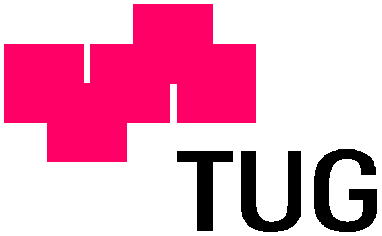
# Comparative Analysis of Human and Mouse Transcriptomes

Alexander Sturn




Doctoral Thesis

Graz University of Technology
Institute for Genomics and Bioinformatics
Petersgasse 14, 8010 Graz, Austria

Graz, February 2005

To my parents and Susanne
In gratitude

# Abstract

The functional annotation and identification of genes involved in the development and progression of complex diseases is a cumbersome and non trivial task. Exploiting the comparisons of the human genome with other genomes at both the distal and proximal evolutionary edges of the vertebrate tree is expected to represent a powerful tool in the puzzle of decoding molecular mechanisms underlying development or disease.

The main objective of this thesis was to develop a comprehensive and efficient bioinformatics platform for large-scale transcriptomic studies. It facilitates comparative analyses of human diseases and corresponding mouse models by integrating gene expression data with genome sequence information.

The specific achievements of the systematic approach represented here are threefold: First, a set of representative transcriptomic datasets describing mouse embryo fibroblasts and human multipotent adipose-derived stem cells during adipocyte differentiation has been produced, annotated, as well as stored in an organized and easily accessible way within a microarray database management system. Second, sophisticated computational tools are provided within a bioinformatics platform for large-scale comparative transcriptomic analyses to distinguish the similar from the dissimilar and to analyze these data in a straightforward, efficient, and reliable way. Several methods are proposed to derive meaningful biological information and distributed high-performance computing is used to facilitate these types of large-scale data analyses in reasonable time. Third, comparative analyses of the human and mouse datasets described above have been conducted with contingent new insights into the universality as well as the specialization between the most important model organism mouse and the designation of all clinical research, the human.

Finally and ultimately these investigations attempt to provide the research community with a markedly improved repertoire of computational tools that facilitate the translation of accumulated information from comparative transcriptomic studies into novel biological insights.

**Keywords:** comparative genomics, transcriptomics, microarray, adipogenesis, transcriptional profiling, functional annotation, cluster analysis, bioinformatics, high-performance computing.

# Publications

This thesis was based on the following publications, as well as upon unpublished observations:

## Papers

Maurer M, Molidor R, **Sturn A**, Hackl H, Hartler J, Stocker G, Prokesch A, Scheideler M, Trajanoski Z. MARS: A Microarray Analysis, Retrieval, and Storage System. BMC Bioinformatics (submitted).

Hackl H, Sanchez Cabo F, **Sturn A**, Wolkenhauer O, Trajanoski Z. Analysis of DNA Microarray Data. Curr Top Med Chem. 2004, 4(13):1355-1368.

Molidor R, **Sturn A**, Maurer M, Trajanoski Z. New trends in bioinformatics: from genome sequence to personalized medicine. Exp Gerontol. 2003 Oct;38(10):1031-6.

**Sturn A**, Mlecnik B, Pieler R, Rainer J, Truskaller T, Trajanoski Z. Client-Server Environment for High-Performance Gene Expression Data Analysis. Bioinformatics. 2003 Apr;19(6):772-773.

Saeed AI, Sharov V, White J, Li J, Liang W, Bhagabati N, Braisted J, Klapa M, Currier T, Thiagarajan M, **Sturn A**, Snuffin M, Rezantsev A, Popov D, Ryltsov A, Kostukovich E, Borisovsky I, Liu Z, Vinsavich A, Trush V, Quackenbush J. TM4: a free, open-source system for microarray data management and analysis. Biotechniques. 2003 Feb;34(2):374-8.

**Sturn A**, Quackenbush J, Trajanoski Z. Genesis: Cluster analysis of microarray data. Bioinformatics. 2002 Jan;18(1):207-8. PMID: 11836235

## Book Chapters

**Sturn A**, Maurer M, Molidor R, Trajanoski Z. Systems for Management of Pharmacogenomic Information. Pharmacogenomics: Methods and Protocols. Humana Press, Totowa, USA 2004, in press

## Conference Proceedings and Abstracts

**Sturn A**, Mlecnik B, Pieler R, Rainer J, Truskaller T, Trajanoski Z. MARS: Microarray Analysis and Retrieval System. OEGBMT Meeting, Graz, Austria. 2004 Nov 12-13.

**Sturn A**, Maurer M, Molidor R, Pieler R, Rainer J, Mlecnik B, Truskaller T, Zeller D, Trost E, Sanchez-Cabo F, Hackl H, Thallinger GG and Trajanoski Z. MARS: Microarray Analysis and Retrieval System. GEN-AU Evaluation Conference, Vienna, Austria. 2004 Nov 8-10.

Hackl H, Burkhard T, Krogsdam A, Paar Ch, **Sturn A**, Ramsauer T, Jorgensen C, Kristiansen K, Gaspard RM, Quackenbush J, Trajanoski Z. Transcriptional Profiling of Adipocyte Differentiation in 3T3-L1 Cell Line and Mouse Embryo Fibroblasts (MEFs). GEN-AU GOLD Meeting, Bad St. Leonhard, Austria. 2004 Jul 1-3.

**Sturn A**, Maurer M, Molidor R, Pieler R, Rainer J, Mlecnik B, Truskaller T, Zeller D, Trost E, Sanchez-Cabo F, Hackl H, Thallinger GG, and Trajanoski Z. MARS: Microarray Analysis and Retrieval System. Biological Discovery Using Diverse High-Throughput Data, Steamboat Springs, Colorado, USA. 2004 Mar 20 – Apr 4.

Mlecnik B, Scheideler M, Galon J, Amri Z, Maurer M, Molidor M, **Sturn A**, Prokesch A, Trajanoski Z. Immunogenomics of Adipocyte Progenitor Cells. Austrian Academy of Sciences, Vienna 2004.

Hackl H, Burkard T, Paar C, Fiedler R, **Sturn A**, Stocker G, Rubio RM, Quackenbush J, Schleiffer A, Eisenhaber A, and Trajanoski Z. Large Scale Expression Profiling and Functional Annotation of Adipocyte Differentiation. Molecular Control of Adipogenesis and Obesity, Fairmont Banff Springs , Banff, USA. 2004 Mar 4-10.

Maurer M., Molidor R., **Sturn A**. and Trajanoski Z. MARS: Microarray Analysis and Retrieval System. GEN-AU Evaluation Conference, Vienna, Austria. 2003 Oct 22-24.

Maurer M., Molidor R., **Sturn A**. and Trajanoski Z. MARS: Microarray Analysis and Retrieval System. 6th International Meeting of the Microarray Gene Expression Data Society, Aix-en-Provence, France. 2003 Sep 3-6.

Molidor R., Galon J., Hackl H., Maurer M., Ramsauer T., **Sturn A**., Trajanoski Z.: Identifying Gene Expression Patterns in Anergic T-Cells and Tumoral Micro-Environment T-Cells, Austrian Academy of Sciences, Vienna 2002.

**Sturn A**, Hackl H, Seed A, Quackenbush J, Trajanoski Z. Data visualization and analysis tool for large-scale gene expression studies. Manchester Microarray Meeting, Manchester, Great Britain. 2002 Mar 27-28.

**Sturn A**, Quackenbush J, Trajanoski Z. Data Visualization and Analysis Tool for Gene Expression Data Mining. 9th International Conference on Intelligent Systems for Molecular Biology (ISMB), Copenhagen, Denmark. 2001 Jul 21-25.

Hackl H, **Sturn A**, Michopoulos V, Quackenbush J, Trajanoski Z. Potential binding sites for PPARg in promoters and upstream sequences. 9th International Conference on Intelligent Systems for Molecular Biology (ISMB), Copenhagen, Denmark. 2001 Jul 21-25.

Saeed A, **Sturn A**, Quackenbush J. TIGR MultipleExperimentViewer: Tools for large-scale gene expression studies. MGED III Meeting, Stanford University, CA, USA. 1001 Mar 29-31.

**Sturn A**, Hackl H, Saeed A, Quackenbush J, Trajanoski Z. Data Visualization and Analysis Tool for Large-Scale Gene Expression Studies. Cambridge Healthtech Institute's 3rd Annual Integrated Bioinformatics Conference, Zurich, Switzerland. 2001 Jan 24-26.

Hackl H, **Sturn A**, Michopoulos V, Quackenbush J, Trajanoski Z. In Silico Identification of PPARg Target Genes. The Institute for Genomic Research: 4th Annual Conference on Computational Genomics , Baltimore, MD, November 2000. Journal of Computational Biology. Volume 7, Numbers 3/4, 2000: 639.

**Sturn A**, Hackl H, Saeed A, Quackenbush J, Trajanoski Z. Java Suit for Gene Expression Data Mining. The Institute for Genomic Research: 4th Annual Conference on Computational Genomics , Baltimore, MD, November 2000. Journal of Computational Biology. Volume 7, Numbers 3/4, 2000: 639.

Saeed IA, **Sturn A**, Quackenbush J. Data Visualization and Analysis Tools for High Density Microarrays. Proceedings 12th International Genome Sequencing and Analysis Conference. 2000 Sep 12-14: 105.

# Contents

# List of Figures

# List of Tables

# Glossary

| | |
|---|---|
| AA | Amino Acid |
| AAS | Authentication and Authorization System |
| ACADM | Acyl-Coenzyme A Dehydrogenase Medium |
| AG | Corporation which is limited by shares, i.e. owned by shareholders |
| AMP | Adenosine Monophosphate |
| ANOVA | Analysis Of Variance |
| API | Application Programming Interface |
| ASN | Abstract Syntax Notation |
| BAC | Bacterial Artificial Chromosome |
| BIN | Bioinformatics Integration Network |
| BLAST | Basic Local Alignment Search Tool |
| BLAT | Blast Like Alignment Tool |
| BMAP | Brain Molecular Anatomy Project |
| bp | Base Pair |
| BSA | Bovine Serum Albumin |
| CA | Correspondence Analysis |
| CAP | Contig Assembly Program |
| CD | Conserved Domain |
| CDART | Conserved Domain Architecture Retrieval Tool |
| cDNA | Complementary DNA |
| CDS | Coding Sequence |
| Cenancestor | The most recent common ancestor of the taxa under consideration |
| CF | Clustering Factor |
| COG | Clusters of Orthologous Groups |
| CPU | Central Processing Unit |
| C/EBP | CCAAT/Enhancer Binding Protein |
| DAG | Directed Acyclic Graphs |
| DB | Database |
| DBA | DNA Block Aligner |
| DBMS | Database Management System |
| DDBJ | DNA Data Bank of Japan |
| DEX | Dexamethasone |
| DMEM | Dulbecco's Modified Eagle's Medium |
| DMSO | Dimethylsulfoxide |
| DNA | Deoxyribonucleic Acid |
| DOM | Document Object Model |
| DTD | Document Type Definition |
| DTT | Dithiothreitol |

| | |
|---|---|
| **EACI** | External Application Connector Interface |
| **EBI** | European Bioinformatics Institute |
| **EGO** | Eukaryotic Gene Orthologs |
| **EJB** | Enterprise Java Bean |
| **EMBL** | European Molecular Biology Laboratory |
| **EST** | Expressed Sequence Tag |
| **FANTOM** | Functional Annotation of Mouse cDNA |
| **FBS** | Fetal Bovine Serum |
| **FGF** | Fibroblast Growth Factor |
| **FOM** | Figure of Merit |
| **FTP** | File Transfer Protocol |
| **GB** | Gigabyte |
| **GEN-AU** | Genome Research in Austria |
| **GEO** | Gene Expression Omnibus |
| **GHz** | Gigahertz |
| **GO** | Gene Ontology |
| **GOA** | GO Annotation @ EBI |
| **HCL** | Hierarchical Clustering |
| **hMADS** | Human Multipotent Adipose-derived Stem Cells |
| **HMM** | Hidden Markov Model |
| **HOC** | Human Oligonucleotide Chip |
| **Homology** | The relationship of any two characters that have descended, usually with divergence, from a common ancestral character |
| **HSP** | High Scoring Sequence Pairs |
| **HTML** | Hyper Text Markup Language |
| **HTTP** | Hypertext Transfer Protocol |
| **HTTPS** | Secure Hypertext Transfer Protocol |
| **IBMX** | Methylisobutylxanthine = MIX |
| **ID** | Identity or Identifier |
| **IIOP** | Internet Inter-Orb Protocol |
| **Inparalogs** | Paralogs in a given lineage that all evolved by gene duplications that happened after the radiation (speciation) event |
| **IPI** | International Protein Index |
| **IT** | Information Technology |
| **J2EE** | Java 2 Enterprise Edition |
| **JCS** | Java Cluster Service |
| **JDBC** | Java Database Connectivity |
| **JNDI** | Java Naming and Directory Interface |
| **JSP** | Java Server Page |
| **KMC** | k-means Clustering |
| **LIMS** | Laboratory Information Management System |
| **LPA** | Lysophosphatidic Acid |

| | |
|---|---|
| **MAGE-ML** | Microarray Gene Expression Markup Language |
| **MARS** | Microarray Analysis and Retrieval System |
| **MARS-QM** | Microarray Analysis and Retrieval System - Quality Management |
| **MB** | Megabyte |
| **MCC** | Mouse cDNA Chip |
| **MDI** | Methylisobutylxanthine + Dexamethasone + Insulin |
| **ME** | Mouse Embryos |
| **MEF** | Mouse Embryo Fibroblast |
| **MGED** | Microarray Gene Expression Data Consortium |
| **MIAME** | Minimum Information About a Microarray Experiment |
| **MIX** | Methylisobutylxanthine |
| **MMDB** | Molecular Modeling Database |
| **mRNA** | Messenger Ribonucleic Acid |
| **NCBI** | National Center for Biotechnology Information, a division of the US National Library of Medicine |
| **NHS** | N-Hydroxysuccinimide |
| **NIA** | National Institute on Aging, a division of the US National Institutes of Health |
| **NIH** | US National Institutes of Health, Bethesda, Maryland, USA |
| **NLM** | National Library of Medicine, a division of the US National Institutes of Health |
| **NQP** | NCBI Query Processor |
| **OMIM** | Online Mendelian Inheritance in Man |
| **ORF** | Open Reading Frame |
| **Orthology** | The relationship of any two homologous characters whose common ancestor lies in the cenancestor of the taxa from which the two sequences were obtained |
| **Outparalogs** | Paralogs in the given lineage that evolved by gene duplications that happened before the radiation (speciation) event. |
| **Palindrome** | Sequence of units (like a strand of DNA) which has the property of reading the same in either direction |
| **Paralogy** | The relationship of any two homologous characters arising from a duplication of the gene for that character |
| **PBS** | Portable Batch System |
| **PC** | Principal Component |
| **PCA** | Principal Component Analysis |
| **PCR** | Polymerase Chain Reaction |
| **PD** | Population Doublings |
| **PDB** | Protein Data Bank |
| **PIR** | Protein Information Resource |
| **PMT** | Photo Multiplier Voltage |
| **PPAR** | Peroxisome Proliferator Activated Receptor |
| **PRF** | Protein Research Foundation |
| **PWM** | Position Weight Matrix |
| **RAM** | Random Access Memory |

| | |
|---|---|
| **RefSeq** | Reference Sequence |
| **RMI** | Remote Method Invocation |
| **RNA** | Ribonucleic Acid |
| **RXR** | Retinoid X Receptor |
| **SAGE** | Serial Analysis of Gene Expression |
| **SD** | Standard Deviation |
| **SDS** | Sodium Dodecyl Sulfate |
| **SEG** | Low complexity region identifying program developed by Wootton and Federhen at the National Center for Biotechnology Information (NCBI) |
| **SFB** | Special Research Area |
| **SOAP** | Simple Object Access Protocol |
| **SOM** | Self Organizing Maps |
| **SREBP** | Sterol Regulatory Element Binding Protein |
| **SRS** | Sequence Retrieval System |
| **SSC** | Saline Sodium Citrate |
| **SSH** | Secure Shell, network protocol providing secure encrypted communication |
| **SVD** | Singular Value Decomposition |
| **SVM** | Support Vector Machines |
| **Syntenic** | All loci on the same chromosome |
| **Taxon** | A taxon (plural taxa) is an element of a taxonomy, e.g. in the scientific classification in biology. Taxa form a hierarchical scheme, each being broken down into subtaxa |
| **TB** | Terabyte |
| **TF** | Transcription Factor |
| **TIFF** | Tagged Image File Format |
| **TIGR** | The Institute for Genomic Research |
| **TNF** | Tumor Necrosis Factor |
| **trEST** | Translated EST |
| **TSS** | Transcription Start Site |
| **TU-Graz** | Graz University of Technology |
| **URL** | Uniform Resource Locator |
| **UTR** | Untranslated Regions |
| **WSDL** | Web Service Description Language |
| **Xenology** | The relationship of any two homologous characters whose history, since their common ancestor, involves an interspecies (horizontal) transfer of the genetic material for at least one of those characters |
| **XML** | Extensible Markup Language |
| **XSLT** | eXtensible Stylesheet Language Transformations |

# 1. Introduction

## 1.1 Background

Our understanding of the basis for human disease is evolving rapidly; we are coming to realize that traditional classification of human disease into chromosomal, monogenic (mendelian), and multifactorial categories is an oversimplification. The evolution towards seeing single-gene traits as versions of complex traits has been under way for some time, as illustrated by widely different phenotypes that can be accounted for by allelic variation in a single gene, the blurring of predicted relationships between genotype and phenotype in several monogenic disorders, or modifier genes and non-genetic factors that contribute to the phenotypes of monogenic disorders.

Disease phenotypes arise from complex interactions of organisms with their environments. While we have a long history of associating genes and gene defects with a large array of diseases, a growing body of data suggests that many disease phenotypes arise from gene-gene interactions and the interactions of genes with their environments – including the genetic background in which those genes are expressed. With the sequencing of the human genome [1,2] and the development of high-throughput technologies, we have now for the first time the means to dissect complex phenotypes and develop novel diagnostic, prognostic, therapeutic, and preventive strategies. To utilize the tremendous potential of high-throughput technologies, model organisms in general, and mouse mutant lines in particular are extremely valuable.

The human and mouse genomes are remarkably similar not only in the structure of their chromosomes but also at the level of DNA sequence. Comparison of human and mouse and human and primate expression data has been successfully performed previously and revealed surprising insights into molecular and physiological gene function, mechanisms of transcriptional regulation, and disease etiology. Hence, the appliance of large-scale comparative analysis of mouse vs. human expression data is highly valuable.

One of the major goals of postgenomic research is to relate variation in the human genome to common complex (multifactorial) diseases. Enabled by the increasing amount of public and internal available microarray studies, comparative analysis of the transcriptome of different cell types, treatments, tissues or even among two or more model organisms promise to significantly enhance the fundamental understanding of the universality as well as the specialization of molecular biological mechanisms.

### 1.1.1 Comparative Genomics

"Know then thyself, presume not God to scan, the proper study of mankind is man" wrote Alexander Pope in 1733. What better reason could there have been to sequence the human genome? But the planners of the Human Genome Project realized that the data could not be fully understood, or used to advance biomedicine, in isolation. Indeed, many of the "lessons learned and promises kept" [3] have been derived from the study of model organisms [4].

At the foundation of the evolutionary relationship of all vertebrates is conserved genetic information in the form of DNA sequence, which is assumed to underline homologous functional and anatomical similarities between species. These genetic sequences - strings of nucleotide bases - are "documents of evolutionary history" [5], from which much information can be inferred from their conservation or divergence and rearrangement relative to a common ancestor.

The sudden wealth of sequence data has allowed whole genome alignments to compare and contrast the evolution and content of available genomes. Genome analysts have applied such comparative strategies at many levels, from multi-megabase rearrangements reflected in chromosome structure down to single nucleotide changes between orthologous genes. This enabled the identification of pockets of DNA sequences conserved over evolutionary time, and such evolutionary conservation has been a powerful guide in sorting functional from non-functional DNA. [6-11].

For this reason, annotators of the human genome are increasingly exploiting comparisons with genomes at both the distal and proximal evolutionary edges of the vertebrate tree and there are good reasons to continue the endeavor to accumulate genome sequence data from the passengers of Noah's Ark. Despite the sequence similarity between primates, comparisons among members of this clade are beginning to identify primate as well as human-specific functional elements. At the distal evolutionary extreme, comparing the human genome to that of non-mammal vertebrates such as fish has proved to be a powerful filter to prioritize sequences that most probably have significant functional activity in all vertebrates [12].

Choosing species to be used in comparative genomics represents a compromise, with benefits and limitations that need to be recognized and weighed. If the compared species are too closely related, then the high degree of similarity between the orthologous sequences will obscure the functional elements within them; in contrast, if the compared species are too distantly related, then the functional elements will have diverged too much to be readily

identifiable (Figure 1.1). The principle of steering a middle course has certainly proved to be successful: comparative analyses of species that are separated by moderate evolutionary times, such as the human and mouse genomes, which diverged from each other approximately 75 million years ago, allowed much better annotation of both these genomes than would have been possible had only one been available [9,13].

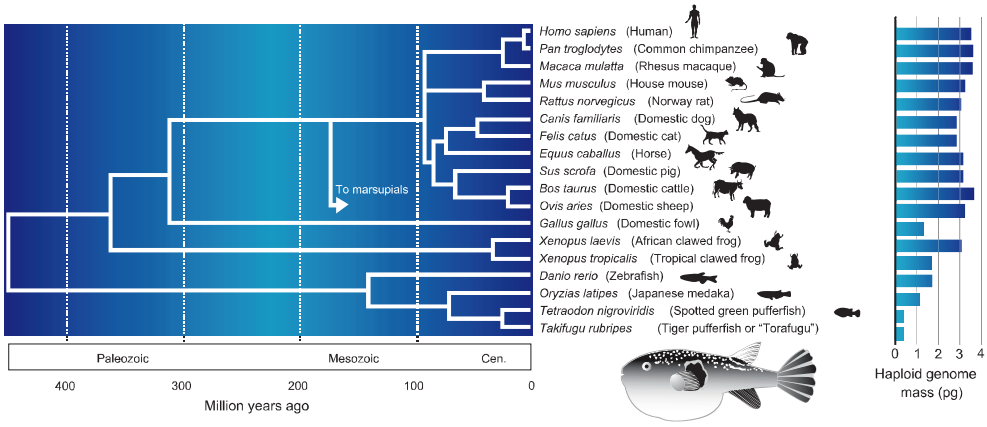



**Figure 1.1: Evolution of vertebrate genomes**: The evolutionary tree shows relationships, times of divergence, and genome sizes (in picograms of DNA, pg) of vertebras whose genomes have been selected for sequencing. Classically, 1 pg of DNA has been considered equivalent to roughly 1 billion base pairs [14].

*Mus musculus*, a species of mouse, has been one of the key model organisms sequenced since the beginnings of the Human Genome Project [15]. There can scarcely be a major area of mammalian biology or medicine to which mouse studies have not contributed in some way, often as surrogates for human studies. For genetics and development, for immunology and pharmacology, for cancer and heart disease, even for behavior, learning and memory as well as psychiatric disorders, the laboratory mouse has become a pre-eminent and indispensable tool for two fundamental reasons [16]:

First, it is a mammal, and so it has many physiological, anatomical and metabolic parallels with humans. Although the anatomical differences between humans and mice appear striking, they reflect alterations in size and shape. Detailed analysis of organs, tissues and cells reveals many similarities, extending to whole-organ systems, physiological homeostasis, reproduction, behavior and disease. The mouse is an excellent surrogate for exploring human biology and disease processes in the animal can accurately reflect those in humans. This explains why the mouse is widely used to investigate diverse aspects of mammalian biology and pathology, ranging from embryonic development to metabolic disease, behavior, and cancer in adults.

Second, although it is certainly true that mammalian biology is available in other species that in some cases are closer to humans, the mouse has one feature that has been uniquely

developed compared with all other mammals: genetic tractability. The similarities in biology and pathology between mouse and human are reflected in the genomes. For virtually every gene in the human genome, a counterpart can readily be identified in the mouse. Genetic manipulation within the living mouse is routine and can these days be done with extraordinary precision. Consequently, many mouse strains have been generated with genetic lesions that echo those observed in human genetic disease. Furthermore, this means that with having the mouse genome sequence at hand it is now truly possible to determine the function of each and every component gene by experimental manipulation and evaluation, in the context of the whole organism.

The evolutionary distance between human and mice places these species at strategic position for the identification of shared functionally conserved sequences. It has been estimated that the rate of divergence in independently evolving vertebrate genomes is on average 0,1-0,5% per million years, supporting the premise that the ~80 million years separating humans and mice from their last common ancestor is sufficient for functionally important sequences to be identified [17]. Moreover, the utility of the mouse as a model for reverse-genetic studies is immediately apparent with the revelation that 99% of mouse genes have a detectable human homolog (and vice versa), with about 80% of mouse genes having a single identifiable human ortholog [18].

It has been well established that the degree of sequence conservation is heterogeneous among different genomic segments in human and mouse. Such interspecies variation is due to wide-ranging differences in human-mouse nucleotide substitution rates across the genome. The result is a set of genomic regions with vast amount of conservation (though probably not functional) and a set lacking significant conservation (though still containing functional elements). Such an observation carries significant implications for cross-species sequence comparisons since this strategy assumes that natural selection has constrained functional sequences to evolve at slower rates than non-functional sequence [19].

The sequencing consortium was able to align long segments of mouse and human chromosomes in which the linear orders of genes in the common ancestral sequences were conserved. It is estimated that the mouse genome contains 27.000-30.500 protein-coding genes. Ninety-nine per cent of these genes have a sequence match in the human genome and 96% of these lie within syntenic regions of mouse and human chromosomes.

Based on pairwise alignments of nearly 13.000 (out of about 28.000) human-mouse orthologous gene pairs, the mouse genome sequencing consortium found that the encoded proteins had a median amino-acid sequence identity of 78,5%. In comparison, orthologous mouse and rat proteins are, on average, 97% identical [20], and a sample of human and

*Caenorhabditis elegans* (nematode) proteins have an average of 49% of their amino acids in common [21].

The comprehensiveness and precision afforded by the genome sequences will allow effective cross-reference of the locations of any genetically mapped traits in the mouse with genes in the orthologous regions of the human genome (and vice versa). This will greatly accelerate the isolation of disease genes. It will also be important for precise deletion (knockout) of mouse genes to study their functions and for targeting human sequences to their syntenic locations in the mouse genome, allowing the mice to be 'humanized' for various traits.

By cataloguing all of the orthologous proteins encoded by the human genome, one will eventually be able to move almost effortlessly back and forth between clinical observations in humans and experiments with mouse models of disease. For physiological and pharmacological studies, the rat (not the mouse) has been the long-standing model organism, primarily because of its larger size. A high-quality draft of the rat genome has just become available [22] and a proposed 'triangulation' strategy [23] should powerfully leverage the advantages of all three organisms (mice, rats and humans) for studies of human disease.

Human-Mouse sequence comparisons are thus expected to represent a powerful tool in the puzzle of the decoding of gene-regulatory sequence. Furthermore it will be possible to combine data on gene expression with data on amino-acid sequence differences to provide a complete picture of the evolutionary changes that distinguish us from other species and made us human.

### 1.1.2 Homology and Homology Subsets

With more and more complete genome sequences becoming available, the genomics community is becoming aware that 'homology' is not a sufficiently well defined term to describe the evolutionary relationships between genes. Emphasis is instead shifting towards identifying orthologs, which are evolutionary and, typically, functional counterparts in different species. Conversely, analysis of paralogs, particularly inparalogs, is important for detecting lineage-specific adaptations. This is particularly relevant for identifying functions of human genes by studying orthologs in model organisms.

Homology is the relationship of two characters that have descended, usually with divergence, from a common ancestral character. Characters can be any genic, structural or behavioral feature of an organism [24].

There are three disjoint subtypes of homology (Figure 1.2):

**Orthology** is that relationship where sequence divergence follows speciation, that is, where the common ancestor of the two genes lies in the most recent common ancestor (cenancestor [25]) of the taxa from which the two sequences were obtained [26]. This gives rise to a set of sequences whose true phylogeny is exactly the same as the true phylogeny of the organisms from which the sequences were obtained. Only orthologous sequences have this property.

**Paralogy** is defined as that condition where sequence divergence follows gene duplication [26]. Such genes might descend and diverge while existing side by side in the same lineage. Mixing paralogs with orthologous sequences can lead to a tree that has the correct phylogeny for the sequences but not for the taxa from which they derive; a gene tree is not necessarily a species tree.

**Xenology** is defined as that condition (horizontal transfer) where the history of the gene involves an interspecies transfer of genetic material [27]. It does not include transfer between organelles and the nucleus and is the only form of homology in which the history has an episode where the descent is not from parent to offspring but, rather, from one organism to another.

Relative to a given speciation event, paralogs derive either from an ancestral duplication and do not form orthologous relationships, or they derive from a lineage-specific duplication, giving rise to co-orthologous relationships. As logical terms therefore 'outparalog' and 'inparalog' have been proposed [28], explicitly denoting that they are subtypes of paralogs and when they branched relative to the given speciation event.

**Inparalogs:** paralogs in a given lineage that all evolved by gene duplications that happened after the radiation (speciation) event that separated the given lineage from the other lineage under consideration.

**Outparalogs:** paralogs in the given lineage that evolved by gene duplications that happened before the radiation (speciation) event.

Phylogenetic reconstructions of organisms created using information from the nucleotide sequences of genes require orthologous, rather than paralogous, genes, so the distinction between these two gene classes is important for practical reasons. More fascinating is the observation that orthologs and paralogs usually have very different evolutionary fates. Orthologs often take over the function of the precursor gene in the species of origin and thus

tend to be conserved. In contrast, young paralogs have redundant functions, which is an evolutionarily unstable situation. Thus, in the long run - with a few exceptions - paralogs either diverge functionally, or all but one of the versions is lost [29].

Prevalent is the romantic, but incorrect, idea that every gene in a species genome has exactly one ortholog in the genome of another species. Genes can be lost during evolution, meaning that a given gene may or may not have surviving orthologs in other species. Moreover, gene duplications can follow a speciation event, generating orthologous 'clades' of paralogs. Orthology between individual genes does therefore not exist; rather, one-to-many or many-to-many orthologous relationships are formed. The situation is further complicated by the fact that gene duplication, gene loss and speciation can be frequent events in the history of a group of organisms. Thus, complex gene relationships are established which cannot be described in simple terms.

Orthology and Paralogy differ in that one proceeds from speciation and the other from gene duplication, but either evolutionary course of divergence has the same potential for acquisition of new properties.

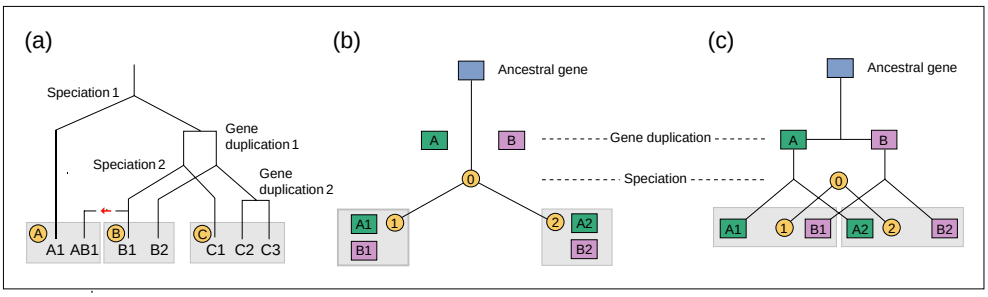



**Figure 1.2: Subtypes of homology: (a)** The idealized evolution of a gene (lines) is shown from a common ancestor in an ancestral population, descending to three populations labeled A, B and C. There are two speciation events, each occurring at the junctions shown as an upside down Y. There are also two gene-duplication events, depicted by a horizontal bar. Two genes whose common ancestor resides at a Y junction (speciation) are orthologs. Two genes whose common ancestor resides at a horizontal bar junction (gene duplications) are paralogs. Thus, C2 and C3 are paralogous to each other but are orthologous to B2. Both are paralogous to B1 but orthologous to A1. The red arrow denotes the transfer of the B1 gene from species B to species A. As a result, the AB1 gene is xenologous to all six other genes. All three subtype relationships are reflexive, that is, A1 → B1 implies B1 → A1 where → should be read, for example, as 'is orthologous to.' However, the relationships are not transitive. Thus, C2 → A1 → C3 might be true, but it is not necessarily therefore true that C2 → C3, as indeed it is not in the figure if → is read as 'is orthologous to.' A different non-transitivity occurs for 'is paralogous to' with B2 → C1 → C2 [26]. **(b,c)** Two different representations for the same evolutionary descent of an ancestral gene to paralogs and orthologs following gene duplication in species 0, and then speciation to yield species 1 and 2 [30,31]. Genes A1 and A2 are orthologs, and so are B1 and B2. A1 and B1 as well as A2 and B2 are paralogs, just as A and B were paralogs in the ancestral species.

### 1.1.2 Transcriptome Analysis (Microarrays)

Distinct profiles of gene expression mirror the complex molecular mechanisms that regulate cellular behavior during development and throughout life. Microarray techniques [32-37] using cDNAs or oligonucleotides are high throughput approaches for large-scale gene expression analysis and enable the investigation of mechanisms of fundamental processes and the molecular basis of diseases on a genomic scale.

It all began more than a quarter century ago, with Ed Southern's key insight that labeled nucleic acid molecules could be used to interrogate nucleic acid molecules attached to a solid support [38]. Today, thousands or even tens of thousands of genes can be spotted on a microscope slide and relative expression levels of each gene can be determined by measuring the fluorescence intensity of labeled mRNA hybridized to the arrays, facilitating the measurement of RNA levels for the complete set of transcripts of an organism.

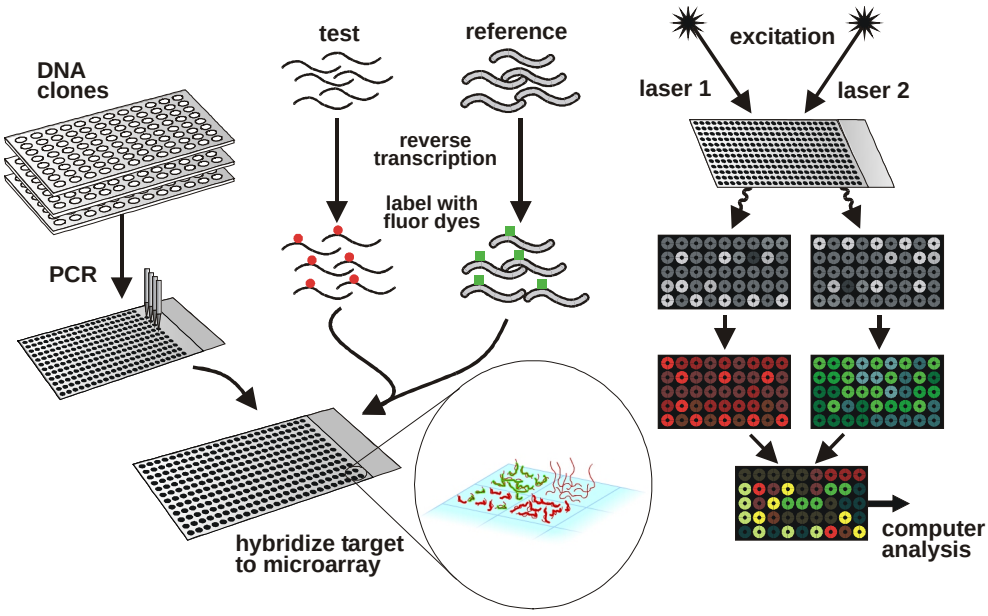



**Figure 1.3: cDNA microarray schema:** First, DNA clones are spotted onto a microscopic glass slide. After hybridization the slide is scanned using laser excitation resulting in two images as a basis for further analysis [42].

Applied to functional genetics and mutation screening, microarrays give us the opportunity to determine thousands of expression values in hundreds of different conditions [39], allowing the contemplation of genetic processes on a whole genomic scale to determine genetic contributions to complex polygenic disorders and to screen for important changes in potential disease genes [40].

cDNA microarrays exploit the preferential binding of complementary, single stranded nucleic acid sequences. Basically, a microarray is a specially coated glass microscope slide to which cDNA molecules are attached at fixed locations, called spots. With up to date computer controlled high-speed robots 30.000 and more spots can be printed on a single slide, each representing a single gene. RNA from control and sample cells is extracted. Fluorescently labeled cDNA probes are prepared by incorporating either Cy3 or Cy5-dUTP using a single round of reverse transcription, usually taking the red dye for RNA from the sample cells and the green dye for that from the control population. Both extracts are simultaneously incubated on the microarray, enabling the gene sequences to hybridize under stringed conditions to their complementary clones attached to the surface of the array (Figure 1.3) [41,42].

The production and hybridization of slides is just one pace in a pipeline of many steps necessary to gain meaningful information from microarray experiments. Because of the vast amount of data produced by a microarray experiment, sophisticated software tools are used to normalize and analyze the data [43,44].

The scanned images are analyzed using image analysis software, which evaluates the expression of a gene by quantifying the ratio of the fluorescence intensities of a spot. The quantified intensities provide information about the activity of a specific gene in a studied cell or tissue. High intensity means high activity, low intensity indicates low or no activity.

### 1.1.3 Comparative Transcriptomics

From their inception, DNA microarrays have been touted as having the potential to shed light on cellular processes by identifying groups of genes that appear to be coexpressed [45]. Because genes that encode proteins that participate in the same pathway or are part of the same protein complex are often coregulated, clusters of genes with related functions often exhibit expression patterns that are correlated under a large number of diverse conditions in DNA microarray experiments [39,46-48]. Unfortunately, coregulation does not necessarily imply that genes are functionally related. For example, cis-regulatory DNA motifs are predicted to occur by chance in the genome and might lead to serendipitous transcriptional regulation of nearby genes. In experiments limited to a single species, it would be difficult or even impossible to distinguish accidentally regulated genes from those that are physiologically important. However, evolutionary conservation is a powerful criterion to identify genes that are functionally important from a set of coregulated genes. Coregulation of a pair of genes over large evolutionary distances implies that the coregulation confers a selective advantage, most likely because the genes are functionally related. Because small and subtle changes in

fitness can confer selective advantage during evolution, the test for related gene function using evolutionary conservation in the wild is more sensitive than scoring the phenotype resulting from strong loss-of-function mutants in the laboratory. The recent availability of large sets of DNA microarray data for humans and various model organisms makes it possible to measure evolutionarily conserved coexpression on a genomewide scale [49-51].

## 1.2 Objectives

The main objective of this thesis was to develop a comprehensive, efficient, and easy to use bioinformatics platform for large-scale transcriptomic studies. It should facilitate comparative analyses of human diseases and corresponding mouse models by integrating gene expression data with genome sequence information.

The specific aims of the systematic approach represented here were threefold. First, a set of representative transcriptomic datasets should be produced, annotated, as well as stored in an organized and easily accessible way within a microarray database management system. Second, sophisticated computational tools should be provided to analyze these data in a straightforward, efficient, and reliable way. Third, comparative analyses of human and mouse cell lines should be conducted with contingent new insights into the universality as well as the specialization between the most important model organism mouse and the designation of all clinical research, the human. The objective was to develop computational tools that are able to distinguish the similar from the dissimilar among two or more large-scale data sets. Finally and ultimately these investigations should attempt to provide the research community with a markedly improved repertoire of database query and accompanying computational tools that facilitate the translation of accumulated information from comparative transcriptomic studies into novel biological insights.

Consequently, the specific aims were to compose a bioinformatics platform for large-scale comparative transcriptomics comprising the following components:

- Tools for automated, high-performance sequence retrieval.
- A fully automated pipeline for finding corresponding protein sequences to any given nucleotide sequence in a high-performant and reliable way.
- A computational pipeline for fully automated retrieval of putative orthologous and inparalogous relations between two arbitrary organisms in general and human-mouse in particular.

- A pipeline for fully automated gene ontology (GO) annotation and tools to display the annotation in context with gene expression data.

- A gene expression analysis and visualization environment providing (a) filtering and sorting of data, (b) a variety of hierarchical and non-hierarchical clustering and classification algorithms (Hierarchical Clustering (HCL), Self Organizing Maps (SOM), k-means Clustering (KMC), Principal Component Analysis (PCA), Correspondence Analysis (CA), One-Way-ANOVA, Support Vector Machines (SVM), Figure of Merit (FOM), and Gene Expression Terrain Maps), (c) a comprehensive set of similarity distance measurements, (d) mapping of gene expression data onto chromosomes, and (e) outsourcing of computational intensive calculations to in-house or remote servers.

- Tools for promoter sequence retrieval and analysis.

All tools and pipelines should be combined and integrated into one single environment for large-scale comparative transcriptomic analyses.

Distributed high-performance computing should be used to facilitate these types of large-scale data analyses in reasonable time. Protocols of communication should be chosen in a way that enables the access of the computation environment also from distant locations and through firewalls.

A major objective was to accomplish program control as well as visualization and handling of data and results in a user friendly and intuitive way. The software suite should be tailored to meet the specific needs and skills of researchers with biological or chemical but not necessarily with computer science background.

# 2 Methods

## 2.1 Biological Sequence Databases

The remarkable achievement of the complete, high-accuracy sequencing of the human genome, often compared to landing a man on the moon, lays the groundwork for a fundamental shift in how biological and biomedical research will be performed in the future. The free, widespread availability of a wide variety of data beyond the human genome sequence (sequence variation data, model organism sequence data, expression data, and proteomic data, to name a few) will provide a fertile playground for biologists in all disciplines to better design and interpret their laboratory and clinical experiments, hopefully accelerating the pace of biological discovery. Along with the data available from numerous completed model genomes, the major public databases contain a phenomenal amount of sequence data. Database efforts have kept pace with the furious rate at which sequence data is being generated, providing investigators access to all public data in a practically instantaneous fashion. There is an important distinction between primary (archival) and secondary (curated) databases. The most important contribution that the sequence databases make to the scientific community is making the sequences themselves accessible. The primary databases represent experimental results (with some interpretation) but are not a curated review. Curated reviews are found in so called secondary databases [52,53]. Combined efforts in the form of providing these curated views of the data in specialized databases have been taking place for many years now. These endeavors afford tremendous value to the biological researcher since they, in essence, reduce the massive 'sequence space' to specific, tractable areas of inquiry and, by doing so, allow for the inclusion of many more types of data than are found in the larger data repositories. These databases often provide not just sequence-based information, but additional data such as gene expression, macromolecular interactions, or biological pathway information, data that might not fit neatly onto a large physical map of a genome. Most importantly, data in these smaller, specialized databases tends to be reliably curated by experts in a particular specialty and are often experimentally-verified, meaning that they represent the best state of knowledge in that particular area. All this databases are associated with information technology in a symbiotic relationship. It is encouraging that the use of high-throughput generated data in combination with computational analysis so far has revealed a wealth of information about important biological mechanisms. While the fruits of sequencing the human genome may not be known or appreciated for another hundred years, the implications to the basic way in which medicine will be practiced in the future by employing the vast amount of available information stored in public databases is staggering [53-55].

### 2.1.1 GenBank

The GenBank® sequence database [52,56-59] is a comprehensive annotated collection of all publicly available nucleotide sequences and their protein translations. GenBank comprises DNA sequences and supporting bibliographic and biological annotation for more than 140.000 named organisms and is built and distributed by the National Center for Biotechnology Information (NCBI) [60], a division of the National Library of Medicine (NLM) [61], located on the campus of the US National Institutes of Health (NIH) [62] in Bethesda, Maryland, USA. The database incorporates also sequences submitted to the European Molecular Biology Laboratory (EMBL) Data Library (EBI, Hinxton, UK) [63-65] and the DNA Data Bank of Japan (DDBJ, Mishima, Japan) [66,67] as part of a long-standing international collaboration (International Nucleotide Sequence Database Collaboration [68]) between the three databases in which data is exchanged daily to ensure a uniform and comprehensive collection of sequence information.

Today, GenBank (release 145.0) includes more than 44 billion base pairs from over 40 million reported sequences and continues to grow at an exponential rate, doubling every 10 months (Figure 2.1). Presently, all records in GenBank are generated from direct submissions to the DNA sequence databases from the original authors, who volunteer their records to make the data publicly available or do so as part of the publication process.

GenBank files are grouped into divisions; some of these divisions are phylogenetically based, whereas others are based on the technical approach that was used to generate the sequence information. A record within GenBank represent, in most cases, single, contiguous stretches of DNA or RNA with annotations. Each entry includes a concise description of the sequence, the scientific name and taxonomy of the source organism, bibliographic references and a table of features [69] listing areas of biological significance, such as coding regions and their protein translations, transcription units, repeat regions, and sites of mutation or modification.

Every GenBank record, consisting of both a sequence and its annotations, is assigned a stable and unique identifier, the accession number, which remains constant over the lifetime of the record even when there is a change to the sequence or annotation. The DNA sequence within a GenBank record is also assigned a unique identifier, called a 'GI', that appears on the VERSION line of GenBank flatfile records following the accession number. A third identifier of the form 'accession.version', also displayed on the VERSION line of flatfile records, consolidates the information present in the GI and accession numbers. When a change is made to a sequence given in a GenBank record, a new GI number is issued to the sequence and the version extension of the 'accession.version' identifier is incremented.

NCBI distributes the GenBank releases in the traditional flatfile format as well as in the Abstract Syntax Notation (ASN.1) format used for internal maintenance. The full bimonthly GenBank release and the daily updates, which also incorporate sequence data from EMBL and DDBJ, are available by anonymous FTP from the NCBI at ftp.ncbi.nih.gov.

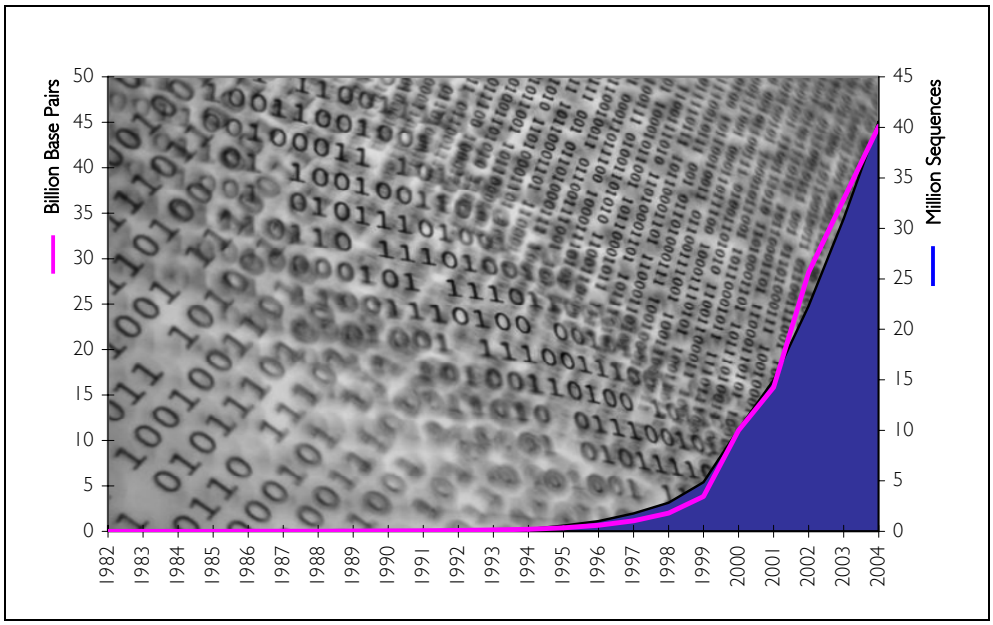



**Figure 2.1: GenBank growth:** From 1982 to the present, the number of base pairs in GenBank has doubled approximately every 14 months. Today, GenBank (release 145.0) includes more than 44 billion base pairs from over 40 million reported sequences.

The DDBJ/EMBL/GenBank database is the most commonly used nucleotide and protein sequence database. It is the primary database for, or at least important to, all databases described in the following chapters.

### 2.1.2 RefSeq

Whereas GenBank is an archival repository of all sequences, the goal of the NCBI Reference Sequence (RefSeq) project [70-74] is to provide the single best non-redundant and comprehensive collection of naturally occurring biological molecules, representing the central dogma. Nucleotide and protein sequences are explicitly linked on a residue-by-residue basis in this collection. The RefSeq database aims to provide a comprehensive, integrated, non-redundant collection of reference standards that includes chromosomes, complete genomic molecules (organelle genomes, viruses, plasmids), intermediate assembled genomic contigs, curated genomic regions, mRNAs, RNAs, and proteins. NCBI provides RefSeqs for taxonomically diverse organisms including eukaryotes, bacteria, and viruses. Additional records are added to the collection as data become publicly available.

RefSeq standards serve as the basis for medical, functional, and diversity studies; they provide a stable reference for gene identification and characterization, mutation analysis, expression studies, polymorphism discovery, and comparative analyses. RefSeqs are used as a reagent for the functional annotation of some genome sequencing projects, including those of human and mouse.

Each RefSeq represents a single, naturally occurring molecule from a particular organism. RefSeqs are frequently based on GenBank records but differ in that each RefSeq is a synthesis of information, not a piece of primary research data in itself. All RefSeq records include attribution to the original sequence data. When a molecule is represented by multiple GenBank sequences, an effort is made to select the 'best' sequence to instantiate as a RefSeq. The goal is to avoid mutations, sequencing errors and cloning artifacts. The RefSeq collection may include alternatively spliced transcripts that share some identical exons, or identical proteins expressed from these alternatively spliced transcripts, close paralogs or homologs.

RefSeq records are provided using several processes:

1. Entrez Genomes processing provides genomic, RNA, and protein records for numerous organisms as data becomes available. This pipeline provides all of the bacterial, viral, organelle, and plasmid RefSeq records and also provides some of the records for larger genomes including plants and fungi.

2. The NCBI Annotation process, an automated computational method, provides intermediate assembled contigs and some records representing potential transcripts and proteins.

3. The LocusLink-supported RefSeq pipeline supplies genomic regions, RNA, and protein sequence records for which a significant level of additional curation and review effort is provided.

4. Collaboration. Both the Entrez Genomes and LocusLink supported pipelines include collaborations that supply a fully annotated genome, gene family, or single gene. Collaborations with official nomenclature groups and organism-specific database groups are also established.

RefSeq records include annotation that provides a general indication of their curation status and reliability. Records generated via the automated annotation pipeline are annotated as 'Model RefSeq' whereas those contributed by the curation pipelines may be annotated as provisional (not yet reviewed), predicted (transcript or protein is not fully supported), or

reviewed (Table 2.1). Reviewed records are the most highly curated and effort has been made to ensure the quality and comprehensive coverage of the sequence itself as well as to apply descriptive information that leads the user to functionally relevant data (publications, names, summaries). Similar to a review article in the literature, a RefSeq is an interpretation by a particular group at a particular time and represents a compilation of our current knowledge of a gene and its transcripts.

| Accession prefix | Molecule | Pipeline | Status category |
|---|---|---|---|
| NT_ | Genomic | Computed annotation | Model |
| NW_ | Genomic | Computed annotation | Model |
| XM_ | mRNA | Computed annotation | Model |
| XR_ | RNA | Computed annotation | Model |
| XP_ | Protein | Computed annotation | Model |
| NC_ | Genomic | Entrez genomes | Provisional, Reviewed |
| NG_ | Genomic | LocusLink | Provisional, Reviewed |
| NM_ | mRNA | LocusLink, Entrez genomes | Provisional, Predicted, Reviewed |
| NR_ | RNA | LocusLink | Provisional, Reviewed |
| NP_ | Protein | LocusLink, Entrez genomes | Provisional, Predicted, Reviewed |

Table 2.1: RefSeq accession prefixes, molecule types, originating pipeline, and annotated status categories [72].

RefSeqs can be retrieved in several different ways: by searching the Entrez nucleotide or protein database, by BLAST searching, by FTP, or through links from other NCBI resources.

### 2.1.3 Ensembl

Ensembl [75-79] is a joint project between EMBL-EBI and the Sanger Institute to develop a software system which produces and maintains automatic annotation on metazoan genomes. Ensembl annotates known genes and predicts new ones, with functional annotation from InterPro, OMIM, SAGE and gene families. It is a comprehensive source of stable automatic annotation of individual genomes as well as of the synteny and orthology relationships between them. Additionally, Ensembl includes confirmed gene predictions that have been integrated with external data sources.

The Ensembl site is one of the leading sources of human genome sequence annotation and provided much of the analysis for publication of the draft genome by the international human genome project.

With a total of currently fourteen genome sequences available from Ensembl and more genomes to follow, recent developments have focused mainly on closer integration between

genomes and external data. Therefore, it is also a framework for integration of any biological data that can be mapped onto features derived from the genomic sequence. Ensembl's aims are to continue to "widen" the biological integration to include other model organisms relevant to understand human biology as they become available, to "deepen" the integration to provide an ever more seamless linkage between equivalent components in different species, and to provide further classification of functional elements in the genome that have been previously elusive.

Ensembl is nowadays a comprehensive source of stable automatic annotation of human, mouse, and other genome sequences and also integrates manually annotated gene structures from external sources where available. With both human and mouse genome sequences available and more vertebrate sequences to follow, many of the recent developments in Ensembl have also been focusing on developing automatic comparative genome analysis and visualization.


**Figure 2.2: Ensembl overview.** (left) Ensembl services. (right) Ensembl pipeline: Ensembl is fed raw DNA sequence taken from public DNA databases and puts it into a large tracking database (the "Ensembl" database). The pipeline joins the sequences into their proper place in the genome and automatically finds genes and other features in theses sequences. The results are presented free of charge to the research community via an interactive website, web services, and a set of flat files.

As well as being one of the leading sources of genome annotation, Ensembl is also an open source software engineering project to develop a portable system able to handle very large genomes and associated requirements from sequence analysis to data storage and visualization. Due to this attitude, Ensembl is available not only as an interactive website, a set of flat files, and via web services, but also as a complete, portable open source software system. The facilities of the system range from sequence analysis to data storage and visualization. All data are provided without restriction, and the source code is freely available. This culture of openness enabled that Ensembl installations exist around the world both in companies and at academic sites on machines ranging from laptops to supercomputers.

### 2.1.4 trEST

TrEST (translated EST) [80-82] is a regularly generated database of hypothetical protein sequences predicted from expressed sequence tag (EST) sequences. It is an attempt to produce contigs from clusters of ESTs and to translate them into proteins. This is a three-step process:

1. ESTs are grouped into clusters that correspond to a single transcript. UniGene clusters are used for that purpose if available; otherwise the clustering is performed using trEST in-house software.

2. ESTs of one cluster are assembled into one or several contigs by appliance of a script by Christian Iseli, that makes use of the contig assembly programs Phrap and CAP [83]. Frequently, more than one contig is produced from a single cluster and these contigs can be either disjoint or overlapping. In the latter case, they can either describe splice variants of reflect ambiguities in the contig assembly process.

3. Detection of the coding regions in the assembled contigs and translation of these regions into protein sequences is performed by the program ESTscan [84].

TrEST entries receive an accession number derived from that of the parent UniGene cluster, supplemented with a number which accounts for the different reconstructed contigs.

The program ESTscan was written to recognize the coding frame in ESTs, and a fortiori in assembled contigs. It makes use of the bias that exists in the distribution of hexanucleotides along the three reading frames of DNA, to detect which is the coding one. ESTscan has to be trained on a set of trusted sequences before being able to produce reliable predictions. These training data are the same table as for the program Genscan [85] and are available for the human, the mouse, and the rat. ESTscan corrects most frame shift errors and predicts their position with an error of a few amino acids. Benchmark experiments have indicated that ca. 95% of true coding regions longer than 30 amino acids are detected.

It must be stressed that trEST entries are NOT real protein sequences. They are hypothetical and are known to contain errors. These data is provided because it is believed that they might help biologists to find which UniGene cluster(s) may be relevant for their work.

### 2.1.5 UniGene

UniGene [86-91] is a system for automatically partitioning GenBank sequences, including ESTs, into a non-redundant set of gene-oriented clusters. UniGene starts with entries in the appropriate organism division of GenBank, combines these with ESTs of that organism and creates clusters of sequences that share virtually identical 3' untranslated regions (3' UTRs). 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.

In addition to sequences of well-characterized genes, hundreds of thousands of novel expressed sequence tag sequences have been included. In the human UniGene database, over 3,6 million human ESTs in GenBank have been reduced 35-fold in number to over 104.000 sequence clusters. Consequently, the collection may be of use to the community as a resource for gene discovery. UniGene has also been used by experimentalists to select reagents for gene mapping projects and large-scale expression analysis.

For each nucleotide sequence in UniGene, a search is made for sequence similarity to known proteins from several organisms. This is done using Blastx. Blastx compares the six-frame conceptual translation products of a nucleotide query sequence (both strands) against a protein sequence database. Blastx has 'in-frame' gapped alignments and uses sum statistics to link alignments from different frames. The protein databases exclude mitochondrial proteins and are screened for redundancy. The nucleotide sequence is considered to match the protein sequence if the Blastx E-value is less than $10^{-6}$. For each of the databases, the best-hit is that alignment with the lowest bit score.

Currently, sequences from human, rat, mouse, cow, zebrafish, clawed frog, fruitfly, mosquito, wheat, rice, barley, maize, cress and others have been processed. These species were chosen because they have the greatest amounts of EST data available and represent a variety of species. Additional organisms may be added in the future.

It should be noted that the procedures for automated sequence clustering are still under development and the results may change from time to time as improvements are made. It should also be noted that no attempt has been made to produce contigs or consensus sequences. There are several reasons why the sequences of a set may not actually form a single contig. For example, all of the splicing variants for a gene are put into the same set. Moreover, EST-containing sets often contain 5' and 3' reads from the same cDNA clone, but these sequences do not always overlap.

UniGene databases are updated weekly with new EST sequences, and bimonthly with newly characterized sequences. UniGene clusters may be searched in several ways: by gene name, chromosomal location, cDNA library, accession number, and ordinary text words. Cluster sequences may also be downloaded by FTP.

### 2.1.6 Fantom II

Even the genomics revolutionaries realize that the challenges only begin as the final base pair of each organism is sequenced. The genes must be identified and their functions determined and placed in the context of essential metabolic processes. In eukaryotes, even finding the genes remains a significant challenge. Although gene prediction programs have advanced significantly in recent years, the best evidence for the actual structure of a gene remains the sequence of a full-length cDNA clone. The most concerted effort to generate such a resource has been carried out in the mouse by collaborators and colleagues at the Riken Genomic Sciences Center in Tskuba, Japan [92].

The RIKEN Mouse Gene Encyclopedia Project [93-95] is a large-scale effort to collect full-length enriched mouse cDNA clones from various tissues, determine the full-length nucleotide sequence, infer their chromosomal locations by computer, and characterize gene expression patterns.

The FANTOM (Functional Annotation of Mouse) DB [96,97] consists of functional annotations for the RIKEN full-length mouse cDNA clones. It provides not simply sequencing analysis results and curated functional annotation information themselves, but also many other informative and useful data describing functional information about the used clones, including Gene Ontology terms. The FANTOM DB has been developed to facilitate the Gene Encyclopedia Project and to facilitate functional genomic studies such as positional candidate cloning, cDNA microarrays, and protein interaction analyses.

FANTOM DB 2 was outcome of the FANTOM 2 collaborative research and contains rich curated functional annotation for 60.770 full length sequenced RIKEN cDNA clones. It consists of 21.076 FANTOM 1 clones and 39.694 additionally sequenced clones. This covered 90% of genes discovered at release date and is the largest and highest quality data set of a mammalian transcriptome. The set includes also many non-protein-coding genes.

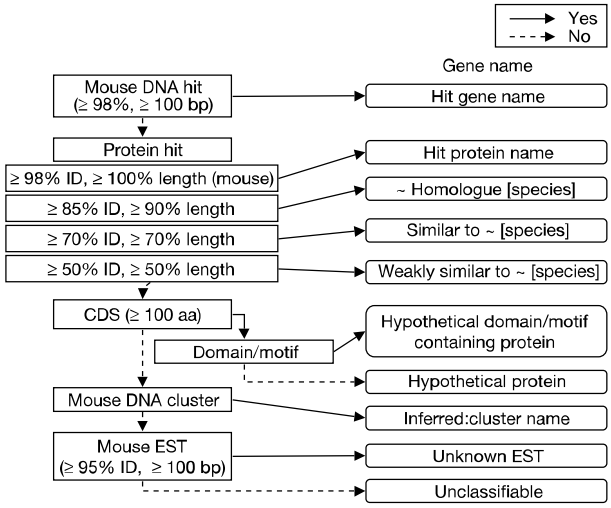



**Figure 2.3: Flow chart of the FANTOM 2 gene-name annotation pipeline:** Nomenclature and criteria used in the gene name are described on the right-hand side. Gene names were annotated (right-hand side) from the gene function descriptor of the reference (left-hand side) according to the nomenclature. Priority was given to reference descriptors from which functional information could be inferred, even if references with less informative descriptors were more similar to the clones. CDS…Coding Sequence, ID…Identity, aa…amino acid, bp…base pair, EST…Expressed Sequence Tag [93].

Functional annotation was carried out using many qualifiers (Figure 2.3). The major advantages of the annotation in FANTOM 2 are as follows:

- Gene names were decided according to the improved pipeline for FANTOM 2. Detailed evidence for gene name annotation was also recorded in database.

- Coding sequence (CDS) regions were determined by human curation. Not only start or stop position of CDS, but also the statuses of clone problems were described.

Mouse has been used in the biomedical field as the best human disease model. The RIKEN mouse transcriptome, supported by all this physical clones, provides the most comprehensive knowledge of gene function. It will contribute to the research for causes of cancer and other common diseases and also contribute to pharmaceutical drug discoveries.

### 2.1.7 International Protein Index (IPI)

Despite the complete determination of the genome sequence of several higher eukaryotes, their proteomes remain relatively poorly defined. Information about proteins identified by different experimental and computational methods is stored in different databases, meaning that no single resource offers full coverage of known and predicted proteins. IPI (the International Protein Index) [98,99] has been developed to address these issues and offers species-specific, complete and nonredundant data sets representing the human, mouse, and rat proteomes, built from the Swiss-Prot [100-103], TrEMBL [102-104], Ensembl, and RefSeq databases. Its sequence- and identifier- based construction eliminates the need for manual filtering of redundant results in protein identification, while maintaining cross-references to the source data.

An automatic and pragmatic approach is taken to assemble IPI data sets: Clusters are built through combining knowledge already present in the primary data sources (and in the cross-references between them) with the results of protein sequence similarity comparisons. After a cluster is assembled, a master entry from among the cluster members is chosen, which supplies the IPI entry with its sequence and annotation. Finally, an identifier is chosen for each cluster.

Each source database identifier can only appear once in an IPI set. Separate sequences are provided for alternatively spliced isoforms (with cross-references to isoform-specific identifiers in the source databases). Proteins with identical sequence but differential post-translational modification are not yet individually represented within IPI, as these are not yet generally well identified in the source databases.

Recapitulatory it can be said, that IPI provides a top level guide to the main databases that describe the human, mouse, and rat proteomes: UniProt (Swiss-Prot plus TrEMBL), RefSeq and Ensembl. IPI…

1. Effectively maintains a database of cross references between the primary data sources.
2. Provides minimally redundant yet maximally complete sets of human, mouse, and rat proteins (one sequence per transcript).
3. Maintains stable identifiers (with incremental versioning) to allow the tracking of sequences in IPI between IPI releases.

IPI is updated monthly in accordance with the latest data released by the primary data sources.

### 2.1.8 NCBI Entrez Proteins

Entrez [105-108] integrates the scientific literature, DNA and protein sequence databases, 3D protein structure and protein domain data, population study datasets, expression data, assemblies of complete genomes, and taxonomic information into a tightly interlinked system. The NCBI Entrez Protein database contains sequence data of the translated coding regions from DNA sequences in GenBank, EMBL, and DDBJ, as well as protein sequences submitted to Protein Information Resource (PIR), SWISS-PROT, Protein Research Foundation (PRF), and Protein Data Bank (PDB) (sequences from solved structures).

What makes Entrez more powerful than many services is that most of its records are linked to other records, both within a given database (such as Nucleotide) and between databases (Figure 2.4). Links within a database are called "neighbors" (e.g., Nucleotide neighbors).

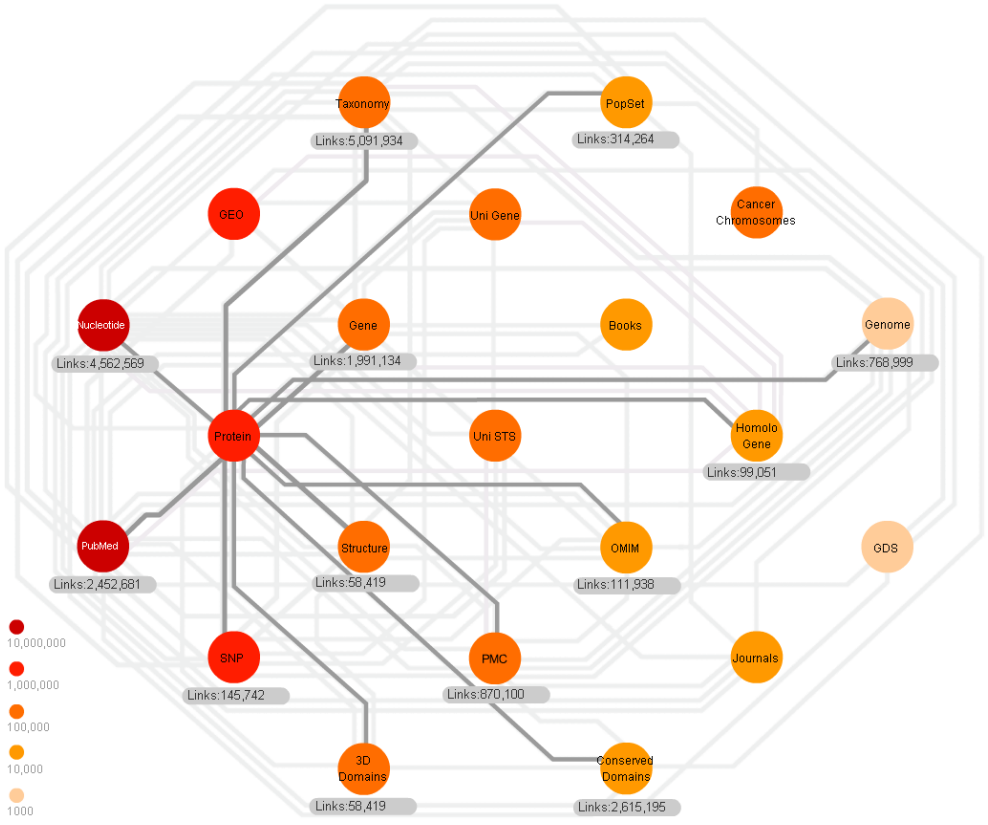



**Figure 2.4: Entrez databases and the connections between them:** Each database is represented by a colored circle, where the color indicates the approximate number of records in the database. Databases linked to the Entrez protein database are selected and the numbers of links that exist between those databases are displayed. The Entrez Protein database contains currently (January 2005) almost 5.7 million proteins.

Protein and Nucleotide neighbors are determined by performing similarity searches using the BLAST algorithm to compare the entry amino acid or DNA sequence to all other amino acid or DNA sequences in the database. Protein sequence records are linked to the nucleotide sequence from which the protein was translated. Nucleotide sequence records in the Nucleotide database are linked to the PubMed citation of the article in which the sequences were published.

## 2.2 Sequence Retrieval

The current data explosion is intractable without advanced data management systems. The integration of heterogeneous databases is also critically important. The numerous data sets become really useful when they are interconnected under a uniform interface, representing the domain knowledge.

### 2.2.1 The Entrez Search and Retrieval System

Entrez [105-108] is a robust and flexible database search and retrieval system used at the NCBI for all major databases and provides accesses to DNA and protein sequence data from more than 130.000 organisms, along with 3D protein structures, genome mapping data, population sets, phylogenetic sets, environmental sample sets, gene expression data, the NCBI taxonomy, protein domain information, protein structures from the Molecular Modeling Database MMDB [109], MEDLINE references via PubMed, and more. Entrez is at once an indexing and retrieval system, a collection of data from many sources, and an organizing principle for biomedical information. Two unique features of Entrez are:

1. Pre-computed similarity searches for each database record, identifying the related records ("neighbors") within that database. The algorithm used to identify related records depends upon the database.

2. Links from a record in one database to associated records in the other Entrez databases, providing integrated access across the various databases. For example, if a MEDLINE record cites a GenBank nucleotide sequence record, which in turn is linked to a protein translation, then there will be a link between those three records.

A Data Model provides a schematic illustration of the connections between the many data types in Entrez (Figure 2.4).

Entrez can be searched with a wide variety of text terms such as author name, journal name, gene or protein name, organism, unique identifier (e.g. accession number, sequence ID or PubMed ID), and other terms, depending on the database being searched.

The Entrez sequence databases are taken from a variety of sources - including GenBank, EMBL, DDBJ, RefSeq, PIR-International, PRF, Swiss-Prot, as well as PDB - and therefore include more sequence data than are available within the GenBank DNA sequence database alone.

External resources can be linked to Entrez records using the Linkout service. Entrez also allows users to store search strategies and select a customized subset of LinkOut links through the NCBI Cubby service.

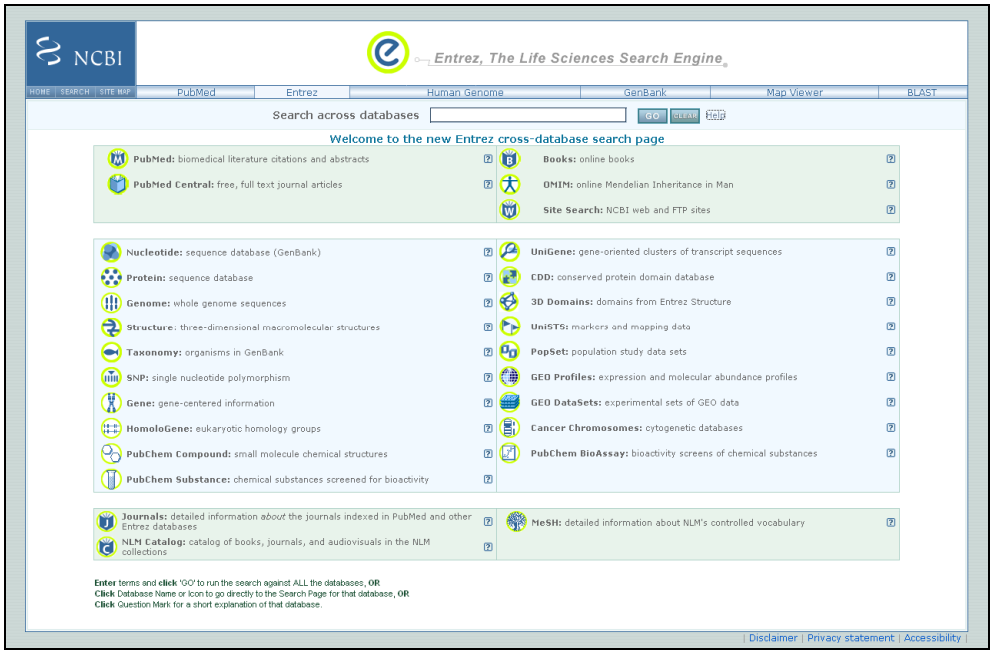



**Figure 2.5. Entrez Query Interface** showing the cross-database search engine with links to the 27 Entrez databases covered.

Entrez Is a Discovery System. A data-retrieval system succeeds when you can retrieve the same data you put in. A discovery system is intended to let you find more information than appears in the original data. By making links between selected nodes and making computed associations within the same node, Entrez is designed to infer relationships between different data that may suggest future experiments or assist in interpretation of the available information, although it may come from different sources.

The records retrieved by an Entrez search can be displayed in a wide variety of formats and downloaded singly or in large batches. Formatting options vary for records of different types. For example, display formats for GenBank records include the GenBank flatfile, FASTA, XML, ASN.1, and others. Graphical display formats are offered for some types of records, including genomic records.

#### Batch Entrez

Batch Entrez allows researchers and bioinformatics people to retrieve a large number of nucleotide sequences or protein sequences from Entrez, in a batch mode, by importing a file containing a list of the desired GI or accession numbers. Search results are saved directly to a local disk file on the computer sending the request.

### 2.2.2 Sequence Retrieval System (SRS)

The Sequence Retrieval System (SRS) [110-113] from LION bioscience AG [114] is a proven, scalable and robust data integration platform that provides fast access to diverse life science data - genetic, protein, cellular, molecular and clinical - from public and proprietary sources, regardless of the data format. SRS facilitates the rapid development of applications and algorithms, as well as bioinformatics portals for the Internet or Intranet, making the data efficiently available to entire organizations. It has been developed into an integration system for both data retrieval and sequence analysis applications. It provides capabilities to search multiple datasets by shared attributes and to query across databases fast and efficiently. LION bioscience actively encourages the use of SRS in the academic community by making the core product, SRS, freely available to non-profit organizations. That is why SRS gained impressive popularity in the bioinformatics community and is used in over 300 commercial and academic sites, delivering genomic data daily to thousands of users around the globe.

SRS is designed to retrieve data directly from text files. Text files are still widely used to exchange and distribute information. In fact, formatted text files are the *de-facto* standard for biological databases. While relational database management systems (RDBMS) are highly advanced for data management, SRS has advantages as a retrieval system: First, it is much faster (10-100 times) than retrieving whole records from large databases with complex data schemas. Second, since it retrieves data directly from flat files, it is less demanding in terms of storage space requirements than RDBMS tables. The average difference of 2-5 times is significant in the case of large databases with various GB of storage requirements. Third, it is reasonable easy to integrate new data with basic retrieval capabilities and extend it further to a more sophisticated data schema. The integrating power of SRS benefits from sharing the

definitions of conceptually equal attributes amongst different data sets. This enforces uniformity and allows multiple database queries. Searchable links between databases and customizable data representation are original features of SRS.

Data becomes more valuable in the context of other data. Besides enriching the original data by providing HTML linking, one of the original features of SRS is the ability to define indexed links between databases. These links reflect equal values of named entry attributes in two databases. The links are bi-directional, operate on sets of entries, can be weighted and can be combined with logical operators (AND, OR and NOT). This is analogous to a table of relations in a relational database schema that follows querying of one table with conditions applied to others. The user can search not only the data contained in a particular database but also any conceptually related databases and then link to the desired data. Using the linking graph, SRS makes it possible to link databases that do not contain direct references to each other. Highly cross-linked data sets become a kind of domain knowledge base.


**Figure 2.6: Sequence Retrieval System (SRS) Interface** from LION bioscience AG. Displayed are the databases available through the SRS server of the Institute for Genomics and Bioinformatics, Graz University of Technology. This SRS is installed on a 48 processor (2.6 GHz Pentium 4) Linux cluster with an attached one Terabyte high-performance disk array for data storage. This enables high-performance indexing of databases and the very efficient retrieval of large batches of sequences.

The introduction of the biosequence object in SRS allows the integration of various sequence analysis tools such as FASTA [115,116], or CLUSTALW [117]. This integration allows the treatment of text output of these applications like any other database. This enables linking to other databanks and user-defined data representations. By expanding in this direction SRS becomes not only a data retrieval system but also a data analysis application server. Recent advances in application integration include different levels of user control over application parameters, support for different UNIX queuing systems, and parallel threading.

SRS provides flexible and up-to-date access to many major databases in the field of molecular biology, produced and maintained at various institutions around the globe. The databases are grouped in specialized sections, which include for instance nucleic acid and protein sequences, mapping data, macromolecular structures, sequence variations, protein domains and metabolic pathways (Figure 2.6). SRS servers contain a selection of today more than 1000 available biological databases and can integrate almost 200 different applications. SRS is a constantly evolving system. New databases are being added and the interfaces to the old ones are always being enhanced. Most institutions maintain a robust hardware solution for running SRS that guarantees optimal uptime and load tolerance in a wide variety of usage conditions.

# 2.3 Sequence Analysis Tools

## 2.3.1 Repeat Masker

In the mid 1960's scientists discovered that genomic DNA of higher organisms contains stretches of highly repetitive DNA sequences of various sorts. These include many short repeated nucleotides (e.g. a dozen or more consecutive CA dinucleotides), longer tandem repeats, and various classes of short and long "interspersed" repeats scattered throughout the genome with various frequency. In general, these sequences are characterized and placed into five categories:

1. Simple Repeats - Duplications of simple sets of DNA bases (typically 1-5 bp) such as A, CA, CGG, etc.
2. Tandem Repeats - Typically found at the centromeres and telomeres of chromosomes these are duplications of more complex sequences with 100-200 bases.
3. Segmental Duplications - Large blocks of 10-300 kilobases that have been copied to another region of the genome.

4. **Processed Pseudogenes** - Pseudogenes are sequences of genomic DNA with such similarity to normal genes that they are regarded as non-functional copies or close relatives of genes.
5. **Transposons** – Transposons, "jumping genes" or "transposable genetic elements" are sequences of DNA that can move around to different positions within the genome of a single cell.

Currently up to 50% of the human genome is repetitive in nature and as improvements are made in detection methods this number is expected to increase.

Besides the obvious desire to label all biological relevant portions of a sequence, identification and subsequent masking of repeat regions and other regions of low sequence complexity is critical to avoid false homology reports in sequence database searches.

RepeatMasker [118] is a program that screens DNA sequences for interspersed repeats and low complexity DNA sequences. The output of the program is a detailed annotation of the repeats that are present in the query sequence as well as a modified version of the query sequence in which all the annotated repeats have been masked (default: replaced by Ns). Sequence comparisons in RepeatMasker are performed by the program cross_match [119], an efficient implementation of the Smith-Waterman-Gotoh [120,121] algorithm developed by Phil Green at the University of Washington.

The RepeatMasker program performs a useful function in the analysis of DNA sequences at the genomic level. The masking of repetitive or low-complexity sequences prior to performing homology-based screens helps to greatly reduce the number of spurious results that are obtained. The masked sequence and alignment files are also useful as an aid in the selection of primers for further experimentation when avoiding low-complexity DNA or repetitive sequences is desirable.

### 2.3.2 BLAST

A sequence itself is not informative, it must be analyzed by comparative methods against existing databases to develop hypothesis concerning relatives and function. Sequence alignments provide a powerful way to compare novel sequences with previously characterized genes. Both functional and evolutionary information can be inferred from well designed queries and alignments. BLAST® (Basic Local Alignment Search Tool) [122-126] is a set of similarity search programs designed to explore all of the available sequence databases

regardless of whether the query is protein or DNA. The BLAST programs have been designed for speed, with a minimal sacrifice of sensitivity to distant sequence relationships. The scores assigned in a BLAST search have a well-defined statistical interpretation, making real matches easier to distinguish from random background hits. BLAST uses a heuristic algorithm that seeks local as opposed to global alignments and is therefore able to detect relationships among sequences that share only isolated regions of similarity. This enables for instance to detect regions of similarity embedded in otherwise unrelated proteins. Both types of similarity may provide important clues to the function of uncharacterized proteins.

The BLAST programs introduced a number of refinements to database searching that improved overall search speed and put database searching on a firm statistical foundation. One innovation introduced in BLAST is the idea of *neighborhood words*. Instead of requiring words to match exactly, a word hit is achieved if the word taken from the subject sequence has a score of at least *T* when a comparison is made using a substitution matrix to the word from the query. This strategy allows the word size (*W*) to be kept high (for speed) without sacrificing sensitivity. Thus, *T* becomes the critical parameter determining speed and sensitivity and *W* is rarely varied. If the value of *T* is increased, the number of background word hits will go down and the program will run faster. Reducing *T* allows more distant relationships to be found.

The occurrence of a word hit is followed by an attempt to find a locally optimal alignment whose score is at least equal to a score cutoff *S*. This is accomplished by iteratively extending the alignment both left and right, with accumulation of incremental scores for matches, mismatches, and the introduction of gaps. In practice, it is more convenient to specify an expectation cutoff *E*, which the program internally converts to an appropriate value of *S* (which would depend on the search context). In regions where matching residues are scare, the cumulative score will begin to drop. As the mismatch and gap penalties mount, it becomes less likely that the score will rebound and ultimately reach *S*. This observation provides the basis for an additional heuristic whereby the extension of a hit is terminated when the reduction on sore (relative to the maximum value encountered) exceeds the score drop-off threshold *X*. Using smaller values of *X* improves performance by reducing the time spent on unpromising hit extensions, at the expense of occasionally missing some true alignments [52].

There are several variants of BLAST, each distinguished by the type of sequence (DNA or protein) of the query and database sequences (Tables 2.2 & 2.3).

| Length ' | Database | Purpose | Program |
|---|---|---|---|
| 20 bp or longer<br>28 bp or above for megablast | Nucleotide | Identify the query sequence | Discontiguous megablast, megablast, or blastn |
| | | Find sequences similar to query sequence | discontiguous megablast or blastn |
| | | Find similar sequence from the Trace archive | Trace megablast, or Trace discontiguous megablast |
| | | Find similar proteins to translated query in a translated database | Translated BLAST (tblastx) |
| | Peptide | Find similar proteins to translated query in a protein database | Translated BLAST (blastx) |
| 7 - 20 bp | Nucleotide | Find primer binding sites or map short contiguous motifs | Search for short, nearly exact matches (blastn) |

Table 2.2: BLAST program selection for nucleotide queries: The appropriate selection of a BLAST program for a given search is influenced by the following three factors (1) the nature of the query, (2) the purpose of the search, and (3) the database intended as the target of the search.

| Length ' | Database | Purpose | Program |
|---|---|---|---|
| 15 residues or longer | Peptide | Identify the query sequence or find protein sequences similar to the query | Standard Protein BLAST (blastp) |
| | | Find members of a protein family or build a custom position-specific score matrix | PSI-BLAST |
| | | Find proteins similar to the query around a given pattern | PHI-BLAST |
| | | Find conserved domains in the query | CD-search (RPS-BLAST) |
| | | Find conserved domains in the query and identify other proteins with similar domain architectures | Conserved Domain Architecture Retrieval Tool (CDART) |
| | Nucleotide | Find similar proteins in a translated nucleotide database | Translated BLAST (tblastn) |
| 5-15 residues | Peptide | Search for peptide motifs | Search for short, nearly exact matches (blastp) |

Table 2.3: BLAST program selection for protein queries.

1: The cut-off is only a recommendation. For short queries, one is more likely to get matches if the "Search for short, nearly exact matches" page at NCBI is used. With default setting, the shortest unambiguous query one can use is 11 for blastn and 28 for megablast.

The significance of each alignment is computed as a P-Value (probability of an alignment occurring with the score in question or better) or an E-Value (expectation value: the number of different alignments with scores equivalent to or better than S that are expected to occur in a database search by chance). Each alignment must be viewed by a critical human eye before being accepted as meaningful. For example high scoring pairs whose similarity is based

on repeated amino acid stretches (e.g. poly glutamine) are unlikely to reflect meaningful similarity between the query and the match. Filters, (e.g. SEG) or RepeatMasker that mask low complexity regions, should be applied to partially alleviate this problem.

## MEGABLAST

The best way to identify an unknown sequence is to see if that sequence already exists in a public database. If the database sequence is a well-characterized sequence, then one will have access to a wealth of biological information. Megablast, discontiguous-megablast, and blastn all can be used to accomplish this goal. However, megablast is specifically designed to efficiently find long alignments between very similar sequences and thus is the best tool to use in order to find the identical match to a query sequence. In addition to the expect value significance cut-off (E-Value), megablast also provides an adjustable percent identity cut-off for the alignment, which overrides the significance threshold set by expect value parameter.

Megablast uses a greedy algorithm [126] for nucleotide sequence alignment search and concatenates many queries to save time spent scanning the database. This program is optimized for aligning sequences that differ slightly as a result of sequencing or other similar "errors". It is up to 10 times faster than more common sequence similarity programs and therefore can be used to swiftly compare two large sets of sequences against each other. Megablast is also able to efficiently handle much longer DNA sequences than the blastn program.

## Discontiguous MEGABLAST

Discontiguous megablast is better at finding nucleotide sequences similar, but not identical, to a nucleotide query. This version of megablast is designed specifically for comparison of diverged sequences, especially sequences from different organisms, which have alignments with low degree of identity, where the original megablast is not very effective. The major difference is in the use of the 'discontiguous word' approach to finding initial offset pairs, from which the gapped extension is then performed.

The BLAST nucleotide algorithm finds similar sequences by breaking the query into short subsequences called words. The program identifies the exact matches to the query words first (word hits) and then extends these word hits in multiple steps to generate the final gapped alignments. One of the important parameters governing the sensitivity of BLAST searches is the length of the initial words, or word size as it is called. The most important reason that blastn is more sensitive than megablast is that it uses a shorter default word size. Because of this, blastn is better than megablast at finding alignments to related nucleotide

sequences from other organisms. The word size is adjustable in blastn and can be reduced from the default value of 11 to a minimum of 7 to increase search sensitivity.

Search sensitivity can be further improved by using the newly introduced discontiguous megablast. Rather than requiring exact word matches as seeds for alignment extension, discontiguous megablast uses non-contiguous words within a longer window of template. In coding mode, the third base wobbling is taken into consideration by focusing on finding matches at the first and second codon positions while ignoring the mismatches in the third position. Searching in discontiguous megablast using the same word size is more sensitive and efficient than standard blastn using the same word size. For this reason, it is now the recommended tool for this type of search. Alternative non-coding patterns can also be specified if desired.

It is important to point out that nucleotide-nucleotide searches are not the best method for finding homologous protein coding regions in other organisms. That task is better accomplished by performing searches at the protein level, by direct protein-protein BLAST searches or by translated BLAST searches. This is because of the codon degeneracy, the greater information available in amino acid sequence, and the more sophisticated algorithm in protein-protein BLAST.

### Blastp

Standard protein BLAST (blastp) is designed for protein searches. It is used for both identifying a query amino acid sequence and for finding similar sequences in protein databases. Like other BLAST programs, blastp is designed to find local regions of similarity. When sequence similarity spans the whole sequence, blastp will also report a global alignment, which is the preferred result for protein identification purposes. Unlike nucleotide BLAST, there is no comparable megablast for protein searches.

### Blastx

Blastx is useful for finding similar proteins to those encoded by a nucleotide query. Translated BLAST services are useful when trying to find homologous proteins to a nucleotide coding region. Blastx compares the translation of the nucleotide query sequence to a protein database. Because blastx translates the query sequence in all six reading frames and provides combined significance statistics for hits to different frames, it is particularly useful when the reading frame of the query sequence is unknown or it contains errors that may lead to frame shifts or other coding errors. Thus blastx is often the first analysis performed with a newly determined nucleotide sequence and is used extensively in analyzing EST sequences.

### Formatdb

Formatdb must be used in order to format protein or nucleotide source databases before these databases can be searched by blastall (blastp, blastn, blastx, tblastn, or tblastx) or megablast. The source database may be in either FASTA or ASN.1 format. Although the FASTA format is most often used as input to formatdb, the use of ASN.1 is advantageous for those who are using ASN.1 as the common source for other formats such as the GenBank report. Once a source database file has been formatted by formatdb it is not needed by BLAST any more. Finally it has to be noted that formatdb does not create non-redundant blast databases.

## 2.3.2 Java Cluster Service (JCS)

Life science is becoming increasingly quantitative as new technologies facilitate collection and analysis of vast amounts of data ranging from complete genomic sequences of organisms to three-dimensional protein structures. As a consequence, biomathematics, biostatistics and computational science are crucial technologies for the study of complex models of biological processes.

The quest for more insight into molecular processes in an organism poses significant challenges on the data analysis and storage infrastructure. Due to the vast amount of available information, data analysis on genomic or proteomic scale becomes impractical or even impossible to perform on commonly used workstations. Computer architecture, CPU performance, amount of addressable and available memory, and storage space are the limiting factors. Today, high performance computing [127] has become the third leg of traditional scientific research, along with theory and experimentation. Advances in genomics are inextricably tied to advances in high-performance computing.

The analysis of the humongous amount of available data for large scale comparative genomics studies requires parallel methods and architectures to solve the computational tasks in reasonable time.

In order to use the parallel features of a high-performance computing facility, the software has to meet parallel demands, too. A certain problem that has to be solved in parallel must be divided into subproblems that can be subsequently delegated to different processors. This partitioning procedure can be done either with so-called *domain decomposition* or *functional decomposition* (Figure 2.7).

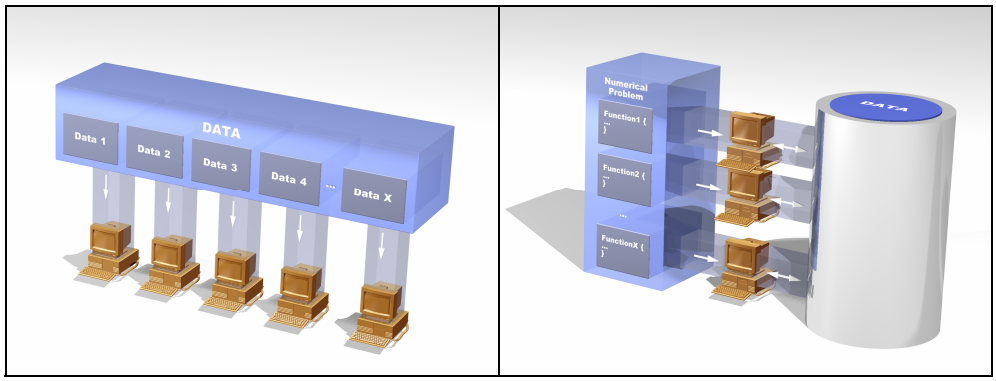



**Figure 2.7: Problem Partitioning:** (left) Domain or data decomposition is a computational paradigm where data to process is distributed and processed on different nodes. (right) Functional decomposition divides the computational problem in functional units, which are distributed onto different working nodes processing the same data.

The term *domain decomposition* describes the approach to partition the input data and to process the same calculation on each available processor. Most of the parallel-implemented algorithms are based on this approach dividing the genomic databases into pieces and calculating e.g. the sequence alignment of a given sequence on a subpart of the database. The second and simplest way to implement the *domain decomposition* on a parallel computing system is to take sequentially programmed applications and execute them on different nodes with different parameters. An example is to run the well known BLAST with different sequences against one database by giving every node another sequence to calculate. This form of application parallelization is called swarming and does not need any adaptation of existing programs. On the other hand *functional decomposition* is based on the decomposition of the computation process. This can be done by discovering disjoint functional units in a program or algorithm and sending these subtasks to different processors. Finally, in some parallel implementations combinations of both techniques are used, so that functional-decomposed units are calculating domain-parallelized sub-tasks.

The Java Cluster Service (JCS) is a Java 2 Enterprise Edition (J2EE) [128] compliant client-server application developed by Gernot Stocker at the Institute for Genomics and Bioinformatics, Graz University of Technology, to simplify distributing and monitoring of bioinformatics applications on Linux cluster systems. The modular concept enables the integration of command line oriented programs into the JCS application framework and facilitates the integration of different applications accessed through one interface and executed on a distributed cluster system.

The package is based on freely available, state-of-the-art technologies like JBoss [129] as application server, Tomcat [130] as web-container (integrated in JBoss), SOAP (Simple Object Access Protocol) as communication protocol between server and clients, and OpenPBS [131] as queuing system.

The delivery to the queuing system and collection of the results are managed by the framework in the background. Hereby, single process oriented applications are supported as well as real parallel applications. The client-server environment consists of a client for data preprocessing and results visualization and the JCS as an application server for distributing and handling of the jobs on the calculation cluster (Figure 2.8).

Data analysis is prepared in the client and the jobs are transferred to the JCS via SOAP, where jobs are distributed to available nodes of the calculation cluster using OpenPBS as queuing system, and results are stored until they are fetched by the client. At all times the client can get information about status and progress of the calculation task. Nevertheless, all server jobs are completely independent from the client, so that the client may be turned off during calculation and restarted again later to retrieve the computed results.

The user management system of the server warrants that only enrolled users have the rights to submit jobs and get their progress information and results.

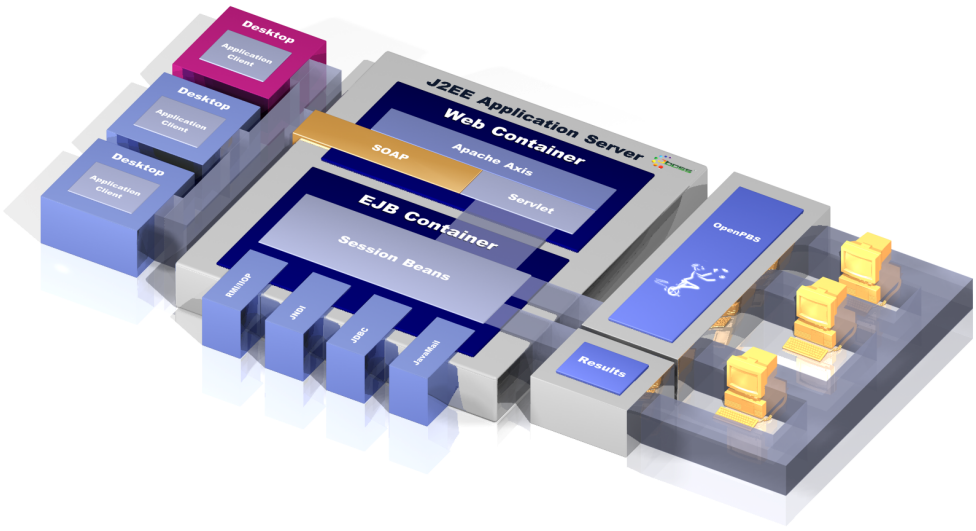



**Figure 2.8: The Java Cluster Service (JCS) architecture:** Desktop clients access the JCS via a SOAP interface. Java 2 Enterprise Edition Session Beans embedded in an EJB container of the JBoss application server handle job description and results. Jobs are submitted to the calculation nodes by the OpenPBS queuing system. The results are written back to the file system and can be fetched by the clients again through the SOAP interface and session beans. The SOAP interface is facilitated by the Apache Axis project, which runs in the web container of the application server.

# 2.4 Expression Profiling

## 2.4.1 Mouse Embryo Fibroblasts (MEFs) during Adipocyte Differentiation

**Growth and adipocyte differentiation of MEF cells:** Cells from mouse embryos (ME) at day 13 of gestation were prepared by culture of small tissue explants in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% fetal bovine serum (FBS; Gibco) and 1% antibiotics. The outgrowing primary cell population was passaged by trypsinization at a ratio of 1:3 to 1:4 upon confluency and continuously cultured in DMEM with 10% serum + 1% antibiotics to favor growth of fibroblastic cells [132]. Mouse embryo fibroblasts (MEFs) were differentiated according to the protocol described in [133]. Briefly, cells were induced to differentiation by the MDI protocol (1 mM Dexamethasone (DEX), 0,5 mM isobutyl methylxanthine (MIX) and 5 mg/ml Insulin). DEX and MIX were omitted after day 2, but Insulin was added throughout the differentiation. Cells were passaged and differentiated in AmnioMax (Gibco) supplemented with 7,5% fetal bovine serum. Nutrition media were changed every second day.

**Isoliation and analysis of RNA:** Three independent time series differentiation experiments were performed. For each time point RNA from three dishes were pooled. Cells were harvested at the preconfluent stage and at 7 time points (0 d, 1 d, 2 d, 3 d, 4 d, 6 d, and 10 d) (Figure 2.9) after hormonal induction total RNA was isolated with TRIzol reagent (Invitrogen-Life Technologies) based on the method described previously [134]. The quality of the RNA was checked by Agilent 2100 Bioanalyzer RNA assays through inspection of the 28S and 18S ribosomal RNA intensity peaks.

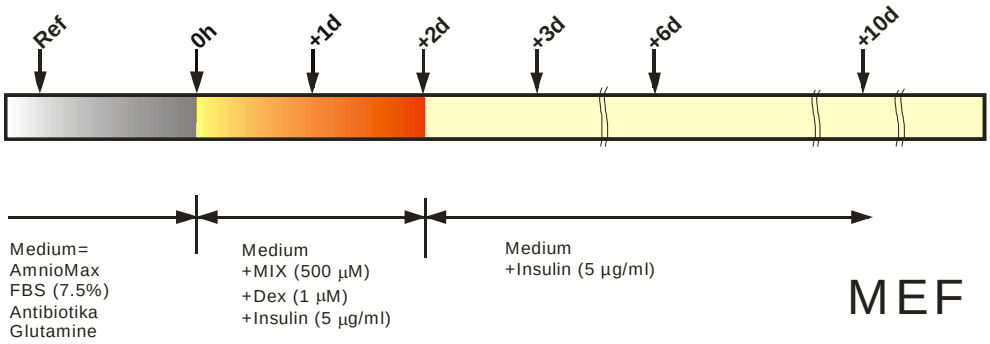



**Figure 2.9: MEF differentiation study experimental design:** Experimental design of expression profiling of Mouse Embryo Fibroblasts (MEFs) during Adipocyte Differentiation.

**Microarrays for expression profiling:** Mouse cDNA chips (MCC) were used for expression profiling. MCC is an aminosilane-coated glass slide (Corning) spotted with 27.648 PCR amplicons in 50% DMSO spotting buffer representing the 15K NIA mouse cDNA clone set, 11K BMAP clone set, and 627 adipose tissue related reporters.

**Probe labeling and microarray hybridization:** The used labeling and hybridization procedures were based on protocols developed at The Institute for Genomic Research [135]. Briefly, 20 µg of total RNA were indirectly labeled with Cy3 and Cy5, respectively. The Random Hexamer (Invitrogen) primed first strand cDNA synthesis was carried out using Superscript Reverse Transcriptase II (Invitrogen) in the presence of amino allyl dUTP (Sigma), dATP, dGTP, dCTP, dTTP (Invitrogen), DTT (Invitrogen), and 1X first strand buffer (Invitrogen) overnight at 42°C. cDNA was purified with QIAquick columns (Qiagen) according manufacturer's directions, but using potassium phosphate wash and elution buffer instead of supplied buffers PE and EB. N-hydroxy succinimide (NHS) esters of Cy3 and Cy5 (Amersham) were coupled to the amino allyl dUTPs incorporated in the cDNA. Coupling reactions were quenched by 0,1 M sodium acetate (pH=5,2) and unincorporated dyes were removed using QIAquick columns (Qiagen). Slides were prehybridized in 1% BSA, 5xSSC, 0,1% SDS for 45 min at 42°C and washed in MilliQ water and 2-Propanol and dried in a centrifuge. Fluorescent cDNA samples were dried in a SpeedVac, resuspended in 12 µl hybridization buffer (50% formamide, 5XSSC, 0,1% SDS) and combined. 20 µg mouse Cot1 DNA and 20 µg poly(A) DNA were added, denatured at 95°C for 3 min. and snap cooled on ice for 1 min. Sample with a final volume of 26 µl was applied to the prehybridized slide, covered with a glass cover slip (Roth) and hybridized in a humidified chamber for 20 hours at 42°C in the dark. Slides were washed 2 min in 1xSSC, 0,2% SDS solution (42°C) to remove the cover slip, 4 min agitating in 1xSSC, 0,2% SDS at room temperature, 4 min agitating in 0,1xSSC, 0,2% SDS at room temperature, 4 min agitating in 0,1xSSC at room temperature, and 2,5 min agitating in 0,1xSSC at room temperature, dipped twice in MilliQ water, and dried 2 min in a centrifuge at 1.500 rpm.

**Microarray image analysis:** Slides were scanned with a GenePix 4000B microarray scanner (Axon Instruments) at 10 µm resolution. Photo multiplier voltages (PMT) were selected in order that the histogram of the red channel (635 nm) and the green channel (532 nm) were overlapping to a large extend and few spots were saturated. Identical settings were used for the scanning of the corresponding dye-swapped hybridized slides.

**Microarray data analysis:** The resulting TIFF images for each of the two fluorophors were analyzed with GenePix Pro 4.1 (Axon Instruments) to get relative gene expression levels for each gene. Data were filtered for low intensity, inhomogeneity, and saturated spots by following criteria. If the median of the pixel intensities within a spot differs more than 20%

from the mean of the pixel intensities spots were considered as inhomogeneous and were filtered out. Spots with more than 10% of saturated pixels (fluorescence intensity >65.535) were excluded from further analysis. Genes with a very low expression value are often removed in order not to confound their signal with the background intensity. Low intensity spots were defined as those where the sum of the medians/means of the pixel intensities in both channels was lower than 1000 or not more than 55% of the pixels within a spot had intensities higher than the intensity of the surrounding background plus one standard deviation of the background pixels. All spots of both channels were background corrected, by estimation and subtraction of the local background.

Microarray data normalization: There are different sources of systematic (sample effect, array effect, dye effect, and gene effect) and random errors associated with microarray experiments [136]. Due to the different physical properties of the fluorescent dyes, the major portion of this bias is introduced by the dye effect. Therefore it is indispensable to normalize the data, which is known as removing of all non-biological variation introduced in the measurement and minimizing the random error to get reliable results [137,138]. After global median normalization, features (genes) showing substantial differences in the intensity ratios between technical replicates (dye-swapped slides) were excluded from further analysis, based on a threshold of 1 for the absolute difference between log ratios. Subsequently, dye-swap normalization was performed, data were log2 transformed, and averaged over the three independent experiments. All normalization steps were carried out using ArrayNorm [139].

### 2.4.2 Human Multipotent Adipose-derived Stem Cells (hMADS) during Adipocyte Differentiation.

hMADS characterization: The establishment and characterization of human multipotent adipose-derived stem (hMADS) cells as well as their ability to differentiate into cells with key features of human adipocytes are described previously [140]. Briefly, hMADS cells were isolated from white adipose tissue removed from surgical scraps of infants undergoing surgery, did not enter senescence while exhibiting a diploid karyotype, were non-transformed though they expressed significant telomerase activity, did not show any chromosomal abnormalities after 140 population doublings (PDs), and maintained their differentiation properties after 160-200 PDs. hMADS cells were able to withstand freeze/thaw procedures and their differentiation could be directed under different culture conditions into various lineages.

**Experimental design:** Three independent time series differentiation experiments were performed for each cell model as biological replicates. Cells were harvested at the preconfluent stage as reference and after hormonal induction 8 time points (-2 d, 0 d, 8 h, 24 h, 48 h, 5 d, 10 d, 15 d) (Figure 2.10). Expression profiling was carried out using DNA microarray technology. Technical replicates were performed by repeating each hybridization with swapped dyes resulting in 48 microarray hybridizations.

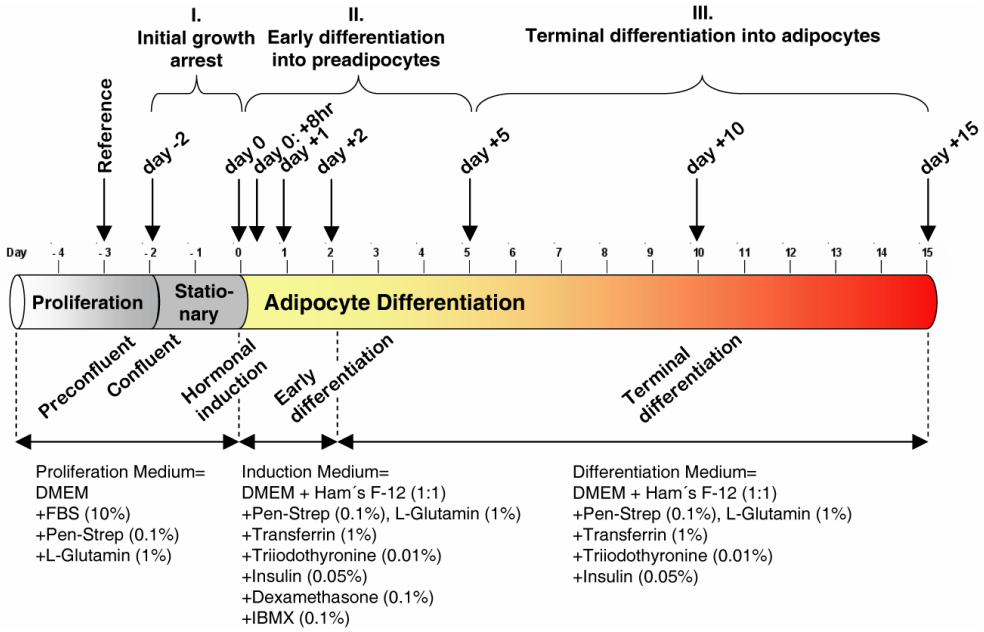



**Figure 2.10: hMADS differentiation study experimental design:** Experimental design of expression profiling of human Multipotent Adipose-derived Stem Cells (hMADS) during Adipocyte Differentiation.

**Microarrays for expression profiling:** Human oligonucleotide chips (HOC) were used for expression profiling. HOC is an epoxy-coated glass slide (Nexterion) spotted with 33.456 features in 3x SSC and 1,5 M Betaine spotting buffer representing 29.552 reporters. Reporter molecules consist of 50mer single stranded oligonucleotides [141,142] with Human Oligo Set A, B, and C with a size of 10K each representing genes with known function or clearly defined protein domains, additional Ensemble annotated open reading frames (ORFs) with experimental sequence evidence, and Ensemble Genscan predicted ORFs.

**Growth and adipocyte differentiation of hMADS cells:** Cells were seeded in triplicate at a density of 25 x $10^3$ cells/100 mm dish in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% heat-inactivated FBS, 2 ng/ml hFGF-2, 5 U/ml penicillin, and 5 µg/ml streptomycin. The medium was changed 2 days later in the absence of hFGF-2 until cells reached confluence. At day 2 post-confluence (designated day 0), cells were then induced to differentiate in the presence of DMEM/Ham's F12 media supplemented with 0,86 µM insulin, 0,2 nM triiodothyronine, 1 µM dexamethasone (DEX), 100 µM isobutyl-methylxanthine

(IBMX), 10 µg/ml transferrin, and 0,5 µM rosiglitazone. This induction medium was changed three days later (DEX and IBMX omitted). The media were then changed every second day and cells were used at the indicated timepoints.

**Isoliation and analysis of RNA:** Total RNA was isolated with TRI Reagent (Euromedex) and TRIzol reagent (Invitrogen-Life Technologies) according to manufacturer´s instruction and based on the method described previously [134]. For each time point RNA from at least two dishes were pooled. The quality of the RNA was checked by Agilent 2100 Bioanalyzer RNA assays through inspection of the 28S and 18S ribosomal RNA intensity peaks.

**Probe labeling and microarray hybridization:** The used labeling and hybridization procedures were based on protocols developed at The Institute for Genomic Research [135]. Briefly, 20 µg of total RNA was indirectly labeled with Cy3 and Cy5, respectively. The Random Hexamer (Invitrogen) primed first strand cDNA synthesis was carried out using Superscript Reverse Transcriptase II (Invitrogen) in the presence of amino allyl dUTP (Sigma), dATP, dGTP, dCTP, dTTP (Invitrogen), DTT (Invitrogen), and 1x first strand buffer (Invitrogen) overnight at 42°C. cDNA was purified with QIAquick columns (Qiagen) according manufacturer's directions, but using potassium phosphate wash and elution buffer instead of supplied buffers PE and EB. N-hydroxy succinimide (NHS) esters of Cy3 and Cy5 (Amersham) were coupled to the amino allyl dUTPs incorporated in the cDNA. Coupling reactions were quenched by 0,1 M sodium acetate (pH=5,2) and unincorporated dyes were removed using QIAquick columns (Qiagen). Slides were prehybridized in 1% BSA, 5xSSC, 0,1% SDS for 45min at 42°C and washed in MilliQ water and 2-Propanol and dried in a centrifuge. Fluorescent cDNA samples were dried in a SpeedVac, resuspended in 12 µl hybridization buffer (50% formamide, 5xSSC, 0,1% SDS) and combined. 20 µg mouse or human Cot1 DNA and 20µg poly(A) DNA were added, denatured at 95°C for 3 min and snap cooled on ice for 1 min. Sample with a final volume of 26 µl was applied to the prehybridized slide, covered with a glass cover slip (Roth) and hybridized in a humidified chamber for 20 h at 42°C in the dark. Slides were washed 2 min in 2xSSC, 0,1% SDS solution (42°C) to remove the cover slip, 5 min agitating in 2x SSC, 0,1% SDS (30°C), 5 min agitating in 1xSSC (30°C), 5 min agitating in 0,5xSSC (30°C), dipped twice in MilliQ water, and dried 2 min in a centrifuge at 1.500 rpm.

**Microarray image analysis:** Slides were scanned with a GenePix 4000B microarray scanner (Axon Instruments) at 10 µm resolution. Photo multiplier voltages (PMT) were selected in order that the histogram of the red channel (635 nm) and the green channel (532 nm) were overlapping to a large extend and few spots were saturated. Identical settings were used for the scanning of the corresponding dye-swapped hybridized slides.

**Microarray data analysis:** The resulting TIFF images for each of the two fluorophors were analyzed with GenePix Pro 4.1 (Axon Instruments) to get relative gene expression levels for each gene. Data were filtered for low intensity, inhomogeneity, and saturated spots by following criteria. If the median of the pixel intensities within a spot are differing more than 20% from the mean of the pixel intensities spots were considered as inhomogeneous and were filtered out. Spots with more than 50% of saturated pixels (fluorescence intensity >65.535) were excluded from further analysis. Genes with a very low expression value are often removed in order not to confound their signal with the background intensity. Low intensity spots were defined as those where the sum of the medians/means of the pixel intensities in both channels was lower than 750 or not more than 55% of the pixels within a spot had intensities higher than the intensity of the surrounding background plus one standard deviation of the background pixels. There are different sources of systematic (sample effect, array effect, dye effect and gene effect) and random errors associated with microarray experiments [136]. Due to the different physical properties of the fluorescent dyes, the major portion of this bias is introduced by the dye effect. Therefore it is indispensable to normalize the data, which is known as removing of all non-biological variation introduced in the measurement and minimizing the random error to get reliable results [137,138,143]. After global mean normalization, features (genes) showing substantial differences in the intensity ratios between technical replicates (dye-swapped slides) were excluded from further analysis, based on a threshold of 1 for the absolute difference between log ratios. Subsequently, dye-swap normalization was performed, data were log2 transformed, and averaged over the three independent experiments. All normalization steps were carried out using ArrayNorm [139].

## 2.5 Transcriptome Data Retrieval

Microarray analysis is a very complex, multi step technique involving array fabrication, labeling, hybridization and data analysis. All aforementioned steps generate a wealth of data spanning tenth of megabytes and each of them leaves a lot of room where errors may occur or protocols might need optimization to improve results. Moreover, information on details of the bench work, typically kept in lab notebooks or scattered files, as well as information regarding spotting, reliable tracking of the spotted molecules, scanning, and image quantification settings, is very relevant to the computational analysis and to reproduce experiments. All these information must be archived according to accepted scientific standards, which then allow scientists to share common information and to make valid comparisons among experiments.

For this reasons, we have developed a multi color Microarray Analysis and Retrieval System (MARS) database suite [144] that allows to store, manage, organize, and query the wealth of

data generated during the microarray production, analysis, and quality control process and that is providing various platform dependent and independent application and programming interfaces. MARS provides: (1) an integrated lab notebook to store the necessary information during biomaterial manipulation, (2) a laboratory information management system (LIMS) to keep track of the information that occurs during the microarray production, (3) a quality management application storing necessary quality control parameters, (4) an application and programming interface, and (5) web services to connect several well established tools such as normalization, clustering and pathway annotation applications.

MARS has been developed as a web-based and MIAME (Minimum Information About a Microarray Experiment) [145-149] compliant microarray database that allows several institutions the acquisition, management, and retrieval of all necessary parameters for microarray production and experiments in a scalable and performant way. In general MARS is accessible via a standard web browser.

MARS is based on a three-tier architecture using the Java 2 Platform, Enterprise Edition (J2EE) [128], which defines a standard for developing multi-tier enterprise applications. The J2EE platform simplifies the development of enterprise applications by basing them on standardized, modular components like Enterprise JavaBeans (EJB), Java Servlets, Java Server Pages (JSP), and XML technology [150], by providing a complete set of services to those components, and by handling many details of application behavior automatically.

A relational database (Oracle [151] or PostgreSQL [152]) builds the data tier. Relational databases organize the data items as a set of formally-described tables, which allow to access data very efficiently or reassemble it in many different ways without having to reorganize database tables.

In the middle tier, JBoss [129] a J2EE compliant application server is situated (Figure 2.11). It manages the access to the relational database as well as the interaction between the data, the so called business logic. The web-server including a servlet-container is responsible for the presentation tier. Within this container all the servlets and JSPs are executed to handle the input and output of the application and to manage the applications workflow logic. An advantage of this multitier architecture is that the different tiers can be deployed to different servers and therefore load distribution as well as scalability can be guaranteed.

The Microarray Gene Expression Markup Language (MAGE-ML) [153] has emerged as a language to describe and exchange information about microarray based experiments [154]. MAGE-ML is based on XML (eXtensible Markup Language) and is used to describe microarray designs, microarray manufacturing information, microarray experiment setups and

execution information, gene expression data, and data analysis results. MARS is able to export samples, extracts, labeled extracts, array designs, raw datasets, and whole experiments including several hybridizations. This feature facilitates to easily share and publish high quality, well annotated data within the life science community by uploading these generated files to public repositories like ArrayExpress [155].

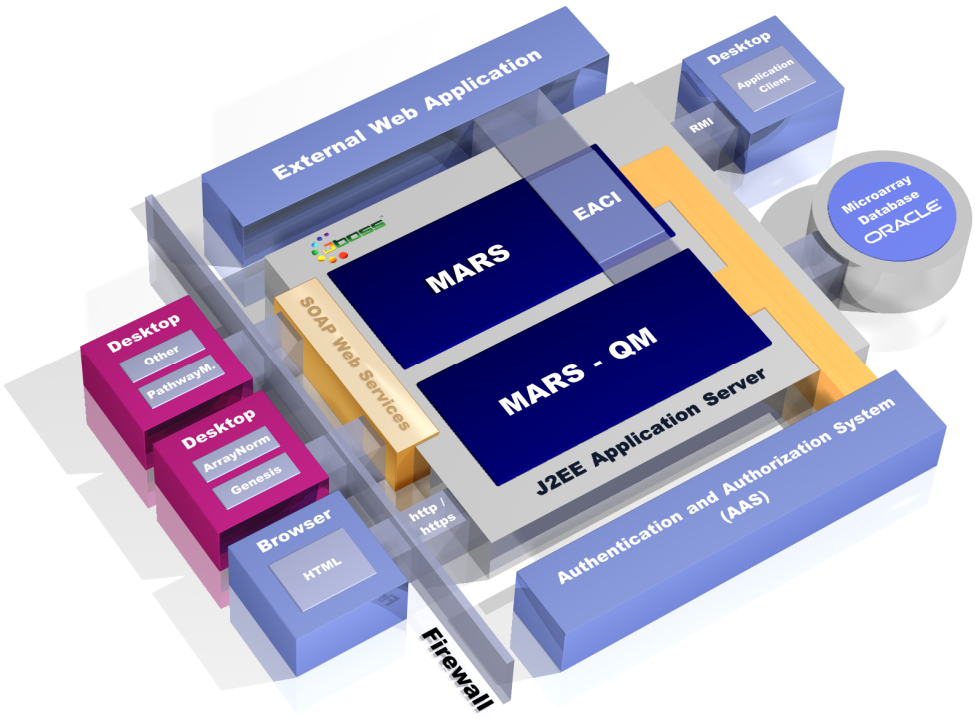



**Figure 2.11: MARS system interactions:** MARS and MARS-QM are deployed in a J2EE compliant application server. Interaction is possible either with a standard web browser or an application supporting the SOAP or RMI protocol. The External Application Connector Interface (EACI) facilitates the connection to data from additional web applications. SOAP and http/https enable MARS access also through firewalls.

The implementation of other web based applications and more importantly the usage and correct linkage of its stored data has been addressed by an External Application Connector Interface (EACI). Additional applications like supplementary quality checks can be added without having to amend the MARS source code. The MARS user interface is dynamically displaying links to all formerly registered applications.

In order to assure high-quality data and to understand or optimize lower value data it is important to be able to trace back all conducted quality control steps. MARS integrates several quality measurements performed during the microarray production as well as during

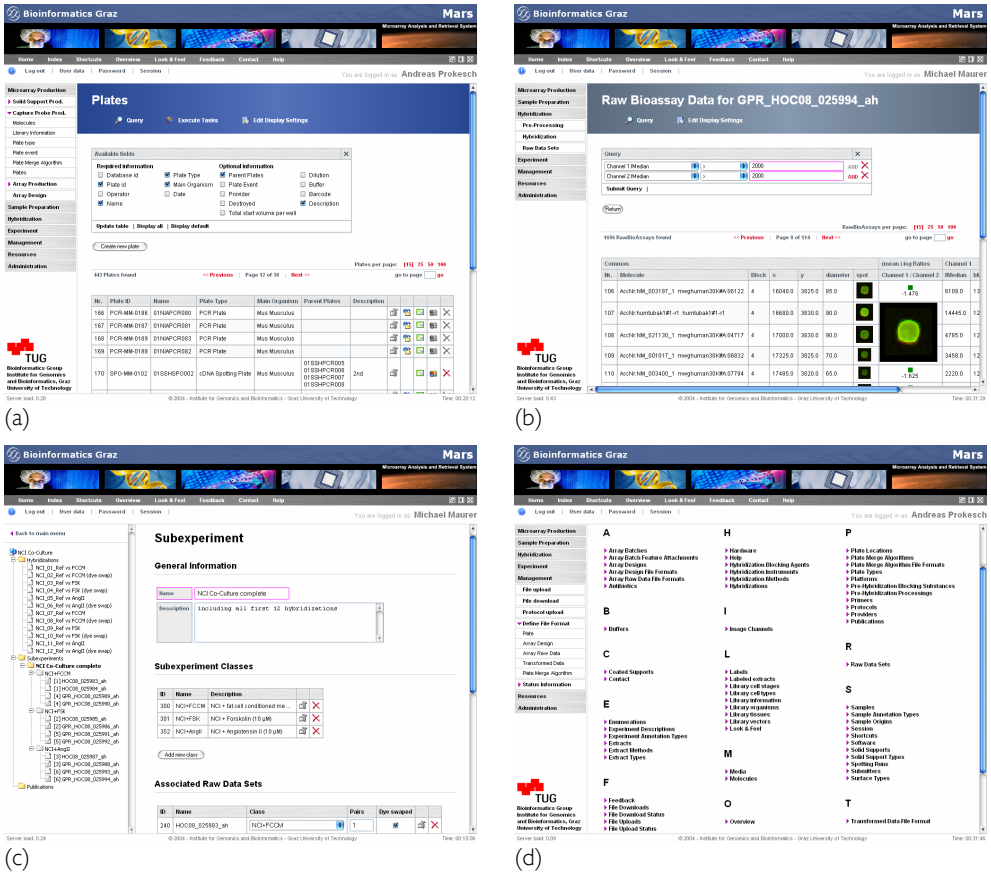



**Figure 2.12: MARS web-based user interface:** (a) Plate information page. Plates can be merged and inherit properties from parent plates. (b) Raw data sets are displayed including an image of each spot. These spot images are automatically retrieved from the scanned microarray image. (c) Whole experiments can be defined, consisting of a bunch of hybridizations. These experiments and all containing data can be imported into ArrayNorm for normalization in a single step. (d) Hundreds of parameters are collected and displayed using more than 70 pages.

the sample preparation, extraction, and hybridization process. As mentioned before these quality checks are implemented as an additional application called MARS-QM which is tightly integrated into MARS utilizing the EACI.

Since microarray as well as the corresponding quality control data may contain highly sensitive data, we have integrated an authentication and authorization system (AAS) to provide authentication and fine grained authorization mechanisms. The combination of AAS and EACI provides through a single sign on mechanism and dynamic linkage of data the possibility to assemble heterogeneous applications to one powerful suite. MARS enables users to share their data sets with other users. Supplementary to the user oriented data management an institution oriented level has been introduced. This amelioration allows several institutes to store their data into one data repository without having to share common settings and resources like scanners, but offering the possibility to share the data among institutions.

Summarizing, the MARS database design, state-of-the-art software technology, well designed user interface, and its powerful application interfaces provide a capable tool for storing, retrieving, and analyzing multi color microarray data. The flexibility, platform independency, well designed user interface, and unique affiliation of using web-based and standalone applications connected to the latest powerful application server technology provides MARS with the potential to become a valuable tool for the pursuit of future biological discoveries.

## 2.6 Gene Expression Data Analysis

An important step in gene expression analysis is to extract the fundamental patterns of gene expression inherent in the data in a mathematical process called clustering, which organizes the genes into biological relevant clusters with similar expression patterns (coexpressed genes).

### 2.6.1 Hierarchical Clustering (HCL)

Hierarchical clustering [40,46,156-158] is an unsupervised procedure of transforming a distance matrix, which is a result of pair wise similarity measurement between elements of a group, into a hierarchy of nested partitions. The hierarchy can be represented with a tree-like dendrogram in which each cluster is nested into the next cluster.

Hierarchical clustering is the most commonly used clustering strategy for gene expression analysis at the moment. The biggest advantage is that aside from a choice of the amalgamation rule and the type of similarity distance measurement, no further parameters have to be specified. The result is a reordered set of genes and/or experiments, where similar vectors are close to each other in the tree structure and the distance between vectors and clusters is encoded in the branch length of a subtree. This not only allows estimation of the similarity of neighboring genes, but also of the distance between distant vectors. This is helpful if someone is more interested in distances rather than similarities between two or more investigated conditions.

### 2.6.2 Self Organizing Maps (SOM)

One of the most popular neural network models today is the principle of a Self-Organizing Map (SOM) [159-163], developed by professor Kohonen at the University of Helsinki. A SOM is basically a multidimensional scaling method, which projects data from input space to a lower dimensional output space. The SOM algorithm is based on unsupervised competitive

learning, which means that the training is entirely data-driven and needs no further information.

### 2.6.3 k-means Clustering (KMC)

k-means [164-167] is a commonly used clustering method because it is based on a very simple principle and provides good results. It is very similar to SOM, unsupervised, and can be seen as a Bayesian (maximum likelihood) approach to clustering.

The basic idea is to maintain two estimates:

1. An estimate of the center location for each cluster and
2. A separate estimate of the partition of the data points according to which one goes into which cluster.

One estimate can be used to refine the other. If we have an estimate of the center locations, then (with reasonable prior assumptions) the maximum likelihood solution is that each data point should belong to the cluster with the nearest center. Hence, we can compute a new partition from a set of center locations, i.e. make one cluster from the set of vectors in each Voronoi cell (The Voronoi region of unit is defined as the union of all input vectors to which it is the closest). For this reason, the k-means algorithm proceeds by a sequence of phases in which it alternates between moving data points to the cluster of the nearest center, and moving all cluster centers to the mean of their Voronoi sets.

### 2.6.4 Figure of Merit (FOM)

Figure of Merit [168] is motivated by the jackknife approach and a method for assessing the quality of clustering results. A clustering algorithm is applied to all but one experimental condition in a dataset, and the left-out condition is used to assess the predictive power of the clustering algorithm. A scalar quantity called the Figure of Merit (FOM) is defined, which is an estimate of the predictive power of a clustering algorithm. The Figure of Merit can be defined as the root mean square deviation in the left-out condition of the individual gene expression levels relative to their cluster means. The adjusted Figure of Merit is the figure of merit divided by a factor that compensates for a statistical bias with many clusters. A small Figure of Merit indicates a clustering algorithm having high predictive power. The Figure of Merit can for instance be used to estimate the number of cluster for k-means clustering or SOM.

### 2.6.5 Principal Component Analysis (PCA)

Principal Component Analysis (PCA), also known as Singular Value Decomposition (SVD) [169-172] is an exploratory multivariate statistical technique that allows the identification of key variables (or combinations of variables) in a multidimensional data set that best explains the differences between observations. Given $m$ observations (experiments) on $n$ variables (genes), the goal of PCA is to reduce the dimensionality of the data matrix by finding $r \leq n$ new variables. These $r$ principal components account together for as much of the variance in the original $n$ variables as possible while remaining mutually uncorrelated and orthogonal. The goal is to reduce dimensionality while filtering noise in the process, making the data more accessible for visualization and analysis.

### 2.6.6 Correspondence Analysis (CA)

Correspondence Analysis [173] is an explorative computational method for the study of associations between variables. Much like principal component analysis, it displays a low-dimensional projection of the data, e.g., into a plane. It does this, though, for two variables simultaneously, thus revealing associations between them. Like other projection methods, CA represents variables such as transcription intensities of genes as vectors in a high dimensional space. In our case, the dimensionality of the space would be the number of hybridizations involved. Both PCA and CA reveal the principal axes of this high-dimensional space, enabling projection into a subspace of low dimensionality that accounts for the main variance in the data. Unlike PCA, CA is able to account for the genes in hybridization-dimensional space and the hybridizations in gene-dimensional space at the same time. Both representations of the data matrix will be projected into the same low-dimensional subspace, for example, a plane, revealing associations both within and between these two variables.

### 2.6.7 One-Way-ANOVA

One-way ANOVA tests [174] differences in a single interval dependent variable among two, three, or more groups formed by the categories of a single categorical independent variable. Also known as univariate ANOVA, simple ANOVA, single classification ANOVA, or one-factor ANOVA, this design deals with one independent variable and one dependent variable. It tests whether the groups formed by the categories of the independent variable seem similar (specifically that they have the same pattern of dispersion as measured by comparing estimates of group variances). If the groups seem different, then it is concluded that the independent variable has an effect on the dependent (e.g. if different treatment groups have different health outcomes).

## 2.6.8 Gene Expression Terrain Maps

In Gene Expression Terrain Maps [47,175-177], co-regulated genes are grouped together and visualized in a three-dimensional expression map that displays correlations of gene expression profiles as distances in two dimensions and gene density in the third dimension. The expression data are used to calculate correlations between every pair-wise combination of genes. For each gene, the similarity between it and the *k* genes with the strongest correlations were used to assign that gene to an x-y coordinate in a two-dimensional scatter plot with the use of force-directed placement. In this x-y ordination step, genes are positioned relative to each other under the influence of attractive and repulsive forces. Each gene is attracted to other genes with a force proportional to their similarity in gene expression, but a constant force also repels each gene from groups of other genes. The spatial distribution of the genes is visualized, resulting in a display in which genes with a high correlation are placed near to each other. As a further visual cue, the scatter plot is converted to a gene expression terrain map showing the gene correlations in three dimensions, where the altitude of a mountain corresponds to density of the genes (Figure 2.13).

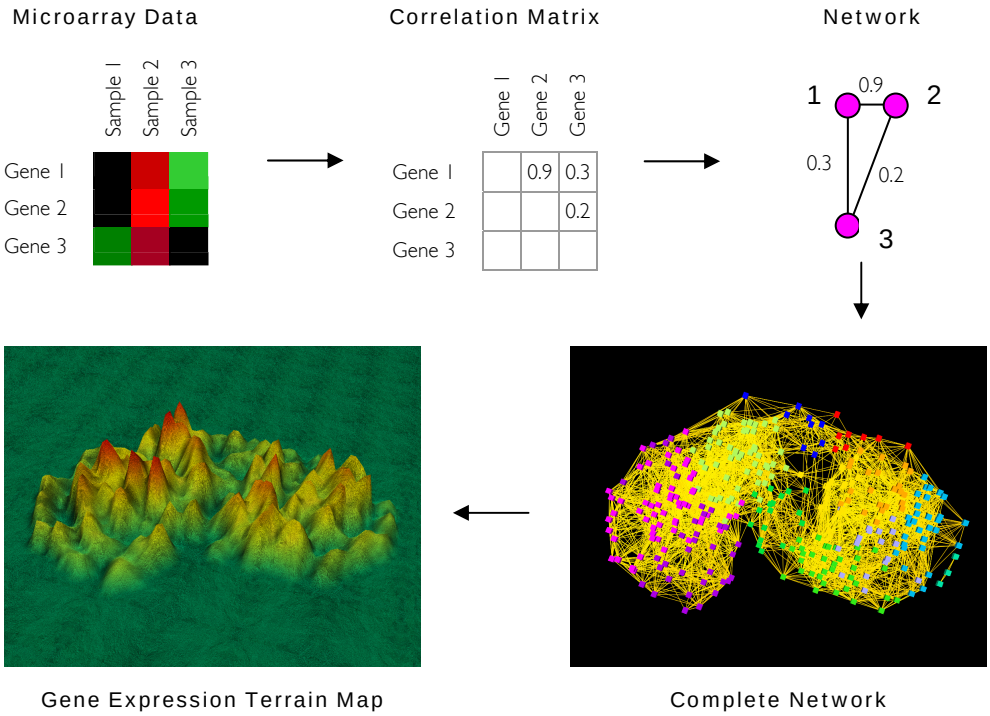



**Figure 2.13: Construction of a gene expression terrain map:** In the expression matrix, red denotes increased relative gene expression and green denotes decreased gene expression. Only three genes and three experiments are shown for simplicity. The expression data are used to calculate correlations between every pair-wise combination of genes. The most correlated genes in the correlation matrix are used to construct a two-dimensional scatter plot. The scatter plot is converted to a gene expression terrain map showing the gene correlations in three dimensions, where the altitude of a mountain corresponds to density of the genes, denoted by red, yellow, and green.

# 2.7 Genome Annotation

## 2.7.1 Gene Ontology (GO)

An ontology [178,179] has two primary pragmatic purposes: The first is to facilitate communication between people and organizations. The second is to improve interoperability between systems.

The exponential growth in the volume of accessible biological information has generated a confusion of voices surrounding the annotation of molecular information about genes and their products. The Gene Ontology (GO) [180-183] project seeks to provide a set of shared, structured vocabularies adequate for the annotation of molecular characteristics across organisms in a consistent way even as knowledge of gene and protein roles in cells is accumulating and changing. This work includes building three extensive ontologies to describe molecular function, biological process, and cellular component. These particular classifications were chosen because they represent information sets that are common to all living forms and are basic to the annotation of information about genes and gene products. Briefly, molecular function describes what a gene product does at the biochemical level. Biological process describes a broad biological objective. Cellular component describes the location of a gene product, within cellular structures and within macromolecular complexes.

This effort parallels work in the computational biology community to provide (1) tools for implementing biological ontologies and (2) a community database resource that supports the use of these ontologies. The ontologies and gene annotations have been loaded into a relational database for robust representation and query capabilities. Implemented in MySQL [184], the data model incorporates the relationships between terms and includes versioning of terms, their synonyms, and definitions. The association files of organism-specific gene-product annotations are also part of the database representation.

The strength of the GO approach lies in its focus on the specifics of the biological vocabularies and on the establishment of precise, defined relationships between them. The ontologies are structured in the form of directed acyclic graphs (DAGs) that represent a network in which each term may be a "child" (more specialized term) of one or more than one "parent" (less specialized term). Relationships of child to parent can be of the "is a" type or the "part" type. Each term in the ontology is an accessible object in the GO data resource. Every term has a unique identifier to be used as a database cross-reference.

The creation of the ontologies and the association of ontology terms with gene products are two independent operations. A gene product is a physical entity: a protein or a functional RNA. Gene products may assemble into entities that function as complexes, or gene product groups. Genes, gene products, gene product precursors, and gene product complexes can each and all be associated with one or more GO terms. Each gene product can be described in this system as having one or more functions, being involved in one or more biological processes, and as occurring in one or more cellular locations.

The shared development of this molecular annotation is believed to contribute to the unification of biological information. The use of GO in analysis of experimental data from high throughput methods enables integration of biological background data in a controlled manner.

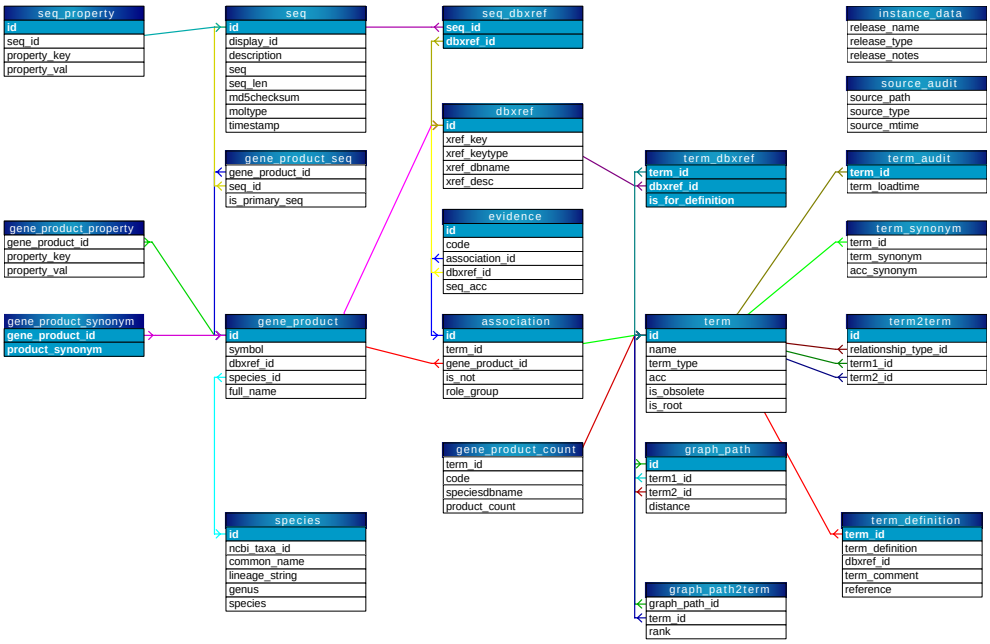



**Figure 2.14: General structure and primary/foreign key relationships of the GO database:** The GO schema has a modular design and the GO database schema modules are designed for housing any GO ontology together with associated annotations and other auxiliary data. The most important are: go-general (general tables, not specific to GO), go-graph (tables for representing directed graphs, the central concept in a GO style ontology. Nodes are terms, arcs are relationships between terms), go-meta (metadata about nodes in the graph; for example, synonyms, links to external databases, comments, and definitions), go-associations (annotations of gene products using GO terms. Stores metadata about the gene product itself, as well as data on the actual association between GO term and gene product, such as evidence), and go-seq (biological sequences attached to gene products).

# 2.8 Promoter Analysis

## 2.8.1 PromoSer Database

Many transcription control elements are within regions close to the Transcription Start Sites (TSSs) of expressed sequences, whether they are for coding or non-coding mRNA [185]. The common prerequisite for all computational analysis methods of the sequence based nature of this regulation is the availability of proximal promoter sequences for large sets of co-regulated genes. It is well known that enhancer and suppressor control elements can exist at sites tens of thousands of bases upstream or even downstream of the transcription start site [186]. In many cases however, the essential control elements are present within the proximal promoter a few hundred to a couple of thousand bases upstream of the TSS [185,187,188]. This region still remains an elusive target for the exact characterization of its structure. With microarray experiments generating data for many thousands of transcribed sequences, an efficient method to obtain the proximal promoters of these transcripts became highly desirable. PromoSer [189-191] is a web service that was designed specifically for this purpose.

PromoSer is a freely accessible web-based service to facilitate the batch extraction of user specified regions around the transcription start of a large number of proximal promoter sequences from mammalian genomes. By providing (1) a list of GenBank accession ids to identify the genes of interest and (2) the range required flanking the TSS, PromoSer will process the input and return the required regions as a multi-FASTA format text file.

A central concept in PromoSer is the prediction of TSSs based on purely experimental transcript data and no computational gene predictions, thus maximizing the confidence in the identified TSSs. The TSSs are identified computationally by considering alignments of a large number of partial and full-length mRNA and EST sequence data to genomic DNA, with provision for alternative promoters. Overlapping alignments are tracked (denoted as a cluster) to determine the furthest possible extension to these sequences and hence determine the TSS. In many cases, the PromoSer data set is enriched with full-length mRNA sequences produced by cap-trapping and oligo-capping methods, providing higher confidence in predictions. By utilizing nearly all major public data sets of full-length cDNA sequence information, PromoSer achieves both coverage and accuracy.

Using a powerful cluster of 128 dual-processor compute nodes and the efficient BLAT tool [192], each of these more than 10.000.000 mRNA or EST sequences were aligned to their

corresponding genomes and localized to specific chromosomal regions. To improve sensitivity, the standalone version of BLAT was used to compare all sequences against each chromosome. The mapping results of every sequence were compared to determine the best alignment, or in some cases where several mappings were nearly equally good, the best alignment and those within 1% of its score were retained.

To improve performance and retain accuracy, the filtering process was performed in three stages. First, a low-stringency (80% identity over at least 100 bases) initial filter was applied to select alignments used in the clustering. Once clustered, a second, stringent filter (90% identity for EST and 95% otherwise, and no more than 5 unaligned bases at the start of the sequence) was used to select alignments that may be used to predict the TSS. Finally, a third pass heuristically assigns the alignments that did not pass the initial filter to the smallest cluster that fully overlaps that alignment. Those guessed alignments are marked and reported as being a guess when the user searches for them.

To cluster sequences, all alignments that overlap a genomic region and are transcribed in the same orientation are collected. They are then separated into sub-clusters based on the sharing of transcribed regions. Finally, each sub-cluster is examined to determine if it contains independent subcomponents that are connected through a single EST. If so, the sub-cluster is broken up into its subcomponents. After clustering, candidate TSSs are identified as the 5´-most position of transcripts passing the stringent filter (see above) plus the 5´-most position in the cluster overall. Sites within 20 bp are grouped and the 5´-most one is retained.

Individual TSSs are assigned a quality and a support score. The cluster is finally annotated based on its largest RefSeq or mRNA sequence. The locations of gaps longer than 500 bases and other clusters upstream are noted as possible boundaries to promoter sequence extraction.

A fully functional SOAP server for PromoSer that captures most of the web user interface functionality has recently been implemented. The service is described with a Web Services Description Language (WSDL) document [193]. The service works using a job-ticket model, in which a reference ID is immediately returned upon successful request submission. A user can then poll the server to check for the availability of the results for that job ticket. When found, PromoSer will return an XML-formatted results file that contains both the web-based report and the promoters. The SOAP interface greatly facilitates access by advanced users and computer scripts. This feature is essential for integrating diverse biological data and applications [194].

### 2.8.2 Distribution of all Octamer DNA Sequences

To identify DNA sequences that cluster relative to the TSS, the distribution of all dinucleotides and 8-mers can be determined in a set promoter sequences [195]. Because both strands of complementary DNA are examined, the number of independent 8-mers is reduced from 65.536 to 32.896 (32.640 nonpalindromic 8-mers + 256 palindromic 8-mers).

First, the promoter sequence is divided into a number of bins, each bin contains 20 bp. To determine if a DNA sequence clusters, the mean ($\bar{x}$) and standard deviation ($\sigma$) is determined based on its abundance in each of the bins. Those bin values that are $\geq 2\sigma$ above the mean are considered to be part of the cluster and a new mean ($\bar{x}'$) and standard deviation ($\sigma'$) is calculated excluding these bin values. A clustering factor (CF) is calculated based on this corrected mean and standard deviation:

$$CF = \frac{x_{\max} - \bar{x}'}{\sigma'}$$

To display the results, the CF values for all 32.896 8-mers are plotted against the bin with the maximum value.

### 2.8.3 Wise DNA Block Aligner

The DNA Block Aligner (DBA) [196,197] was developed by Niclas Jareborg, Richard Durbin, and Ewan Birney for characterizing shared regulatory regions of genomic DNA, either in upstream regions or introns of genes. DBA aligns two sequences under the assumption that the sequences share a number of co-linear blocks of conservation (perhaps with one or two insertions or deletions) separated by potentially large and varied lengths of DNA in the two sequences. The conserved blocks may be regions of importance for the regulation of a gene. This is a very sensible thing to do with syntenous regions of non coding DNA between mouse and human (for example, the upstream regions of a gene from mouse and human or the conserved intron of a human-chicken gene).

The subsequent model was a 3 state model, which was a log-odd'd ratio to a null model of their being no examples of a motif in the two sequences. The final model is a probabilistic finite state machine (or pair-HMM) which aligns the two sequences. Each block can choose one of 4 different parameter sets, roughly being conservation at 65, 75, 85 or 95 percent identity. Linear gaps (gaps where the gap open is the same as the extension) have been

modeled in the blocks at a fixed probability of 0.05 and each block is expected around 1% of the DNA sequence.

### 2.8.4 PromoterWise

PromoterWise [198] is a sort of next generation DBA. It is designed for comparisons between two promoter sequences or realistically any two orthologous regulatory regions (or homologous for that matter, but in theory it should work better for orthologous regulatory regions, depending on how much active change paralogous regulatory regions are expected to have). PromoterWise reports alignments between these two sequences assuming that alignments cannot overlap in both sequences, but "not" assuming that the alignments have to be co-linear or on the same strand.

PromoterWise works by taking the two sequences and then finds all common exact 7mers between them, in both the forward and reverse strands. These are then merged such that close high scoring sequence pairs (whose centers are within the window size of each other) are considered as one region. These regions then have a local version of the DBA algorithm run over them, which has a model of DNA similarity of small regions of similarity, potentially with small gaps separated by large pieces of unknown DNA.

The resulting set of alignments is then sorted by score, and a simple greedy algorithm is used to discard "bad" subsequent alignments. By default this is to discard alignments which overlap on the query coordinate with alignments of a higher score (this can be changed). The alignments are then outputted with bits score.

# 3 Results

## 3.1 Overview

A bioinformatics platform for large-scale comparative transcriptomics has been composed comprising the following components:

1. Tools for automated high-performance sequence retrieval.
2. A fully automated pipeline for finding corresponding protein sequences to any given nucleotide sequence in a high-performant and reliable way.
3. A pipeline for fully automated retrieval of putative orthologous and inparalogous relations between two arbitrary organisms in general and human-mouse in particular.
4. A pipeline for fully automated gene ontology (GO) annotation and tools to display the annotation in context with gene expression data.
5. A gene expression analysis and visualization environment providing (a) filtering and sorting of data, (b) a comprehensive set of similarity distance measurements, (c) a variety of hierarchical and non-hierarchical clustering and classification algorithms (Hierarchical Clustering (HCL), Self Organizing Maps (SOM), k-means Clustering (KMC), Principal Component Analysis (PCA), Correspondence Analysis (CA), One-Way-ANOVA, Support Vector Machines (SVM), Figure of Merit (FOM), and Gene Expression Terrain Maps), (d) mapping of gene expression data onto chromosomes, and (e) outsourcing of computational intensive calculations to in-house or remote servers.
6. Tools for promoter sequence retrieval and analysis.

All tools have been incorporated into the gene expression analysis suite Genesis [199-201]. Genesis has been written in Java (Java 2 Standard Edition 5.0 [128]) for optimal platform independency and performance. Extensive work has been undertaken to accomplish program control as well as visualization and handling of data and results in a user friendly and intuitive way. Java3D [202] was used to render very informative three-dimensional representations of results for PCA, CA, and Terrain Maps.

Genesis uses the Java Cluster Service (JCS) and SOAP communication for the three pipelines, sequence retrieval, and promoter analysis tools. The JCS has been installed on the master of the Myrinet Linux Cluster of the Institute for Genomcis and Bioinformatics, Graz University of Technology and has access to 24 calculation nodes (Dual-Xeon 2.6GHz CPUs, 1 GB RAM) with 48 CPUs in total. Calculation results are stored on a 1 TB NetApp Filer attached to the cluster master.

First, the main software components are introduced to get an impression of their functionality. For demonstration purposes these methods are applied to conduct a comparative transcriptomics study of human multipotent adipose-derived stem cells and mouse embryo fibroblasts during adipocyte differentiation.

# 3.2 Sequence Retrieval

Automated high-performance sequence retrieval is an inevitable instrument in order to conduct comparative genomic or transcriptomic studies. The gene expression analysis suite Genesis has been expanded in order to facilitate these needs.

## 3.2.1 NCBI Entrez Sequence Retrieval

Genesis is using a NCBI service called *e-utilities* [203] that provides a programmatic interface to the Entrez search engine. *e-utilities* receive HTTP GET requests from Genesis and returns XML that represents search results. Two of NCBIs e-utilities are accessed:

- *esearch* performs a search and returns a list of IDs.
- *efetch* fetches requested documents in a variety of formats.

Genesis converts its input parameters into GET URLs and uses them to retrieve data from NCBI. The first step in retrieving document information from the Entrez server is to perform an *esearch* with several parameters collected by the input dialog (Figure 3.2). The *esearch* service executes the query on the Entrez server and returns data corresponding to the results of the query. Since the HTTP GET query string includes the parameter "*usehistory=y*", the Entrez server also returns two additional items of data: a *WebEnv* string and a *QueryKey*. A *WebEnv* string is a unique identifier for user state within a session in the Entrez server. This state includes a history of previous queries and their result sets. A *QueryKey* is a small integer that identifies the specific query within the session. Together, the *WebEnv* and the *QueryKey* represent the query result set on the server. A result set is retrieved as described in Fig. 3.1:

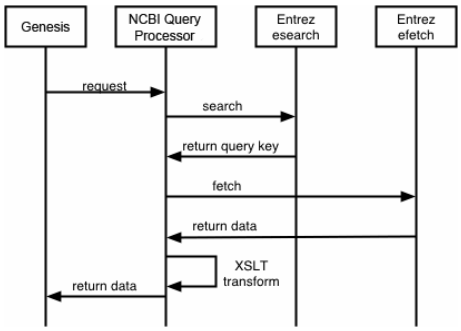
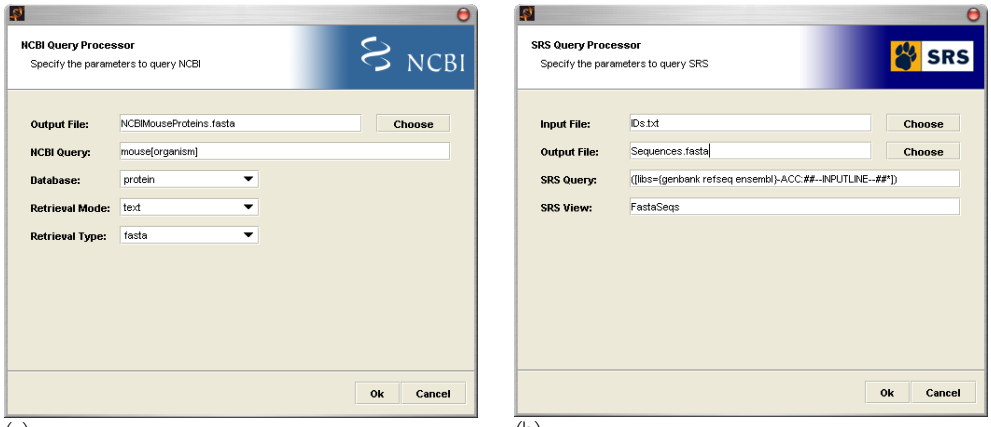



**Figure 3.1: Sequence diagram of Entrez data retrieval:** (1) The NCBI query processor (NQP) receives request parameters from Genesis and executes an HTTP GET with the parameters required by *esearch*. (2) *esearch* sends back XML containing data that identifies the server-side result set. (3) NQP parses the XML returned by *esearch*, and uses DOM API calls to retrieve the *WebEnv* and *QueryKey*. It then uses these values to build a URL to get the data. (4) NQP executes another HTTP GET, this time to *efetch*. It indicates the result set, and includes formatting parameters specified in the original request. (5), NQP receives the requested document data from *efetch*, and transforms the data before returning the results to Genesis.

*efetch* is able to return only parts of the result set enabling the retrieval of large sequence information in multiple steps. This approach is used by Genesis and enables the retrieval of reasonable large numbers of sequences (hundreds of thousands) from NCBI in reasonable time by avoiding connection problems due to large return data sets. A user can specify the output file, NCBI query (e.g. mouse[organism] for all mouse sequences), database (protein, nucleotide, genome, pubmed, pmc, journals, taxonomy, popset), retrieval mode (xml, html, text, asn.1) as well as retrieval type (native, fasta, gb, gbc, gbwithparts, est, gss, gp, gpc, uilist, chr, flt, rsr, brief, docset) rendering this tool to a very powerful and easy to use NCBI information retrieval device.

### 3.2.2 SRS Sequence Retrieval

Input sequence IDs are segmented equally to create a variable number of query tasks, which are distributed to calculation nodes using the JCS. The SRS command line tool *getz* is used to execute the query within SRS. After all jobs have been completed, results are fetched from the cluster, combined, and stored into an output file. SRS query and output view can be specified by the user (Figure 3.2). Due to the parallel execution of the SRS query, this tool is able to retrieve a large number of sequences very efficiently.




(a) (b)

**Figure 3.2: Sequence Retrieval Tool Input Dialogs:** (a) NCBI Query Input Dialog. All proteins of the organism mouse are retrieved into the text file NCBIMouseProteins.fasta using the FASTA format. (b) SRS Query Input dialog. All sequences described by IDs in the IDs.txt file are searched in the databases GenBank, RefSeq, and Ensemble. Sequences are stored into the file Sequences.fasta using the SRS view *FastaSeqs* (FASTA sequence format).

The SRS Sequence Retrieval instrument is usually due to the parallel processing of queries much more performant than the NCBI Sequence Retrieval but relies on the data up-to-dateness of the used SRS system. The NCBI Sequence Retrieval contains of course always the latest data which may outbalance the performance drawback.

## 3.3 Protein Finding Pipeline

In order to find orthologous or paralogous relationships between genes of different organisms, comparisons have to be conducted on the protein level to get reliable result. Unfortunately protein sequences are frequently not available for genes of interest or an investigation is based on EST or oligonucleotide sequences. Therefore it is mandatory to provide a procedure to retrieve the protein sequence of any given nucleotide sequence in a high-throughput, high-performance, and reliable way.

The Protein-Finding-Pipeline described here takes a FASTA file as input and performs very efficiently billions of sequence comparisons to retrieve the optimal protein sequence for a given DNA sequence using the sequence information and annotation of seven renowned sequence databases (Figure 3.3).

The temporal sequence of instructions is performed in the following way:

1. The sequences in FASTA file are indexed and divided into parts, depending on the number of CPUs available for calculation.
2. Input files are transmitted to the cluster (via SSH2 for performance issues).
3. JCS jobs are created for each part and submitted to the calculation cluster. Jobs are immediately sent into the queuing system by the JCS and executed as soon as a CPU is available.
4. The status of all jobs is constantly logged (running, done, queued, failed, undefined, and not determinable) until all jobs are completed.
5. Local output directories are cleaned and jobs are fetched from the JCS via SOAP.
6. NCBI-BLAST DTD (Document Type Definitions) files are copied to each output directory for parsing of the blast results.
7. Jobs (result files) are deleted on the cluster.
8. BLAST xml result files are parsed and hits are analyzed. If a hit DNA sequence has a corresponding protein sequence in the protein part of the currently investigated database the protein sequence is extracted.

The pipeline can be controlled and configured within Genesis. All parameters concerning the pipeline (blast parameters, input and output files, settings how to distribute jobs on the cluster, etc) are defined in an xml file. All results are represented graphically in an intuitive user interface incorporated into the Genesis gene expression analysis environment. All log files are saved for later validation of the results.

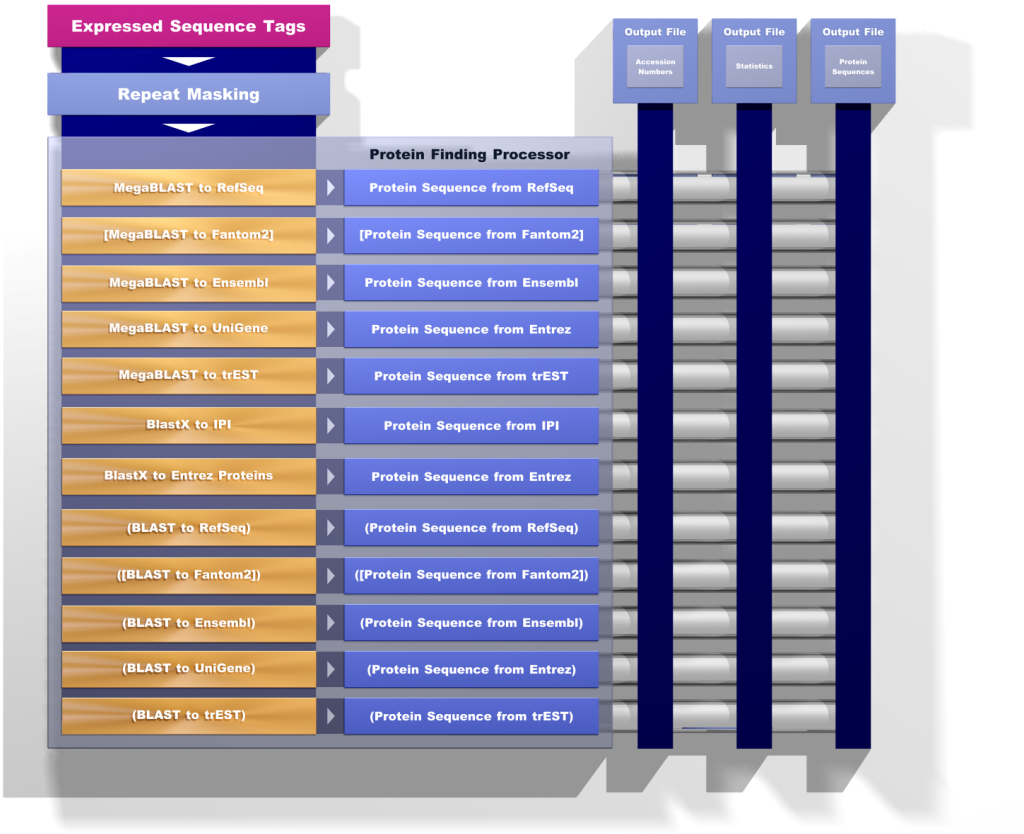



**Figure 3.3: Protein-Finding-Pipeline architecture:** MegaBLAST and BLAST searches against Fantom 2 are performed for mouse sequences only. BLAST searches (bottom five) are used only if performance is not an issue and better sensitivity of blastn is required.

The output of the pipeline comprises:

1. FASTA file with all found protein sequences. The ID of the query nucleotide sequence is stored in the FASTA header for later mapping purposes.
2. Table of corresponding sequence IDs found (e.g. RefSeq IDs, Ensembl IDs, etc).
3. Statistics describing the overall performance of the pipeline for a specific search (number and percentage of nucleotide sequence hits and proteins found, contribution of each database to the overall result, number of sequence comparisons conducted).
4. All blast results organized in a tree structure and graphically processed.
5. Histogram of hit sequence lengths for each database separately.

Currently the pipeline uses seven renowned sequence databases (RefSeq, Fantom 2, Ensembl, trEST, UniGene, IPI, Entrez protein database) and is designed for human and mouse protein searches. However, the pipeline can easily be adapted (using the xml configuration file) in order to handle any other organisms for which sequence databases are available. Fantom 2 is used of course for *mus musculus* only.

# 2.4. Comparative Genomics Pipeline

This Comparative-Genomics-Pipeline has been designed for fully automated retrieval of putative orthologous and inparalogous relations between two arbitrary organisms in general and human-mouse in particular. The pipeline described here takes two FASTA files containing protein sequences (one for each organism) as input and performs an all-versus-all BLAST search to detect the mutually best hits as candidates for putative orthologous relationships. Additional orthologs (inparalogs) are clustered together with each pair of potential orthologs (Figure 3.4). The temporal sequence of instructions is performed in the following way:

1. The protein sequence files of both organisms are transferred to the calculation cluster using SSH2 (for performance issues).
2. Formatdb is executed remotely on the cluster in order to format the protein sequence files (required by blast).
3. The sequences are indexed, divided into parts (depending on the number of CPUs available for calculation) and are transmitted to the cluster (via SSH2). JCS jobs are created for each part and submitted to the calculation cluster. Jobs are immediately sent into the queuing system by the JCS and executed as soon as a CPU is available.
4. The status of all jobs is constantly logged (running, done, queued, failed, undefined, and not determinable) until all jobs are completed.
5. Results are compressed on the calculation cluster in order to optimize transfer time.
6. Job information data (error logs) are fetched from the JCS via SOAP.
7. Job results are fetched from the JCS via SSH2 (compressed files).
8. Jobs (result and information files) are deleted on the cluster.
9. Job result files are decompressed locally.
10. NCBI-BLAST DTD (Document Type Definitions) files are copied to each output directory and blast results are preprocessed to get valid xml files (blastn puts multiple xml files into one file) for parsing. These xml results are parsed and hits are analyzed to extract the necessary information for subsequent comparative analyses.
11. Reciprocal best blast hits are searched and marked as potential orthologs. Moreover, additional orthologs (inparalogs) are added for each main orthologous pair.

The output of the pipeline comprises a table of main orthologous pairs and corresponding inparalogs (columns show sequence IDs, E-Values, and type of relation), as well as all blast results organized in a tree structure and graphically processed. All log files are saved for later validation of the results.

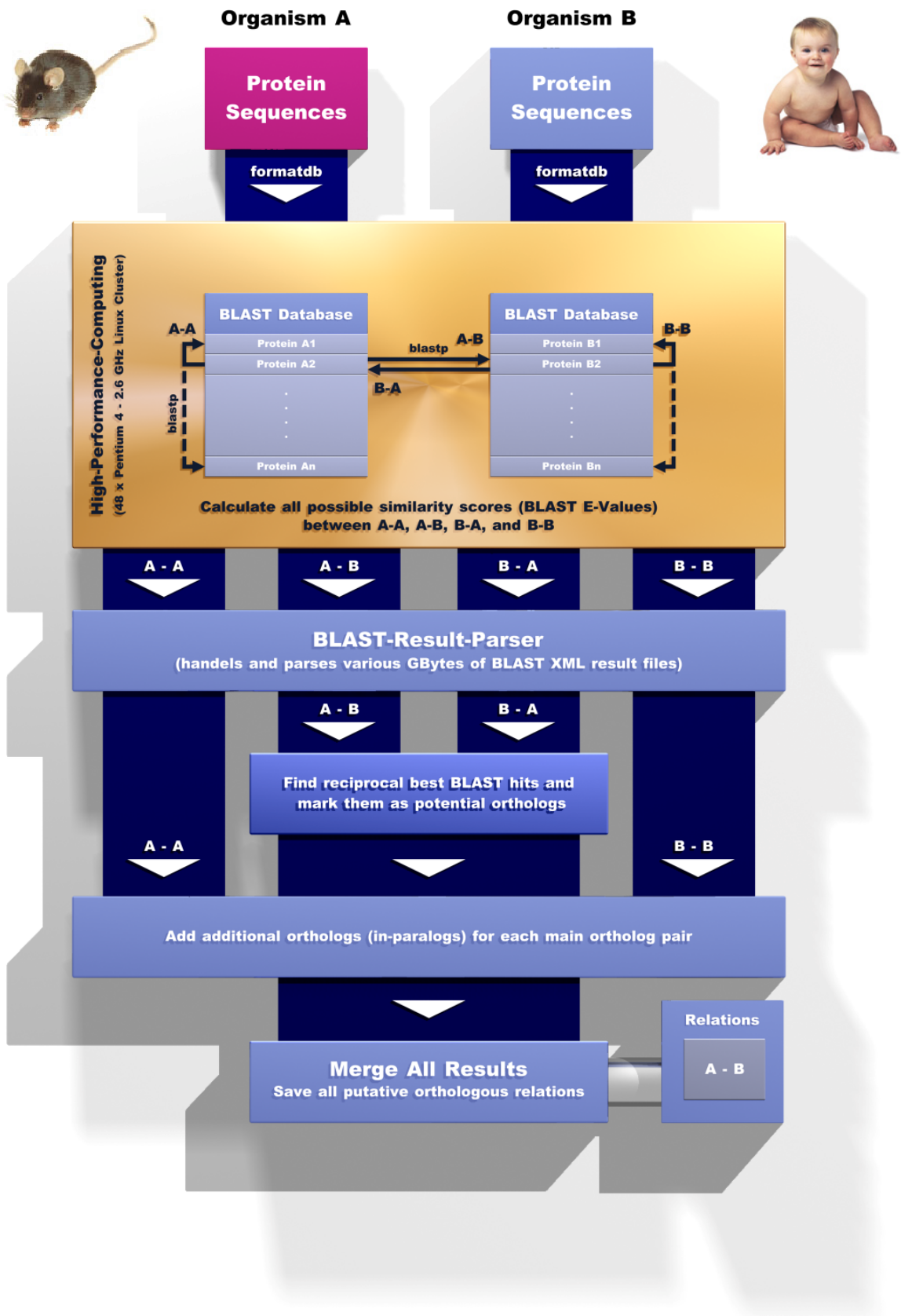



**Figure 3.4: Comparative-Genomics-Pipeline architecture.** The pipeline requires two FASTA sequence files A and B containing protein sequences. An all-versus-all BLAST search is executed and sequence pairs with mutually best hits are detected. Additional orthologs (inparalogs) are clustered together with each remaining pair of potential orthologs.

## 3.5 GO Annotation Pipeline

Comparative transcriptomics additionally requires the incorporation of annotation like gene ontology to facilitate the comparison among genes from different organisms. Mapping, comparing, and combining of gene IDs and annotations from databases containing versatile information about different organisms is a challenging but integral part for conducting these kinds of studies.

The GO-Annotation-Pipeline has been specifically designed to meet these demands by facilitating fully automated gene ontology (GO) annotation for a given set of proteins. The GO-Annotation-Pipeline described here takes a FASTA file containing protein sequences as input and performs very efficiently millions of sequence comparisons to retrieve the optimal GO protein sequence for a given query protein sequence. Once the GO protein has been found, the GO database is queried for the complete GO annotation of this particular protein. This information is used to annotate the query protein sequence (Figure 3.5).

The temporal sequence of instructions is performed in the following way:

1. All sequences stored in the "seq" table of the GO database are retrieved using Genesis and saved as a fasta file.
2. The protein and GO sequence files are transferred to the calculation cluster using SSH2 for performance issues (instead of SOAP).
3. Formatdb is executed remotely on the cluster in order to format the GO sequence file (required by blast).
4. The query-sequences are indexed, divided into parts (depending on the number of CPUs available for calculation) and are transmitted to the cluster (using SSH2 for performance issues). JCS jobs are created for each part and submitted to the computing cluster. Jobs are immediately sent into the queuing system by the JCS and executed as soon as a CPU is available.
5. The status of all jobs is constantly logged (running, done, queued, failed, undefined, and not determinable) until all jobs are completed.
6. Results are compressed on the calculation cluster in order to optimize transfer time.
7. Job information data (error logs) are fetched from the JCS via SOAP.
8. Job results are fetched from the JCS via SSH2 (compressed files).
9. Jobs (result and information files) are deleted on the cluster.
10. Job result files are decompressed locally.

11. NCBI-BLAST DTD (Document Type Definitions) files are copied to each output directory and blast results are preprocessed to get valid xml files (blastn puts multiple xml files into one file) for parsing. These xml results are parsed and hits are analyzed to extract the necessary information for subsequent analyses, as well as filtered (min. 95% HSP identity and min. 50% total hit identity).
12. Best blast hits are searched and the GO database is queried for the complete GO annotation of this particular protein. The annotations are filtered (only species specific hits are allowed) and the information is saved as a GO mapping file.

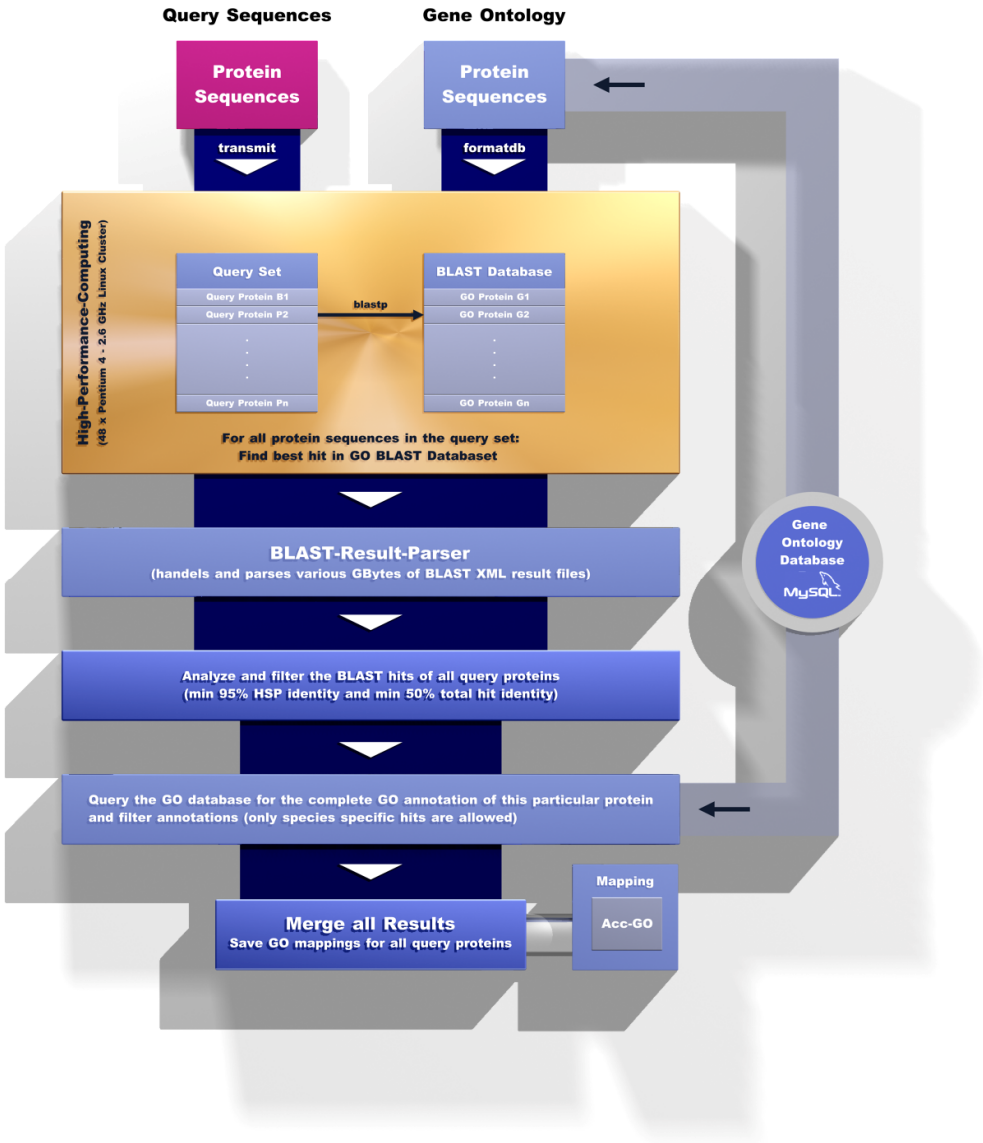



**Figure 3.5: GO-Annotation-Pipeline architecture:** The pipeline requires a FASTA formatted sequence file as input and fully automatically annotates these sequences based on the blast hits found by searching each protein sequence in all GO proteins sequences. Additional filtering is done to gain high-confidence annotation.

The output of the pipeline comprises a tab delimited file containing the GO mapping of query protein sequence identifiers (e.g. accession numbers) and a set of GO term accession numbers associated to that sequence. These mappings can be used to annotate gene expression datasets with GO terms facilitating the much easier evaluation of results. It enables to investigate the dataset on more general functional or biological domains rather than looking at each gene separately.

The GO-Annotation-Pipeline uses the same framework of tools as described earlier for the Comparative-Genomics-Pipeline. The pipeline can be controlled and configured within Genesis and all parameters concerning the pipeline (e.g. filtering parameters) are defined in an xml file. Again, all log files are saved for later validation of the results. The GO database is accessed via JDBC (Java Data Base Connectivity) for performance issues. This enables Genesis to load the complete term tree and additional required information in a matter of a few seconds. Additionally, Genesis has also been endowed with tools to display the GO annotation in context with gene expression data (Figure 3.6).

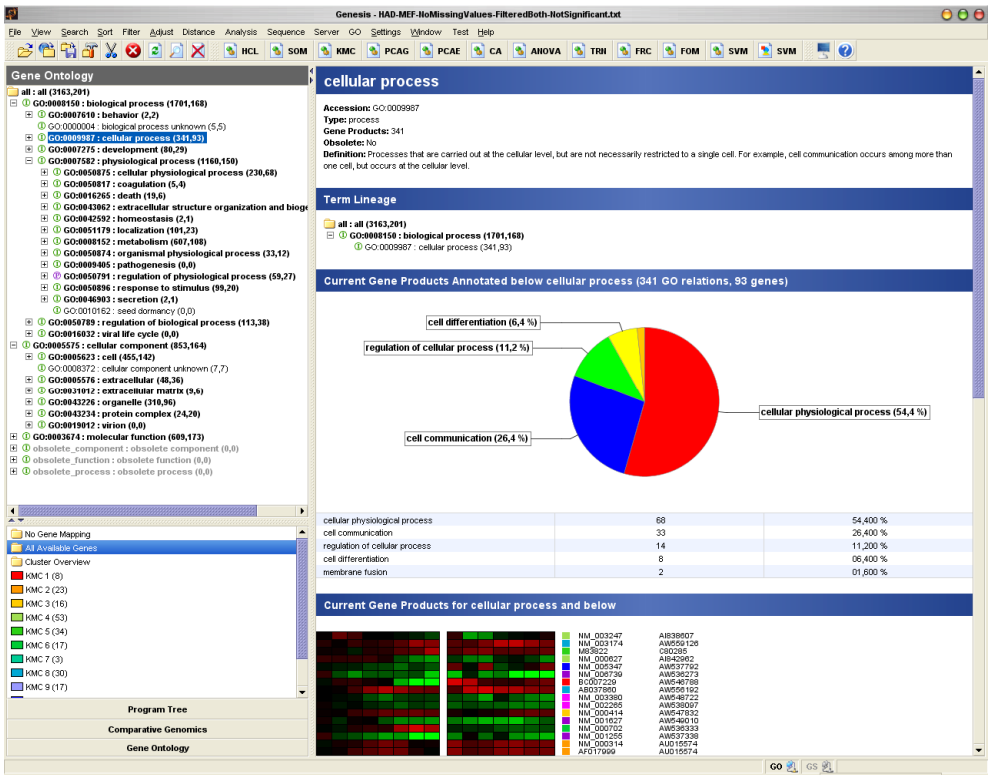



**Figure 3.6: Gene Ontology environment in Genesis:** The top left panel contains the GO tree for biological process, cellular component, and molecular function. Numbers in brakes denote the number of mapped GO terms and unique genes associated with this node. The right panel shows details on a selected term like: general description and GO accession number, term lineage, a pie chart and table showing the distribution of mapped gene products in relation to the children of the selected term, as well as an expression matrix of all genes associated with this node and all sub children. The bottom left panel is used for GO mapping. All available genes or each gene expression cluster separately can be mapped onto the gene ontology.

# 3.6 Comparative Transcriptomics Study

## 3.6.1 Nucleotide Sequence Retrieval

25.569 EST sequences for the TIGR Mouse cDNA chip (MCC) were retrieved from TIGR (The Institute for Genomic Research [204]). 29.550 50mer single stranded oligonucleotide sequences for the Human Oligonucleotide Chip (HOC) were retrieved from data provided by MWG [142]. Both sequence sets were masked for repeats by distributed appliance or RepeatMasker [118] using the JCS and Genesis. Subsequently, sequences were filtered by a minimum length of 50 nucleotides, leading to 25.134 masked sequences for the MCC and 29.127 masked sequences for the HOC. That means that 435 EST and 429 oligonucleotide sequences were filtered out due to a sequence length less than 50 nucleotides (after repeat masking).

## 3.6.2 Protein Sequence Retrieval

Protein sequences were searched using the Protein-Finding-Pipeline described earlier: 29.127 human sequences where searched against six human databases (RefSeq, Ensembl, UniGene, trEST, IPI, and Entrez) comprising 672.364 sequences. 25.134 mouse sequences where searched against seven mouse databases (RefSeq, Fantom2, Ensembl, UniGene, trEST, IPI, and Entrez) comprising 911.660 sequences. Blast parameters were set according to table 3.1 (for the whole configuration file listing see appendix A). 22.626 human nucleotide sequences (77,68% of all oligonucleotide sequences) and 22.498 mouse nucleotide sequences (89,51% of all ESTs) could be found in the databases using a blast expectation value of $10^{-9}$, meaning that the alignment similarity has a 1 to 1.000.000.000 chance of occurring by chance alone. From these nucleotide sequence hits 22.400 human protein sequences (76,90% of all oligonucleotide sequences) and 21.257 mouse protein sequences (84,57% of all ESTs) could be retrieved. Database contributions and more detailed results are shown in Tables 3.2 and 3.3. The 19.583.946.228 sequence comparisons for human and 22.913.662.440 sequence comparisons for mouse have been conducted each in notedly less than one hour on a 48 processor Linux cluster (Myrinet Cluster: 24 Nodes & 1 Master, Dual-Xeon 2.6GHz CPUs, 1GB RAM per Node, 4GB RAM in Master, 127 GFlops peak performance), that are far more than 1.000.000 sequence comparisons per second.

The result is described in a table listing all found database relations with nucleotide and protein sequence accession numbers (nucleotide and protein sequence) as well as corresponding E-Values for each query sequence (Figure 3.7).

| BLAST Type | Blast Parameters |
|---|---|
| Megablast (discontiguous) | -e "1e-9" -F "F" -W "11" -t "18" -p "90" -v "5" -b "5" -m "7" -D "2" -a "1" |
| Blastx (blastall) | -p "blastx" -e "1e-9" -F "F" -v "5" -b "5" -a "1" -m "7" |
| Blastn (blastall) | -p "blastn" -e "1e-9" -F "F" -v "5" -b "5" -a "1" -m "7" |

**Table 3.1: Blast parameters for the Protein-Finding-Pipeline:** (specified in an xml file, parameters except parameters for input (-i), database (-d), and output (-o), which are set automatically by the pipeline processor). MegaBLAST: –e: cutoff expectation value, -F: filtering (false because sequences have already been repeat masked), -W: word size (has to be 11 or 12 for discontiguous megablast), -t: discontiguous word template length (supported template lengths are 16, 18, and 21), -p: cutoff by percentage of identity, -v: maximal number of database sequences to report alignments from, -b: maximal number of reported alignments for a given database sequence, -m: alignment view options: 7 denotes XML Blast output, -D: type of output: 2 denotes traditional BLAST output, -a: number of processors to use. BLASTALL: -p: program name, -e: expectation value, -F: Filter query sequence (DUST with blastn, SEG with others) (false because sequences have already been repeat masked), -v: number of database sequences to show one-line descriptions for, -b: Number of database sequence to show alignments for, -a: number of processors to use, -m: alignment view options: 7 denotes XML Blast output.

| | Blast Search | | Protein Mapping | | Contribution | | BLAST Comparisons | |
|---|---|---|---|---|---|---|---|---|
| Task Name | Found | Fraction | Found | Fraction | Found | Fraction | DBSize | Comparisons |
| HumanRefSeq | 19.375 | 66,52% | 19.334 | 66,38% | 19.334 | 86,31% | 28.898 | 841.712.046 |
| HumanEnsembl | 19.934 | 68,44% | 19.658 | 67,49% | 1.181 | 5,27% | 35.838 | 1.043.853.426 |
| HumanUnigene | 18.120 | 62,21% | 16.782 | 57,62% | 696 | 3,11% | 52.803 | 1.537.992.981 |
| HumanTrest | 6.750 | 23,17% | 6.528 | 22,41% | 282 | 1,26% | 71.277 | 2.076.085.179 |
| HumanIPI | 13.365 | 45,89% | 13.365 | 45,89% | 318 | 1,42% | 46.941 | 1.367.250.507 |
| HumanEntrez | 8.493 | 29,16% | 8.302 | 28,50% | 488 | 2,18% | 247.791 | 7.217.408.457 |
| HumanRefSeqBlastn | 19.457 | 66,80% | 19.413 | 66,65% | 41 | 0,18% | 28.898 | 841.712.046 |
| HumanEnsemblBlastn | 20.008 | 68,69% | 19.727 | 67,73% | 11 | 0,05% | 35.838 | 1.043.853.426 |
| HumanUnigeneBlastn | 17.869 | 61,35% | 16.548 | 56,81% | 17 | 0,08% | 52.803 | 1.537.992.981 |
| HumanTrestBlastn | 6.995 | 24,02% | 6.750 | 23,17% | 32 | 0,14% | 71.277 | 2.076.085.179 |
| | **22.626** | **77,68%** | **22.400** | **76,90%** | **22.400** | **100,00%** | **672.364** | **19.583.946.228** |

**Table 3.2: Protein-Finding-Pipeline result for human repeat masked sequences** of the Human Oligo Chip. Comparison numbers are based on 29.127 query sequences.

| | Blast Search | | Protein Mapping | | Contribution | | BLAST Comparisons | |
|---|---|---|---|---|---|---|---|---|
| Task Name | Found | Fraction | Found | Fraction | Found | Fraction | DBSize | Comparisons |
| MouseRefSeq | 15.916 | 63,32% | 15.908 | 63,29% | 15.908 | 74,84% | 26.195 | 658.385.130 |
| MouseFantom2 | 15.866 | 63,13% | 11.155 | 44,38% | 945 | 4,45% | 60.770 | 1.527.393.180 |
| MouseEnsembl | 15.670 | 62,35% | 15.574 | 61,96% | 560 | 2,63% | 35.247 | 885.898.098 |
| MouseUnigene | 17.047 | 67,82% | 11.774 | 46,84% | 536 | 2,52% | 45.772 | 1.150.433.448 |
| MouseTrest | 21.839 | 86,89% | 20.302 | 80,78% | 3.030 | 14,25% | 198.776 | 4.996.035.984 |
| MouseIPI | 6.828 | 27,17% | 6.828 | 27,17% | 46 | 0,22% | 42.469 | 1.067.415.846 |
| MouseEntrez | 6.810 | 27,09% | 6.665 | 26,52% | 0 | 0,00% | 135.671 | 3.409.954.914 |
| MouseRefSeqBlastn | 16.148 | 64,25% | 16.139 | 64,21% | 138 | 0,65% | 26.195 | 658.385.130 |
| MouseFantom2Blastn | 16.180 | 64,37% | 11.355 | 45,18% | 15 | 0,07% | 60.770 | 1.527.393.180 |
| MouseEnsemblBlastn | 15.906 | 63,28% | 15.812 | 62,91% | 9 | 0,04% | 35.247 | 885.898.098 |
| MouseUnigeneBlastn | 17.456 | 69,45% | 12.051 | 47,95% | 13 | 0,06% | 45.772 | 1.150.433.448 |
| MouseTrestBlastn | 22.049 | 87,73% | 20.502 | 81,57% | 57 | 0,27% | 198.776 | 4.996.035.984 |
| | **22.498** | **89,51%** | **21.257** | **84,57%** | **21.257** | **100,00%** | **911.660** | **22.913.662.440** |

**Table 3.3: Protein-Finding-Pipeline result for mouse repeat masked sequences** of the Mouse EST Chip. Comparison numbers are based on 25.134 query sequences.

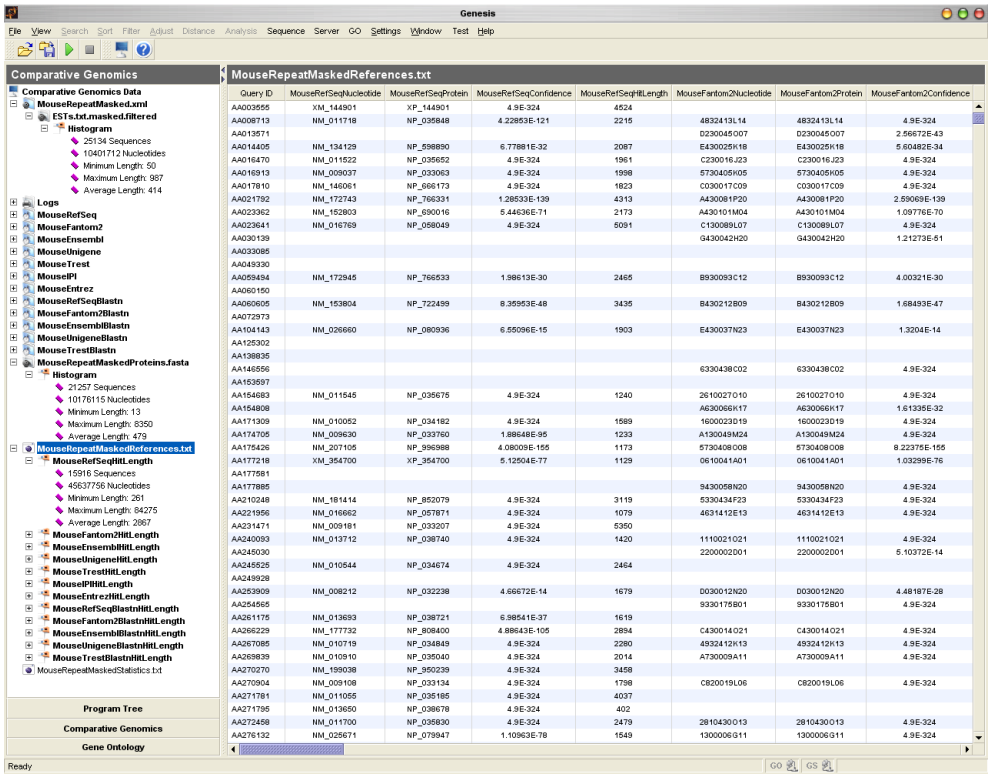



**Figure 3.7: Genesis Protein-Finding-Pipeline result:** Table of found database relations containing accession numbers (nucleotide and protein sequences) and corresponding E-Values for each query sequence.

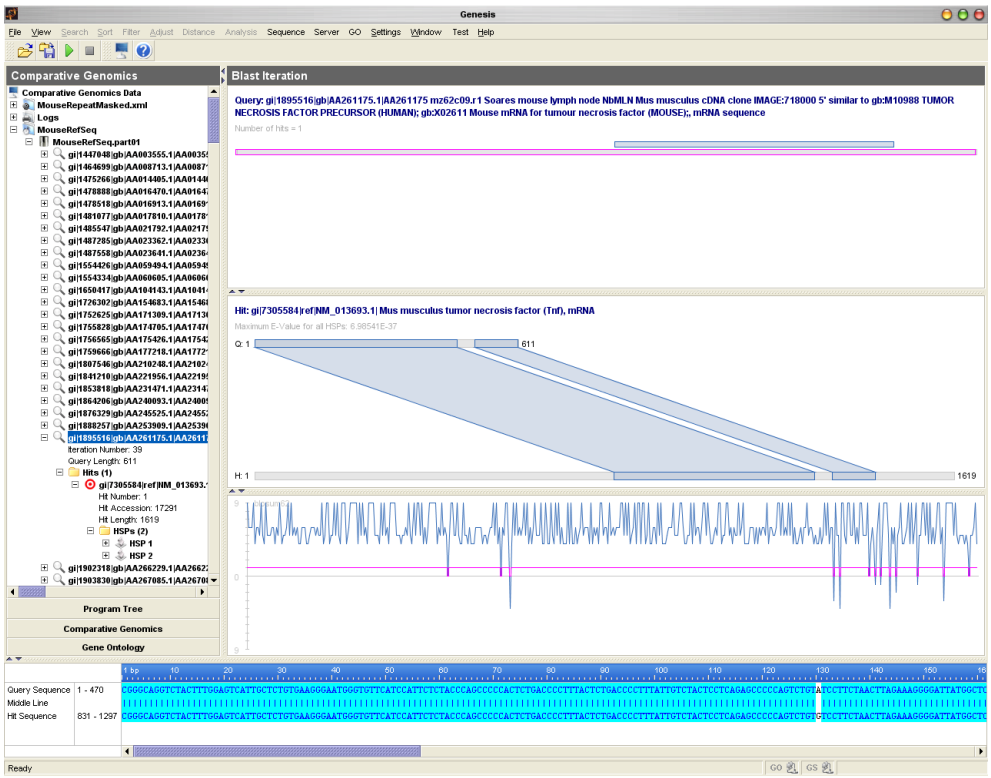



**Figure 3.8: Genesis Blast results viewer:** Mouse RefSeq hit for *mus musculus* tumor necrosis factor (Tnf), mRNA. All blast results are archived and can be visually evaluated very intuitively using this view.

All blast results can be viewed interactively using the Genesis Blast viewer (Figure 3.8). This enables the visual evaluation of blast results, which may be important for quality control issues, e.g. if a promising orthologous relationship has been discovered one can go back and check the blast results the relation is based on.

### 3.6.3 Detection of Orthologous Relations

Putative orthologous sequence relations where analyzed using the Comparative-Genomics-Pipeline described above. 22.400 human and 21.257 mouse sequences from the Protein-Finding-Pipeline were used as input files together with an xml file describing pipeline parameters (see Appendix A). 8.824 putative orthologous and 27.071 putative inparalogous relations could be retrieved leading to a total of 35.895 human-mouse relations for the two microarrays described above. Less than one hour has been required to compute the 952.313.600 protein sequence comparisons on the Linux cluster described above.

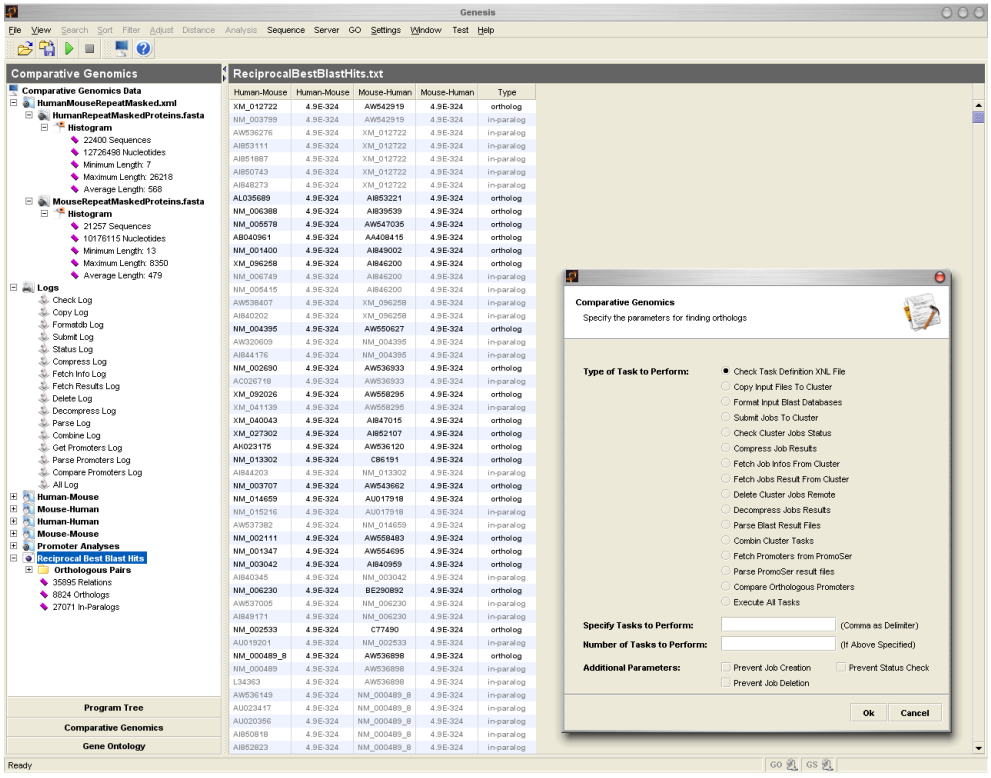



**Figure 3.9: List of orthologous relations:** The list displays all reciprocal best blast hits (black) and corresponding putative inparalogous (gray) relations between the HOC and MCC chips. In this case 8.824 putative orthologous and 27.071 putative inparalogous relations could be retrieved leading to a total of 35.895 human-mouse relations for the two microarrays. Displayed are sequence IDs, the E-Values of the blast hits, and the type of relation (ortholog or inparalog). On the right side the dialog to control the pipeline is displayed. Basically the only input required is to state which task has to be computed (first radio button list). After pressing Ok the task is calculated completely automatically.

| BLAST Type | Blast Parameters |
|---|---|
| Blastp (blastall) | -p "blastp" -e "1e-9" -F "m S" -a "1" -m "7" |

**Table 3.4: Blast parameters for the Comparative-Genomics-Pipeline:** (specified in an xml file, parameters except parameters for input (-i), database (-d), and output (-o), which are set automatically by the pipeline processor). -p: program name, -e: expectation value, -F: Filter query sequence (SEG in this case), -a: number of processors to use, -z: -m: alignment view options: 7 denotes XML Blast output.

## 3.6.4 Gene Expression Data Analysis

Data has been stored in and extracted from MARS. 14.613 human genes of the HOC and 20.220 mouse gene of the MCC had at least one expression value. 20.262 genes of the HOC and MCC arrays could be mapped using the 35.895 human-mouse relations. We denote this genes metagenes. Genes were first filtered and cluster analysis was conducted on the filtered data using: Hierarchical Clustering, Self Organizing Maps, k-means Clustering, Principal Component Analysis, Correspondence Analysis, One-Way-ANOVA, Figure of Merit, and Gene Expression Terrain Maps. All these algorithms have been incorporated into Genesis. Some results are represented here:

### Filtering

10.516 metagenes could be found with at least one expression value in each organism. 4.794 had gene expression values present for all samples in human and in mouse. 425 metagenes had at least one value 2-fold up- or down-regulated in each organism. The rationale of this filtering process is that genes that exhibit little or no variation across investigated conditions do not contribute valuable information for distinguishing among specimens.

### One-Way-ANOVA

One-Way-ANOVA with Euclidian Distance has been calculated using two groups (equal to t-test) containing the HOC or MCC arrays. A p-Value of 0.001 was used to separate these 425 metagenes in 119 metagenes significantly differentially expressed in both groups (human and mouse) and 306 metagenes not significantly differentially expressed in both groups (similar expression in human and mouse) (Figure 3.10).

### Correspondence Analysis

Correspondence Analysis (CA) has been calculated to visualize the dataset and the relation of human and mouse arrays. HOC and MCC arrays are close together and perfectly separated in direction of the x-axis. CA-lines are lines starting at the origin of the graph and intercepting the center of gravity of the two groups of array, HOC and MCC respectively.

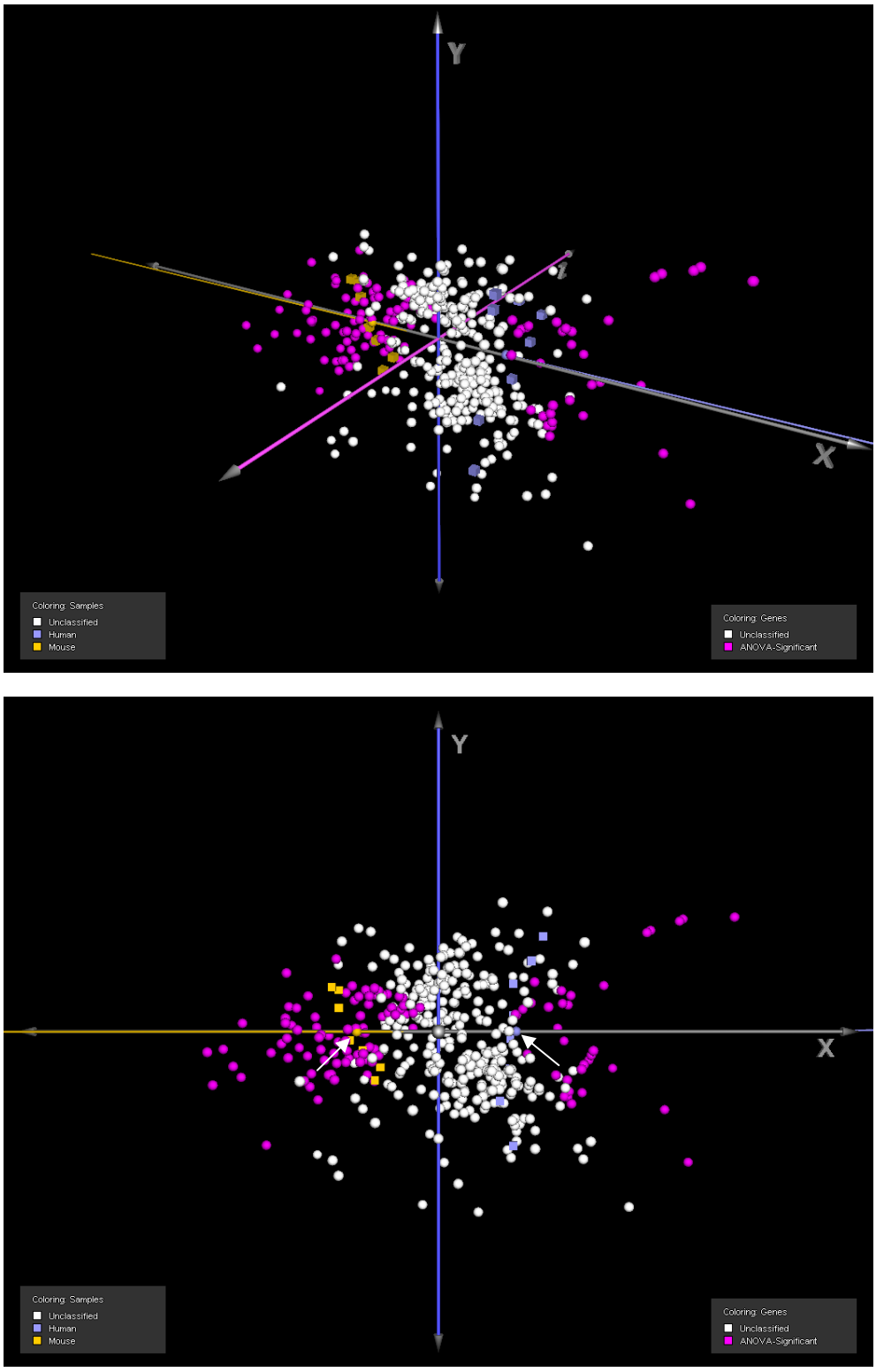



**Figure 3.10: Correspondence Analysis:** Two different view angles of the same result. HOC experiments are displayed as blue boxes, MCC experiments as yellow boxes. The center of gravity for the HOCs and MCCs are displayed as blue or yellow spheres (white arrows, bottom image). Significant differentially expressed genes in Mouse and Human calculated by One-Way-ANOVA with a p-Value of 0.001 are colored magenta. The correspondence analysis lines are almost parallel and very close to the x-axis, separating the two organisms almost perfectly.

The 119 significantly differentially expressed metagenes denoted by ANOVA has been colored to display their location in CA-space (Figure 3.10).

### Figure of Merit

The Figure of Merit (FOM) for k-means clustering of the 306 metagenes has been computed using Euclidean Distance, a clustering range of 1 to 30, and 10 iterations (Figure 3.11). First level off is at 12, second at 17 clusters. However, due to the fact that 17 clusters resulted in some very small clusters, 12 clusters has been chosen for a parsimonious k-means clustering. Additionally, a FOM for each available distance measurement procedure has been conducted in order to obtain the optimal distance procedure for k-means and this dataset. Euclidian Distance was superior to all other distance functions.

### k-means Clustering

The data were analyzed by k-means clustering using Euclidian Distance, which groups genes based on the similarity of their patterns of gene expression. The information of the FOM result indicates that the selected 306 metagenes can be grouped parsimoniously into 12 temporally distinct patterns, each containing between 3 and 59 genes (Figure 3.12).

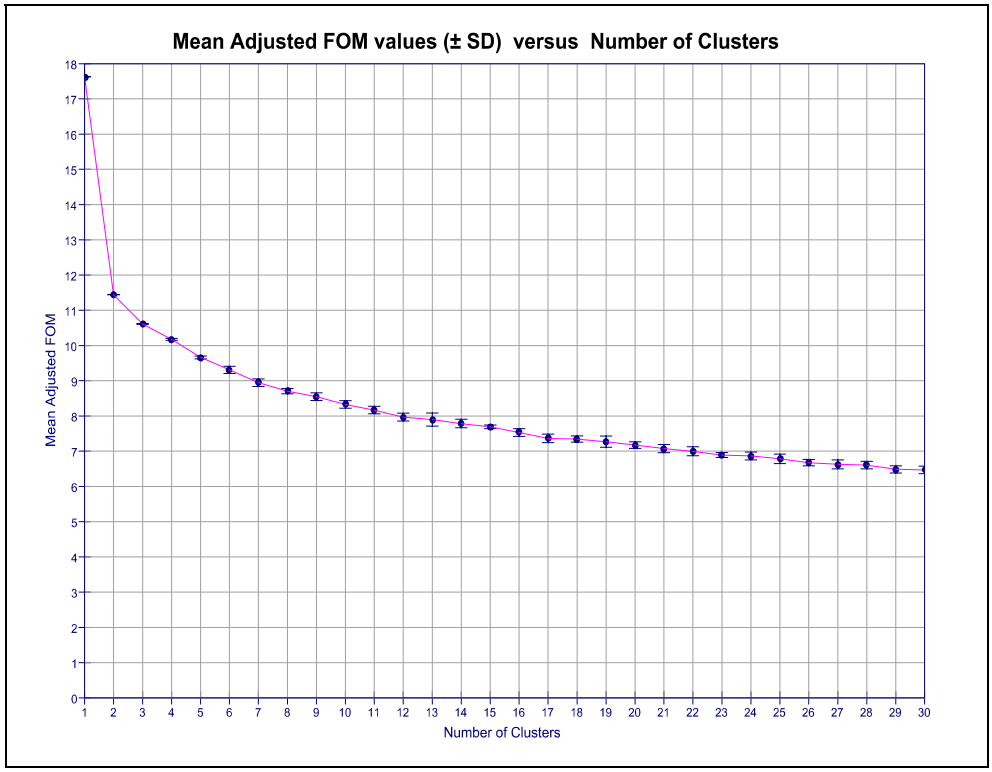



**Figure 3.11: Figure of Merit:** Figure of Merit for k-means clustering of the 306 metagenes using Euclidean Distance, a clustering range of 1 to 30, and 10 iterations. First level off is at 12, second at 17 clusters. The standard deviation of the mean adjusted FOM is also very small for 12 clusters.

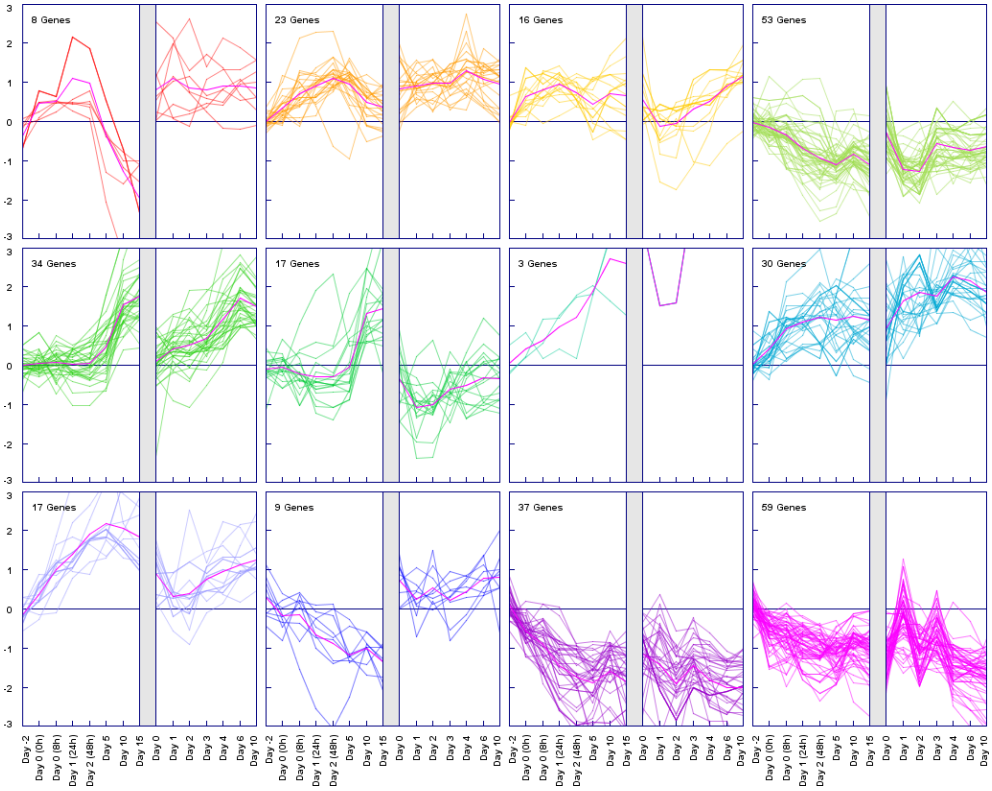



**Figure 3.12: k-means clustering:** Gene Expression Plots (log2 gene expression intensity ratios) and number of metagenes of all 12 k-means clusters obtained using Euclidian Distance.

K-means converged very well to a good solution. Clusters were colored automatically and made public to all other clustering results. Of special interest is k-means cluster number 5: The adipocyte phenotype can be defined by the induction of genes in the late phase of differentiation. In the current experiment clusters comprising genes, which are more abundant in the late time points (5d, 10d, 15d in human and d4, d6 and d10 in mouse) than in the early phases are relevant. Many key players in adipogenesis can be found in this cluster, e.g. the peroxisome proliferator activated receptor gamma (PPARγ), the sterol regulatory element binding protein 1 (SREBP-1), the CGI-45 protein, the pre-B-cell colony-enhancing factor (visfatin), or ponsin (Sorbin and SH3 domain containing 1).

### Self Organizing Maps (SOM)

A 3x4 SOM clustering (12 clusters) using Euclidian Distance was calculated but showed no significant change in comparison to the k-means clustering.

### Principal Component Analysis

A principal component analysis (PCA) has been conducted to evaluate the 12 k-means clusters in PC-Space (space spanned by the first 3 principal components) containing more

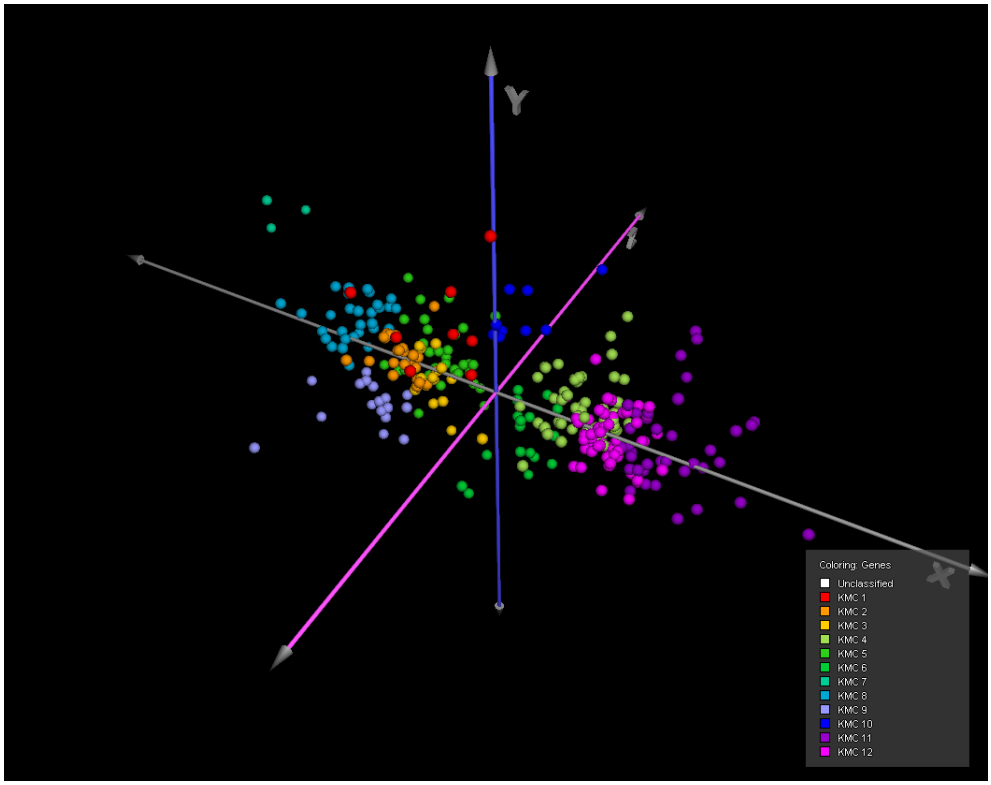


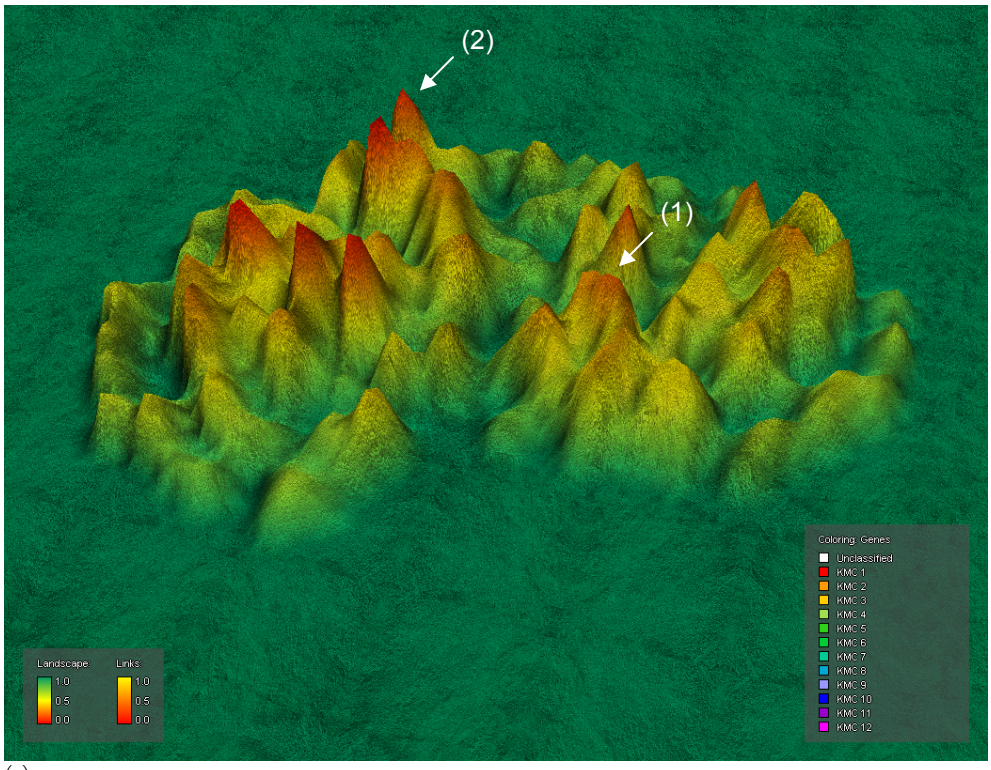

**Figure 3.13: Principal Component Analysis:** 3D-visualization of the first 3 principal components (X=PC1, Y=PC2, Z=PC3), Genes are colored according to previous k-means clustering.

than 89% of the variance inherent in the data. K-means clusters are compact and clearly separated in space supporting the results from FOM and implicating the reliability of the k-means clustering (Figure 3.13).

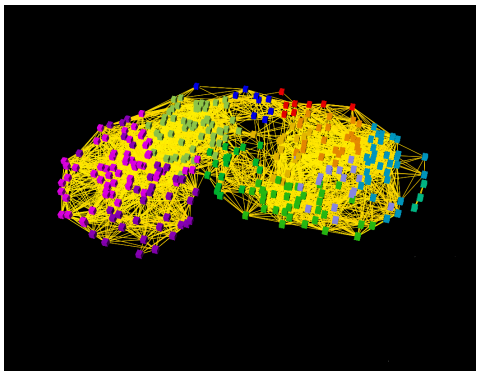
## Gene Expression Terrain Map

A gene expression terrain map has been calculated using Euclidian Distance and 20 neighbors. Metagenes with small distance to each other (indicating a higher significance of conserved coexpression) were placed close to each other, whereas metagenes with larger distances among each other are placed farther apart. The altitude in the final visualization indicates the local density of genes.

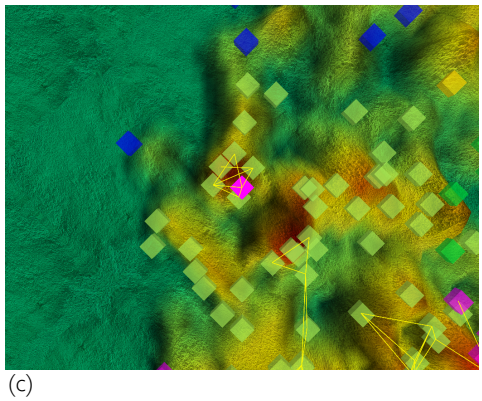
The terrain maps are represented in an interactive, hardware accelerated 3D environment enabling the navigation through the landscape using different behaviors (fly, hover, orbit, etc), selection of genes by clicking onto them in order to get the name and description of this gene, as well as definition of clusters by specifying the number of nearest neighbors of a selected gene. The gene expression terrain map can be displayed in various ways, e.g. using user defined textures, using transparency to see the underlying network, displaying just a wire-frame of the map, displaying the links between genes, and color genes according to other clustering results (Figure 3.14).




(a)



(b)



(c)

**Figure 3.14: Gene Expression Terrain Map:** (a) A gene expression terrain map calculated using Euclidian Distance and 20 neighbors. (1) Genes involved in adipogenesis, e.g. peroxisome proliferator activated receptor gamma (PPARγ), the sterol regulatory element binding protein 1 (SREBP-1), the CGI-45 protein, the pre-B-cell colony-enhancing factor (visfatin), or the Sorbin and SH3 domain containing 1. (2) Many proteins involved in actin binding, e.g. filamin b beta actin binding protein 278, filamin 1 (actin-binding protein-280), or thyroid autoantigen truncated actin-binding protein. (b) The network underlying the gene expression terrain map. (c) Zoom in on mountain (2): genes are very close together and linked to each other. Only links above a certain threshold are displayed, in this case 0.97. All 3 genes mentioned above are present in this sub network.

Distinct mountains for genes involved in adipogenesis or actin binding can be observed by navigating through the terrain or network (Figure 3.14). The adipogenesis mountain contains many genes of k-means clusters 5 but also unknown genes very closely connected to the adipogenesis regulating genes. In general it can be observed that genes, which are known to act together, are often also connected in the gene expression terrain map network.

### 3.6.5 Functional Annotation

22.400 human and 21.257 mouse protein sequences were submitted to the GO-Annotation-Pipeline described above, resulting in 22.398 GO sequence blast hits for human and 21.256 GO sequence blast hits for mouse. Only hits with a HSP identity of 95% and a hit identity of 50% were considered for the annotation. After filtering, 15.503 human proteins (91.480 GO terms) and 13.838 mouse proteins (75.186 GO terms) could be annotated. The human and mouse GO mappings where used to annotate the 306 metagenes described above. 226 (74%) of the 306 metagenes could be annotated with a total of 4.018 GO terms using the human GO mapping, 201 (66%) metagenes with a total of 3.163 GO terms using the mouse GO mapping.

Currently the GO database contains 221.696 protein sequences. The 4.965.990.400 sequence comparisons for human and 4.712.591.872 sequence comparisons for mouse could be conducted each in a little bit more than one hour on the 48 processor computing cluster, that are again more than 1.300.000 sequence comparisons per second.

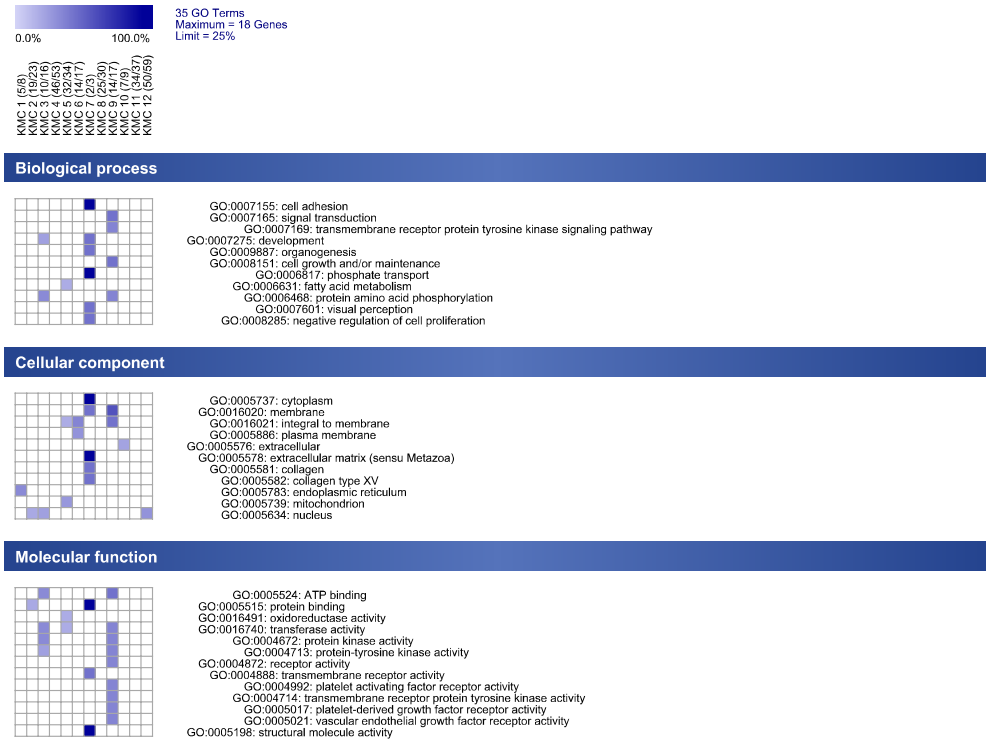



**Figure 3.15: GO mapping overview:** All clusters and the corresponding significant GO annotation are shown. For instance: k-means cluster 5 has 34 metagenes. 32 of them could be annotated. 8 of 32 (25%) metagenes are involved in fatty acid metabolism (GO:0006131). Only GO terms with more than 25% representation in a cluster are displayed here (the limit is variable). The position of GO term descriptions represents their position in the GO tree. "Visual perception" (GO:0007601) for instance is very deep in the tree and so more specific in comparison to GO term "extracellular" (GO:0005578). However, the position of the annotation does not represent a relation to a GO term above the term!

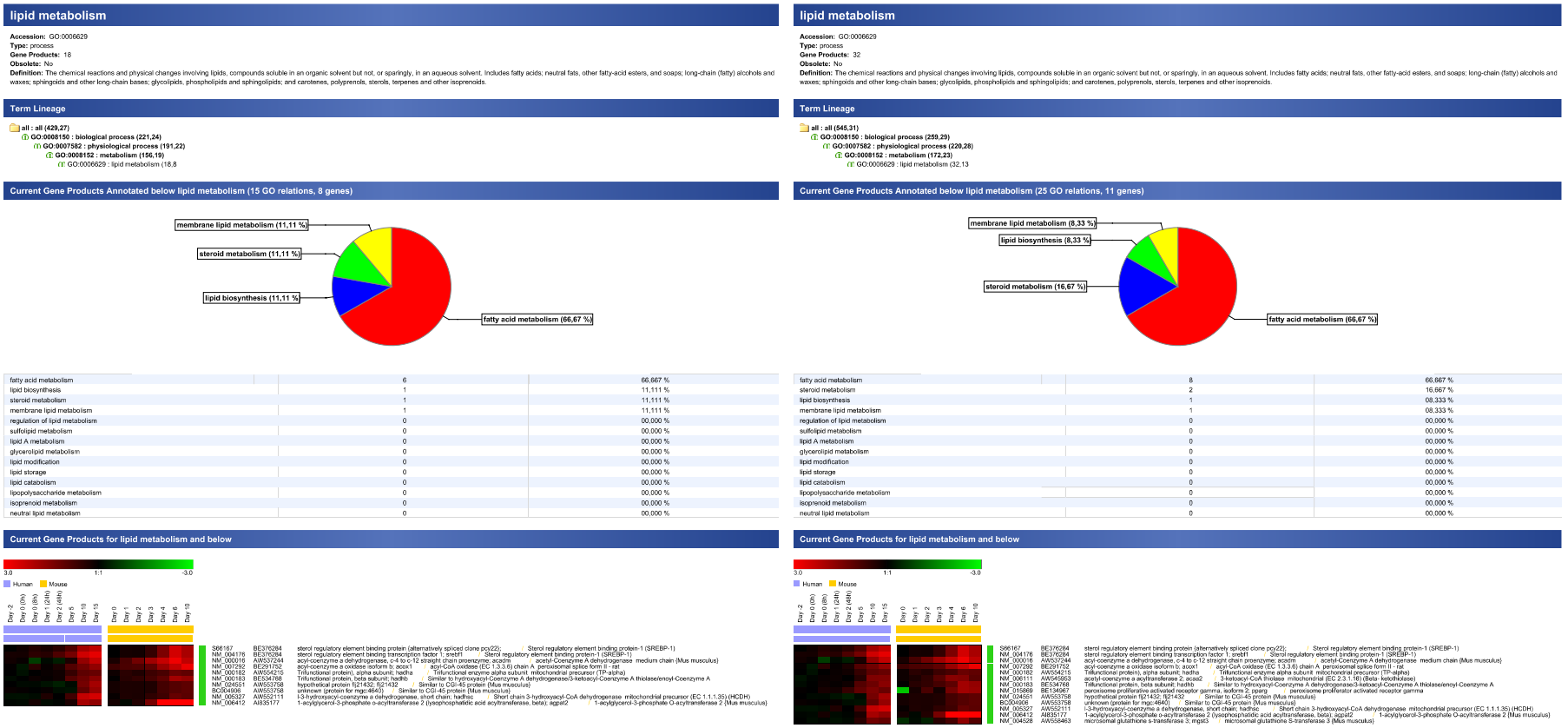



**Figure 3.16: Human (right/top) and mouse (left/bottom) GO mapping for the lipid metabolisms of 34 metagenes from k-means cluster 5.** 32 metagenes could be annotated for human (31 unique human genes), 30 for mouse (27 unique mouse genes). The GO-Annotation-Pipeline for human and mouse genes found to a great extent the same metagenes for lipid metabolism. Genes in k-means cluster 5 show a significant representation of genes annotated for fatty acid metabolism (8 of 32 annotated metagenes for human = 25%, 7 of 31 annotated metagenes for mouse = 23 %).

The k-means clustering showed a good separation of the datasets into different expression patterns. To investigate the function of genes inducing theses patterns, Genesis was used to determine if GO annotated genes in a cluster show an overrepresentation of specific functional behavior, i.e. do specific GO terms occur more often in a certain cluster.

Most of the clustered genes are involved in lipid metabolism as indicated by the GO annotations. Acyl-coenzyme A dehydrogenase is the rate limiting step in the beta oxidation of medium, short, and branched chain fatty acids (Figure 3.17). In the current analysis, besides the Acyl-coenzyme A dehydrogenase medium chain (ACADM), a number of other enzymes important in the decomposition from acyl-CoA to acetyl-CoA during beta oxidation of fatty acids in the mitochondria are highly expressed in the late stage of adipocyte differentiation:

Acyl-CoA oxidase, which uses in opposite to Acetyl-coenzyme A dehydrogenase $O_2$ as substrate and is regulated by PPARs, hydroxyacyl-coenzyme A dehydrogenase, and acetyl-coenzyme A acyltransferase 2.

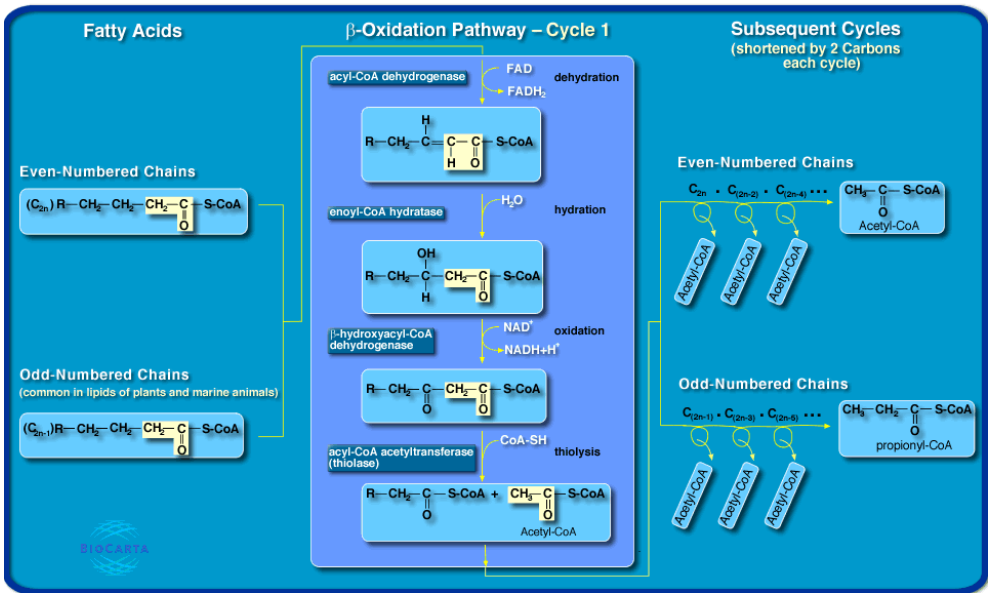



**Figure 3.17: Beta-Oxidation of Fatty Acids Pathway** adopted from Biocarta [205].

Moreover, microsomal glutathione S-transferase 3 is involved in the detoxification of many substrates including fatty acid peroxides derived from lipid oxidation. Lysophosphatidic acid (LPA) is a phospholipid with diverse biological activities and serves as intermediate in membrane phospholipid metabolism. LPA is converted to phosphatidic acid, itself a lipid mediator, by an LPA acyltransferase (1-acylglycerol-3-phosphate O-acyltransferase) [206]

Interestingly, the CGI-45 protein is also expressed in the late phase of human and mouse models for adipocyte differentiation. CGI-45 was recently identified as adiponectin receptor 1 (AdipoR1), which is abundantly expresses in muscle, and that it mediates increased AMP kinase and PPARα ligand activities, as well as fatty-acid oxidation and glucose uptake by adiponectin, a hormone secreted by adipocytes that acts as an antidiabetic and anti-atherogenic adipokine [207].

Adipocyte differentiation is regulated by a transcriptional cascade. A number of transcription factors including members of the CCAAT/enhancer binding protein (C/EBP) family (C/EBPα, C/EBPβ, C/EBPδ), SREBP-1c, and peroxisome proliferators activated receptor γ are play a pivotal role in this process, and it was confirmed that these molecules are necessary for adipogenesis *in vivo*. The latter shown in the present study to be upregulated in late and terminal stage of adipocyte differentiation.

SREBP-1 could potentially be involved in a mechanism that links lipogenesis and adipogenesis, since SREBP-1 can activate a broad program of genes involved in fatty acid and triglyceride metabolism in both fat and liver and can accelerate adipogenesis [208]. The activation of the adipogenesis process by SREBP-1 could be affected via direct activation of PPARγ or through generation of endogenous ligands for PPARγ [209].

PPARγ is a nuclear receptor, which binds as heterodimer together with RXRα to the DNA, regulates and initiates the expression of many genes involved in lipid metabolism, like lipoprotein lipase [210]. Therefore PPARγ (cooperatively with C/EBPα) is known as the key player in the terminal adipocyte differentiation which is responsible for adipocyte phenotype with inclusion of large lipid droplets, demonstrated by Oil-Red-O staining.

The result renders the GO-Annotation-Pipeline superior in comparison to the ID-based mapping strategy of the SOURCE [211] database used before the pipeline has been established. Additionally, while ID-based GO-mappings usually are applicable only to a hand of model organisms (e.g. human, mouse, and rat in SOURCE), the sequence based approach works for all organisms annotated in the GO database.

| BLAST Type | Blast Parameters |
|---|---|
| Blastp (blastall) | -p "blastp" -e "1e-9" -F "m S" -v "20" -b "20" -a "1" -m "7" |

**Table 3.5: Blast parameters for the GO-Annotation-Pipeline:** (specified in an xml file, parameters except parameters for input (-i), database (-d), and output (-o), which are set automatically by the pipeline processor). -p: program name, -e: expectation value, -F: Filter query sequence (SEG in this case), -v: number of database sequences to show one-line descriptions for, -b: Number of database sequence to show alignments for, -a: number of processors to use, -z: -m: alignment view options: 7 denotes XML Blast output.

## 3.4.6 Promoter Analysis

### Promoter Sequence Retrieval

Promoters for 22.785 unique human and 13.776 unique mouse IDs were searched in the PromoSer database via Genesis using the PromoSer SOAP Web Service. 6.282 human and 7.086 mouse promoter sequences (500 bp downstream and 5.000 bp upstream of the TSS, 146 MB fasta files, 152 MB xml information) could be fetched in about 4 hours.

### Dinucleotides and Octamer Analysis

Dinucleotide and octamer distribution of all human and mouse promoter sequences have been calculated and displayed. Both show significant change close to the transcription start site (Figure 3.18 and 3.19).

Initially, the distributions of each dinucleotide (2-mer) in the set of 6.282 human and 7.086 mouse promoter sequences were investigated. The positions of each dinucleotide on the DNA coding strand across the 5.500 bp, between -5.000 and 500 bp, were determined and the results were plotted as a frequency histogram.

Three general distributions can be observed: (1) a peak near the TSS for the 2-mers containing G and/or C (GC, CG, GG, and CC), (2) a valley near the TSS for the 2-mers containing A and/or T (AT, TA, TT, and AA), and (3) no preference for the remaining 2-mers (Figures 3.18 and 3.19). In general, peaks and valley are more significant in human promoters compared to mouse promoters. Although peaking around the TSS, the CG sequence, which can be methylated on the C base, is the least abundant outside the promoter region, as is observed in genomic DNA [212].

To identify DNA sequences that cluster relative to the TSS, the distribution of all 8-mers were determined in this set of human and mouse promoter sequences. There is a clear preference for the clustering factor (CF) to be higher near the putative TSS (Figure 3.20). Additionally the distribution of all 8-mers for the 215 (213 unique) human and 216 (194 unique) mouse promoter sequences of the 306 metagenes analyzed above were calculated and displayed (Figure 3.20).

For the human promoter sequences there is a clear preference for the clustering factor (CF) to be higher near the putative TSS. However, for mouse promoters this behavior could not be observed.

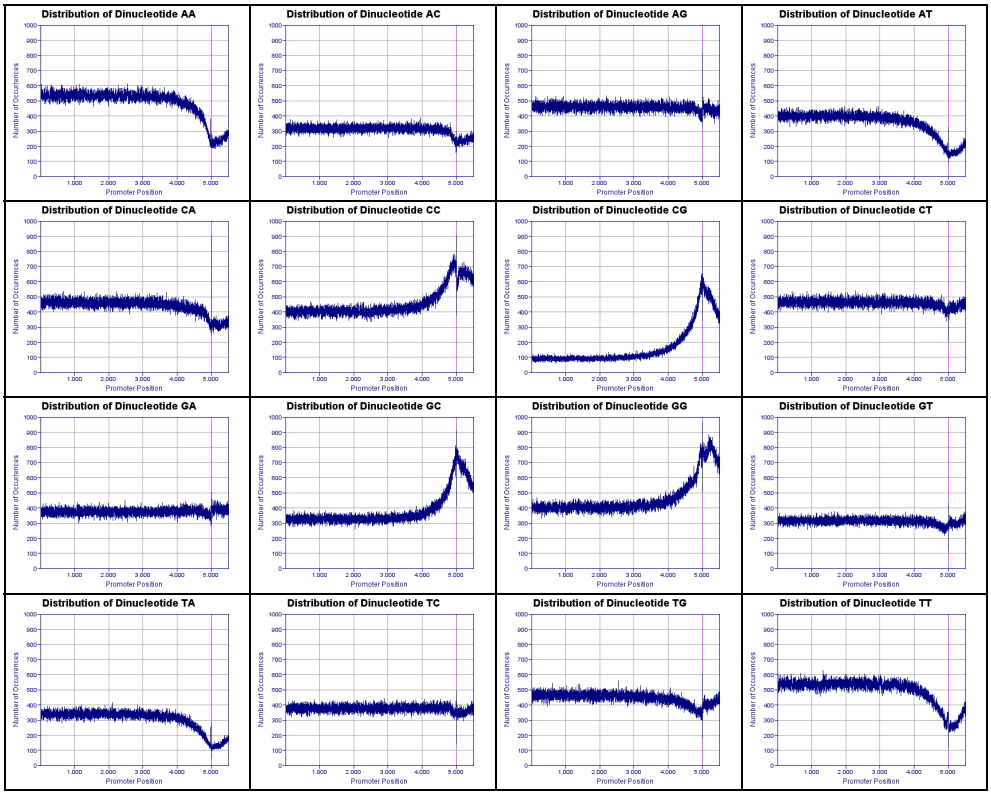



**Figure 3.18: Human Dinucleotides Distribution:** Distribution of all dinucleotides calculated from 6.282 human promoter sequences (500 bp downstream and 5000 bp upstream of the transcription start site).

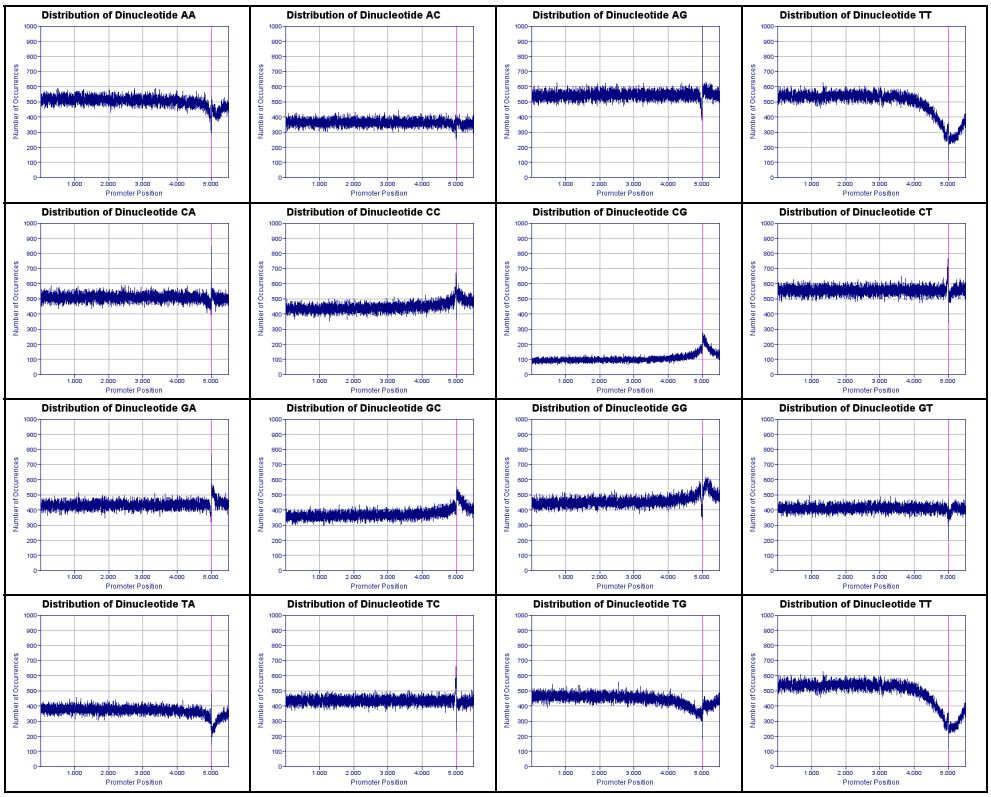



**Figure 3.19: Mouse Dinucleotides Distribution:** Distribution of all dinucleotides calculated from 7.086 mouse promoter sequences (500 bp downstream and 5000 bp upstream of the transcription start site).

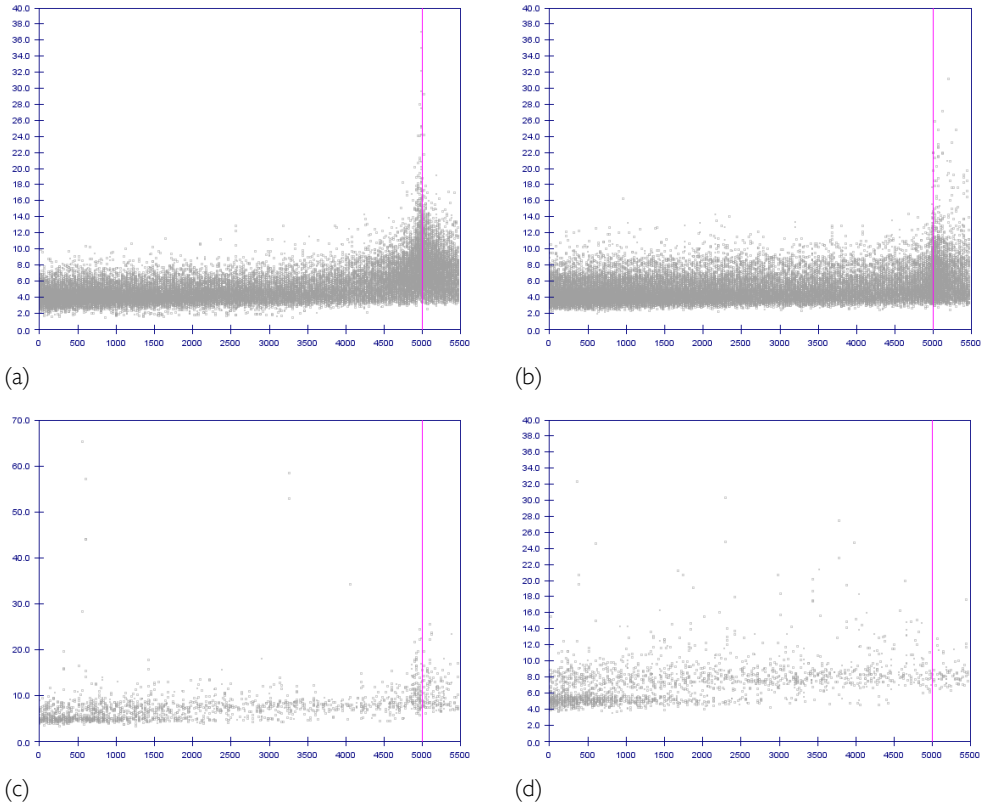



**Figure 3.20: Distribution of all octamers for the human and mouse promoters**: Clustering factor of each 8-mer DNA sequence plotted at the position of the most populated bin: (a) 6.282 human promoters of the HOC chip, (b) 7.086 mouse promoters of the MCC chip, (c) 215 human promoters for the 306 metagenes, and (d) 216 mouse promoters for the 306 metagenes.

All octamers from the promoter sequences of the 306 metagenes with a clustering factor of more than 10 have been extracted. Exactly 250 human and 185 mouse octamers passed this criterion (Table 3.6). All significant octamers of have been aligned to known transcription factor (TF) position weight matrices (PWM). Many octamers show good alignments to TF binding sites known to be active in adipogenesis (PPARγ, E2F, C/EBPα, C/EBPβ, etc.). Sequence logos based on these octamer alignments show considerable similarity with the original PWMs (Table 3.7 and 3.8 for a selection) retrieved from TRANSFAC® [213,214] and Jaspar [215,216].

Subsequently, the total number of occurrences of abundant octamers has been counted within bins of 100 nucleotides in the set containing all related promoter sequences. Figure 3.21 shows the 15 octamers with the highest appearance. Especially the mouse promoter sequences contain many regions of low complexity sequence composed of A and C (CACACACA, ACACACAC) or C and T (TCTCTCTC, CTCTCTCT). In human promoters this characteristic is less distinctive. Also short motifs like GCC, CGG, GAAA or TATA, occur frequently within the promoter sequences. The latter is predicted to be bound by the TATA binding protein [217] that recruits the basal machinery to initiate transcription [218]. Although

repetitive sequences are frequently ignored (masked) in promoter analysis, they may actually contain active control elements by virtue of their specific location [219].

In human the octamer CCCCGCCC can be grouped into the binding sites for the SP1 family of three-zinc finger proteins [195]. The central G is critical, changing it to C results in a sequence that did not occur significantly.

| 215 human promoter sequences | | | | | | 216 mouse promoter sequences | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Nr | Sequence | CF | Nr | Sequence | CF | Nr | Sequence | CF | Nr | Sequence | CF |
| 01 | **GATGGATG** | 65,55 | 51 | **CGGGCGGC** | 15,609 | 01 | **GCGCGCGC** | 32,41 | 51 | **AGAAAGAA** | 13,86 |
| 02 | **TATCTATC** | 58,69 | 52 | **TGGTCCCA** | 15,554 | 02 | **CATTCATT** | 30,45 | 52 | **AGCATTCA** | 13,69 |
| 03 | **GGATGGAT** | 57,37 | 53 | **CGGCGGCG** | 15,544 | 03 | **ATTATTAT** | 27,56 | 53 | **AGGCAAGA** | 13,67 |
| 04 | **ATCTATCT** | 53,06 | 54 | **TTGTTTGT** | 15,469 | 04 | **TCATTCAT** | 24,94 | 54 | **TGGTCCTG** | 13,43 |
| 05 | **ATGGATGG** | 44,20 | 55 | **CCCGCTCC** | 15,295 | 05 | **AAGCGGAG** | 24,78 | 55 | **GTGACCCT** | 13,43 |
| 06 | **TGGATGGA** | 44,17 | 56 | **GCCGGCGC** | 15,281 | 06 | **GTTGTTGT** | 24,66 | 56 | **CCTCTCAG** | 13,41 |
| 07 | **GGGGGGGG** | 34,31 | 57 | **GGCGCTGC** | 14,963 | 07 | **TTATTATT** | 22,90 | 57 | **GAAAGAAA** | 13,40 |
| 08 | **TAATAATA** | 28,47 | 58 | **CCGCCCCC** | 14,956 | 08 | **CAGCAGCA** | 21,43 | 58 | **ATGTATGT** | 13,26 |
| 09 | **GCGCCGCG** | 25,67 | 59 | **CGGGGCCG** | 14,946 | 09 | **TAATAATA** | 21,29 | 59 | **TGTATGTA** | 13,22 |
| 10 | **GGCGCGGG** | 24,58 | 60 | **AGAGCAGC** | 14,940 | 10 | **CCGCCGCC** | 20,83 | 60 | **CCACCACC** | 13,21 |
| 11 | **GCCGCCGC** | 23,74 | 61 | **GCCCGCGC** | 14,868 | 11 | **TAGATAGA** | 20,81 | 61 | **CTTCATCC** | 13,19 |
| 12 | **CGCCGCGC** | 23,45 | 62 | **CCCCGCGG** | 14,856 | 12 | **AAGAAGAA** | 20,75 | 62 | **GGGAGGCA** | 13,04 |
| 13 | **CCGCCGCC** | 23,43 | 63 | **GGGGCGGG** | 14,670 | 13 | **GGATGGAT** | 20,29 | 63 | **TTTATTTA** | 13,03 |
| 14 | **GGCGCCGC** | 22,54 | 64 | **GGGGGGCG** | 14,590 | 14 | **CGGCGGCG** | 20,07 | 64 | **CTTCTTCT** | 13,00 |
| 15 | **GCCGCGCC** | 22,27 | 65 | **CGGGGCGC** | 14,435 | 15 | **ATAGATAG** | 19,64 | 65 | **CCAGAACA** | 13,00 |
| 16 | **CGGCGCGG** | 22,27 | 66 | **GCTGCCGC** | 14,435 | 16 | **TGGTTTGG** | 19,50 | 66 | **CCAGAAGC** | 12,98 |
| 17 | **GCCAATCA** | 21,67 | 67 | **GCAGCCGC** | 14,400 | 17 | **TTTTTTAG** | 19,21 | 67 | **AAGGAAGG** | 12,97 |
| 18 | **CAGCCAAT** | 19,74 | 68 | **GAGCGGGG** | 14,272 | 18 | **TGGATGGA** | 18,74 | 68 | **AGGAGGAG** | 12,95 |
| 19 | **CGGCGGCT** | 19,67 | 69 | **GCCGCCGG** | 14,246 | 19 | **CTGCTGCT** | 18,52 | 69 | **GGGGCGGG** | 12,92 |
| 20 | **GTTGTTGT** | 19,64 | 70 | **CGCCGGGG** | 14,170 | 20 | **CGGAGGCC** | 18,33 | 70 | **TGCAGGGA** | 12,91 |
| 21 | **CCGCCCCG** | 19,29 | 71 | **CCGGGGCG** | 14,170 | 21 | **AAGAAAGA** | 18,00 | 71 | **CATCCCCT** | 12,76 |
| 22 | **GGGCGGAG** | 18,48 | 72 | **TTTGTTTG** | 14,102 | 22 | **CAACAACA** | 17,69 | 72 | **ACATCTGT** | 12,73 |
| 23 | **GCGCGCGC** | 18,07 | 73 | **CTGCTGCT** | 14,024 | 23 | **GATGGATG** | 17,57 | 73 | **TGCTCTTG** | 12,69 |
| 24 | **GCGCCGGG** | 17,99 | 74 | **CTCTCTCT** | 14,015 | 24 | **ATGGATGG** | 17,51 | 74 | **ATTTATTT** | 12,64 |
| 25 | **CAACAACA** | 17,98 | 75 | **CGCGGGCC** | 13,998 | 25 | **AATTAATT** | 16,33 | 75 | **GGACTCTG** | 12,57 |
| 26 | **GCGCAGCG** | 17,95 | 76 | **AGGCGGGG** | 13,973 | 26 | **TTGTTTAA** | 16,30 | 76 | **CCTTCTCA** | 12,55 |
| 27 | **CCCCGCCC** | 17,90 | 77 | **AGGAGGAG** | 13,970 | 27 | **CACTCACT** | 16,11 | 77 | **GTCCTTCT** | 12,53 |
| 28 | **CCAATCAG** | 17,85 | 78 | **ATATATAT** | 13,947 | 28 | **TGGCTGGC** | 16,08 | 78 | **TCTTGTGT** | 12,51 |
| 29 | **AGAAAAGA** | 17,85 | 79 | **CGCCCCGC** | 13,904 | 29 | **GCTGCTGC** | 15,86 | 79 | **CCATCCCC** | 12,46 |
| 30 | **CGCTCCCC** | 17,57 | 80 | **TATATAAA** | 13,542 | 30 | **TCTTCTTC** | 15,72 | 80 | **CCACCAGG** | 12,37 |
| 31 | **GCCCGGGG** | 17,11 | 81 | **ACACACAC** | 13,525 | 31 | **GAGGCCTC** | 15,63 | 81 | **CCCAGTGC** | 12,35 |
| 32 | **CGCGCCGC** | 17,07 | 82 | **CCGGGGCC** | 13,499 | 32 | **CCAGTGGG** | 15,61 | 82 | **CACCACAG** | 12,35 |
| 33 | **GCGCGCGG** | 16,98 | 83 | **GGCTCCGG** | 13,408 | 33 | **GGCCAAAG** | 15,28 | 83 | **TCTTCAGC** | 12,19 |
| 34 | **GGGGCCCG** | 16,87 | 84 | **CTTTTCAT** | 13,384 | 34 | **TGACTGGG** | 15,19 | 84 | **TCTTTTTC** | 12,18 |
| 35 | **GGCGGGGG** | 16,77 | 85 | **CACACACA** | 13,383 | 35 | **TTGTTGTT** | 15,07 | 85 | **TTCCTGAG** | 12,16 |
| 36 | **ATCCTCTC** | 16,65 | 86 | **TATATATA** | 13,364 | 36 | **TTGGTTTG** | 14,95 | 86 | **AGGAAGGA** | 12,16 |
| 37 | **GCCCCGCC** | 16,59 | 87 | **CCCGCGGG** | 13,360 | 37 | **GACAGGTG** | 14,94 | 87 | **CAGTCAGT** | 12,14 |
| 38 | **AGGGGAGG** | 16,51 | 88 | **GGGGCCGC** | 13,289 | 38 | **GTATGTAT** | 14,87 | 88 | **CCTCTTTG** | 12,03 |
| 39 | **CGGGCCGC** | 16,51 | 89 | **GCGGAGGG** | 13,289 | 39 | **GGAAAGGA** | 14,81 | 89 | **AAACCCCA** | 12,01 |
| 40 | **GCCCGCCC** | 16,14 | 90 | **GGCAGCCC** | 13,208 | 40 | **CCCTGGGT** | 14,81 | 90 | **TTTCTAGA** | 11,95 |
| 41 | **CCCGGCGG** | 16,06 | 91 | **CCAGCCCC** | 13,176 | 41 | **ACCAAACC** | 14,62 | 91 | **GCCCCTGG** | 11,86 |
| 42 | **TATTTTAG** | 16,06 | 92 | **CCGCGGGC** | 13,162 | 42 | **GGAGGCCA** | 14,54 | 92 | **CAAAAAAG** | 11,86 |
| 43 | **CCGCGGCC** | 15,95 | 93 | **CAGCAGCA** | 13,103 | 43 | **AGAGCTTG** | 14,36 | 93 | **TTCTCTTT** | 11,84 |
| 44 | **CGCGGGGC** | 15,92 | 94 | **GCCTGTAG** | 13,100 | 44 | **TATATATA** | 14,32 | 94 | **GTGTTTAT** | 11,84 |
| 45 | **GAAAAGAA** | 15,83 | 95 | **GAGAGCAG** | 13,072 | 45 | **GCACACAC** | 14,30 | 95 | **TGGGGGTG** | 11,83 |
| 46 | **TTGTTGTT** | 15,78 | 96 | **AAGAAGAA** | 13,051 | 46 | **GTTTGGTT** | 14,22 | 96 | **ACACAAAA** | 11,80 |
| 47 | **CGCGCGCG** | 15,74 | 97 | **CGGGCCGG** | 13,045 | 47 | **TATTTATT** | 14,10 | 97 | **ATTAAACA** | 11,80 |
| 48 | **GGCGGCGC** | 15,70 | 98 | **GGCCGCGC** | 13,045 | 48 | **GAAAGGAA** | 14,02 | 98 | **CAGCAGTG** | 11,78 |
| 49 | **CGCCGGCC** | 15,68 | 99 | **GCTGGGCT** | 13,009 | 49 | **TATGTATG** | 13,95 | 99 | **CACCACCA** | 11,76 |
| 50 | **TTTATAAA** | 15,65 | 100 | **CCCACCGC** | 12,999 | 50 | **GAGGCCAA** | 13,92 | 100 | **GCCTGCTT** | 11,74 |

**Table 3.6: List of the 100 most significant octamer sequences** in the 215 human and 216 mouse promoter sequences of the 306 metagenes. CF = Clustering Factor. Palindromic octamers are printed bold.

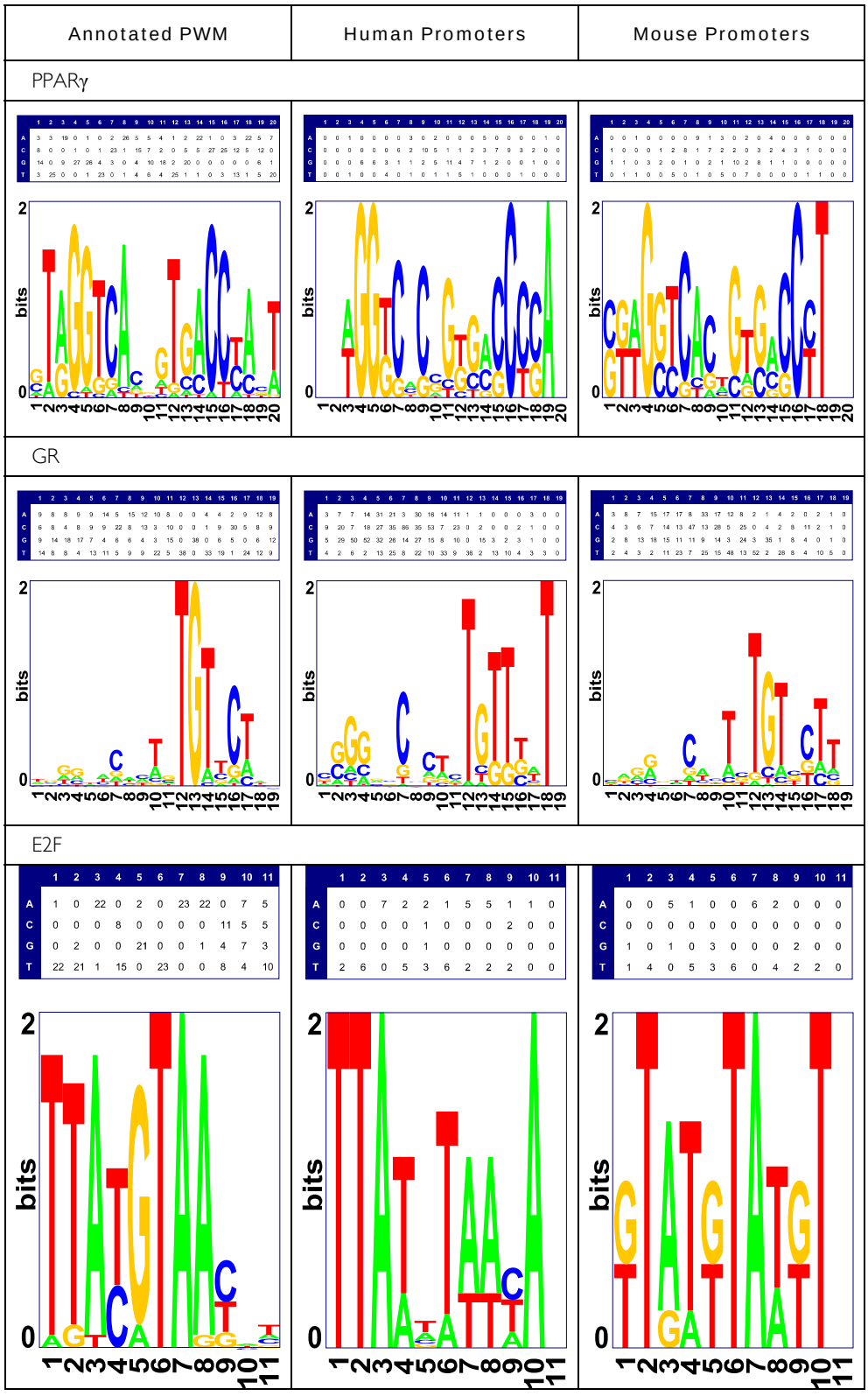



**Table 3.7: Sequence Logos of transcription factor PWMs for PPARγ, GR, and E2F** found in the 215 human and 216 mouse promoter sequences of the 306 metagenes. Many octamers show good alignments to transcription factors known to be active in adipogenesis. Sequence logos based on these octamer alignments have been rendered using Genesis and show considerable similarity with the original PWMs (left column).

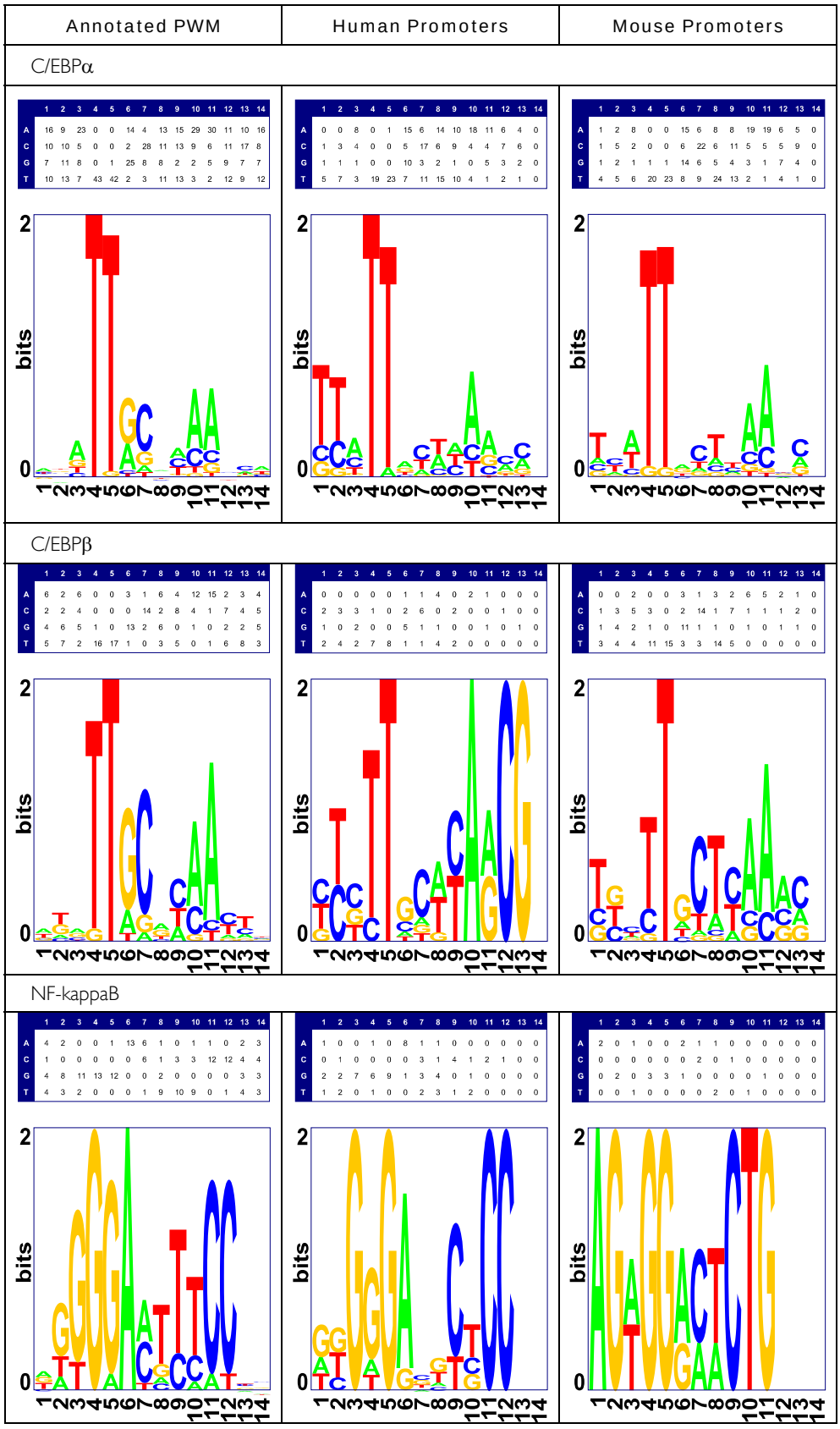



**Table 3.8: Sequence Logos of transcription factor PWMs for C/EBPα, C/EBPβ, and NF-kappaB** found in the 215 human and 216 mouse promoter sequences of the 306 metagenes.

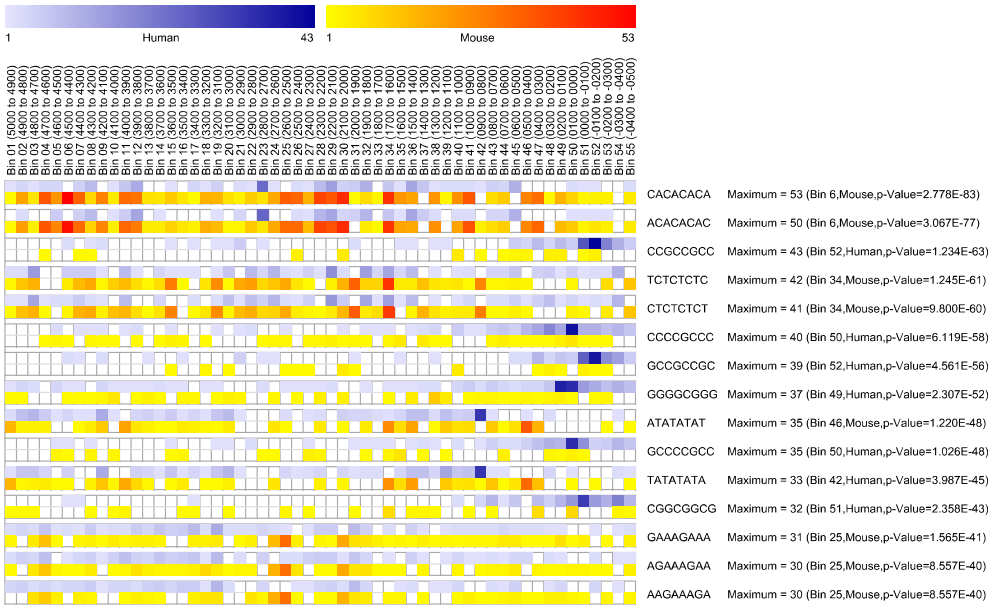



**Figure 3.21: Octamer Location:** 15 most significant octamer occurrences of the 215 human and 216 mouse promoter sequences of the 306 metagenes.

To circumvent the detection of low complexity sequence, octamers which are repetitive in themselves (e.g. TATATATA, GAAAGAAA, GCCCCGCC, etc) have been excluded from successive analyses. The remaining 50 most significant octamers are displayed in Figure 3.22.

Three of the first 16 non repetitive octamers contain an invariant 5-mer CCAAT (CCAATCAG, GCCAATCA and CAGCCAAT).

The CCAAT [220-222] box is a prototypical promoter element, almost invariably found between -60 and -100 upstream of the major transcription start site (TSS), which is clearly indicated by all 3 octamers accumulated in bin 50 (100 to 000 = TSS). It is bound and activated by the histone fold trimer NF-Y, a protein complex that bends DNA by using the histone fold motif [223]. The CCAAT box acts in concert with neighboring elements, and its bending by NF-Y is thought to be a major mechanism required for transcription activation.

The CCAAT/enhancer binding protein (C/EBP) family also plays an important role in the transcriptional regulation of adipogenesis. Regulated expression is seen for several C/EBP family members during adipogenesis (Figure 3.23), and recent gain- and loss-of-function studies indicate that these proteins have a profound impact on fat cell development. In cultured preadipocytic cell lines that have been induced to differentiate, C/EBPβ and δ mRNA and protein levels rise early and transiently [224,225]. C/EBPα, on the other hand, is induced later in the differentiation process, slightly preceding the induction of most of

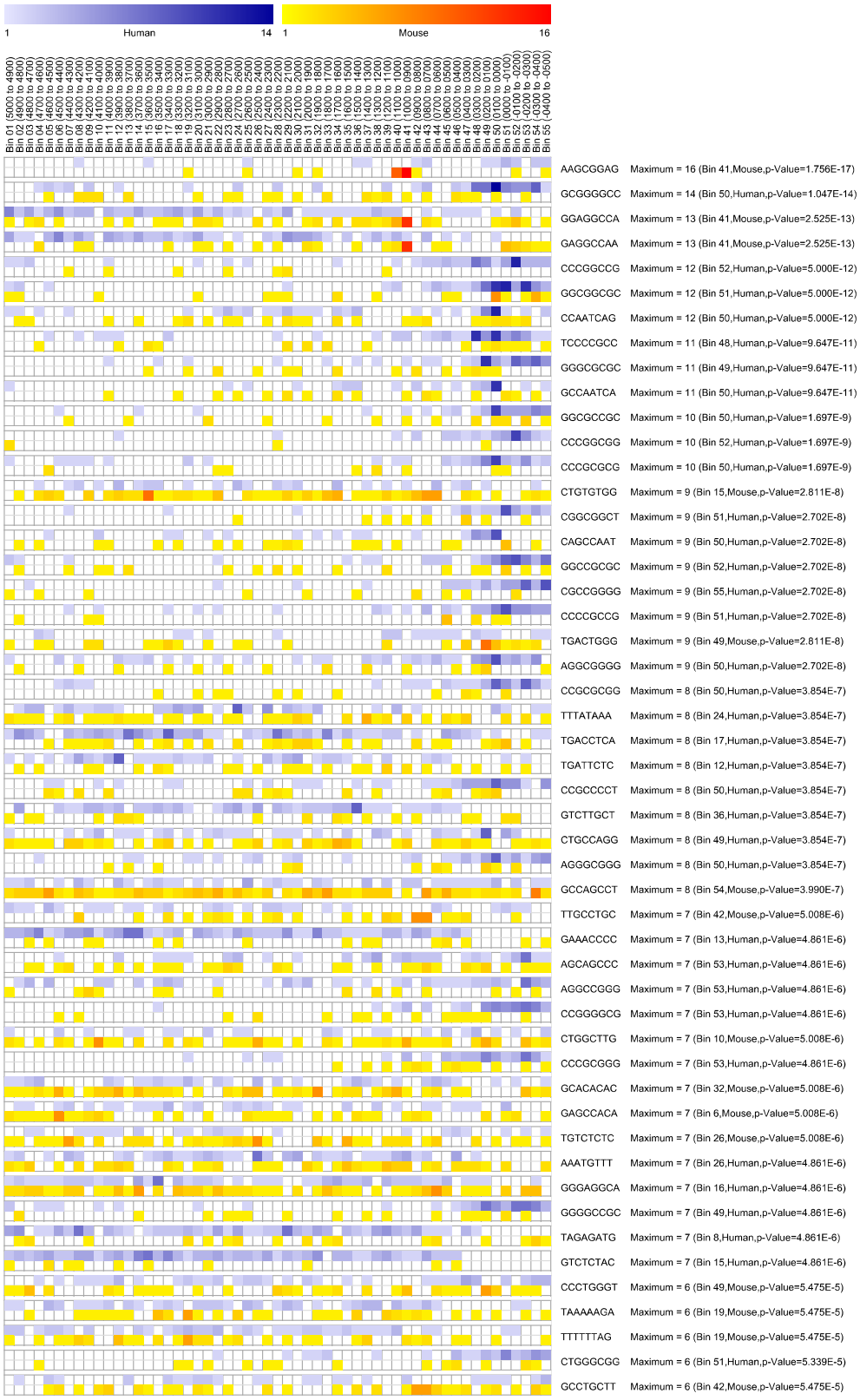



**Figure 3.22: Octamer Location excluding repetitive sequences:** 50 most significant octamer occurrences of the 215 human and 216 mouse promoter sequences of the 306 metagenes. There is a good visible accumulation of augmented occurrences of certain octamers around the transcription start site of the human promoters. Mouse has a very significant accumulation at bin 41 (1000 to 900 upstream) of the octamers AAGCGGAG, GGAGGCCA, and GAGGCCA.

the end-product genes of fat cells. The inhibitory C/EBPζ (also known as CHOP10 or GADD153), on the other hand, is suppressed during the induction of differentiation, but returns when differentiation has progressed almost to completion [226]. This isoform may therefore act as a brake on the adipogenic program after important events have been initiated [227].

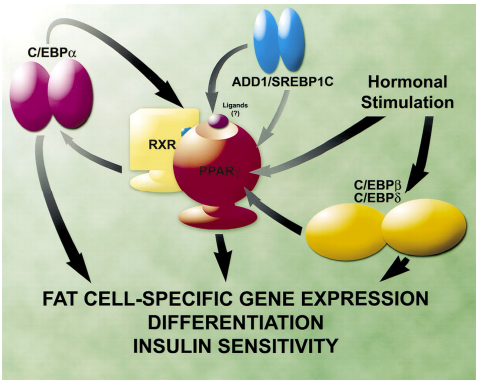



**Figure 3.23. The transcriptional control of adipogenesis** involves the activation of several families of transcription factors. These proteins are expressed in a network in which C/EBPβ and C/EBPδ are detected first, followed by PPARγ, which in turn activates C/EBPα. C/EBPα exerts positive feedback on PPARγ to maintain the differentiated state. ADD1/SREBP1 can activate PPARγ by inducing its expression as well as by promoting the production of an endogenous PPARγ ligand. All of these factors contribute to the expression of genes that characterize the terminally differentiated phenotype [227].

## DNA Block Alignment

4.695 metagenes have promoter sequences for human and mouse. DNA block alignments have been calculated using the DNA block aligner (DBA) of the wise package from EBI. Due to the high calculation costs of the DBA more than 7 h were required to calculate the results on 48 processors. However, if not done in parallel, calculation would have required more than two weeks on a single CPU computer.

123 of the 306 metagenes had promoter sequences for human and mouse and a block alignment could be calculated. Many metagenes show good alignment results. 42 genes had at least 5 DNA blocks. In this group, the histone family member promoters are for instance very well conserved in human and mouse (Figure 3.24).

DNA block alignments can be displayed with the expression values, cluster affiliation, and gene annotations in one graph. The alignments can be sorted according to (1) number of blocks (representing evolutionary conservation), (2) gene expression similarity between human and mouse arrays (representing coexpression), and (3) cluster affiliation (representing general expression patterns).

**Figure 3.24: DNA Block Alignments of Human-Mouse Promoters:** Evolutionary conservation of promoter regions is illustrated by this figure. Metagenes are sorted according to k-means cluster affiliation. Only promoter alignments containing at least 5 DNA Block Alignments are displayed. Each DNA Block Alignment has its own color. The magenta line represents the transcription start site. Histone family member promoters are for instance very well conserved in human and mouse. First numbers represent the number of blocks.

**Figure 3.25: PromoterWise Alignments of Human-Mouse Promoters:** Evolutionary conservation of promoter regions is illustrated by this figure. Metagenes are sorted according to k-means cluster affiliation. Only promoter alignments containing at least 7 block alignments are displayed. Each block has its own color. The magenta line represents the transcription start site. The first number represents the number of blocks for these two promoter sequences. NM_000183: *homo sapiens* hydroxyacyl-coenzyme a dehydrogenase/3-ketoacyl-coenzyme a thiolase/enoyl-coenzyme a hydratase (trifunctional protein), alpha subunit; hadha, AW554215: *mus musculus* Hydroxyacyl-Coenzyme A dehydrogenase/3-ketoacyl-Coenzyme A thiolase/enoyl-Coenzyme A hydratase (trifunctional protein), alpha subunit (Hadha)

### PromoterWise

4.695 metagenes have promoter sequences for human and mouse. Block alignments of all these sequence pairs have been calculated in parallel using Ewan Birney´s PromoterWise program, the JCS, and Genesis. The calculation costs for this computation are considerably lower compared to the DNA Block Aligner. 3.079 promoter pairs had at least one block alignment.

Eighty-six of the 306 metagenes had promoter sequences for human and mouse and a block alignment with at least one block. Many metagenes show 10 or more alignments, 42 genes had at least 7 blocks.

The genes NM_000182 (*homo sapiens* Hydroxyacyl-Coenzyme A dehydrogenase / 3-ketoacyl-Coenzyme A thiolase / enoyl-Coenzyme A hydratase (trifunctional protein), alpha subunit (HADHA)) and AW 554215 (mus musculus Hydroxyacyl-Coenzyme A dehydrogenase / 3-ketoacyl-Coenzyme A thiolase / enoyl-Coenzyme A hydratase (trifunctional protein), alpha subunit (Hadha)) for instance are almost identically expressed and share 17 blocks (third largest number). Both genes are involved in the beta-oxidation of the fatty acid pathway.

Results are displayed in the same way as described above for DNA block alignments results (Figure 3.25).

# 4 Discussion

The functional annotation and identification of genes involved in the development and progression of complex diseases is a cumbersome and non trivial task. DNA microarrays allow generating a composite picture of the expression profile of a cell and are widely used in basic research as well as in clinical medicine and pharmacogenomics. Additionally, evolutionary conservation is a powerful criterion to identify coregulated genes that are functionally important. Coregulation of a pair of genes over large evolutionary distances implies that the coregulation confers a selective advantage, most likely because the genes are functionally related. Exploiting the comparisons of the human genome with other genomes at both the distal and proximal evolutionary edges of the vertebrate tree is expected to represent a powerful tool in the puzzle of decoding molecular mechanisms underlying development or disease.

Therefore, a comprehensive, efficient, and easy to use bioinformatics platform for large-scale transcriptomic studies has been developed. It facilitates comparative analyses of human diseases and corresponding mouse models by integrating gene expression data with genome sequence information.

The specific achievements of the systematic approach represented here are threefold: First, a set of representative transcriptomic datasets describing mouse embryo fibroblasts and human multipotent adipose-derived stem cells during adipocyte differentiation has been produced, annotated, as well as stored in an organized and easily accessible way within a microarray database management system. Second, sophisticated computational tools are provided within a bioinformatics platform for large-scale comparative transcriptomic analyses to distinguish the similar from the dissimilar and analyze these data in a straightforward, efficient, and reliable way. Several methods are proposed to derive meaningful biological information and distributed high-performance computing is used to facilitate these types of large-scale data analyses in reasonable time. Third, comparative analyses of the human and mouse cell lines described above have been conducted with contingent new insights into the universality as well as the specialization between the most important model organism mouse and the designation of all clinical research, the human.

### Sequence Retrieval

Automated high-performance sequence retrieval is an inevitable instrument in order to conduct comparative genomic or transcriptomic studies. Sequences can be retrieved using NCBI Entrez or the local SRS service. The SRS sequence retrieval instrument is due to the

parallel processing of queries an order of magnitude faster but relies on the refresh period of the databases incorporated into SRS. On the other hand, Entrez sequence retrieval is due to the nature of Entrez always up to date, which may outbalance the performance drawback. However, problems may result from restrictions applied by NCBI for very large or frequent sequence retrievals.

### Protein-Finding-Pipeline

In order to solve the problem of finding the corresponding protein sequences for any given nucleotide sequence, two ways are present to choose from. (1) Using already existing sequence databases filled with annotated sequence information (e.g. GenBank, Ensemble, etc.) and using these annotations and links to other databases to retrieve a protein sequence (hopping from one database to another, using conjoint accession numbers or ids) or (2) using strategies based on sequence comparisons. For this thesis the latter approach was chosen because of the following reasons:

1. Databases mostly contain information for a limited number of frequently used model organisms only. Moreover, many databases are specialized to serve a specific scientific topic and contain only a certain set of required data. Sequence based approaches are not limited to organisms or annotations stored in these databases and therefore present a more general approach. Relationships can be found between any sequenced organisms stored in locally or publicly available databases.

2. Databases are evolving constantly and these evolution leads to a lot of versioning problems: file formats change, ids change, entries may be removed from the database resulting in broken links, etc. This requires a constant adaptation of software relying on these databases. Sequence search based approaches face theses problems in a lot alleviated form, since sequences change not so often considerably, and so they can be found easily using a search algorithm insensitive to little changes (e.g. BLAST).

3. Databases usually present a static state of the knowledge present during the last update. If databases are not updated regularly they may not present state-of-the-art data. Unfortunately one has no influence on the update cycle of a database. Having a sequence based approach in house enables the researcher to update their data as soon as new sequence data is available. The automatic course of action of the pipeline also helps to do this regularly.

For these reasons the author believes that the sequence based approach presented here is more suitable to the needs of the pipelines presented in this thesis and justify the drawbacks, namely: high computational costs, challenging handling of large amounts of data and large storage requirements for intermediate results, as well as uncertainties due to the statistical mode of operation. Results are always based on statistical significant thresholds (e.g. BLAST E-Values); they are not curated and therefore represent no proven facts. The latter issue is definitely the most critical one and represents the greatest drawback of this approach. However, many visualization tools have been incorporated to check the results a promising hypothesis might be based on. BLAST alignments with an E-Value of less than $10^{-300}$ may even be more reliable and transparent than id-based results from public databases which also may be based on sequence comparisons done by the database vendor.

Although BLAST is fast and detects biologically relevant homologies reliably, it should be used with caution. The main problem for the presented ortholog detection algorithm is that BLAST reports local similarities. The orthologs are expected to share sequence similarity over the entire length, or at least over the majority of their length. Additional problems with the BLAST output appear with sequence pairs that have two or three separate regions of sequence similarity. This happens with many sequences whose N terminus and C terminus are conserved, but the conservation in the middle of the sequence is too low to be reported by BLAST. The BLAST output segmentation could be addressed by setting the dropoff value for gapped alignment (-X) on the command line, which would cause the BLAST program to report longer segments of similarity. However, currently it has been found that ignoring the non-conserved areas and summing only the conserved segment scores is more realistic [234].

An important point to consider is the length of found hit sequences. Finding a 200kbp BAC (bacterial artificial chromosome) sequence and extracting a protein related to that sequence does not make sense and has to be prevented. Although, databases containing mainly mRNA sequences are used, the pipeline implements methods to filter hit sequences based on BAC or genomic DNA sequence. Additionally histograms of sequence length are drawn for each database separately enabling the examination of the hit sequence lengths distribution.

### Comparative Genomics Pipeline

In order to solve the problem to generate sets of orthologous relationships between two or more organisms, there are three ways to choose from: (1) Using already existing databases filled with annotated orthologous genes (e.g. Clusters of Orthologous Groups (COGs) defined by NCBI [228-231], or Eukaryotic Gene Orthologs (EGO) defined by TIGR [232,233]), (2) detecting orthologs by construction of phylogenetic trees, and (3) using strategies based on all-versus-all sequence comparisons [175,234].

Automatic detection of orthologs and inparalogs from full genomes is an important but challenging problem. As orthologs, by definition, are related through evolutionary history, phylogenetic trees are the most natural way to detect orthologs. Unfortunately, construction of phylogenetic trees involves some poorly automatable steps and demands large resources of computing power. Carrying out this approach for all genes of two or more genomes would require clustering of homologs, generation of correct multiple alignments for each group of homologous domains, construction of a phylogenetic tree for each group, and finally extraction of orthologs from these trees. Approaches for automating the final step exist [235], but current methods for automatic generation of multiple alignments of domains still yield sub-standard quality output, which makes subsequent orthology analysis unreliable.

Kim et al [175] compared the sequence based approach to the COG and EGO databases. They gained the insight that they could not use the definition of orthologs found in the COG database because some sets of orthologs contained a large number of genes from a single organism. For example, in some cases over 100 genes from *C. elegans* were grouped together. Having a large number of genes from a single organism complicates the gene correlations: a single human gene would have correlations to each of the 100 worm genes in the same orthologous group. Additionally, they did not use the EGO database because the same gene was sometimes assigned to separate orthologous groups. For example, tentative orthologs 336024, 350993 and 402694 each contain the same yeast gene encoding nuclear transport factor 2. Having multiple orthologous groups again complicates gene correlations since a gene from one organism would have correlations for each group. These findings as well as the drawbacks mentioned for the Protein-Finding-Pipeline can also be applied to the work in this thesis.

Therefore, for this paper the third approach was chosen because of the reasons equivalent to the articles mentioned above. The idea is that if sequences are orthologs, they should score higher with each other than with any other sequence in the other genome. This approach does not use multiple alignments or phylogenetic trees and therefore avoids potential errors that might be introduced at these steps. The method represented here is very similar to the approach proposed by Remm et al [234] and is designed for inparalog and ortholog identification. Outparalogs are not reported. The methodology can be seen as an extension of the all-versus-all technique, but with special rules for cluster analysis in order to extract all in-paralogs.

Most orthology detection approaches simply identify mutually best matches. The author believes that such an approach is too limited for eukaryotic genomes, and that it is important to identify additional orthologs (inparalogs). One approach that does include paralogs is the COG system, which has much in common with the method presented here. However, in the

final Clusters of Orthologous Groups (COGs), no distinction is made between inparalogs and outparalogs. The main reason for this is that the COG database strives to flatly group orthologs from all species together, while the approach described in this thesis considers only two species or lineages at the time. In fact, a COG must consist of at least three species. Sequences with unidirectional best hits to members of a COG are added later, representing potential inparalogs. However, because the orthology in a COG is not defined to a particular evolutionary point, both inparalogs and out-paralogs may be added.

In contrast to COGs, this approach is limited explicitly to two species or lineages only, which allows us to define the evolutionary point of the orthology precisely, and to separate in-paralogs from outparalogs. Defining the evolutionary point of the orthology is important, because the number of branches increases during evolution due to sequence duplication. As a result, the number of orthologous groups between closely related species is expected to be greater than the number of orthologous groups between distantly related species. That is the reason why the pipeline is preferred to be limited to the comparison of two species or lineages and do not attempt to create flat groups of orthologs covering many lineages or the whole tree of life. Groups of orthologs from more than two species can still achieved by considering a lineage of multiple species as a kind of "superspecies".

### GO Annotation Pipeline

The result presented earlier renders the GO annotation pipeline superior in comparison to the id-based mapping strategy of the SOURCE [211] database used before the pipeline has been established. For instance, while the GO mapping retrieved with the pipeline shows many genes of k-means cluster 5 assigned to the fatty acid metabolism, and the gene annotation supports this results (see 3.6.5), the SOURCE GO mapping did not lead to this insight. Additionally, while id-based GO-mappings usually are applicable only to a hand of model organisms (e.g. human, mouse, and rat in SOURCE), the sequence based approach works for all organisms annotated in the GO database.

### Comparative Transcriptomics Study

Although this work is dedicated to develop a bioinformatics platform for large-scale comparative transcriptomics and was mainly focused on the computer science side, its power has been demonstrated by conducting a comparative transcriptomics study of mouse embryo fibroblasts and human multipotent adipose-derived stem cells during adipocyte differentiation.

Even though the potential of the information contained in large and diverse genome-wide expression profiles is well reorganized, the extraction of meaningful biological knowledge from such data remains a challenging task.

The multi species gene-coexpression analysis differs from previous gene-expression compendiums in two major ways: (1) the multiple-species results only pap those genes that have orthologs in other species and thus can focus strongly on core, conserved biological processes, (2) interaction in multiple-species results imply a functional relationship based on evolutionary conservation, whereas interactions using data from single species only indicate correlated gene expression.

An important and valuable feature of this software is to calculate and compare clustering results from different algorithmic approaches. One of the challenges in analyzing microarray data is the fact that there is no biological definition of a gene cluster. Moreover, due to the different underlying assumptions for the clustering techniques and the necessity to adjust various parameters, the clustering results can differ substantially. Thus, it is an imperative to apply several clustering techniques on the same data set and to compare the results. The comparison of clusters obtained using several clustering techniques enables the researchers to identify genes and/or experiments that have been rated similar in all clustering results. For example, one can begin with a Hierarchical Clustering or FOM to get a first impression on the number of patterns hidden in the dataset and then use this information to adjust the parameters for k-means and SOM clustering. PCA and CA can be used to visualize these clusters in 3D space and to get an impression on cluster size, integrity, and distribution, and to retrieve the most significant patterns in a study. It can also reveal some information about the number of clusters in the dataset, provided that data clouds of genes in the principal component space representing a cluster can be distinguished. All these clustering procedures enable us to get an impression of what subset of genes or experiments represents the most significant information for a given investigated condition and therefore provides the opportunity for researchers to concentrate on particular target genes or experiments.

The user input in all this mathematical approaches is very important, since different parameters can lead to different results and normalization can change the data in an unintentional way. For that reason the investigating researcher should be familiar with the mathematical procedures and the effects they can have to the biological information gained from such analysis methods. Improper analysis can and will yield to false results and wrong conclusions. It is also essential that results received from clustering analysis can only be seen as an indication of possible relationships. Verifying these presumptions in biological experiments is indispensable!

### Promoter Sequence Analysis

An enigma in eukaryotic promoter analysis is that not all DNA sequences that can be bound by a transcription factor are biologically relevant. However, it can be suggested, that if a

particular DNA sequence is observed in the same position relative to the TSS, it is likely that the individual DNA sequences that comprises a cluster are important for regulating gene expression of their promoters. Many of the most significant octamers can be aligned to known transcription factor consensus sequences. Although this approach permits us to identify sequences that are likely to be biologically relevant, it does not necessarily imply that related DNA sequences are not important. It could simply be that the related sequences are not sufficiently abundant to form a peak.

Eight of 100 human and 6 of 100 mouse octamers are palindromic although only 0.3% (256/65.536) of all octamers are palindromic, with a lot more being palindromic in all but one nucleotide position. According to Vinson et al [195] two properties of palindromes may explain their predominance as important transcription factor (TF) binding sites. First, palindromes can be bound on either strand of DNA, thus doubling their concentration and increasing the number of productive encounters between the TF and the DNA. Second, palindromes can be bound by dimeric proteins. Monomers dimerize to double their local concentration and now bind palindromic DNA that, again, is in a higher concentration because it can be "viewed" on both strands of DNA. Both of these effects make palindromic sequences "attractive" structures for TFs to bind.

The author is aware of that this promoter analysis is by no means comprehensive and a more in-depths analysis has to be conducted to get further insights into the regulatory mechanisms of the selected coexpressed genes. Transcription factor binding sites for instance have to be put in context to each other to check if there are significant modules of transcription factor binding sites working together. However, it gives a first glance on the constitution and conservation of promoter sequences retrieved from coexpressed and orthologous genes.

### Implementation

Distributed high-performance computing has been used and is mandatory to facilitate these types of large-scale data analyses in reasonable time. Using the computing power of the local Linux cluster it is possible to conduct a complete comparative genomics study (finding protein sequences, retrieving orthologs and inparalogs, annotate the data with GO terms, and perform promoter analyses) within one day. Conducting theses elaborate tasks on a single CPU would require days or weeks of calculation and do not really present a practically feasible option. Therefore, protocols of communication have been chosen in a way that enables the access of the computation environment also from distant locations and through firewalls, enabling researchers with a lack of local computer power to use the equipment provided by core facilities.

Challenging is also the handling of the vast amount of data generated by this kind of studies. The survey presented in this work accumulated almost 8.000 files representing a data volume of more than 29.4 GB. Very efficient parsers for BLAST xml results and sophisticated algorithms have been implemented to perform the analysis as efficiently as possible. For instance, by using efficient hash map structures, the octamer promoter study could be improved by a factor of more than 32.000, finishing after less than a minute instead of more than 100 hours.

A major objective of this work was to accomplish program control as well as visualization and handling of data and results in a user friendly and intuitive way. The software suite has been tailored to meet the specific needs and skills of researchers with biological or chemical but not necessarily with computer science background. All results are graphically represented, using two or three dimensional graphics often coupled with an interactive environment to survey the results. Examples are for instance the 3D representation of PCA, CA, or gene expression terrain maps, the graphical BLAST result viewer, visual illustration of DNA block alignments or the drawing of sequence logos. All graphic can be saved as pixel images in various formats or as vector graphics for incorporation in documents describing the data.

The program has been developed using the latest Java and Java3D technology in order to maintain platform independency, which is very important in life science since researchers often use different platforms like WindowsXP, Linux, or MacOS X. Java enables the use of very advanced computer hardware like multiprocessor servers and high performance workstations with large memory resources, which may be necessary for the analysis of large datasets. The software has been tested on PCs under WindowsXP, Linux, and MacOS X, as well as on multiprocessor servers under Solaris and Linux without any adaptation of the code. Although Java is not as fast as native code, the time consumed for calculating a result is little on state of the art computers. Java3D is based on OpenGL [236] or DirectX [237] and uses the hardware acceleration of modern graphic cards, enabling the very efficient rendering of complex three dimensional structures. Additionally, smart compilers, well-tuned interpreters, and just-in-time byte-code compilers brings performance close to that of native code without loosing the advantage of portability.

### Outlook

The techniques introduced in this thesis have proven to be useful to unveil meaningful biological information but they can be just the starting point in the analysis of transcriptomic data. Some issues are at the moment not fully addressed and require further investigation. Substantial investment and effort will be necessary to develop new and improve existent analytic methods that elucidate more complex correlations and dissimilarities between

transcriptomes of different organisms. The algorithms presented here can only be seen as a further step towards a more extensive and proper analysis system of the complex processes of gene expression and regulation in living cells.

The field of microarray technology and transcriptomic research in general is developing at an astonishing rate. It will be necessary to continuously improve and adapt the developed software to the newly gained knowledge and standards to maintain the status of a valuable and flexible software package in functional genomics. Some possible improvements for the near future include:

- A broader statistical environment for the analysis tools: Results should be corroborated by P-Values, error models, permutation or bootstrap approaches, comparisons with random data, etc.

- Incorporation of additional filters for pipelines, e.g. domain-level matches should be avoided by forcing the matched area to be longer than a certain percentage of the longer sequence. This should avoid finding hit sequences that share only short domains.

- Application of clustering tools to resolve overlapping groups of orthologs in the Comparative-Genomics-Pipeline.

- Incorporation of the GOA database [238,239] from EBI into the GO-Annotation-Pipeline in order to further improve results for human, mouse, and rat.

- Incorporation of additional promoter analysis and visualization tools to facilitate a more in-depths analysis of promoter sequences.

- Export of all results into the gene expression database MARS.

Although these results suggest the potential of systematic comparative analysis in functional genomics, the author expects that future work will improve these results. For example, the development of methods to systematically assign genes to "regulons" [48,240] may make possible regulon-based measures of correlation that could be more sensitive and specific in their identification of analogous biological programs. The integrative use of expression data from different species is an emerging area of research [48-51,241-244], and elements of these different approaches might be combined to develop additional analytical tools.

Comparative functional genomics could be a powerful way to distinguish the essential from the species-specific features of biological processes, such as disease, stress and development. Aided by growing repositories of expression data (e.g. MARS, ArrayExpress [155], or Gene Expression Omnibus [245]) and conventions for reporting genomic experiments [146], measures of correlation in searchable databases could identify new analogies among disease states, mutant strains, and drug responses in diverse organisms.

## Conclusion

The software described in this thesis called Genesis has been tailored to become a comprehensive, versatile, user-friendly, and platform independent Java suite for comparative transcriptomic data handling and analysis. It has been developed using state-of-the-art software technology, providing Genesis with the best possible performance, usability, and scalability. The flexibility, platform independency, well designed user interface, and the variety of analysis and data visualizations tools will provide Genesis with the potential to become a valuable tool for the pursuit of future biological discoveries.

In summary, the software suite described here facilitates the integration of gene expression data with genome sequence information in order to (1) provide a more complete picture of the transcriptomic behavior of a cell and the varieties that distinguish us from other species and make us human and (2) enable comparative analyses of human diseases and corresponding mouse models.

Finally and ultimately these investigations attempt to provide the research community with a markedly improved repertoire of database query and accompanying computational tools that facilitate the translation of accumulated information from comparative transcriptomic studies into novel biological insights.

# Acknowledgments

Major parts of this work were supported by the Austrian Science Foundation, SFB project Biomembranes (F718) and the GEN-AU: BIN, Bioinformatics Integration Network. I would like to express my deepest gratitude to my mentor Zlatko Trajanoski for his encouragement, visions, and believing in me. Many thanks also go to Dr. Frank Eisenhaber for rendering an expert opinion on my thesis. I want to express my appreciation to my colleagues and friends Michael Maurer and Robert Molidor for their assiduousness and fervor in developing MARS with me and for the many, many fruitful discussions. Sincere thanks go to all present and past members of the Institute for Genomics and Bioinformatics for valuable comments, support, and their friendship. A special acknowledgment is dedicated to the people, that have contributed to this work: Hubert Hackl for realizing the MEF gene expression profiling and for explaining and discussing umpteen issues I did not know or understand due to my limited biological knowledge, Marcel Scheideler for realizing the hMADS study, Fatima Sanchez Cabo for helping me with the mathematics, and last but definitely not least Gernot Stocker for the development of the Java Cluster Service, the perfect management of our computing facility, and for granting me access to his extensive IT knowledge. Without the genius of all these people, this work would not have been possible. I am indebted to my parents for their unfailing support and my companion in life Susanne for accompanying me, her love, encouragement, and understanding.

# Bibliography

## Introduction

### Background

[01] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature. 2001 Feb 15;409(6822):860-921.Erratum in: Nature 2001 Aug 2;412(6846):565. Nature 2001 Jun 7;411(6838):720.

[02] Venter JC et al. The sequence of the human genome. Science. 2001 Feb 16;291(5507):1304-51.Erratum in: Science 2001 Jun 5;292(5523):1838.

### Comparative Genomics

[03] Tilghman SM. Lessons learned, promises kept: a biologist's eye view of the Genome Project. Genome Res. 1996 Sep;6(9):773-80.

[04] Boguski MS. Comparative genomics: the mouse that roared. Nature. 2002 Dec 5;420(6915):515-6.

[05] Zuckerkandl E, Pauling L. Molecules as documents of evolutionary history. J Theor Biol. 1965 Mar;8(2):357-66.

[06] Duret L, Bucher P. Searching for regulatory elements in human noncoding sequences. Curr Opin Struct Biol. 1997 Jun;7(3):399-406.

[07] Hardison RC, Oeltjen J, Miller W. Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. Genome Res. 1997 Oct;7(10):959-66.

[08] Hardison RC. Conserved noncoding sequences are reliable guides to regulatory elements. Trends Genet. 2000 Sep;16(9):369-72.

[09] Loots GG, Locksley RM, Blankespoor CM, Wang ZE, Miller W, Rubin EM, Frazer KA. Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. Science. 2000 Apr 7;288(5463):136-40.

[10] Pennacchio LA, Rubin EM. Genomic strategies to identify mammalian regulatory sequences. Nat Rev Genet. 2001 Feb;2(2):100-9.

[11] Gottgens B, Barton LM, Chapman MA, Sinclair AM, Knudsen B, Grafham D, Gilbert JG, Rogers J, Bentley DR, Green AR. Transcriptional regulation of the stem cell leukemia gene (SCL) - comparative analysis of five vertebrate SCL loci. Genome Res. 2002 May;12(5):749-59.

[12] Boffelli D, Nobrega MA, Rubin EM. Comparative genomics at the vertebrate extremes. Nat Rev Genet. 2004 Jun;5(6):456-65.

[13] Pennacchio LA, Olivier M, Hubacek JA, Cohen JC, Cox DR, Fruchart JC, Krauss RM, Rubin EM. An apolipoprotein influencing triglycerides in humans and mice revealed by comparative sequencing. Science. 2001 Oct 5;294(5540):169-73.

[14] Hedges SB, Kumar S. Genomics. Vertebrate genomes compared. Science. 2002 Aug 23;297 (5585): 1283-5.

[15] Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. Nature. 2002 Dec 5;420(6915):520-62.

[16] Bradley A. Mining the mouse genome. Nature. 2002 Dec 5;420(6915):512-4.

[17] Tautz D. Evolution of transcriptional regulation. Curr Opin Genet Dev. 2000 Oct;10(5):575-9.

[18] Sands AT. The master mammal. Nat Biotechnol. 2003 Jan;21(1):31-2.

[19] Nobrega MA, Pennacchio LA. Comparative genomic analysis as a tool for biological discovery. J Physiol. 2004 Jan 1;554(Pt 1):31-9. Review.

[20] Makalowski W, Boguski MS. Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. Proc Natl Acad Sci U S A. 1998 Aug 4;95(16):9407-12.

[21] Wheelan SJ, Boguski MS, Duret L, Makalowski W. Human and nematode orthologs--lessons from the analysis of 1800 human genes and the proteome of Caenorhabditis elegans. Gene. 1999 Sep 30;238(1):163-70.

[22] Rat Genome Sequencing Project Consortium. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature. 2004 Apr 1;428(6982):493-521.

[23] Jacob HJ, Kwitek AE. Rat genetics: attaching physiology and pharmacology to the genome. Nat Rev Genet. 2002 Jan;3(1):33-42.

## Homology and Homology Subsets

[24] Fitch WM. Homology a personal view on some of the problems. Trends Genet. 2000 May;16(5):227-31.

[25] Fitch WM, Upper K. The phylogeny of tRNA sequences provides evidence for ambiguity reduction in the origin of the genetic code. Cold Spring Harb Symp Quant Biol. 1987;52:759-67.

[26] Fitch WM. Distinguishing homologous from analogous proteins. Syst Zool. 1970 Jun;19(2):99-113.

[27] Gray GS, Fitch WM. Evolution of antibiotic resistance genes: the DNA sequence of a kanamycin resistance gene from Staphylococcus aureus. Mol Biol Evol. 1983 Dec;1(1):57-66.

[28] Sonnhammer EL, Koonin EV. Orthology, paralogy and proposed classification for paralog subtypes. Trends Genet. 2002 Dec;18(12):619-20.

[29] Theissen G. Secret life of genes. Nature. 2002 Feb 14;415(6873):741.

[30] Koonin EV. An apology for orthologs - or brave new memes. Genome Biol. 2001;2(4):COMMENT1005. Epub 2001 Apr 06.

[31] Jensen RA. Orthologs and paralogs - we need to get it right. Genome Biol. 2001;2(8):INTERACTIONS1002. Epub 2001 Aug 03.

## Transcriptome Analysis (Microarrays)

[32] Fodor SP, Read JL, Pirrung MC, Stryer L, Lu AT, Solas D. Light-directed, spatially addressable parallel chemical synthesis. Science. 1991 Feb 15;251(4995):767-73.

[33] Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science. 1995 Oct 20;270(5235):467-70.

[34] DeRisi JL, Iyer VR, Brown PO. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science. 1997 Oct 24;278(5338):680-6.

[35] Lockhart DJ, Winzeler EA. Genomics, gene expression and DNA arrays. Nature. 2000 Jun 15;405(6788):827-36. Review.

[36] Young RA. Biomedical discovery with DNA arrays. Cell. 2000 Jul 7;102(1):9-15. Review.

[37] Zhang MQ. Large-scale gene expression data analysis: a new challenge to computational biologists. Genome Res. 1999 Aug;9(8):681-8. Review.

[38] Southern EM. Detection of specific sequences among DNA fragments separated by gel electrophoresis. J Mol Biol. 1975 Nov 5;98(3):503-17.

[39] Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD, Kidd MJ, King AM, Meyer MR, Slade D, Lum PY, Stepaniants SB, Shoemaker DD, Gachotte D, Chakraburtty K, Simon J, Bard M, Friend SH. Functional discovery via a compendium of expression profiles. Cell. 2000 Jul 7;102(1):109-26.

[40] Collins FS. Microarrays and macroconsequences. Nat Genet. 1999 Jan;21(1 Suppl):2.

[41] Brazma A, Robinson A, Cameron G, Ashburner M. One-stop shop for microarray data. Nature. 2000 Feb 17;403(6771):699-700.

[42] Duggan DJ, Bittner M, Chen Y, Meltzer P, Trent JM. Expression profiling using cDNA microarrays. Nat Genet. 1999 Jan;21(1 Suppl):10-4.

[43] Ermolaeva O, Rastogi M, Pruitt KD, Schuler GD, Bittner ML, Chen Y, Simon R, Meltzer P, Trent JM, Boguski MS. Data management and analysis for gene expression arrays. Nat Genet. 1998 Sep;20(1):19-23.

[44] Bassett DE Jr, Eisen MB, Boguski MS. Gene expression informatics--it's all in your mine. Nat Genet. 1999 Jan;21(1 Suppl):51-5.

### Comparative Transcriptomics

[45] Quackenbush J. Genomics. Microarrays--guilt by association. Science. 2003 Oct 10;302(5643):240-1.

[46] Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A. 1998 Dec 8;95(25):14863-8.

[47] Kim SK, Lund J, Kiraly M, Duke K, Jiang M, Stuart JM, Eizinger A, Wylie BN, Davidson GS. A gene expression map for Caenorhabditis elegans. Science. 2001 Sep 14;293(5537):2087-92.

[48] Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet. 2003 Jun;34(2):166-76.

[49] Alter O, Brown PO, Botstein D. Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. Proc Natl Acad Sci U S A. 2003 Mar 18;100(6):3351-6. Epub 2003 Mar 11.

[50] van Noort V, Snel B, Huynen MA. Predicting gene function by conserved co-expression. Trends Genet. 2003 May;19(5):238-42.

[51] Teichmann SA, Babu MM. Conservation of gene co-regulation in prokaryotes and eukaryotes. Trends Biotechnol. 2002 Oct;20(10):407-10.

## Methods

[52] Baxevainis AD, Ouellette BFF. Bioinformatics – A Practical Guide to the Analysis of Genes and Proteins. Second Edition. Wiley-Interscience. 2001.

[53] Baxevanis AD. The Molecular Biology Database Collection: an updated compilation of biological database resources. Nucleic Acids Res. 29: 1-10 (2001).

[54] Baxevanis AD. The Molecular Biology Database Collection: 2003 update. Nucleic Acids Res. 31: 1-12 (2003).

[55] Galperin MY. The Molecular Biology Database Collection: 2004 update. Nucleic Acids Res. 2004 Jan 1;32 Database issue:D3-22.

## GenBank

[56] URL: http://www.ncbi.nlm.nih.gov/Genbank/

[57] Mizrachi I. GenBank: The Nucleotide Sequence Database. In: The NCBI Handbook [http://ncbi.nlm.nih.gov/entrez/query.fcgi?db=Books]. National Library of Medicine (US), National Center for Biotechnology Information. 2003 Aug.

[58] Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank. Nucleic Acids Res. 2003 Jan 1;31(1):23-7.

[59] Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank: update. Nucleic Acids Res. 2004 Jan 1;32 Database issue:D23-6.

[60] URL: http://www.ncbi.nih.gov/

[61] URL: http://www.nlm.nih.gov/

[62] URL: http://www.nih.gov/

[63] URL: http://www.ebi.ac.uk/embl/

[64] Stoesser G, Baker W, van den Broek A, Garcia-Pastor M, Kanz C, Kulikova T, Leinonen R, Lin Q, Lombard V, Lopez R, Mancuso R, Nardone F, Stoehr P, Tuli MA, Tzouvara K, Vaughan R. The EMBL Nucleotide Sequence Database: major new developments. Nucleic Acids Res. 2003 Jan 1;31(1):17-22.

[65] Kanz C, Aldebert P, Althorpe N, Baker W, Baldwin A, Bates K, Browne P, van den Broek A, Castro M, Cochrane G, Duggan K, Eberhardt R, Faruque N, Gamble J, Diez FG, Harte N, Kulikova T, Lin Q, Lombard V, Lopez R, Mancuso R, McHale M, Nardone F, Silventoinen V, Sobhany S, Stoehr P, Tuli MA, Tzouvara K, Vaughan R, Wu D, Zhu W, Apweiler R. The EMBL Nucleotide Sequence Database. Nucleic Acids Res. 2005 Jan 1;33 Database Issue:D29-33.

[66] URL: http://www.ddbj.nig.ac.jp

[67] Tateno Y, Imanishi T, Miyazaki S, Fukami-Kobayashi K, Saitou N, Sugawara H, Gojobori T. DNA Data Bank of Japan (DDBJ) for genome scale research in life science. Nucleic Acids Res. 2002 Jan 1;30(1):27-30.

[68] URL: http://www.ncbi.nlm.nih.gov/projects/collab/

[69] URL: http://www.ncbi.nlm.nih.gov/collab/FT/

## Refseq

[70] URL: http://www.ncbi.nlm.nih.gov/RefSeq/

[71] Pruitt K, Tatusova T, Ostell J. The Reference Sequence (RefSeq) Project. In: The NCBI Handbook [http://ncbi.nlm.nih.gov/entrez/query.fcgi?db=Books]. National Library of Medicine (US), National Center for Biotechnology Information. 2003 Aug.

[72] Pruitt KD, Tatusova T, Maglott DR. NCBI Reference Sequence project: update and current status. Nucleic Acids Res. 2003 Jan 1;31(1):34-7.

[73] Pruitt KD, Maglott DR. RefSeq and LocusLink: NCBI gene-centered resources. Nucleic Acids Res. 2001 Jan 1;29(1):137-40.

[74] Pruitt KD, Katz KS, Sicotte H, Maglott DR. Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. Trends Genet. 2000 Jan;16(1):44-7.

## Ensembl

[75] URL: http://www.ensembl.org

[76] Birney E, Andrews TD, Bevan P, Caccamo M, Chen Y, Clarke L, Coates G, Cuff J, Curwen V, Cutts T, Down T, Eyras E, Fernandez-Suarez XM, Gane P, Gibbins B, Gilbert J, Hammond M, Hotz HR, Iyer V, Jekosch K, Kahari A, Kasprzyk A, Keefe D, Keenan S, Lehvaslaiho H, McVicker G, Melsopp C, Meidl P, Mongin E, Pettett R, Potter S, Proctor G, Rae M, Searle S, Slater G, Smedley D, Smith J, Spooner W, Stabenau A, Stalker J, Storey R, Ureta-Vidal A, Woodwark KC, Cameron G, Durbin R, Cox A, Hubbard T, Clamp M. An overview of Ensembl. Genome Res. 2004 May;14(5):925-8. Epub 2004 Apr 12. Review.

[77] Birney E, Andrews D, Bevan P, Caccamo M, Cameron G, Chen Y, Clarke L, Coates G, Cox T, Cuff J, Curwen V, Cutts T, Down T, Durbin R, Eyras E, Fernandez-Suarez XM, Gane P, Gibbins B, Gilbert J, Hammond M, Hotz H, Iyer V, Kahari A, Jekosch K, Kasprzyk A, Keefe D, Keenan S, Lehvaslaiho H, McVicker G, Melsopp C, Meidl P, Mongin E, Pettett R, Potter S, Proctor G, Rae M, Searle S, Slater G, Smedley D, Smith J, Spooner W, Stabenau A, Stalker J, Storey R, Ureta-Vidal A, Woodwark C, Clamp M, Hubbard T. Ensembl 2004. Nucleic Acids Res. 2004 Jan 1;32 Database issue:D468-70.

[78] Clamp M, Andrews D, Barker D, Bevan P, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T, Durbin R, Eyras E, Gilbert J, Hammond M, Hubbard T, Kasprzyk A, Keefe D, Lehvaslaiho H, Iyer V, Melsopp C, Mongin E, Pettett R, Potter S, Rust A, Schmidt E, Searle S, Slater G, Smith J, Spooner W, Stabenau A, Stalker J, Stupka E, Ureta-Vidal A, Vastrik I, Birney E. Ensembl 2002: accommodating comparative genomics. Nucleic Acids Res. 2003 Jan 1;31(1):38-42.

[79] Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T, Durbin R, Eyras E, Gilbert J, Hammond M, Huminiecki L, Kasprzyk A, Lehvaslaiho H, Lijnzaad P, Melsopp C, Mongin E, Pettett R, Pocock M, Potter S, Rust A, Schmidt E, Searle S, Slater G, Smith J, Spooner W, Stabenau A, Stalker J, Stupka E, Ureta-Vidal A, Vastrik I, Clamp M. The Ensembl genome database project. Nucleic Acids Res. 2002 Jan 1;30(1):38-41.

## trEST

[80] URL: http://hits.isb-sib.ch/

[81] Sperisen P, Iseli C, Pagni M, Stevenson BJ, Bucher P, Jongeneel CV. trome, trEST and trGEN: databases of predicted protein sequences. Nucleic Acids Res. 2004 Jan 1;32 Database issue:D509-11.

[82] Pagni M, Iseli C, Junier T, Falquet L, Jongeneel V, Bucher P. trEST, trGEN and Hits: access to databases of predicted protein sequences. Nucleic Acids Res. 2001 Jan 1;29(1):148-51.

[83] Huang X, Madan A. CAP3: A DNA sequence assembly program. Genome Res. 1999 Sep;9(9):868-77.

[84] Iseli C, Jongeneel CV, Bucher P. ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. Proc Int Conf Intell Syst Mol Biol. 1999:138-48.

[85] Burge CB, Karlin S. Finding the genes in genomic DNA. Curr Opin Struct Biol. 1998 Jun;8(3):346-54.

## UniGene

[86] URL: http://www.ncbi.nlm.nih.gov/UniGene

[87] Pontius JU, Wagner L, Schuler GD. UniGene: a unified view of the transcriptome. In: The NCBI Handbook [http://ncbi.nlm.nih.gov/entrez/query.fcgi?db=Books]. National Library of Medicine (US), National Center for Biotechnology Information. 2003 Aug.

[88] Wheeler DL, Church DM, Federhen S, Lash AE, Madden TL, Pontius JU, Schuler GD, Schriml LM, Sequeira E, Tatusova TA, Wagner L. Database resources of the National Center for Biotechnology. Nucleic Acids Res. 2003 Jan 1;31(1):28-33.

[89] Schuler GD. Pieces of the puzzle: expressed sequence tags and the catalog of human genes. J Mol Med. 1997 Oct;75(10):694-8. Review.

[90] Schuler GD, Boguski MS, Stewart EA, Stein LD, Gyapay G, Rice K, White RE, Rodriguez-Tome P, Aggarwal A, Bajorek E, Bentolila S, Birren BB, Butler A, Castle AB, Chiannilkulchai N, Chu A, Clee C, Cowles S, Day PJ, Dibling T, Drouot N, Dunham I, Duprat S, East C, Hudson TJ, et al. A gene map of the human genome. Science. 1996 Oct 25;274(5287):540-6. Review.

[91] Boguski MS, Schuler GD. ESTablishing a human transcript map. Nat Genet. 1995 Aug;10(4):369-71.

## Fantom II

[92] Quackenbush J. Viva la revolution! A report from the FANTOM meeting. Nat Genet. 2000 Nov;26(3):255-6.

[93] The FANTOM Consortium and the RIKEN Genome Exploration Research Group Phase I & II Team. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature. 2002 Dec 5;420(6915):563-73.

[94] The RIKEN Genome Exploration Research Group Phase II Team and the FANTOM Consortium. Functional annotation of a full-length mouse cDNA collection. Nature. 2001 Feb 8;409(6821):685-90.

[95] Hayashizaki Y. RIKEN mouse genome encyclopedia. Mech Ageing Dev. 2003 Jan;124(1):93-102.

[96] URL: http://fantom.gsc.riken.go.jp/

[97] Bono H, Kasukawa T, Furuno M, Hayashizaki Y, Okazaki Y. FANTOM DB: database of Functional Annotation of RIKEN Mouse cDNA Clones. Nucleic Acids Res. 2002 Jan 1;30(1):116-8.

## International Protein Index (IPI)

[98] URL: http://www.ebi.ac.uk/IPI/

[99] Kersey PJ, Duarte J, Williams A, Karavidopoulou Y, Birney E, Apweiler R. The International Protein Index: an integrated database for proteomics experiments. Proteomics. 2004 Jul;4(7):1985-8.

[100] URL: http://www.expasy.org/sprot/

[101] URL: http://www.ebi.ac.uk/swissprot/

[102] Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. 2003 Jan 1;31(1):365-70.

[103] O'Donovan C, Martin MJ, Gattiker A, Gasteiger E, Bairoch A, Apweiler R. High-quality protein knowledge resource: SWISS-PROT and TrEMBL. Brief Bioinform. 2002 Sep;3(3):275-84.

[104] URL: http://www.ebi.ac.uk/trembl/

## Entrez

[105] URL: http://www.ncbi.nlm.nih.gov/Entrez/

[106] Ostell J. The Entrez Search and Retrieval System. In: The NCBI Handbook [http://ncbi.nlm.nih.gov/entrez/query.fcgi?db=Books]. National Library of Medicine (US), National Center for Biotechnology Information. 2003 Aug.

[107] Schuler GD, Epstein JA, Ohkawa H, Kans JA. Entrez: molecular biology database and retrieval system. Methods Enzymol. 1996;266:141-62.

[108] Tatusova TA, Karsch-Mizrachi I, Ostell JA. Complete genomes in WWW Entrez: data representation and analysis. Bioinformatics. 1999 Jul-Aug;15(7-8):536-43.

[109] Chen J, Anderson JB, DeWeese-Scott C, Fedorova ND, Geer LY, He S, Hurwitz DI, Jackson JD, Jacobs AR, Lanczycki CJ, Liebert CA, Liu C, Madej T, Marchler-Bauer A, Marchler GH, Mazumder R, Nikolskaya AN, Rao BS, Panchenko AR, Shoemaker BA, Simonyan V, Song JS, Thiessen PA, Vasudevan S, Wang Y, Yamashita RA, Yin JJ, Bryant SH. MMDB: Entrez's 3D-structure database. Nucleic Acids Res. 2003 Jan 1;31(1):474-7.

## Sequence Retrieval System (SRS)

[110] URL: http://srs.ebi.ac.uk

[111] Etzold T, Ulyanov A, Argos P. SRS: information retrieval system for molecular biology data banks. Methods Enzymol. 1996;266:114-28.

[112] Zdobnov EM, Lopez R, Apweiler R, Etzold T. The EBI SRS server-new features. Bioinformatics. 2002 Aug;18(8):1149-50.

[113] Zdobnov EM, Lopez R, Apweiler R, Etzold T. The EBI SRS server--recent developments. Bioinformatics. 2002 Feb;18(2):368-73.

[114] URL: http://www.lionbioscience.com/

[115] Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A. 1988 Apr;85(8):2444-8.

[116] Pearson WR. Rapid and sensitive sequence comparison with FASTP and FASTA. Methods Enzymol. 1990;183:63-98.

[117] Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994 Nov 11;22(22):4673-80.

## RepeatMasker

[118] Smit AFA, Green P. RepeatMasker at http://repeatmasker.org. Unpublished.

[119] URL: http://www.phrap.org/

[120] Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol. 1981 Mar 25;147(1):195-7.

[121] Gotoh O. An improved algorithm for matching biological sequences. J Mol Biol. 1982 Dec 15;162(3):705-8.

## Blast

[122] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990 Oct 5;215(3):403-10.

[123] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997 Sep 1;25(17):3389-402. Review.

[124] Karlin S, Altschul SF. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc Natl Acad Sci U S A. 1990 Mar;87(6):2264-8.

[125] Gish W, States DJ. Identification of protein coding regions by database similarity search. Nat Genet. 1993 Mar;3(3):266-72.

[126] Zhang Z, Schwartz S, Wagner L, Miller W. A greedy algorithm for aligning DNA sequences. J Comput Biol. 2000 Feb-Apr;7(1-2):203-14.

## Java Cluster Service (JCS)

[127] Buyya R. High Performance Cluster Computing: Architectures and Systems (Vol. 1 & 2). Prentice Hall, NJ, USA. 1999.

[128] URL: http://java.sun.com

[129] URL: http://www.jboss.org

[130] URL: http://jakarta.apache.org/tomcat/

[131] URL: http://www.openpbs.org/

## Mouse Embryo Fibroblasts (MEFs)

[132] Lukas J, Bartkova J, Rohde M, Strauss M, Bartek J: Cyclin D1 is dispensable for G1 control in retinoblastoma gene-deficient cells independently of cdk4 activity. Mol Cell Biol 1995, 15: 2600-2611.

[133] Hansen JB, Petersen RK, Larsen BM, Bartkova J, Alsner J, Kristiansen K: Activation of peroxisome proliferator-activated receptor gamma bypasses the function of the retinoblastoma protein in adipocyte differentiation. J Biol Chem 1999, 274: 2386-2393.

[134] Chomczynski P, Sacchi N: Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction Anal Biochem 1987, 162: 156-159.

[135] Hegde P, Qi R, Abernathy K, Gay C, Dharap S, Gaspard R et al.: A concise guide to cDNA microarray analysis. Biotechniques 2000, 29: 548-556.

[136] Kerr MK, Martin M, Churchill GA: Analysis of variance for gene expression microarray data. J Comput Biol 2000, 7: 819-837.

[137] Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J et al.: Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res 2002, 30: e15.

[138] Quackenbush J: Microarray data normalization and transformation. Nat Genet 2002, 32 Suppl: 496-501.

[139] Pieler R, Sanchez-Cabo F, Hackl H, Thallinger GG, Trajanoski Z: ArrayNorm: comprehensive normalization and analysis of microarray data. Bioinformatics 2004.

## Human Multipotent Adipose-derived Stem Cells (hMADS)

[140] Rodriguez A.M., Elabd C., Delteil F., Astier J., Vernochet C., Saint-Marc P., Guesnet J., Guezennec A., Amri E.Z., Dani C., Ailhaud G. 2004. Adipocyte differentiation of multipotent cells established from human adipose tissue. Biochemical and Biophysical Research Communications 315: 255–263.

[141] Kane M.D., Jatkoe T.A., Stumpf C.R., Lu J., Thomas J.D., Madore S.J. 2000. Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. Nucleic Acids Res. 28:4552-4557.

[142] URL: http://www.mwg-biotech.com

[143] Hackl H., Cabo F.S., Sturn A., Wolkenhauer O., Trajanoski Z. 2004. Analysis of DNA microarray data. Curr Top Med Chem. 4:1357-70.

## MARS

[144] URL: https://mars.genome.tugraz.at

[145] Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet. 2001 Dec;29(4):365-71.

[146] Stoeckert C J, Jr., Causton H C, Ball C A . Microarray databases: standards and ontologies . Nat Genet. 32 Suppl: 469-473 (2002).

[147] Ball CA, Brazma A, Causton H, Chervitz S, Edgar R, Hingamp P, Matese JC, Parkinson H, Quackenbush J, Ringwald M, Sansone SA, Sherlock G, Spellman P, Stoeckert C, Tateno Y, Taylor R, White J, Winegarden N. Submission of microarray data to public repositories. PLoS Biol. 2004 Sep;2(9):E317. Epub 2004 Aug 31.

[148] URL: http://www.mged.org

[149] URL: http://www.mged.org/Workgroups/MIAME/

[150] URL: http:// http://www.w3c.org/XML/

[151] URL: http://www.oracle.com

[152] URL: http://www.postgresql.org

[153] Spellman PT, Miller M, Stewart J, Troup C, Sarkans U, Chervitz S, Bernhart D, Sherlock G, Ball C, Lepage M, Swiatek M, Marks WL, Goncalves J, Markel S, Iordan D, Shojatalab M, Pizarro A, White J, Hubley R, Deutsch E, Senger M, Aronow BJ, Robinson A, Bassett D, Stoeckert CJ Jr, Brazma A. Design and implementation of microarray gene expression markup language (MAGE-ML). Genome Biol. 2002 Aug 23;3(9):RESEARCH0046. Epub 2002 Aug 23.

[154] Quackenbush J. Data standards for 'omic' science. Nat Biotechnol. 22: 613-614 (2004).

[155] Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, Holloway E, Kapushesky M, Kemmeren P, Lara GG, Oezcimen A, Rocca-Serra P, Sansone SA. ArrayExpress--a public repository for microarray gene expression data at the EBI. Nucleic Acids Res. 2003 Jan 1;31(1):68-71.

## Hierarchical Clustering (HCL)

[156] Gedina G. Praktische Methodenlehre Hintergrund-Material, lecture notes for "Praktische Methodenlehre", WS 1999/2000, University of Osnabrueck, Germany.

[157] Fisher D. Optimization and Simplification of Hierarchical Clusterings, KDD-95:118-123.

[158] Wen X, Fuhrman S, Michaels GS, Carr DB, Smith S, Barker JL, Somogyi R. Large-scale temporal gene expression mapping of central nervous system development. Proc Natl Acad Sci U S A. 1998 Jan 6;95(1):334-9.

## Self Organizing Maps (SOM)

[159] Vesanto J. Usisng SOM in Data Mining. Licentiate's thesis. Helsinki University of Technology, Department of Computer Science and Engineering. 2000 Apr 10.

[160] Vesanto J. SOM-Based Data Visualization Methods. Helsinki University of Technology, Laboratory of Computer and Information Science. 1999 Nov 19.

[161] Vesanto J. Data Mining Techniques Based on the Self-Organizing Map. Thesis for degree of Master of Science in Engineering. Helsinki University of Technology, Department of Engineering Physics and Mathematics. 1997 May 26.

[162] Kohonen T, Hynninen J, Kangas J, Laaksonen J. SOM_PAK. The Self-Organizing Map Program Package. Manual Version 3.1. Helsinki University of Technology. Laboratory of Computer and Information Science. 1995 April 7.

[163] Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc Natl Acad Sci U S A. 1999 Mar 16;96(6):2907-12.

## k-means Clustering (KMC)

[164] Zhexue Huang. A fast clustering algorithm to cluster very large categorical data sets in data mining. In Proc. SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, 1997.

[165] Zhexue Huang. Clustering large data sets with mixed numerical and categorical values. Proceedings of the First Pacific-Asia Conference on Knowledge Discovery and Data Mining, Singapore, Word Scientific, 1997.

[166] Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. Systematic determination of genetic network architecture. Nat Genet. 1999 Jul;22(3):281-5.

[167] Soukas A, Cohen P, Socci ND, Friedman JM. Leptin-specific patterns of gene expression in white adipose tissue. Genes Dev. 2000 Apr 15;14(8):963-80.

## Figure of Merit (FOM)

[168] Yeung KY, Haynor DR, Ruzzo WL. Validating clustering for gene expression data. Bioinformatics. 2001 Apr;17(4):309-18.

## Principal Component Analysis (PCA)

[169] Basilevsky A. Statistical factor analysis and related methods, Theory and Applications. Wiley series in probability and mathematical statistics. John Wiley & Sons.

[170] Raychaudhuri S, Stuart JM, Altman RB. Principal components analysis to summarize microarray experiments: application to sporulation time series. Pac Symp Biocomput. 2000;455-66.

[171] Holter NS, Mitra M, Maritan A, Cieplak M, Banavar JR, Fedoroff NV. Fundamental patterns underlying gene expression profiles: simplicity from complexity. Proc Natl Acad Sci U S A. 2000 Jul 18;97(15):8409-14.

[172] Alter O, Brown PO, Botstein D. Singular value decomposition for genome-wide expression data processing and modeling. Proc Natl Acad Sci U S A. 2000 Aug 29;97(18):10101-6.

## Correspondence Analysis (CA)

[173] Fellenberg K, Hauser NC, Brors B, Neutzner A, Hoheisel JD, Vingron M. Correspondence analysis applied to microarray data. Proc Natl Acad Sci U S A. 2001 Sep 11;98(19):10781-6. Epub 2001 Sep 04.

## One-Way-ANOVA

[174] Zar J.H. Biostatistical Analysis. 4th ed. Prentice Hall, NJ. 1999.

## Gene Expression Terrain Maps

[175] Stuart JM, Segal E, Koller D, Kim SK. A gene-coexpression network for global discovery of conserved genetic modules. Science. 2003 Oct 10;302(5643):249-55. Epub 2003 Aug 21.

[176] Werner-Washburne M, Wylie B, Boyack K, Fuge E, Galbraith J, Weber J, Davidson G. Comparative analysis of multiple genome-scale data sets. Genome Res. 2002 Oct;12(10):1564-73.

[177] Davidson, G.S., Wylie, B.N. & Boyack, K.W. Cluster stability and the use of noise in interpretation of clustering. Proceedings of the IEEE Symposium on Information Visualization 2001 (INFOVIS'01). 2001.

## Gene Ontology (GO)

[178] Gruber TR. A translation approach to portable ontology specifications. Knowledge Acquisition, Vol.5, No.2, 1993, pp. 199-220.

[179] Gruber TR. Towards principles for the design of ontologies used for knowledge sharing. International Journal of Human-Computer Studies, 43, pp 907-928, 1995.

[180] Gene Ontology Consortium. Gene ontology: tool for the unification of biology. Nat Genet. 2000 May;25(1):25-9.

[181] Gene Ontology Consortium. Creating the gene ontology resource: design and implementation. Genome Res. 2001 Aug;11(8):1425-33.

[182] Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. 2004 Jan 1;32 Database issue:D258-61.

[183] URL: http://www.geneontology.org/

[184] URL: http://www.mysql.com/

## PromoSer

[185] Novina CD, Roy AL. Core promoters and transcriptional control. Trends Genet. 1996 Sep;12(9):351-5.

[186] Pilpel Y, Sudarsanam P, Church GM. Identifying regulatory networks by combinatorial analysis of promoter elements. Nat Genet. 2001 Oct;29(2):153-9.

[187] McKnight SL, Kingsbury R. Transcriptional control signals of a eukaryotic protein-coding gene. Science. 1982 Jul 23;217(4557):316-24.

[188] Mitchell PJ, Tjian R. Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. Science. 1989 Jul 28;245(4916):371-8.

[189] URL: http://biowulf.bu.edu/zlab/PromoSer/

[190] Halees AS, Leyfer D, Weng Z. PromoSer: A large-scale mammalian promoter and transcription start site identification service. Nucleic Acids Res. 31: 3554-3559 (2003).

[191] Halees AS, Weng Z. PromoSer: improvements to the algorithm, visualization and accessibility. Nucleic Acids Res. 32: W191-W194 (2004).

[192] Kent WJ. BLAT - the BLAST-like alignment tool. Genome Res. 2002 Apr;12(4):656-64.

[193] URL: http://biowulf.bu.edu/zlab/promoser/promoser.wsdl

[194] Stein L. Creating a bioinformatics nation. Nature. 2002 May 9;417(6885):119-20.

### Distribution of all octamer DNA Sequences

[195] FitzGerald PC, Shlyakhtenko A, Mir AA, Vinson C. Clustering of DNA sequences in human promoters. Genome Res. 14: 1562-1574 (2004).

### Wise

[196] URL: http://www.ebi.ac.uk/Wise2/dbaform.html

[197] Birney E, Clamp M, Durbin R. GeneWise and Genomewise. Genome Res. 2004 May;14(5):988-95.

[198] URL: http://www.ebi.ac.uk/Wise2/promoterwise.html

## Results

### Overview

[199] Sturn A, Quackenbush J, Trajanoski Z . Genesis: cluster analysis of microarray data . Bioinformatics. 18: 207-208 (2002).

[200] Sturn A, Mlecnik B, Pieler R, Rainer J, Truskaller T, Trajanoski Z. Client-Server environment for high-performance gene expression data analysis. Bioinformatics. 19: 772-773 (2003).

[201] URL: http://genome.tugraz.at/Software/

[202] URL: http://java.sun.com/products/java-media/3D/

### Sequence Retrieval

[203] URL: http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html

### Comparative Transcriptomics Study

[204] URL: http://www.tigr.org

[205] URL: http://www.biocarta.com/

[206] Eberhardt C, Gray PW, Tjoelker LW. Human lysophosphatidic acid acyltransferase. cDNA cloning, expression, and localization to chromosome 9q34.3. J Biol Chem. 1997 Aug 8;272(32):20299-305.

[207] Yamauchi T, Kamon J, Ito Y, Tsuchida A, Yokomizo T, Kita S, Sugiyama T, Miyagishi M, Hara K, Tsunoda M, Murakami K, Ohteki T, Uchida S, Takekawa S, Waki H, Tsuno NH, Shibata Y, Terauchi Y, Froguel P, Tobe K, Koyasu S, Taira K, Kitamura T, Shimizu T, Nagai R, Kadowaki T. Cloning of adiponectin receptors that mediate antidiabetic metabolic effects. Nature. 2003 Jun 12;423(6941):762-9.

[208] Kim JB, Spiegelman BM. ADD1/SREBP1 promotes adipocyte differentiation and gene expression linked to fatty acid metabolism. Genes Dev. 1996 May 1;10(9):1096-107.

[209] Kim JB, Wright HM, Wright M, Spiegelman BM. ADD1/SREBP1 activates PPARgamma through the production of endogenous ligand. Proc Natl Acad Sci U S A. 1998 Apr 14;95(8):4333-7.

[210] Gregoire FM, Smas CM, Sul HS. Understanding adipocyte differentiation. Physiol Rev. 1998 Jul;78(3):783-809.

[211] Diehn M, Sherlock G, Binkley G, Jin H, Matese JC, Hernandez-Boussard T, Rees CA,Cherry JM, Botstein D, Brown PO, Alizadeh AA. SOURCE: a unified genomic resource of functional annotations, ontologies, andgene expression data. Nucleic Acids Res. 31: 219-223 (2003).

## Promoter Analysis

[212] Hapgood JP, Riedemann J, Scherer SD. Regulation of gene expression by GC-rich DNA cis-elements. Cell Biol Int. 2001;25(1):17-31.

[213] URL: http://www.gene-regulation.com/

[214] Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DU, Land S, Lewicki-Potapov B, Michael H, Munch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S, Wingender E. TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res. 2003 Jan 1;31(1):374-8.

[215] URL: http://jaspar.cgb.ki.se

[216] Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D91-4.

[217] Kim Y, Geiger JH, Hahn S, Sigler PB. Crystal structure of a yeast TBP/TATA-box complex. Nature. 1993 Oct 7;365(6446):512-20.

[218] Geiger JH, Hahn S, Lee S, Sigler PB. Crystal structure of the yeast TFIIA/TBP/DNA complex. Science. 1996 May 10;272(5263):830-6.

[219] Suzuki Y, Yamashita R, Shirota M, Sakakibara Y, Chiba J, Mizushima-Sugano J, Nakai K, Sugano S. Sequence comparison of human and mouse genes reveals a homologous block structure in the promoter regions. Genome Res. 2004 Sep;14(9):1711-8.

[220] Dynan WS, Tjian R. Control of eukaryotic messenger RNA synthesis by sequence-specific DNA-binding proteins. Nature. 1985 Aug 29-Sep 4;316(6031):774-8.

[221] Shao D, Lazar M A . Peroxisome proliferator activated receptor gamma, CCAAT/enhancer- binding protein alpha, and cell cycle status regulate the commitment to adipocyte differentiation . J Biol Chem. 272: 21473-21478 (1997).

[222] Mantovani R. The molecular biology of the CCAAT-binding factor NF-Y. Gene. 1999 Oct 18;239(1):15-27. Review.

[223] Romier C, Cocchiarella F, Mantovani R, Moras D. The NF-YB/NF-YC structure gives insight into DNA binding and transcription regulation by CCAAT factor NF-Y. J Biol Chem. 2003 Jan 10;278(2):1336-45. Epub 2002 Oct 24.

[224] Cao Z, Umek RM, McKnight SL. Regulated expression of three C/EBP isoforms during adipose conversion of 3T3-L1 cells. Genes Dev. 1991 Sep;5(9):1538-52.

[225] Yeh WC, Cao Z, Classon M, McKnight SL. Cascade regulation of terminal adipocyte differentiation by three members of the C/EBP family of leucine zipper proteins. Genes Dev. 1995 Jan 15;9(2):168-81.

[226] Darlington GJ, Ross SE, MacDougald OA. The role of C/EBP genes in adipocyte differentiation. J Biol Chem. 1998 Nov 13;273(46):30057-60. Review.

[227] Rosen ED, Walkey CJ, Puigserver P, Spiegelman BM. Transcriptional regulation of adipogenesis. Genes Dev. 2000 Jun 1;14(11):1293-307. Review.

## Discussion

[228] URL: http://www.ncbi.nlm.nih.gov/COG/

[229] Tatusov RL, Galperin MY, Natale DA, Koonin EV. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. 2000 Jan 1;28(1):33-6.

[230] Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV. The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res. 2001 Jan 1;29(1):22-8.

[231] Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA. The COG database: an updated version includes eukaryotes. BMC Bioinformatics. 2003 Sep 11;4(1):41.

[232] URL: http://www.tigr.org/tdb/tgi/ego/

[233] Lee Y, Sultana R, Pertea G, Cho J, Karamycheva S, Tsai J, Parvizi B, Cheung F, Antonescu V, White J, Holt I, Liang F, Quackenbush J. Cross-referencing eukaryotic genomes: TIGR Orthologous Gene Alignments (TOGA). Genome Res. 2002 Mar;12(3):493-502.

[234] Remm M, Storm CE, Sonnhammer EL. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J Mol Biol. 2001 Dec 14;314(5):1041-52.

[235] Yuan YP, Eulenstein O, Vingron M, Bork P. Towards detection of orthologues in sequence databases. Bioinformatics. 1998;14(3):285-9.

[236] URL: http://www.opengl.org/

[237] URL: http://www.microsoft.com/windows/directx/

[238] URL: http://www.ebi.ac.uk/GOA/

[239] Camon E, Magrane M, Barrell D, Lee V, Dimmer E, Maslen J, Binns D, Harte N, Lopez R, Apweiler R. The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. Nucleic Acids Res. 2004 Jan 1;32 Database issue:D262-6.

[240] Wang W, Cherry JM, Botstein D, Li H. A systematic approach to reconstructing transcription networks in Saccharomycescerevisiae. Proc Natl Acad Sci U S A. 2002 Dec 24;99(26):16893-8. Epub 2002 Dec 13.

[241] Whitfield ML, Sherlock G, Saldanha AJ, Murray JI, Ball CA, Alexander KE, Matese JC, Perou CM, Hurt MM, Brown PO, Botstein D. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. Mol Biol Cell. 2002 Jun;13(6):1977-2000.

[242] Lee JS, Chu IS, Mikaelyan A, Calvisi DF, Heo J, Reddy JK, Thorgeirsson SS. Application of comparative functional genomics to identify best-fit mouse models to study human cancer. Nat Genet. 2004 Dec;36(12):1306-11.

[243] Kho AT, Zhao Q, Cai Z, Butte AJ, Kim JY, Pomeroy SL, Rowitch DH, Kohane IS. Conserved mechanisms across development and tumorigenesis revealed by a mouse development perspective of human cancers. Genes Dev. 2004 Mar 15;18(6):629-40.

[244] McCarroll SA, Murphy CT, Zou S, Pletcher SD, Chin CS, Jan YN, Kenyon C, Bargmann CI, Li H. Comparing genomic expression patterns across species identifies shared transcriptional profile in aging. Nat Genet. 2004 Feb;36(2):197-204. Epub 2004 Jan 18.

[245] Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002 Jan 1;30(1):207-10.

# A Supplementary Information

## A.1 Protein Finding Pipeline

```
<?xml version="1.0" encoding="UTF-8"?>
<BLAST-Project version="1.0">
  <InputFile>HumanOligos.fasta.masked.filtered</InputFile>
  <InputDirectory>Z:/PhD/Data/HumanRepeatMasked</InputDirectory>
  <OutputDirectory>Z:/PhD/Data/HumanRepeatMasked</OutputDirectory>
  <IndexingDirectory>Z:/PhD/Data/Indexing</IndexingDirectory>
  <LogDirectory>Z:/PhD/Data/HumanRepeatMasked/Logs</LogDirectory>
  <ProteinFilesDirectory>Z:/PhD/Data/Databases</ProteinFilesDirectory>
  <dtdDirectory>D:/Java/marsCGM</dtdDirectory>
  <QuerySequencesIndex>HumanRepeatMaskedQuerySequences.index</QuerySequencesIndex>
  <QuerySequencesFastaParserClass>at.tugraz.genome.genesis.fasta.FastaParserAll</QuerySequencesFastaParserClass>
  <JobStorageFile>HumanRepeatMaskedJobs.dat</JobStorageFile>
  <NumberOfJobs>48</NumberOfJobs>
  <MaxNumberOfPartsToProcess>1000</MaxNumberOfPartsToProcess>
  <UpdateInterval>1000</UpdateInterval>
  <HostName>mcluster.tu-graz.ac.at</HostName>
  <UserName>xxxxxxxxx</UserName>
  <Password>xxxxxxxxx</Password>
  <RemoteBlastDatabaseDirectory>/netapp/BioInfo/sturn/blast</RemoteBlastDatabaseDirectory>
  <RemoteInputDirectory>/netapp/BioInfo/sturn/blast/Input</RemoteInputDirectory>
  <RemoteResultDirectory>/home/jcluster/J2EE/pbs/jobs</RemoteResultDirectory>
  <Task index="0" name="HumanRefSeq">
     <DatabaseType>REFSEQ</DatabaseType>
     <BlastType>NCBI-MEGABLAST</BlastType>
     <BlastParameters>-e "1e-9" -F "F" -W "11" -t "18" -p "90" -v "5" -b "5" -m "7" -D "2" -a "1"</BlastParameters>
  </Task>
  <Task index="1" name="HumanEnsembl">
     <DatabaseType>ENSEMBL</DatabaseType>
     <BlastType>NCBI-MEGABLAST</BlastType>
     <BlastParameters>-e "1e-9" -F "F" -W "11" -t "18" -p "90" -v "5" -b "5" -m "7" -D "2" -a "1"</BlastParameters>
  </Task>
  <Task index="2" name="HumanUnigene">
     <DatabaseType>UNIGENE</DatabaseType>
     <Organism>Homo sapiens</Organism>
     <BlastType>NCBI-MEGABLAST</BlastType>
     <BlastParameters>-e "1e-9" -F "F" -W "11" -t "18" -p "90" -v "5" -b "5" -m "7" -D "2" -a "1"</BlastParameters>
  </Task>
  <Task index="3" name="HumanTrest">
     <DatabaseType>TREST</DatabaseType>
     <BlastType>NCBI-MEGABLAST</BlastType>
     <BlastParameters>-e "1e-9" -F "F" -W "11" -t "18" -p "90" -v "5" -b "5" -m "7" -D "2" -a "1"</BlastParameters>
  </Task>
  <Task index="4" name="HumanIPI">
     <DatabaseType>IPI</DatabaseType>
     <BlastType>NCBI-BLAST</BlastType>
     <BlastParameters>-p "blastx" -e "1e-9" -F "F" -v "5" -b "5" -a "1" -m "7"</BlastParameters>
  </Task>
  <Task index="5" name="HumanEntrez">
     <DatabaseType>ENTREZ</DatabaseType>
     <BlastType>NCBI-BLAST</BlastType>
     <BlastParameters>-p "blastx" -e "1e-9" -F "F" -v "5" -b "5" -a "1" -m "7"</BlastParameters>
  </Task>
  <Task index="6" name="HumanRefSeqBlastn">
     <DatabaseType>REFSEQ</DatabaseType>
     <DatabaseToUse>HumanRefSeq</DatabaseToUse>
     <BlastType>NCBI-BLAST</BlastType>
     <BlastParameters>-p "blastn" -e "1e-9" -F "F" -v "5" -b "5" -a "1" -m "7"</BlastParameters>
  </Task>
  <Task index="7" name="HumanEnsemblBlastn">
     <DatabaseType>ENSEMBL</DatabaseType>
     <DatabaseToUse>HumanEnsembl</DatabaseToUse>
     <BlastType>NCBI-BLAST</BlastType>
     <BlastParameters>-p "blastn" -e "1e-9" -F "F" -v "5" -b "5" -a "1" -m "7"</BlastParameters>
  </Task>
  <Task index="8" name="HumanUnigeneBlastn">
     <DatabaseType>UNIGENE</DatabaseType>
     <DatabaseToUse>HumanUnigene</DatabaseToUse>
     <Organism>Homo sapiens</Organism>
     <BlastType>NCBI-BLAST</BlastType>
     <BlastParameters>-p "blastn" -e "1e-9" -F "F" -v "5" -b "5" -a "1" -m "7"</BlastParameters>
  </Task>
  <Task index="9" name="HumanTrestBlastn">
     <DatabaseType>TREST</DatabaseType>
     <DatabaseToUse>HumanTrest</DatabaseToUse>
     <BlastType>NCBI-BLAST</BlastType>
     <BlastParameters>-p "blastn" -e "1e-9" -F "F" -v "5" -b "5" -a "1" -m "7"</BlastParameters>
  </Task>
</BLAST-Project>
```

**Listing A.1: Human Oligo Chip XML Definition File** for the Protein-Finding-Pipeline.

```
<?xml version="1.0" encoding="UTF-8"?>
<BLAST-Project version="1.0">
  <InputFile>ESTs.txt.masked.filtered</InputFile>
  <InputDirectory>Z:/PhD/Data/MouseRepeatMasked</InputDirectory>
  <OutputDirectory>Z:/PhD/Data/MouseRepeatMasked</OutputDirectory>
  <IndexingDirectory>Z:/PhD/Data/Indexing</IndexingDirectory>
  <LogDirectory>Z:/PhD/Data/MouseRepeatMasked/Logs</LogDirectory>
  <ProteinFilesDirectory>Z:/PhD/Data/Databases</ProteinFilesDirectory>
  <dtdDirectory>D:/Java/marsCGM</dtdDirectory>
  <QuerySequencesIndex>MouseRepeatMaskedQuerySequences.index</QuerySequencesIndex>
  <QuerySequencesFastaParserClass>at.tugraz.genome.genesis.fasta.FastaParserGenBank</QuerySequencesFastaParserClass>
  <JobStorageFile>MouseRepeatMaskedJobs.dat</JobStorageFile>
  <NumberOfJobs>48</NumberOfJobs>
  <MaxNumberOfPartsToProcess>1000</MaxNumberOfPartsToProcess>
  <UpdateInterval>1000</UpdateInterval>
  <HostName>mcluster.tu-graz.ac.at</HostName>
  <UserName>xxxxxxxxx</UserName>
  <Password>xxxxxxxxx</Password>
  <RemoteBlastDatabaseDirectory>/netapp/BioInfo/sturn/blast</RemoteBlastDatabaseDirectory>
  <RemoteInputDirectory>/netapp/BioInfo/sturn/blast/Input</RemoteInputDirectory>
  <RemoteResultDirectory>/home/jcluster/J2EE/pbs/jobs</RemoteResultDirectory>
  <Task index="0" name="MouseRefSeq">
     <DatabaseType>REFSEQ</DatabaseType>
     <BlastType>NCBI-MEGABLAST</BlastType>
     <BlastParameters>-e "1e-9" -F "F" -W "11" -t "18" -p "90" -v "5" -b "5" -m "7" -D "2" -a "1"</BlastParameters>
  </Task>
  <Task index="1" name="MouseFantom2">
     <DatabaseType>PHANTOM2</DatabaseType>
     <BlastType>NCBI-MEGABLAST</BlastType>
     <BlastParameters>-e "1e-9" -F "F" -W "11" -t "18" -p "90" -v "5" -b "5" -m "7" -D "2" -a "1"</BlastParameters>
  </Task>
  <Task index="2" name="MouseEnsembl">
     <DatabaseType>ENSEMBL</DatabaseType>
     <BlastType>NCBI-MEGABLAST</BlastType>
     <BlastParameters>-e "1e-9" -F "F" -W "11" -t "18" -p "90" -v "5" -b "5" -m "7" -D "2" -a "1"</BlastParameters>
  </Task>
  <Task index="3" name="MouseUnigene">
     <DatabaseType>UNIGENE</DatabaseType>
     <Organism>Mus musculus</Organism>
     <BlastType>NCBI-MEGABLAST</BlastType>
     <BlastParameters>-e "1e-9" -F "F" -W "11" -t "18" -p "90" -v "5" -b "5" -m "7" -D "2" -a "1"</BlastParameters>
  </Task>
  <Task index="4" name="MouseTrest">
     <DatabaseType>TREST</DatabaseType>
     <BlastType>NCBI-MEGABLAST</BlastType>
     <BlastParameters>-e "1e-9" -F "F" -W "11" -t "18" -p "90" -v "5" -b "5" -m "7" -D "2" -a "1"</BlastParameters>
  </Task>
  <Task index="5" name="MouseIPI">
     <DatabaseType>IPI</DatabaseType>
     <BlastType>NCBI-BLAST</BlastType>
     <BlastParameters>-p "blastx" -e "1e-9" -F "F" -v "5" -b "5" -a "1" -m "7"</BlastParameters>
  </Task>
  <Task index="6" name="MouseEntrez">
     <DatabaseType>ENTREZ</DatabaseType>
     <BlastType>NCBI-BLAST</BlastType>
     <BlastParameters>-p "blastx" -e "1e-9" -F "F" -v "5" -b "5" -a "1" -m "7"</BlastParameters>
  </Task>
  <Task index="7" name="MouseRefSeqBlastn">
     <DatabaseType>REFSEQ</DatabaseType>
     <DatabaseToUse>MouseRefSeq</DatabaseToUse>
     <BlastType>NCBI-BLAST</BlastType>
     <BlastParameters>-p "blastn" -e "1e-9" -F "F" -v "5" -b "5" -a "1" -m "7"</BlastParameters>
  </Task>
  <Task index="8" name="MouseFantom2Blastn">
     <DatabaseType>PHANTOM2</DatabaseType>
     <DatabaseToUse>MouseFantom2</DatabaseToUse>
     <BlastType>NCBI-BLAST</BlastType>
     <BlastParameters>-p "blastn" -e "1e-9" -F "F" -v "5" -b "5" -a "1" -m "7"</BlastParameters>
  </Task>
  <Task index="9" name="MouseEnsemblBlastn">
     <DatabaseType>ENSEMBL</DatabaseType>
     <DatabaseToUse>MouseEnsembl</DatabaseToUse>
     <BlastType>NCBI-BLAST</BlastType>
     <BlastParameters>-p "blastn" -e "1e-9" -F "F" -v "5" -b "5" -a "1" -m "7"</BlastParameters>
  </Task>
  <Task index="10" name="MouseUnigeneBlastn">
     <DatabaseType>UNIGENE</DatabaseType>
     <DatabaseToUse>MouseUnigene</DatabaseToUse>
     <Organism>Mus musculus</Organism>
     <BlastType>NCBI-BLAST</BlastType>
     <BlastParameters>-p "blastn" -e "1e-9" -F "F" -v "5" -b "5" -a "1" -m "7"</BlastParameters>
  </Task>
  <Task index="11" name="MouseTrestBlastn">
     <DatabaseType>TREST</DatabaseType>
     <DatabaseToUse>MouseTrest</DatabaseToUse>
     <BlastType>NCBI-BLAST</BlastType>
     <BlastParameters>-p "blastn" -e "1e-9" -F "F" -v "5" -b "5" -a "1" -m "7"</BlastParameters>
  </Task>
</BLAST-Project>
```

**Listing A.2: Mouse EST Chip XML Definition File** for the Protein-Finding-Pipeline.

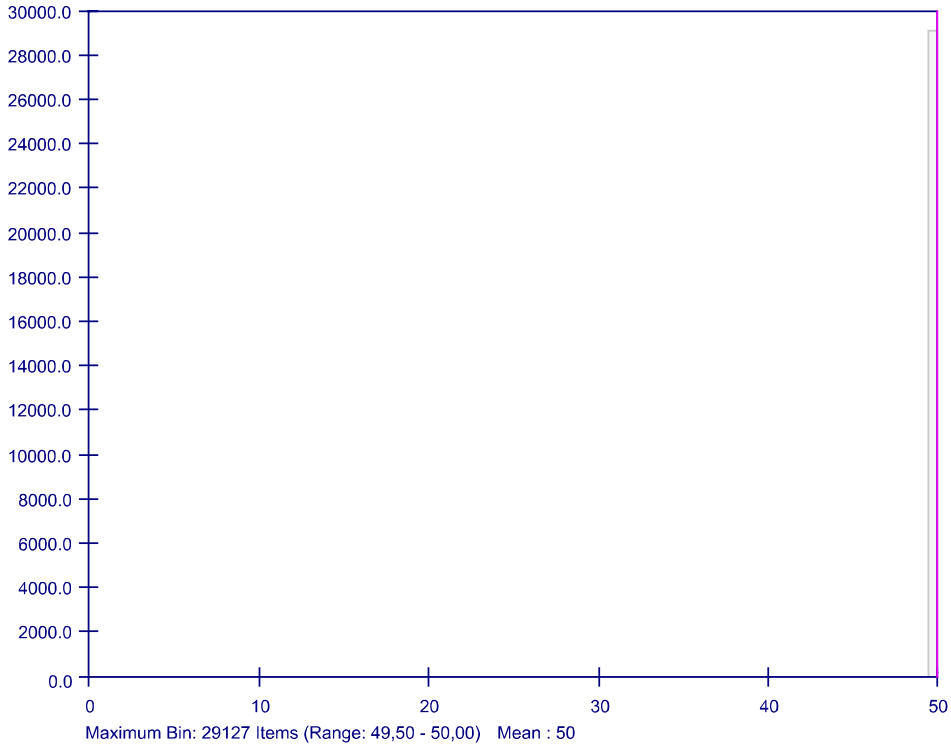



**Figure A.1: Histogram of human oligo lengths** (29.127 sequences, 1.456.350 nucleotides, minimum length: 50, maximum length: 50, average length: 50, due to strict filtering of min length: 50).

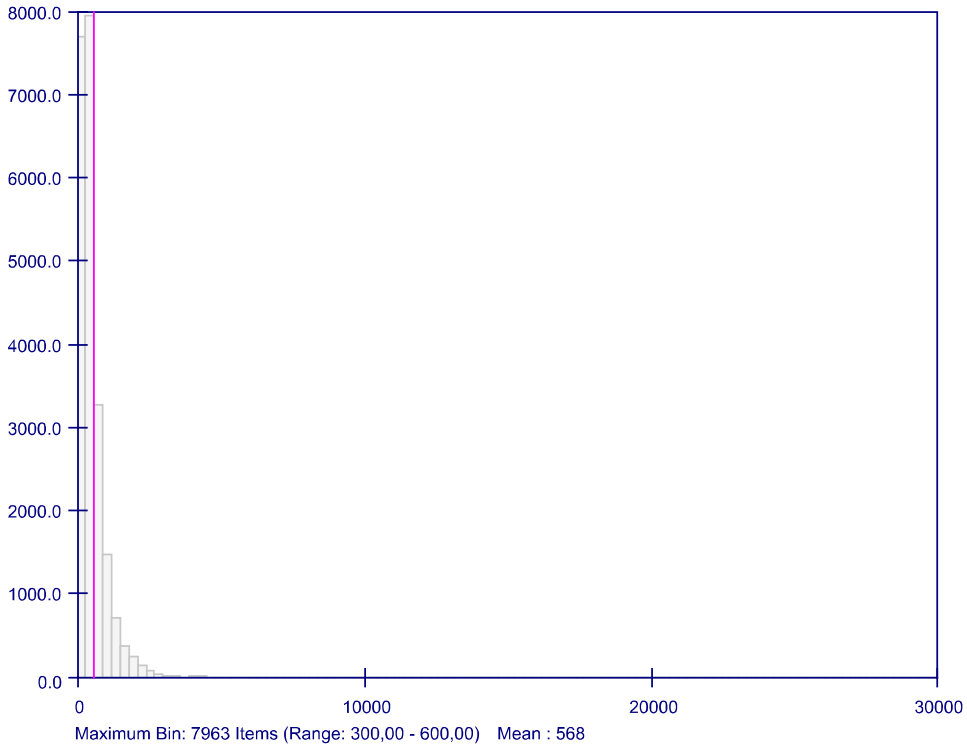
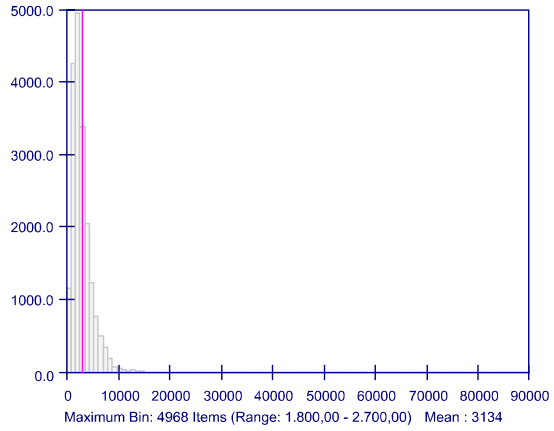
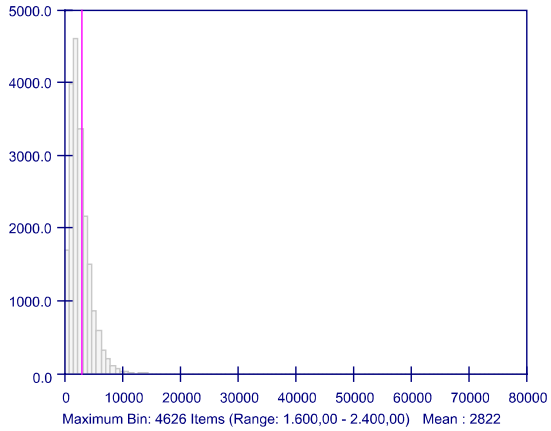


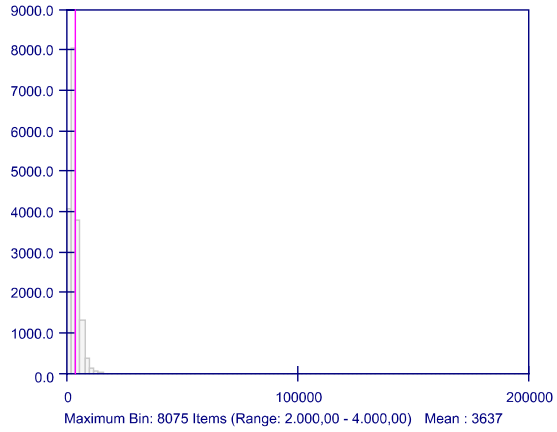

**Figure A.2: Histogram of human protein sequence lengths** (22.400 sequences, 12.726.498 peptides, minimum length: 7, maximum length: 26.218, average length: 568).



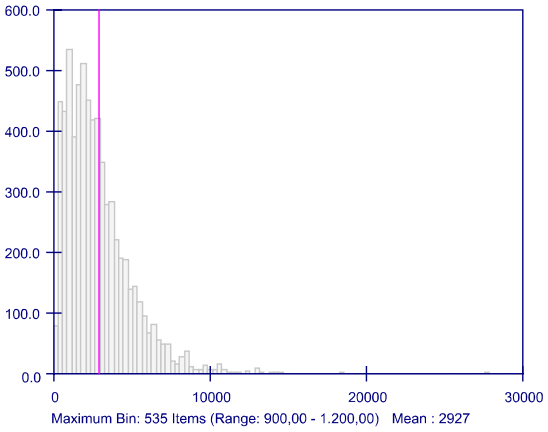

RefSeq: 19.375 sequences, min length=213, max length=82.603



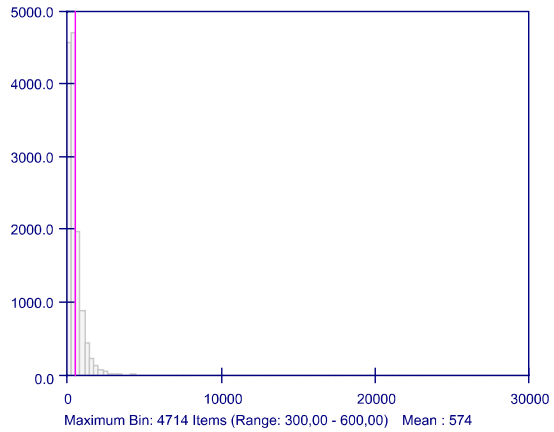

Ensembl: 19.934 sequences, min length=63, max length=79.818



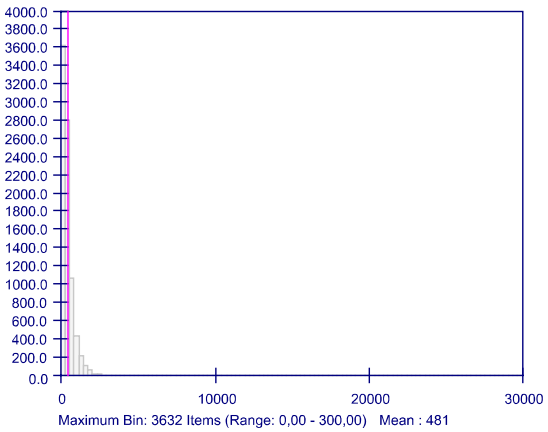

UniGene: 18.120 sequences, min length=17, max. =101.518




trEST: 6.750 sequences, min length=104, max length=27.675




IPI: 13.365 sequences, min length=18, max length=26.926




Entrez: 8.493 sequences, min length=16, max length=26.926

Table A.1 Histograms of human BLAST hit length.

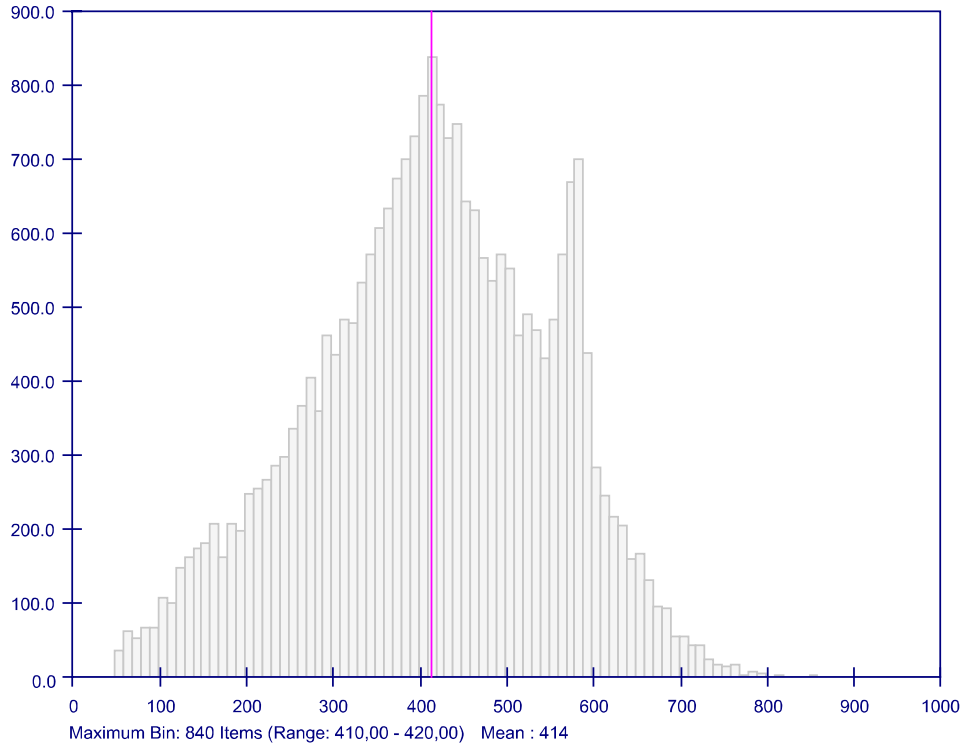



**Figure A.3: Histogram of mouse EST lengths** (25.134 sequences, 10.401.712 nucleotides, minimum length: 50, maximum length: 987, average length: 414).

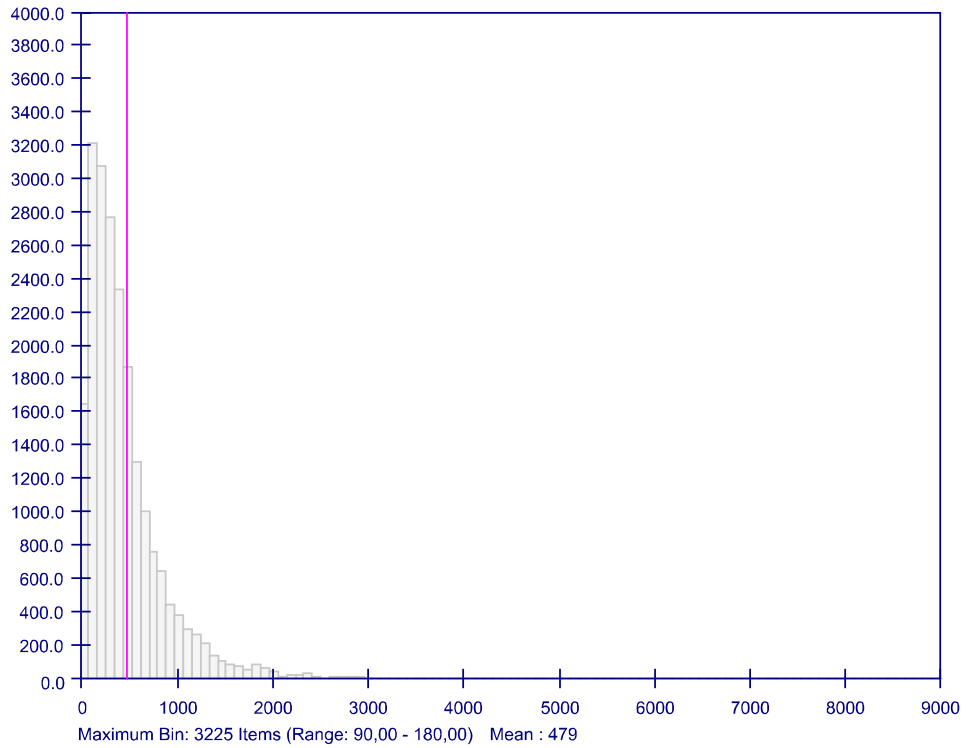
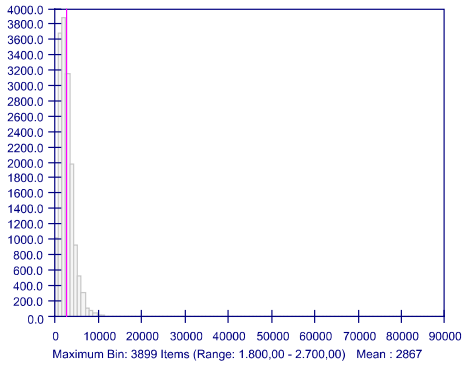
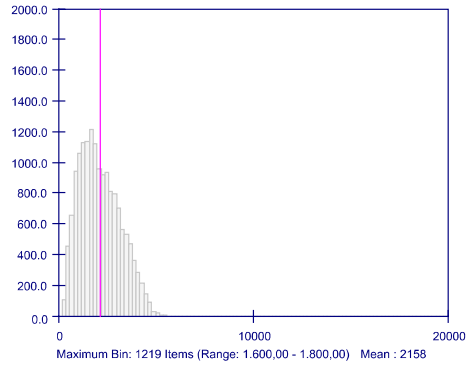


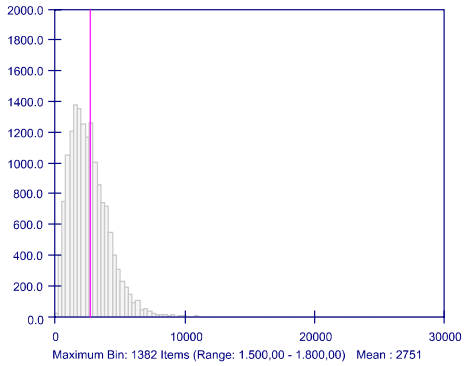

**Figure A.4: Histogram of mouse protein sequence lengths** (21.257 sequences, 10.176.115 peptides, minimum length: 13, maximum length: 8350, average length: 479).



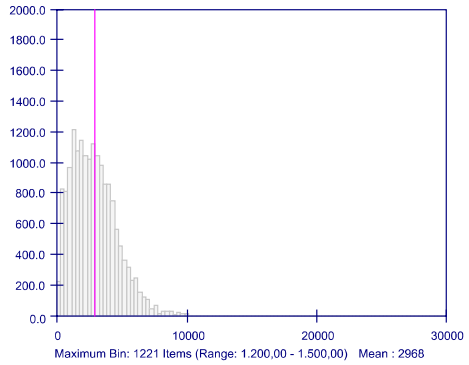

RefSeq: 15.916 sequences, min length=261, max length=84.275



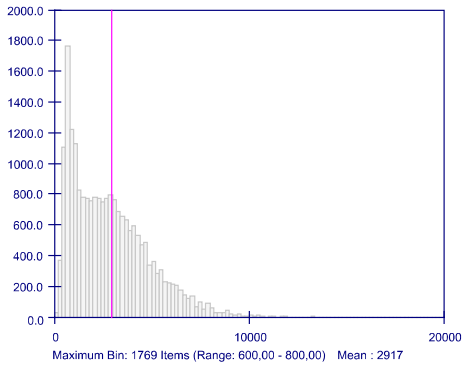

Fantom2: 15.866 sequences, min length=84, max length=10.481



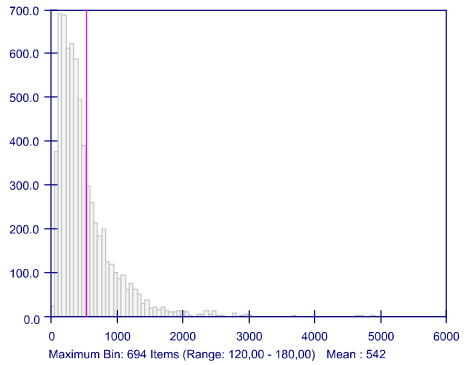

Ensembl: 15.670 sequences, min length=78, max length=23.218



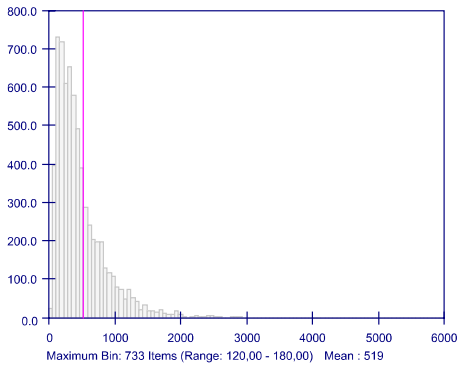

UniGene: 17.047 sequences, min length=144, max. =25.839




trEST: 21.839 sequences, min length=47, max length=18.297




IPI: 6.828 sequences, min length=29, max length=5.321




Entrez: 6.810 sequences, min length=29, max length=5.327

Table A.2 Histograms of mouse BLAST hit length.

# A.2 Comparative Genomics Pipeline

```
<?xml version="1.0" encoding="UTF-8"?>
<BLAST-Project version="1.0">
  <OrganismName1>Human</OrganismName1>
  <OrganismName2>Mouse</OrganismName2>
  <InputFile1>Z:/PhD/Data/HumanRepeatMasked/HumanRepeatMaskedProteins.fasta</InputFile1>
  <InputFile2>Z:/PhD/Data/MouseRepeatMasked/MouseRepeatMaskedProteins.fasta</InputFile2>
  <InputDirectory>Z:/PhD/Data/HumanMouseRepeatMasked</InputDirectory>
  <OutputDirectory>Z:/PhD/Data/HumanMouseRepeatMasked</OutputDirectory>
  <IndexingDirectory>Z:/PhD/Data/Indexing</IndexingDirectory>
  <LogDirectory>Z:/PhD/Data/HumanMouseRepeatMasked/Logs</LogDirectory>
  <dtdDirectory>D:/Java/marsCGM</dtdDirectory>
  <PromoserDirectory>Z:/PhD/Data/Databases/Promoser</PromoserDirectory>
  <QuerySequencesIndex>HumanMouseRepeatMaskedQuerySequences.index</QuerySequencesIndex>
  <QuerySequencesFastaParserClass1>at.tugraz.genome.genesis.fasta.FastaParserAll</QuerySequencesFastaParserClass1>
  <QuerySequencesFastaParserClass2>at.tugraz.genome.genesis.fasta.FastaParserAll</QuerySequencesFastaParserClass2>
  <JobStorageFile>HumanMouseRepeatMaskedJobs.dat</JobStorageFile>
  <NumberOfJobs>48</NumberOfJobs>
  <MaxNumberOfPartsToProcess>1000</MaxNumberOfPartsToProcess>
  <UpdateInterval>1000</UpdateInterval>
  <HostName>mcluster.tu-graz.ac.at</HostName>
  <UserName>xxxxxxxxx</UserName>
  <Password>xxxxxxxxx</Password>
  <RemoteBlastDatabaseDirectory>/netapp/BioInfo/sturn/blast</RemoteBlastDatabaseDirectory>
  <RemoteBlastExecutableDirectory>/usr/local/bioinf/ncbi-blast</RemoteBlastExecutableDirectory>
  <RemoteInputDirectory>/netapp/BioInfo/sturn/blast/Input</RemoteInputDirectory>
  <RemoteResultDirectory>/netapp/BioInfo/sturn/blast/Output</RemoteResultDirectory>
  <BlastParameters>-p "blastp" -e "1e-9" -F "m S" -a "1" -m "7"</BlastParameters>
</BLAST-Project>
```

**Listing A.3: Comparative-Genomics-Pipeline XML Definition File** for the human-mouse comparison.

# A.3 GO Annotation Pipeline

```
<?xml version="1.0" encoding="UTF-8"?>
<BLAST-Project version="1.0">
  <OrganismName>Human</OrganismName>
  <InputFile>Z:/PhD/Data/HumanRepeatMasked/HumanRepeatMaskedProteins.fasta</InputFile>
  <GOSequencesFile>Z:/PhD/Data/HumanMouseRepeatMasked/GO/GO-Sequences.fasta</GOSequencesFile>
  <InputDirectory>Z:/PhD/Data/HumanMouseRepeatMasked</InputDirectory>
  <OutputDirectory>Z:/PhD/Data/HumanMouseRepeatMasked/GO</OutputDirectory>
  <IndexingDirectory>Z:/PhD/Data/Indexing</IndexingDirectory>
  <LogDirectory>Z:/PhD/Data/HumanMouseRepeatMasked/GO/Logs</LogDirectory>
  <dtdDirectory>D:/Java/marsCGM</dtdDirectory>
  <QuerySequencesIndex>HumanRepeatMaskedGOQuerySequences.index</QuerySequencesIndex>
  <QuerySequencesFastaParserClass>at.tugraz.genome.genesis.fasta.FastaParserAll</QuerySequencesFastaParserClass>
  <JobStorageFile>HumanMouseRepeatMaskedGOJobs.dat</JobStorageFile>
  <NumberOfJobs>48</NumberOfJobs>
  <MaxNumberOfPartsToProcess>1000</MaxNumberOfPartsToProcess>
  <UpdateInterval>10000</UpdateInterval>
  <HostName>mcluster.tu-graz.ac.at</HostName>
  <UserName>xxxxxxxxx</UserName>
  <Password>xxxxxxxxx</Password>
  <RemoteBlastDatabaseDirectory>/netapp/BioInfo/sturn/blast</RemoteBlastDatabaseDirectory>
  <RemoteBlastExecutableDirectory>/usr/local/bioinf/ncbi-blast</RemoteBlastExecutableDirectory>
  <RemoteInputDirectory>/netapp/BioInfo/sturn/blast/Input</RemoteInputDirectory>
  <RemoteResultDirectory>/netapp/BioInfo/sturn/blast/Output</RemoteResultDirectory>
  <BlastParameters>-p "blastp" -e "1e-9" -F "m S" -v "20" -b "20" -a "1" -m "7"</BlastParameters>
</BLAST-Project>
```

**Listing A.4: GO-Annotation-Pipeline XML Definition File** for the human oligo chip GO annotation.

```
<?xml version="1.0" encoding="UTF-8"?>
<BLAST-Project version="1.0">
  <OrganismName>house mouse</OrganismName>
  <InputFile>Z:/PhD/Data/MouseRepeatMasked/MouseRepeatMaskedProteins.fasta</InputFile>
  <GOSequencesFile>Z:/PhD/Data/HumanMouseRepeatMasked/GO/GO-Sequences.fasta</GOSequencesFile>
  <InputDirectory>Z:/PhD/Data/MouseRepeatMasked</InputDirectory>
  <OutputDirectory>Z:/PhD/Data/MouseRepeatMasked/GO</OutputDirectory>
  <IndexingDirectory>Z:/PhD/Data/Indexing</IndexingDirectory>
  <LogDirectory>Z:/PhD/Data/MouseRepeatMasked/GO/Logs</LogDirectory>
  <dtdDirectory>D:/Java/marsCGM</dtdDirectory>
  <QuerySequencesIndex>MouseRepeatMaskedGOQuerySequences.index</QuerySequencesIndex>
  <QuerySequencesFastaParserClass>at.tugraz.genome.genesis.fasta.FastaParserAll</QuerySequencesFastaParserClass>
  <JobStorageFile>MouseRepeatMaskedGOJobs.dat</JobStorageFile>
  <NumberOfJobs>48</NumberOfJobs>
  <MaxNumberOfPartsToProcess>1000</MaxNumberOfPartsToProcess>
  <UpdateInterval>10000</UpdateInterval>
  <HostName>mcluster.tu-graz.ac.at</HostName>
  <UserName>xxxxxxxxx</UserName>
  <Password>xxxxxxxxx</Password>
  <RemoteBlastDatabaseDirectory>/netapp/BioInfo/sturn/blast</RemoteBlastDatabaseDirectory>
  <RemoteBlastExecutableDirectory>/usr/local/bioinf/ncbi-blast</RemoteBlastExecutableDirectory>
  <RemoteInputDirectory>/netapp/BioInfo/sturn/blast/Input</RemoteInputDirectory>
  <RemoteResultDirectory>/netapp/BioInfo/sturn/blast/Output</RemoteResultDirectory>
  <BlastParameters>-p "blastp" -e "1e-9" -F "m S" -v "20" -b "20" -a "1" -m "7"</BlastParameters>
</BLAST-Project>
```

**Listing A.5: GO-Annotation-Pipeline XML Definition File** for the mouse EST chip GO annotation.

# B Publications

## Papers

Maurer M, Molidor R, **Sturn A**, Hackl H, Hartler J, Stocker G, Prokesch A, Scheideler M, Trajanoski Z. MARS: A Microarray Analysis, Retrieval, and Storage System. BMC Bioinformatics (submitted).

Hackl H, Sanchez Cabo F, **Sturn A**, Wolkenhauer O, Trajanoski Z. Analysis of DNA Microarray Data. Curr Top Med Chem. 2004, 4(13):1355-1368.

Molidor R, **Sturn A**, Maurer M, Trajanoski Z. New trends in bioinformatics: from genome sequence to personalized medicine. Exp Gerontol. 2003 Oct;38(10):1031-6.

**Sturn A**, Mlecnik B, Pieler R, Rainer J, Truskaller T, Trajanoski Z. Client-Server Environment for High-Performance Gene Expression Data Analysis. Bioinformatics. 2003 Apr;19(6):772-773.

Saeed AI, Sharov V, White J, Li J, Liang W, Bhagabati N, Braisted J, Klapa M, Currier T, Thiagarajan M, **Sturn A**, Snuffin M, Rezantsev A, Popov D, Ryltsov A, Kostukovich E, Borisovsky I, Liu Z, Vinsavich A, Trush V, Quackenbush J. TM4: a free, open-source system for microarray data management and analysis. Biotechniques. 2003 Feb;34(2):374-8.

**Sturn A**, Quackenbush J, Trajanoski Z. Genesis: Cluster analysis of microarray data. Bioinformatics. 2002 Jan;18(1):207-8. PMID: 11836235

## Book Chapters

**Sturn A**, Maurer M, Molidor R, Trajanoski Z. Systems for Management of Pharmacogenomic Information. Pharmacogenomics: Methods and Protocols. Humana Press, Totowa, USA 2004, in press.