES	AAF24476	Sticky ch1	$1.7e^{-11}$	3	56
14	7799	2	YES	AAF48990	CG12238-PA	$1.7e^{-22}$	3	64
15	19264	1	YES	AAF55415	CG5407-PA	$6.3e^{-198}$	3	10
16	5529	1	YES	AAF57639	CG15093-PA	$2.2e^{-45}$	1	24
17	CD327522		NO	AAK06753	roughoid/rhomboid-3	$1.1e^{-29}$	8	26
18	22774		NO	AAK14073	DNA-binding transcription factor	$8.6e^{-10}$	11	158
19	1415	445	YES	AAL86442	slamdance	$5.5e^{-70}$	26	194
20	BU911996		NO	AAN74533	transcription factor fruitless	$7.9e^{-10}$	28	459
21	CD329851		NO	BAA78210	white protein	$2.2e^{-36}$	17	54
22	21321	1	YES	CAA33450	glass protein	$2.2e^{-45}$	21	1739
23	2426	1	YES	CAA38746	neurotactin	$2.4e^{-24}$	103	706
24	9209	1	YES	CAA52934	Drosophila cyclin E type I	$2.5e^{-56}$	2	25
25	18485	1	NO	CAA76941	UNC-13 protein	$2.7e^{-165}$	1	14
26	17438	1	NO	NP.523928	CG7525-PA	$8.7e^{-24}$	101	1508
27	570	1	YES	NP.524354	CG4236-PA	0.0	1	17
28	BI349728		NO	NP.573095	CG9170-PA	$2.7e^{-17}$	1	7
29	1761	1	NO	NP.609033	CG9536-PA	$1.2e^{-21}$	1	6
30	12008	1	YES	NP.609545	CG14946-PA	$9.1e^{-25}$	8	63
31	440	2	YES	NP.610108	CG8663-PA	$5.7e^{-17}$	5	90
32	9019	1	YES	NP.611013	CG11798-PA	$1.4e^{-7}$	156	2411
33	10147	2	YES	NP.648269	CG5653-PA	$1.9e^{-16}$	5	48
34	3752	1	YES	NP.649919	CG9427-PA	$4.1e^{-13}$	1	32
35	20081	1	YES	NP.725617	CG5522-PF	$7.1e^{-49}$	1	18
36	2636	2	YES	NP.729075	CG10625-	$1.7e^{-28}$	16	1185
37	8386		YES	O18381	Eyeless protein	$3.9e^{-70}$	7	75
38	11614	1	YES	P00528	Tyrosine-protein kinase Src64B	$4.3e^{-152}$	3	150
39	4073	1	NO	P10181	Homeobox protein rough	$3.1e^{-14}$	13	165
40	919		NO	P20483	String protein (Cdc25-like protein)	$3.3e^{-40}$	3	43
41	1777	1	YES	P36872	Twins protein (PR55)	0.0	2	41
42	9517	1	YES	P48554	Ras-related protein Rac2	$1e^{-109}$	1	22
43	7661	1	YES	Q01070	E(spl) mgamma	$5.5e^{-19}$	5	52
44	7661	1	YES	Q01071	E(spl) mdelta	$1.1e^{-15}$	7	63
45	4146	2	YES	Q23989	Villin-like protein quail	$6.3e^{-23}$	9	138
46	10061	1	YES	Q27324	Derailed protein	$1.2e^{-45}$	23	400
47	14903	1	YES	Q27350	Sine oculis protein	$3.9e^{-87}$	1	20

Sequences without significant homology

#	Accession	Description	#	Accession	Description
48	O77459	transcription factor Ken	60	NP.651346	CG11849-PA
49	AAF46666	CG10527-PA	61	Q23997	Chitinase-like protein DS47 precursor
50	NP.728586	CG9134-PA	62	AAD09748	Gasp precursor
51	NP.609450	CG17124-PA	63	AAF63503	SP2523
52	CG140595	Zea mays genomic	64	AAF47412	CG13897-PA
53	NP.570064	CG10803-PA	65	AAL27368	zinc finger C2H2 protein sequoia
54	NP.650785	CG5835-PA	66	NP.730444	CG32209: CG32209-PB
55	AAF51847	CG11370-PA	67	NP.723827	CG18507-PA
56	AAG46059	SKELETOR	68	NP.611728	CG13532-PA
57	AAN61340	BcDNA:GH10711	69	NP.651343	CG13651-PA
58	NP.729183	CG10121-PA	70	NP.995997	CG12605-PA
59	AAO39528	RE22242p	71	NP.610067	CG9335-PA

Table 5.6.: *Xenopus* matches to Pax6/ey Regulated Genes identified by [106]

were obtained from the NCBI Gene Expression Omnibus (accession: GSE271) and used to query the XenDB database to identify 47 potential homologues of the *Drosophila* Pax6/ey regulated genes and included 32 predicted full length sequences (Table 5.6). As these sequences are available from commercial sources, they can be readily obtained and tested using the various experimental approaches available to *Xenopus* such as gain of function studies by microinjection.

5.5.5. Application of the IsoSVM classifier to *X. laevis* EST data

The *IsoSVM* tool introduces an automated approach to identifying isoforms on the protein level using a support vector machine (SVM) classifier. Based on three specific features used as input of the SVM classifier, it is possible to automatically identify isoforms with an accuracy of more than 97%. As an example application *IsoSVM* has been used to estimate that a subset of the XenDB EST clusters consists of approximately 81% cases where sequences are each other's paralogs and 19% cases where sequences are each other's isoforms [140].

To assess whether the splitting of clusters by *CAP3* into several contigs was caused by grouping isoforms into the same cluster, or whether the splitting was due to paralogs, we extracted 722 clusters that have multiple contigs (2,243 contigs total), and for which each contig has a full length protein match in the protein NR database. *X. laevis*, as an allotetraploid species, has undergone a genome wide duplication. Therefore, many genes are represented by two paralogs. Isoforms of *X. laevis* proteins have not been studied in any systematic way before.

Most of the 722 clusters consist of only two contigs and only a fraction features three or more contigs. Treating each contig consensus as a sequence, 5,459 sequence pairs were compared by *IsoSVM* within clusters; 986 of these samples (19.3%) were classified as isoforms and 4,125 as paralogs (80.7%). 348 samples were left out, representing contigs with almost no overlap, i.e. sequence pairs of low (<1%) similarity. These results were also used as a further check to assess the accuracy of *IsoSVM*. 290 randomly chosen samples were reviewed manually: an accuracy of 97.93% and a precision of 99.23% was found.

5.6. Summary

We have clustered more than 350,000 *X. laevis* ESTs using the previously described pipeline into 25,971 clusters and 40,877 singlets. The subsequent CAP3 assembly split some clusters into several contigs, such that 31,353 contig sequences, 4,801 CAP3 singlets and the 40,877 singlets from the first clustering step were subject to a comprehensive sequence analysis. 19,721 sequences could be assigned a GO annotation, 17,624 were functionally classified according to KOG categories. The identification of full length ORF containing contigs revealed 5,139 contigs which had a FASTY match against a protein in the NR database including start and stop codons. The FASTY analysis identified ~20% more full length ORF containing contigs than a similar BLASTX analysis. 10,500 IMAGE clones could be identified as containing a full insert and therefore provide easy access to complete cDNA clones of the corresponding genes.

In a variety of examples we have demonstrated the utility of XenDB, the resulting database, available through a user-friendly web interface at <http://bibiserv.techfak.uni-bielefeld.de/xendb/>. It allows as a unique functionality a rapid mapping of sequences from other model organisms to the *X. laevis* contig sequences and clones.

Computational Identification of miRNAs in *X. laevis* EST clusters

In Chapter 5 we have described the construction of a *X. laevis* gene index and results of a thorough sequence analysis which tried to identify homologues to most of the resulting contigs. However, a class of sequences remains: those without significant hits to known proteins. In our analysis we have used an E-value cutoff of $10e^{-6}$, which is of course necessarily arbitrary. Based on this value, we remain with 43,753 sequences that neither have a BLASTX nor a FASTY hit to a known model organism sequence. The lack of similarity could be due to significant divergence of the sequence, the lack of an appropriate homologue in the public dataset, sequencing errors inherent in the EST data or due to the presence of non-coding, presumably regulatory sequences, in the EST clone set.

These unmatched sequences mirror the situation in the UniGene set for both mouse and human with more than 4.3 and 7 million EST sequences in 65,000 and 84,000 clusters respectively while fewer than 30,000 coding sequences have been recognized [115, 69, 144]. The source of these discrepancies are currently unclear, but may arise from non coding RNA (ncRNA) [111], incompletely or unspliced transcripts [169]. In particular, ncRNAs are a likely source for a large fraction of the discrepancy based on estimates of

a 40-fold greater number of non-coding transcription units than protein coding genes in human and mouse. In mammals, open reading frames constitute <2% of the genome, this number being higher in less complex organisms like insects (18%), nematodes (25%), or fungi (60-60%) [146]. It has been estimated that around 98% of transcription is non-coding [104].

Much of the analysis and identification of ncRNA relies on the availability of genomic sequence which is currently unavailable for *X. laevis* and incomplete for *X. tropicalis*, the highly homologous diploid species.

6.1. microRNAs: Biogenesis and Prediction

One of the in recent years most intensively studied class of non-coding RNAs is the class of microRNAs (miRNAs). miRNAs play an important role as gene expression regulators in diverse organisms including animals and plants. Initially identified by Lee *et al.* [94] as *lin-4* RNA in *C. elegans* in 1993, the number of miRNAs registered in the miRBase database [56] has grown to 4361 entries (Release 9.0, October 2006) since then, including 474 for human and 373 for mouse.

miRNA Biogenesis

miRNAs are small (~22 nucleotides) noncoding RNA gene products. They are derived from long primary transcripts (pri-miRNAs), which can contain one or more miRNA precursors (pre-miRNAs). The pre-miRNAs are the products of the Drosha enzyme, which cuts the pri-miRNAs into ~70 nt long sequences which can form stable stem-loop (hairpin) structures, the mature miRNA present in one arm of the stem, which lacks large bulges or internal loops. Sometimes the pri-miRNA transcripts contain multiple hairpins, where different hairpins give rise to different miRNAs (polycistronic miRNA transcripts).

After transport to the cytoplasm by exportin-5, the RNase-III-type enzyme Dicer cuts the pre-miRNA into the active ~22 nt long mature miRNA. Dicer also initiates the formation of the RNA-induced silencing complex (RISC), which is responsible for gene silencing.

miRNAs function as gene regulators. In animals, the mature miRNA binds to specific sites in the 3' UTR of the target mRNA. The binding sites are not fully complementary,

which allows one miRNA to recognize multiple target sequences by forming imperfect duplexes with internal loops and bulges. These miRNA-mRNA duplexes cause either cleavage of the mRNA or inhibits protein translation and are crucial for the miRNA's regulatory activity.

miRNA Prediction

While the first miRNAs were discovered by positional cloning of sequences from RNA samples fractionated by size, nowadays a great portion of known miRNAs have been identified by pure computational approaches or in combination with biochemical methods. Ambros *et al.* [9] suggest the following annotation criteria for miRNAs:

Expression criteria (A) Hybridization of a distinct ~22 nt RNA transcript to a size-fractionated RNA sample.

(B) Identification of the ~22 nt sequence in a cDNA library made from size-fractionated RNA. Sequence must exactly match genomic sequence of source organism.

Biogenesis criteria (C) Prediction of a potential stem-loop precursor structure, containing the ~22 nt miRNA sequence within one arm. The hairpin must be the folding with the lowest free energy, as predicted by `mfold` [172] or another conventional RNA-folding program, and must include 16 bp involving the first 22 nt of the miRNA and the other arm of the hairpin. It should not contain large internal loops or bulges. These hairpins are usually ~60-80 nt long in animals.

(D) Phylogenetic conservation of the ~22 nt miRNA sequence and its predicted hairpin secondary structure.

(E) Detection of increased precursor accumulation in organisms with reduced Dicer function.

None of the above criteria on its own is sufficient for annotating a candidate gene as miRNA, evidence of both expression and biogenesis characteristic of miRNAs are required. However, Ambros *et al.* state that *“homologs of previously validated miRNAs need not meet as stringent criteria to be annotated as additional miRNA loci. Very close homologs*

6. Computational Identification of miRNAs in *X. laevis* EST clusters

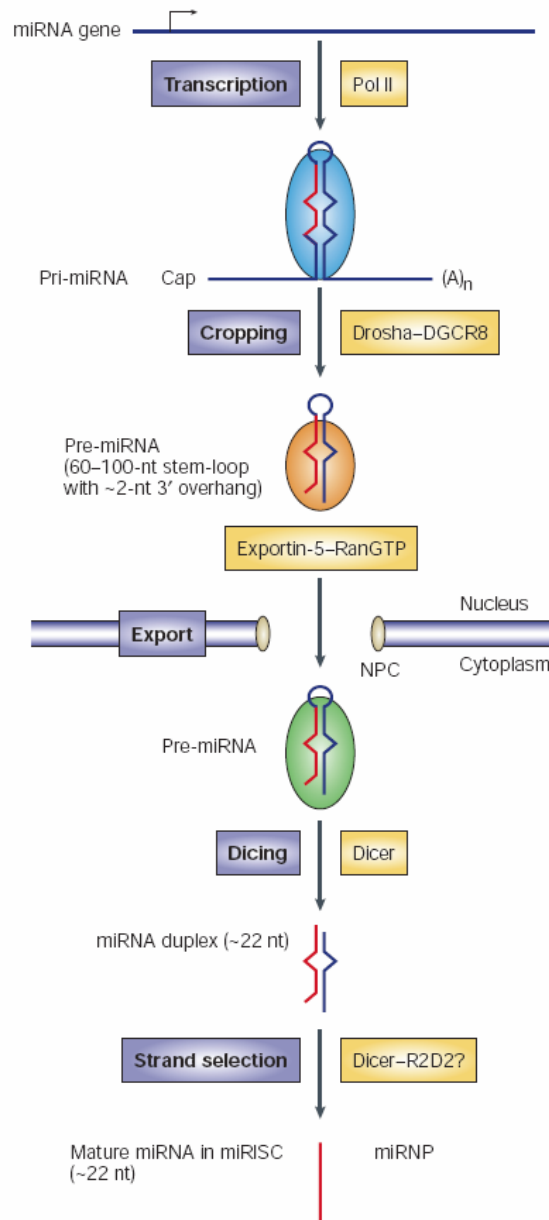


Figure 6.1.: miRNA biogenesis: miRNA genes are transcribed to pri-miRNAs and then cut by Drosha to ~70 nt pre-miRNAs, forming stable stem-loop structures. The pre-miRNAs are exported from the nucleus by exportin-5 and further processed in the cytoplasm. The RNase-III-type Dicer enzyme ~22 nt long miRNA duplexes from which one strand is selected as mature miRNA. (Adopted from [81]).

in other species can be annotated as miRNA orthologs without experimental validation, provided they satisfy criterion D.”

As the time of writing, out of 4361 entries in release 9.0 of miRBase 2024 (46.4%) have experimental evidence, 2337 entries (53.6%) have none. In the major model organisms hundreds of miRNAs have been discovered (e.g. human: 474, mouse: 373, zebrafish: 337, rat: 234). For *X. laevis* however, only 7 miRNAs are known, all identified by Watanabe *et al.* [161]. In the next section, we will describe a homology based approach to identify miRNAs in the *X. laevis* EST clusters described in Chapter 5.

6.2. Computational Identification of miRNAs

Following the recommendations by Ambros *et al.* we tried to computationally identify miRNAs in the *X. laevis* XenDB data set based on similarity to known miRNA sequences. ESTs are partial cDNA sequences of expressed genes. We count this as an expression criteria, however, further analysis (e.g. Northern blotting) should be applied to verify the identified genes as miRNA candidates.

We first downloaded all known metazoan mature miRNA sequences from the miRBase database. Additionally, information about the position of the mature miRNA in the precursor sequence was obtained. Next, the following strategy was used to identify homologue miRNAs in the XenDB data set:

1. Search XenDB contigs using *Vmatch* to identify matches against highly conserved mature miRNAs.
2. Shorten contig sequence to length of pre-miRNA as found in miRBase, preserving relative position of mature miRNA in pre-miRNA sequence.
3. Perform secondary structure prediction of pre-miRNA by shape analysis using *RNA-shapes* [142].
4. Filter candidate matches based on shape probability and minimum free energy (mfe) cutoff.

Vmatch was used as matching tool as it allows to specify the desired kind of matches much easier as e.g. BLAST. Mature miRNAs had to match with a minimum length of 20 nt

(forward or reverse complemented). Additionally, a seedlength of 8 and edit distance of 1 was chosen for the search.

Next, contig sequences matching mature miRNAs with the above parameters were trimmed to a length of ~60-80 nt, depending on the length of the corresponding pre-miRNA. The relative position of the mature miRNA sequence in the pre-miRNA as obtained from the miRBase entry was preserved. Assuming a high conservation of not only the mature sequence itself but also the position of the mature sequence in the precursor stem-loop structure, the putative precursor sequence was cut out from the contig sequence. This step makes the subsequent structure analysis easier, as candidate pre-miRNAs should fold into a stem-loop structure, containing the ~22 nt miRNA sequence within one arm of the hairpin.

The next step was to predict the secondary structure of the putative precursor sequence. The secondary structure analysis was performed by a shape analysis using *RNASHapes*. *RNASHapes* allows the analysis of shape representatives and the computation of accumulated shape probabilities. An RNA shape is an abstract representation of an RNA secondary structure. We used the most abstract representation of structure: shape type 5. It is an abstraction from loop and stack lengths, where stacking regions are represented by a pair of squared brackets (nested helices are combined) and unpaired regions are not included.

The probability of a shape is the sum of the probabilities of all structures that fall into this shape. As we are especially interested in hairpin shapes, we calculate the probability for this particular shape (*RNASHapes* options: `-m []` and `-q`). For the hairpin shape we also keep the shape representative (*shrep*), which is the structure with the minimum free energy (mfe) inside this shape class.

The initial matching step returns hundreds of matches in the EST data set. The subsequent shape analysis folds all of the resulting precursor sequences into a hairpin, albeit with poor probabilities. Therefore, the final step is a filter applied to the *RNASHapes* results to remove sequences that have low probabilities of forming the desired hairpin structure or only with unacceptable energy values.

6.3. Results: *X. laevis* miRNAs

We applied the described approach to identify miRNAs in the *X. laevis* XenDB data set. We used quite stringent matching criteria for *Vmatch* in order to keep the number of false positives in the predicted miRNA candidates to a minimum and ensure high sequence

conservation (1 mismatch) in the mature sequence. The search of all metazoan miRNAs in the *X. laevis* contigs resulted in 587 matches in 89 contigs, using a minimum match length of 20 nt and a maximal edit distance of 1. These included matches to 98 distinct metazoan miRNAs from 31 different species.

miRNA	Cluster	Pos	Shape	mfe	Probability
miR-1a	cluster:24915 contig:1	684..704	□	-30.00	0.9996044
miR-15a	cluster:11697 contig:1	199..219	□	-34.60	0.9999891
miR-17	vm_singlet:44527	94..114	□	-35.80	0.9975913
miR-18b	cluster:16044 contig:1	725..747	□	-29.70	0.9996439
<i>miR-19b</i>	cluster:16044 contig:1	424..446	□	-37.20	0.9997667
<i>miR-20</i>	cluster:16044 contig:1	571..592	□	-28.20	0.9999950
miR-23a	vm_singlet:187885	210..229	□	-25.70	0.9944582
miR-24b	vm_singlet:263804	95..115	□	-24.40	0.9998323
miR-27b	vm_singlet:263804	226..246	□	-45.60	0.9999043
miR-92	cluster:16044 contig:1	293..313	□	-29.90	0.9999795
miR-92a	vm_singlet:60513	69..89	□	-36.00	0.9999902
miR-106a	vm_singlet:44527	94..113	□	-32.50	0.9846771
<i>miR-133a</i>	cluster:6980 contig:1	258..279	□	-33.30	0.9999983
miR-133b	cluster:19810 contig:1	465..485	□	-32.70	0.9999486
miR-133d	cluster:24915 contig:1	266..287	□	-30.42	0.9999716
miR-194a	vm_singlet:181117	38..58	□	-34.50	0.9999917
miR-205a	vm_singlet:196915	231..252	□	-35.52	1.0000000
miR-223	cluster:17235 contig:1	512..532	□	-36.10	0.9876003
miR-363	cluster:16044 contig:1	162..181	□	-26.40	0.9999319
<i>miR-427</i>	cluster:2756 contig:1	356..377	□	-25.00	0.9947582
miR-689	vm_singlet:115768	194..214	□	-78.60	0.9998615

Table 6.1.: 21 *X. laevis* miRNAs identified in XenDB set with strong homology to known miRNAs. All corresponding pre-miRNAs fold confidently into a stem-loop structure. The table shows for each mature miRNA the position within the contig, mfe, and hairpin shape probability for the pre-miRNA as computed by *RNAshapes*. Previously known *X. laevis* miRNAs are shown in italics.

Next, the probability of the hairpin shape and the minimum free energy of the corresponding shrep of each candidate precursor sequence were obtained using *RNAshapes*. Again, we used stringent criteria for the following filtering: matches were discarded if the probability for a precursor candidate to form a hairpin shape was below 95% or the minimum free energy of the shrep was >-20 kcal/mol. 59 matches were discarded because of too low probabilities, 6 because of too high mfe values, and 101 because of both probability and mfe.

The remaining matches included 21 distinct miRNA candidates as shown in Table 6.1.

Only 4 of these are already known in *X. laevis* according to Release 9.0 of the miRBase database, leaving 17 new candidates identified in this study. Table 6.1 includes for each miRNA the contig containing the miRNA gene, the position of the mature sequence, the probability of the hairpin shape and the mfe of the corresponding shrep. The structures of the pre-miRNA candidates are shown in Appendix A, positions of the mature miRNA highlighted.

miR-17 cluster

Several miRNA genes have been shown to exist as clusters of 2 or more genes [114], suggesting that the transcription is controlled by common regulatory elements. If a single promoter drives the transcription of the clustered miRNA genes, the transcript must be polycistronic [95]. The two contigs 16044 and 24915 are examples of polycistronic transcripts (see Table 6.1).

Figure 6.2 shows the secondary structure of the sequence of contig 16044 as predicted by RNAfold [63]. It contains the precursor miRNAs of *miR-18b*, *miR-19b*, *miR-20*, *miR-92*, and *miR-363*. This cluster appears to be the *miR-17 cluster* as described by Tanzer *et al.* [147], who state that the miR-17 cluster had arose through a complex history of duplication and loss of individual member as well as duplications of entire clusters. In the case of contig 16044, *miR-17* is missing, while *miR-18b*, *miR-19b*, *miR-20*, and *miR-92* are members of the described *miR-17 cluster*.

6.4. Summary

The described method of a computational identification of miRNAs based on sequence similarity and secondary structure prediction works well in the case of *X. laevis* EST clusters. Although we used very stringent matching criteria, we successfully identified 17 new miRNA genes in the data set. Relaxing these criteria will probably identify more candidates, however one has to try to keep false positives at a minimum. It seems promising that ESTs are an invaluable resource not only for prediction of protein coding genes but also for noncoding RNAs, as ESTs are products of actually expressed genes. The analysis of the *X. laevis* ESTs gave information which did not seem to be available for species that lack genomic sequence.

Further analysis including a comparison to the *X. tropicalis* genome, which starts to

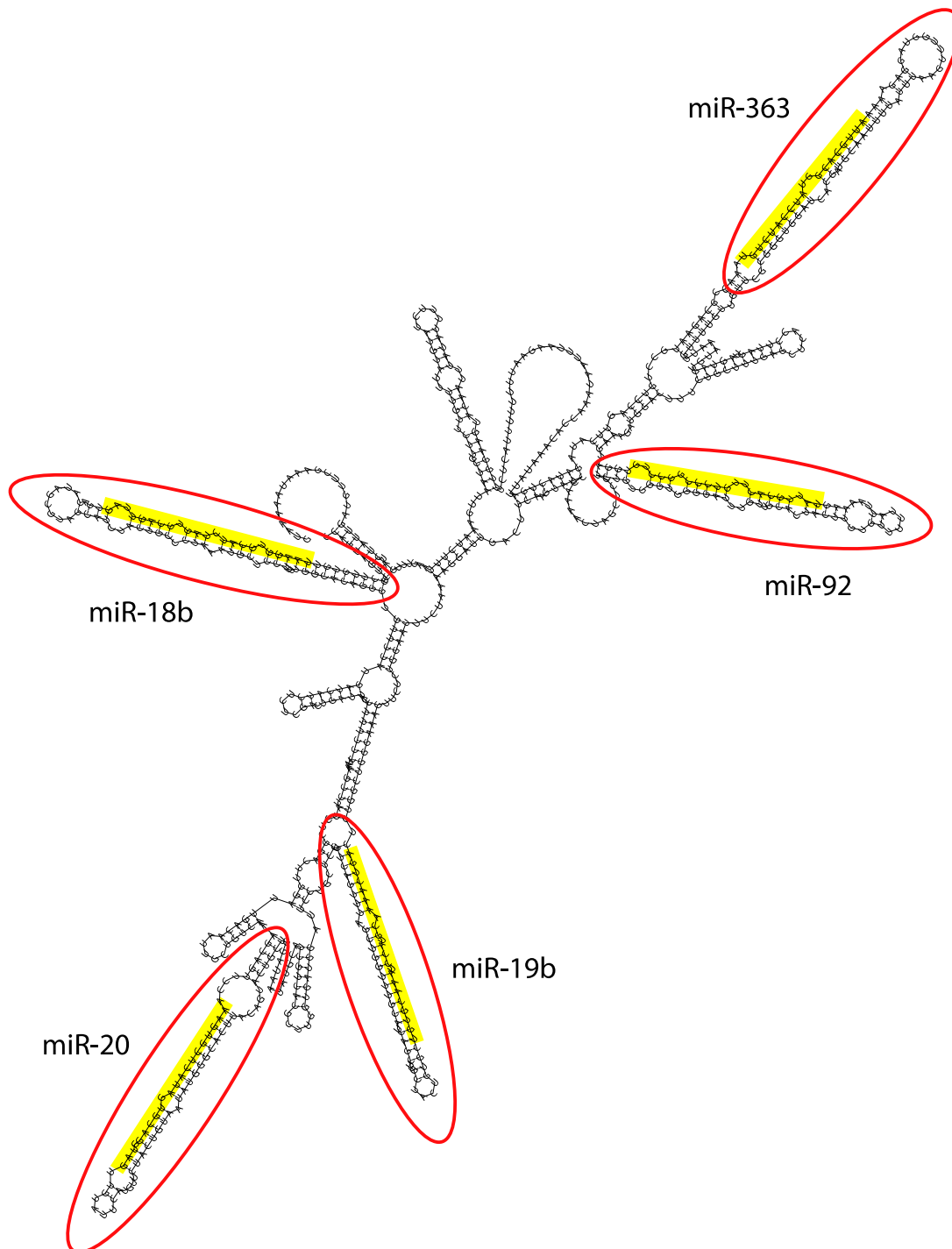


Figure 6.2.: Polycistronic transcript containing the miR-17 cluster of miRNAs identified in contig 16044 of clustered *X. laevis* EST data set.

6. Computational Identification of miRNAs in *X. laevis* EST clusters

become available, seems promising as part of future work emerging from this study. This includes the prediction of the targets for the miRNAs. The analysis of the XenDB clusters identified more than 5,700 full length *CAP3* contigs which have 3' UTR sequences available. Even more 3' UTR sequences are available for only partially reconstructed genes, in consequence of the EST sequencing protocol which favours 3' ends during the poly(dT) priming step. The UTRs are possible targets for the predicted miRNAs and an analysis with tools like *RNAhybrid* [128, 86] can be applied to predict target genes. The combination of the *X. laevis* and *X. tropicalis* data can enhance the quality of the predictions as *RNAhybrid* can not only calculate the statistical significance of individual binding sites, but also of binding sites in comparative analyses of orthologous sequences across species.

Conclusion and Outlook

With EST sequences remaining the only available information about the gene content of some species, EST mapping and clustering are important applications for genome analysis. We have shown that algorithms based on enhanced suffix arrays are well suited for the growing amounts of data to be expected in the future. Qualitatively, the clustering results are at least comparable to those of well known clustering tools, at the same time reducing the running time by up to two orders of magnitude. We therefore have achieved the design goal of being able to frequently update this aspect of the analysis.

The implemented clustering pipeline accomplished a comprehensive analysis of the *X. laevis* EST data. The resulting XenDB database provides a resource of gene-oriented EST clusters and transcript oriented contigs, enriched with various information from heterogeneous sources, that would be of value to the biology community and the Xenopus community in particular. Using the XenDB system, the biologist can identify sequences of interest using simple gene name queries, accessions, or gene ontologies. The identified sequences have been mapped to public resources like NCBI's UniGene and TIGR Gene Indices. Over 10,000 publicly available IMAGE clones were identified that maximize the 5' sequence to provide a full length construct when possible.

All the *X. laevis* sequences have been compared to the human and mouse protein sets

to identify conserved proteins. An obvious question is how complete is the *Xenopus* EST set and what percentage of genes have been identified assuming that the vast majority of protein coding sequences have been evolutionarily conserved. Of the 40,000 sequences in the IPI databases, more than 7,000 human and mouse sequences do not have a strong match in our data set. Thus, there is a considerable effort remaining to develop a complete *Xenopus* protein coding set. On the other hand, there are a large number of singletons remaining in the clustered EST set, and also a number of contigs do not have a match against any of the model organisms. These sequences are thought to represent rarely expressed genes, which may be of great biological interest.

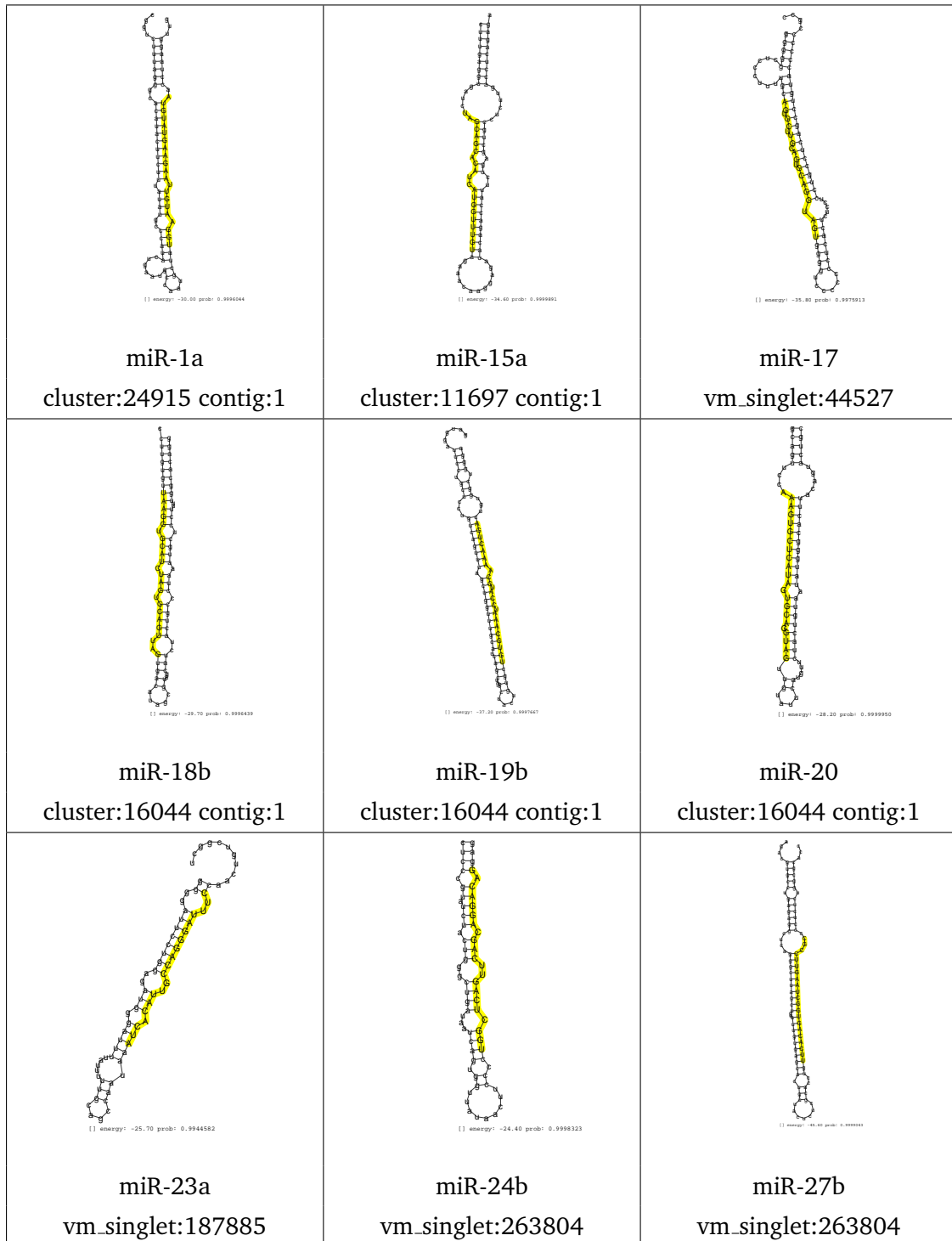
The identification of full-length cDNAs facilitate the production of many tools such as microarrays that provide a wealth of gene expression data in a single experiment. Since our first *X. laevis* microarray experiments [6] a number of groups have used this tool to study gene expression patterns in both global and more focused ways. The abundance of more full-length cDNAs enables the design of microarrays that can be used to complement existing gain of function experiments by identifying genes changing their expression patterns. The annotation of these sequences will speed up the analysis of the microarray data.



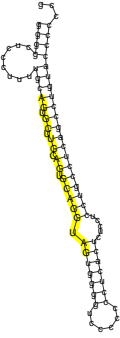

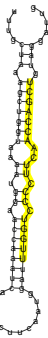
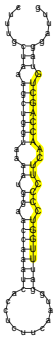

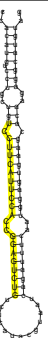
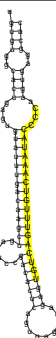
As with all ongoing high throughput sequencing efforts, certain aspects of the results change in proportion to the total number of sequences. A complete gene set for *Xenopus* will require additional sequencing. The difference in ploidy makes *X. laevis* distinct from all of the other organisms for which similar analysis have been performed. The generation of tetra, octo and dodecaploid species of *Xenopus* between 80 and 10 million years ago offers opportunities in the field of evolutionary biology. For example, comparisons of 3' UTR regions between in-paralogs of *X. laevis* and their counterpart diploid *tropicalis* species may improve statistical models of molecular evolution. In the course of our analysis we note the high degree of similarity between the *laevis* and *tropicalis* *Xenopus* species. This conservation may allow sequences from both species to be combined to generate a more complete set.

At the genome level, the potential availability of genome data from the polyploid species may provide insight into questions of chromosome segregation and silencing. The selection of *Xenopus* as a model organism by the NIH and the establishment of the Trans-NIH *Xenopus* Initiative have directly led to the support of EST and genome sequencing efforts. Among the priorities identified is the establishment and funding of a *Xenopus* Database which will integrate sequence, expression and other *Xenopus* data.

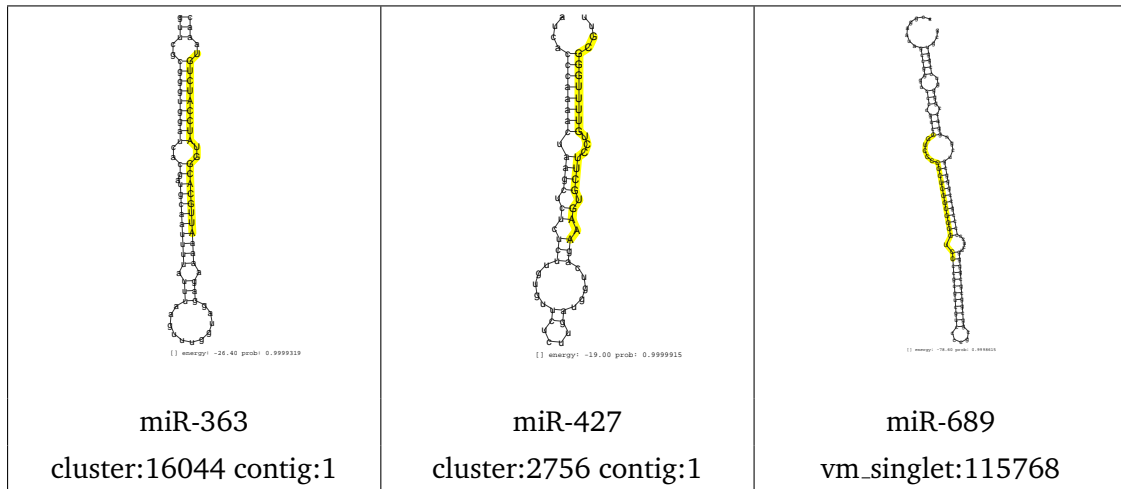
Structures of Predicted miRNA Precursors

A. Structures of Predicted miRNA Precursors



 <p>miR-92 cluster:16044 contig:1</p>	 <p>miR-92a vm_singlet:60513</p>	 <p>miR-106a vm_singlet:44527</p>
 <p>miR-133a cluster:6980 contig:1</p>	 <p>miR-133b cluster:19810 contig:1</p>	 <p>miR-133d cluster:24915 contig:1</p>
 <p>miR-194a vm_singlet:181117</p>	 <p>miR-205a vm_singlet:196915</p>	 <p>miR-223 cluster:17235 contig:1</p>

A. Structures of Predicted miRNA Precursors



Bibliography

- [1] J. Aaronson, B. Eckman, R. Blevins, J. Borkowski, J. Myerson, S. Imran, and K. Elliston. Toward the development of a gene index to the human genome: an assessment of the nature of high-throughput EST sequence data. *Genome Res*, 6(9):829–845, 1996.
- [2] M. Abouelhoda, S. Kurtz, and E. Ohlebusch. Replacing Suffix Trees with Enhanced Suffix Arrays. *Journal of Discrete Algorithms*, 2:53–86, 2004.
- [3] M. Adams, M. Dubnick, A. Kerlavage, R. Moreno, J. Kelley, T. Utterback, J. Nagle, C. Fields, and J. Venter. Sequence identification of 2,375 human brain genes. *Nature*, 355(6361):632–634, 1992.
- [4] M. Adams, J. Kelley, J. Gocayne, M. Dubnick, M. Polymeropoulos, H. Xiao, C. Merrill, A. Wu, B. Olde, R. Moreno, A. Kerlavage, W. McCombie, and J. Venter. Complementary DNA Sequencing: Expressed Sequence Tags and Human Genome Project. *Science*, 252(5013):1651–6, 1991.
- [5] M. Adams, A. Kerlavage, R. Fleischmann, R. Fuldner, C. Bult, N. Lee, E. Kirkness, K. Weinstock, J. Gocayne, and O. White *et al.* Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature*, 377(6547 Suppl):3–174, 1995.
- [6] C. R. Altmann, E. Bell, A. Sczyrba, J. Pun, S. Bekiranov, T. Gaasterland, and A. H. Brivanlou. Microarray-based analysis of early development in *Xenopus laevis*. *Dev Biol*, 236(1):64–75, Aug 2001.
- [7] S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–10, 1990.
- [8] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–3402, Sep 1997.
- [9] V. Ambros, B. Bartel, D. P. Bartel, C. B. Burge, J. C. Carrington, X. Chen, G. Dreyfuss, S. R. Eddy, S. Griffiths-Jones, M. Marshall, M. Matzke, G. Ruvkun, and T. Tuschl. A uniform system for microRNA annotation. *RNA*, 9(3):277–279, Mar 2003.

Bibliography

- [10] K. Arima, J. Shiotsugu, R. Niu, R. Khandpur, M. Martinez, Y. Shin, T. Koide, K. W. Y. Cho, A. Kitayama, N. Ueno, R. A. S. Chandraratna, and B. Blumberg. Global analysis of RAR-responsive genes in the *Xenopus* neurula using cDNA microarrays. *Dev Dyn*, 232(2):414–431, Feb 2005.
- [11] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25(1):25–29, May 2000.
- [12] D. Baltimore. Viral RNA-dependent DNA Polymerase: RNA-dependent DNA polymerase in virions of RNA tumour viruses. *Nature*, 226(5252):1209–1211, 1970.
- [13] T. Barrett, T. O. Suzek, D. B. Troup, S. E. Wilhite, W.-C. Ngau, P. Ledoux, D. Rudnev, A. E. Lash, W. Fujibuchi, and R. Edgar. Ncbi geo: mining millions of expression profiles—database and tools. *Nucleic Acids Res*, 33(Database issue):D562–D566, Jan 2005.
- [14] M. Beckstette, J. Mailänder, R. Marhöfer, A. Sczyrba, E. Ohlebusch, R. Giegerich, and P. Selzer. Genlight: Interactive high-throughput sequence analysis and comparative genomics. *Journal of Integrative Bioinformatics*, 0008():, 2004.
- [15] M. Beckstette, A. Sczyrba, and P. Selzer. Genlight: Interactive high-throughput sequence analysis and comparative genomics. In *Proceedings of the German Conference on Bioinformatics*, volume P-53, pages 179–186. GI Lecture Notes in Informatics, 2004.
- [16] J. Bishop, J. Morton, M. Rosbash, and M. Richardson. Three abundance classes in HeLa cell messenger RNA. *Nature*, 250(463):199–204, 1974.
- [17] M. Boguski. The turning point in genome research. *Trends Biochem Sci*, 20(8):295–296, 1995.
- [18] M. Boguski, T. Lowe, and C. Tolstoshev. dbEST - database for "expressed sequence tags". *Nat Genet*, 4(4):332–333, 1993.
- [19] M. Boguski and G. Schuler. ESTablishing a human transcript map. *Nat Genet*, 10(4):369–71, 1995.
- [20] M. Boguski, C. Tolstoshev, and D. J. Bassett. Gene discovery in dbEST. *Science*, 265(5181):1993–1994, 1994.
- [21] M. Bonaldo, G. Lennon, and M. Soares. Normalization and subtraction: two approaches to facilitate gene discovery. *199*, 6(9):791–806, 1996.
- [22] K. Boon, E. C. Osorio, S. F. Greenhut, C. F. Schaefer, J. Shoemaker, K. Polyak, P. J. Morin, K. H. Buetow, R. L. Strausberg, S. J. D. Souza, and G. J. Riggins. An anatomy of normal and malignant gene expression. *Proc Natl Acad Sci U S A*, 99(17):11287–11292, Aug 2002.
- [23] S. Brenner. The human genome: the nature of the enterprise. In *Human Genetic Information: Science, Law and Ethics*, volume 149 of *Ciba Found Symp*, pages 6–12, 1990.
- [24] D. Brett, H. Pospisil, J. Valrcel, J. Reich, and P. Bork. Alternative splicing and genome complexity. *Nat Genet*, 30(1):29–30, Jan 2002.
- [25] C. Burge and S. Karlin. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, 268(1):78–94, 1997.

- [26] J. Burke, D. Davison, and W. Hide. d2_cluster: A Validated Method for Clustering EST and Full-Length cDNA Sequences. *Genome Res*, 9(11):1135–42, 1999.
- [27] S. Burkhardt, A. Crauser, P. Ferragina, H.-P. Lenhof, E. Rivals, and M. Vingron. q-gram based database searching using a suffix array (QUASAR). In *RECOMB '99: Proceedings of the third annual international conference on Computational molecular biology*, pages 77–83. ACM Press, 1999.
- [28] E. Camon, M. Magrane, D. Barrell, V. Lee, E. Dimmer, J. Maslen, D. Binns, N. Harte, R. Lopez, and R. Apweiler. The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res*, 32(Database issue):D262–D266, Jan 2004.
- [29] M. Cariaso, P. Folta, M. Wagner, T. Kuczumarski, and G. Lennon. IMAGEne I: clustering and ranking of I.M.A.G.E. cDNA clones corresponding to known genes. *Bioinformatics*, 15(12):965–73, 1999.
- [30] P. Carninci, K. Waki, T. Shiraki, H. Konno, K. Shibata, M. Itoh, K. Aizawa, T. Arakawa, Y. Ishii, D. Sasaki, H. Bono, S. Kondo, Y. Sugahara, R. Saito, N. Osato, S. Fukuda, K. Sato, A. Watahiki, T. Hirozane-Kishikawa, M. Nakamura, Y. Shibata, A. Yasunishi, N. Kikuchi, A. Yoshiki, M. Kusakabe, S. Gustinich, K. Beisel, W. Pavan, V. Aidinis, A. Nakagawara, W. A. Held, H. Iwata, T. Kono, H. Nakauchi, P. Lyons, C. Wells, D. A. Hume, M. Fagiolini, T. K. Hensch, M. Brinkmeier, S. Camper, J. Hirota, P. Mombaerts, M. Muramatsu, Y. Okazaki, J. Kawai, and Y. Hayashizaki. Targeting a Complex Transcriptome: The Construction of the Mouse Full-Length cDNA Encyclopedia. *Genome Res.*, 13(6b):1273–1289, 2003.
- [31] K.-M. Chao, W. Pearson, and W. Miller. Aligning two sequences within a specified diagonal band. *Comput. Appl. Biosci.*, 8(5):481–487, 1992.
- [32] B. Chevreux, T. Pfisterer, B. Drescher, A. Driesel, W. Muller, T. Wetter, and S. Suhai. Using the miraEST Assembler for Reliable and Automated mRNA Transcript Assembly and SNP Detection in Sequenced ESTs. *Genome Res.*, 14(6):1147–1159, 2004.
- [33] A. Christoffels, A. van Gelder, G. Greyling, R. Miller, T. Hide, and W. Hide. STACK: Sequence Tag Alignment and Consensus Knowledgebase. *Nucleic Acids Res*, 29(1):234–8, 2001.
- [34] H. A. Chung, J. Hyodo-Miura, A. Kitayama, C. Terasaka, T. Nagamune, and N. Ueno. Screening of FGF target genes in *Xenopus* by microarray: temporal dissection of the signalling pathway using a chemical inhibitor. *Genes Cells*, 9(8):749–761, Aug 2004.
- [35] L. Clarke and J. Carbon. A colony bank containing synthetic CoI EI hybrid plasmids representative of the entire *E. coli* genome. *Cell*, 9(1):91–99, 1976.
- [36] E. Coward, S. Haas, and M. Vingron. SpliceNest: visualizing gene structure and alternative splicing based on EST clusters. *Trends Genet*, 18(1):53–55, 2002.
- [37] W. G. Cox and A. Hemmati-Brivanlou. Caudalization of neural fate by tissue recombination and bfgf. *Development*, 121(12):4349–4358, Dec 1995.
- [38] D. Crump, K. Werry, N. Veldhoen, G. V. Aggelen, and C. C. Helbing. Exposure to the herbicide acetochlor alters thyroid hormone-dependent gene expression and metamorphosis in *Xenopus Laevis*. *Environ Health Perspect*, 110(12):1199–1205, Dec 2002.
- [39] D. Davison and J. Burke. Brute force estimation of the number of human genes using EST clustering as a measure. *IBM J Res Dev*, 45(3-4):439–447, 2001.

- [40] C. Del Val, K. Glatting, and S. Suhai. cDNA2Genome: A tool for mapping and annotating cDNAs. *BMC Bioinformatics*, 4(1):39, 2003.
- [41] C. Drepper. *Identifizierung und Charakterisierung des krankheitsauslösenden Gens in einem Mausmodell für humane Motoneuronerkrankungen*. PhD thesis, Fakultät für Biologie, Universität Bielefeld, Germany, 2005.
- [42] B. Eckman, J. Aaronson, J. Borkowski, W. Bailey, K. Elliston, A. Williamson, and R. Blevins. The Merck Gene Index browser: an extensible data integration system for gene finding, gene characterization and EST data mining. *Bioinformatics*, 14(1):2–13, 1998.
- [43] R. Edgar, M. Domrachev, and A. E. Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*, 30(1):207–210, Jan 2002.
- [44] E. E. Eichler. Masquerading Repeats: Paralogous Pitfalls of the Human Genome. *Genome Res.*, 8(8):758–762, 1998.
- [45] L. Florea, G. Hartzell, Z. Zhang, G. Rubin, and W. Miller. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res*, 8(9):967–74, 1998.
- [46] E. Fowlkes and C. Mallows. A method for comparing two hierarchical clusterings, (with comments and rejoinder). *Journal of the American Statistical Association*, 78:553–584, 1983.
- [47] M. Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Am. Stat. Assoc.*, 32(200):675–701, 1937.
- [48] W. J. Gehring. The genetic control of eye development and its implications for the evolution of the various eye-types. *Int J Dev Biol*, 46(1):65–73, Jan 2002.
- [49] W. J. Gehring, M. Affolter, and T. Brglin. Homeodomain proteins. *Annu Rev Biochem*, 63:487–526, 1994.
- [50] W. J. Gehring and K. Ikeo. Pax 6: mastering eye morphogenesis and eye evolution. *Trends Genet*, 15(9):371–377, Sep 1999.
- [51] C. Gemund, C. Ramu, B. Altenberg-Greulich, and T. J. Gibson. Gene2EST: a BLAST2 server for searching expressed sequence tag (EST) databases with eukaryotic gene-sized queries. *Nucl. Acids. Res.*, 29(6):1272–1277, 2001.
- [52] T. Glaser, D. S. Walton, and R. L. Maas. Genomic structure, evolutionary conservation and aniridia mutations in the human PAX6 gene. *Nat Genet*, 2(3):232–239, Nov 1992.
- [53] G. Gonnet, R. Baeza-Yates, and T. Snider. *Information Retrieval: Data Structures and Algorithms*, chapter New indices for text: PAT trees and PAT arrays, pages 66–82. Prentice Hall, Englewood Cliffs, N.J., 1992.
- [54] P. Green. Phrap website. <http://www.phrap.org/>.
- [55] R. E. Green, J. Krause, S. E. Ptak, A. W. Briggs, M. T. Ronan, J. F. Simons, L. Du, M. Egholm, J. M. Rothberg, M. Paunovic, and S. Pbo. Analysis of one million base pairs of neanderthal dna. *Nature*, 444(7117):330–336, Nov 2006.
- [56] S. Griffiths-Jones, R. J. Grocock, S. van Dongen, A. Bateman, and A. J. Enright. miR-Base: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res*, 34(Database issue):D140–D144, Jan 2006.
- [57] D. Gusfield. *Algorithms on Strings, Trees, and Sequences*. Cambridge University Press, New York, 1997.

- [58] J. Hancock and J. Armstrong. SIMPLE34: an improved and enhanced implementation for VAX and Sun computers of the SIMPLE algorithm for analysis of clustered repetitive motifs in nucleotide sequences. *Comput Appl Biosci*, 10(1):67–70, 1994.
- [59] M. A. Harris, J. Clark, A. Ireland, J. Lomax, M. Ashburner, R. Foulger, K. Eilbeck, S. Lewis, B. Marshall, C. Mungall, J. Richter, G. M. Rubin, J. A. Blake, C. Bult, M. Dolan, H. Drabkin, J. T. Eppig, D. P. Hill, L. Ni, M. Ringwald, R. Balakrishnan, J. M. Cherry, K. R. Christie, M. C. Costanzo, S. S. Dwight, S. Engel, D. G. Fisk, J. E. Hirschman, E. L. Hong, R. S. Nash, A. Sethuraman, C. L. Theesfeld, D. Botstein, K. Dolinski, B. Feierbach, T. Berardini, S. Mundodi, S. Y. Rhee, R. Apweiler, D. Barrell, E. Camon, E. Dimmer, V. Lee, R. Chisholm, P. Gaudet, W. Kibbe, R. Kishore, E. M. Schwarz, P. Sternberg, M. Gwinn, L. Hannick, J. Wortman, M. Berriman, V. Wood, N. de la Cruz, P. Tonellato, P. Jaiswal, T. Seigfried, R. White, and G. O. Consortium. The gene ontology (go) database and informatics resource. *Nucleic Acids Res*, 32(Database issue):D258–D261, Jan 2004.
- [60] E. Hartuv, A. Schmitt, J. Lange, S. Meier-Ewert, H. Lehrach, and R. Shamir. An algorithm for clustering cDNAs for gene expression analysis. In *RECOMB '99: Proceedings of the third annual international conference on Computational molecular biology*, pages 188–197. ACM Press, 1999.
- [61] W. Hide, J. Burke, and D. Davison. Biological evaluation of d2, an algorithm for high-performance sequence comparison. *J Comput Biol*, 1(3):199–215, 1994.
- [62] L. Hillier, G. Lennon, M. Becker, M. Bonaldo, B. Chiapelli, S. Chissoe, N. Dietrich, T. DuBuque, A. Favello, W. Gish, M. Hawkins, M. Hultman, T. Kucaba, M. Lacy, M. Le, N. Le, E. Mardis, B. Moore, M. Morris, J. Parsons, C. Prange, L. Rifkin, T. Rohlfling, K. Schellenberg, and M. Marra. Generation and analysis of 280,000 human expressed sequence tags. *Genome Res.*, 6(9):807–828, 1996.
- [63] I. Hofacker, W. Fontana, P. Stadler, L. Bonhoeffer, M. Tacker, and P. Schuster. Fast Folding and Comparison of RNA Secondary Structures. *Monatsh. Chem.*, 125:167–188, 1994.
- [64] R. Houlgatte, R. Mariage-Samson, S. Duprat, A. Tessier, S. Bentolila, B. Lamy, and C. Auffray. The Genexpress Index: a resource for gene discovery and the genic map of the human genome. *Genome Res.*, 5(3):272–304, 1995.
- [65] X. Huang and A. Madan. CAP3: A DNA sequence assembly program. *Genome Res*, 9(9):868–77, 1999.
- [66] L. Hubert and P. Arabie. Comparing Partitions. *Journal of Classification*, 2:193–218, 1985.
- [67] T. Imanishi, T. Itoh, Y. Suzuki, C. O'Donovan, S. Fukuchi, K. Koyanagi, R. Barrero, T. Tamura, Y. Yamaguchi-Kabata, and M. Tanino *et al.* Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biol*, 2:E162, 2004.
- [68] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.
- [69] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, Oct 2004.
- [70] H. V. Isaacs, M. E. Pownall, and J. M. Slack. Regulation of hox gene expression and posterior development by the xenopus caudal homologue xcad3. *EMBO J*, 17(12):3413–3427, Jun 1998.
- [71] P. Jaccard. Nouvelles recherches sur la distribution florale. *Bull. Soc. Vaudoise Sci. Nat.*, 44:223–270, 1908.

Bibliography

- [72] A. Jain and R. Dubes. *Algorithms for Clustering Data*. Prentice Hall, 1988.
- [73] J. Jurka. Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet*, 16(9):418–20, 2000.
- [74] J. Jurka, V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, and J. Walichiewicz. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res*, 110:462–467, 2005.
- [75] A. Kalyanaraman, S. Aluru, V. Brendel, and S. Kothari. Space and Time efficient parallel algorithms and software for EST clustering. *IEEE Transactions on Parallel and Distributed Systems*, 14(12):1209–1221, 2003.
- [76] A. Kalyanaraman, S. Aluru, S. Kothari, and V. Brendel. Efficient clustering of large EST data sets on parallel computers. *Nucl. Acids. Res.*, 31(11):2963–2974, 2003.
- [77] J. Kärkkäinen and P. Sanders. Simple linear work suffix array construction. In *Proc. 30th International Colloquium on Automata, Languages and Programming (ICALP '03)*, volume 2719 of *Lecture Notes in Computer Science*, pages 943–955. Springer, 2003.
- [78] W. Kent. BLAT-The BLAST-Like Alignment Tool. *Genome Res.*, 12(4):656–664, 2002.
- [79] P. J. Kersey, J. Duarte, A. Williams, Y. Karavidopoulou, E. Birney, and R. Apweiler. The International Protein Index: an integrated database for proteomics experiments. *Proteomics*, 4(7):1985–1988, Jul 2004.
- [80] L. Kim, J. Sim, H. Park, and K. Park. Linear-time construction of suffix arrays. In *Combinatorial Pattern Matching: 14th Annual Symposium (CPM 2003)*, volume 2676 of *Lecture Notes in Computer Science*, pages 186–199. Springer, 2003.
- [81] V. N. Kim. MicroRNA biogenesis: coordinated cropping and dicing. *Nat Rev Mol Cell Biol*, 6(5):376–385, May 2005.
- [82] R. Knig, D. Baldessari, N. Pollet, C. Niehrs, and R. Eils. Reliability of gene expression ratios for cDNA microarrays in multiconditional experiments with a reference design. *Nucleic Acids Res*, 32(3):e29, 2004.
- [83] M. Ko, X. Wang, J. Horton, M. Hagen, N. Takahashi, Y. Maezaki, and J. Nadeau. Genetic mapping of 40 cDNA clones on the mouse genome by PCR. *Mamm Genome*, 5(6):349–355, 1994.
- [84] P. Ko and S. Aluru. Space efficient linear time construction of suffix arrays. In *Combinatorial Pattern Matching: 14th Annual Symposium (CPM 2003)*, volume 2676 of *Lecture Notes in Computer Science*, pages 200–210. Springer, 2003.
- [85] A. A. Komar and M. Hatzoglou. Internal ribosome entry sites in cellular mRNAs: mystery of their existence. *J Biol Chem*, 280(25):23425–23428, Jun 2005.
- [86] J. Krüger and M. Rehmsmeier. RNAhybrid: microRNA target prediction easy, fast and flexible. *Nucleic Acids Res*, 34(Web Server issue):W451–W454, Jul 2006.
- [87] J. Krüger, A. Sczyrba, S. Kurtz, and R. Giegerich. e2g: an interactive web-based server for efficiently mapping large EST and cDNA sets to genomic sequences. *Nucl. Acids. Res.*, 32(suppl_2):W301–304, 2004.
- [88] S. Kurtz. Reducing the Space Requirement of Suffix Trees. *Software—Practice and Experience*, 29(13):1149–1171, 1999.

- [89] S. Kurtz. *The Vmatch large scale sequence analysis software: A Manual*. Center for Bioinformatics, University of Hamburg, September 2005.
- [90] S. Kurtz, J. Choudhuri, E. Ohlebusch, C. Schleiermacher, J. Stoye, and R. Giegerich. RE-PUTer: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res*, 29(22):4633–42, 2001.
- [91] A. Lal, A. E. Lash, S. F. Altschul, V. Velculescu, L. Zhang, R. E. McLendon, M. A. Marra, C. Prange, P. J. Morin, K. Polyak, N. Papadopoulos, B. Vogelstein, K. W. Kinzler, R. L. Strausberg, and G. J. Riggins. A public database for gene expression in human cancers. *Cancer Res*, 59(21):5403–5407, Nov 1999.
- [92] A. E. Lash, C. M. Tolstoshev, L. Wagner, G. D. Schuler, R. L. Strausberg, G. J. Riggins, and S. F. Altschul. SAGEmap: a public gene expression resource. *Genome Res*, 10(7):1051–1060, Jul 2000.
- [93] B. T. K. Lee, T. W. Tan, and S. Ranganathan. MGAlignIt: a web service for the alignment of mRNA/EST and genomic sequences. *Nucl. Acids. Res.*, 31(13):3533–3536, 2003.
- [94] R. C. Lee, R. L. Feinbaum, and V. Ambros. The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell*, 75(5):843–854, Dec 1993.
- [95] Y. Lee, K. Jeon, J.-T. Lee, S. Kim, and V. N. Kim. MicroRNA maturation: stepwise processing and subcellular localization. *EMBO J*, 21(17):4663–4670, Sep 2002.
- [96] G. Lennon, C. Auffray, M. Polymeropoulos, and M. Soares. The I.M.A.G.E. Consortium: an integrated molecular analysis of genomes and their expression. *Genomics*, 33(1):151–152, 1996.
- [97] F. Liang, I. Holt, G. Pertea, S. Karamycheva, S. Salzberg, and J. Quackenbush. An optimized protocol for analysis of EST sequences. *Nucleic Acids Res*, 28(18):3657–65, 2000.
- [98] W. Lin, H. Yang, and M. Lee. Allelic variation in gene expression identified through computational analysis of the dbEST database. *Genomics*, 2005.
- [99] K. Liolios, N. Tavernarakis, P. Hugenholtz, and N. C. Kyrpides. The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide. *Nucleic Acids Res*, 34(Database issue):D332–D334, Jan 2006.
- [100] A. J. Lopez. Alternative splicing of pre-mRNA: developmental consequences and mechanisms of regulation. *Annu Rev Genet*, 32:279–305, 1998.
- [101] K. Malde, E. Coward, and I. Jonassen. Fast sequence clustering using a suffix array algorithm. *Bioinformatics*, 19(10):1221–1226, 2003.
- [102] K. Malde, E. Coward, and I. Jonassen. A graph based algorithm for generating EST consensus sequences. *Bioinformatics*, 21(8):1371–1375, 2005.
- [103] U. Manber and E. Myers. Suffix arrays: a new method for on-line string searches. *SIAM J. Comput.*, 22(5):935–948, 1993.
- [104] J. S. Mattick. Non-coding RNAs: the architects of eukaryotic complexity. *EMBO Rep*, 2(11):986–991, Nov 2001.
- [105] R. Medzhitov, P. Preston-Hurlburt, and C. J. Janeway. A human homologue of the *Drosophila* Toll protein signals activation of adaptive immunity. *Nature*, 388(6640):394–397, 1997.

Bibliography

- [106] L. Michaut, S. Flister, M. Neeb, K. P. White, U. Certa, and W. J. Gehring. Analysis of the eye developmental pathway in *Drosophila* using DNA microarrays. *Proc Natl Acad Sci U S A*, 100(7):4024–4029, Apr 2003.
- [107] R. Miller, A. Christoffels, C. Gopalakrishnan, J. Burke, A. Ptitsyn, T. Broveak, and W. Hide. A comprehensive approach to clustering of expressed human gene sequence: the sequence tag alignment and consensus knowledge base. *Genome Res*, 9(11):1143–55, 1999.
- [108] R. Miller, M. Phillips, I. Jo, M. Donaldson, J. Studebaker, N. Addleman, S. Alfisi, W. Ankener, H. Bhatti, C. Callahan, B. Carey, C. Conley, J. Cyr, V. Derohannessian, R. Donaldson, C. Elosua, S. Ford, A. Forman, C. Gelfand, N. Grecco, S. Gutendorf, C. Hock, M. Hozza, S. Hur, S. In, D. Jackson, S. Jo, S. Jung, S. Kim, K. Kimm, E. Kloss, D. Koboldt, J. Kuebler, F. Kuo, J. Lathrop, J. Lee, K. Leis, S. Livingston, E. Lovins, M. Lundy, S. Maggan, M. Minton, M. Mockler, D. Morris, E. Nachtman, B. Oh, C. Park, C. Park, N. Pavelka, A. Perkins, S. Restine, R. Sachidanandam, A. Reinhart, K. Scott, G. Shah, J. Tate, S. Varde, A. Walters, J. White, Y. Yoo, J. Lee, M. Boyce-Jacino, P. Kwok, and T. S. C. A. F. Project. High-density single-nucleotide polymorphism maps of the human genome. *Genomics*, 86(2):117–126, 2005.
- [109] G. Milligan and M. Cooper. A Study of the Comparability of External Criteria for Hierarchical Cluster Analysis. *Multivariate Behav Res*, 21:441–458, 1986.
- [110] G. Milligan and D. Schilling. Asymptotic and finite sample characteristics of four external criterion measures. *Multivariate Behavioral Research*, 20:97–109, 1985.
- [111] C. Morey and P. Avner. Employment opportunities for non-coding rnas. *FEBS Lett*, 567(1):27–34, Jun 2004.
- [112] L. Morey and A. Agresti. The measurement of classification agreement: An adjustment to the rand statistic for chance agreement. *Educational and Psychological Measurement*, 44:33–37, 1984.
- [113] R. Mott. EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. *CABIOS*, 13(4):477–8, 1997.
- [114] Z. Mourelatos, J. Dostie, S. Paushkin, A. Sharma, B. Charroux, L. Abel, J. Rappsilber, M. Mann, and G. Dreyfuss. miRNPs: a novel class of ribonucleoproteins containing numerous microRNAs. *Genes Dev*, 16(6):720–728, Mar 2002.
- [115] Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–562, Dec 2002.
- [116] I. Muoz-Sanjuan, E. Bell, C. R. Altmann, A. Vonica, and A. H. Brivanlou. Gene profiling during neural induction in *Xenopus laevis*: regulation of BMP signaling by post-transcriptional mechanisms and TAB3, a novel TAK1-binding protein. *Development*, 129(23):5529–5540, Dec 2002.
- [117] T. Nakamura, G. Morin, K. Chapman, S. Weinrich, W. Andrews, J. Lingner, C. Harley, and T. Cech. Telomerase catalytic subunit homologs from fission yeast and human. *Science*, 277(5328):955–9(5328):955–959, 1997.
- [118] National Institute of Science and Technology (NIST), USA. Secure hash standard. Federal Information Processing Standard FIPS PUB 180-2, August 2000.
- [119] S. Ohno. *Evolution by gene duplication*. Springer, 1970.

- [120] W. R. Pearson, T. Wood, Z. Zhang, and W. Miller. Comparison of DNA sequences with protein sequences. *Genomics*, 46(1):24–36, Nov 1997.
- [121] D. A. Peiffer, A. V. Bubnoff, Y. Shin, A. Kitayama, M. Mochii, N. Ueno, and K. W. Y. Cho. A *Xenopus* DNA microarray approach to identify novel direct BMP target genes involved in early embryonic development. *Dev Dyn*, 232(2):445–456, Feb 2005.
- [122] G. Pertea, X. Huang, F. Liang, V. Antonescu, R. Sultana, S. Karamycheva, Y. Lee, J. White, F. Cheung, B. Parvizi, J. Tsai, and J. Quackenbush. TIGR Gene Indices clustering tools (TG-ICL): a software system for fast clustering of large EST datasets. *Bioinformatics*, 19(5):651–652, 2003.
- [123] J. Pontius, L. Wagner, and G. Schuler. *The NCBI Handbook*, chapter UniGene: a unified view of the transcriptome. National Center for Biotechnology Information (NCBI), Bethesda (MD), 2003.
- [124] A. Ptitsyn and W. Hide. CLU: a new algorithm for EST clustering. *BMC Bioinformatics*, 6(Suppl 2):S3, 2005.
- [125] S. Putney, W. Herlihy, and P. Schimmel. A new troponin T and cDNA clones for 13 different muscle proteins, found by shotgun sequencing. *Nature*, 302(5910):718–721, 1983.
- [126] J. Quackenbush, J. Cho, D. Lee, F. Liang, I. Holt, S. Karamycheva, B. Parvizi, G. Pertea, R. Sultana, and J. White. The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Res*, 29(1):159–64, 2001.
- [127] W. M. Rand. Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.
- [128] M. Rehmsmeier, P. Steffen, M. Höchsmann, and R. Giegerich. Fast and effective prediction of microRNA/target duplexes. *RNA*, 10(10):1507–1517, Oct 2004.
- [129] S. Rudd. Expressed sequence tags: alternative or complement to whole genome sequences? *Trends Plant Sci*, 8(7):321–329, 2003.
- [130] J. Sambrook and D. Russell. *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, 3rd edition, 2001.
- [131] F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*, 74(12):5463–5467, Dec 1977.
- [132] T. Schmitt-John, C. Drepper, A. Mussmann, P. Hahn, M. Kuhlmann, C. Thiel, M. Hafner, A. Lengeling, P. Heimann, J. M. Jones, M. H. Meisler, and H. Jockusch. Mutation of *vps54* causes motor neuron disease and defective spermiogenesis in the wobbler mouse. *Nat Genet*, 37(11):1213–1215, Nov 2005.
- [133] G. Schuler. Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J Mol Med*, 75(10):694–8, 1997.
- [134] G. D. Schuler, M. S. Boguski, E. A. Stewart, L. D. Stein, G. Gyapay, K. Rice, R. E. White, P. Rodriguez-Tom, A. Aggarwal, E. Bajorek, S. Bentolila, B. B. Birren, A. Butler, A. B. Castle, N. Chiannikulchai, A. Chu, C. Clee, S. Cowles, P. J. R. Day, T. Dibling, C. East, N. Drouot, I. Dunham, S. Duprat, C. Edwards, J.-B. Fan, N. Fang, C. Fizames, C. Garrett, L. Green, D. Hadley, M. Harris, P. Harrison, S. Brady, A. Hicks, E. Holloway, L. Hui, S. Hussain, C. Louis-Dit-Sully, J. Ma, A. MacGilvery, C. Mader, A. Maratukulam, T. C. Matise, K. B. McKusick, J. Morissette, A. Mungall, D. Muselet, H. C. Nusbaum, D. C. Page, A. Peck, S. Perkins, M. Piercy, F. Qin, J. Quackenbush, S. Ranby, T. Reif, S. Rozen, C. Sanders, X. She, J. Silva,

- D. K. Slonim, C. Soderlund, W.-L. Sun, P. Tabar, T. Thangarajah, N. Vega-Czarny, D. Vollrath, S. Voyticky, T. Wilmer, X. Wu, M. D. Adams, C. Auffray, N. A. R. Walter, R. Brandon, A. Dehejia, P. N. Goodfellow, R. Houlgatte, J. Hudson, J. R., S. E. Ide, K. R. Iorio, W. Y. Lee, N. Seki, T. Nagase, K. Ishikawa, N. Nomura, C. Phillips, M. H. Polymeropoulos, M. Sandusky, K. Schmitt, R. Berry, K. Swanson, R. Torres, J. C. Venter, J. M. Sikela, J. S. Beckmann, J. Weissenbach, R. M. Myers, D. R. Cox, M. R. James, D. Bentley, P. Deloukas, E. S. Lander, and T. J. Hudson. A Gene Map of the Human Genome. *Science*, 274(5287):540–546, 1996.
- [135] A. Sczyrba, M. Beckstette, A. Brivanlou, R. Giegerich, and C. Altmann. XenDB: Full length cDNA prediction and cross species mapping in *Xenopus laevis*. *BMC Genomics*, 6:123, 2005.
- [136] A. Sczyrba, M. Beckstette, R. Giegerich, and C. R. Altmann. Identification of 10,500 *Xenopus laevis* Full Length Clones through EST Clustering and Sequence Analysis. In *Proceedings of the German Conference on Bioinformatics, Discovery Note*. GI Lecture Notes in Informatics, 2004. Discovery Note.
- [137] Y. Shin, A. Kitayama, T. Koide, D. A. Peiffer, M. Mochii, A. Liao, N. Ueno, and K. W. Y. Cho. Identification of neural genes using *Xenopus* DNA microarrays. *Dev Dyn*, 232(2):432–444, Feb 2005.
- [138] T. Smith and M. Waterman. Identification of common molecular subsequences. *J Mol Biol*, 147(1):195–197, 1981.
- [139] M. Soares, M. Bonaldo, P. Jelene, L. Su, L. Lawton, and A. Efstratiadis. Construction and Characterization of a Normalized cDNA Library. *PNAS*, 91(20):9228–9232, 1994.
- [140] M. Spitzer, S. Lorkowski, P. Cullen, A. Sczyrba, and G. Fuellen. IsoSVM - Distinguishing isoforms and paralogs on the protein level. *BMC Bioinformatics*, 7:110, 2006.
- [141] J. E. Stajich, D. Block, K. Boulez, S. E. Brenner, S. A. Chervitz, C. Dagdigian, G. Fuellen, J. G. R. Gilbert, I. Korf, H. Lapp, H. Lehvslaiho, C. Matsalla, C. J. Mungall, B. I. Osborne, M. R. Pocock, P. Schattner, M. Senger, L. D. Stein, E. Stupka, M. D. Wilkinson, and E. Birney. The bioperl toolkit: Perl modules for the life sciences. *Genome Res*, 12(10):1611–1618, Oct 2002.
- [142] P. Steffen, B. Voss, M. Rehmsmeier, J. Reeder, and R. Giegerich. RNASHapes: an integrated RNA analysis package based on abstract shapes. *Bioinformatics*, 22(4):500–503, Feb 2006.
- [143] R. L. Strausberg, K. H. Buetow, S. F. Greenhut, L. H. Grouse, and C. F. Schaefer. The cancer genome anatomy project: online resources to reveal the molecular signatures of cancer. *Cancer Invest*, 20(7-8):1038–1050, 2002.
- [144] A. I. Su, T. Wiltshire, S. Batalov, H. Lapp, K. A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, M. P. Cooke, J. R. Walker, and J. B. Hogenesch. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A*, 101(16):6062–6067, Apr 2004.
- [145] Sun Microsystems, Inc., Santa Clara, CA 95054 U.S.A. *Sun ONE Grid Engine Administration and User's Guide*, October 2002.
- [146] M. Szymanski and J. Barciszewski. RNA regulation in mammals. *Ann N Y Acad Sci*, 1067:461–468, May 2006.
- [147] A. Tanzer and P. Stadler. Molecular evolution of a microRNA cluster. *J Mol Biol*, 339(2):327–335, 2004.

- [148] R. L. Tatusov, N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, D. M. Krylov, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya, B. S. Rao, S. Smirnov, A. V. Sverdlov, S. Vasudevan, Y. I. Wolf, J. J. Yin, and D. A. Natale. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, 4:41, Sep 2003.
- [149] T. M. P Team. The Status, Quality, and Expansion of the NIH Full-Length cDNA Project: The Mammalian Gene Collection (MGC). *Genome Res.*, 14(10b):2121–2127, 2004.
- [150] H. Temin and S. Mizutani. Viral RNA-dependent DNA Polymerase: RNA-dependent DNA polymerase in virions of Rous Sarcoma virus. *Nature*, 226(5252):1211–1213, 1970.
- [151] The International SNP Map Working Group. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, 409(6822):928–933, 2001.
- [152] D. Torney, C. Burkes, D. Davison, and K. Sirkin. Computation of d2: A measure of sequence dissimilarity. In G. Bell and T. Marr, editors, *Computers and DNA*, volume VII of *SFI studies in the sciences of complexity*. Addison-Wesley, New York, NY, 1990.
- [153] P. H. Tran, D. A. Peiffer, Y. Shin, L. M. Meek, J. P. Brody, and K. W. Y. Cho. Microarray optimizations: increasing spot accuracy and automated identification of true microarray signals. *Nucleic Acids Res*, 30(12):e54, Jun 2002.
- [154] S. Tugendreich, J. Bassett, D.E., V. McKusick, M. Boguski, and P. Hieter. Genes conserved in yeast and humans. *Hum. Mol. Genet.*, 3(90001):1509–1517, 1994.
- [155] E. Ukkonen. On-line construction of suffix trees. *Algorithmica*, 14:249–260, 1995.
- [156] F. Useche, G. Gao, M. Harafey, and A. Rafalski. High-throughput identification, database storage and analysis of SNPs in EST sequences. *Genome Inform Ser Workshop Genome Inform*, 12:194–203, 2001.
- [157] J. Usuka and V. Brendel. Gene structure prediction by spliced alignment of genomic DNA with protein sequences: increased accuracy by differential splice site scoring. *J Mol Biol*, 297(5):1075–1085, Apr 2000.
- [158] J. Usuka, W. Zhu, and V. Brendel. Optimal spliced alignment of homologous cDNA to a genomic DNA template. *Bioinformatics*, 16(3):203–211, 2000.
- [159] V. E. Velculescu, L. Zhang, B. Vogelstein, and K. W. Kinzler. Serial analysis of gene expression. *Science*, 270(5235):484–487, Oct 1995.
- [160] J. Venter *et al.* The sequence of the human genome. *Science*, 291:1304–51, 2001.
- [161] T. Watanabe, A. Takeda, K. Mise, T. Okuno, T. Suzuki, N. Minami, and H. Imai. Stage-specific expression of microRNAs during *Xenopus* development. *FEBS Lett*, 579(2):318–324, 2005.
- [162] J. D. Watson and F. H. Crick. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, Apr 1953.
- [163] P. Weiner. Linear Pattern Matching Algorithms. In *Proceedings of the 14th IEEE Annual Symposium on Switching and Automata Theory*, pages 1–11, The University of Iowa, 1973.
- [164] J. Williams. *Genetic Engineering*, volume 1, chapter The preparation and screening of a cDNA clone bank, pages 1–59. Academic Press, 1981.
- [165] A. Williamson. The Merck Gene Index project. *Drug Discov Today*, 4(3):115–122, 1999.

Bibliography

- [166] T. Wolfsberg and D. Landsman. A comparison of expressed sequence tags (ESTs) to human genomic sequences. *Nucl. Acids Res.*, 25(8):1626–1632, 1997.
- [167] T. Wolfsberg and D. Landsman. *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*, chapter Expressed Sequence Tags (ESTs), pages 283–301. Wiley-Liss, Inc., 2001.
- [168] C. V. Wright, E. A. Morita, D. J. Wilkin, and E. M. D. Robertis. The xenopus xihbox 6 homeo protein, a marker of posterior neural induction, is expressed in proliferating neurons. *Development*, 109(1):225–234, May 1990.
- [169] R. Yelin, D. Dahary, R. Sorek, E. Y. Levanon, O. Goldstein, A. Shoshan, A. Diber, S. Biton, Y. Tamir, R. Khosravi, S. Nemzer, E. Pinner, S. Walach, J. Bernstein, K. Savitsky, and G. Rotman. Widespread occurrence of antisense transcription in the human genome. *Nat Biotechnol*, 21(4):379–386, Apr 2003.
- [170] Z. Zhang, S. Schwartz, L. Wagner, and W. Miller. A greedy algorithm for aligning DNA sequences. *J Comput Biol*, 7:203–14, 2000.
- [171] W. Zhu, S. D. Schlueter, and V. Brendel. Refined Annotation of the Arabidopsis Genome by Complete Expressed Sequence Tag Mapping. *Plant Physiol.*, 132(2):469–484, 2003.
- [172] M. Zuker, D. Mathews, and D. Turner. Algorithms and Thermodynamics for RNA Secondary Structure Prediction: A Practical Guide. In J. Barciszewski and B. Clark, editors, *RNA Biochemistry and Biotechnology*, NATO ASI Series, pages 11–43,. Kluwer Academic Publishers, 1999.